



Titre: Prise de décision liée à la conception et aux opérations en milieu industriel complexe par l'étude des données en temps réel
Title: industrial complex by the study of real time data

Auteur: Émilie Thibault
Author:

Date: 2023

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: Thibault, É. (2023). Prise de décision liée à la conception et aux opérations en milieu industriel complexe par l'étude des données en temps réel [Ph.D. thesis, Polytechnique Montréal]. PolyPublie. <https://publications.polymtl.ca/53340/>
Citation:

 **Document en libre accès dans PolyPublie**
Open Access document in PolyPublie

URL de PolyPublie: <https://publications.polymtl.ca/53340/>
PolyPublie URL:

Directeurs de recherche: Paul R. Stuart, & Milan Korbel
Advisors:

Programme: Génie chimique
Program:

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

**Prise de décision liée à la conception et aux opérations en milieu industriel
complexe par l'étude des données en temps réel**

ÉMILIE THIBAUT

Département de génie chimique

Thèse présentée en vue de l'obtention du diplôme de *Philosophiae Doctor*

Génie chimique

Mars 2023

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Cette thèse intitulée :

Prise de décision liée à la conception et aux opérations en milieu industriel complexe par l'étude des données en temps réel

présentée par **Émilie THIBAUT**

en vue de l'obtention du diplôme de *Philosophiae Doctor*

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

Michel PERRIER, président

Paul STUART, membre et directeur de recherche

Milan KORBEL, membre et codirecteur de recherche

René HOFMANN, membre

Miguel J. BAGAJEWICZ, membre externe

DÉDICACE

À toutes celles et tous ceux qui ont cru moi et en ce projet du premier au dernier moment.

À celui à qui je dois tout, Steve.

REMERCIEMENTS

Ce travail fut réalisé grâce au support du Conseil de Recherches en Sciences Naturelles et en Génie (CRSNG), Aurel Systems, AVEVA (OSIsoft), Fortress Specialty Cellulose et 3C Software.

Je voudrais tout d’abord remercier Paul Stuart de m’avoir inspiré, guidé et motivé dans la réalisation de ce travail d’envergure. Merci pour tous vos précieux conseils et votre soutien inlassable au cours de ces dernières années. Je vous remercie de m’avoir donné l’occasion d’interagir avec de nombreuses personnes impliquées dans l’industrie des pâtes et papier, des membres de la communauté étudiante ainsi que des membres du corps professoral.

Merci énormément à mon co-directeur, Milan Korbel pour vos conseils judicieux, pour l’aide avec la direction, et surtout le démarrage de ma recherche et de m’avoir aidé à rester sur la bonne voie à de nombreuses reprises. Vous avez toujours été une source d’inspiration et de motivation pour moi!

Je tiens à souligner le travail remarquable de la Fondation Arbour et remercier de tout cœur Diane de Champlain, ma mentore, Marine Hadengue et Joëlle Dorais, pour tous vos encouragements, votre support (moral et financier) et pour toutes les opportunités que vous m’avez offertes. Merci de votre confiance!

Je remercie grandement Bruno Poulin pour son support continu, son aide précieuse, ses conseils et ses idées. Nos échanges étaient toujours très enrichissants et pertinents. Merci pour cet encadrement et l’assistance technique avec EXPLORE.

Un grand merci à Moncef Chioua pour l’aide, le soutien et les discussions fructueuses tout au long de ce projet.

Je tiens à remercier Jeffrey Dean Kelly pour sa vision pratique, son support, son dévouement et son aide avec IMPL.

Merci à tous les étudiants et toutes les étudiantes de notre groupe pour nos discussions et les bons moments, notamment, un chaleureux merci à Caroline Brucel et Émanuelle Trottier pour tous nos échanges, pratiques de présentations et pour votre aide; à Francis Lebreux Désilets pour nos longs débats, échanges et débogage de code et à Alireza Moussavi pour tous les conseils et pour toujours avoir été le premier à accepter de réviser mes travaux. Un chaleureux merci aux filles pour nos chalets de rédaction – Laura Ahunon et Fiona Milano !

Je remercie Maxime McKay et Philippe Mack pour votre aide et contribution, ainsi que toutes les personnes impliquées dans la réalisation des entrevues effectuées dans le cadre de ce projet.

Je remercie chaleureusement ma famille pour leurs encouragements, leur intérêt et leur support continu, Lucie Vallée, Alain Michaud, Danielle, Florent et Stéphanie Thibault, ainsi que mes meilleurs amis Améni Abid et Sylvain Gagnon.

Finalement, je suis immensément reconnaissante envers toi qui as rendu ce projet possible, pour ta patience infaillible, ton support inconditionnel et ton amour à travers ces années, merci Steve Michaud! Ta confiance, ton soutien et tes encouragements ont toujours été la source la plus profonde de mon inspiration.

NOTE D'INTENTION

Ce texte utilisera un langage inclusif, afin de donner une visibilité à toutes les identités de genre.

RÉSUMÉ

Aujourd'hui, dans les industries de transformation, de nombreuses données sont mesurées par des capteurs en temps réel, stockées selon un protocole, puis interprétées pour la prise de décision. Ces données contiennent plusieurs sources et types d'erreurs et, par conséquent, des techniques de traitement des données doivent être appliquées avant d'incorporer ces données dans toute analyse. De plus, l'interprétation des données n'est pas évidente, car les données représentent des opérations complexes. La mise en œuvre d'approches de traitement des données (ou des signaux) permet de corriger les erreurs de mesure afin d'améliorer la fidélité des données et de disposer d'une base plus fiable pour la prise de décision concernant les procédés.

Dans la pratique, la plupart des décisions liées au procédé sont prises en tenant compte de l'état d'équilibre du procédé, mais très peu considèrent qu'un procédé opère sous plusieurs régimes permanents distincts. Une revue de la littérature sur les techniques de traitement des signaux utilisées dans le contexte des opérations en régime permanent a révélé que différentes applications de prise de décision nécessitent des étapes de traitement des signaux distinctes, ces dernières pouvant inclure par exemple le nettoyage des données, la détection des régimes permanents, la réconciliation des données et la détection des régimes d'opération. Pour mieux comprendre les pratiques industrielles liées aux données de procédé, des entrevues sont menées avec des répondants de l'industrie afin d'explorer comment les données peuvent être utilisées et traitées dans différents contextes de prise de décision, quelle est la valeur ajoutée des données de procédé, et les principaux obstacles qui empêchent les usines de les utiliser à leur plein potentiel. Ces entrevues sont menées avec (1) des expertes et experts en usine ainsi que (2) des développeuses et développeurs de logiciels travaillant en soutien à l'industrie des pâtes et papiers.

Les entrevues ont révélé que le nettoyage des données est effectué en relation avec l'application. Les expertes et experts de procédé sont minutieux en ce qui concerne cette étape, mais des approches rudimentaires sont encore utilisées par rapport à ce qui est possible et facilement disponible. De plus, les entrevues ont confirmé que la plupart des décisions dans l'industrie sont basées sur un paradigme d'état d'équilibre. Par exemple, les personnes en industrie font généralement la moyenne des données de procédé représentant des opérations "bonnes et normales" pour la majorité de leurs analyses. La littérature propose de nombreuses techniques basées sur les statistiques pour détecter les périodes en régime permanent au sein d'un procédé, mais celles-ci ne sont généralement pas utilisées par les expertes

et experts de l'usine. En outre, les entrevues ont mis en évidence que les données de procédé auraient une plus grande valeur si les régimes d'opération étaient pris en compte.

Le potentiel des données d'usine n'est souvent pas exploité en raison de la complexité du traitement des signaux. Plusieurs techniques existantes de traitement des données, dont le filtre de Kalman, le filtre à moyenne mobile pondérée exponentielle, la transformée de Fourier à court terme et la transformée en ondelettes, sont comparées en fonction de leurs performances sur des signaux synthétiques de débit et de température, c'est-à-dire en fonction de leur rapport signal/bruit et de leur erreur quadratique moyenne. La transformée en ondelettes s'avère être l'approche la plus robuste et la plus performante pour réduire le bruit des signaux de procédés industriels. Cette approche est connue pour fournir un moyen approprié d'examiner les événements multi-échelles, car elle présente une bonne résolution à la fois en temps et en fréquence. Elle décompose un signal en ses composantes de basse et haute fréquence pour l'analyser. La transformation en ondelettes réduit les composantes haute fréquence qui représentent le bruit du signal et reconstruit les composantes basse fréquence qui reflètent la tendance fondamentale du signal.

Les usines étant lourdement instrumentées par des milliers de capteurs, le réglage des paramètres des filtres pour traiter les signaux industriels individuellement est fastidieux, exigeant et prend du temps. Comme la sélection des paramètres d'une technique de traitement des signaux affecte la performance globale du débruitage, un algorithme est conçu pour déterminer les valeurs optimales des paramètres qui réduisent les composantes haute fréquence indésirables tout en conservant la tendance du signal. Sur la base des caractéristiques des signaux, des heuristiques de traitement des signaux peuvent être extraites, facilitant la sélection des paramètres d'une technique de traitement des signaux.

Un ensemble d'heuristiques permettant de sélectionner les paramètres de réglage d'une approche de traitement des signaux pour le débruitage d'une grande quantité de signaux de procédés industriels de manière systématique et efficace est ainsi proposé. Les heuristiques sont obtenues par un algorithme qui a identifié les paramètres optimaux de la transformation en ondelettes pour des centaines de signaux. Les paramètres optimaux sont déterminés en appliquant une fonction objective qui minimise l'amplitude des pics des composantes dont la fréquence est supérieure à la fréquence d'intérêt ainsi que l'erreur quadratique moyenne entre l'amplitude des pics des composantes dont la fréquence est inférieure à la fréquence d'intérêt pour le signal brut et le signal filtré. Les centaines de paramètres optimaux de la transformation en ondelettes des signaux, les caractéristiques des signaux (variance, taux de variation, asymétrie, aplatissement, coefficient de variation et fréquence de tendance) et les paramètres de

classification (type de mesure, nature du fluide, type de capteur et configuration du procédé) sont évalués dans une analyse en composantes principales pour identifier leur corrélation et leur relation. Par conséquent, sur la base des caractéristiques des signaux, des heuristiques aident à sélectionner les paramètres optimaux de la transformation en ondelettes. L'utilisation de ces heuristiques est une première étape pour rendre les signaux de procédés industriels plus adaptés à la prise de décision.

Pour une analyse ou une application ciblée, une fois que les erreurs aléatoires ont été traitées et que tous les signaux de procédé sont synchronisés, sans décalage et sans valeur aberrante, les périodes de régime permanent sont détectées et la réconciliation des données est effectuée pour détecter, identifier et éliminer les erreurs grossières persistantes des mesures. Cependant, les procédés n'opèrent pas dans une seule condition de régime permanent, et les régimes d'opération doivent être pris en compte. La détection et l'identification des régimes d'opération est un moyen de réaliser le plein potentiel des données de procédé pour la prise de décision, elle ouvre la porte à de nombreuses nouvelles opportunités. Les données de procédé interprétées selon de régimes d'opération permettent de prendre des décisions plus éclairées en matière de conception et d'opération.

Les régimes d'opération "*top-down*" sont déterminés par les modifications de la stratégie de contrôle causées par différentes conditions d'opération (grades, saisons, changements de matières premières, maintenance, utilisation de différentes lignes de production, etc.). D'autre part, les régimes d'opération "*bottom-up*" sont identifiés par une analyse de clustering (analyse en composantes principales et algorithme *k-means*), en plus de la connaissance du procédé. Par exemple, les opérateurs agissent lorsqu'un événement nécessitant un ajustement du procédé se produit. Dans ce cas, différentes personnes opérant le procédé n'effectueront pas nécessairement le même ajustement.

La reconnaissance des périodes en régime permanent et l'interprétation des régimes d'opération qui leur sont associés peuvent apporter des avantages substantiels dans la prise de décision concernant les procédés. La combinaison des régimes d'opération définis par des conditions de procédé prédéterminées (recettes ou analyse "*top-down*") avec ceux détectés par l'analyse guidée par les données (analyse "*bottom-up*") fournit un ensemble de régimes d'opération qui permet une interprétation plus concrète du fonctionnement du procédé et donc de meilleures décisions de conception et d'opération du procédé. Par exemple, grâce à l'utilisation de la modélisation des coûts axée sur les opérations et des données traitées correspondant à un fonctionnement en régime permanent, un profit est attribué à chaque régime d'opération afin d'identifier le plus rentable.

La définition de l'étendue d'une analyse, la définition des régimes d'opération "*top-down*", le traitement des signaux, la détection des opérations en régime permanent, la réconciliation des données, ainsi que la détection et l'identification des régimes d'opération "*bottom-up*" constituent un cadre méthodologique de traitement des données permettant de prendre des décisions éclairées en vue d'améliorer les procédés. Toutes les étapes de ce cadre méthodologique basé sur les régimes d'opération nécessitent une combinaison de connaissances du procédé et d'approches axées sur les données. La valeur ajoutée et le potentiel de ce cadre sont démontrés à l'aide d'une étude de cas; en effectuant une analyse de coût basée sur les activités dans le département de lavage d'une usine de pâte à dissoudre. La performance du procédé de lavage est évaluée pour douze conditions d'opération en régime permanent et une analyse d'optimisation a exploré et comparé ces scénarios pour mettre en évidence une façon améliorée d'opérer (combinaison optimale des valeurs de point de consigne), et reconnu les variables contrôlées qui sont responsables de cette haute performance. Par conséquent, l'analyse d'optimisation a identifié les régimes d'opération qui présentent une performance améliorée, c'est-à-dire lorsque le procédé présente le débit de pâte le plus élevé, une conductivité de pâte minimale (performance de lavage plus élevée) et un coût minimal.

ABSTRACT

Today in the process industries, extensive data are measured by sensors in real-time, stored according to a protocol, and then interpreted for decision-making. These data contain several sources and types of error and, therefore, data processing techniques must be applied before incorporating these data into any analysis. Furthermore, data interpretation is not obvious because the data represent complex operations. The implementation of data (or signal) processing approaches enables the correction of measurement errors to improve data fidelity, and a more reliable basis for process decision-making.

In practice, most process decisions are made considering process steady-state, but very few consider that a process operates under many distinct steady-state regimes. A literature review of signal processing techniques employed for the context of steady-state operations revealed that different decision-making applications require distinct signal processing steps, the latter of which can include for example data cleaning, steady-state detection, data reconciliation, and operating regime detection. To better grasp industrial practice related to process data, interviews are completed with industry respondents to explore how data might be used and processed in different decision-making contexts, what the added value of process data is, and the major barriers that prevent plants from using them to their full potential. These interviews are conducted with (1) mill experts as well as (2) with software developers working in support of the pulp and paper industry.

The interviews revealed that data cleaning is performed in relation to the application. Process experts are thorough regarding this step, but rudimentary approaches are still used compared to what is possible and readily available. Moreover, the interviews confirmed that most decisions in industry are based on a steady-state paradigm. For example, industry practitioners generally average process data representing “good and normal” operations for the majority of their analyses. The literature proposes numerous statistically-based techniques to detect steady state periods in a process, but these are generally not used by plant experts. Furthermore, the interviews highlighted that process data would have greater value if operating regimes were considered.

The potential of plant data is often not realized due to the complexity of signal processing. Various existing data processing techniques including Kalman filter, Exponential Weighted Moving Average filter, short-time Fourier transform, and wavelet transform are compared based on their performance on synthetic flow and temperature signals, i.e., according to their signal-to-noise ratio and root mean square error. The wavelet transform is found to be the most robust and performant approach to reduce noise in

industrial process signals. This approach is known to provide a suitable means of examining multi-scale events, as they have good resolution in both time and frequency. It decomposes a signal into its low and high frequency components for analysis. The wavelet transform reduces the high frequency components that represent the noise of the signal and reconstructs the low frequency components that reflect the fundamental trend of the signal.

As plants are heavily instrumented with many thousands of sensors, tuning filter parameters to process industrial signals individually is tedious, labor-intensive, and time-consuming. As the selection of a signal processing technique's parameters affect the overall denoising performance, an algorithm is tailored to determine the optimal parameter values that reduce unwanted high frequency components while retaining the signal's trend. Based on signal characteristics, signal processing heuristics can be extracted, facilitating the selection of a signal processing technique's parameters.

A set of heuristics to select signal processing approach tuning parameters for denoising a large amount of industrial process signals in a systematic and efficient way is thus proposed. Heuristics are obtained through an algorithm that identified the optimal wavelet transform parameters for hundreds of signals. The optimal parameters are determined by applying an objective function that minimizes the peaks amplitude of components whose frequency is higher than the frequency of interest as well as the root-mean-square error between the peaks amplitude of components whose frequency is lower than the frequency of interest for the raw signal and the filtered signal. The hundreds of signals' optimal wavelet transform parameters, features (variance, rate of change, skewness, kurtosis, coefficient of variation, and trend frequency) and classification parameters (measurement type, fluid nature, sensor type and process configuration) are assessed in a Principal Component Analysis to identify their correlation and relationship. Therefore, based on the signal characteristics, heuristics help selecting the optimal wavelet transform parameters. Using these heuristics is a first step in making industrial process signals more suitable for decision-making.

For a targeted analysis or application, once random errors are reduced from process signals, and that all signals are synchronized, lag-free and outlier-free, then steady-state periods are detected, and data reconciliation is performed to detect, identify and eliminate measurements persistent gross errors. However, processes don't operate at a single steady-state condition, and operating regimes must be accounted for. Operating regime detection and identification is a means to achieve the full potential of

process data for decision-making, it opens the door to many new opportunities. Process data interpreted in terms of operating regimes allows for more informed design and operating decisions.

“Top-down” operating regimes are driven by changes in the control setpoint strategy caused by different operating conditions (grades, seasons, feedstock changes, maintenance, the usage of different production lines, etc.). On the other hand, “bottom-up” operating regimes are identified through clustering analysis (principal component analysis and *k*-means algorithm), alongside process knowledge. For instance, operators act when an event occurs that requires a process adjustment. In this case, different operators won’t necessarily make the same adjustment.

Recognizing steady-state conditions and interpreting the operating regimes associated with these can lead to substantive benefits in process decision-making. Merging operating regimes defined by predetermined process conditions (recipes or “top-down” analysis) with those detected by data-driven analysis (“bottom-up” analysis) provides a set of operating regimes that allow for a more concrete interpretation of process operation and thus better design and operational decisions about the process. For instance, with the use of operations-driven cost modeling and processed data corresponding to steady-state operation, profit is assigned to each operating regime in order to identify the most cost-efficient one.

Defining the scope of an analysis, defining “top-down” operating regimes, performing signal processing, detecting steady-state operations, performing data reconciliation, and detecting and identifying “bottom-up” operating regimes comprise a data processing framework for informed decision-making for process improvement. All steps of this operating regime-based data processing framework require a combination of process knowledge and data-driven approaches. The added value and potential of this framework is proven using a case study; performing an activity-based costing analysis in the washing department of a dissolving pulp mill. The washing process performance is assessed for twelve steady-state operating conditions, and an optimization analysis explored and compared these scenarios to highlight the preferred way to operate (optimal combination of set point values), and recognized which controlled variables are responsible for this high performance. Therefore, the optimization analysis identified which operating regimes present an improved performance, i.e., when the process is showing the highest pulp throughput, minimal pulp conductivity (higher washing performance) and minimal cost.

TABLE DES MATIÈRES

DÉDICACE	iii
REMERCIEMENTS	iv
NOTE D'INTENTION	v
RÉSUMÉ.....	vi
ABSTRACT	x
TABLE DES MATIÈRES.....	xiii
LISTE DES TABLEAUX	xvi
LISTE DES FIGURES.....	xvii
LISTE DES SIGLES ET ABRÉVIATIONS.....	xix
LISTE DES ANNEXES.....	xx
CHAPITRE 1 INTRODUCTION	1
1.1 Problématique.....	1
1.2 Questions de recherche	9
1.3 Organisation de la thèse	10
CHAPITRE 2 REVUE DE LA LITTÉRATURE	12
2.1 Type d'erreurs dans les données de procédé.....	12
2.2 Retrait du bruit dans les données.....	14
2.2.1 Filtre de Kalman.....	15
2.2.2 Moyenne mobile à pondération exponentielle	17
2.2.3 Transformée de Fourier à court terme	18
2.2.4 Transformée en ondelettes	20
2.2.5 Comparaison des approches	26
2.2.6 Sélection des paramètres de la transformée d'ondelette	30

2.3	Caractérisation des signaux.....	34
2.4	Formulation systématique d’heuristiques	35
2.5	Traitement des données dans le cadre d’applications et de prises de décisions industrielles.....	38
2.6	Détection des régimes d’opération des procédés industriels	48
2.7	Approches utilisées pour l’analyse des coûts d’opération	50
2.8	Analyse d’optimisation dans le secteur des pâtes et papier	54
2.9	Lacunes dans le corps de connaissance	56
CHAPITRE 3 OBJECTIFS.....		61
CHAPITRE 4 APPROCHE MÉTHODOLOGIQUE GLOBALE		63
4.1	Survol de la méthodologie	63
4.2	Mener des entrevues au sujet de l’utilisation des données en industrie	63
4.3	Comparaison d’approches permettant le retrait du bruit	66
4.4	Établissement des heuristiques pour le traitement des données	69
4.5	Développement d’un cadre méthodologique de manipulation des données pour la prise de décision	72
4.6	Application du cadre méthodologique pour l’optimisation d’un procédé industriel.....	74
CHAPITRE 5 ÉTUDE DE CAS		78
5.1	Contexte	78
5.2	Données industrielles.....	80
5.3	Régimes d’opération.....	81
5.4	Coût de production	82
CHAPITRE 6 SOMMAIRES EXÉCUTIFS DES PUBLICATIONS		84
6.1	Présentation des publications	84
6.2	Relation entre les publications	84

6.3	Synthèse des résultats	85
6.3.1	Revue des pratiques de l'industrie en matière de traitement et d'utilisation des données dans la prise de décision	86
6.3.2	Comparaison des approches de traitement des signaux.....	90
6.3.3	Établissement d'heuristiques basées sur les transformées d'ondelette.....	94
6.3.4	Établissement d'un cadre méthodologique de traitement des données qui considère les régimes d'opération d'un procédé industriel	98
6.3.5	Application du cadre méthodologique a un problème d'optimisation	108
CHAPITRE 7	DISCUSSION GÉNÉRALE.....	114
7.1	Prises de décisions en industrie : régime permanent et régime d'opération	114
7.2	Heuristiques basées sur les transformées en ondelettes pour un traitement plus efficace des données	119
7.3	L'établissement d'un cadre méthodologique de traitement des données : régimes d'opération et coûts d'opération pour une interprétation plus concrète.....	123
CHAPITRE 8	CONCLUSION.....	127
RÉFÉRENCES		133
ANNEXES		145

LISTE DES TABLEAUX

Tableau 2.1 – Comparaison des approches prenant en compte la sélection des paramètres.....	56
Tableau 2.2 – Recensement des articles qui considèrent la sélection des paramètres optimaux des TO	58
Tableau 4.1 – Paramètres de débruitage couramment utilisés	67
Tableau 4.2 – Plages des paramètres de débruitage par ondelettes couramment utilisés	70
Tableau 4.3 – Variables utilisées dans l'analyse en composantes principales.....	72
Tableau 6.1 – Résumé de la manière dont l'industrie utilise ses données pour la prise de décision et comment elles sont traitées	87
Tableau 6.2 – Moyennes des critères (RMSE et SNR) avec leurs écarts-types pour toutes les approches	92
Tableau 6.3 – Paramètres clés pour tous les régimes d'opération détectés.....	110
Tableau 7.1 – Validation des heuristiques sur sept signaux	120

LISTE DES FIGURES

Figure 2.1 – Représentation de la précision (erreur aléatoire et valeurs aberrantes) et du biais (erreur systématique)	14
Figure 2.2 – Représentation en ondelettes multi-échelles de mesures en temps réel, inspiré de (Korbel, Milan et al., 2014)	23
Figure 2.3 – Exemples de familles d'ondelettes (Taspinar, 2018).....	24
Figure 2.4 – Décomposition et synthèse du signal à trois échelles en utilisant des transformées en ondelettes	25
Figure 2.5 – Fenêtre unique et fixe de la TFCT et celles variables de la TO (Ganesan et al., 2004).....	28
Figure 2.6 – Algorithme de traitement des données pour l'optimisation et le contrôle, inspiré de Abu-el-zeet et al. (2001).....	40
Figure 2.7 – Cadre méthodologique de traitement des données pour l'analyse des coûts, inspiré de Korbel, Milan (2011)	43
Figure 2.8 – Stratégie d'amélioration de la qualité des données pour la calibration et la validation des simulations (Jiang, Taiwen, Chen, Jasim, et al., 2003).....	44
Figure 2.9 – Méthodologie pour améliorer la qualité des données proposée par Bellec et al. (2007) ...	46
Figure 2.10 – Schéma de la chaîne de traitement des données (Heikkinen et al., 2011).....	49
Figure 2.11 – Présentation de la méthodologie présentée par Korbel et al. (Korbel, M. et al., 2007) ...	52
Figure 2.12 – Échange d'informations entre le modèle de procédé et les données économiques (Laflamme-Mayer et al., 2011).....	53
Figure 2.13 – Comparaison des méthodes d'analyse des coûts (Korbel, M. & Stuart, 2011)	54
Figure 4.1 – Répartition des rôles des personnes interrogées	65
Figure 4.2 – Simulation d'un réseau d'échangeurs de chaleur: signaux de débit et de température	66
Figure 4.3 – Méthodologie employée pour formuler des heuristiques de traitement des signaux en fonction de leurs caractéristiques	69
Figure 4.4 - Présentation du cadre méthodologique de traitement des données.....	73

Figure 5.1 - Département de lavage de la pâte	79
Figure 6.1 – Représentation visuelle des différentes publications et activités doctorales	85
Figure 6.2 – Réduction de l'erreur quadratique moyenne pour les signaux synthétiques considérant le KF, la MMPE, la TFCT et la TO	90
Figure 6.3 – Amélioration du rapport signal/bruit pour les signaux synthétiques considérant le KF, la MMPE, la TFCT et la TO.....	90
Figure 6.4 – Résultats du débruitage du signal TI20 en utilisant a) FK, b) MMPE, c) TFCT, et d) TO	93
Figure 6.5 – <i>Scores chart</i> pour tous les paramètres de classification du signal et de TO.....	95
Figure 6.6 – Contributions des clusters présents dans le <i>scores chart</i>	95
Figure 6.7 – Scores chart considérant uniquement les mesures des capteurs dans les boucles de contrôle	96
Figure 6.8 – Paramètres d'ondelettes optimaux en fonction des caractéristiques de classification du signal	98
Figure 6.9 – Cadre combinant la connaissance du procédé et les approches axées sur les données	99
Figure 6.10 – Représentation du démarrage, de l'arrêt et de l'étendue des périodes en régime permanent	104
Figure 6.11 – Exemple de schéma bloc pour présenter les courants permanents et temporaires.....	105
Figure 6.12 – Contribution des variables pour les régimes d'opération 2 et 12.....	111
Figure 7.1 – Paramètres d'ondelettes optimaux appliqués à de nouveaux signaux pour évaluer leur applicabilité.....	121

LISTE DES SIGLES ET ABRÉVIATIONS

ABC	<i>Activity-based costing</i>
ACP	Analyse en composante principale
FFT	<i>Fast Fourier transform</i>
FK	Filtre de Kalman
MMPE	Moyenne mobile à pondération exponentielle
ODCA	<i>Operations-driven costing approach</i>
RMSE	<i>Root-mean-square error</i>
SNR	<i>Signal-to-noise ratio</i>
TF	Transformée de Fourier
TFCT	Transformée de Fourier à court terme
TFCTI	Transformée de Fourier à court terme inverse
TFR	Transformée de Fourier rapide
TO	Transformée en ondelettes
TOC	Transformée en ondelettes continues
TOD	Transformée en ondelettes discrètes

LISTE DES ANNEXES

Annexe A Configuration du prodécé de pâte dissoute.....	145
Annexe B Article #1 Experimental methods in chemical engineering : data processing and data usage in decision-making	146
Annexe C Article #2 Comparison of signal processing methods considering their optimal parameters using synthetic signals in a heat exchanger network simulation	174
Annexe D Article #3 Signal processing heuristics for the selection of wavelet transform parameters considering signal characteristics derived from a case study on a chemical pulp mill.....	200
Annexe E Article #4 An operating regime-based data-driven industrial framework for improved process decision-making – Part 1: Framework development	225
Annexe F Article #5 An operating regime-based data-driven industrial framework for improved process decision-making – Part 2: Framework application considering activity-based costing concepts	272

CHAPITRE 1 INTRODUCTION

1.1 Problématique

L'industrie forestière en Amérique du Nord fait face à de nombreux défis. Sa culture particulière, définie plus bas, accroît la difficulté à survivre dans le marché actuel, qui est très concurrentiel voire, en déclin. En effet, elle connaît une diminution de son chiffre d'affaires et de ses activités causée principalement par la baisse de consommation de ses produits. Les usines de pâtes et papiers cherchent à se rajeunir, suivre le mouvement écologique et à améliorer leur situation financière par l'implémentation de nouveaux modèles d'affaires plus durables. Une des stratégies employées est de diversifier leur portefeuille de produits. Ainsi, de plus en plus d'usines de papier kraft se convertissent à la pâte à dissoudre et se tournent vers le concept du bioraffinage – traitement de la biomasse forestière pour la production intégrée de produits de pâtes et papiers conventionnels et de biocarburants, biochimiques, et biomatériaux. Toutefois, la modernisation des usines de pâtes et papiers est ardue en raison de leur nature. En effet, plusieurs faits ressortent de la culture du secteur des pâtes et papiers.

Premièrement, les usines, souvent très vieilles, comptent une grande quantité de conduites et de boucles de recirculation; l'interprétation des opérations de ces procédés industriels est assez complexe. En outre, les mesures rapportées par les différents capteurs sont inexactes et le très faible niveau de redondance des instruments de mesure à travers ces usines n'aide pas le traitement des données récoltées. Également, seule une très faible proportion d'usines se dotent d'une simulation de leur procédé. Lorsqu'une telle simulation est réalisée, elle n'est souvent utilisée qu'à court terme, sans représentation des différents régimes d'opération (conditions d'opération caractérisant un procédé incluant les points de consignes appliqués, les saisons, les périodes de maintenance, le niveau de production, etc.) et n'est pas utilisée au quotidien pour la prise de décision afin de gérer les opérations.

Deuxièmement, notons que des millions de données sont collectées quotidiennement – une quantité importante de ces données se trouvent sous la forme de signaux¹ émis par des capteurs situés

¹ Dans les procédés industriels, la plupart des données proviennent de capteurs qui génèrent des signaux. En effet, l'échantillonnage des données se fait dans le domaine discret; cependant, le temps très court entre les échantillons crée un signal qui semble être continu. Ainsi, une séquence de données constitue un signal "continu" de données de

stratégiquement dans l'usine – mais elles ne sont pas utilisées à leur plein potentiel. Dans la majorité des cas, aucune information économique n'en est déduite; la relation entre les données de procédé et les coûts d'opération est sous-exploitée. Pourtant, cela permettrait d'identifier les coûts par produits et par régimes d'opération. En effet, les marges bénéficiaires des produits individuels et des régimes sont souvent inconnues, seule la marge globale de l'usine l'est et cette dernière ne permet pas d'identifier les produits et régimes non rentables. La totalité des coûts encourus dans l'année et les revenus annuels permettent de calculer la marge globale, mais pas le coût et le revenu marginal de chaque produit et régime d'opération. Effectuer cette analyse pourrait s'avérer fort pertinent étant donné que le secteur forestier est régi par une culture de commodités qui implique de faibles marges économiques. Ainsi, connaître les coûts associés aux régimes d'opération de l'usine permettrait d'identifier ceux qui s'avèrent non rentables pour l'usine et qui, sans le savoir, font diminuer la marge globale, car les contributions individuelles sont inconnues. Cependant, les régimes d'opérations, qui ouvriraient la possibilité d'interpréter les données plus efficacement pour une meilleure prise de décision, ne sont que très peu pris en compte dans les différentes analyses.

Les données de procédé, qui rappelons-le, ne sont pas utilisées à leur juste valeur et dont le potentiel passe souvent inaperçu, offrent pourtant, de nombreuses possibilités. Plus particulièrement, les données permettent une meilleure compréhension du procédé; tant qu'une infrastructure de collecte des données n'est pas mise en place, il n'est pas possible de savoir ce qui se passe au sein du procédé, autant au moment présent que dans le passé. Les données de procédé fournissent des informations tangibles que les usines peuvent exploiter; elles permettent de faire des découvertes importantes, entre autres, en analysant les relations entre les variables du procédé.

En outre, disposer de données historiques représentatives du procédé permet de conduire à de meilleures décisions pour les opérations futures. Ces prédictions se basent sur la surveillance du procédé en temps réel². On y observe, entre autres, les changements au sein des opérations, la façon dont les opérateurs

capteurs de procédé industriels. La convention utilisée dans ce document consiste généralement à utiliser indifféremment le traitement des données et le traitement des signaux.

² En temps réel par rapport à la dynamique du procédé analysé. Les données en temps réel signifient que la vitesse d'analyse des données se rapproche du temps réel à l'échelle du procédé. Si un procédé ne varie que toutes les heures et qu'autrement rien ne se passe, et qu'une valeur est échantillonnée toutes les minutes, les données sont en temps

gèrent l'usine ainsi que les impacts saisonniers. De plus, en utilisant des données historiques, une analyse post-mortem peut être effectuée. Il est possible d'analyser des événements passés qui ont conduit à un problème et d'en identifier la cause. Par la suite, utiliser les données pour reconnaître qu'une situation déjà survenue est en train de se reproduire et trouver comment anticiper et prévenir les problèmes avant qu'ils ne se produisent est primordial. Par conséquent, il est extrêmement pertinent d'identifier les problèmes d'opération, de les diagnostiquer pour tenter de les éviter et d'identifier de nouvelles opportunités. Globalement, les données de procédé font partie intégrante du processus d'amélioration continue des usines.

Également, la performance d'une usine peut être évaluée grâce aux données en utilisant des indicateurs clés de performance tels que le rendement du procédé et la qualité du produit. De plus, les données permettent d'évaluer la disponibilité des équipements par l'intermédiaire d'une analyse de désengorgement. Conjointement, les données offrent la possibilité de faire de la gestion des temps d'arrêt afin de comprendre leurs causes, et par le fait même, de réduire ces temps, et permettent également l'analyse des écarts de qualité au niveau des produits afin de les corriger. Grâce aux données de procédé, il est possible de viser l'excellence opérationnelle (efficacité) par des études d'optimisation, d'effectuer des analyses d'efficacité énergétique et d'évaluer la consommation de ressources ainsi que la conformité aux réglementations environnementales. De même, les données permettent une maintenance prédictive par le biais d'analyses de la performance et l'état des équipements. Enfin, les données de procédé peuvent être intégrées à d'autres parties de l'organisation par l'utilisation d'un logiciel de gestion des ressources tel que SAP, afin d'améliorer la gestion de l'inventaire, de déterminer quelle matière première commander, d'établir des rapports, d'effectuer un suivi et, surtout, de réaliser une analyse de coûts. Il suffit de dire que la plus grande valeur est de disposer de données exploitables à partir desquelles de véritables décisions peuvent être prises.

réel, car tout ce qui va varier sera capturé avec cet échantillon. Si un autre procédé varie toutes les heures, mais qu'une valeur n'est prélevée que toutes les 24 heures, ces données ne sont pas en temps réel. Le "temps réel" est toujours en relation avec quelque chose. Pour être en temps réel, il faut suffisamment de points pour reconnaître tous les événements du procédé.

Les données sont sous-utilisées pour de nombreuses raisons. Les défis qui empêchent l'utilisation des données à leur plein potentiel incluent le manque de ressources, autant humaines que temporelles. Il y a peu de ressources dans l'usine qui sont dédiées à l'analyse des données. Les usines sont en pénurie de main-d'œuvre en général, et donc leur demander d'implanter un nouvel outil ou d'analyser des données, même si cela engendre une valeur substantielle, n'est pas réaliste puisque toutes leurs ressources sont déployées dans le maintien d'une production efficace. En outre, l'extraction de grandes quantités de données prend beaucoup de temps. L'analyse des données n'est pas une priorité; les usines se concentrent sur les opérations quotidiennes et la livraison des produits de haute qualité.

Plus spécifiquement, il existe un fossé entre les expertes et experts de procédé qui n'ont pas forcément les aptitudes nécessaires pour comprendre l'étendue de ce qui est possible de faire avec les données et les personnes qui ont ces compétences, mais qui ont peu de connaissances au sujet du procédé. Toutefois, cet obstacle n'est pas propre à l'industrie forestière; il se produit dans la plupart, voire la totalité, des industries de transformation.

Également, un obstacle fort important est la culture du secteur forestier. Les usines ont tendance à être très figées dans leur façon de faire les choses et elles fonctionnent à effectifs réduits; il n'y a souvent pas assez de personnel pour accomplir l'essentiel. En particulier, elles n'ont pas le temps d'innover, d'explorer et d'essayer de faire les choses différemment ou de voir les choses sous un angle différent, car leur principal objectif est d'atteindre les cibles de production et de qualité. Ainsi, les usines n'ont pas cette culture qui leur permettrait de considérer les données d'une manière différente, d'essayer de nouveaux outils et de nouvelles méthodologies de traitement des données. C'est un défi considérable pour ces dernières de continuer à évoluer dans leur utilisation des données de procédé.

Conjointement vient la résistance au changement; pour certaines personnes, les changements se produisent plus vite qu'elles ne peuvent les absorber. Dans le même ordre d'idée, même si la technologie est présente de manière prépondérante dans nos vies, peu d'individus ont comme réflexe de mettre en œuvre des solutions technologiques permettant d'utiliser les données pour la prise de décision.

De plus, collecter une quantité importante de données engendre son lot de soucis. L'usine doit se demander à quoi ces données serviront-elles et pourquoi autant ont été collecté, et particulièrement, si la collecte a été faite correctement. Les usines doivent éviter « le syndrome de la boîte noire » qui consiste à collecter un maximum de données pour être sûr d'être en mesure de trouver l'information nécessaire. Elles stockent ensuite toutes ces données parce qu'elles ne savent jamais quand elles en auront besoin, mais si cela n'a

pas été fait de manière rationnelle et systématique, les usines se rendront compte qu'elles ont stocké des données de mauvaise qualité. Par conséquent, il est probable que dans le système d'historisation des données, ces dernières ne soient pas organisées correctement, qu'elles ne soient pas à la bonne résolution – ce qui signifie que les données sont disponibles, mais pas utilisables en soi – et il y ait des données manquantes. De ce fait, il s'agit d'une question de qualité des données; la qualité se mesure par le fait d'avoir le moins d'erreurs possible, le moins de données manquantes et incohérentes. Ainsi, le défi consiste à transformer ces données stockées, qui ont été recueillies sans savoir à quoi elles serviraient, en un ensemble de données utilisables.

Cependant, les usines ne possèdent pas nécessairement l'expertise requise pour comprendre les données (ce qu'elles signifient), les interpréter et les traiter (corriger) afin de les rendre utilisables. Beaucoup passent directement à l'analyse en utilisant des données brutes sans effectuer de traitement de données (rendre les données utilisables) au préalable pour s'assurer que les données introduites dans les différentes applications ont été conditionnées pour ces dernières. Également, d'autres aspects à considérer pour une utilisation efficace des données sont le décalage temporel des données pour refléter le décalage au sein du procédé – il faut 20 heures pour passer du digesteur à la machine à papier – ainsi que la nature transitoire des usines. Comprendre la dynamique du procédé pourrait constituer un défi majeur en ce qui concerne le traitement des données.

L'importante quantité de données collectées rend également difficile l'identification de ce qu'il faut utiliser; il s'agit d'examiner les données et de comprendre ce qu'elles signifient. L'utilisation de ces données offre tellement de possibilités qu'il est fastidieux de statuer sur les applications à prioriser et quelles données sont pertinentes.

En outre, un autre défi en lien avec l'utilisation des données est de choisir le bon taux de compression, car il ne faut pas stocker trop d'informations (ce qui pourrait être considéré comme du bruit) ni surcomprimer les données (ce qui signifierait que certaines données ne seraient pas disponibles rapidement). Même si la mémoire de stockage dans l'infonuagique est de moins en moins chère, le prix est souvent fonction du nombre de variables (tags), ainsi les considérations de coût de stockage ne sont donc pas totalement exclues. Ainsi, la compression des données est toujours pratiquée puisqu'elle impacte fortement la performance de l'extraction des données *online* en temps réel, et particulièrement dans le cas où des millions de données sont stockées en continu. Il existe un équilibre entre la qualité de l'information (données obtenues) et la performance de l'infrastructure de collecte des données. Avec un bon ratio de

compression, les usines peuvent maintenir la qualité de l'information sans consommer trop d'espace. Elles devraient sélectionner les tags à compresser et la manière de le faire. L'algorithme de compression devrait idéalement être spécifique à chaque variable, mais les usines ont tendance à privilégier des approches plus globales (à large spectre) étant donné le grand volume de variables.

Une autre raison qui pourrait justifier la sous-utilisation des données est qu'il est difficile de savoir quels outil ou méthode de traitement des données appliquer pour chaque application, car les usines n'ont pas nécessairement le temps d'analyser cela en détail. Dans le même ordre d'idée, il n'est pas toujours facile de convaincre les usines de déployer les efforts substantiels requis pour mettre en œuvre un nouvel outil, malgré les avantages évidents. En général, elles s'interrogent sur la pertinence de la solution, les aspects techniques, la sécurité, le coût, le déploiement et l'hébergement, la satisfaction de leurs clients, les aspects juridiques, etc.

Un dernier frein à l'utilisation des données de procédé est le fait qu'elles présentent généralement diverses sources d'erreurs et qu'elles doivent être traitées en fonction de l'application visée. De ce fait, il est ardu d'interpréter et de travailler avec ces données; une bonne connaissance du procédé est primordiale afin de pouvoir les traiter et en retirer les erreurs. Ainsi, le traitement des données demande beaucoup de ressources, et ce, surtout si les usines ne disposent pas des bons outils.

Si les données ne sont pas de bonne qualité (taux de variation hors des limites normales, données non synchronisées, espacement irrégulier, changements brusques, valeurs hors limites, présence de bruit, des valeurs aberrantes, absence de données, etc.) et ne sont pas fiables, les résultats des analyses qui émanent de l'utilisation de ces données et les décisions qui en découlent ne le seront pas non plus. Ainsi, les données caractérisant un procédé affichant des erreurs peuvent nuire à la prise de décision. En effet, les erreurs dans les mesures peuvent entraîner une détérioration des opérations de l'usine. Elles peuvent dégrader les performances du système de contrôle, masquer les améliorations apportées au procédé, entraîner une défaillance ou une détérioration de l'état équipements, empêcher l'optimisation du procédé ou faire en sorte que le procédé opère dans un régime non rentable ou non sécuritaire. Plus précisément, le bruit (erreurs aléatoires) affecte directement la précision du système de contrôle, l'évaluation des performances, le diagnostic des fautes et la fiabilité des données extraites.

Il existe diverses approches mathématiques permettant de retirer les erreurs des signaux, et particulièrement les composantes haute fréquence indésirables (erreurs aléatoires) telles que le filtre de Kalman, le filtre exponentiel à moyenne mobile pondérée, la transformée de Fourier à court terme et la

transformé d'ondelette. Toutes ces approches ne présentent pas nécessairement la même efficacité, surtout lorsqu'il s'agit de filtrer des signaux de capteurs de procédé qui présentent une large gamme de composantes fréquentielles. La performance de ces approches mathématiques dépend fortement de la sélection de ses paramètres. Toutefois, la justification de ce choix n'est souvent pas explicite et ce dernier peut même être fait sur une base arbitraire à partir d'expérience personnelle. Ainsi, advenant une comparaison de la performance de ces techniques de filtrage³, il est essentiel de s'assurer que les paramètres utilisés soient optimaux. En plus de la performance des filtres, la façon dont leurs paramètres sont sélectionnés doit être prise en compte, c'est-à-dire la facilité de sélectionner les paramètres appropriés pour chaque approche, et surtout en considérant la nécessité d'effectuer un traitement pour un grand volume de données industrielles. Dans ce cas, il devient fastidieux d'appliquer ces méthodes (régler leurs paramètres) un signal à la fois. Par conséquent, une approche permettant de réduire le temps consacré au retrait du bruit, ainsi que l'intervention des expertes et experts de procédé, pourrait être bénéfique.

Outre le retrait des erreurs aléatoires, dans cette thèse, le traitement des signaux est assimilé à l'ensemble des étapes nécessaires pour traiter les données brutes afin de les rendre adaptées à diverses applications décisionnelles. Parmi les techniques permettant de préparer les signaux (données) en vue d'une analyse ultérieure, citons la réduction du bruit (erreurs aléatoires), mais également la détection des valeurs aberrantes, la détection des périodes en régime permanent, la détection des régimes d'opération et la réconciliation des données.

De nombreuses décisions liées à la gestion des opérations se font sur la base de l'analyse de données historiques qui représentent le procédé en régime permanent, que ce soit de façon directe ou indirecte, explicite ou implicite. Ainsi, la détection de ces périodes qui représentent les "opérations normales souhaitables" est souvent nécessaire et requise. Les données en régime permanent permettent de mieux comprendre le fonctionnement d'un procédé, elles sont plus faciles à interpréter et apporte une grande valeur; elles fournissent des informations essentielles à la prise de décision.

Dans un procédé, il existe de nombreux régimes permanents différents qui peuvent être caractérisés par des changements au niveau des intrants (nature ou qualité), différents produits, saisons, recettes, etc. Le

³ Les termes débruitage, suppression du bruit, filtrage et traitement sont utilisés tout au long du document pour désigner la même chose.

processus de prise de décision pourrait bénéficier grandement de l'identification de ces régimes d'opération. En effet, ceux-ci conduisent à de nombreuses décisions relatives au procédé (davantage que lorsqu'ils ne sont pas considérés) et permettent d'effectuer des analyses plus poussées. Par exemple, la détection des régimes d'opération offre la possibilité d'évaluer la consommation d'énergie (ou de ressources), l'efficacité, la performance, le coût, la qualité du produit et le rendement de chaque régime. Cette analyse pourrait indiquer à une usine quelle est la meilleure façon d'opérer et aider à désengorger le procédé. Par conséquent, la détection des régimes d'opération pourrait avoir un impact sur la planification de la production. Ainsi, la définition des régimes d'opération aide à comprendre le procédé de façon générale.

Ainsi, le traitement des données est ce qui permet de tirer profit des données de procédé; en d'autres mots, pour extraire des informations pertinentes des données, elles doivent être traitées. Subséquemment, les données traitées peuvent être utilisées pour de nombreuses applications et permettent de prendre des décisions éclairées liées aux opérations; pour exploiter les données industrielles pour la prise de décision, une série d'étapes de traitement, qui varie selon l'application, est suivie. En effet, les données brutes n'ont qu'une valeur limitée tant qu'elles ne sont pas traitées. Cela est renforcé par la fameuse citation "*garbage in, garbage out*" qui souligne le fait que le traitement des données est une étape cruciale et essentielle. De ce fait, cette étape augmente la valeur ajoutée et améliore la qualité des données recueillies ainsi que la prise de décision. Les données traitées sont utilisées pour effectuer des tâches liées à l'opération des procédés chimiques, telles que la configuration et l'amélioration des systèmes de contrôle, la détection et le diagnostic de fautes, la planification des opérations, la résolution de problèmes d'opérations, l'identification des besoins de maintenance, l'optimisation et l'analyse des coûts.

Tout le potentiel des données industrielles ne peut être obtenu que si la connaissance du procédé est liée aux approches mathématiques pour leur traitement. En effet, la complexité des procédés industriels exige la prise en compte de la connaissance des procédés lors de l'analyse des données industrielles pour la prise de décision. Ainsi, il est critique de ne pas négliger cet aspect lors de la gestion des données de procédé.

En dernier lieu, dans le secteur forestier, les études économiques sont basées sur la comptabilité par allocation (au prorata). Cette méthode n'est pas fiable et peut induire en erreur; il est courant de voir les produits, annoncés comme étant les plus rentables, s'avérer être vendus à perte par l'entreprise. Une mauvaise allocation des coûts aux différents produits peut entraîner de grandes pertes financières. Ainsi, à l'instar de la comptabilité par allocation, la comptabilité par activité (*activity-based costing* – ABC)

permet de déterminer la marge bénéficiaire individuelle de chaque produit en retraçant les ressources qu'elle requiert. Sachant qu'une analyse de coûts est aussi juste que les données qui la composent, en couplant les données en régime permanent rectifiées aux données de coûts issues du modèle ABC, il est possible d'obtenir le coût unitaire (marges économiques) des produits et d'évaluer les coûts engendrés par chaque régime d'opération. Ceci permettrait donc d'optimiser la prise de décision, mitiger les risques de conception et améliorer les performances d'un site industriel forestier.

À l'égard des problématiques décrites ci-haut, des questions de recherches ont été déduites.

1.2 Questions de recherche

Le contexte présenté plus haut soulève les questions suivantes :

- Pour quelles raisons les données de procédé ne sont-elles pas utilisées à leur plein potentiel dans les usines de transformation pour la prise de décision au niveau des opérations ? En comparant les étapes de traitement des données dans le cadre d'applications industrielles proposées dans la littérature avec ce qui est utilisé en industrie, y a-t-il des éléments (étapes, techniques de traitement) qui devraient être plus présents et mieux utilisés dans le secteur industriel ? Ces méthodes disponibles dans la littérature prennent-elles en compte les préoccupations de l'industrie et répondent-elles à des problèmes industriels réels ?
- Parmi les méthodes de traitement des signaux existantes, laquelle est la plus performante et robuste face à la réduction du bruit contenu dans les signaux de procédés industriels ? Est-ce que l'exploitation de la relation entre les caractéristiques des signaux et les paramètres de la transformée d'ondelette permet de générer des heuristiques afin de traiter de nombreux signaux de manière systématique et efficace ? De quelle façon des facteurs caractérisant les signaux (variance, taux de variation, aplatissement, asymétrie, etc.) sont-ils influencés par la nature et le type de mesure et la configuration du procédé autour du capteur ?
- Est-ce que l'établissement d'un cadre méthodologique de traitement des données qui combine l'interprétation et la connaissance des opérations du procédé aux approches mathématiques d'analyse des données permettent l'identification de régimes d'opérations interprétable pour la prise de décision ?
- Est-ce que l'utilisation des données de procédé étudiées par un cadre méthodologique ainsi que l'analyse des coûts associés aux régimes d'opération permettraient d'optimiser les opérations de sorte

que des améliorations rentables du procédé et des opportunités de désengorgement pourraient être identifiées et donc, en général, d'améliorer la prise de décision liée aux opérations ? Quels sont les avantages de traiter les données à l'aide d'un cadre méthodologique – est-ce que l'amélioration de la qualité et de la fiabilité des données et la détection des régimes d'opération en régime permanent justifient l'effort de mise en œuvre d'un cadre méthodologique ? Comment le traitement des données et la connaissance des régimes d'opération améliorent la capacité à évaluer les coûts ? Quelle application concrète du cadre méthodologique mettrait en évidence un cas où il est intéressant et justifiable d'utiliser le cadre pour résoudre un problème opérationnel ?

1.3 Organisation de la thèse

Le présent document a pour objectif de présenter l'étendue des résultats et des analyses effectuées dans le cadre de la thèse intitulée « Prise de décision liée à la conception et aux opérations en milieu industriel complexe par l'étude des données en temps réel ». Pour ce faire, le document est organisé comme suit. Le chapitre 2 examine la littérature pertinente en lien avec les questions de recherche afin d'identifier les lacunes dans le corps de connaissances. Plus particulièrement, les sections 2.1 à 2.4 couvrent la matière en lien avec les secondes questions, i.e., les types d'erreurs dans les données de procédé, les approches employées pour retirer le bruit des données, la caractérisation des signaux ainsi que la formulation systématique d'heuristiques. Par la suite, la section 2.5 relate des approches et cadres méthodologiques de traitement des données pour la prise de décision industrielle afin de répondre aux premières questions. La section 2.6 synthétise l'identification de régimes d'opération des procédés industriels. Ces deux dernières sections répondent à la question abordée au 3^e point. Finalement, les dernières questions sont couvertes par les sections 2.7 et 2.8. La première réfère aux approches utilisées pour l'analyse des coûts d'opération et la seconde couvre les analyses d'optimisation dans le secteur des pâtes et papier. La section 2.9 introduit l'analyse critique de la revue de la littérature et identifie les lacunes dans le corps de connaissance.

Le chapitre 3 énonce les objectifs et les hypothèses du projet doctoral. Par la suite, le chapitre 4 présente l'approche méthodologique globale (4.1) ainsi que la méthodologie requise pour mener des entrevues au sujet de l'utilisation des données en industrie (4.2), l'établissement des heuristiques pour le traitement des données (4.3), le développement d'un cadre méthodologique de manipulation des données pour la prise de décision (4.4) et l'application du cadre méthodologique pour l'optimisation d'un procédé industriel (4.5).

Dans le cadre de cette thèse, une usine de pâte à dissoudre est étudiée. Ainsi, le chapitre 5 introduit le contexte de l'étude de cas (5.1) ainsi que les détails relatifs aux données industrielles (5.2), les régimes d'opération (5.3) et aux coûts de production (5.4). Le chapitre 6 aborde les sommaires exécutifs des publications. La section 6.1 présente les publications, la suivante explique la relation entre elles puis 6.3 fait une synthèse des résultats. Les sections 6.3.1 à 6.3.5 décrivent respectivement les résultats des cinq publications.

Le chapitre 7 expose la discussion générale. Plus précisément, les sous-sections 7.1 à 7.4 se réfèrent aux implications des différentes contributions au savoir, soit les hypothèses/objectifs. Finalement, le chapitre 8 conclut la thèse.

CHAPITRE 2 REVUE DE LA LITTÉRATURE

2.1 Type d'erreurs dans les données de procédé

Les données de procédé provenant de capteurs sont corrompues par plusieurs types et sources d'erreur. Il existe deux grandes catégories d'erreurs : les erreurs aléatoires (qui comprennent le bruit et les valeurs aberrantes) et les erreurs systématiques (Figure 2.1).

Premièrement, les valeurs aberrantes sont des observations qui ne se conforment pas au comportement attendu (normal) (Chandola et al., 2009). Elles sont, par définition, des valeurs situées en dehors de la distribution caractérisant les (la majorité des) mesures. Ce type d'erreur ne peut pas être calculé ou prédit. Les valeurs aberrantes peuvent se présenter comme des pics dans l'ensemble des données. Ces anomalies se caractérisent par un changement soudain de très grande amplitude et de courte durée. Toutes les valeurs aberrantes ne sont pas mauvaises, et certaines ne doivent pas être supprimées. Il n'est pas recommandé de supprimer des données simplement pour produire un modèle mieux ajusté ou des résultats statistiquement significatifs. Il est important de comprendre comment les valeurs aberrantes se produisent. Dans les procédés industriels, diverses raisons peuvent être à l'origine de valeurs aberrantes. Elles peuvent être attribuées à une observation incorrectement enregistrée ou saisie dans le système d'historisation des données ou à un traitement inapproprié des données manquantes. C'est notamment le cas pour les données saisies manuellement, comme les données de laboratoire. Dans ces cas, la valeur erronée doit être corrigée si possible; si ce n'est pas possible, le point doit être supprimé. Une deuxième cause de valeurs aberrantes dans les ensembles de données, qui se résout par la suppression des données problématiques, sont les problèmes d'échantillonnage et les conditions inhabituelles, une perturbation du procédé – notamment l'ouverture/la fermeture d'une vanne, un coup de bélier dans le procédé, une baisse ou une hausse soudaine de la température, une bulle de gaz – un prélèvement d'échantillon, une erreur humaine, une intervention manuelle dans le procédé, un flash de vapeur, une déviation dans le procédé, des conditions expérimentales anormales, le démarrage et l'arrêt d'équipements, une défaillance ou dysfonctionnement d'un instrument de mesure ou d'un équipement. Enfin, les valeurs aberrantes peuvent être assimilées à une variabilité naturelle, il n'y a donc rien d'anormal avec ces données. En fait, toutes les distributions de données présentent une dispersion des valeurs, et si la taille de l'échantillon est suffisamment grande, il y aura forcément des valeurs inhabituelles. Des valeurs extrêmes peuvent se produire, mais elles sont moins probables. Ces valeurs inhabituelles sont une partie normale de la distribution des données. Par

conséquent, il n'y a aucune raison justifiable de supprimer ces points. Ultimement, les anomalies peuvent être classées comme tout ce qui se situe en dehors des opérations dites normales et peuvent être considérées comme des observations qui s'écartent de manière significative de la majorité des observations.

Il existe trois catégories différentes dans lesquelles les valeurs aberrantes se classent. (1) Valeurs aberrantes globales – données qui sont très éloignées de la totalité de l'ensemble de données dans lequel elles se trouvent. Pour prendre en considération ces valeurs, l'évaluation du modèle le plus approprié du comportement normal et de la distribution sous-jacente pour chaque série chronologique est à envisager. (2) Valeurs aberrantes contextuelles – données qui s'écartent considérablement du reste de l'ensemble de données dans un contexte ou une condition spécifique, mais qui ne seraient pas considérées comme aberrantes si elles se produisaient dans un contexte différent. En d'autres termes, les valeurs ne sont pas globalement en dehors de la plage normale, mais sont anormales par rapport au contexte. Pour les données de séries chronologiques, le contexte est presque toujours temporel (par exemple, les saisons). Le contexte (saisonnalité ou tendances) doit être spécifié dans les modèles décrivant les opérations normales, cela aidera pour les valeurs aberrantes contextuelles. (3) Valeurs aberrantes collectives – un sous-ensemble de données qui, en tant que groupe, s'écartent de manière significative du reste de l'ensemble de données, mais les valeurs de ces points individuels ne sont pas elles-mêmes anormales, que ce soit dans un sens global ou contextuel. La détection des valeurs aberrantes collectives implique la compréhension des relations entre les différentes variables de procédé.

Les mesures des procédés contiennent du bruit et cela peut être potentiellement nuisible aux actionneurs de procédé si le rapport signal/bruit atteint des niveaux élevés ingérables. Les erreurs aléatoires (bruit) sont causées par les vibrations physiques des éléments du procédé (pompe, compresseur), les conditions ambiantes variables et d'autres perturbations interceptées par les capteurs; elles peuvent affecter les signaux et entraver les analyses du procédé. Ces phénomènes ne sont pas reproductibles, et leur ampleur ne peut donc pas être prédite. L'hypothèse selon laquelle les erreurs aléatoires sont caractérisées par une distribution normale avec une moyenne nulle et une variance connue est généralement acceptée (Bagajewicz, M. J., 2000). Ces erreurs ne peuvent pas être complètement éliminées, elles sont toujours présentes, mais elles doivent être retirées autant que possible de l'ensemble des données pour améliorer les performances, la précision et la robustesse des prédictions (Chiang et al., 2017). Les erreurs aléatoires sont représentées dans les signaux par des fréquences élevées de faible ordre de grandeur (Bellec, 2004; Bellec et al., 2007; Romagnoli & Sanchez, 2000).

Enfin, les erreurs brutes (systématiques) sont reproductibles et peuvent être calculées. Ces erreurs sont principalement associées aux incohérences entre les mesures des capteurs et les bilans de masse et d'énergie. Elles peuvent être évitées grâce à de bonnes procédures d'installation et de maintenance des capteurs. Deux types de causes expliquent ces erreurs : les instruments de mesure, et les erreurs inhérentes au procédé. Dans le premier cas, les erreurs sont généralement causées par un instrument défectueux (une mauvaise installation ou maintenance du capteur) ou par un biais systématique dû à une mauvaise calibration des instruments de mesure, à la corrosion ou à l'encrassement du capteur. Pour le deuxième type, les erreurs sont associées à des fuites ou à des dépôts solides dans le procédé (Bellec, 2004). Les erreurs systématiques sont plus importantes en magnitude, mais moins fréquentes que les erreurs aléatoires (Bazin et al., 1998; Romagnoli & Sanchez, 2000). Enfin, le défi n'est pas seulement d'identifier la présence de ces erreurs, mais aussi d'estimer leur ampleur afin qu'une correction puisse être appliquée et que le traitement des données puisse continuer jusqu'à ce que l'instrument soit recalibré et l'erreur corrigée (Bazin et al., 1998; Romagnoli & Sanchez, 2000).

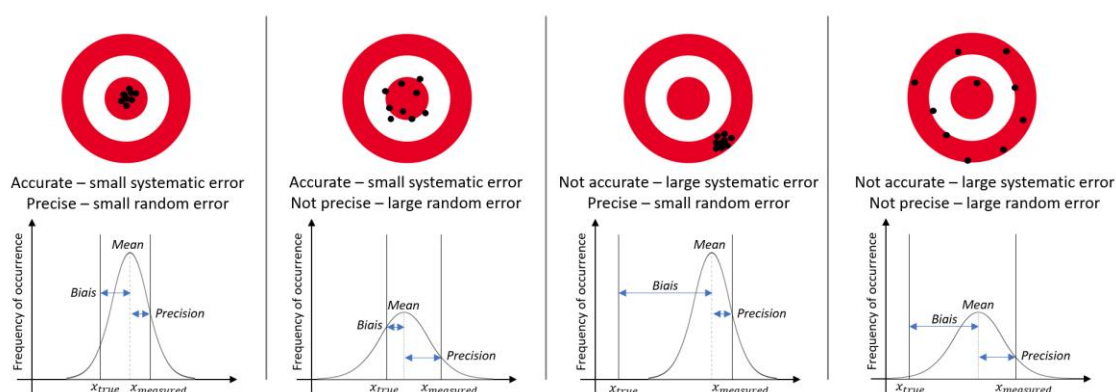


Figure 2.1 – Représentation de la précision (erreur aléatoire et valeurs aberrantes) et du biais (erreur systématique)

2.2 Retrait du bruit dans les données

Dans les industries de transformation, la plupart des données proviennent de capteurs générant des signaux susceptibles d'être dégradés par leur environnement, c'est-à-dire contaminés par du bruit à haute fréquence. Ce bruit peut masquer des caractéristiques et des événements ayant lieu au sein d'un procédé industriel. Ces événements se produisent à diverses fréquences (échelles) et sont localisés à différents points dans le temps. En d'autres termes, les variations inhérentes au procédé sont souvent multi-échelles avec différentes caractéristiques localisées dans le temps et en fréquence. Le processus de réduction de ce

bruit tout en conservant au maximum les informations de procédé sous-jacentes est une étape préliminaire importante et est appelé nettoyage, lissage, filtrage ou débruitage des données.

Il existe diverses méthodes de traitement des données permettant de retirer les composantes haute fréquence indésirables des signaux, telles que les filtres à réponse impulsionnelle finie (filtre à moyenne mobile, filtre polynomial) et ceux à réponse impulsionnelle infinie (filtre exponentiel, filtre de Kalman) (Narasimhan & Jordache, 2000). En plus des filtres, des méthodes d'analyse fréquentielle, notamment la transformée de Fourier à court terme et la transformée en ondelettes, peuvent être utilisées (Ganesan et al., 2004). Les sous-sections suivantes arborent quatre approches couramment utilisées : le filtre de Kalman, la moyenne mobile à pondération exponentielle, la transformée de Fourier à court terme et la transformée en ondelettes.

2.2.1 Filtre de Kalman

Le filtre de Kalman (FK) est un outil qui fournit une estimation fiable de l'état d'un procédé à partir d'observations bruitées grâce à un algorithme itératif de prédiction-correction. Les valeurs réelles de l'état sont calculées à l'aide de mesures entrantes et d'un modèle de procédé. Ce filtre est discret; il repose sur des échantillons de mesures prélevés à un rythme constant. Cet algorithme de filtrage non-supervisé ajuste les données actuelles de capteur en tenant compte des valeurs passées. De meilleurs résultats sont attendus avec le FK étant donné que comparativement à une méthode de filtrage basée uniquement sur les mesures entrantes, les données historiques sont également utilisées de manière récursive (Park, S. et al., 2019). Le principe de ce filtre récursif est de toujours corriger le modèle de procédé de telle sorte que l'erreur d'estimation soit minimisée. Autrement dit, il prédit les valeurs futures en combinant les observations avec les informations du modèle, en les comparant aux mesures réelles, en minimisant la différence entre l'état réel et l'état filtré et en ajustant les estimations (Harkat et al., 2016). Par conséquent, l'ensemble d'équations qui compose le filtre prend en charge les estimations des états passés, présents et futurs. Le FK est largement utilisé pour la réduction du bruit, la correction des données bruitées de capteurs, la prédiction des prochaines valeurs de séries de données temporelles et l'estimation de variables non-mesurées à partir de données mesurées (Park, S. et al., 2019). Il est souvent mis en œuvre dans les systèmes de contrôle, car pour contrôler un procédé, il faut d'abord une estimation précise des variables du procédé (Simon, 2001). Le filtre itère sur deux phases distinctes : la prédiction et la mise à jour. La dernière prédit l'estimation initiale de l'état et de la matrice de covariance. Elle projette dans le temps l'état actuel et les estimations de covariance ($t - 1$) pour obtenir des estimations *a priori* pour le prochain pas de temps t (Welch &

Bishop, 2006). L'estimation de la prochaine valeur à mesurer est basée sur les données historiques des capteurs, l'estimation de l'état précédent (Park, S. et al., 2019). La phase de prédiction est responsable de la rétroaction, elle ajuste (raffine) les estimations projetées obtenues à l'étape précédente par la mesure réelle à ce moment-là. Ainsi, cette phase incorpore les nouvelles mesures aux estimations *a priori* pour obtenir une meilleure estimation - l'*a posteriori* (Welch & Bishop, 2006). La phase de mise à jour calcule le gain de Kalman et génère une estimation de l'état *a posteriori* et une estimation de la covariance des erreurs. Ainsi, les phases de prédiction et de mise à jour sont répétées en utilisant les estimations *a posteriori* précédentes comme prédiction pour les nouvelles estimations *a priori* (Welch & Bishop, 2006).

Pour réduire le bruit d'un signal avec le filtre de Kalman, il faut pouvoir décrire ou approximer le procédé par un système linéaire qui est l'équation d'état. Le modèle de procédé est représenté par la matrice de transition d'état F_t , la matrice de contrôle B_t appliquée au vecteur de contrôle u_t et la matrice d'observation H_t qui multiplie la mesure entrante pour la convertir en un état. L'algorithme du filtrage de Kalman est résumé ci-dessous (Narayan et al., 2013).

FILTRE DE KALMAN

Entrée : Équations d'état linéaire et ensemble de mesures $y_1, y_2, \dots, y_N$

Sortie : Estimés des états $x_1, x_2, \dots, x_N$ aux pas de temps $n = 1, 2, \dots, N$ et covariance mise à jour

Pour les pas de temps $t = 1, 2, \dots, N$, exécuter les équations suivantes:

Équations de prédiction

1. Prédiction de l'état (estimé initial)	$\hat{x}_{t t-1} = F_t \hat{x}_{t-1 t-1} + B_t u_t$
2. Prédiction de la covariance , erreur du système (estimé initial)	$P_{t t-1} = F_t P_{t-1 t-1} F_t^T + Q_t$

Équations de mise à jour

3. Vecteur de mesure résiduel	$z_t = y_t - H_t \hat{x}_{t t-1}$
4. Matrice de variance résiduelle	$S_t = H_t P_{t t-1} H_t^T + R_t$
5. Gain de Kalman	$K_t = P_{t t-1} H_t^T S_t^{-1}$
6. Mise à jour de l'état , état actuel (nouvel estimé)	$\hat{x}_{t t} = \hat{x}_{t t-1} + K_t z_t$
7. Mise à jour de la variance , erreur du système (nouvel estimé)	$P_{t t} = (I - K_t H_t) P_{t t-1}$

La covariance des erreurs de mesure peut être déterminée à l'aide des données antérieures du capteur. Cette dernière varie pour chaque mesure puisqu'elle provient de l'appareil de mesure. Cependant, la détermination de l'erreur de procédé est généralement plus difficile, car il n'est pas possible de l'observer

directement et le procédé n'est pas toujours connu parfaitement (Welch & Bishop, 2006). Les performances de filtrage du FK varient en fonction des paramètres du bruit d'entrée (Park, S. et al., 2019). Finalement, le filtre de Kalman est connu comme un estimateur optimal dans la mesure où il minimise l'erreur quadratique moyenne des états estimés $E[(x_t - \hat{x}_{t|t})^2]$. Il fournit une estimation avec la plus petite variance d'erreur possible.

2.2.2 Moyenne mobile à pondération exponentielle

La moyenne mobile pondérée exponentiellement (MMPE) est un filtre passe-bas linéaire populaire classé comme un filtre à réponse impulsionnelle infinie qui atténue les composantes haute fréquence des signaux (Nounou & Bakshi, 2000). Il s'agit en fait de la forme de filtrage la plus utilisée dans les industries de procédés chimiques (Harrison & Stuart, 2011). Ce filtre passe-bas classique est l'un des types de filtres les plus élémentaires que l'on puisse mettre en œuvre. Comme le filtre de Kalman, ce filtre est mis en œuvre de manière récursive, c'est-à-dire qu'il lisse les données en faisant la moyenne de la valeur d'entrée actuelle avec toutes les entrées précédentes. Ainsi, la MMPE calcule la moyenne pondérée de la dernière valeur mesurée et de la précédente valeur lissée en fonction de leurs poids respectifs (Perry, 2010) :

$$S_t = \alpha x(t) + (1 - \alpha)S_{t-1}$$

où le coefficient α est le coefficient de lissage représentant le poids, $x(t)$ est la mesure d'entrée actuelle au pas de temps t et S_t est la valeur de sortie du filtre.

Le facteur alpha varie entre 0 et 1 et est sélectionné en fonction de la vitesse de changement de la moyenne du procédé; il peut être considéré comme la fréquence de coupure au-dessus de laquelle les caractéristiques à haute fréquence sont retirées (Perry, 2010). Ce coefficient indique dans quelle mesure l'entrée actuelle affecte le signal lissé. Un grand coefficient de lissage donnera plus de poids, plus d'importance aux nouvelles mesures entrantes (plus récentes), ce qui est approprié pour une variation rapide de la moyenne. Cela garantit que les caractéristiques de haute fréquence seront conservées dans le signal (Perry, 2010). Un petit paramètre de lissage mettra davantage l'accent sur les observations précédentes, sur la moyenne plutôt que sur l'observation actuelle, ce qui convient mieux à une variation moyenne moins fréquente. Cela permet d'abaisser la fréquence de coupure et de retirer les composantes à basse fréquence (Perry, 2010).

2.2.3 Transformée de Fourier à court terme

La transformée de Fourier (TF) est basée sur le principe que toute fonction variant dans le temps peut être divisée en signaux périodiques simples, et donc considérée comme une succession de portions périodiques – une somme de fonctions oscillantes simples. Cet algorithme décompose les signaux en différentes composantes de fréquence; il transforme un signal de son domaine temporel à son domaine fréquentiel. Les signaux sont multipliés par une série d'ondes sinusoïdales de différentes fréquences pour déterminer les fréquences présentes dans le signal. Les pics dans le spectre de fréquence indiquent les fréquences les plus présentes dans un signal – plus ils sont grands et nets, plus la fréquence est prévalente. La TF est largement utilisée, car le contenu en fréquence est d'une grande importance. Cette dernière a une haute résolution dans le domaine des fréquences, mais une résolution nulle dans le domaine temporel; elle indique quelles fréquences sont présentes dans un signal, mais pas le moment où ces fréquences sont apparues. Les transformées de Fourier décomposent le signal en composantes de fréquence, mais l'information temporelle est perdue.

Ainsi, les résultats de la transformation de Fourier sont décents dans le cas de signaux stationnaires, c'est-à-dire lorsque le spectre de fréquences (contenu fréquentiel) ne change pas dans le temps – les fréquences présentes dans le signal ne dépendent pas du temps, mais présentent un inconvénient majeur pour les signaux non-stationnaires. En effet, en transformant les signaux dans le domaine des fréquences, l'information temporelle est perdue. Plus un signal est dynamique ou non-stationnaire, plus les résultats seront mauvais. Ainsi, en regardant la TF d'un signal non-stationnaire, il n'est pas possible de dire quand un événement particulier s'est produit. Sachant que la plupart des signaux de procédé contiennent de nombreuses caractéristiques non-stationnaires ou transitoires telles que des dérives, changements brusques, variations d'état, etc., qui sont souvent l'aspect le plus important d'un signal, cette analyse n'est pas adaptée pour les détecter et les localiser dans le temps.

Pour surmonter cette limitation et obtenir à la fois l'information fréquentielle et temporelle, il est possible de diviser le temps total en plusieurs intervalles très courts et de prendre la transformée de chacun; il s'agit de la transformée de Fourier à court terme (TFCT) (Abramovich, 1998; Fugal, 2009). La TFCT est un compromis entre les informations temporelles et fréquentielles; elle fournit des informations sur le moment et la fréquence d'un événement du signal. Elle partitionne le signal non-stationnaire en segments temporels de longueur égale en multipliant le signal discret par une fonction de fenêtre glissante d'analyse et pour chaque temps différent, un spectre différent est obtenu. La TFCT applique séparément des

transformées de Fourier rapide (TFR) conventionnelles à ces segments, qui peuvent être considérés stationnaires (Patil, 2015). Par conséquent, la TFCT est utilisée pour obtenir d'un signal le spectre variable dans le temps; c'est une version de la TFR comprenant une variante temporelle. En ce qui concerne la fenêtre d'analyse, une grande fenêtre fournira une excellente résolution dans le domaine des fréquences, mais une mauvaise résolution dans le domaine temporel, et vice-versa (Patil, 2015). Le problème des fenêtres de la TFCT est qu'elles se heurtent aux limites théoriques du principe d'incertitude : plus la taille de la fenêtre est petite, plus il est possible d'en savoir sur l'endroit où une fréquence est apparue dans le signal, mais moins sur la valeur de la fréquence elle-même, alors que plus la taille de la fenêtre est grande, plus nous en saurons sur la valeur de la fréquence et moins sur le temps.

La TFCT réduit le bruit en trois étapes : **(1)** calcul de la TFCT du signal bruité, **(2)** seuillage de la TFCT, et **(3)** calcul de la TFCT inverse. Pour la première étape, il faut sélectionner une fonction de fenêtre d'analyse et multiplier le signal par cette fenêtre. Une bonne fonction doit équilibrer la localisation temporelle et fréquentielle. Quelques exemples de fonctions sont la fenêtre *block* qui a une transition nette et donc une mauvaise localisation de la fréquence, la fenêtre Hamming qui est plus lisse et la fenêtre Hanning qui a des dérivés continues. Une fois la taille de la fenêtre sélectionnée, la résolution en fréquence est fixée pour l'ensemble du spectre. La transformée de Fourier convertit les signaux dont le contenu fréquentiel change au fil du temps du domaine temporel en un domaine temps-fréquence. Pour obtenir cela, le signal est divisé en sections plus petites. Ces segments sont pris dans le domaine de fréquence en utilisant la TFR. En sélectionnant des fenêtres successives se chevauchant, c'est-à-dire que pour un chevauchement de 50 %, les fenêtres sont décalées de la moitié de la longueur de la fenêtre, et en répétant cette procédure jusqu'à la fin du signal, on obtient une représentation temps-fréquence du signal d'entrée.

La deuxième étape consiste à appliquer des filtres sur les données du domaine fréquentiel pour retirer les caractéristiques indésirables comme le bruit. Toutes les valeurs inférieures au seuil en valeur absolue sont mises à zéro. Le choix du seuil est quelque peu délicat; si le seuil est trop grand, le signal peut être déformé et s'il est trop petit, le résultat peut encore contenir du bruit (Selesnick, 2009). Le seuil peut également être une fonction du temps ou de la fréquence.

Enfin, la troisième étape consiste à appliquer une transformée de Fourier à court terme inverse sur les données du domaine fréquentiel pour obtenir des données du domaine temporel. La transformée de Fourier à court terme inverse (TFCTI) retourne au domaine temporel avec une fenêtre de synthèse en utilisant un ajustement par la méthode des moindres carrés pour produire un signal traité qui correspond autant que

possible aux données d'entrée. La plupart des données d'une TFCT sont redondantes; la TFCT produit plus de données qu'il n'y en a dans le signal original. Par conséquent, l'opération inverse (TFCTI) doit retirer les données supplémentaires. Autrement, il est peu probable que le signal traité corresponde au signal initial. La TFCTI reconstruit un signal à partir de sa représentation dans le domaine fréquentiel. Pour permettre l'inversion d'une TFCT, le fenêtrage du signal doit obéir à l'une des contraintes existantes telles que la procédure OLA (*overlap-add*), COLA (constant), WOLA (pondéré), NOLA (non nul), etc.

2.2.4 Transformée en ondelettes

La transformée en ondelettes (TO) est un outil mathématique puissant, car elle permet de mettre en évidence des informations contenues dans les signaux et elle est adaptée à l'analyse de signaux non-stationnaires présentant des caractéristiques multi-résolution. La TO décompose un signal en ses éléments constitutifs dans le domaine temps-fréquence en utilisant une version dilatée (mise à l'échelle) et translatée (décalée) de l'ondelette $\psi(t)$, qui sont des fonctions caractérisées par une moyenne nulle, c'est-à-dire :

$$\int_{-\infty}^{+\infty} \psi(t) dt = 0$$

Ces dernières sont irrégulières, souvent asymétriques et utilisent la dilatation s et le décalage τ pour faire varier leur résolution temps-fréquence.

La fonction ondelette ou fenêtre est caractérisée par le paramètre d'échelle s qui est lié à la fréquence du signal (f^{-1}) et le paramètre de translation τ qui indique la position temporelle de la fenêtre (Polikar, NA). Le facteur d'échelle correspond à la taille de la fenêtre. Une échelle élevée, c'est-à-dire une ondelette dilatée/étirée, analyse les composantes à basse fréquence des signaux, tandis qu'une échelle plus petite, c'est-à-dire une ondelette comprimée, évalue les composantes à haute fréquence et trouve les changements rapides dans le signal. La fonction ondelette est définie mathématiquement comme suit :

$$\psi_{\tau,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right)$$

La transformée en ondelettes continue (TOC) d'un signal $x(t)$ est donnée par le produit scalaire suivant :

$$Wx(\tau, s) = \langle x(t), \psi_{\tau,s}(t) \rangle = \int_{-\infty}^{+\infty} x(t) \psi_{\tau,s}(t) dt$$

Cette TOC est obtenue en mettant à l'échelle la fonction ondelette, en la décalant dans le temps et en la convoluant avec un signal. La transformée en ondelettes analyse les signaux non-stationnaires en les comparant à une variété d'ondelettes étirées et décalées (Fugal, 2009). Cette étape est appelée mise à l'échelle. Elle implique la superposition des événements cachés dans le signal en fonction de leur fréquence et de leur emplacement dans le temps. La fonction ondelette peut être étirée à la même fréquence que l'événement et décalée dans le temps pour s'aligner sur lui. Plus l'ondelette est étirée, plus la fréquence est basse. Ainsi, l'information sur le temps et la fréquence de l'événement est donnée en identifiant la translation et la dilatation ou la compression de l'ondelette.

La fonction ondelette est considérée comme un filtre passe-haut; par conséquent, la sortie de la TOC est constituée de coefficients correspondant aux détails – composantes haute fréquence – intégrés dans le signal d'entrée. La fonction ondelette est, la plupart du temps, associée à une fonction d'échelle $\Phi_{j,i}$. Cette dernière fournit les coefficients d'approximation; elle est complémentaire et assez similaire à l'équation (1) et elle est considérée comme un filtre passe-bas (Mallat, S. G., 2009).

$$\int_{-\infty}^{+\infty} \Phi(t) dt \neq 0$$

La transformée en ondelettes dyadiques discrètes (TOD) est presque identique à la TOC à l'exception que la TOD utilise des valeurs discrètes pour le facteur d'échelle et de translation. Le paramètre d'échelle est discrétisé pour permettre des implémentations numériques rapides; il est échantillonné suivant une séquence dyadique 2^j , mais pas le paramètre de translation (Fugal, 2009; Mallat, S. & Zhong, 1992). De son côté, la TOC peut fonctionner à n'importe quelle échelle. Ainsi, ce qui distingue la TOC de la TOD, ce sont les échelles auxquelles elles opèrent. Dans la plupart des cas, l'ondelette ne nécessite pas une valeur de corrélation pour chaque décalage de chaque ondelette étirée comme c'est le cas pour la TOC – il s'agit d'une représentation hautement redondante. En outre, la TOD permet de manipuler les données avec un seuil, puis de reconstruire le signal en calculant l'inverse de la transformée. Ce dernier point n'est pas possible avec la TOC qui ne dispose pas d'une transformée inverse viable et dont la reconstruction serait une opération numérique beaucoup moins stable (Fugal, 2009; MathWorks, 2019). Ainsi, la TOD offre une analyse plus efficace, elle supprime la redondance de la TOC, tout en offrant la propriété de réversibilité. La TOD n'est discrète que dans le domaine de l'échelle et de la translation, et non dans le domaine temporel. Par conséquent, pour un signal discret, il faut utiliser la transformée en ondelettes en temps discret ou la transformée en ondelettes continues en temps discret.

La TOD est définie par

$$Wx(\tau, 2^j) = \langle x(t), \psi_{\tau, 2^j}(t) \rangle = \frac{1}{\sqrt{2^j}} \int_{-\infty}^{+\infty} x(t) \psi \left(\frac{t - \tau}{2^j} \right) dt$$

Dans le domaine discret, les ondelettes continues et la fonction de mise à l'échelle sont remplacées respectivement par des filtres discrets passe-haut $g[n]$ et passe-bas $h[n]$ appelés filtres miroirs conjugués. Ces filtres passe-bas et passe-haut de décomposition (h et g), ainsi que leurs filtres de reconstruction associés (h' et g') forment les filtres miroirs en quadrature.

$$g[n] = \left\langle \frac{1}{\sqrt{2}} \psi \left(\frac{t}{2} \right), \phi(t - n) \right\rangle \text{ et } h[n] = \left\langle \frac{1}{\sqrt{2}} \phi \left(\frac{t}{2} \right), \phi(t - n) \right\rangle$$

$$g[n] = (-1)^{1-n} h[1 - n]$$

Comme l'ondelette et la fonction d'échelle, les filtres passe-haut et passe-bas sont liés et correspondent respectivement aux détails et à l'approximation du signal. Les coefficients d'approximation (cA) et les coefficients de détail (cD) du niveau i sont décrits comme suit (Alyasseri, Z. A. A. et al., 2020) :

$$cA_i(t) = \sum_{k=-\infty}^{\infty} cA_{i-1}(\tau) \phi_i(t - \tau) \text{ et } cD_i(t) = \sum_{k=-\infty}^{\infty} cD_{i-1}(\tau) \psi_i(t - \tau)$$

Ainsi, effectuer la TOD équivaut à filtrer un signal avec deux filtres, un filtre passe-haut pour analyser les composantes de haute fréquence (détails) et un filtre passe-bas pour les composantes de basse fréquence (approximation).

En pratique, les mesures des capteurs contiennent des caractéristiques présentes dans plus d'une gamme de fréquences. Ces caractéristiques sont causées par des perturbations ou sont inhérentes au procédé. Afin de représenter correctement ces caractéristiques, les ondelettes peuvent être utilisées. L'idée principale des ondelettes est de représenter les mesures dans une séquence d'approximations successives du signal appelées niveaux. La TO décompose les données en plusieurs niveaux différents possédant une gamme de fréquences spécifiques (Figure 2.2) et permet de bien localiser les événements dans le domaine temps-fréquence et les caractériser (Mallat, S. & Hwang, 1992; Mallat, S. & Zhong, 1992). En effet, les niveaux sont utiles pour représenter les caractéristiques du signal, car ils ne contiennent que les mesures présentes sous une fréquence spécifique. La TO permet de représenter le signal $f(t)$ par une somme d'approximations (fonction d'échelle $\Phi_{j,i}$) qui forment le signal lissé et une somme de détails (bruit) représentés par des fonctions d'ondelettes $\psi_{j,k}$, selon différents niveaux. Ces derniers sont utilisés pour afficher les caractéristiques du signal dans le temps et à différentes fréquences (Bakshi, Bhavik R., 1999).

Le signal est une combinaison linéaire d'une fonction d'échelle et d'une ondelette mise à l'échelle et translatée en temps et en fréquence :

$$f(t) = \sum_{i \in I_j} c_{j,i} \Phi_{j,i} + \left(\sum_{j=1}^J \sum_{k \in K_L} d_{j,k} \psi_{j,k} \right)$$

Le premier terme correspond au signal lisse (approximation) qui contient les éléments de basse fréquence du signal original présent au niveau j (Figure 2.2, gauche), tandis que le second terme représente le signal détaillé contenant les éléments de haute fréquence (bruit) du signal original présent à ce niveau (Figure 2.2, droite) (Jiang, Taiwen, Chen, He, et al., 2003; Korbel, Milan et al., 2014). La différence entre le signal original et la première approximation a_1 est le détail d_1 , la différence entre a_1 et a_2 est le détail d_2 . Ainsi, la différence entre les signaux lisses de deux niveaux successifs permet d'identifier la contribution des composantes haute fréquence (Bakshi, B. R. & Stephanopoulos, 1994).

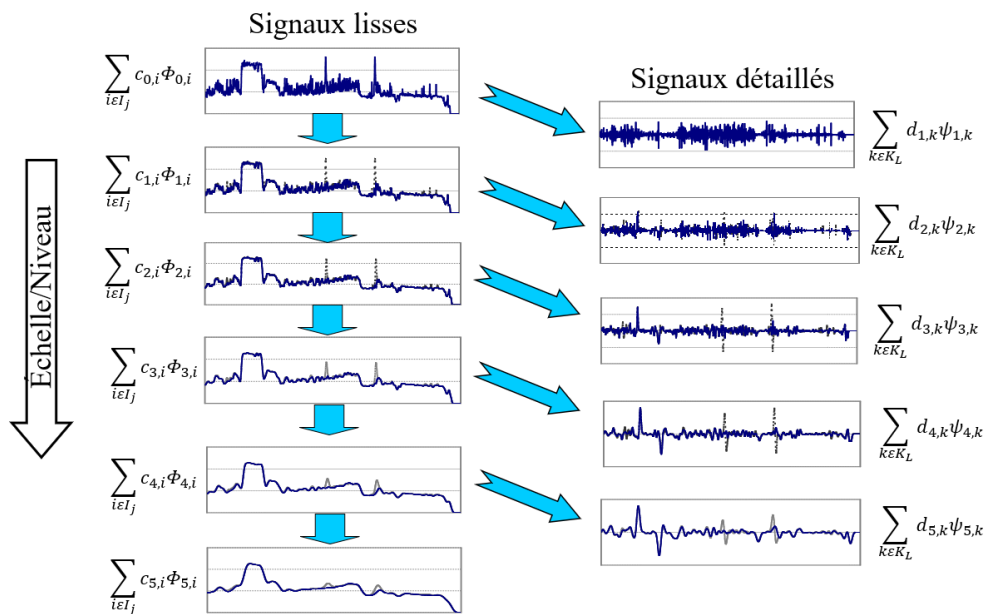


Figure 2.2 – Représentation en ondelettes multi-échelles de mesures en temps réel, inspiré de (Korbel, Milan et al., 2014)

La représentation en ondelettes multi-échelles (Figure 2.2) permet d'examiner sélectivement les composantes du signal à chaque échelle et d'identifier de manière adéquate les caractéristiques des signaux localisées à différentes fréquences et de sélectionner l'échelle retenue en fonction de l'objectif, c'est-à-dire la compression, le lissage, le filtrage, la détection de caractéristiques, etc. (Nounou & Bakshi, 1999).

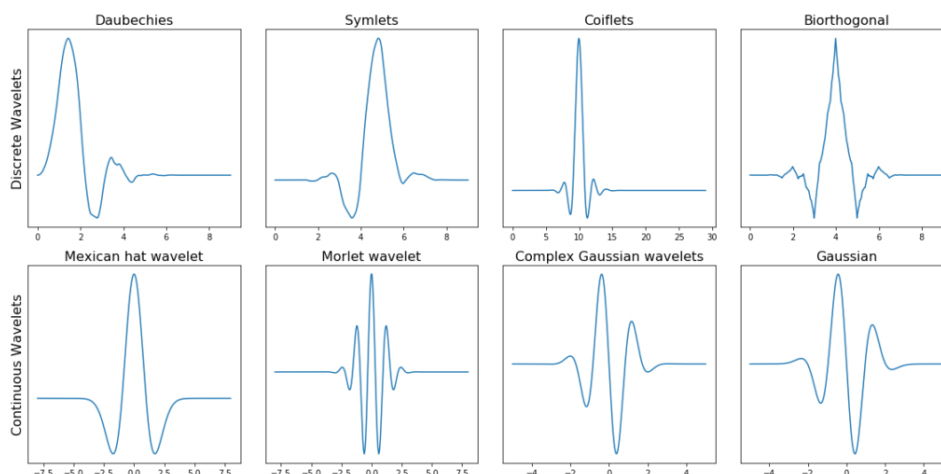


Figure 2.3 – Exemples de familles d'ondelettes (Taspinar, 2018)

Il existe plusieurs familles de TO qui présentent un compromis entre la forme, le lissage et la compacité et, selon son adéquation avec les caractéristiques du signal, l'une d'entre elles peut être préférée à une autre. Parmi les propriétés souhaitables des fonctions d'ondelettes, citons de bonnes localisations temps-fréquence, divers degrés de lissage, c'est-à-dire le nombre de dérivées continues et un grand nombre de *vanishing moments* (Ganesan et al., 2004). Chaque fonction est associée à un chiffre qui fait référence au nombre de coefficients – les ondelettes ont un certain nombre de *vanishing moments* égal à la moitié du nombre de coefficients. Plus le nombre de coefficients augmente, plus les fonctions deviennent lisses. Les ondelettes peuvent être orthogonales ou non-orthogonales, bi-orthogonales ou non, symétriques ou pas, complexes ou réelles. Quelques exemples de diverses familles d'ondelettes sont présentés à la Figure 2.3. Ces différentes familles ont des propriétés mathématiques distinctes et conduisent à des décompositions de signaux différentes (Ranchin & Wald, 1992). L'ondelette Haar est la plus simple. Il s'agit d'une fonction orthogonale discontinue, également connue sous le nom de Daubechies db1. La fonction Haar n'a pas une bonne localisation temps-fréquence (Daubechies, 1992a). Les Daubechies sont des ondelettes orthogonales à support compact, elles sont largement les plus appliquées. Les ondelettes biorthogonales ont la propriété d'avoir deux ondelettes, une pour la décomposition et une autre pour la reconstruction, au lieu d'utiliser la même ondelette. La fonction ondelette Coiflets a $2N$ (N étant le niveau) moments égaux à 0 alors que la fonction d'échelle en a $2N-1$. Toutes deux ont un support de longueur $6N-1$. Les ondelettes Symlets sont une modification des ondelettes Daubechies, elles sont presque symétriques. Ces deux familles ont des propriétés similaires. L'ondelette Morlet n'a pas de fonction d'échelle. Le chapeau mexicain n'a pas non plus de fonction d'échelle et est découlé d'une fonction qui est proportionnelle à la seconde dérivée de la fonction de densité de probabilité gaussienne. En somme, certaines ondelettes

opèrent exclusivement dans le domaine continu, comme les ondelettes de chapeau mexicain, Morlet, Gaussienne, Meyer, Shannon, B-Spline à fréquence complexe, Morlet complexe et Gaussienne complexe (Figure 2.3) (Fugal, 2009).

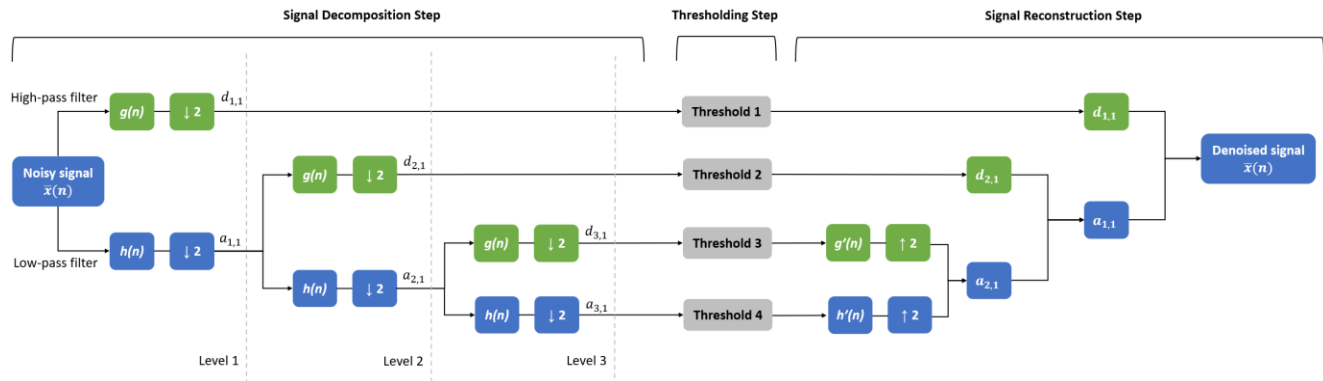


Figure 2.4 – Décomposition et synthèse du signal à trois échelles en utilisant des transformées en ondelettes

La Figure 2.4 présente une décomposition de signal multi-échelle. La TOD décompose un signal brut contenant du bruit en trois niveaux en utilisant des filtres passe-haut et passe-bas qui donnent respectivement le coefficient de détail (cD) et le coefficient d'approximation (cA) à chaque niveau. Cette opération conduit à deux fois plus de données que ce que contenait le signal d'entrée. Pour corriger cela, les coefficients doivent être sous-échantillonnés de moitié ($\downarrow 2$). La composante basse fréquence (cA) est à nouveau divisée pour obtenir d'autres détails et une approximation du signal d'entrée dans les niveaux suivants (Pérez & Barros, 2012). Ce processus est répété pour le nombre souhaité de niveaux (Moridani & Pouladian, 2009). La sortie de ces filtres représente les détails contenus dans les données (les composantes haute fréquence) et l'approximation (les composantes basse fréquence). Ces dernières sont souvent les plus importantes, car elles représentent la forme globale d'un signal, sa tendance. Ensuite, certains coefficients sont réduits en utilisant un seuil choisi et le processus de reconstruction du signal lissé commence en utilisant des coefficients suréchantillonnés ($\uparrow 2$) – le suréchantillonnage est le processus d'allongement du signal en insérant des zéros entre chaque échantillon. Ces coefficients sont convolués avec un filtre de reconstruction passe-bas pour l'approximation et un filtre de reconstruction passe-haut pour les détails.

Ainsi, les erreurs aléatoires peuvent être réduites en trois étapes. **(1)** La décomposition du signal brut en coefficients d'approximation et de détail, ou signal lissé (composante basse fréquence) et signal détaillé (composante haute fréquence), en utilisant respectivement des filtres passe-bas et passe-haut déterminés

par une fonction ondelette et une fonction d'échelle correspondante. Pour ce faire, une fonction d'ondelette appropriée pour l'analyse du signal d'intérêt est sélectionnée (l'ondelette doit avoir des caractéristiques et des traits qui correspondent à ceux à détecter et à extraire du signal d'entrée). En outre, un niveau de décomposition garantissant la réduction des composantes haute fréquence, mais qui n'affecte pas les tendances, est sélectionné. **(2)** Une fois les coefficients obtenus, diverses modifications des coefficients de détail peuvent être entreprises en fonction du but, i.e., débruitage, compression, extraction de caractéristiques, analyse des tendances globales et localisées, etc. Ainsi, pour la réduction du bruit, une valeur seuil est définie et appliquée aux coefficients de détail. **(3)** La reconstruction du signal débruité se fait en calculant la TOD inverse en utilisant les coefficients d'approximation du niveau sélectionné et les coefficients de détail modifiés de chaque niveau (Donoho, David L. et al., 1995; Flehmig et al., 1998; Nounou & Bakshi, 1999). Ces étapes permettent d'obtenir un signal lisse qui représente avec précision les opérations d'un procédé.

2.2.5 Comparaison des approches

Toutes les approches détaillées aux sous-sections 2.2.1 à 2.2.4 ne présentent pas nécessairement la même efficacité, surtout lorsqu'il s'agit de filtrer des signaux de capteurs de procédé qui présentent une large gamme de composantes de fréquence. Cette section présente les avantages et les inconvénients de chaque technique.

Tout d'abord, le filtre MMPE est largement utilisé dans les industries de transformation. Un aspect positif de cette approche est qu'elle est facile à mettre en œuvre et qu'elle nécessite un faible effort de calcul (mémoire réduite et calcul rapide). En revanche, une connaissance suffisante des données du procédé est nécessaire pour définir le paramètre α – le poids de la dernière valeur entrée dans l'ensemble de données. Certains tests peuvent être nécessaires pour trouver la valeur la plus appropriée en fonction de l'application.

Deuxièmement, le filtre de Kalman est théoriquement un estimateur optimal; il peut se corriger entre la phase de prédiction et de mise à jour. Il estimera avec précision les états affectés par une perturbation externe. Un avantage majeur du filtre de Kalman est sa capacité à fournir à l'estimation un indicateur de la précision du filtre, c'est-à-dire la covariance des erreurs de prédiction (Harkat et al., 2016). De tous les filtres existants, le filtre de Kalman est le seul qui minimise la covariance de l'erreur d'estimation (Simon, 2001). D'autre part, ce filtre nécessite la connaissance de la variance du bruit du procédé ainsi que de la

variance du bruit des observations (mesures). La sélection de ces valeurs nécessite une grande connaissance du procédé et des capteurs (temps de réaction des capteurs, largeur de bande des capteurs).

De plus, afin d'utiliser le FK, la matrice de transition d'état doit être établie. Pour ce faire, un modèle parfait est rarement disponible, surtout dans les industries de transformation. Les équations qui composent le modèle (bilans de masse et d'énergie) ne sont jamais vraiment exactes en raison des nombreuses hypothèses sous-jacentes. Une revue de la littérature a démontré que personne n'a utilisé le FK comme réducteur de bruit pour les lectures des capteurs de procédé. Ce qui s'en rapproche le plus, ce sont les publications relatives à l'estimation d'états non-mesurables ou non-mesurés – un des principaux objectifs du filtre – pour des réacteurs basés sur un modèle (MacGregor et al., 1986).

Les filtres linéaires classiques sont efficaces pour réduire les erreurs aléatoires au-dessus ou au-dessous d'une certaine fréquence (hors bande). Il s'agit de méthodes simples, faciles à utiliser et à mettre en œuvre. Cependant, leur action est limitée pour les signaux *in-band* puisque les données ne sont représentées que pour un seul niveau de fréquence. En d'autres termes, ces filtres sont à échelle unique par nature, car les fonctions de base ont une localisation temps-fréquence fixe (Nounou & Bakshi, 2000). En effet, lorsqu'ils sont appliqués à des signaux *in-band*, ils suppriment également le signal d'intérêt. Par conséquent, les filtres linéaires doivent trouver un équilibre entre la suppression efficace de l'erreur aléatoire globale dans le temps et la représentation précise des caractéristiques localisées dans le temps conservées dans le signal (Nounou & Bakshi, 2000). Les données de mesure des capteurs sont bien connues pour être non-stationnaires et multi-échelles étant donné la variété des événements qui se produisent dans un procédé, et sont donc moins adaptées à une analyse à échelle unique.

Ce problème est partiellement résolu par la transformation de Fourier qui permet l'analyse fréquentielle des données. En revanche, elle est inefficace pour extraire des informations temporelles des caractéristiques des signaux (Fugal, 2009). De plus, quelques méthodes spectrales reposent sur l'hypothèse fondamentale implicite que les signaux sont périodiques et stationnaires, ce qui n'est pas le cas des signaux provenant des capteurs de procédé (Ganesan et al., 2004). En effet, la TF n'est pas adaptée à l'analyse des signaux non-stationnaires et notamment ceux qui contiennent des changements soudains. Pour les signaux non-stationnaires, la TF prend la moyenne des composantes de fréquence sur toute la durée du signal (Ganesan et al., 2004). Cependant, ces inconvénients peuvent être surmontés avec la TFCT qui est adéquate pour les signaux non-stationnaires. La TFCT, en plus d'une bonne résolution

fréquentielle, permet d'évaluer à quel moment un événement fréquentiel particulier s'est produit dans le signal.

Toutefois, la résolution de la TFCT reste constante même si le temps change, c'est-à-dire que la TFCT utilise une fenêtre d'analyse unique et fixe, ce qui signifie que la largeur de la fenêtre temporelle et fréquentielle est constante tout au long du processus (Figure 2.5). La résolution temps-fréquence est donc assez limitée, principalement en raison de l'utilisation d'une taille de fenêtre fixe; une fois fixée, cette fenêtre est la même pour toutes les fréquences. Il est difficile de savoir quelle taille serait adéquate pour le signal; une fenêtre large aurait une bonne résolution en fréquence, mais une mauvaise résolution temporelle et vice versa pour une fenêtre trop étroite. Cette dernière n'est pas adéquate pour les événements de basse fréquence et la première n'est pas nécessaire pour les composantes de haute fréquence. Ainsi, la précision de la méthode de filtrage est fortement limitée par la taille de la fenêtre. Par conséquent, la TFCT n'est pas aussi bien adaptée aux signaux avec une gamme de fréquences et pour localiser les singularités ou les événements de courte durée de haute fréquence (Fugal, 2009; Ganesan et al., 2004). Cet aspect rend difficile l'obtention de caractéristiques appropriées à partir des signaux (Park, S.-M. et al., 2016). Ainsi, un inconvénient majeur de la TFCT est le compromis de résolution entre le temps et la fréquence.

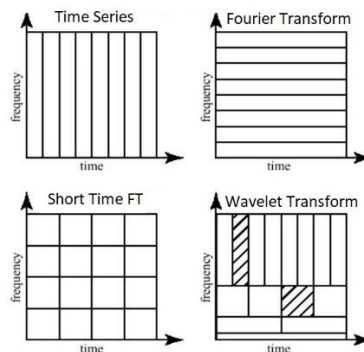


Figure 2.5 – Fenêtre unique et fixe de la TFCT et celles variables de la TO (Ganesan et al., 2004)

Avec la TO, la notion d'échelle est introduite comme alternative à la fréquence; le facteur d'échelle est défini comme l'inverse de la fréquence. Contrairement à la TFCT, qui utilise une seule fenêtre d'analyse, la TO utilise des fenêtres de tailles variables qui s'adaptent pour représenter les caractéristiques importantes du signal : des fenêtres étroites pour les hautes fréquences et des fenêtres plus larges pour les basses fréquences (Figure 2.5). De ce fait, la TO atteint un équilibre idéal de résolution échelle-temps (elle a plus de flexibilité dans la résolution temporelle et d'échelle). En outre, par rapport à la TFCT, elle est plus apte à révéler les changements ou les caractéristiques à court terme des signaux se produisant à haute fréquence.

La transformation en ondelettes est connue pour être la technique la plus courante, puissante et efficace pour le débruitage des signaux non stationnaires (Alyasseri, Z. A. A. et al., 2020). La TOD est la méthode de filtrage temps-fréquence la plus généralement utilisée (Ben Slama et al., 2018). En examinant un signal à différentes échelles et en l'analysant avec différentes résolutions (étant donné ses capacités multi-résolution inhérentes), la TO devient très efficace pour détecter et représenter les éléments des signaux (bruit, anomalies, caractéristiques et tendances). Cette approche permet de supprimer les caractéristiques indésirables qui se superposent au signal d'intérêt grâce à la décomposition de la fréquence et à l'utilisation d'un seuillage statistique, sans l'affecter. En d'autres termes, la mise en œuvre de la transformée en ondelettes entraîne une distorsion minimale des données filtrées (elle est capable de décomposer des signaux complexes puis de les reconstruire sans perdre d'informations pertinentes pour le procédé) et permet d'extraire des caractéristiques qui sont associées uniquement à la dynamique du capteur (Upadhyaya et al., 2014).

Cette méthode fournit d'excellentes informations localisées en temps et en fréquence et a la capacité de fournir une représentation multi-résolution d'un signal pour l'examen d'événements survenant à plusieurs échelles. Elle constitue également un excellent moyen d'analyser des données non-gaussiennes non-stationnaires et auto-corrélées puisqu'elle génère des coefficients d'ondelettes gaussiens stationnaires et distribués de manière indépendante et identique, quelle que soit la distribution des données d'entrée. Par conséquent, cette analyse permet d'appliquer aux données décomposées par ondelettes les procédures statistiques qui nécessitent des données distribuées de manière gaussienne.

En outre, les signaux présentant des changements brusques peuvent être mieux analysés avec une ondelette irrégulière qu'avec une sinusoïde lisse fournie par la TFCT. Les ondelettes sont utilisées dans différents domaines d'application, car elles peuvent traiter des signaux de formes particulières grâce aux différentes familles d'ondelettes. En effet, la TFCT décompose les signaux en sinus et en cosinus alors que la TO offre de nombreuses fonctions de base différentes pour la décomposition qui ont des formes variées (Figure 2.3). Elles peuvent donc s'adapter à diverses caractéristiques du signal, ce qui donne à la TO la capacité de débruiter des signaux particuliers.

Finalement, une sélection appropriée des paramètres des méthodes de traitement des données est essentielle afin de réduire le bruit de façon optimale. Ainsi, la section suivante aborde comment ces paramètres sont choisis dans le cas particulier de la transformée en ondelette.

2.2.6 Sélection des paramètres de la transformée d'ondelette

Le traitement des signaux à l'aide de la transformée en ondelettes requiert la sélection de divers paramètres, notamment le type de fonction d'ondelette et son ordre (*vanishing moment*), le niveau de décomposition, la méthode de débruitage et la règle du seuil, car ils affectent fortement les performances de débruitage. À la connaissance des auteurs, il n'existe pas de règle universelle pour la sélection de ces paramètres puisqu'ils dépendent du problème (Katunin & Przysławka, 2015). Bien que l'efficacité de la réduction du bruit dépende grandement de la sélection des paramètres des ondelettes, la détermination de ces paramètres peut être ambiguë et peu pratique. En effet, compte tenu du nombre de paramètres, les TO sont plus difficiles à mettre en œuvre que les autres filtres, lourds à régler et à maintenir. Ceci est particulièrement vrai pour les signaux industriels, car ils ne sont pas périodiques, ils sont sujets à des valeurs aberrantes, des arrêts et des événements anormaux. En analysant l'état actuel de la littérature, il est possible de noter qu'il n'existe aucune recommandation sur la fonction ondelette à utiliser pour analyser les signaux de procédé industriel, sur le niveau de décomposition auquel cette analyse doit être effectuée et sur la règle de seuillage à appliquer. Les paragraphes suivants expliquent comment les paramètres des ondelettes sont réglés dans diverses applications.

Tous domaines confondus, donc pas uniquement liés au génie chimique et aux procédés industriels, quelques articles ont utilisé des algorithmes (principalement l'algorithme génétique) pour sélectionner les paramètres optimaux des ondelettes afin de maximiser les performances de filtrage (Abbaszadeh & Alipour, 2017; Cao et al., 2022; El-Dahshan, 2011; Hasan et al., 2011; Katunin & Przysławka, 2015), d'autres s'appuient sur des méthodes empiriques pour sélectionner les paramètres des ondelettes (Liu, W. et al., 2011; Shafri & Yusof, 2009; Zhou et al., 2009), tandis que d'autres encore utilisent une approche par essais et erreurs (Tan et al., 2008) ou leur expérience personnelle (Smallwood & Cap, 1998). Dans Wang, P. et al. (2021), les auteurs ont déterminé les paramètres optimaux de la TO en comparant le coefficient de corrélation de Pearson entre le spectre de réponse au choc corrigé et un spectre de référence à basse fréquence à différents niveaux de décomposition et *vanishing moment*. Il est constaté que Db7 avec un niveau de décomposition à sept est la combinaison optimale pour traiter les signaux d'essai de choc, car le coefficient de corrélation atteint 0,9995.

Hasan et al. (2011) présente un algorithme de débruitage des données d'un système de positionnement global et d'un système de navigation inertielle. La sélection des paramètres d'ondelettes est vitale pour le débruitage de ce type de données. Hasan et al. (2011) ont testé une gamme de paramètres d'ondelettes

dans leur algorithme d'optimisation : les types de fonction d'ondelettes – Daubechies (Db2-Db10), Coiflets (Coif1-Coif5), Symlets (Sym2-Sym8) et Biortogonal (Bior1.3-Bior6.8), la règle de sélection du seuil – *minimax*, *heursure*, *rigrsure*, *universal*, et le niveau de décomposition – 1 à 10. En mettant en œuvre l'algorithme proposé, les paramètres suivants sont identifiés comme les paramètres optimaux pour la position dans les axes x, y et z : fonction d'ondelette Db3, Db4, Sym4, et Bior4.4, seuil *universal* et *minimax* et 5-6 comme niveau de décomposition.

La sélection des paramètres d'ondelettes a un impact direct sur la performance du débruitage, c'est pourquoi certains articles ont étudié les paramètres idéaux pour des applications spécifiques. Dans le cas des applications hyperspectrales, Zhou et al. (2009) ont déterminé que les paramètres de TO optimaux pour le spectre de végétation (gazon) provenant de la bibliothèque de spectres de *United States Geological Survey* étaient la fonction ondelette sym6, la règle de sélection du seuil *heursure* et la décomposition à sept niveaux. Liu, W. et al. (2011) ont montré que la fonction ondelette sym8, la méthode de remise à l'échelle du bruit à un seul niveau, la règle de sélection du seuil *heursure* et la décomposition à trois niveaux produisent les meilleures performances de débruitage pour le spectre de la matière organique (sol). En utilisant des signaux synthétiques, Shafri et Yusof (2009) ont indiqué que les paramètres d'ondelettes les plus appropriés pour débruiter le signal de réflectance de l'huile de palme sont la fonction ondelette sym8 ou db9 (après avoir testé db4-30, sym4-15 et coif1-5) et le facteur de remise à l'échelle du seuil de bruit à un seul niveau lors de l'utilisation de la règle de sélection du seuil *universal*, du seuillage doux et de la décomposition à cinq niveaux. Li et al. (2012) ont recherché les paramètres (règle de sélection du seuil, remise à l'échelle du seuil, niveau de décomposition et type d'ondelette) qui donneraient de bons résultats pour n'importe quel spectre de la bibliothèque en utilisant un critère d'angle spectral normalisé.

Dans les applications de la TO à la modélisation des procédés hydrologiques basés sur l'apprentissage machine, l'identification d'un ensemble optimal de paramètres d'ondelettes, tels que le type de fonction d'ondelette, le *vanishing moment* et le niveau de décomposition maximal, est généralement effectuée en utilisant une approche par essais et erreurs (Abbaszadeh & Alipour, 2017). L'inconvénient de ce processus est qu'il prend beaucoup de temps et ne garantit pas que les paramètres d'ondelettes optimaux soient sélectionnés. Abbaszadeh et Alipour (2017) ont développé un cadre méthodologique qui aide à choisir les paramètres de TO les plus appropriés (type de fonction d'ondelette, *vanishing moment* et niveau de décomposition maximum) spécifiquement pour la modélisation des procédés hydrologiques. Les paramètres sont choisis dans le but d'améliorer l'efficacité du modèle développé (Abbaszadeh & Alipour, 2017). Les auteurs ont également fourni une relation significative entre les paramètres d'ondelettes

appropriés et les caractéristiques statistiques des données de bassins versants (précipitations et ruissellement), c'est-à-dire minimum, maximum, moyenne, variance. Dans cette étude, des données quotidiennes et mensuelles sont évaluées.

Dans le contexte de la modélisation des procédés hydrologiques, la sélection appropriée des paramètres est capitale, car elle conduit à un modèle d'apprentissage machine bien calibré. Bien que la TO ait été utilisée avec succès par les chercheuses et chercheurs pour prétraiter les séries chronologiques hydrologiques de manière répétée, personne n'a abordé la manière d'identifier ces paramètres dans leur ensemble (Abbaszadeh & Alipour, 2017). Le choix inapproprié des paramètres de TO peut conduire à un modèle d'apprentissage machine plus exigeant en termes de calcul et peut diminuer son efficacité (Abbaszadeh & Alipour, 2017). Par conséquent, pour contourner ce problème, quelques spécialistes de la recherche ont utilisé une équation pour trouver le niveau de décomposition optimal (Nourani et al., 2009; Nourani et al., 2011; Wang, W. & Ding, 2003), cependant son utilisation est limitée aux signaux représentés par des modèles autorégressifs pour lesquels le niveau de décomposition maximal est uniquement déterminé en fonction de la longueur du signal (Nourani et al., 2011). En outre, Abbaszadeh et Alipour (2017) ont mentionné l'absence de recherche spécifique sur la sélection appropriée de la fonction ondelette, un autre facteur important qui affecte l'efficacité du prétraitement des séries hydrologiques par la TO. La similitude de forme entre la fonction d'ondelette et le signal brut a souvent servi de ligne directrice dans le choix d'une fonction d'ondelette fiable. Enfin, le *vanishing moment* de la fonction d'ondelette a reçu peu d'attention de la part des spécialistes de la recherche dans le développement de modèles hydrologiques (Abbaszadeh & Alipour, 2017).

Abbaszadeh et al. ont considéré quatre types de fonction d'ondelette – Daubechies, Coiflet, Haar, Symlet – avec différents *vanishing moments* et niveaux de décomposition. Sur la base de leur travail, les paramètres d'ondelettes les plus appropriés doivent être déterminés en utilisant les directives suivantes pour capturer les caractéristiques hydrologiques :

1. L'ondelette Daubechies est la fonction d'ondelette la plus appropriée pour extraire les caractéristiques informatives des séries chronologiques hydrologiques en raison de sa similitude de forme avec les signaux originaux.
2. Le *vanishing moment* de cette fonction d'ondelette doit être choisi en fonction du comportement du bassin versant. Pour le bassin versant chaotique en général, db4 est recommandé, tandis que pour le bassin versant doux, db1 est préféré.

3. Le niveau de décomposition, quant à lui, doit être déterminé en fonction de la résolution temporelle des séries chronologiques hydrologiques. L'échelle temporelle quotidienne nécessite un niveau 8 de décomposition pour converger vers un résultat optimal, tandis que l'échelle temporelle mensuelle nécessite un niveau 4 de décomposition pour obtenir une estimation souhaitable. En utilisant ces niveaux, la périodicité des séries hydrologiques peut être correctement capturée.

Alyasseri, Zaid Abdi Alkareem et al. (2018) ont déclaré que les signaux d'électrocardiogramme, étant donné leur non-stationnarité (comme c'est le cas pour les signaux industriels), rencontrent fréquemment du bruit et un faible rapport signal/bruit (RSB). Dans leur article, la transformée en ondelette est utilisée pour le débruitage, mais la sélection de la meilleure combinaison de paramètres est une tâche ardue, car elle ne se fait pas en appliquant une technique donnée, mais plutôt en se référant à des preuves empiriques ou à l'expérience (Alyasseri, Zaid Abdi Alkareem et al., 2018). Par conséquent, en utilisant les signaux d'une base de données largement utilisée comme référence pour comparer les performances de différentes techniques de réduction du bruit, les auteurs ont appliqué un algorithme pour identifier les paramètres d'ondelettes optimaux qui minimisent l'erreur quadratique moyenne entre les signaux d'électrocardiogramme originaux et débruités. Les performances de l'algorithme ont été évaluées en utilisant comme critères la différence quadratique moyenne en pourcentage et le rapport signal/bruit. La méthode proposée a permis de réduire efficacement le bruit blanc gaussien de signaux non-stationnaires puisque le signal de sortie obtient un RSB élevé (qui correspond à un faible niveau de bruit, un signal lisse) et une faible différence quadratique moyenne (qui indique un débruitage efficace du signal original). Alyasseri, Zaid Abdi Alkareem et al. (2018) ont trouvé que les paramètres optimaux des ondelettes pour débruiter les signaux d'électrocardiogrammes sont la fonction d'ondelette *sym7*, le niveau de décomposition 2, la fonction de seuillage doux, la règle de sélection de seuillage *rigrsure* et l'approche de remise à l'échelle à un seul niveau.

De même, Sawant et Patil (2014) ont testé plusieurs paramètres d'ondelettes, notamment db2 à db10 comme fonction d'ondelette, fonction de seuillage doux et dur et *universal*, *minimax*, *heursure* et *rigrsure* comme règles de sélection de seuils, pour trouver les meilleurs qui peuvent obtenir un signal d'électrocardiogramme débruité avec un RSB élevé. Sawant et Patil (2014) ont mené leurs expériences sur des données synthétiques et ont identifié db5, niveau 4, seuillage doux avec seuil *rigrsure* comme étant les paramètres optimaux. Enfin, Mikhled et Khaled (2008) ont recommandé d'utiliser db4, niveau 5 et seuillage doux comme paramètres de débruitage pour minimiser l'erreur entre les signaux originaux et bruités.

2.3 Caractérisation des signaux

La sélection des paramètres d'une technique de traitement des signaux dépend des caractéristiques de la perturbation (bruit) et du signal lui-même (Dambros et al., 2018). La caractérisation des signaux de procédés est abordée dans certains articles, et en particulier pour les signaux de vibration. Les signaux de vibration sont analysés dans le domaine temporel en utilisant l'écart-type et la valeur de Kurtosis comme caractéristiques, et dans le domaine fréquentiel par inspection visuelle (Zuo-wei et al., 2014). L'amplitude et la fréquence d'un signal de vibration d'un générateur de puissance défectueux dans une centrale thermoélectrique sont analysées à l'aide de méthodes de décomposition du signal pour extraire les caractéristiques des fautes qu'il contient (Gao et al., 2008). Dans l'unité d'une centrale hydroélectrique, un signal de vibration est caractérisé par sa fréquence, son amplitude et sa tendance en utilisant un diagramme de forme d'onde et un spectrogramme (Ren et al., 2021). Un signal de niveau d'eau a été caractérisé en évaluant la distribution, le spectre de puissance et la fonction de densité de probabilité du bruit provenant du capteur de niveau d'eau, tandis que les signaux de vibration sont caractérisés par la distribution de fréquences (Yang et al., 2020).

Dans la littérature, les caractéristiques des signaux telles que la moyenne (García Plaza et al., 2018, 2019) [52, 53], l'écart-type (dispersion des données) (Fan et al., 2022; García Plaza et al., 2018, 2019; Malghan & Kumar Hota, 2022), l'amplitude moyenne (force du signal) (Fan et al., 2022), l'amplitude de crête à crête (Fan et al., 2022; García Plaza et al., 2018, 2019), l'énergie (Fan et al., 2022; García Plaza et al., 2018), l'entropie de Shannon (Fan et al., 2022; García Plaza et al., 2018, 2019), la moyenne quadratique (Fan et al., 2022), kurtosis (aplatissement) (Fan et al., 2022; García Plaza et al., 2018, 2019), skewness (asymétrie) (Fan et al., 2022; García Plaza et al., 2018, 2019), la moyenne de la valeur absolue (Fan et al., 2022), l'erreur quadratique moyenne (Malghan & Kumar Hota, 2022) et rapport signal/bruit (Malghan & Kumar Hota, 2022) sont abordés dans la littérature. Le RSB est souvent défini pour étudier les caractéristiques du bruit associé à un signal (Liu, Huan et al., 2017; Tanaka, 1994).

Compte tenu des caractéristiques des signaux, une évaluation de la corrélation entre celles-ci et les paramètres de la transformée en ondelettes mettrait en évidence certaines règles du pouce (lignes directrices ou heuristiques) pour un réglage systématique, efficace et structuré des paramètres de la transformée en ondelettes. À la connaissance des auteurs, il n'existe aucune recherche ayant évalué la corrélation entre les caractéristiques des signaux de capteurs de procédés industriels et les paramètres de la transformée en ondelettes. Cependant, l'idée d'une relation entre les caractéristiques des signaux et les

paramètres de certaines techniques de traitement n'est pas nouvelle. Afin de filtrer efficacement les signaux non-stationnaires, les paramètres du filtre doivent être sélectionnés en fonction des caractéristiques locales des signaux (Harashima et al., 1982; Pomalaza-raez & McGillem, 1984; Rajala & Figueiredo, 1981). Arakawa (1994) a proposé une technique de traitement des signaux basée sur des règles floues pour l'estimation de signaux non-stationnaires contaminés par des bruits aléatoires additifs. Pour cette technique de filtrage, des règles floues concernant la relation entre les caractéristiques des signaux et la conception du filtre sont appliquées pour définir les paramètres du filtre, en tenant compte des caractéristiques locales des signaux (Arakawa, 1994).

Néanmoins, les paramètres de filtres déduits de règles ou de directives ne peuvent pas être utilisés de manière incontestable si les règles décrivant comment définir ces paramètres sont exprimées de manière ambiguë. Par exemple, il est bien connu que la taille de la fenêtre d'un filtre visant à lisser des signaux contenant des bruits aléatoires doit changer de manière adaptative en fonction de la composante fréquentielle du signal d'entrée brut. Cependant, il est difficile de distinguer une grande taille de fenêtre d'une petite par un seuil unique. Une autre directive, spécifique aux TO, stipule que la fonction d'ondelette sélectionnée doit être similaire au signal d'entrée. Ces règles ne sont pas triviales à mettre en œuvre et à prendre en compte. Il est donc nécessaire d'établir des règles plus spécifiques pour le réglage des filtres, et plus particulièrement de la transformée en ondelettes dans ce cas.

2.4 Formulation systématique d'heuristiques

Le traitement des signaux de procédé est souvent le résultat de nombreuses années d'expérience dans le domaine. Par conséquent, cette connaissance tend à ne pas être exprimée explicitement, mais réside plutôt sous la forme de l'expérience du personnel de l'usine. Disposer d'un ensemble de règles (heuristiques) pourrait aider le personnel dans les tâches reliées au traitement des signaux. L'idée derrière les heuristiques est d'extraire ces connaissances et de les rendre ainsi explicites.

Il n'y a pas de définition standard des heuristiques dans la littérature; il existe diverses définitions, car elles ont été étudiées dans de nombreuses disciplines. En ingénierie, les heuristiques sont définies comme des stratégies potentiellement faillibles, qui ne garantissent pas de solutions, mais qui donnent une direction vers la résolution d'un problème (Fillingim, K. B. et al., 2020). Elles sont utilisées par les expertes et experts pour guider, découvrir et révéler (Fillingim, K. B. et al., 2020). Dans le domaine de la conception, une heuristique est une règle empirique qui permet de choisir l'action à entreprendre, compte tenu de l'état

actuel du processus de conception (Fillingim, Kenton B. et al., 2020). Les heuristiques sont, la plupart du temps, spécifiques à un domaine, elles sont basées sur une situation contextuelle. Ainsi, en général, une heuristique est une directive basée sur le contexte (dépendant du contexte), fondée sur l'intuition, la connaissance ou la compréhension expérimentale, qui augmentent les chances d'atteindre une solution satisfaisante, mais pas nécessairement optimale (Fu et al., 2015). Lorsque le contexte ou l'environnement change, les heuristiques peuvent ne plus être bénéfiques; de nouvelles heuristiques doivent être acquises. Comprendre et savoir quand et comment utiliser une heuristique est aussi important que de savoir quoi utiliser et pourquoi (Fillingim, K. B. et al., 2020). Il faut toujours tenir compte du contexte d'applicabilité des heuristiques. Mathématiquement, les heuristiques $h_i = \{s_i, a_i\}$ consistent en un ensemble de situations contextuelles (contexte d'applicabilité), s_i , et un ensemble d'actions possibles, a_i . Elles sont généralement moins formalisées que les principes ou les lignes directrices; ce sont elles qui ont le moins de preuves à l'appui ou de validation expérimentale.

Même si les heuristiques sont parfois utilisées de manière interchangeable dans la littérature avec les principes et les lignes directrices, car elles sont toutes des types de codification des connaissances, elles ne sont pas exactement identiques. Les principes sont considérés comme des règles ou des lois fondamentales, tandis que les heuristiques sont moins validées et formalisées, car elles reposent sur l'intuition et l'expérience (Fu et al., 2015). D'autre part, les lignes directrices reposent sur davantage de preuves empiriques que les heuristiques (Fu et al., 2015).

Il existe différentes méthodes dans la littérature expliquant comment ces heuristiques ont été établies. Les heuristiques sont souvent dérivées empiriquement par les expertes et experts du domaine; elles s'appuient sur l'expérience antérieure, l'intuition et les connaissances pour suggérer des actions. Cependant, les heuristiques proviennent d'une variété de sources, elles ne se limitent pas à l'expérience personnelle. Elles doivent être obtenues à l'aide d'une méthodologie de recherche scientifique répétable. Si les heuristiques peuvent être extraites de manière répétée et cohérente, alors elles peuvent être documentées, étudiées et présentées aux autres. Une revue de la littérature a montré comment les heuristiques ont été systématiquement définies.

Sangelkar et McAdams (2013) ont cherché un moyen d'automatiser le processus de génération d'heuristiques. Leur objectif est de réduire le besoin d'interprétation humaine dans la formulation d'heuristiques de conception en exploitant des méthodes d'apprentissage machine. Les auteurs pensent que les heuristiques de conception fournies par une machine pourraient être plus complètes et rigoureuses

que la génération manuelle d'heuristiques. Leur espoir est d'apprendre à la machine à enseigner aux expertes et experts du domaine. Ils ont extrait des heuristiques statistiquement significatives d'une base de données pour servir de suggestions ou de directives de conception en utilisant l'exploration de données de graphiques. Ils trouvent efficacement de nouvelles heuristiques pour la pratique de la conception. L'exploration de données s'est avérée être un outil utile pour rassembler des heuristiques afin d'améliorer la conception de produits à partir de données discrètes qui ne peuvent pas être comprises analytiquement (Sangelkar & McAdams, 2013).

Peu d'articles ont extrait des heuristiques dans la conception à partir des résultats de sondages et entrevues auprès d'expertes et experts (Fillingim, Kenton B. et al., 2020; Fillingim, K. B. et al., 2020; Lee, B. et al., 2017). Les heuristiques sont le résultat de l'expérience. Les expertes et experts ont évalué la valeur des heuristiques en fonction de leur utilité dans un contexte donné (Lee, B. et al., 2017). Ils peuvent mettre à jour leurs évaluations de la valeur des heuristiques en fonction de l'expérience, d'une meilleure compréhension, de l'apprentissage, de l'observation d'autres expertes et experts qui élargissent leur compréhension de comment, quand et pourquoi une heuristique est utile (Lee, B. et al., 2017). Il fut argumenté que la compréhension du contexte et de l'environnement dans lesquels une heuristique devrait être employée est une étape cruciale dans la détermination de sa valeur (Fillingim, K. B. et al., 2020).

Une autre approche est la méthode d'extraction heuristique (MEH); cette méthode basée sur les données est fondée sur la reconnaissance de modèles dans les solutions existantes (Matthews et al., 2002). Plus précisément, elle utilise l'analyse en composantes principales (ACP). L'ACP identifie les relations potentielles à partir de l'ensemble de données pour générer des heuristiques. En effet, les relations ont été transformées en heuristiques en décrivant la tendance ou les corrélations encodées par les équations linéaires. Cette méthode a été spécifiquement utilisée pour les heuristiques de conception, elle est considérée comme une méthode permettant de guider les concepteurs dans la direction de meilleures solutions lorsqu'une aide est nécessaire. MEH extrait des connaissances pour fournir un soutien dans les premières étapes du processus de conception, il identifie la relation entre les composants de la conception. Il est indiqué que cette méthode est particulièrement utilisée pour les domaines de conception où il existe peu de relations explicitement connues (Matthews et al., 2002). Le MEH cherche à extraire automatiquement des relations pertinentes et plus complexes à partir d'une base de données de conceptions antérieures. L'algorithme collecte les données de conception, génère des heuristiques à l'aide de l'ACP, puis les expertes et experts du domaine les examinent, les vérifient et les valident. L'implication des expertes et experts du domaine dans ce processus est essentielle, car ils peuvent identifier les relations les

plus importantes et les plus pertinentes, donner une interprétation physique aux résultats et évaluer la précision des heuristiques. Somme toute, il a été prouvé que l'ACP pouvait générer des heuristiques en convertissant les relations de conception pertinentes en heuristiques de conception. Cette thèse cherche à extraire des heuristiques, non pas dans le domaine de la conception, mais plutôt concernant les caractéristiques des signaux et les paramètres des transformées en ondelettes utilisées pour en retirer les composantes hautes fréquences (bruit) afin de traiter les signaux industriels de manière efficace et systématique.

Il a été mentionné que l'extraction d'heuristiques peut être basée sur des arbres de décision. Cependant, cette technique a tendance à devenir assez compliquée et difficile à interpréter par un humain lorsque l'ensemble de données devient plus grand (Matthews et al., 2002). L'apprentissage machine peut également générer un ensemble d'heuristiques, mais celles-ci doivent rester simples (heuristiques de type SI-ALORS), car l'ensemble de données ne peut être séparé qu'en fonction d'un seul composant à la fois (Matthews et al., 2002). Un réseau neuronal artificiel pourrait être entraîné à extraire les heuristiques implicites dans les données, mais la complexité de ces heuristiques tend à constituer un défi lorsqu'elles doivent être utilisées par quiconque (Matthews et al., 2002). De ce fait, l'ACP est l'approche retenue pour extraire les heuristiques dans le cadre des travaux présentés dans ce document.

Outre le retrait des erreurs aléatoires dont il était question dans les sections précédentes, plusieurs autres étapes sont requises dans l'optique d'un traitement des données de procédé. Ces étapes sont abordées à la section 2.5.

2.5 Traitement des données dans le cadre d'applications et de prises de décisions industrielles

Le traitement des données (signaux) dans les industries de transformation a comme objectif d'extraire les informations pertinentes des signaux de procédé en les transformant. Il permet une meilleure connaissance de l'usine, une réduction des temps d'arrêt (augmentation du débit), une diminution des coûts de maintenance, une augmentation de la qualité des produits, une réduction de la consommation d'énergie et une amélioration de l'efficacité énergétique (Desbiens et al., 2008). Ces bénéfices renvoient à diverses applications telles que l'optimisation, l'analyse des coûts, la surveillance des conditions d'opération, l'analyse des défaillances (détection et diagnostic de fautes), le dépannage de procédé, l'analyse énergétique et la simulation de procédé, qui requièrent un traitement des données. Les paragraphes

suivants décrivent comment les étapes de traitement des données des industries de transformation varient d'une application (décision) à l'autre, en se concentrant exclusivement sur les analyses hors ligne.

Delou et al. (2021) ont avancé que l'optimisation des procédés doit commencer par une étape d'acquisition des données où il faut spécifier si l'analyse est effectuée hors ligne, c'est-à-dire indépendamment de l'état réel de l'usine, ou en ligne, en lisant les valeurs directement dans la base de données du procédé. À l'étape de traitement des données, chaque variable d'entrée peut subir une transformation linéaire pour être mise à l'échelle ou pour effectuer un changement d'unité, et toute valeur indésirable ou hors limite est exclue. Ces deux premières étapes sont réitérées jusqu'à ce que toutes les erreurs possibles soient éliminées. Ensuite, pour estimer de manière fiable les paramètres du modèle d'optimisation, la détection des périodes de régime permanent doit être effectuée avant l'étape de réconciliation des données. Delou et al. (2021) ont appliqué deux tests statistiques pour la détection du régime permanent. À la quatrième étape, la réconciliation des données est formulée comme un problème de programmation non linéaire. Ensuite, le problème d'optimisation peut être résolu, et la solution peut être validée.

Abu-el-zeet et al. (2001) ont proposé d'effectuer la réconciliation des données au départ parce qu'elle peut réduire considérablement l'inexactitude des données du procédé due aux erreurs de mesure. Abu-el-zeet et al. (2001) affirment que cette opération améliore la connaissance du procédé et la performance du système de contrôle. En effet, les résultats montrent une nette amélioration des performances des algorithmes d'optimisation lorsque des techniques de réconciliation des données et d'estimation des biais sont utilisées pour traiter les données avant l'étape d'optimisation. Pour Abu-el-zeet et al. (2001), la réconciliation des données consiste à réduire ou éliminer, autant que possible, l'effet de l'erreur de mesure aléatoire, à améliorer la confiance dans le calcul des variables non mesurées, et enfin à identifier les pertes ou fuites au sein du procédé et les mesures erronées. Le problème de réconciliation des données est résolu à l'aide de la programmation quadratique séquentielle. Cependant, bien que l'algorithme d'optimisation utilise un modèle de procédé en régime permanent, il n'est pas fait mention de la manière dont les périodes de régime permanent sont identifiées. Les solutions de l'algorithme d'optimisation en régime permanent sont utilisées comme points de consigne dans les boucles de contrôle de bas niveau, comme le montre la Figure 2.6.

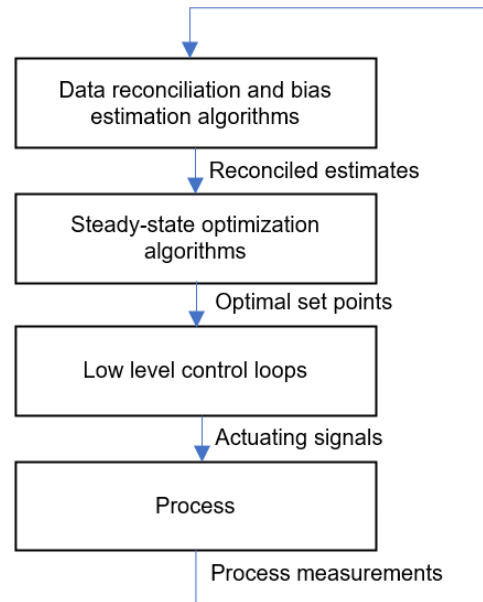


Figure 2.6 – Algorithme de traitement des données pour l'optimisation et le contrôle, inspiré de Abu-el-zeet et al. (2001)

Pour la détection et le diagnostic de fautes dans les équipements rotatifs, la première étape consiste à détecter les segments en régime permanent des signaux de vibration non-stationnaire. Pour ce faire, Su et al. (2006) ont utilisé une méthode de traitement des signaux basée sur les transformées de Fourier. Le traitement des signaux de vibration avant la détection de fautes permet une détection plus précise et réduit le nombre de fausses alarmes et de fautes manquées. Su et al. (2006) mentionnent que pour obtenir des indicateurs de fautes performants, qui ne sont pas influencés par les caractéristiques de l'équipement qui varient dans le temps, la tendance du signal doit être extraite des signaux de vibrations de l'équipement en régime permanent. Kaneko et Funatsu (2017) ont confirmé que l'utilisation de méthodes de lissage (telles que les moyennes mobiles et le filtrage Savitzky-Golay) améliore la précision de la détection et de l'identification de fautes et conduit également à une amélioration des performances de reconnaissance des états au sein du procédé. Kaneko et Funatsu (2017) ont constaté que le bruit des mesures et des capteurs dans les données d'opération rend difficile la détection, l'isolement et le diagnostic des fautes, ainsi que la reconnaissance des états de procédé. En effet, le bruit masque les variations importantes des variables de procédé et il diminue considérablement les performances d'identification des états de procédé (Kaneko & Funatsu, 2017). Par conséquent, il est critique de gérer le bruit avant d'effectuer la détection, l'isolation et le diagnostic de fautes ainsi que la reconnaissance des états (détection des périodes en régime permanent et

des régimes d'opération) (Kaneko & Funatsu, 2017). En effet, afin de fournir des résultats de détection et de diagnostic fiables, des données traitées sont nécessaires (Liu, Hancong et al., 2004).

En ce qui concerne la surveillance des usines, cette application utilise généralement des données de procédé en régime permanent ou quasi-permanent (Upadhyaya et al., 2014). En effet, Aldrich et van Deventer (1993) ont mentionné que la surveillance des usines dépend de manière cruciale d'ensembles de données fiables en régime permanent. L'acquisition de telles données est fondamentale pour une compréhension claire des opérations d'une usine, pour la modélisation et l'optimisation du procédé, ainsi que pour l'identification d'autres phénomènes propres au procédé. Les données de mesure sont généralement sujettes à des bruits aléatoires et à des erreurs grossières qui enfreignent généralement les contraintes du système. Il est donc nécessaire de prétraiter les données et de tenir compte des erreurs systématiques (grossières) avant d'effectuer la surveillance de l'usine. Aldrich et van Deventer (1993) y sont parvenus en utilisant un réseau neuronal à rétropropagation.

De plus, le régime permanent est utilisé dans la conception des procédés, plus particulièrement dans la conception des procédés en continu (Turton et al., 2009). De ce fait, des boucles de contrôle sont implantées, entre autres, afin de maintenir les procédés en régime permanent (Towler & Sinnott, 2008; Turton et al., 2009).

Également, la clé de la résolution de problèmes de dépannage et de désengorgement est d'utiliser des données de procédés collectées pendant les périodes de bon fonctionnement de l'usine (ce qui s'apparente au régime permanent) (Turton et al., 2009).

Korbel, Milan (2011) a proposé un cadre méthodologique de traitement des données dans l'optique d'une analyse de coûts d'opération afin d'identifier des bénéfices à court et à long terme (Figure 2.7). La connaissance du procédé doit être intégrée pour évaluer les coûts d'opération de chaque régime de production (Korbel, Milan, 2011). De ce fait, l'étape initiale consiste à définir les régimes d'opération sur la base de la connaissance du procédé, c'est-à-dire à caractériser les différences d'opération qui se produisent pendant la production de chaque produit. Connaître les différents régimes d'opération permet d'éviter les régimes coûteux et de maintenir le procédé dans la zone d'opération la plus rentable le plus longtemps possible. Les régimes d'opération peuvent être dus à des modifications de la conception du procédé, à l'utilisation d'équipements différents, à des modifications de l'objectif de production ou de la stratégie de contrôle du point de consigne, ainsi qu'à la diversité des matières premières (Korbel, Milan, 2011).

La deuxième étape consiste à identifier le fonctionnement en régime permanent dans le régime d'opération analysé. Pour obtenir des données en régime permanent acceptables, les données mesurées doivent être traitées pour réduire le bruit du signal avant que l'identification des périodes en régime permanent ne puisse être effectuée (Korbel, M. et al., 2007). Ainsi, le bruit et les anomalies sont d'abord extraits des tendances de mesures du procédé en appliquant des paramètres optimaux de transformation en ondelettes qui sont ajustés de manière itérative jusqu'à ce que le prétraitement optimal des données et la détection précise de l'état stationnaire soient perçus comme atteints. Ensuite, les tendances sont analysées pour identifier les occurrences des périodes en régime permanent en utilisant une méthode multi-échelle basée sur les ondelettes présentée dans (Korbel, Milan et al., 2014). La détection des périodes en régime permanent se fait en deux étapes, (1) de manière univariée, et (2) de manière multivariée en comparant l'ensemble des états des variables clés dans le temps.

Ensuite, la rectification des données pour l'ensemble de l'usine évalue la validité des données en les comparant aux valeurs du modèle de procédé sous-jacent, c'est-à-dire une simulation qui tient compte des bilans de masse et d'énergie et d'autres lois de conservation. Les données rectifiées obtenues à la suite de cette dernière étape sont utilisées pour une analyse avancée des coûts en vue de la prise de décision en matière de production. Cette analyse des coûts axée sur les opérations est réalisée par une approche de modélisation des coûts basée sur les activités pour un régime d'opération donné.

En somme, Korbel, Milan (2011) estime que pour atteindre un certain niveau de qualité des données en régime permanent, il faut utiliser des techniques de traitement avancées. Ces dernières constituent le pilier de cette approche de calcul des coûts. Korbel, Milan (2011) a également mentionné que les données de procédé mesurées par des capteurs sont une source de connaissance du procédé, mais qu'elles sont cependant intrinsèquement contaminées par différents types d'erreurs et, par conséquent, des techniques de nettoyage des données doivent être employées avant d'utiliser les données dans toute analyse.

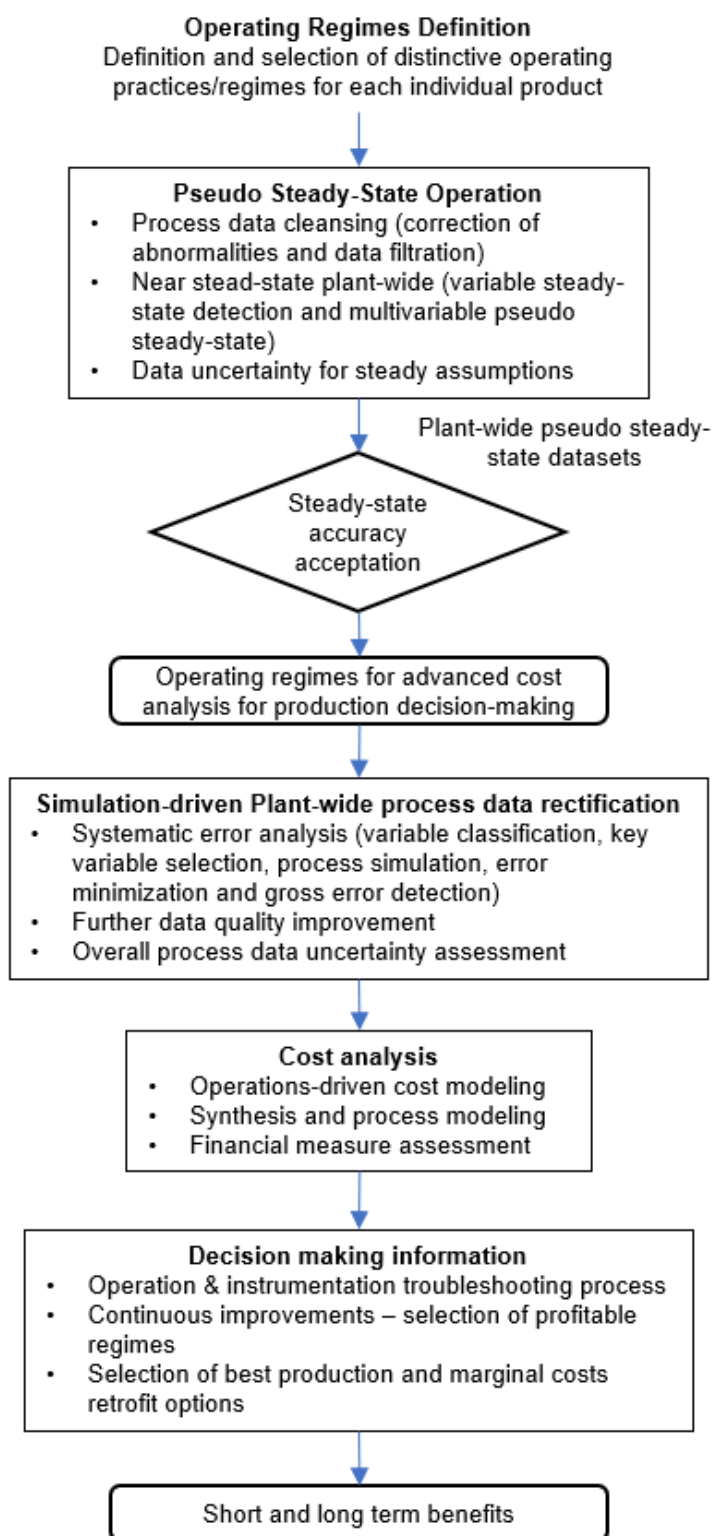


Figure 2.7 – Cadre méthodologique de traitement des données pour l'analyse des coûts, inspiré de Korbel, Milan (2011)

Jiang, Taiwen, Chen, Jasim, et al. (2003) ont proposé une stratégie systématique pour améliorer la qualité des données (Figure 2.8). En partant des données de procédé, l'étape de traitement des données réduit les erreurs et extrait les tendances du procédé. Pour ce faire, Jiang, Taiwen, Chen, Jasim, et al. (2003) ont effectué un traitement des données à plusieurs échelles (niveau) en utilisant la transformation en ondelettes. Le bruit aléatoire et les anomalies sont retirés et les tendances du procédé sont extraites en utilisant l'approche proposée par Jiang, Taiwen et al. (2000). Ensuite, une méthode basée sur la transformation en ondelettes présentée par (Jiang, Taiwen et al., 2000) a été utilisée pour la détection de l'état d'équilibre (régime permanent). Enfin, une étape de réconciliation des données a été réalisée à l'aide du logiciel Sigmafine (Sigmafine, 2022).

Le cadre méthodologique de traitement des données proposé par Jiang, Taiwen, Chen, Jasim, et al. (2003) vise à améliorer, valider et calibrer une simulation en régime permanent d'une usine de recaustification dans une usine de pâte à papier. Les données traitées et réconciliées ont montré que la simulation est raisonnable pour une application pratique telle que l'analyse de la sensibilité du procédé, l'étude de l'augmentation de la production de pâte à papier et pour donner un aperçu des options rentables de modernisation de l'usine.

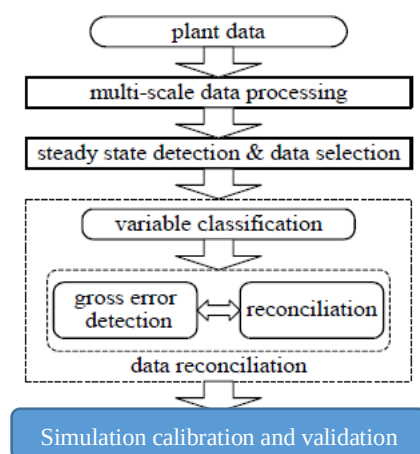


Figure 2.8 – Stratégie d'amélioration de la qualité des données pour la calibration et la validation des simulations (Jiang, Taiwen, Chen, Jasim, et al., 2003)

Plusieurs logiciels offrent la simulation des opérations d'un procédé en régime permanent étant donné que la plupart des procédés opérant en continu ne sont simulés qu'en régime permanent (Towler & Sinnot, 2008). Les simulations en régime permanent pour la conception de procédé sont discutées en détail dans Seider et al. (2009). Également, dans le but de construire un modèle de simulation d'une usine de papier journal, Harrison et Stuart (2007) ont prétraité les données de procédé notamment en identifiant et en

éliminant les valeurs aberrantes à l'aide d'une analyse en composantes principales et en utilisant un filtre basé sur une moyenne mobile à pondération exponentielle.

En se basant sur les travaux de Jiang, Taiwen, Chen, Jasim, et al. (2003), Bellec et al. (2007) ont suggéré une séquence d'étapes afin d'augmenter la valeur des données (Figure 2.9). Le cadre méthodologique de Bellec et al. (2007) vise à améliorer la qualité des mesures, c'est-à-dire l'exactitude et la précision globales des données de procédé. Les techniques de traitement des données sélectionnées sont adaptées aux applications en ligne⁴.

Le cadre méthodologique de traitement des données de Bellec et al. (2007) permet d'identifier les diverses sources d'erreurs présentes dans les mesures en temps réel, notamment le bruit, les anomalies et les incohérences dans les bilans, afin de fournir des données de meilleure qualité. Ces données pourraient aider les ingénieurs et ingénieures de procédé à identifier rapidement les besoins de maintenance des équipements, à dépanner et à diagnostiquer des problèmes d'opération et à améliorer la prise de décision liée au procédé. En ce qui concerne ce dernier point, la planification, l'ordonnancement et la coordination de la production sont des exemples de décisions de gestion de procédé qui sont généralement basées sur des ensembles de données en régime permanent.

La première étape de Bellec et al. (2007) présentée à la Figure 2.9 est la détection et la correction des mesures anormales; ceci est accompli en employant une méthode basée sur les transformées en ondelettes (Jiang, Taiwen, Chen, He, et al., 2003). Ensuite, une technique de filtrage des données (également basée sur la transformation en ondelettes (Jiang, Taiwen, Chen, He, et al., 2003)) est utilisée pour corriger les erreurs aléatoires. Ensuite, afin de réconcilier les mesures, les périodes d'opération en régime permanent sont soigneusement identifiées sur la base de l'approche mentionnée dans Jiang, Taiwen, Chen, He, et al. (2003). Enfin, la réconciliation des données est entreprise pour corriger les erreurs grossières en utilisant une méthode des moindres carrés pondérés.

⁴ L'analyse des données en ligne est effectuée tout en étant connectée au procédé. Elle utilise les données qui sont actuellement dans le procédé, avec peu ou pas de latence par rapport à leur temps d'acquisition. Les données sont mesurées pendant que le procédé continue d'opérer.

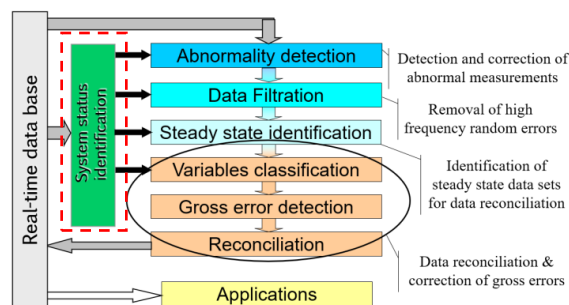


Figure 2.9 – Méthodologie pour améliorer la qualité des données proposée par Bellec et al. (2007)

Le cadre méthodologique de Bellec et al. (2007) inclut une identification de l'état du système où l'objectif principal est d'identifier les conditions d'opération du système étudié et de sélectionner les paramètres de traitement des données associés à l'état du procédé identifié (régime d'opération). La sélection est basée sur une étude hors ligne⁵ de toutes les conditions d'opération en régime permanent et les paramètres sont réglés pour obtenir des résultats optimaux. Bellec et al. (2007) restent vagues quant à l'optimalité des résultats.

La validation du cadre méthodologique présenté par Bellec et al. sur les données historiques d'une machine à papier inclut la notion de régimes d'opération. Bellec et al. (2007) ont souligné qu'un procédé peut avoir de nombreux régimes d'opération différents en régime permanent. Ceux-ci sont identifiés pour le système étudié, puis les caractéristiques associées aux différents régimes d'opération sont compilées et stockées dans une base de données. Ainsi, dans leur exemple, 27 régimes d'opération différents sont combinés en 4 groupes, à savoir la production de papier de grade A, la production de papier de grade B, les arrêts de maintenance et les apparitions d'une rupture dans la feuille de papier. La probabilité d'occurrence, la durée de chaque régime et le délai entre ces régimes sont évalués. En outre, la moyenne et la variance des mesures de procédé dans chaque régime d'opération sont également compilées et stockées dans une base de données. Ces caractéristiques, associées à chaque mesure du procédé dans un régime d'opération, sont utiles pour identifier correctement les conditions d'opération du procédé étudié (Bellec et al., 2007).

En outre, Bellec et al. (2007) s'assurent que les retards dans le procédé sont pris en compte (c'est-à-dire que les données sont synchronisées) en ne retenant que les périodes de régime permanent plus longues qu'une valeur seuil, c'est-à-dire le retard du système avec un facteur de sécurité. Des périodes de régime permanent

⁵ Hors ligne signifie que le procédé est analysé sur la base d'ensembles de données historiques.

suffisamment longues ont été ciblées afin de compenser le fait que les différents départements d'un procédé ne sont pas synchronisés.

Bellec et al. (2007) ont mentionné que l'amélioration de la qualité et de la fiabilité des données (exactitude et précision) justifie l'effort et les frais de mise en œuvre de leur cadre méthodologique. De plus, en compilant les données réconciliées en régime permanent dans différents régimes d'opération, les usines peuvent constituer une base de données précises sur les procédés. Cette dernière pourrait conduire à des avantages liés à l'opération du procédé et au dépannage, tels que l'identification des fuites dans le procédé et des défaillances des instruments pour améliorer le rendement du procédé et le calcul de la conversion. La compréhension du procédé est améliorée par la connaissance des différentes conditions d'opération du procédé. Enfin, une optimisation de l'ensemble ou de plusieurs unités pourrait être réalisée en couplant les données opérationnelles et les données sur les coûts d'opération pour des produits spécifiques dans des régimes d'opération particuliers.

Par conséquent, par rapport au cadre méthodologique proposé par Jiang, Taiwen, Chen, Jasim, et al. (2003), Bellec et al. (2007) ont pris en compte les régimes d'opération. À notre connaissance, les régimes d'opération sont identifiés manuellement par des expertes et experts du procédé. De plus, les décalages dans le procédé sont pris en compte, mais pas de manière systématique et rigoureuse; pour les éliminer, Bellec et al. (2007) n'ont retenu que les périodes de régime permanent qui durent un certain temps.

Ainsi, selon l'application, les étapes de traitement des données peuvent varier, mais elles constituent un préalable nécessaire et maximisent le réalisme et l'utilité des analyses ultérieures (Harrison & Stuart, 2011). Néanmoins, il est possible de conclure que, de façon générale, plusieurs analyses requièrent la connaissance des périodes de régime permanent. La détection de ces périodes est essentielle à l'évaluation de la performance des procédés et à l'application d'autres fonctionnalités telles que l'optimisation et le contrôle de procédé (Moreno, 2010). Certaines analyses (optimisation, détection de fautes, identification de modèles, etc.) requièrent des données de procédé en régime permanent afin d'obtenir des performances optimales (Moreno, 2010).

Outre les périodes de régime permanent, les régimes d'opération doivent être considérés afin de tenir compte de la variabilité de l'usine. La section suivante traite de la détection des régimes d'opération des procédés industriels.

2.6 Détection des régimes d'opération des procédés industriels

Les régimes d'opération sont des stratégies de fonctionnement différentes. Des exemples de régimes d'opération sont différentes matières premières, recettes de produits, saisons, différents grades, taux de production, modes des équipements, etc. Cependant, les changements dans les conditions d'opération, tels que l'évolution à long terme et la transition d'un régime à un autre, ont souvent été ignorés lors de l'analyse des données de procédés industriels (Liukkonen, Hiltunen, et al., 2011). Les systèmes modernes d'acquisition de données automatisées, l'amélioration de la capacité de stockage et de récupération des données, combinés à la prolifération de capteurs et d'instruments de mesure peu coûteux, ont augmenté massivement la disponibilité des données. Cette quantité croissante de données n'est pas facile à interpréter par les moyens conventionnels. Pour exploiter et tirer parti de ces données disponibles, les ingénieurs et ingénieurs chimiques doivent utiliser des outils de science des données, tels que ceux abordés dans ce document, et notamment dans la section 2.6. Les usines industrielles (chimiques) opèrent généralement dans un ensemble fini de régimes d'opération, qui peuvent être classés comme des modes où les variables de procédé varient dans une bande relativement étroite, c'est-à-dire qu'ils sont en régime permanent (Srinivasan et al., 2005). D'autre part, les périodes de transition sont caractérisées par de grands changements dans une ou plusieurs variables (Srinivasan et al., 2005). Ces dernières représentent l'état du procédé entre deux modes, par exemple, un changement de point de consigne, l'ouverture d'une vanne et un changement de configuration de l'équipement. Des recherches ont montré que le comportement d'un procédé peut être très différent selon les modes (Liukkonen, Hiltunen, et al., 2011).

Les données de procédé archivées peuvent être utilisées pour regrouper ou caractériser les états des procédés (Heikkinen et al., 2011). Heikkinen et al. (2011) ont regroupé des variables par l'algorithme *k*-means en clusters qui représentent différents états du procédé. Par conséquent, l'état du procédé est déterminé par l'analyse des données. Les clusters sont analysés pour indiquer les différences entre les états du procédé en examinant leurs propriétés et en utilisant la connaissance du procédé. Un cluster représentant les périodes d'arrêt de l'usine est facilement identifié sur la base d'un débit volumétrique très faible. Leurs résultats montrent que la saisonnalité est l'un des facteurs de distinction les plus clairs entre les clusters. Un groupe est prédominant en été, un autre en hiver, un troisième représente les périodes où l'ensemble du procédé est hors de contrôle ou arrêté, et un quatrième représente les périodes où le procédé est instable (Heikkinen et al., 2011). Les différents états du procédé identifiés ont fourni des informations sur le

comportement du procédé. En outre, ces informations sont utilisées pour déterminer les points de consigne optimaux pour les variables clés du procédé dans un état de procédé. La connaissance du procédé est nécessaire pour la sélection des variables et pour l'association et l'attribution des états de procédé aux clusters (Figure 2.10). Enfin, il y a toujours le bruit de mesure qui affecte la précision; les variables sont donc filtrées par un filtre à moyenne mobile (Figure 2.10).

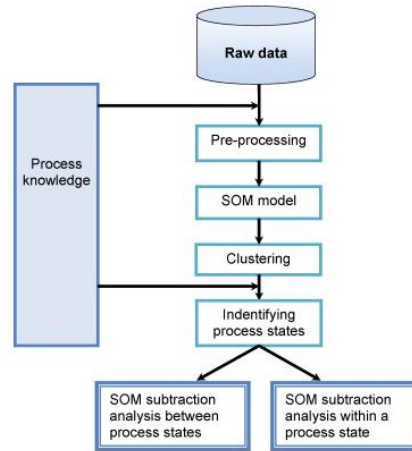


Figure 2.10 – Schéma de la chaîne de traitement des données (Heikkinen et al., 2011)

L'idée d'utiliser le clustering k -means pour créer des sous-ensembles représentant les états du procédé est également reprise par Liukkonen, Heikkinen, et al. (2011); (Liukkonen, Hiltunen, et al., 2011). L'une des méthodes de clustering non-supervisée la plus couramment utilisée est l'algorithme k -means (Jain, 2010). Le problème le plus important qui survient lors de mise en œuvre de cette technique est sûrement la nécessité de fixer à l'avance le nombre de clusters; celui-ci est rarement connu pour des procédés complexes opérant dans des conditions inconnues ou dans un environnement changeant (Saari & Odelius, 2018). Par conséquent, Saari et Odelius (2018) ont utilisé une autre technique de clustering non-supervisée pour détecter les régimes d'opération dans des données de maintenance axée sur la condition, par exemple, les signaux de vibration, appelée *variational Bayesian Gaussian mixture*.

Les techniques traditionnelles de clustering sont coûteuses en calcul et donnent généralement de mauvais résultats pour les données de séries chronologiques, car elles sont destinées à des observations plus indépendantes du temps. Srinivasan et al. (2004) ont proposé une méthode de clustering en deux étapes basée sur l'analyse en composantes principales. Les usines de transformation chimiques opèrent normalement selon un certain nombre d'états (démarrage de l'usine, changement de grade, arrêt), y compris les périodes en régime permanent, et passent fréquemment de l'un à l'autre. Par conséquent, à la

première étape, les états du procédé sont classés en modes, c'est-à-dire que l'unité de traitement fonctionne en régime permanent et en transitions, en utilisant un nouvel algorithme multivarié pour segmenter les données historiques. L'analyse en composante principale est ensuite utilisée à la deuxième étape pour comparer les différents modes et transitions et les regrouper en fonction de leurs similitudes. Le contrôle des procédés, les simulations, la détection de fautes et la gestion des alarmes sont des exemples d'applications où les paramètres du contrôleur doivent être ajustés pour s'adapter à l'état actuel du procédé (Srinivasan et al., 2004).

En pratique, la traduction des clusters obtenus en régimes d'opération nécessite une connaissance adéquate de l'historique du procédé, mais celle-ci est souvent indisponible. Par conséquent, des régimes d'opération inconnus peuvent exister. Ainsi, Beaver et al. (2007) ont utilisé l'analyse de clusters (Everitt, 1993) comme cadre méthodologique pour identifier simultanément les états du procédé, détecter les périodes de transition entre les états et étiqueter les périodes d'occurrence en utilisant des données historiques. Dans le cadre de l'interprétation des clusters, les régimes d'opération les plus souhaités sont identifiés sur la base de la connaissance de l'historique du procédé.

Le cadre méthodologique proposé dans le cadre de cette recherche est couplé à une analyse économique afin de démontrer son applicabilité et son potentiel. La section suivante aborde les différentes approches utilisées pour effectuer des analyses de coûts d'opération.

2.7 Approches utilisées pour l'analyse des coûts d'opération

La méthode de comptabilité par activité (*activity-based costing* – ABC) est orientée vers les procédés et les opérations : elle établit ce qui unit les objets de coûts et les activités et convient à l'analyse des coûts d'opération d'une production continue. Les applications de ce modèle de coût orienté procédé sont nombreuses. Tout d'abord, une approche qui intègre des mesures environnementales aux principes de l'ABC pour une évaluation économique et environnementale pour la prise de décision a été développée; les opportunités liées à l'environnement ont été classées dans le but d'améliorer la compétitivité tout en réduisant l'impact environnemental (Emblemsvåg, J. & Bras, 2000). Deuxièmement, une raffinerie de pétrole a utilisé l'ABC pour évaluer les coûts des sous-produits ainsi que du produit principal. En effet, cette méthode améliore l'évaluation des coûts de chaque produit lorsque plusieurs sont produits simultanément (Wang, C. et al., 2008). De plus, l'approche ABC a été utilisée pour évaluer la

performance des investissements dans les entreprises manufacturières (Sawhney, 1991). La méthode ABC a également été utilisée pour effectuer une analyse comparative des coûts du cycle de vie de différents scénarios de budget d'usine afin d'identifier différents facteurs clés (Emblemsvåg, Jan, 2003; Rivero & Emblemsvåg, 2007). Enfin, un gabarit a été proposé pour assister la prise de décision concernant les opérations afin de permettre une amélioration continue en utilisant la méthode ABC (Greenwood & Reeve, 1992).

En se concentrant maintenant sur le secteur des pâtes et papier, la méthode de comptabilité par activités a été employée pour la modélisation et l'analyse des coûts dans le cas d'une usine virtuelle (Korpunen & Paltakari, 2013). À la suite d'une analyse très approfondie prenant en compte les inducteurs de coûts (*cost drivers*) généraux et spécifiques au procédé (choisis avec soin), les résultats ont montré que la méthode ABC est une approche pragmatique du calcul des coûts, et avec les paramètres utilisés, il a été estimé que la fabrication du papier peut encore être rentable dans les conditions nordiques. Étant donné que l'ABC fournit des informations précises pour la prise de décisions stratégiques et tactiques (Korpunen & Paltakari, 2013), les procédés qui génèrent des coûts plus élevés que la valeur ajoutée ont été identifiés par les directeurs d'usine. Korpunen et Paltakari (2013) ont mentionné que les méthodes traditionnelles de calcul des coûts ne fournissent pas nécessairement des informations basées sur les procédés, et que certains produits marginaux ainsi que des matières premières marginales passent parfois inaperçus, ce qui peut donc conduire à une fausse détermination des prix. Les auteurs ont montré qu'un léger changement dans les prix du marché peut être critique pour la rentabilité et que quelques erreurs dans les estimations de coûts peuvent éventuellement conduire toute entreprise à un résultat préjudiciable.

D'autres études similaires – comptabilité par activités pour le secteur de l'industrie des pâtes et papier – ont été réalisées (Hsieh et al., 2020; Korpunen et al., 2012; Nurminen et al., 2009). Par exemple, Janssen a présenté une approche *bottom-up* pour la modélisation d'affaires (*business modeling*) en exploitant le potentiel des données de procédé en temps réel, des données de coûts ABC et des données environnementales pour la prise de décision concernant les procédés (Janssen, Matty & Stuart, 2004). Le concept du modèle d'affaires a été illustré par une application dans une usine de papier journal.

De même, une approche sophistiquée propose d'intégrer les informations sur les coûts et les procédés pour fournir une approche de calcul des coûts axée sur les opérations (*operations-driven costing approach* – ODCA) (Laflamme-Mayer et al., 2011; Laflamme-Mayer et al., 2008). Cela permet une

compréhension et une interprétation approfondies de la consommation des ressources. Cette méthode relie les coûts aux opérations du procédé en utilisant l'ABC. Cette dernière a été mise au point par des comptables ainsi que des ingénieurs et ingénieurs de procédé dans l'industrie des pâtes et papier, car elle nécessite une connaissance/compréhension approfondie des opérations de procédé (Korbel, Milan, 2011). L'ODCA cherche à minimiser les coûts de production. Elle identifie les régimes d'opération non rentables, et donc à éviter, ainsi que la marge de chaque produit; elle peut être utilisée comme un outil d'aide à la décision pour les opérations. La connaissance des marges de chaque produit permet d'identifier les produits sur lesquels il faut mettre l'accent (El-Halwagi & Stuart, 2013). Enfin, l'ODCA a été appliquée pour évaluer les coûts de production pour différentes campagnes de production; les résultats montrent que le coût de fabrication d'un même produit varie significativement d'une campagne (régime) à l'autre (Laflamme-Mayer et al., 2011). Elle a également été utilisée pour analyser et optimiser la chaîne d'approvisionnement d'une usine de l'industrie forestière (Laflamme-Mayer et al., 2008).

Chaque produit issu des usines de pâtes et de papier est fabriqué selon une recette spécifique. Cependant, chacune de ces recettes peut être produite selon différentes stratégies d'opération, i.e., des régimes d'opération. Ceux-ci sont caractérisés par des conditions d'opération et par des choix dans les opérations du procédé. Des points de consigne différents pour un paramètre, l'utilisation d'une ou deux lignes de production, l'ouverture et la fermeture d'une boucle de recirculation sont des exemples de régimes d'opération. Dans cette optique, Korbel et al. ont proposé d'utiliser l'ODCA pour évaluer le coût de production de différents produits sous différents régimes d'opération (Figure 2.11) (Korbel, M. et al., 2007).

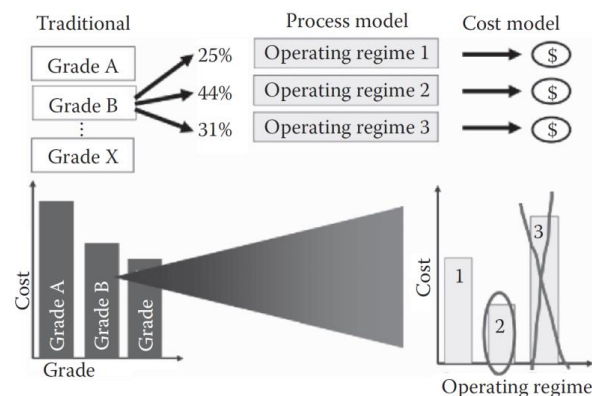


Figure 2.11 – Présentation de la méthodologie présentée par Korbel et al. (Korbel, M. et al., 2007)

Les méthodes d'analyse des coûts plus traditionnelles analysent les marges des différents produits (grades) de manière ad hoc (El-Halwagi & Stuart, 2013). Cependant, l'ODCA couplé aux régimes d'opération est basé sur une perspective d'ingénierie de procédés et fournit ainsi une meilleure compréhension des coûts et des marges des produits pour chaque campagne. En effet, cette technique repose sur l'utilisation d'un modèle de procédé qui présente les opérations en régime permanent. Ensemble, les modèles de coûts et de procédé permettent d'évaluer la probabilité d'occurrence de chaque régime d'opération et leur rentabilité individuelle et respective (El-Halwagi & Stuart, 2013). Ainsi, cette information est utilisée pour prendre des décisions au niveau des opérations de l'usine. Cette approche fonctionne de manière *bottom-up*, i.e., de l'identification des différents régimes d'opération au modèle de procédé, à la collecte des données de coûts, puis au calcul des coûts d'opération. La Figure 2.12 montre l'ensemble des échanges d'informations qui ont lieu avec ODCA.

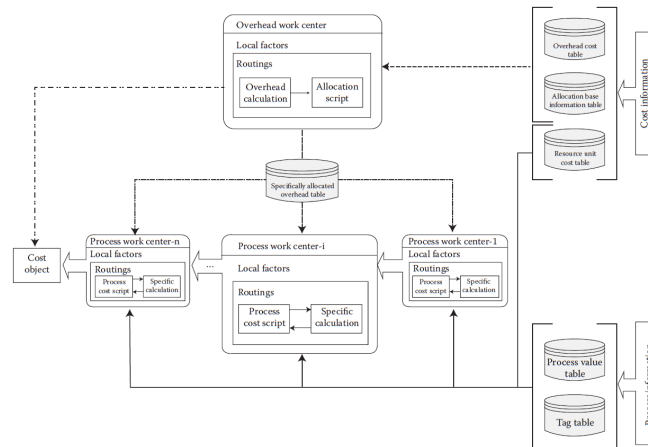


Figure 2.12 – Échange d'informations entre le modèle de procédé et les données économiques (Laflamme-Mayer et al., 2011)

En outre, l'ODCA permet l'évaluation économique de procédés continus générant un éventail de produits. Elle permet l'identification des régimes optimaux en agréant la visualisation de la façon dont les coûts varient avec chaque régime, et donne le coût de production et la marge des différents produits pour chaque régime d'opération. Bien qu'elle aide à identifier les régimes qui ne sont pas rentables, certains d'entre eux sont inhérents au procédé et sont inévitables. La Figure 2.13 montre le contraste entre les méthodes traditionnelles et l'ODCA. Cette dernière commence par l'identification des régimes d'opération.

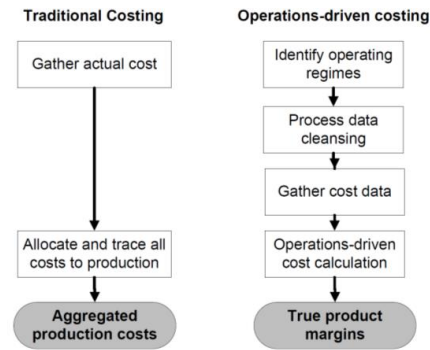


Figure 2.13 – Comparaison des méthodes d'analyse des coûts (Korbel, M. & Stuart, 2011)

L'analyse des coûts va être appliquée pour résoudre un problème d'optimisation dans le secteur des pâtes et papier. Par conséquent, les cas d'optimisation dans cette industrie, en mettant l'accent sur ceux qui considèrent les régimes d'opération dans l'identification de la solution, ont été examinés.

2.8 Analyse d'optimisation dans le secteur des pâtes et papier

Quelques articles traitent de l'analyse d'optimisation liée à l'industrie des pâtes et papier, comme celui sur l'optimisation énergétique (Marinova et al., 2009). Un autre article vise à améliorer l'efficacité du lavage et à réduire la consommation d'eau en déterminant le niveau optimal d'utilisation de l'eau de lavage dans le département de lavage de la pâte – niveau auquel les pertes de lavage sont satisfaisantes (Kopra et al., 2010). En outre, une étude visant à améliorer l'efficacité de la ligne de lavage de la pâte d'une usine de pâte chimique a été réalisée (Kopra et al., 2015). Les résultats indiquent que la quantité de pertes de lavage peut être réduite en contrôlant le facteur de dilution. De plus, un autre article a pour objectif de passer en revue les paramètres de coût du lavage de la pâte et d'explorer leurs sensibilités ainsi que la performance des équipements, le tout afin de présenter une optimisation des coûts (Cantrell, 2019). Enfin, Chew et al., 2013 présentent un modèle d'optimisation visant à économiser simultanément de l'eau et de l'énergie pour le système de lavage dans une usine de pâtes et papier, ainsi qu'une analyse de sensibilité pour explorer la relation entre l'efficacité du lavage et la consommation des utilités, i.e., l'eau et l'énergie (Chew et al., 2013). Toutes ces études ne tenaient pas compte des régimes d'opération.

Néanmoins, un article a abordé l'optimisation des performances des laveurs en calculant le ratio de déplacement (Kopra et al., 2011). Dans ce cas, les analyses indiquent que l'objectif était d'identifier le fonctionnement optimal du filtre sous pression dans le lavage de la pâte en modifiant les conditions expérimentales de traitement et les paramètres de contrôle. Ainsi, le type de pâte de bois varie (pin ou

bouleau), de même que le taux de production, la consistance de la pâte d'alimentation, le facteur de dilution et la pureté du filtrat de lavage (κ). De ce fait est survenue l'idée de considérer différentes recettes ou conditions d'opération, et d'identifier la recette optimale dans le contexte de l'opération des filtres à pression.

Dans un autre article, le concept de régimes d'opération a été mis de l'avant en tenant compte des conditions saisonnières (variations de la charge thermique). De ce fait, dans une usine de pâte chimique, un modèle d'optimisation multi-périodes pour les décisions de conception qui prend en compte les limites opérationnelles des unités de production de vapeur ainsi que les variations de la charge thermique au cours de l'année a été proposé (Svensson & Berntsson, 2013). Il a été mentionné qu'une approche conventionnelle qui considère des moyennes annuelles fixes (période unique) des paramètres du procédé risquerait de conduire à des solutions sous-optimales dans le contexte de la réduction de la demande de chauffage du procédé en raison de la négligence des variations de la demande d'utilités chaudes et des limites opérationnelles (Svensson & Berntsson, 2013). Par conséquent, l'approche suggérée a considéré la modélisation de la demande variable en utilité chaude. Suite à la comparaison de l'approche mentionnée ci-haut et de l'approche conventionnelle, cette dernière présente des erreurs significatives et aboutit à une estimation inadéquate de la solution optimale, ce qui fournit de mauvaises indications concernant la prise de décision financière. Les résultats de l'analyse d'optimisation ne peuvent être correctement déterminés qu'en incluant correctement les données de variation de la charge thermique. La même idée a également été abordée dans (Svensson, 2014). Ces études ne considéraient que les variations saisonnières de la demande de vapeur. Cependant, les différents taux de production, les campagnes de produits et les températures ambiantes sont autant de facteurs/causes supplémentaires des variations de la demande en vapeur du procédé. De plus, les prix de l'électricité et des matières premières (biomasse) varient également dans le temps, et peuvent avoir un impact significatif sur le fonctionnement optimal du système.

Par conséquent, un modèle d'optimisation de la conception des systèmes de distribution de vapeur d'une usine de pâte kraft, soumis à des variations des prix de l'énergie (électricité) et des conditions d'opération, y compris la production et les demandes de vapeur, est présenté dans Svensson et al. (2020). Ce dernier est utilisé pour étudier comment ces variations affectent les décisions d'investissement. Le modèle inclut davantage de sources de variations et tente de surmonter les défis qui accompagnent les tâches de modélisation tels que ceux liés à la fiabilité des données, aux incertitudes des données de coûts, au

manque de documentation et de spécifications techniques, à la faible correspondance entre les modèles issus de la littérature et les données de mesure ainsi qu'au large éventail de sources de variations pour lesquelles les corrélations sont souvent difficiles à identifier (Svensson et al., 2020). L'investissement optimal obtenu en tenant compte des conditions variables a été comparé à celui trouvé lorsqu'un seul scénario d'opération et de prix est considéré, et il a été démontré que le premier peut identifier des solutions avec plus de flexibilité que celles identifiées avec un modèle à période unique.

2.9 Lacunes dans le corps de connaissance

L'état de l'art des différentes approches employées pour retirer le bruit amène à leur comparaison. Les performances des méthodes de traitement du signal abordées aux sous-sections 2.21 à 2.2.4 ont été comparées dans la littérature afin d'identifier la technique la plus adéquate pour diverses applications (Dong et al., 2009; Dyason et al., 2017; Ibrahim et al., 2018; Perez & Barros, 2006; Pérez & Barros, 2012; Wang, Y. & Veluvolu, 2017; Yinfeng et al., 2008; Zhao et al., 2018). Parmi ces applications, aucune n'est spécifiquement axée sur les données de procédé, c'est-à-dire les données collectées par des capteurs de procédé (débit, température, niveau, etc.). Le Tableau 2.1 présente un résumé des articles qui ont abordé la comparaison des différentes techniques de traitement du signal et considère également la façon dont les paramètres des approches sont sélectionnés selon un niveau de 1 à 3. Un niveau 1 signifie que la sélection est triviale ou succincte (ou que les paramètres sont mentionnés sans aucune explication), le niveau 2 indique qu'elle est raffinée, mais non adaptée au type de procédé, enfin le niveau 3 suggère qu'elle est raffinée et adaptée au domaine.

Tableau 2.1 – Comparaison des approches prenant en compte la sélection des paramètres

Référence	Techniques comparées				Paramètres disponibles	Niveau de sophistication concernant la sélection
	FK	MMPE	TFCT	TO		
Yinfeng et al. (2008)	X		X	X	Oui	1
Zhao et al. (2018)	X		X	X	Oui	1
Wang, Y. et Veluvolu (2017)	X		X	X	Oui	1
Dong et al. (2009)	X		X	X	Oui	1
Pérez et Barros (2012)	X		X	X	Oui	1
Ibrahim et al. (2018)	X		X	X	Non	NA
Dyason et al. (2017)	X	X			Oui	2
Perez et Barros (2006)	X		X	X	Oui	1
Proposé ici	X	X	X	X	Oui	3

Les études qui évaluent la capacité ou la performance de ces approches pour supprimer le bruit des données de procédés industriels tout en considérant la sélection de leurs paramètres optimaux manquent dans la littérature. Le Tableau 2.1 montre que les articles comparant les techniques de traitement du signal ne divulguent pas toujours les paramètres de l'approche sélectionnée, et lorsqu'ils sont disponibles, ils sont sélectionnés de manière plutôt arbitraire, ou la méthode de sélection n'est pas explicite. Finalement, aucune étude ne compare les performances des quatre approches, tous domaines confondus. Par conséquent, l'évaluation des paramètres optimaux de chaque méthode de traitement du signal et l'identification de l'approche, montrant la meilleure performance pour des données synthétiques de capteurs de procédé, est abordée dans le cadre de travail.

Pour obtenir de meilleures performances, le choix des paramètres d'une approche de traitement des données doit être adapté en fonction des différents types de signaux traités. Cependant, les paramètres optimaux sont difficiles à déterminer, car ils nécessitent souvent des connaissances empiriques sur le système ou le procédé. Particulièrement pour le cas des transformées en ondelettes, certaines publications ont identifié la combinaison optimale des paramètres de débruitage en ondelettes pour les signaux de système de positionnement global, géologique, hydrologique et les électrocardiogrammes, mais aucun travail équivalent pour les données de capteurs de procédés industriels n'a été trouvé dans la littérature (Tableau 2.2). De plus, les approches employées citées dans la littérature ne mentionnent pas explicitement la réduction des composantes de haute fréquence tout en s'assurant de conserver les informations pertinentes en fonction de l'application visée. Enfin, seuls Abbaszadeh et Alipour (2017) abordent la corrélation entre les caractéristiques des signaux et les paramètres des ondelettes.

Pour pallier les inconvénients du traitement des données, il faut l'aborder de manière systématique avec des solutions permettant de traiter une grande quantité de données. Une revue de l'état de l'art a montré que la sélection des paramètres de la transformée en ondelettes affecte la performance globale du débruitage et que la sélection de valeurs appropriées est une tâche difficile qui est généralement réalisée sur la base de preuves empiriques ou d'expériences ou d'algorithmes complexes. Par conséquent, la littérature ne présente aucune directive ou règle du pouce permettant la sélection des paramètres des transformées en ondelettes de manière robuste, efficace et rapide en fonction du type de signaux et de leurs caractéristiques.

Tableau 2.2 – Recensement des articles qui considèrent la sélection des paramètres optimaux des TO

Référence	Domaine d'application	Paramètres d'ondelettes analysés	Approche employée pour sélectionner les paramètres	Corrélation entre les caractéristiques des signaux et les paramètres
Katunin et Przystańska (2015)	Structures composites	Fonction d'ondelette et son ordre	Algorithme de méta-optimisation multi-objectif	Non
Hasan et al. (2011)	Système de positionnement global	Fonction d'ondelette, règle de sélection de seuils, niveau de décomposition	Algorithme génétique	Non
Cao et al. (2022)	Traitement des images	Nombre de niveau, seuil, type d'ondelette	Algorithme de migration auto-organisé adaptatif	Non
El-Dahshan (2011)	Signaux d'électrocardiogramme	Fonction d'ondelette, niveau de décomposition	Algorithme génétique	Non
Abbaszadeh et Alipour (2017)	Modélisation de procédés hydrologiques	Fonction d'ondelette, <i>vanishing moment</i> , niveau de décomposition	Sélectionnés dans le but d'améliorer l'efficacité du modèle développé	Oui
Shafri et Yusof (2009)	Réflectance hyperspectrale	Règles de seuillage, fonction d'ondelette	Sélectionné à l'aide de données synthétiques	Non
Tan et al. (2008)	Analyse spectroscopique infrarouge	Niveaux de décomposition, fonction d'ondelette	Approche par essais et erreurs	Non
Wang, P. et al. (2021)	Spectre de réponse aux chocs	Niveau de décomposition, <i>vanishing moments</i>	Sélectionné à l'aide du coefficient de corrélation de Pearson	Non
Li et al. (2012)	Applications hyperspectrales	Règle de sélection des seuils, règles de seuillage, rééchantillonnage multiplicatif des seuils, niveau de décomposition, type d'ondelette	Sélectionné en utilisant l'angle spectral normalisé	Non
Alyasseri, Zaid Abdi Alkareem et al. (2018)	Signaux d'électrocardiogramme	Fonction d'ondelette, niveau de décomposition, méthode de seuillage, règle de sélection des seuils, méthode de remise à l'échelle des seuils	Sélectionné en utilisant l'erreur quadratique moyenne minimale entre les signaux original et traité	Non
Sawant et Patil (2014)	Signaux d'électrocardiogramme	Fonction d'ondelette, fonction de seuillage, règles de sélection des seuils	Sélectionné en utilisant le rapport signal/bruit	Non
Cette recherche	Procédé chimique	Fonction d'ondelette, échelle, fonction de seuillage, règle de sélection du seuil	Balayage paramétrique	Oui

Étant donné la revue des approches utilisées pour sélectionner les paramètres des transformées en ondelettes présentée au Tableau 2.2 et le fait que l'état de l'art a démontré qu'il est essentiel d'extraire la tendance d'un signal à partir des données bruitées en fonction de l'application avant de poursuivre avec toute analyse que ce soit, cette recherche propose d'évaluer la corrélation entre les paramètres de réglage de la TO et les caractéristiques des signaux de procédé afin d'accélérer le traitement des signaux.

Outre le retrait du bruit, d'autres étapes sont requises pour le traitement des données dans le cadre de prises de décisions. Il ressort de la littérature que la plupart des cadres méthodologiques de traitement des données pour la prise de décision hors ligne prennent en compte les étapes de traitement des données (suppression du bruit, détection des valeurs aberrantes et réconciliation des données), mais leur champ d'application n'est pas toujours explicite et la détection et l'identification des régimes d'opération n'est pas systématiquement prise en compte. Les régimes d'opération doivent être considérés afin de tenir compte de la variabilité de l'usine.

En considérant le cadre méthodologique proposé par Korbel (2011), la première étape est de définir les régimes d'opération. Ceux-ci sont définis grâce à la connaissance du procédé. Il est intéressant de considérer également les approches axées sur les données afin de compléter la détection et l'identification des régimes d'opération. Également, l'étape de *data cleaning* est effectuée à l'aide de la transformée en ondelettes. Des heuristiques permettant de sélectionner rapidement les paramètres optimaux des ondelettes qui rendent possible le traitement de nombreux signaux de manière systématique et efficace à travers le procédé. La détection des périodes en régime permanent se fait avec la transformée en ondelette, une approche plus pratique qui prend en considération la connaissance du procédé serait bénéfique. De plus, un cadre méthodologique qui considère la synchronisation des données, l'imputation des données ainsi que la correction des décalages serait à privilégier. Finalement, outre la perspective méthodologique du cadre présenté par Korbel (2011), dans une perspective d'application du cadre, une étude de cas portant sur d'une usine de papier journal thermomécanique intégrée a été réalisée. Les usines de pâte à dissoudre se trouvent à opérer selon un procédé plus complexe, et représentent donc un défi considérable en termes de traitement des données pour la prise de décisions liées aux opérations et à la conception. De ce fait, le cadre méthodologique proposé dans le présent document est une extension du cadre proposé par Korbel en 2011, tant au niveau de la méthodologie que de l'application.

Même si la notion de connaissance des procédés est considérée (Heikkinen et al., 2011) ou mentionnée (Beaver et al., 2007), les études, approches ou cadres méthodologiques qui identifient les régimes d'opération en utilisant une combinaison de connaissance du procédé et d'approches axées sur les données (*data-driven*) font défaut dans la littérature. Ce manque est adressé dans cette thèse en mettant en place un cadre méthodologique de traitement des données dans lequel les régimes d'opération sont pris en compte et détectés en utilisant à la fois des connaissances spécifiques au domaine et une analyse de clustering (analyse des composantes principales et *k-means*). Ainsi, une méthode systématique

et structurée de traitement des données inspirée des cadres méthodologiques présentés dans la section 2.5, qui prend en compte la définition de l'analyse et sa portée, le traitement des données, la détection des périodes de régime permanent, la réconciliation des données et la détection des régimes d'opération, est proposée. Le potentiel et la valeur ajoutée de l'approche proposée sont démontrés dans le cadre d'une analyse d'optimisation qui considère les coûts d'opération des différents régimes d'opération.

Comme cela a été démontré à la section 2.8, il n'existe pas d'analyse d'optimisation qui considère les régimes d'opération dans toute leur complexité pour mettre en évidence la manière optimale d'opérer en termes de coût, de débit des produits et de qualité de la pâte dans le département de lavage. Dans le cadre de ce document, un problème d'optimisation est résolu. Il n'y a pas d'algorithme d'optimisation à proprement parler. Il s'agit d'une étude de cas où les opérations sont améliorées par l'identification du régime d'opération le plus approprié dans lequel laver la pâte – celui qui minimise le coût de production, maximise le débit de production et la qualité du produit (pâte).

Ces différentes lacunes amènent à des objectifs de recherches et leurs hypothèses respectives qui sont présentées au chapitre 3.

CHAPITRE 3 OBJECTIFS

Sur la base de la problématique énoncée plus haut et en considérant la revue de la littérature, l'hypothèse générale de cette thèse intitulée « Prise de décision liée à la conception et aux opérations en milieu industriel complexe par l'étude des données en temps réel » est formulée comme suit :

En utilisant un cadre méthodologique de traitement des données qui identifie les régimes d'opération en régime permanent et en considérant leurs coûts d'opération associés, les opérations et la conception des améliorations d'un procédé peuvent être optimisées, de sorte que des modifications rentables et des possibilités de désengorgement soient identifiées.

Cette dernière peut être divisée en trois sous-hypothèses :

1. Bien que la prise de décision dans l'industrie des pâtes et papier soit généralement basée sur un fonctionnement en régime permanent pour la plupart des analyses de procédé, ces périodes stationnaires ne sont pas attribuées à des régimes d'opération spécifiques pour l'interprétation afin d'aborder par exemple l'optimisation du procédé et l'analyse des coûts d'opération.
2. La transformée en ondelettes est performante et robuste pour retirer le bruit de signaux de procédés industriels. L'exploitation de la relation entre ses paramètres et les caractéristiques des signaux permet de définir des heuristiques qui rendent possible le traitement de nombreux signaux de manière systématique et efficace à travers le procédé.
3. La fusion des régimes d'opération définis par des conditions reconnues du procédé (recettes ou analyse "*top-down*") avec ceux détectés par l'analyse des données (analyse "*bottom-up*") fournit un ensemble de régimes d'opération qui permet d'interpréter plus concrètement l'opération du procédé et donc de prendre de meilleures décisions liées à la conception et aux opérations concernant le procédé.

Compte tenu des hypothèses formulées ci-haut, les objectifs suivants ont été déduits, en commençant par l'objectif général :

Établir un cadre méthodologique dans un contexte industriel en résolvant un projet d'optimisation de procédé qui nécessite la reconnaissance des régimes d'opération en régime permanent et ses coûts associés et concrétiser son potentiel.

Cet objectif est divisé en trois sous-objectifs :

1. Mener des entrevues auprès d'usines du secteur des pâtes et papier ainsi qu'auprès d'entreprises développant des logiciels connexes, afin de déterminer en pratique comment les données sont traitées dans le cadre d'analyses de procédé.
2. Relier les paramètres de la transformée en ondelettes avec (1) les paramètres physiques du procédé (type de mesure, nature du fluide, type de capteur et configuration du procédé/localisation du capteur) et (2) les caractéristiques du signal (variance, coefficient de variation, taux de variation, fréquence, aplatissement et asymétrie) et appliquer ces relations à un ensemble de données.
3. Appliquer une méthodologie de traitement des données historiques qui permet une interprétation améliorée des régimes d'opération des procédés industriels, en combinant la connaissance des procédés avec des approches axées sur les données afin d'identifier des opportunités pour optimiser le procédé.

CHAPITRE 4 APPROCHE MÉTHODOLOGIQUE GLOBALE

4.1 Survol de la méthodologie

Dans le cadre de cette recherche doctorale, des entrevues sont menées auprès d'expertes et experts œuvrant dans des usines dans le secteur des pâtes et papier ainsi qu'auprès de développeuses et développeurs de logiciels soutenant l'industrie des pâtes et papiers pour déterminer comment les données sont traitées dans le cadre de l'analyse des procédés. En se concentrant uniquement sur les erreurs aléatoires (bruit) dans les données, la simulation d'un réseau d'échangeurs de chaleur permet de générer des signaux synthétiques qui identifient les paramètres optimaux de différentes méthodes de traitement des données. De ce fait, la performance de ces dernières est comparée. Étant donné la multitude de données à traiter dans le cadre d'une analyse, la possibilité de formuler des heuristiques (règles du pouce) qui fourniraient les paramètres optimaux d'une approche de traitement en fonction des caractéristiques des signaux est abordée. Plusieurs signaux ont donc été caractérisés selon le type de mesure, la nature du fluide, le type de capteur, leur environnement immédiat et selon plusieurs paramètres incluant la variance, taux de variation, fréquence de la tendance, symétrie, aplatissement, coefficient de variation. Considérant les étapes additionnelles au retrait du bruit requises dans le cadre de la prise de décisions industrielles liées à la gestion des opérations, une méthodologie systématique de traitement des données est proposée. Cette méthodologie (cadre méthodologique) considère l'objectif de l'analyse, le traitement des données (retrait des erreurs) et la détection des régimes d'opération. Toutes ces étapes nécessitent à la fois la connaissance du procédé industriel et l'utilisation d'approches mathématiques. Finalement, le cadre méthodologique proposé est appliqué aux données du département de lavage d'une usine de pâte et papier. Cette application du cadre méthodologique se fait conjointement à une analyse économique du procédé afin de résoudre un problème d'optimisation. Les sections suivantes détaillent la méthodologie appliquée dans chacune des publications présentées dans ce document.

4.2 Mener des entrevues au sujet de l'utilisation des données en industrie

Des entrevues avec des expertes et experts sont menées, en collaboration avec AVEVA, concernant l'utilisation et le traitement des données de procédé pour la prise de décision dans le secteur des pâtes et papier. Les entrevues durent environ une heure chacune et visent à identifier les applications pratiques des données de procédé ainsi que les applications des outils de traitement des données. Les personnes

expertes de l'industrie des pâtes et papier ciblées sont autant du côté des opérations en usines, que du développement de logiciels. Par conséquent, les entrevues sont menées auprès de deux groupes (1) le personnel des usines de pâtes et papiers et (2) les entreprises développant et fournissant des logiciels ou des technologies. Au total, cinq usines et huit sociétés de développement de logiciels sont ciblées. Parmi tous les membres du personnel des usines interrogés, plusieurs travaillent pour l'une des plus grandes organisations de produits forestiers au monde.

Pour les usines de pâtes et papier (qui utilisent toutes le système PI), la personne ciblée pour réaliser l'entrevue est soit celle qui collecte les données et essayait de les interpréter, qui est chargée d'améliorer les opérations du procédé en fonction de l'analyse des données ou qui travaille au niveau corporatif d'une organisation dans le secteur forestier (Figure 4.1). Les personnes interrogées dans les usines sont principalement des ingénieures et ingénieurs chimiques. Leurs rôles sont les suivants : experte et expert de projet à l'échelle de l'organisation, spécialiste-procédé et chef de projet de transformation numérique. Les chefs de projet dirigent, entre autres, des projets de transformation numérique (industrie 4.0) dans les usines.

Les spécialistes-procédé travaillent dans le contrôle avancé des procédés et sont responsables des données (système PI) et des technologies de l'information de l'usine. À ce titre, ils conçoivent et publient des tableaux de bord de production portant sur l'opération, la qualité, l'énergie, les coûts et les utilités. Certains supervisent les systèmes d'information de l'usine au niveau de l'organisation.

Enfin, les expertes et experts de projets à l'échelle de l'organisation supervisent toutes les technologies susceptibles d'être appliquées sur l'un de leurs sites, ainsi que tout ce qui a un impact sur l'organisation. Ces personnes dirigent les efforts en matière de meilleures pratiques (en général et avec le système PI), surveillent la performance de l'usine et la santé des équipements (en essayant d'éviter les pannes d'équipement en prédisant quand elles pourraient se produire), améliorent la fiabilité des équipements et soutiennent le personnel de l'usine au niveau du système PI ou des données de procédé. Ces expertes et experts forment un groupe stratégique d'ingénieures et ingénieurs dédié à l'analyse des données à distance de l'environnement de l'usine. Ces personnes évaluent de nouvelles technologies et nouveaux outils d'analyse des données.

Des personnes des départements d'ingénierie, du développement de l'organisation et de la présidence d'organisation sont choisies pour répondre aux questions (Figure 4.1). Les présidentes et présidents ont

généralement la charge du développement de nouvelles fonctionnalités des produits, du déploiement, de la formation, du support, du marketing et des ventes. Les responsables du développement de l'organisation soutiennent les ventes par le biais d'une approche de vente consultative et offrent des conseils sur les meilleures pratiques en matière d'analyse avancée dans différentes industries de transformation. Enfin, les ingénieures et ingénieurs seniors de recherche sont des expertes et experts du domaine qui offrent un soutien au personnel en matière d'analyse de données.



Figure 4.1 – Répartition des rôles des personnes interrogées

Les résultats des entrevues sont agrégés et comparés de manière anonyme. Les meilleures pratiques sur la façon d'utiliser les données de procédé pour la prise de décision sont mises en évidence, permettant aux usines de comparer leurs pratiques avec d'autres usines de pâtes et papier et de se mettre à jour sur la façon dont les autres usines font évoluer leur traitement des données pour s'adapter à l'environnement commercial changeant. D'un autre côté, les développeuses et développeurs de logiciels ont mieux compris les besoins et les défis auxquels sont confrontées les personnes utilisant le système PI. Finalement, une comparaison entre les approches de traitement des données présentées dans la littérature, proposées par les développeuses et développeurs de technologies et celles utilisées dans l'industrie et en usine est effectuée.

Les résultats des entrevues ont fait l'objet d'une analyse qualitative; une analyse critique a été effectuée et des discussions ont eu lieu avec des consultantes et consultants en traitement des données. Ces personnes ont contribué à la compréhension des résultats et ont validé l'analyse critique. Il a été confirmé que les résultats rapportés ici ne sont pas exclusifs à l'industrie des pâtes et papier et qu'ils sont en fait applicables à toutes les industries de transformation continues. Les résultats de ces entrevues sont présentés à la section 6.3.1.

Finalement, exclusivement pour le traitement des erreurs aléatoires, la section 4.3 présente comment la comparaison de certaines approches est proposée dans le cadre de cette recherche doctorale.

4.3 Comparaison d'approches permettant le retrait du bruit

Les performances du filtre de Kalman, du filtre EWMA, de la transformée de Fourier à court terme et de la transformée en ondelettes sont comparées à l'aide de signaux synthétiques. Les signaux synthétiques (simulés) sont obtenus à partir de la simulation d'un réseau d'échangeurs de chaleur (Figure 4.2) inspiré d'un train de préchauffage de pétrole brut, à laquelle sont ajoutées des erreurs aléatoires fictives. Cette simulation fournit 6 signaux de débit et 14 signaux de température. Les courants bleus sont de l'eau, tandis que les flux rouges ont la capacité thermique spécifique de l'huile.

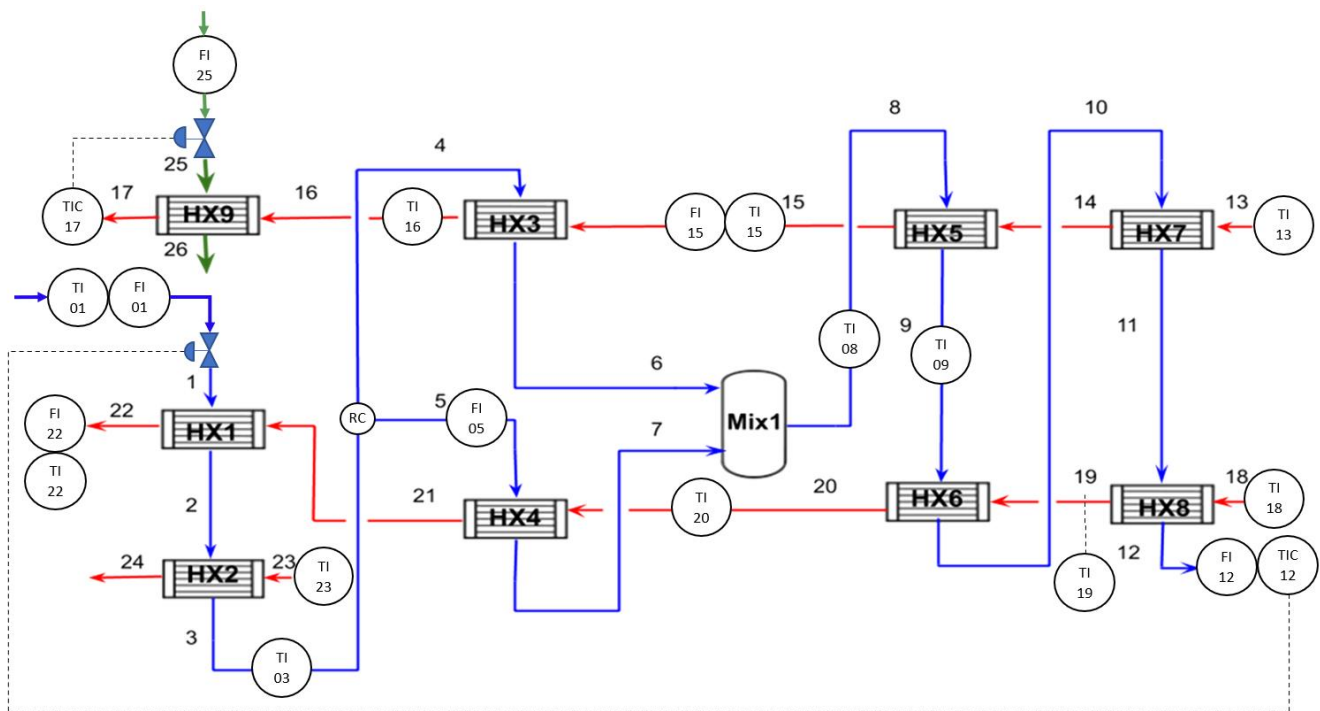


Figure 4.2 – Simulation d'un réseau d'échangeurs de chaleur : signaux de débit et de température

Les signaux de simulation sont générés en tenant compte du taux de variation en utilisant une fonction aléatoire haute fréquence superposée aux variations saisonnières et opérationnelles (tendance d'un procédé). De petites valeurs aléatoires à haute fréquence sont utilisées pour démontrer la variabilité rapide que peut avoir un signal. Ces dernières sont générées uniformément autour de la tendance. Ces petites valeurs aléatoires sont normalement distribuées étant donné que l'hypothèse selon laquelle les

erreurs aléatoires suivent une distribution gaussienne/normale est généralement acceptée, elle constitue une bonne approximation de nombreux environnements du monde réel (Park, S. et al., 2019). Les erreurs aléatoires sont simulées, basées sur la méthode de calcul séquentiel, en considérant l'amplitude de la mesure pour fournir des signaux de procédé réalistes et intéressants.

Dans le cadre de cette comparaison, la sélection des paramètres optimaux de chaque méthode est considérée. Un balayage paramétrique conserve les paramètres qui minimisent l'erreur entre le signal filtré par chacune des approches et le signal d'intérêt (tendance). Les différentes valeurs des paramètres pour chaque technique de filtrage sont présentées dans le Tableau 4.1.

Tableau 4.1 – Paramètres de débruitage couramment utilisés

Méthode	Paramètres spécifiques	Range or value
Filtre de Kalman	Bruit de mesure R	Écart-type du bruit
	Bruit de procédé Q	R/1000; R/100, R/10, R (Ma'arif et al., 2020)
Moyenne mobile à pondération exponentielle	Coefficient de lissage α	0.05 à 0.95 (avec un pas de 0.05)
Transformée de Fourier à court terme	Type de fenêtre	Blackman; Chebyshev; Gaussian; Hamming; Hann; Rectangular; Triangular
	Dimension de la fenêtre	Heure; Jour; Semaine; Mois
	Longueur de la FFT	Deux fois la dimension de la fenêtre
	Taux d'échantillonnage	Un point par minute
	Longueur du chevauchement	1%; 10%; 25; 50; 75 % de chevauchement de la dimension de la fenêtre
Transformée en ondelette	Fonction d'ondelette	Daubechies (db2-db15); Symlet (sym2-sym15); Coiflet (coif1-coif5); Biorthogonal (bior1.1-bior6.8)
	Échelle (niveau de décomposition)	1 à 10
	Fonction de seuillage	Seuil doux ; Seuil dur
	Règle de sélection des seuils	Rigrsure; Universal; Minimax

En utilisant les paramètres optimaux identifiés, la performance à réduire de manière adéquate les erreurs de signaux synthétiques est évaluée à l'aide de deux (2) critères quantitatifs, à savoir l'erreur quadratique moyenne et le rapport signal/bruit entre le signal traité et le signal d'intérêt. Ces derniers sont évalués systématiquement avant et après traitement du signal pour chaque méthode, et la différence est analysée. Ainsi, les erreurs quadratiques moyennes entre les signaux d'intérêt et les signaux bruités mesurés, ainsi qu'entre les signaux d'intérêt et les signaux traités, sont données respectivement par les équations suivantes :

$$RMSE_{avant} = \sqrt{\frac{1}{N} \sum_{i=1}^N [x(i) - \bar{x}(i)]^2}; RMSE_{après} = \sqrt{\frac{1}{N} \sum_{i=1}^N [x(i) - \hat{x}(i)]^2}$$

où $\bar{x}(i)$ est le signal bruité synthétique, $\hat{x}(i)$ est le signal traité, $x(i)$ est le signal d'intérêt, et N est le nombre d'échantillons dans le signal.

Ensuite, en prenant la différence entre $RMSE_{avant}$ et $RMSE_{après}$, l'efficacité d'une méthode de traitement est évaluée.

$$RMSE_{différence} = RMSE_{avant} - RMSE_{après}$$

Finalement, cette différence est divisée $RMSE_{avant}$ pour évaluer la $RMSE_{réduction}$. Plus la valeur est grande, plus la réduction est importante, et plus l'algorithme est efficace. Une réduction de 100% signifierait qu'il n'y a plus d'erreur et que le signal filtré est identique à celui d'intérêt :

$$RMSE_{réduction}[\%] = \frac{RMSE_{différence}}{RMSE_{avant}} \times 100$$

L'efficacité de la réduction du bruit est évaluée à l'aide du rapport signal/bruit. Le rapport signal/bruit est défini comme le rapport entre la puissance du signal et la puissance du bruit en décibel; il compare l'amplitude d'un signal souhaité au niveau du bruit de fond. Cette caractéristique reflète la dynamique du signal. Un rapport supérieur à 0 dB indique qu'il y a plus de signal que de bruit, donc plus le SNR est élevé, c'est-à-dire plus le niveau de bruit est faible, meilleure est la qualité du signal. Les SNR avant et après débruitage sont respectivement définis comme suit :

$$SNR_{avant}(dB) = 10 \log_{10} \frac{\sum_{i=1}^N \bar{x}^2(i)}{\sum_{i=1}^N [\bar{x}(i) - x(i)]^2}$$

$$SNR_{après}(dB) = 10 \log_{10} \frac{\sum_{i=1}^N \hat{x}^2(i)}{\sum_{i=1}^N [\hat{x}(i) - x(i)]^2}$$

$SNR_{amélioration}$ est utilisé pour évaluer l'efficacité de la réduction du bruit. Plus la valeur est élevée, meilleur est le résultat.

$$SNR_{différence}(dB) = SNR_{après}(dB) - SNR_{avant}(dB)$$

$$SNR_{amélioration}[\%] = \frac{SNR_{différence}}{SNR_{avant}} \times 100$$

Les résultats de ces deux critères pour les quatre approches sont présentés à la section 6.3.2. Finalement, considérant le fait que la sélection des paramètres des approches de traitement des signaux soit assez fastidieuse, une méthodologie de formulation d'heuristiques est proposée à cet égard à la section 4.4.

4.4 Établissement des heuristiques pour le traitement des données

L'identification des paramètres optimaux de la TO pour la réduction du bruit des signaux de capteurs industriel est obtenue par balayage paramétrique. Ce dernier teste chaque combinaison de paramètres et garde celle qui a minimisé une fonction objective minimisant la perte d'information pertinente, conservant l'information en dessous de la fréquence d'intérêt, et réduisant les composantes haute fréquence non désirées. Les paramètres de la transformée en ondelettes obtenus, combinés aux caractéristiques du signal, sont utilisés pour générer des heuristiques. Ces heuristiques sélectionneront les paramètres de débruitage et permettront de traiter les signaux en fonction de leurs caractéristiques inhérentes de manière efficace et structurée. Les étapes nécessaires à cette opération sont présentées à la Figure 4.3.

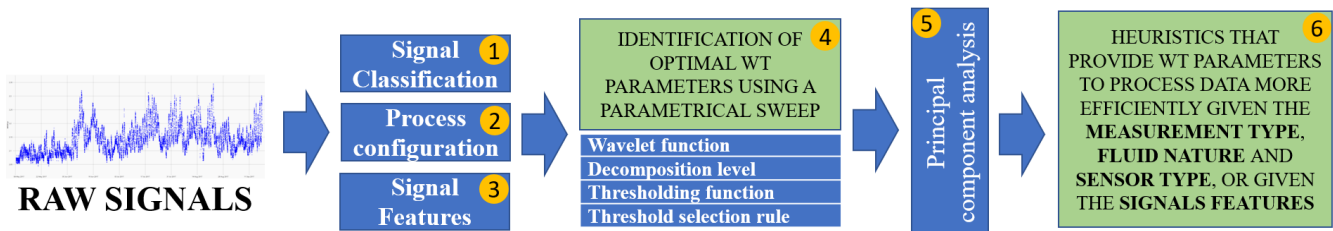


Figure 4.3 – Méthodologie employée pour formuler des heuristiques de traitement des signaux en fonction de leurs caractéristiques

La première étape consiste en la sélection et classification des signaux. Les signaux étudiés proviennent de capteurs situés dans plusieurs départements d'une usine de pâte à dissoudre. Différents types de mesures tels que la température, le niveau, la consistance et le débit ainsi que différentes natures de fluides tels que la vapeur, la pâte, la liqueur noire faible et forte, la liqueur blanche, etc. sont considérés. Un total de 400 signaux provenant de 100 capteurs sont débruités en utilisant la transformée en ondelettes.

Deuxièmement, l'environnement physique de chaque capteur, c'est-à-dire les équipements avant et après le capteur, est décrit. Des schémas-blocs succincts autour de chaque capteur sont dessinés. Ils permettent de voir où se trouvent les mesures dans le procédé et donc de mieux comprendre l'impact de leurs

emplacements. En plus de la configuration du procédé, le type de capteur est également pris en compte, c'est-à-dire si le capteur fait partie d'une boucle de contrôle ou s'il est uniquement un indicateur.

La troisième étape consiste à extraire les caractéristiques des signaux à l'étude dans l'optique d'évaluer leur corrélation avec la nature et le type de mesure. Suite à la revue de la littérature, les caractéristiques prises en compte dans cette analyse sont la variance, le taux de variation, la fréquence de tendance, l'asymétrie (skewness), l'aplatissement (kurtosis) et le coefficient de variation. La fréquence de tendance fait référence à une analyse spectrale des signaux effectuée par l'algorithme de la transformée de Fourier rapide (FFT) (Gao et al., 2008). Elle fournit les fréquences les plus fréquentes, y compris celle de la tendance.

Quatrièmement, les paramètres optimaux des ondelettes qui fournissent la tendance de signaux de procédés industriels afin d'établir un modèle prédictif sont identifiés. Dans cette étude, quatre types de familles de fonctions d'ondelettes, dont Daubechies (db), Symlet (sym), Coiflet (coif) et Biorthogonal (bior), avec différents *vanishing moments* et niveaux de décomposition, sont analysés par un balayage paramétrique. Ce dernier fait partie de l'approche proposée pour formuler des heuristiques, comme le montre la Figure 4.3. Différentes fonctions de seuillage et règles de sélection sont également comparées comme indiqué au le Tableau 4.2.

Tableau 4.2 – Plages des paramètres de débruitage par ondelettes couramment utilisés

Paramètres de la transformée en ondelettes	Intervalle
Fonction d'ondelette	Daubechies (db2-db15); Symlet (sym2-sym15); Coiflet (coif1-coif5); Biorthogonal (bior1.1-bior6.8)
Niveau de décomposition	1-10
Fonction de seuillage	Seuil doux ; Seuil dur
Règle de sélection des seuils	<i>Rigsure; Heursure; Universal; Minimax</i>

Un balayage paramétrique codé en MATLAB est développé pour sélectionner la combinaison optimale de paramètres énumérés dans le Tableau 4.2 qui fournit la tendance quotidienne des données. Ces données seront ensuite utilisées pour réaliser un modèle prédictif pour les décisions de gestion de l'usine pour lesquelles une granularité d'un jour est suffisante. Par conséquent, dans ce cas, une information toutes les secondes est considérée comme du bruit. Le bruit est un concept subjectif qui dépend de l'application, c'est une composante haute fréquence indésirable qui ne contient aucune information pertinente pour une application spécifique, c'est-à-dire que le bruit consiste en une variation plus rapide

qu'un phénomène étudié. Par conséquent, seules les informations quotidiennes sont conservées – les données sont traitées afin d'être synchronisées avec ce qui est intéressant dans ce cas, c'est-à-dire pour correspondre à la fréquence de ce qui est analysé compte tenu de leur contenu fréquentiel initial. L'information quotidienne est d'intérêt dans ce cas, c'est la référence, donc toutes les variations plus rapides qu'une journée ne sont pas nécessaires.

Tous les signaux industriels bruts n'ont pas le même contenu fréquentiel, ainsi les paramètres optimaux qui permettent de conserver les informations pertinentes, c'est-à-dire les informations liées à la fréquence ciblée (la fréquence d'intérêt), sont obtenus à l'aide d'un balayage paramétrique basé sur le rejet des composantes à des fréquences supérieures à un jour (informations indésirables). Cependant, bien que le lissage des données pour faire ressortir la tendance journalière soit important, il est essentiel de s'assurer que le signal n'est pas trop lissé et que des informations utiles ne soient pas perdues. Un critère rigoureux permettant un filtrage adéquat est nécessaire. Ainsi, à partir des composantes fréquentielles des signaux bruts et filtrés issues de l'analyse FFT (ligne 3 du pseudo-code), les paramètres optimaux sont déterminés en appliquant une fonction objectif (ligne 12) qui minimise l'amplitude des pics des composantes dont la fréquence est supérieure à la fréquence d'analyse (fréquence d'intérêt) (ligne 10) ainsi que l'erreur quadratique moyenne (RMSE) entre l'amplitude des pics des composantes dont la fréquence est inférieure à la fréquence d'intérêt pour le signal brut et le signal filtré (ligne 11).

Pseudo-code pour l'identification des paramètres optimaux d'ondelettes pour une analyse de la prédiction de la tendance

```

1 : pour tous les signaux  $i$ 
2 :     initialiser la valeur de la fonction objectif à l'infini (critère d'arrêt)
3 :     calculer la FFT du signal brut (d'entrée)  $i$ 
4 :     pour toutes les fonctions d'ondelettes  $j$ 
5 :         pour tous les signaux de décomposition  $k$ 
6 :             pour toutes les méthodes de débruitage  $l$ 
7 :                 pour toute règle de seuillage  $m$ 
8 :                     débruitage du signal avec la transformée en ondelettes
9 :                     calculer la FFT du signal filtré
10 :                     évaluer la somme de l'amplitude pour la fréquence > fréquence d'analyse
11 :                     évaluer RMSE de l'amplitude pour les fréquences < fréquence d'analyse
12 :                     évaluer la valeur de la fonction objective :  $OBJ = SAHF + RMSE$ 
13 :                     conserver les paramètres de débruitage si  $OBJ_{jklm} < OBJ_{min}$ 
14 :                      $OBJ_{min} = OBJ_{jklm}$ 
15 :     retourne les paramètres optimaux des ondelettes

```


À la cinquième étape, les relations entre les paramètres de classification des signaux, les caractéristiques et les paramètres de la transformée en ondelettes identifiés (Tableau 4.3) sont évaluées à l'aide d'une analyse en composante principale.

Tableau 4.3 – Variables utilisées dans l'analyse en composantes principales

Paramètres de classification	Caractéristiques	Paramètres des transformées en ondelettes
Type de signal	Variance	Fonction d'ondelette
Nature du signal	Coefficient de variation	<i>Vanishing moments</i>
Configuration du procédé	Aplatissement	Niveau de décomposition
Type de capteur	Asymétrie	Fonction de seuillage
	Tendance de la fréquence	Règle de sélection des seuils
	Taux de variation	

À la dernière étape, les heuristiques sont formulées grâce à l'analyse en composante principale qui considère les paramètres de classification des signaux de procédé, les caractéristiques et les paramètres optimaux de la transformée en ondelettes dérivés du balayage paramétrique. Ces heuristiques réduisent le bruit des signaux de procédé, d'une manière plus systématique et structurée, avant de les utiliser pour une analyse de prédiction. Plus précisément, ces heuristiques fournissent les valeurs optimales des paramètres de débruitage des ondelettes en fonction des caractéristiques du signal. Les heuristiques formulées et testées/validées sont présentées à la section 6.3.3.

Finalement, outre le retrait des erreurs aléatoires, d'autres étapes de traitement sont requises afin d'obtenir des données permettant une prise de décision fiable. De ce fait, la section 4.5 présente la méthodologie appliquée pour l'obtention d'un cadre méthodologique de manipulation des données pour la prise de décision.

4.5 Développement d'un cadre méthodologique de manipulation des données pour la prise de décision

Premièrement, la revue de la littérature des cadres méthodologiques de traitement des données existants a permis de souligner les lacunes de ces cadres. De ce fait, une façon de prendre en compte à la fois la connaissance du procédé et l'information contenue dans les données est réfléchie, et ce, pour chacune des étapes du cadre méthodologique.

Particulièrement en ce qui a trait à la détection et l'identification des régimes d'opération, la prise en compte de la connaissance du procédé relève de ce que nous appelons l'analyse *top-down*, tandis que l'approche axée sur les données est qualifiée d'analyse *bottom-up*. Les régimes d'opération classés comme *top-down* sont évidents et facilement détectables; les expertes et experts de procédé les connaissent. Les expertes et experts en procédés les connaissent. Les grades, les saisons, le taux de production, les recettes, les points de consigne, la fermeture des lignes, l'arrêt des équipements, le statut des pompes, etc. en sont des exemples. D'autre part, les régimes d'opération résultant de l'analyse *bottom-up* sont identifiés par une analyse de clustering. Ces régimes d'opération sont détectés grâce à l'analyse en composantes principales et à l'algorithme *k-means*, qui eux-mêmes nécessitent la connaissance du procédé. Cette dernière est également requise dans l'analyse *bottom-up* afin de donner un sens et une interprétation aux résultats des analyses axées sur les données. Ainsi, à partir des données de procédé, il est possible d'en savoir plus sur le fonctionnement de l'usine.

L'analyse *top-down* peut être assimilée à une approche *white-box* puisqu'elle met en avant la façon dont une usine est censée opérer, comment elle est conçue, quels sont les grades produits, quels sont les points de consigne pour ces grades, quelles sont les procédures d'opération standard (SOP), etc. D'autre part, l'approche *bottom-up* peut être perçue comme une approche *black-box*; les données du procédé, par le biais d'une analyse de clustering, peuvent découvrir des éléments inconnus sur le procédé.

Par la suite, outre la détection et l'identification des régimes d'opération du procédé, il faut définir la portée de l'analyse, traiter les données, détecter les périodes de régime permanent et réconcilier les données. Les trois dernières étapes sont assimilées à l'étape de gestion des données. Enfin, les informations obtenues par le cadre méthodologique sont utilisées pour la prise de décision. La Figure 4.4 illustre l'architecture globale du cadre méthodologique.



Figure 4.4 - Présentation du cadre méthodologique de traitement des données

Définition du champ d'application

La première étape de toute analyse de traitement des données consiste à définir les objectifs de l'analyse, par exemple optimiser la séquence de production, réduire le coût d'opération d'un département d'usine,

identifier les régimes d'opération non rentables. Ensuite, les régimes d'opération sont définis en utilisant la connaissance du procédé. Enfin, les données sont collectées sur un horizon représentatif pour l'analyse.

Gestion des données

Dans le cadre de l'étape de gestion des données, les données sont traitées, ce qui signifie que les erreurs aléatoires sont réduites, que les périodes de régime permanent sont détectées, c'est-à-dire les périodes où l'ensemble du système analysé est relativement stable, et que la dernière étape est une réconciliation des données à l'échelle du système pour détecter, identifier et éliminer les erreurs grossières (systématiques).

Détection et identification des régimes d'opération

L'étape de détection et l'identification des régimes d'opération est réalisée par une analyse de clustering basée sur une combinaison de l'analyse en composantes principales, de l'algorithme k -means et de la connaissance du domaine. Les données réconciliées en régime permanent sont classées en régimes d'opération et, enfin, certains de ces régimes sont sélectionnés en fonction de la possibilité de résoudre un problème de gestion de haut niveau (lié à la conception ou aux opérations du procédé).

Somme toute, afin de mettre en place le cadre méthodologique proposé, toutes les étapes présentées à la Figure 4.4 ont été réfléchies afin d'être pratiques et d'incorporer la connaissance du procédé. Suite au développement du cadre, une étude de cas est réalisée sur les données du département de lavage de la pâte brune d'une usine de pâte à dissoudre. La méthodologie employée pour démontrer l'applicabilité de l'approche méthodologique proposée est explicitée à la section 4.6.

4.6 Application du cadre méthodologique pour l'optimisation d'un procédé industriel

Les procédés continus, tels que les usines de pâte à dissoudre, sont souvent caractérisés par la production de produits et régimes d'opération distincts. Il est donc crucial d'évaluer les marges des produits individuels dans chaque régime pour s'assurer qu'ils sont tous rentables. L'amélioration de la compréhension des écarts de coûts entre les différents régimes d'opération (recettes de production) dans le département de lavage d'une usine de pâte à dissoudre se fait grâce à :

- (1) la définition d'un problème à résoudre, c'est-à-dire l'identification d'une étude de cas d'optimisation de procédé;

- (2) appliquer le cadre méthodologique proposé pour obtenir des données réconciliées en régime permanent dans chaque régime d'opération;
- (3) évaluer le coût de chaque régime d'opération en utilisant la comptabilité par activités;
- (4) résoudre le problème d'optimisation;
- (5) proposer une meilleure façon d'opérer le département de lavage.

L'optimisation du département de lavage doit être liée à son objectif principal, c'est-à-dire éliminer les composants indésirables de la pâte. En particulier, cela doit être fait de la manière la plus économique possible en maximisant le profit pour un maximum de solides dissous retirés de la pâte. Ainsi, la meilleure manière d'opérer le département de lavage, c'est-à-dire les régimes d'opération les plus appropriés, est identifiée en affectant les coûts de production à chaque régime. De ce fait, l'objectif du problème d'optimisation est d'améliorer les opérations du département de lavage en identifiant le régime d'opération préféré et les régimes non rentables (ou moins rentables); en d'autres termes, de déterminer quelles conditions (paramètres d'opération) répondraient aux critères suivants :

- (1) maximiser le rendement/capacité globale du département de lavage (quantité de pâte lavée);
- (2) maximiser la qualité/performance du lavage en minimisant la conductivité de la pâte;
- (3) maximiser le profit, en tenant compte de la quantité d'antimousse et de l'utilisation de produits chimiques de blanchiment.

Le premier critère doit tenir compte des contraintes physiques du département, c'est-à-dire des capacités maximales de production de pâte de chaque laveur. L'objectif est d'identifier le régime qui achemine la plus grande quantité de pâte au réservoir de pâte épaisse écrue. Cependant, il existe un compromis entre la capacité et la qualité de la pâte (la performance du lavage).

Le second critère tente de minimiser la conductivité (indicateur de propreté de la pâte). La qualité du lavage de la pâte est indirectement mesurée par sa conductivité à la fin du département; plus sa valeur est faible, plus la qualité est élevée. La conductivité est due à la liqueur noire (lignine) qui reste dans la pâte. Une valeur de conductivité élevée signifie que la pâte contient de la lignine (composante que le département de lavage essaie d'éliminer autant que possible – il s'agit de ce qui consomme les produits chimiques de blanchiment et cause la couleur brune de la pâte). Par conséquent, le coût du blanchiment est lié à la teneur en lignine, une teneur élevée en lignine laissée dans la pâte entraîne des coûts plus

élevés en produits chimiques de blanchiment. Ainsi, le taux de production idéal respecte la capacité de charge des laveurs et tient compte du fait qu'une surcharge entraîne une moins bonne efficacité du lavage.

Pour le troisième critère, le coût d'opération des différents régimes est évalué. La formation de mousse est l'une des principales difficultés qui affectent négativement l'efficacité du lavage de la pâte à dissoudre. La mousse se forme lorsque les copeaux de bois contiennent des concentrations élevées de matières extractibles ou lorsque l'air est entraîné dans le système. La formation de mousse peut entraîner une perte de production et augmenter la consommation de produits chimiques. L'agent antimousse est utilisé pour éliminer la mousse qui empêche un bon lavage; une utilisation suffisante d'agent antimousse augmente la capacité du laveur. La quantité d'antimousse requise augmente proportionnellement avec la production de pâte; les unités sont donc des mL par tonne de pâte.

Les données de procédé nécessaires pour résoudre le problème d'optimisation sont traitées selon les étapes présentées dans le cadre méthodologique. À savoir, afin d'effectuer l'analyse de coûts par activité, les informations du procédé sont acquises en utilisant des techniques de traitement du signal, de détection de périodes en régime permanent et de réconciliation des données pour chaque régime d'opération (détecté dans le cadre méthodologique susmentionné).

L'analyse des coûts s'inspire du modèle de coûts axé sur les opérations proposé par Laflamme-Mayer et al. (2011). La première étape consiste à caractériser les opérations du procédé à partir des données du procédé afin d'identifier les différentes ressources et activités du département. Ensuite, la consommation des ressources est affectée aux différentes activités, puis les activités aux objets de coûts. Ainsi, les activités servent de lien entre la consommation de ressources et les objets de coûts (Cooper & Kaplan, 1991). En bref, l'information sur les coûts est extraite, et les inducteurs de coûts sont définis et organisés en matrices qui correspondent aux équations fondamentales de masse et d'énergie sous-jacentes. Le coût de production par tonne de produit (objet de coût) est évalué.

Le résultat de cette analyse d'optimisation est l'identification de la meilleure façon (idéale) d'opérer le département de lavage. La solution ne prend pas la forme d'une modification de la conception nécessitant un investissement en capital. En effet, le secteur des pâtes et papier essaie actuellement de faire le maximum avec le moins d'investissement possible, ainsi les solutions nécessitant peu ou pas d'investissement sont prioritaires. Les différents régimes d'opération vont être classés en fonction de la valeur ajoutée qu'ils apportent et de leur coût, en considérant leur faisabilité. L'analyse des coûts par

activité fournira le coût marginal de la production dans chaque régime d'opération. Le temps pendant lequel le procédé peut opérer selon les différents régimes d'opération sera également être déterminé.

Une analyse plus poussée pour comprendre les causes des performances des régimes d'opération est effectuée avec une analyse en composantes principales.

CHAPITRE 5 ÉTUDE DE CAS

5.1 Contexte

Les publications 3, 4 et 5 utilisent des données provenant d'une usine de pâte à dissoudre. Cette dernière s'alimente principalement de bouleaux, produit en moyenne entre 475 tonnes de pâte par jour, et compte les départements (unités) suivantes : mise en pâte (lessiveurs), lavage, blanchiment, machine à papier, évaporation, chaudières de récupération, caustification et four à chaux. Le procédé de pâte à dissoudre élimine l'hémicellulose de la pâte afin de garantir un plus haut contenu en cellulose. La pâte obtenue a un haut niveau de brillance et peut être utilisée dans la production de rayonne, viscose, billet de banque, cellophane, film de photographie, produits chirurgicaux, etc. L'annexe A présente une version simplifiée du procédé.

Plus particulièrement, les deux dernières publications sont orientées vers le département de lavage. La pâte, qui vient d'être cuite, est composée de fibres (cellulose) et d'impuretés en suspension dans la liqueur noire. Cette dernière contient des composants dissous du bois, dont la lignine (l'indice Kappa est un indicateur de la teneur en lignine de la pâte), et les produits chimiques initialement présents dans la liqueur de cuisson. L'objectif principal du lavage de la pâte brune est de séparer les matières organiques dissoutes et les produits chimiques de cuisson usés de la pâte afin qu'elle puisse être blanchie et que la liqueur soit envoyée aux évaporateurs. En d'autres termes, le but est de récupérer les produits chimiques de cuisson usés aussi efficacement que possible et de minimiser le transfert des solides de la liqueur noire vers le département de blanchiment afin de réduire l'utilisation excessive de produits chimiques de blanchiment.

Si la pâte est mal lavée, certains composants dissous resteront et auront un impact sur le procédé de blanchiment qui nécessitera donc plus de produits chimiques. D'un autre côté, une pâte trop lavée signifie une liqueur noire trop diluée, ce qui augmente la demande de vapeur aux évaporateurs; la quantité d'eau à évaporer augmente et donc la demande d'énergie augmente proportionnellement.

Les principes de lavage de la pâte brune (pour les laveurs à tambour) sont la dilution, l'extraction et le déplacement. Tout d'abord, la pâte est diluée dans la cuve du laveur, ce qui diminue la concentration de solides dissous dans la liqueur noire. Le facteur de dilution représente la masse d'eau ajoutée moins la masse de liquide partant avec la pâte, le tout divisé par la masse de pâte. Ensuite, le fait de retirer du

liquide de la suspension pâte/liqueur réduit la quantité totale de solides dissous, mais la concentration reste la même. Enfin, le courant de la douche traverse le tapis de pâte; il remplace le liquide déjà présent dans la pâte (qui est éliminé par le vide du tambour). Le rapport de déplacement est une mesure de l'efficacité du lavage.

D'autres indicateurs de la performance du lavage sont la viscosité et la brillance qui sont des mesures indirectes de la fraction de cellulose dans la pâte. La vitesse de rotation du tambour peut affecter l'épaisseur de la feuille qui se forme sur le tambour; pour obtenir une épaisseur optimale de feuille, la meilleure fourchette est de 2 à 3,5 rotations/min. En outre, la température de l'eau de lavage peut avoir un impact sur l'efficacité du lavage, car une température plus élevée augmente l'élimination des solides dissous. Enfin, les consistances de l'alimentation et de la décharge affectent la consommation d'énergie de l'équipement, et peuvent donc également affecter les performances.

Dans le cadre de cette étude de cas, le département de lavage utilise trois équipements pour retirer différents composants de la pâte (Figure 5.1). Tout d'abord, un système en cascade à deux noueurs est utilisé pour éliminer les fibres de bois non cuites et les nœuds. Ensuite, le lavage est effectué dans deux lignes parallèles de laveurs rotatifs disposés dans une configuration classique à contre-courant. Les deux différentes lignes de lavage ne fonctionnent pas toujours simultanément. Cela se produit lorsqu'un problème d'opération se produit soit en amont, dans les lessiveurs, soit en aval, dans le département de blanchiment. Enfin, un système de cascade à trois tamis est utilisé pour éliminer les autres débris (plastique, sable, etc.).

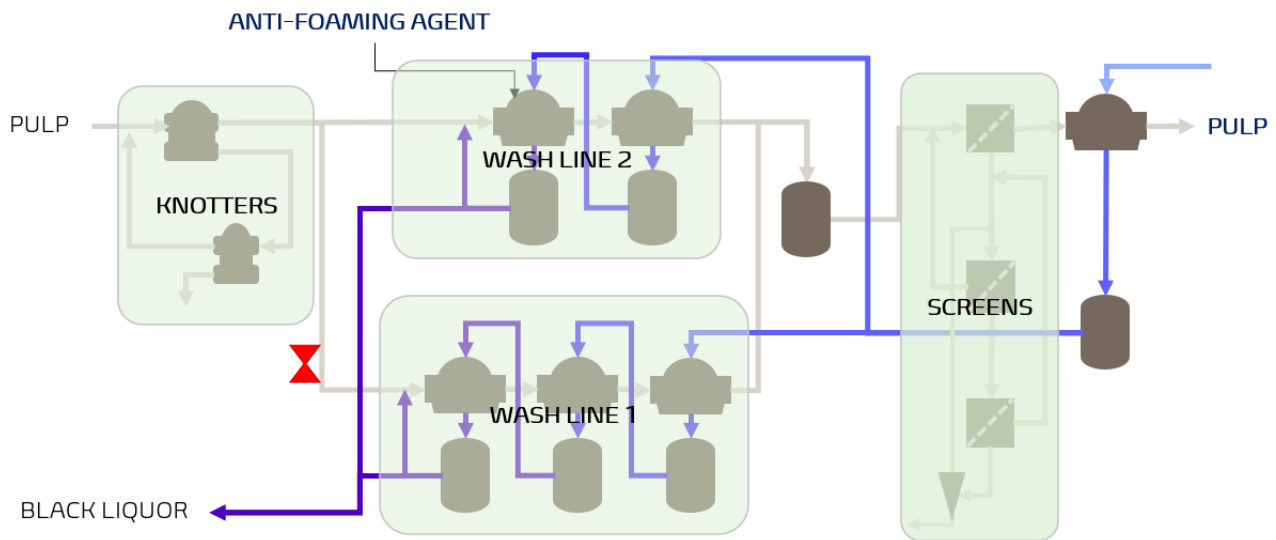


Figure 5.1 - Département de lavage de la pâte

Ce département, essentiel au bon fonctionnement de l'usine, présente plusieurs défis d'opération; c'est l'un des secteurs du procédé qui varie le plus, il présente d'importantes fluctuations. En effet, ce département, qui opère en continu, est assez instable étant donné le département de mise en pâte qui le précède qui opère en lot (décharge à chaque 90 minutes). Le réservoir de soufflage à l'entrée du département n'est pas suffisant pour contrer les fluctuations – ces fluctuations se reflètent sur le niveau du réservoir. Également, la faible capacité du département de lavage fait en sorte qu'afin de laver toute la pâte sortant des lessiveurs, les laveurs sont opérés à vitesse élevée, ce qui les rend plus difficiles à contrôler – la vitesse à laquelle les lessiveurs fonctionnent influence la vitesse d'opération des laveurs; ces derniers ne doivent pas ralentir les lessiveurs. Le département de lavage se trouve donc à être un important goulot d'étranglement. Cependant, opérer les laveurs à leur capacité maximale pourrait faire perdre le contrôle des opérations et ainsi réduire la qualité du lavage de la pâte, faire en sorte qu'une trop grande quantité de lignine reste dans la pâte et que cette dernière se retrouve au blanchiment et que les coûts des agents de blanchiment soient beaucoup plus élevés.

Aussi, les opératrices et opérateurs ont peu de marge de manœuvre dans ce département; afin de laver la pâte correctement, il faut mettre tout juste assez d'eau afin de ne pas avoir trop de filtrat (et donc de liqueur noire, qui va impacter les autres départements), mettre la bonne quantité d'antimousse et tenter de suivre la cadence des lessiveurs (trouver la bonne vitesse d'opération).

Ainsi, une stratégie employée pour optimiser de façon incrémentale les opérations du département de lavage est d'identifier des opportunités nécessitant peu ou pas d'investissement en capital, basé sur l'analyse des données (section 5.2), la considération des régimes d'opération (section 5.3) et les coûts de production (section 5.4).

5.2 Données industrielles

Le type d'analyse et son objectif premier vont dicter la fréquence de stockage des données de capteurs collectées à travers le procédé. Les données utilisées dans cette étude de cas sont stockées dans le système PI d'AVEVA – un serveur de base de données conçu spécifiquement pour le stockage des données de procédé qui permet un accès facile et fiable aux données par le personnel de l'usine – sur la base de moyennes aux 10 minutes.

La fréquence de stockage dans le système PI ne doit pas être assimilée à la fréquence d'échantillonnage; certaines variables sont échantillonnées à un intervalle supérieur à 10 minutes alors que d'autres ont été interpolées pour s'adapter à la fréquence de stockage de 10 minutes – cela se produit lorsque la fréquence d'échantillonnage est inférieure à la fréquence de stockage. En ce qui concerne les variables échantillonnées plus rapidement qu'aux 10 minutes, la compression est utilisée pour économiser l'espace de stockage des données en n'enregistrant que les valeurs qui diffèrent significativement de la valeur stockée précédente (au-dessus d'un pourcentage spécifié) ou en prenant la moyenne sur un laps de temps donné (10 minutes dans cette étude).

D'autre part, sachant que les expertes et experts du procédé abordé dans cette étude de cas considèrent l'opération du procédé comme stable lorsque les valeurs des variables de références restent sensiblement identiques pendant au moins 12 heures, effectuer des analyses avec des moyennes aux 10 minutes est amplement suffisant (cela permet de ne manquer aucune information). De façon générale, les expertes et experts vont commencer leurs analyses avec des moyennes aux heures et lorsqu'un phénomène intéressant est observé, une analyse plus approfondie sera faite à l'aide de moyennes aux 10 minutes (laps de temps le plus court considéré).

5.3 Régimes d'opération

Étant donné les changements subtils qui ont lieu au sein des opérations et la difficulté à les identifier, la combinaison d'une bonne connaissance du procédé et l'utilisation des approches basées sur les données sont de mises pour la détection des régimes d'opération. Avant de procéder à l'analyse des données, la connaissance du procédé de lavage de la pâte permet d'avancer que durant l'été, l'usine fonctionne à un taux de production plus élevé et le département du lavage est le facteur limitant. Ainsi, la capacité maximale des équipements de ce département est souvent atteinte et cela engendre de moins bonnes performances de lavage. En hiver, les taux de production sont plus bas, cela amène une meilleure stabilité aux opérations – le débit de lavage est plus facile à contrôler; le département de lavage n'est pas le goulot d'étranglement. Ainsi, le fait d'avoir moins de pâte à traiter amène le département de lavage à avoir une meilleure performance en hiver. Cependant, avoir une plus faible production n'est jamais souhaitable au niveau économique.

Ainsi, de façon générale au sein de l'usine, les saisons jouent un rôle critique pour l'opération du procédé. Cela va impacter le nombre de lessiveurs utilisés (lots produits) chaque jour (à l'hiver, moins de lessiveurs sont utilisés) – le nombre de lessiveurs est directement lié au taux de production de pâte, et la température de l'eau de procédé qui est plus froide à l'hiver qu'à l'été.

Outre les variations saisonnières et périodiques (tendances cycliques), le procédé de pâte à dissoudre considéré dans le cadre de cette recherche fabrique principalement un seul produit. Cependant, différents grades sont produits, mais les changements sont uniquement au blanchiment; pour les lessiveurs et laveurs, une recette unique est considérée.

Finalement, le département de lavage comporte deux lignes de lavage et selon les opérations en amont et en aval, elles peuvent être utilisées simultanément ou en alternance. Cela impacte le taux de production du département. Lorsqu'une seule ligne est en opération, cela s'explique soit par le fait qu'une des deux lignes est brisée ou que le procédé opère au ralenti à cause d'un autre secteur de l'usine – ce n'est donc jamais une situation désirable. Étant donné que la ligne A est plus difficile à arrêter et à redémarrer, que ses laveurs sont plus récents et efficaces, et qu'elle a la plus grande capacité (environ 60% de la capacité), c'est souvent elle qui continue de rouler.

5.4 Coût de production

Afin de concrétiser l'approche systématique proposée dans le cadre de cette recherche, les coûts d'opération de chaque régime sont considérés. Ces derniers ne sont pas évalués de façon absolue, mais plutôt par rapport à la production moyenne de pâte à dissoudre, qui est le référentiel ou la base de référence. Plus précisément, sachant que l'objectif de l'usine est de produire le plus de tonnes de pâte possible, l'analyse économique s'attarde au profit engendré pour les tonnes additionnelles (de plus que la moyenne) de pâtes produites. Également, il a été démontré par l'usine que les recettes engendrées par la vente de la pâte produite couvrent toujours la consommation additionnelle d'antimousse et de produits chimiques nécessaires au blanchiment. Chaque tonne de pâte produite en haut de la moyenne amène un profit incrémental. Ainsi, le directeur d'usine aimerait connaître la meilleure façon d'opérer afin de faire le plus d'argent possible. Il s'agit donc de trouver le régime d'opération qui est le plus rentable.

Chaque façon différente d'opérer se voit apposer un profit obtenu par rapport à la base de référence, qui est estimée à 475 tonnes/jour. Les ressources prises en compte dans cette analyse de coûts axée sur les

opérations sont la quantité moyenne d'antimousse et de produits chimiques de blanchiment consommés au cours des différents régimes d'opération; le service de lavage n'étant pas énergivore, le coût de l'énergie n'a pas été pris en compte. Les activités qui consomment ces ressources sont l'utilisation des laveurs et les réactions qui se produisent dans les tours de blanchiment. Le *cost object* est le département de lavage.

Également, en considérant les coûts de production absolus par tonne de pâte produite et leurs répartitions en coûts directs et indirects, il est possible d'identifier que pour chaque tonne de pâte supplémentaire produite par rapport à la moyenne, l'usine fait un gain de 500 \$. De ce fait, en soustrayant les coûts à ce gain, il est possible d'obtenir le profit engendré par chaque régime d'opération en considérant la quantité de pâte produite au-dessus de la moyenne et de la quantité de matière première consommée.

CHAPITRE 6 SOMMAIRES EXÉCUTIFS DES PUBLICATIONS

6.1 Présentation des publications

Les publications suivantes (présentées respectivement aux Annexes B à F de cette thèse) ont été soumises à des revues scientifiques évaluées par des pairs :

- ❖ Thibault, É., Chioua, M., McKay, M., Patience, G., Stuart, P. *Experimental Methods in Chemical Engineering : Data Processing and Data Usage in Decision-Making*. Canadian Journal of Chemical Engineering, 2023.
- ❖ Thibault, É., Lebreux Désilets, F., Poulin, B., Chioua, M., Stuart P. *Comparison of signal processing methods considering their optimal parameters using synthetic signals in a heat exchanger network simulation*. Computers and Chemical Engineering, 2023.
- ❖ Thibault, É., Lebreux Désilets, F., Poulin, B., Chioua, M., Stuart P. *Signal processing heuristics for the selection of wavelet transform parameters considering signal characteristics derived from a case study on a chemical pulp mill*. Computers and Chemical Engineering, 2023.
- ❖ Thibault, É., Kelly, J. D., Lebreux Désilets, F., Chioua, M., Poulin, B., Stuart, P. *An operating regime-based data-driven industrial framework for improved process decision-making – Part 1: Framework development*. Industrial and Engineering Chemistry Research, 2023.
- ❖ Thibault, É., Ledoux, C., Stuart, P. *An operating regime-based data-driven industrial framework for improved process decision-making – Part 2: Framework application considering activity-based costing concepts*. Industrial and Engineering Chemistry Research, 2023.

6.2 Relation entre les publications

Des entrevues menées auprès d'expertes et experts dans le domaine des pâtes et papier (en usines et en entreprises fournissant des technologies d'analyse des données) ont donné lieu à une première publication qui aborde la façon dont les données de procédés sont utilisées et traitées dans le cadre d'applications distinctes. Différentes consultations et webinaires ont été effectués en lien avec ce premier rendu. Une attention plus particulière sur le retrait des erreurs aléatoires dans les données a mené à la comparaison des quatre approches fréquemment utilisées pour le retrait du bruit, c'est-à-dire le filtre de Kalman, la moyenne mobile à pondération exponentielle, la transformée de Fourier à court terme et la transformée

en ondelettes (seconde publication). S'en suit une troisième publication qui s'intéresse à la formulation d'heuristiques afin de traiter les signaux de façon systématique et efficace. En considérant les autres étapes de traitement des données (autre que le retrait des erreurs aléatoires), un cadre méthodologique de traitement des données de procédé qui met de l'avant la connaissance du domaine et des approches axées sur les données est proposé dans une quatrième publication. Finalement, un dernier article, qui se trouve à être la seconde partie de l'article #4, décrit une étude de cas qui présente un problème d'optimisation du département de lavage qui nécessite la connaissance des régimes d'opération et des coûts d'opération. Ces derniers sont évalués avec une étude des coûts basée sur les activités (*activity-based costing*). La Figure 6.1 présente les différentes publications et les activités connexes réalisées dans le cadre de ce projet doctoral.

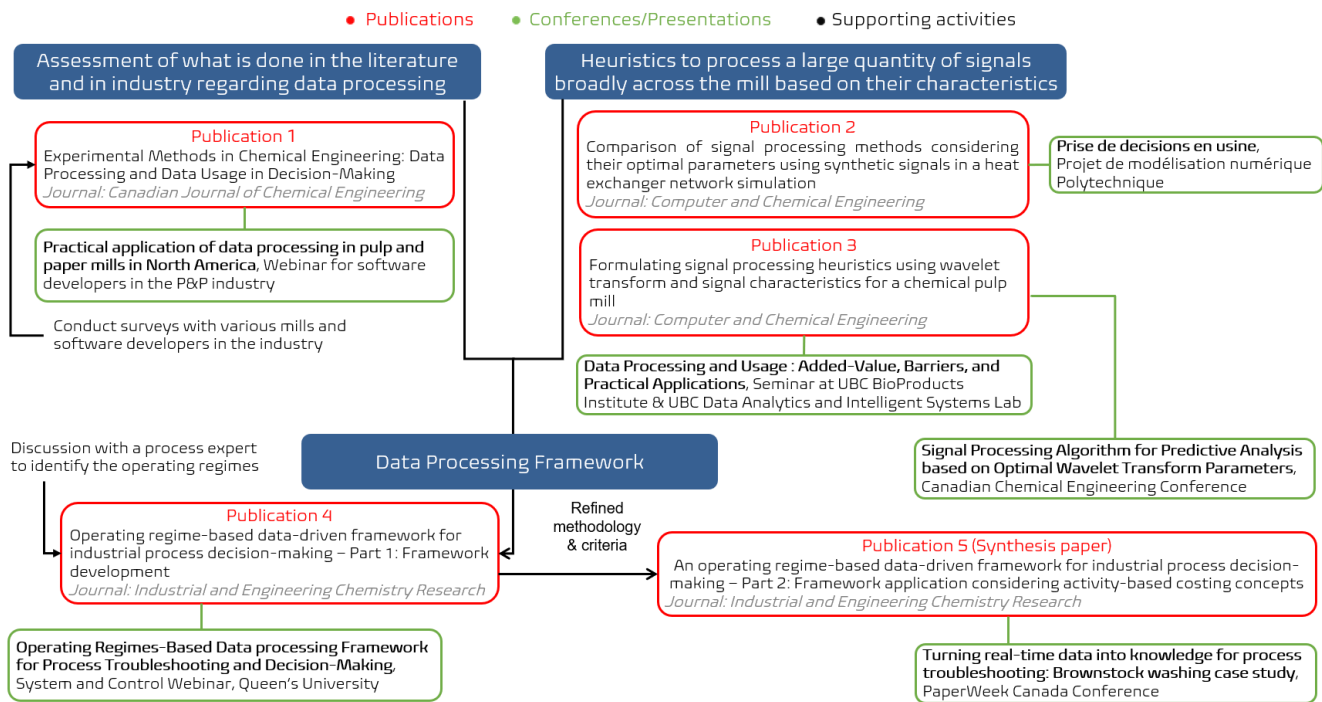


Figure 6.1 – Représentation visuelle des différentes publications et activités doctorales

6.3 Synthèse des résultats

Cette section présente une vue d'ensemble des résultats les plus pertinents des travaux effectués dans le cadre de ce projet de recherche doctorale afin de mettre en évidence les contributions majeures au domaine.

6.3.1 Revue des pratiques de l'industrie en matière de traitement et d'utilisation des données dans la prise de décision

En fonction des objectifs de l'analyse et de l'application, le traitement des données diffère. Ainsi, dans le cadre d'une application spécifique, sans traitement préalable des données, il est probable que les ingénieures et ingénieurs ne parviennent pas à la bonne décision prise sur la base des données de procédé. Au cours des entrevues, une personne spécialiste au niveau des usines de pâte et papier a mentionné que sans un traitement des données soigneusement planifié, des ressources importantes sont gaspillées.

Les personnes expertes de l'usine ont été interrogées sur la façon dont elles utilisent les données de procédé et comment elles les traitent pour chaque application. Les applications visées étaient l'analyse comparative, l'analyse des défaillances (détection et diagnostic de fautes), la surveillance des conditions, l'analyse des coûts, l'optimisation des procédés, l'analyse énergétique et la simulation des procédés. Lorsqu'il leur fut demandé de classer les applications en fonction de leur valeur ajoutée, il a été observé que le personnel des usines voit clairement la valeur de l'analyse des défaillances, ce qui est cohérent avec le fait qu'il se concentre principalement sur la production. Bien que la simulation de procédé ne soit pas une application en soi, mais plutôt un outil, elle apporte une valeur considérable à l'usine, surtout si elle inclut l'analyse des coûts. Cette dernière est l'application de données la plus importante pour les usines; c'est ce qui les maintient en activité. Le Tableau 6.1 présente les détails de différentes applications et exemples où les données de procédé sont utilisées et comment elles ont été traitées.

Tableau 6.1 – Résumé de la manière dont l'industrie utilise ses données pour la prise de décision et comment elles sont traitées

Catégorie	Application (exemple en usine)	Traitement des données	Considération des régimes d'opération
Benchmark/ Analyse des défaillances/ Surveillance des conditions	Identifier les changements au sein du procédé par rapport aux objectifs et aux performances historiques. S'assurer que les indicateurs clés de performance opèrent dans les limites de tolérance. Prédire les performances et les échecs futurs afin de prendre des mesures correctives proactives.	Collecte des données de procédé, en utilisant le 90e percentile des trois dernières années pour définir le point de référence par grade, et une fois par jour, comparaison des données réelles moyennées à ces valeurs de référence. Les moyennes sur une longue période sont utilisées pour détecter un changement.	Aucune
Analyse des coûts	Coût de production et de maintenance. Tableaux de bord. Coût des produits chimiques en fonction du grade produit. Comparer la valeur du coût avec la valeur budgétée.	Réalisé manuellement dans Excel; suppression des valeurs aberrantes et des conditions perturbées, puis calcul de moyennes mobiles de 15 minutes sur 30 jours (filtrage du bruit). La moyenne est calculée sur de bonnes données nettoyées afin qu'elle puisse être considérée comme représentative. Les filtres sont les suivants: vérification des limites, indicateur d'exécution (si le procédé est en cours d'exécution), filtres conditionnels.	Aucune
Analyse énergétique (efficacité)	Consommation et distribution d'énergie. Rendement des chaudières. Suivi de la consommation de vapeur dans le temps dans Power BI. Analyse autour des échangeurs de chaleur.	Téléchargement manuel dans Excel, filtrage des mauvaises données, filtres conditionnels. Utilisation des filtres dans Asset Framework pour la vérification des limites, indicateur d'exécution (si le procédé est en cours d'exécution), s'agit-il d'une opération normale. Prenez une moyenne de 5 minutes.	Identification des régimes d'opération afin de prendre des décisions pour la gestion de l'énergie (saisons).
Optimisation des procédés	Optimisation des coûts lors des audits. Réaliser des bilans énergétiques et massiques pour la vapeur et l'électricité afin d'optimiser les sources d'énergie et de gérer les changements saisonniers.	Dans Excel: application de filtres, tri et nettoyage des données, suppression des valeurs nulles, erreurs dans l'interface de saisie (timeout, mauvaise saisie, code d'erreur, perte de connexion, etc.) Fait manuellement par une personne experte. Les filtres sont les suivants : vérification des limites, indicateur d'exécution (si le procédé est en cours d'exécution), s'agit-il d'une opération normale, filtres conditionnels.	Le coût de fabrication du produit est plus difficile à estimer, car il existe différents grades et spécifications.
Simulation de procédés	Bilan de masse et d'énergie pour la prise de décision. Surveiller les différences entre le modèle et les valeurs réelles et signaler les défaillances. Fournir des capteurs virtuels.	Effectué manuellement dans Excel; suppression des valeurs aberrantes et des conditions perturbées, puis moyennes de 15 minutes sur 30 jours (filtrage du bruit). Filtres conditionnels dans Asset Framework également. Les filtres sont la vérification des limites, indicateur d'exécution (si le procédé est en cours d'exécution), s'agit-il d'une opération normale, les filtres conditionnels.	Les personnes expertes en usine prennent en compte les différents grades et les différents types de copeaux de bois (feuillus ou résineux), et elles simulent à la fois les taux de production normaux et les taux de production minimums.

L'analyse comparative, l'analyse des défaillances ou la surveillance de l'état des équipements comprennent l'examen rétrospectif (sur une longue période), la comparaison de l'opération aux valeurs de performance optimales en régime permanent, la vérification que les équipements fonctionnent dans leur meilleure plage d'opération et l'indication de la dérive des capteurs à l'aide d'Excel, de PI ProcessBook et de PARCview. Ces applications détectent si les moyennes sur certaines périodes sont décalées par rapport à ce qu'elles ont été historiquement, ou en d'autres termes, si elles correspondent au modèle de régime permanent. Elles utilisent les données du système PI d'AVEVA pour calculer des indicateurs clés de performance quotidiens et les comparent à ceux basés sur les trois meilleurs mois consécutifs ou sur le 90^e percentile des trois dernières années. Les types d'analyse qui portent sur plusieurs années de données sont généralement effectués davantage dans le cadre d'audits de procédé que pour la surveillance en continu.

En outre, avant l'analyse des coûts, comme c'était le cas pour la simulation de procédé, le personnel de l'usine effectue une analyse visuelle, en recherchant ce qu'il considère comme une bonne durée d'exécution à l'aide de la fonction de tri et en supprimant les données indésirables. Les filtres d'Asset Framework sont également utilisés pour créer de nouvelles variables dans le PI System d'AVEVA avec les valeurs filtrées. Ainsi, toute analyse ultérieure peut passer de l'utilisation des valeurs erronées provenant des mesures à ces valeurs filtrées lorsqu'elles sont disponibles. D'autre part, l'analyse énergétique est effectuée pour s'assurer que les échangeurs de chaleur ne s'encrassent pas, ou pour évaluer s'ils doivent être nettoyés parce qu'ils ont commencé à s'encrasser.

Avant d'introduire les données dans un algorithme d'optimisation hors ligne, le personnel de l'usine filtre les données indésirables. Il utilise principalement des données moyennées, car il a peu ou pas de confiance envers les données en temps réel et il cherche à ce que les données de procédé soient à la même fréquence que les informations sur les coûts qui ne sont pas disponibles en temps réel. D'autre part, les simulations de procédé sont des outils d'aide à la décision pour l'opération des procédés. Certaines usines simulent l'ensemble de l'usine dans CADSIM Plus pour évaluer les opérations, les rejets, le rendement et la performance des équipements. La plupart du temps, la simulation compare la performance actuelle à une performance simulée pour voir si quelque chose ne va pas. Pour la simulation de procédé, la moyenne des données est effectuée sur de bonnes données nettoyées, de sorte que ces dernières puissent être considérées comme représentatives. Pour ce faire, le personnel de l'usine recherche une période de relativement stable (fonctionnement relativement constant), par exemple, 24h de fonctionnement stable

et sans aberration. Les données sont prétraitées par une analyse visuelle, en utilisant également la fonction de tri et en supprimant les mauvaises données. Pour la simulation, le personnel de l'usine choisit une période (plutôt longue) sur laquelle il fait une moyenne basée sur les tendances. Les simulations de procédé en régime permanent sont également utilisées pour la conception et considèrent généralement au moins deux conditions d'opération : l'été et l'hiver. Si l'usine réalise un projet de modernisation (rétro-installation), la définition du régime d'opération est essentielle pour réduire les risques pour le procédé existant et pour obtenir une meilleure estimation du retour sur investissement.

Dans CADSIM Plus, les données bruitées sont prises en charge par des techniques de réconciliation dynamique des données et, en ce qui concerne les valeurs aberrantes, il utilise une valeur moyenne attendue, des conditions limites, et tout ce qui dépasse cette valeur ne sera pas pris en compte. Plus une mesure s'écarte de cette valeur typique, moins cette mesure a d'importance en termes de poids; il est possible d'ajouter un poids aux données qui est lié à la confiance accordée à la mesure et à sa dégradation dans le temps. C'est à ce moment que les connaissances des ingénieures et ingénieurs de procédés deviennent importantes, pour tout mettre en place correctement. CADSIM Plus a développé sa propre méthode basée sur les principes fondamentaux pour la réconciliation des données. La détection des périodes en régime permanent se fait manuellement, et la détection des régimes d'opération se fait à l'aide de la logique.

En guise d'ouverture pour une analyse plus approfondie, il est intéressant de constater que les usines sont principalement en mode dépannage, résolution de problèmes, et que les répondantes et répondants au sein des usines ne mentionnent encore que très peu les fonctionnalités de prédiction. Les données sont encore perçues comme la boîte noire d'un avion; après un écrasement, elles sont utilisées, mais pas de façon systématiquement pour anticiper les événements ou les dérives de performance. La valeur des données pourrait être étendue à des analyses prédictives et prescriptives. Anticiper que quelque chose d'anormal va se produire et donner des instructions pour optimiser les opérations pourrait avoir un impact économique énorme sur le procédé. Les capacités technologiques existent aujourd'hui, mais elles n'ont pas été largement adoptées par l'industrie des pâtes et papiers.

Parallèlement à ce qui est fait en industrie pour traiter les données de procédé, la littérature propose différentes approches pour le retrait des erreurs aléatoires qui sont comparées à la section 6.3.2.

6.3.2 Comparaison des approches de traitement des signaux

Les résultats du débruitage par le filtre de Kalman, la moyenne mobile à pondération exponentielle, la transformée de Fourier à court terme et la transformée en ondelettes sont comparés selon l'erreur quadratique moyenne et le ratio signal/bruit. La technique de traitement du signal idéale serait celle qui présente la valeur la plus élevée pour les deux critères, c'est-à-dire qui affiche la plus grande amélioration du SNR et la plus grande réduction du RMSE. Les Figure 6.2 et Figure 6.3 présentent les résultats de RMSE et de SNR, notés respectivement comme $RMSE_{réduction}$ et $SNR_{amélioration}$ pour les quatre méthodes de filtrage pour tous les signaux de température et de débit.

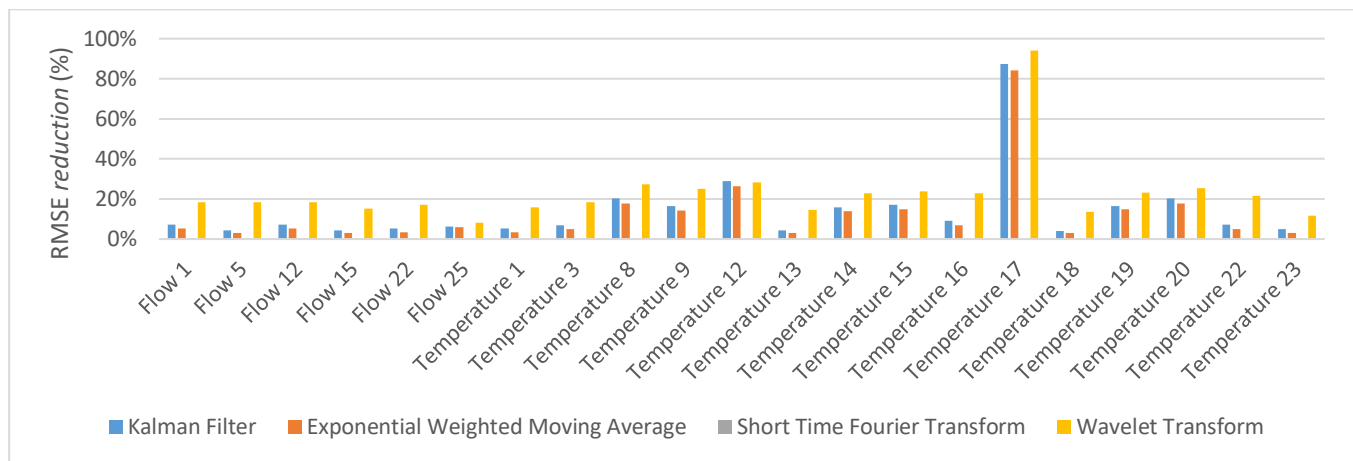


Figure 6.2 – Réduction de l'erreur quadratique moyenne pour les signaux synthétiques considérant le KF, la MMPE, la TFCT et la TO

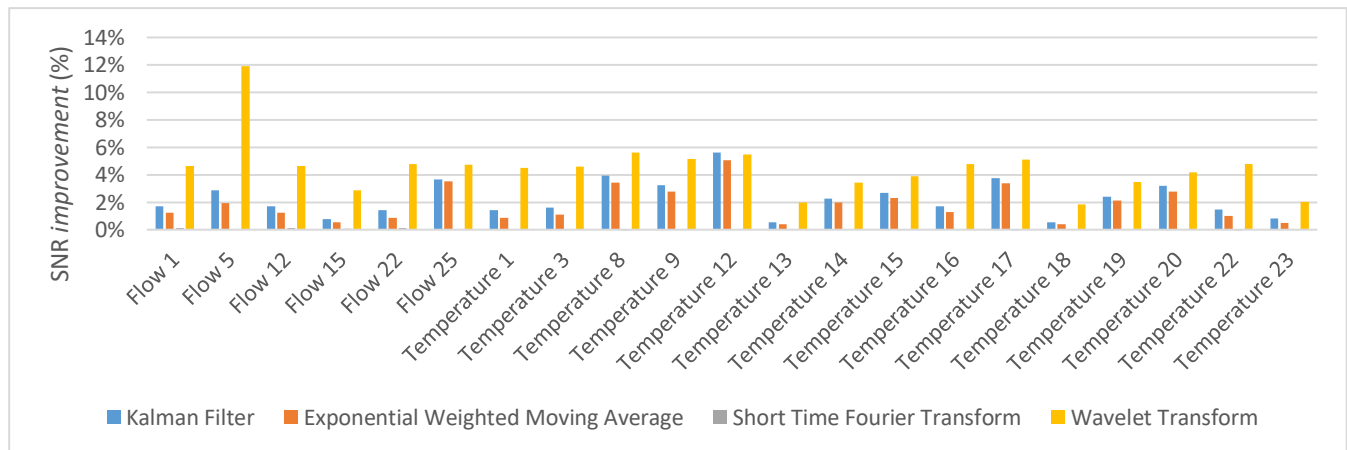


Figure 6.3 – Amélioration du rapport signal/bruit pour les signaux synthétiques considérant le KF, la MMPE, la TFCT et la TO

Les Figure 6.2 et Figure 6.3 présentent respectivement la comparaison de différentes techniques de traitement du signal en termes de réduction de l'erreur quadratique moyenne et d'amélioration du rapport signal/bruit en pourcentage lors du traitement de signaux synthétiques de température et de débit provenant d'une simulation de réseau d'échangeurs de chaleur. La TO domine en termes de réduction de RMSE, c'est la méthode la plus performante pour la plupart des signaux, à l'exception de quelques signaux (température 12 et débit 25) où la TO est aussi performante que les autres techniques. En ce qui concerne l'amélioration du SNR, la TO est l'approche à privilégier pour tous les signaux, à l'exception de la température 12 où l'amélioration du SNR est de 5,60% pour le FK et de 5,45% pour la TO. La TFCT semble être systématiquement la technique la moins efficace pour la réduction du RMSE et l'amélioration du SNR.

La réduction de l'erreur quadratique moyenne varie sur une large plage, allant de valeurs nulles pour la TFCT à près de 100 % pour la TO. Cela signifie que les performances des techniques de traitement du signal variaient fortement, et il en va de même pour l'amélioration du SNR. Enfin, les deux critères ont conduit aux mêmes conclusions : la transformée en ondelettes est la plus efficace pour réduire le bruit gaussien des signaux de procédés simulés comparée au filtre de Kalman, à la moyenne mobile à pondération exponentielle et à la transformée de Fourier court terme. Le débruitage avec la TO a obtenu l'erreur quadratique moyenne la plus faible et le rapport signal/bruit le plus élevé la plupart du temps, d'où la plus grande différence avec le signal original.

Toutes les techniques de traitement ont présenté des valeurs de réduction de l'erreur quadratique moyenne supérieures aux valeurs d'amélioration du rapport signal/bruit, ce qui signifie qu'elles ont toutes réduit la différence entre les signaux originaux et ceux d'intérêt, mais que les signaux sont toujours bruités. Il n'est pas possible de reconstruire parfaitement un signal sans bruit, cependant un filtre approprié donnera la tendance des données plutôt que leur variation instantanée. Par conséquent, dans ce cas, les signaux filtrés doivent ressembler au signal d'intérêt.

Les deux critères sont toujours en accord; un signal qui a une valeur élevée de réduction de l'erreur quadratique moyenne a également une valeur élevée d'amélioration du rapport signal/bruit. La transformée en ondelettes a montré son efficacité en obtenant des résultats presque aussi bons pour les deux critères (RMSE et SNR), quel que soit le signal; la réduction de l'erreur quadratique moyenne est d'environ 23% et l'amélioration du rapport signal/bruit d'environ 5%. Le Tableau 6.2 présente les valeurs

moyennes de la réduction de l'erreur quadratique moyenne et de l'amélioration du rapport signal/bruit, ainsi que leurs écarts-types, pour tous les signaux (débits, températures et tous deux confondus). Les résultats montrent sans ambiguïté que les TO sont globalement significativement meilleures que les autres approches selon les deux critères. En fait, les TO montrent en moyenne une réduction de 14,14% de l'erreur quadratique moyenne et une amélioration de 3,12% du rapport signal/bruit par rapport au filtre de Kalman, la moyenne mobile à pondération exponentielle et la transformée de Fourier court terme pour tous les signaux.

Tableau 6.2 – Moyennes des critères (RMSE et SNR) avec leurs écarts-types pour toutes les approches

	Capteurs	FK	MMPE	TFCT	TO
Moyenne RMSE	Débit	5.79% ± 1.33%	4.31% ± 1.18%	0.21% ± 0.21%	15.88% ± 4.04%
	Température	17.62% ± 20.66%	15.55% ± 20.32%	0.01% ± 0.01%	25.88% ± 19.52%
	Tous	14.24% ± 18.15%	12.34% ± 17.79%	0.06% ± 0.14%	23.03% ± 17.09%
Moyenne SNR	Débit	2.02% ± 1.05%	1.56% ± 1.07%	0.08% ± 0.06%	5.59% ± 3.18%
	Température	2.35% ± 1.42%	1.96% ± 1.34%	0.01% ± 0.01%	4.05% ± 1.26%
	Tous	2.26% ± 1.31%	1.84% ± 1.26%	0.03% ± 0.05%	4.94% ± 2.03%

La Figure 6.4 montre, à gauche, le signal original TI20 (bleu), son signal d'intérêt (noir) et le signal filtré (vert) pour toutes les méthodes et, à droite, l'erreur de mesure (rouge) et l'erreur d'estimation (noir) de la méthode respective. La première évalue la différence ou la distance entre le signal original et le signal d'intérêt, et la seconde est la différence entre le signal original et le signal filtré.

La Figure 6.4a) montre une forte diminution du niveau de bruit et une erreur plus faible que celle présente à l'origine dans le signal. Le FK présente une amélioration considérable – une réduction de l'erreur quadratique moyenne de 20,17% et une amélioration du rapport signal/bruit de 3,21%. Même si les résultats présentés dans la Figure 6.4b) sont assez similaires à ceux obtenus avec le FK, le débruitage n'est pas aussi bon que celui obtenu avec le FK. En effet, la MMPE présente une réduction de l'erreur quadratique moyenne de 17,82% et une amélioration du rapport signal/bruit de 2,80%. La Figure 6.4c) présente une mauvaise exécution du filtrage avec la TFCT; il n'y a presque aucune différence avec l'erreur avant le filtrage (la réduction de l'erreur quadratique moyenne et l'amélioration du rapport signal/bruit sont respectivement de 0,04% et 0,03%). Dans la Figure 6.4d), il ne reste presque plus de bruit; le signal d'intérêt est pratiquement superposé au signal filtré. La TO offre le plus faible pourcentage d'erreur par rapport au signal d'intérêt; la réduction de l'erreur quadratique moyenne est de 25,39% et l'amélioration du rapport signal/bruit de 4,79%. L'amélioration de l'erreur d'estimation est clairement visible et se distingue de toutes les autres approches.

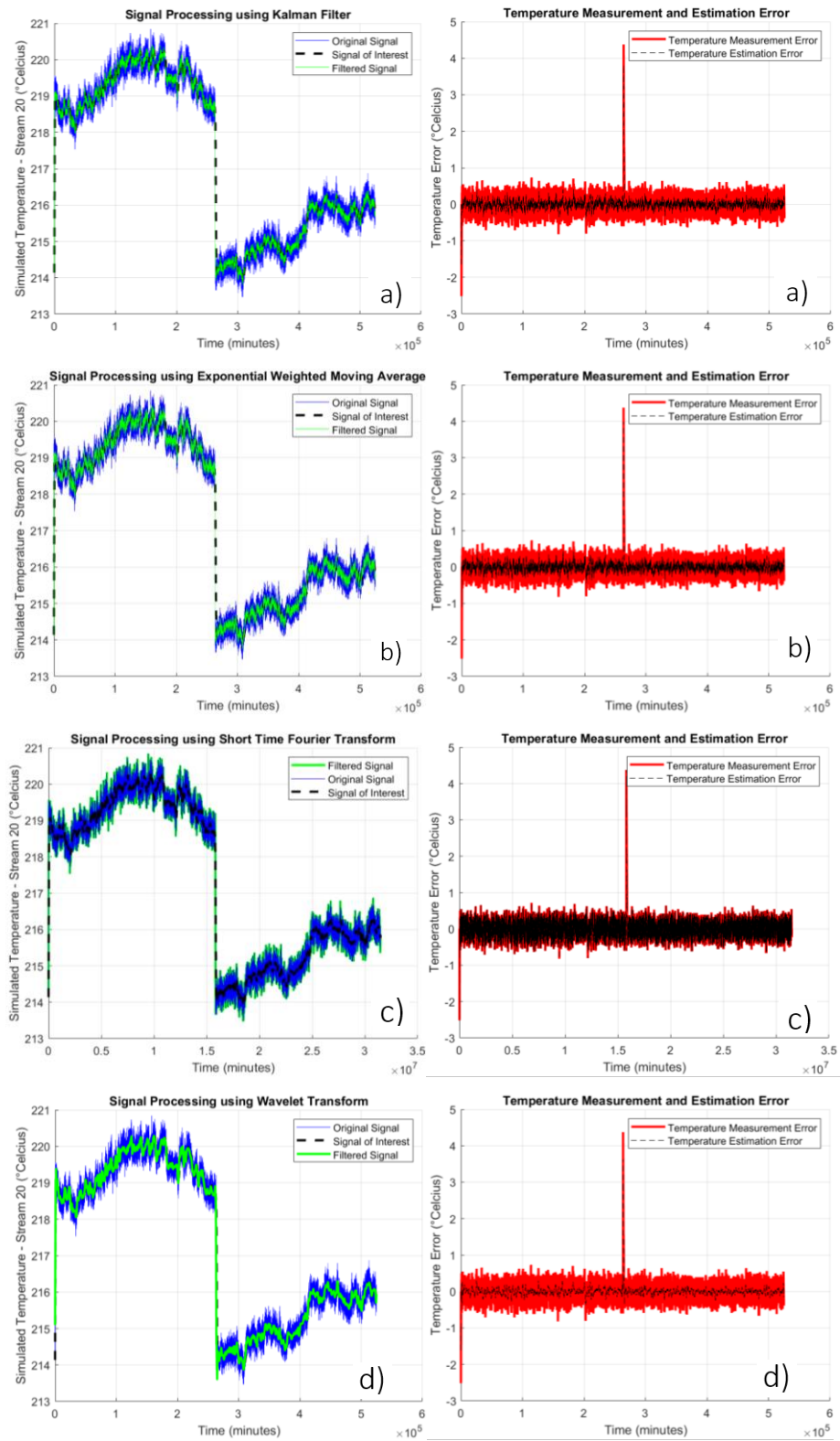


Figure 6.4 – Résultats du débruitage du signal TI20 en utilisant a) FK, b) MMPE, c) TFCT, et d) TO

6.3.3 Établissement d'heuristiques basées sur les transformées d'ondelette

Les valeurs optimales des paramètres de la transformée en ondelettes permettant d'obtenir la tendance quotidienne de chaque signal, ainsi que les paramètres de classification du signal et leurs caractéristiques sont étudiés dans une analyse en composantes principales afin d'en extraire les relations pertinentes.

En considérant une à une les caractéristiques des signaux ainsi que l'ensemble des paramètres de classification des signaux, l'analyse en composantes principales permet d'identifier et d'expliquer la variabilité ainsi que les variables qui influencent la caractéristique étudiée. Les observations suivantes sont ainsi obtenues :

1. Les signaux de température présentent des valeurs de coefficient de variation faibles alors que les signaux de niveau et de débit présentent des valeurs élevées. Par conséquent, la moyenne des signaux de la température est plus importante que son écart-type. Ceci s'explique par l'inertie thermique; le signal de température présente une variation plus lente.
2. Les signaux de débit et de consistance présentent des valeurs de variance élevées. Les signaux de débit et de consistance sont constitués de valeurs très éloignées de leurs valeurs moyennes, particulièrement comparativement aux signaux de niveau et de température. Ceci est justifié par le fait qu'ils présentent plus de variations – les mesures de consistance présentent généralement moins de variations que les mesures de débit.
3. Les signaux de débit semblent présenter des valeurs d'aplatissement élevées. Cela implique une probabilité plus élevée de présence de valeurs extrêmes (plus extrêmes que les $3\sigma \pm \mu$ habituels prévus par la distribution normale). À l'inverse, les distributions à faible aplatissement présentent généralement des données moins extrêmes que celles d'une distribution normale. Les signaux de débit sont très affectés et sensibles aux événements qui se produisent dans le procédé, et sont donc plus enclins aux valeurs aberrantes.
4. Les signaux de niveau présentent des valeurs d'asymétrie élevées, tandis que les signaux de température présentent des valeurs d'asymétrie faibles. Les valeurs élevées d'asymétrie positive des signaux de niveau indiquent que les valeurs extrêmes des données sont plus importantes, ce qui se caractérise par une valeur moyenne importante pour l'ensemble des données et qui conduit à une distribution inclinée vers la gauche. D'autre part, les faibles (négatives) valeurs d'asymétrie

des signaux de température indiquent que les valeurs de données extrêmes sont plus petites, que la moyenne diminue et que la distribution est inclinée vers la droite.

5. Rien de significatif ne peut être observé en ce qui concerne la fréquence de la tendance et le taux de variation.

Par conséquent, les signaux ont des caractéristiques distinctes en fonction de leur type de mesure. Les relations entre les paramètres de classification du signal et ceux de la transformée en ondelettes sont également étudiées. Le *score chart* de ces variables est considéré pour observer la présence de clusters (Figure 6.5). En outre, les contributions des variables sont explorées (Figure 6.6); les variables qui présentent une valeur de contribution élevée (valeur absolue) expliquent le plus la variabilité entre les variables.

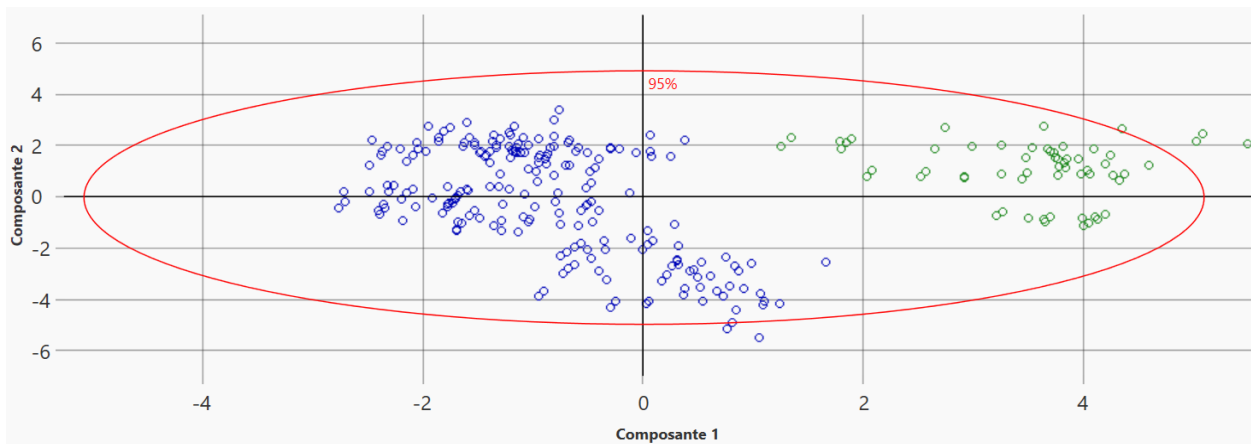


Figure 6.5 – Scores chart pour tous les paramètres de classification du signal et de TO

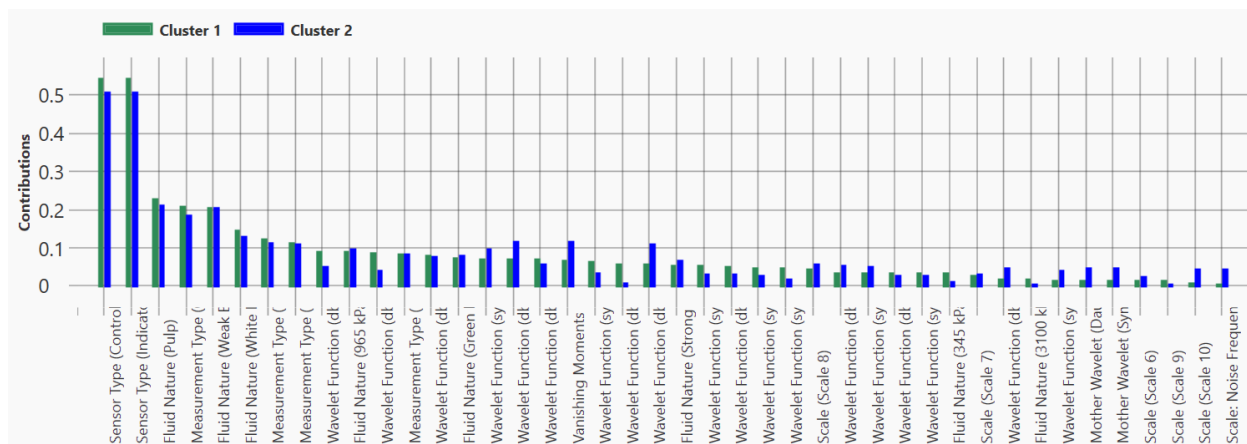


Figure 6.6 – Contributions des clusters présents dans le scores chart

Le *score chart* des paramètres de classification des signaux et de ceux de la transformée en ondelettes révèle la présence de deux groupes distincts (Figure 6.5) qui peuvent être expliqués par la Figure 6.6. Une analyse de contribution (Figure 6.6) a montré que le fait qu'un capteur fasse partie d'une boucle de contrôle ou qu'il soit uniquement un indicateur explique en grande partie ce phénomène. Ainsi, en se concentrant uniquement sur les capteurs se trouvant dans les boucles de contrôle, il est possible de détecter que le *score chart* présente deux clusters (Figure 6.7) qui sont principalement expliqués par le type de mesure. Cela peut s'expliquer par le fait que les différents types de mesure ont un impact différent sur le niveau de bruit qui affecte le signal, car leur dynamique est différente. Par exemple, en raison de l'inertie thermique, les signaux de température présentent une variation plus lente (ils ont des valeurs de fréquence plus basses) que les signaux de débit; les débits varient plus rapidement et présentent plus de composantes de haute fréquence (bruit) que les températures. Ensuite, en considérant uniquement les mesures de débit, il est possible de tirer des conclusions sur les paramètres d'ondelettes optimaux à utiliser pour chaque nature de fluide. Le même processus peut être accompli pour tous types de mesures.

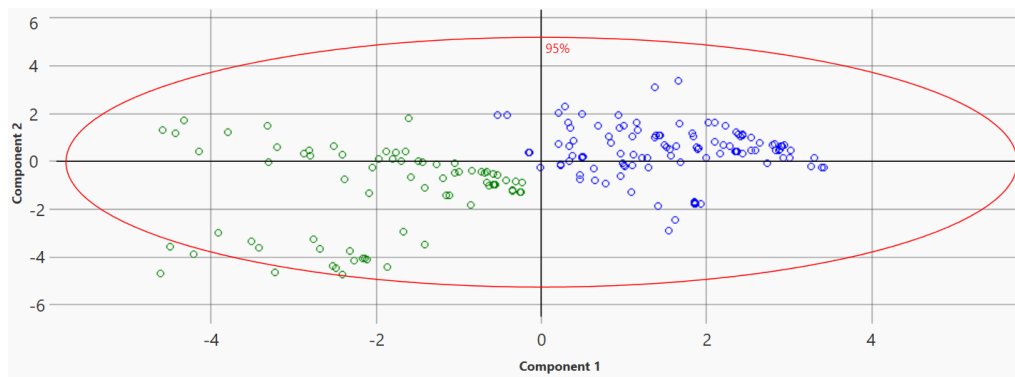


Figure 6.7 – Scores chart considérant uniquement les mesures des capteurs dans les boucles de contrôle

L'analyse en composantes principales a mis en évidence le fait que le type de capteur, le type de mesure et la nature du fluide sont identifiés comme les facteurs clés qui influencent la façon dont les signaux sont traités, c'est-à-dire le choix des paramètres du filtre (Figure 6.5, Figure 6.6, Figure 6.7). Il s'avère que les signaux provenant d'un type de capteur spécifique, tel qu'un indicateur, présentant un type de mesure spécifique, tel que la température, et mesurant un fluide particulier, la pâte par exemple, avaient des caractéristiques communes; les signaux sont caractérisés différemment en fonction de cela, ils doivent donc être traités en conséquence. De ce fait, des heuristiques qui capturent les corrélations entre les caractéristiques de classification du signal et les paramètres de filtrage sont formulées. Les heuristiques établies permettent le traitement d'une grande quantité de données de façon efficace et

systematique. En effet, ils fournissent une approche rapide pour déterminer les paramètres optimaux de la transformée en ondelettes qui pourraient être utilisés à grande échelle, et donc éviter le réglage manuel fastidieux de milliers de signaux industriels. Ainsi, les heuristiques minimisent le temps d'ingénierie accordé au traitement des données et faciliteront dans l'avenir le réglage des filtres, c'est-à-dire lors du traitement de tous signaux émis par les capteurs de l'usine.

Les heuristiques dérivées dans le cadre de cette recherche guideraient les expertes et experts de procédé dans la procédure de traitement du signal; elles pourraient être transmises au personnel de l'usine pour servir de point de départ à leur analyse, pour améliorer les processus de formation et raccourcir la courbe d'apprentissage liée au traitement des signaux. Ainsi, les heuristiques fournissent un encadrement rapide et peu coûteux vers des actions prometteuses et utiles; elles réduisent la complexité des tâches.

Les heuristiques sont formulées pour inclure le contexte dans lequel elles sont applicables, suivi d'une suggestion d'action à entreprendre. Par conséquent, dans le contexte du traitement des données de procédé industriel d'une usine de pâte à dissoudre à l'aide de la transformées en ondelettes,

1. la fonction de seuillage doux doit être considérée par défaut pour la suppression du bruit.
2. *universal threshold* doit être considérée par défaut pour la suppression du bruit.
3. l'arbre de classification de la Figure 6.8 doit être utilisé pour sélectionner la fonction d'ondelette et le niveau de décomposition. De ce fait, les paramètres optimaux de la transformée en ondelettes sont sélectionnés en fonction de la combinaison type de capteur/type de mesure/nature du fluide.

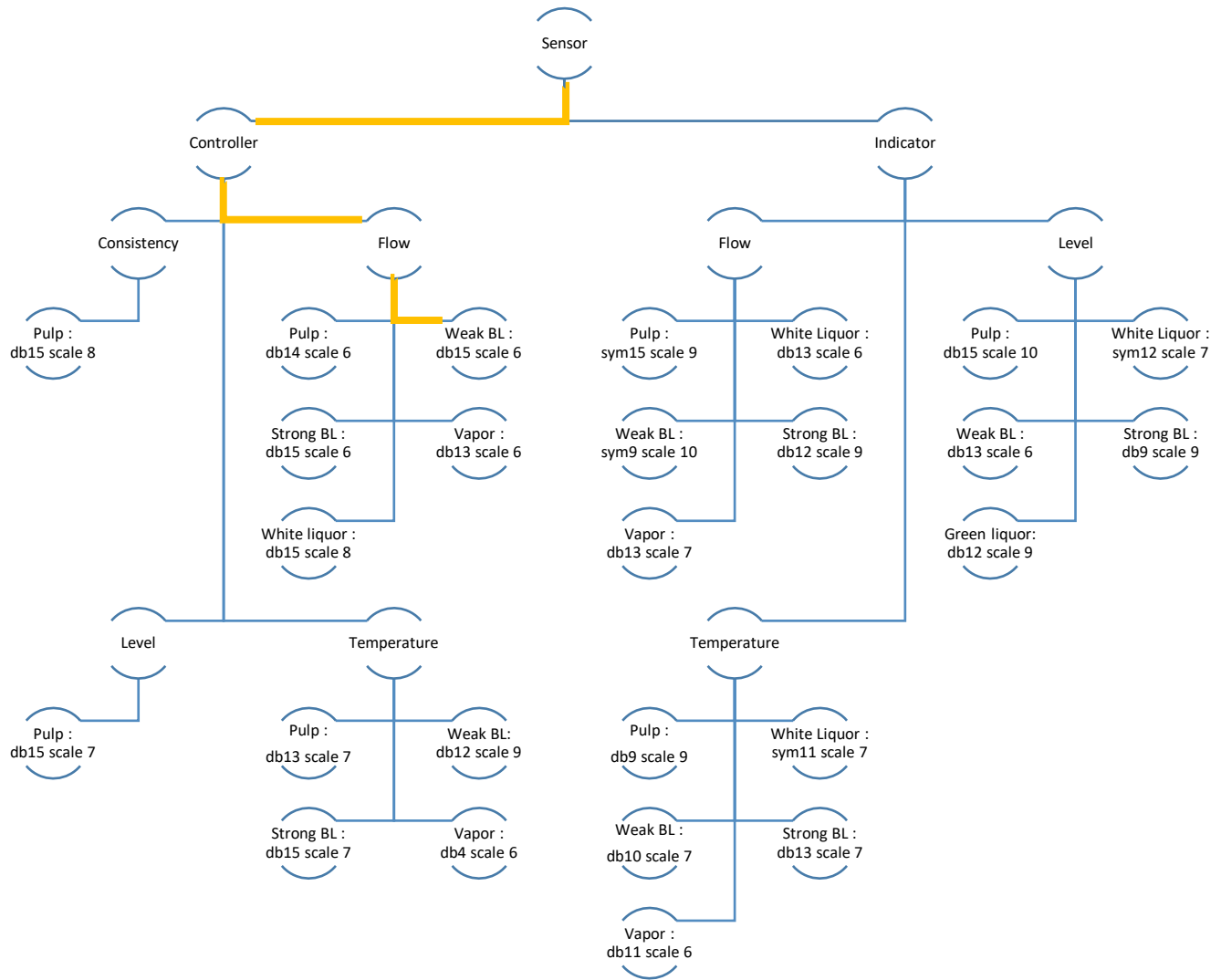


Figure 6.8 – Paramètres d'ondelettes optimaux en fonction des caractéristiques de classification du signal

6.3.4 Établissement d'un cadre méthodologique de traitement des données qui considère les régimes d'opération d'un procédé industriel

La Figure 6.9 illustre le cadre méthodologique proposé. Toutes les étapes de l'analyse des données considèrent la combinaison de la connaissance du domaine et des approches guidées par les données. Tout d'abord, étant donné que les données sont toujours traitées en fonction d'un objectif spécifique et ciblé, l'objectif de l'analyse (le champ d'application) est défini, et les données sont collectées. Ensuite, les données sont traitées, c'est-à-dire qu'elles subissent une étape de synchronisation, une correction du

décalage, les opérations anormales sont supprimées et les données sont filtrées pour réduire les composantes haute fréquence indésirables. La troisième étape consiste à détecter les séries de données en régime permanent qui permettront l'étape de réconciliation des données. Enfin, les différents régimes d'opération sont détectés et identifiés. Ce cadre méthodologique se limite aux données d'opération historiques (par opposition aux données de procédé en ligne), car elles constituent une source importante d'informations sur les procédés.

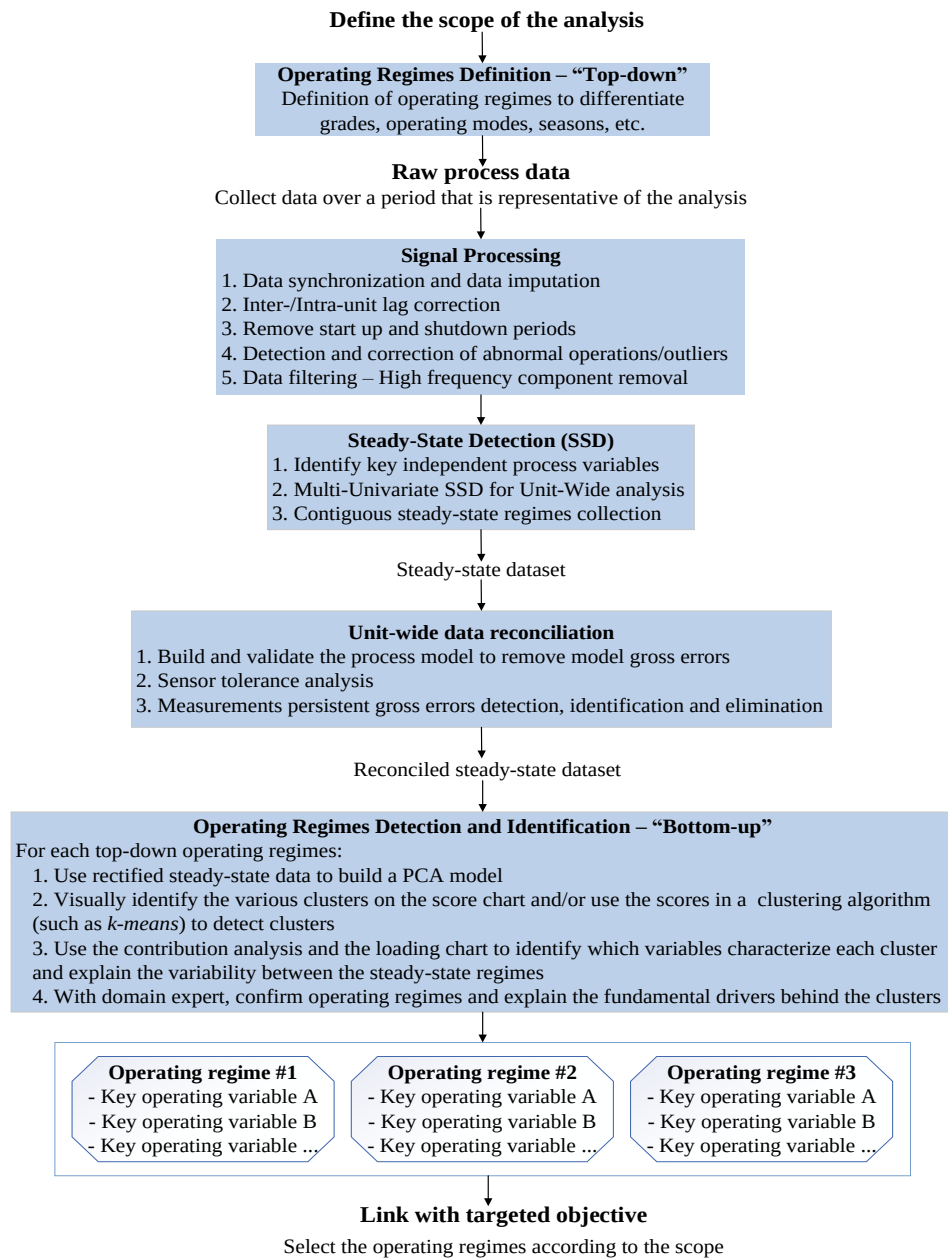


Figure 6.9 – Cadre combinant la connaissance du procédé et les approches axées sur les données

Ce cadre méthodologique met de l'avant la combinaison de la compréhension du procédé et des approches guidées par les données pour obtenir des connaissances enrichies sur le procédé et les utiliser pour prendre des décisions de gestion (hors ligne) éclairées. Les approches guidées par les données sont, entre autres, utilisées pour découvrir certains régimes d'opération et extraire des connaissances, alors que la connaissance du domaine contient des informations qui ne sont pas nécessairement visibles dans les données. En résumé, la méthodologie proposée exploite les avantages des approches axées sur les données et utilise la compréhension du procédé comme sources d'information complémentaires.

Définition de l'étendue et de la portée de l'analyse

Toute analyse de traitement des données commence par la définition des objectifs. Il est essentiel de mettre en avant les raisons pour lesquelles cette analyse est réalisée, à quoi servent les données et ce qui est étudié. Il faut donc considérer les régimes d'opération reconnus par les expertes et experts du procédé, c'est-à-dire les différentes recettes exprimées dans les procédures opérationnelles standard, qui sont liées à l'objectif de l'analyse. Les régimes d'opération sont utilisés pour répondre à des questions, résoudre des problèmes et prendre des décisions. En fonction de la question, de la problématique ou de la décision à prendre, des régimes d'opération spécifiques seront nécessaires.

Les régimes d'opération "*top-down*" sont déterminés par les changements de la stratégie de contrôle dû à différentes conditions d'opération (saisons, matières premières différentes, maintenance, utilisation de différentes lignes de production, etc.) Cependant, ces points de consigne ne tiennent pas compte de tout ce qui se passe au sein des procédés industriels. En fait, même si les régimes d'opération sont définis par des procédures d'opération standard, les points de consigne sont modifiés par les opératrices et opérateurs, car une usine n'est pas totalement automatisée. Le procédé peut être opéré en mode automatique ou manuel; certaines variables sont laissées en permanence en contrôle automatique, tandis que d'autres ne le sont pas. De plus, il est possible d'observer des différences au sein de certaines variables lorsqu'un équipement a besoin de maintenance, donc avant et après la maintenance. Également, les performances des équipements changent tout le temps dû à l'encrassement et la détérioration, et la recette doit donc être ajustée en conséquence. Les opératrices et opérateurs agissent donc lorsque quelque chose change au sein du procédé. En outre, deux personnes différentes opérant le procédé n'agiront pas forcément de la même façon. Ces considérations sont prises avec la détection des régimes d'opération "*bottom-up*" par une analyse de clustering et les causes sous-jacentes à ces régimes sont évaluées (la connaissance du procédé est requise pour comprendre pourquoi les variables agissent comme elles le

font). Dans la définition du champ d'application, seuls les régimes d'opération "*top-down*" sont considérés; l'idée est d'aller aussi loin que possible en utilisant les recettes selon l'objectif de l'analyse, puis ces régimes seront sous-catégorisés en utilisant les clusters.

Enfin, les données brutes sont collectées sur une période représentative pour l'analyse ciblée.

Traitement des données

Les données doivent être synchronisées au même pas de temps et de préférence, à la fréquence d'intérêt, car elles sont souvent stockées sur la base du changement de valeur, qui n'est pas une fréquence régulière, et les variables utilisées dans une analyse ne sont pas nécessairement échantillonnées au même rythme. Ainsi, si la fréquence d'échantillonnage d'une variable est inférieure à la fréquence d'intérêt pour une analyse, une imputation des données doit être effectuée (interpolation linéaire des données, répétition de la dernière valeur valide) (Huang, 2021). En revanche, si la fréquence d'échantillonnage est supérieure à la fréquence d'intérêt, il faut alors considérer la moyenne, la valeur la plus proche, la valeur minimale ou la valeur maximale sur un intervalle d'échantillonnage égal au pas de temps souhaité, selon l'analyse.

Une fois que le processus de synchronisation pour fixer les données sur un intervalle de temps régulier est effectué, le décalage est pris en compte. Les décalages tiennent compte du temps nécessaire aux opérations, à la rétention et au transport, et varient en fonction du taux de production. En ignorant les décalages, certaines corrélations peuvent être perdues. L'analyse de correction des décalages est réalisée sur la base d'une analyse de Fourier (un logiciel comme EXPLORE (Canada, 2022) offre cette fonctionnalité).

Dans le cadre de la phase de traitement des données, les périodes d'arrêt et de démarrage sont supprimées. En utilisant des variables de référence (variables indépendantes qui décrivent le procédé), des périodes spécifiques sont détectées visuellement et supprimées de l'ensemble de données par validation numérique, c'est-à-dire toutes les valeurs inférieures à une certaine limite qui ne sont pas représentatives des conditions de fonctionnement normales et souhaitables. Les valeurs limites sont spécifiées par une personne experte. La suppression des périodes d'arrêt et de démarrage du procédé basée sur les variables de référence est répercutée sur les autres variables utilisées dans le cadre de l'analyse.

Additionnellement à la suppression des arrêts, une étape importante du traitement des signaux consiste à détecter et à supprimer les valeurs aberrantes (Liu, Hancong et al., 2004). Le fait que les régimes d'opération "*top-down*" soient définis plus tôt aide à l'identification des valeurs aberrantes, car la

segmentation des données du procédé réduit la variabilité (et l'écart-type), et facilite ainsi la détection des déviations du procédé et évite les rejets inutiles. Les variables représentant le système analysé (variables clés du procédé) sont utilisées pour détecter les valeurs aberrantes. L'analyse des quartiles (Dunn, 2022a) est utilisée pour identifier et éliminer les valeurs aberrantes de manière univariée dans les variables clés du procédé. Cette suppression se répercute par la suite sur toutes les autres variables. L'analyse des quartiles est basée sur la médiane et est donc indépendante de la distribution et plus robuste (et efficace) pour la détection des valeurs aberrantes.

En outre, une analyse de détection des valeurs aberrantes basée sur la corrélation spatiale est effectuée à l'aide de l'analyse en composantes principales pour valider les résultats de l'analyse des quartiles (Barton & Lennox, 2022; Joe Qin et al., 2021; Krishnamurthi et al., 2020). Une fois identifiés et supprimés, les points de données anormaux peuvent être corrigés par imputation. Cette valeur peut être prédite à l'aide des données précédentes, par interpolation ou en utilisant la valeur moyenne.

Ensuite, afin de réduire le bruit qui affecte les signaux en les rendant moins représentatifs de l'état réel du procédé et, par conséquent, de distinguer la tendance du procédé, les signaux sont filtrés. Pour cela, les transformées en ondelettes sont utilisées. Cependant, la qualité du débruitage par transformée en ondelettes dépend de l'ajustement de ses paramètres de contrôle : niveau de coupure du signal (échelle), fonction d'ondelette et paramètres de seuil. L'approche permettant d'identifier les paramètres d'ondelettes optimaux pour un procédé industriel afin de ne retenir que la tendance associée au fonctionnement du procédé est présentée à la Section 4.4. Enfin, pour s'assurer que le filtrage n'induit pas de décalage indésirable, une seconde analyse de décalage est effectuée.

De même que pour la détection des valeurs aberrantes, la prise en compte de la variabilité inhérente des données à partir de la définition "*top-down*" des régimes d'opération avant la suppression du bruit peut être bénéfique pour cette étape, car l'amplitude du bruit de procédé peut fluctuer en fonction des conditions d'opération, et les paramètres du filtre doivent donc être adaptés (Rhinehart, 2013).

Détection du régime permanent

Afin de détecter quand un procédé fonctionne dans des conditions de régime permanent, la première étape consiste à identifier les variables critiques qui seront utilisées comme base pour la détection de l'état d'équilibre. Cette identification est faite en utilisant la connaissance du procédé. Les variables doivent être celles qui caractérisent le mieux la stabilité du système et doivent en outre être

indépendantes, c'est-à-dire contrôlées par des points de consigne ou des variables différentes, et non corrélées, c'est-à-dire ne pas varier ensemble au même moment. Lorsque toutes les variables clés du procédé sont simultanément en état d'équilibre, il est possible de dire que le système a atteint les conditions de régime permanent. Une autre solution serait de déclarer l'état d'équilibre si un certain pourcentage des variables critiques sont en état d'équilibre. La valeur de ce pourcentage varierait en fonction de la complexité du procédé.

L'algorithme de détection du régime permanent employé dans ce cadre est inspiré de celui présenté par Kelly, J. et Hedengren (2013). Cet algorithme utilise le *Student-t test* pour déterminer si la différence entre la valeur du signal du procédé et sa moyenne est supérieure ou inférieure à l'écart-type multiplié par sa valeur critique statistique. Une valeur inférieure signifierait que l'instant est stable tandis qu'une valeur supérieure indiquerait qu'il est instable. Cette approche offre un moyen pratique, efficace et robuste de détecter les conditions en régime permanent. La connaissance du procédé peut et doit être combinée avec cette approche, car elle utilise des paramètres ajustables.

L'algorithme de détection des périodes en régime permanent est exécuté sur toutes les variables de référence, de manière univariée, mais tient toujours compte de la présence des autres variables. Pour gérer plusieurs signaux de procédé qui déterminent collectivement si un système est stable, l'intervalle de confiance est modifié (réduit) selon l'équation de Sidak (Kelly, J. & Hedengren, 2013). La validité de cette équation repose sur l'indépendance de toutes les variables clés du procédé. Ainsi, dans le cas d'une analyse multivariable, moins de conditions d'équilibre sont rejetées.

Après la détection des périodes en régime permanent, un critère seuil pour la période minimale de temps pour que l'opération en régime permanent soit considérée comme représentative est pris en compte. La période minimale de régime permanent est établie sur la base de la connaissance du procédé; ce critère est souvent fixé comme le temps moyen nécessaire pour atteindre l'état d'équilibre dans le procédé. Il s'agit d'une justification fréquemment acceptée pour le seuil, car elle est généralisable, reproductible et robuste.

Dans le cadre de cet algorithme de détection des états stationnaires, le début, l'arrêt et l'étendue sont calculés pour chaque période en régime permanent (Figure 6.10). L'algorithme est capable de garder la trace de la date à laquelle les périodes de régime permanent ont été détectées en utilisant un vecteur d'indice temporel y_t . L'indice de départ est renvoyé lorsque le vecteur y_t devient stable, puis l'indice

d'arrêt est renvoyé lorsque le dernier $y_t = 1$ contigu est obtenu, c'est-à-dire lorsque le vecteur y_t passe de 1 à 0, et l'étendue renvoie le nombre de données ou d'indices de temps entre le début et la fin. Enfin, l'algorithme de détection des périodes de régime permanent renvoie certaines mesures récapitulatives basées sur le nombre d'indices par étendue, telles que le nombre maximum et minimum de données dans une étendue, la moyenne, le mode, la médiane et la variance des étendues. Toutes les étendues jugées acceptables (qui ont atteint une période de temps minimale) sont ensuite réconciliées. Cette idée est cohérente avec le fait que dans les opérations de procédé, il existe des périodes d'état stationnaire sporadiques et persistantes. En cherchant les opérations qui sont stables de façon continue, il faut considérer une période minimale acceptable d'opération en régime permanent.

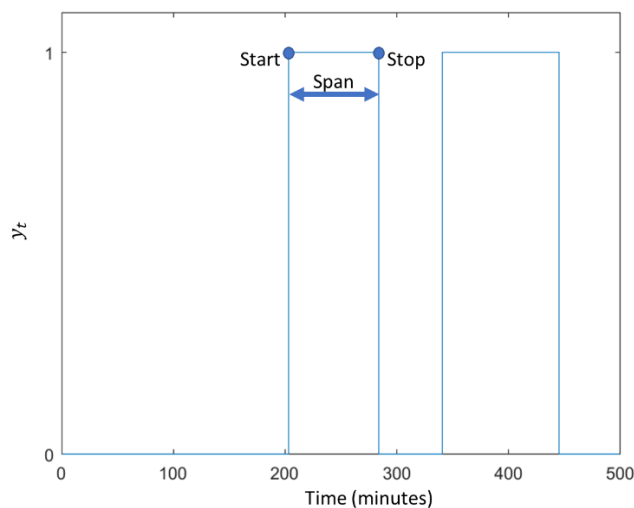


Figure 6.10 – Représentation du démarrage, de l'arrêt et de l'étendue des périodes en régime permanent

Réconciliation des données

La réconciliation des données consiste à identifier les erreurs systématiques (brutes) des mesures, c'est-à-dire les incohérences par rapport aux lois de conservation, et plus particulièrement les erreurs brutes persistantes. Cela se fait en ajustant les données aux bilans théoriques de matière et d'énergie. Par conséquent, la réconciliation des données conduit à l'identification d'instruments de mesure défectueux, défaillants et incohérents. En outre, elle fournit des estimations des variables non mesurées.

Dans le cadre méthodologique proposé ici, la réconciliation des données est effectuée à l'aide de IMPL-DATA (Kelly, Jeffrey Dean, 2023).

Un modèle de réconciliation des données basé sur des modèles physiques théoriques (modèles thermodynamiques, de bilan matière et d'énergie) est construit pour identifier les incohérences. Ce modèle représente le diagramme d'écoulement du département, c'est un sous-ensemble des équations utilisées dans la simulation de procédé. Cette étape nécessite une profonde compréhension du procédé. Le modèle de réconciliation des données est lié aux lois fondamentales de la conservation de la matière, de l'énergie et de la quantité de mouvement, et il dépend également des opérations (modes d'opération, grades, réglages des équipements, démarrages, arrêts, boucles de recyclage/recirculation, etc.). Du point de vue de la réconciliation des données, ces variations dans les opérations peuvent être exprimées mathématiquement en utilisant la notion de courants temporaires (versus permanents) qui peuvent être ouverts/fermés ou actifs/inactifs en utilisant des interrupteurs de courants. Les courants temporaires peuvent être activés ou désactivés en fonction de l'état du procédé. Habituellement, les personnes à l'usine savent *a priori* quand les changements de mode, de grade et autres se produiront via leur département de planification, d'ordonnancement et de coordination. Ainsi, des indicateurs manuels et automatiques tels que les positions des actionneurs de vannes, les conditions clés du procédé, les démarrages/arrêts des pompes/compresseurs, etc. seront disponibles pour aider à déterminer quel changement s'est produit et à quel moment. Le modèle de réconciliation des données représente tous les courants possibles, qu'ils soient permanents ou temporaires. La plupart des courants sont permanents dans les procédés industriels, ils sont représentés par des lignes pleines à la Figure 6.11. Néanmoins, lorsque les conditions du procédé changent (passage à un autre mode d'opération), il y a des changements de courants ou des courants temporaires (représentés par des lignes pointillées à la Figure 6.11). Ainsi, lorsqu'un procédé passe à un autre mode d'opération, certains courants ne sont plus nécessaires. Par conséquent, la manipulation de ces configurations de courants peut indiquer de manière fiable un certain mode, grade ou changement qui s'est produit, réduisant ainsi la possibilité, la plausibilité et la probabilité d'occurrence des erreurs brutes du modèle.

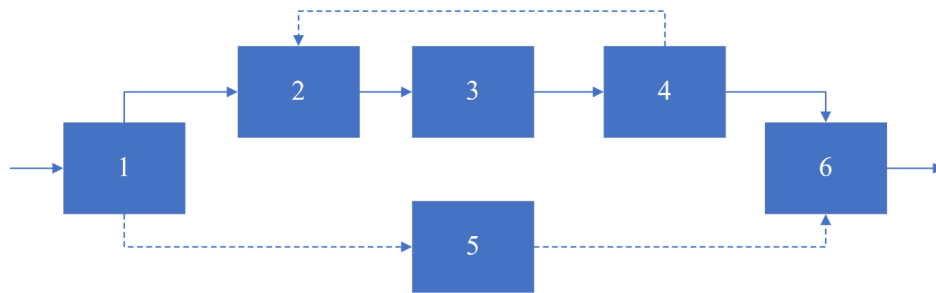


Figure 6.11 – Exemple de schéma bloc pour présenter les courants permanents et temporaires

L'étape suivante consiste à valider le modèle, c'est-à-dire à s'assurer qu'il n'y a pas d'erreurs de modèle. De plus, il doit être validé pour chaque régime d'opération "*top-down*". En effet, afin de valider correctement le modèle, les régimes d'opération doivent être reconnus, car pour chaque mode de fonctionnement, les opérations changent, il peut y avoir différents courants (courants temporaires). Par conséquent, si le mode d'opération change, le modèle pourrait être erroné s'il n'est pas modifié correctement. Ainsi, les courants temporaires dus aux changements de mode d'opération sont correctement pris en compte dans la structure du modèle.

Enfin, avant de valider le modèle, il est attendu de trouver à la fois des erreurs brutes au sein du modèle et des mesures. Cependant, une fois la validation effectuée, il ne reste plus que les erreurs brutes des mesures; il ne devrait plus y avoir d'erreurs dans le modèle. De plus, lorsque le procédé est instable, il y aura des erreurs brutes au sein du modèle, car les équations de bilans massiques ne seront pas égales à zéro. Par conséquent, afin de n'identifier que les erreurs brutes des mesures, les périodes instables sont exclues de l'analyse, c'est-à-dire que c'est l'une des raisons pour lesquelles la détection des périodes de régime permanent est nécessaire.

Une fois le modèle de procédé validé, il est utilisé pour analyser des données de procédé en régime permanent, typiquement moyennées ou agrégées. Toutes les périodes qui s'avèrent être statistiquement stables sont réconciliées. La réconciliation des données en régime permanent, qui est basée sur des analyses statistiques, est calculée sur la base de la moyenne de la période.

La réconciliation des données prend en compte l'impact de l'instrumentation de mesure du procédé en utilisant une métrique de tolérance (poids), c'est-à-dire que les variances brutes du capteur sont fournies. Cette métrique implique la précision, l'exactitude et la fiabilité du capteur. Elle prend en compte le type et la marque de l'instrumentation et la qualité des données de mesure. La tolérance de l'erreur de mesure est égale à deux fois l'écart-type. La variance est donc égale au carré de la moitié de la tolérance.

Dans le cadre de l'algorithme de réconciliation des données, une distinction est faite entre une erreur grossière sporadique (transitoire) et une erreur grossière soutenue (persistante) (Tong & Crowe, 1996) en associant le contexte temporel au problème de réconciliation. Au début, lorsqu'une erreur grossière est détectée, il est impossible d'identifier si elle est sporadique ou soutenue. Afin de faire la distinction entre les deux, le nombre de fois (la probabilité) où un capteur est en contradiction avec les bilans de matière et d'énergie sur un horizon de surveillance est calculé; l'occurrence est enregistrée et consignée.

Cette mesure identifie les erreurs grossières persistantes en classant la probabilité d'occurrence des erreurs. IMPL-DATA détecte quels sont les capteurs qui sont constamment en erreur par rapport à ceux qui le sont sporadiquement. Généralement, le personnel de l'usine n'agira que sur les erreurs persistantes. Ainsi, la réconciliation des données est utilisée pour dépister les erreurs grossières persistantes dans les données et pour que le personnel de l'usine les corrige au fil du temps en recalibrant le capteur, par exemple.

Il est intéressant de constater que la détection du régime permanent est un concept relatif étant donné qu'une période est jugée en état d'équilibre par rapport à ses voisins immédiats et qu'il en est de même pour la réconciliation des données. En effet, il est impossible de savoir si un capteur est fautif dans l'absolu, il est uniquement possible de déterminer s'il est cohérent avec les autres capteurs (qu'il obéit aux contraintes du procédé).

Détection et identification des régimes d'opération

La dernière étape du cadre méthodologique proposé consiste à détecter les régimes d'opération. En partant d'une perspective "*top-down*", les données traitées en régime permanent sont utilisées pour construire un modèle d'analyse en composantes principales. L'ACP exploite les données historiques du procédé et découvre des événements cachés qui peuvent être utiles pour détecter des conditions d'opération inattendues, nouvelles ou inconnues. Ensuite, les scores des composantes principales sont utilisés comme variables d'entrée dans l'algorithme de clustering *k*-means. Seules les composantes qui expliquent une proportion importante (90 % par exemple) de la variation entre les variables d'opération sont sélectionnées. L'approche essai-erreurs est utilisée pour identifier le nombre de clusters détectés par l'algorithme *k*-means.

En outre, l'analyse de contribution et le *loading chart* (Dunn, 2022b) sont utilisés pour identifier les variables les plus influentes (qui caractérisent chaque cluster) et expliquent la variabilité entre les régimes d'opération en régime permanent; par conséquent, les variables qui sont importantes pour un régime particulier, c'est-à-dire les variables qui sont affectées par les différents clusters. Enfin, les expertes et experts confirment les régimes d'opération et expliquent les facteurs fondamentaux sous-jacents, c'est-à-dire que les variables d'intérêt qui aident à identifier les clusters.

Ces régimes d'exploitation "*bottom-up*" basés sur une approche axée sur les données ainsi que sur la connaissance du domaine offrent la possibilité d'interpréter et de comprendre le procédé de manière plus

approfondie. Ces régimes pourraient être liés à la façon dont l'usine est gérée par les opératrices et opérateurs, puisque ces personnes peuvent l'exploiter différemment lors de la production d'un même produit.

Les clusters détectés sont évalués par des expertes et experts du domaine pour mettre en évidence ce qui différencie un cluster d'un autre et déterminer s'ils représentent tous des régimes d'opération distincts. Les régimes d'opération identifiés doivent être mutuellement exclusifs. Par conséquent, les valeurs spécifiques de certaines variables doivent être attribuées uniquement à un régime d'opération. En d'autres termes, il existe un ensemble unique de valeurs de paramètres pour distinguer chaque régime.

La détection et l'identification des régimes d'opération permettent d'améliorer la connaissance des procédés, car l'analyse de chaque cluster peut conduire à la découverte d'un mode de fonctionnement inconnu. D'autre part, la connaissance du domaine est nécessaire pour interpréter les clusters détectés par l'analyse en composantes principales. Par conséquent, les résultats de l'ACP doivent être interprétés sur la base d'une compréhension des principes fondamentaux du procédé. Il s'agit d'une tâche pour les expertes et experts de procédé; les régimes de fonctionnement "*bottom-up*" sont détectés et identifiés en fusionnant l'approche axée sur les données et la connaissance des procédés.

6.3.5 Application du cadre méthodologique a un problème d'optimisation

Afin de concrétiser l'approche systématique proposée à la sous-section 6.3.4, et sa valeur, les coûts d'opération de chaque régime sont considérés. Ceux-ci ne sont pas évalués de manière absolue, mais plutôt par rapport à la production moyenne de pâte à dissoudre, qui constitue la référence. Plus précisément, sachant que l'objectif de l'usine est de produire le plus de pâte possible, l'analyse économique se concentre sur le profit généré par les tonnes supplémentaires (supérieures à la moyenne) de pâte produites. En outre, l'usine a démontré que les revenus générés par la vente de la pâte produite couvrent toujours les coûts de la consommation supplémentaire d'antimousse et de produits chimiques de blanchiment. Chaque tonne de pâte produite au-dessus de la moyenne apporte un bénéfice. Par conséquent, l'objectif est d'identifier la meilleure façon d'opérer pour gagner le plus d'argent; trouver les régimes d'opération les plus rentables.

Chaque mode d'opération différent se voit attribuer un bénéfice obtenu par rapport à la base de référence, qui est estimée à 475 tonnes/jour. Les ressources prises en compte dans cette analyse des coûts induits par les opérations sont la quantité moyenne d'antimousse et de produits chimiques de blanchiment

consommés au cours des différents régimes d'opération; le département de lavage n'étant pas une activité à forte intensité énergétique, les coûts énergétiques n'ont pas été pris en compte. Ainsi, les activités qui consomment ces ressources sont l'utilisation des laveurs et les réactions qui se produisent dans les tours de blanchiment. Le *cost object* est le département de lavage.

En considérant les coûts de production absolus par tonne de pâte produite et leur ventilation en coûts directs et indirects, il est possible d'identifier que pour chaque tonne supplémentaire de pâte produite par rapport à la moyenne, l'usine réalise un gain de 500 \$. Par conséquent, en considérant la quantité de pâte produite par rapport à la moyenne et la quantité de matière première consommée, en soustrayant les coûts du gain, il est possible d'obtenir le profit engendré par chaque régime d'opération.

Cette analyse du coût marginal fournit les économies par régime d'opération pour une année de fonctionnement (355 jours). Cependant, la faisabilité doit être prise en compte; il s'agit d'identifier combien de jours par an il est possible de fonctionner de cette manière et de réduire les coûts d'opération. Ces coûts peuvent être réduits dans des conditions spécifiques lorsque le lavage est en fait le goulot d'étranglement de l'usine.

L'application ou l'analyse ciblée dicte quels régimes d'opération sont nécessaires et comment ils doivent être définis (le niveau d'agrégation ou de granularité des données). Une piste possible serait de se concentrer sur l'identification des régimes d'opération qui sont moins rentables et d'évaluer s'ils sont essentiels et évitables. Cependant, les régimes d'opération identifiés pour cette étude de cas sont ceux de l'été (où l'usine fonctionne à plein régime), et surtout lorsque le procédé fonctionne bien (production de plus de 475 tonnes de pâte par jour). Par conséquent, les jours où moins de 15 lots ont été produits n'ont pas été considérés; moins de 15 lots indique un problème dans les opérations, ces données ne sont pas représentatives de bonnes opérations.

Pendant l'été, l'usine fonctionne à un taux de production plus élevé et le département de lavage est le facteur limitant. Ainsi, la capacité maximale des équipements de ce département est souvent atteinte, ce qui se traduit par une baisse des performances de lavage. En hiver, les taux de production sont plus faibles, ce qui apporte plus de stabilité à l'opération – le débit de lavage est plus facile à contrôler; le département de lavage n'est pas le goulot d'étranglement. Ainsi, le fait d'avoir moins de pâte à traiter permet au département de lavage de mieux fonctionner en hiver. Cependant, une production plus faible n'est jamais souhaitable d'un point de vue économique. Ainsi, en général dans l'usine, les saisons jouent

un rôle critique dans les opérations du procédé. Cela a un impact sur le nombre de lessiveurs utilisés (lots produits) chaque jour (en hiver, moins de lessiveurs sont utilisés) – le nombre de lessiveurs est directement lié au taux de production de pâte, et la température de l'eau du procédé est plus froide en hiver qu'en été.

Les régimes d'opération en régime permanent identifiés à l'aide du cadre proposé à la sous-section 6.3.4 devraient durer au moins 12 heures. Il s'agit de la durée minimale pour que l'état d'équilibre soit représentatif et compte tenu du fait que les rotations de quarts de travail ont lieu deux fois par jour. Au cours de chaque quart de travail, chaque équipe est chargée d'atteindre les objectifs du département. Étant donné que certains régimes d'opération ont duré plus de 12 heures, ces régimes montrent un chevauchement entre les équipes. Les régimes d'opération sont comparés en fonction de l'équipe qui travaille pendant la durée du régime d'opération, car les différentes équipes ne réagiront pas nécessairement de la même façon face à un événement survenant dans l'usine. Les régimes d'opération sont également distingués par des paramètres clés, notamment la production moyenne de pâte et la qualité (conductivité) de la pâte, ainsi que le profit par rapport à la production moyenne.

Le Tableau 6.3 présente les valeurs des paramètres clés moyennes sur la période (durée de chaque régime d'opération). La dernière colonne indique le profit en millions de dollars par an pour chaque régime d'opération. Étant donné les différents régimes d'opération identifiés sur la base du cadre méthodologique présenté à la sous-section 6.3.4, les plus favorables peuvent être reconnus en fonction des valeurs des paramètres qui définissent de manière unique chacun des régimes d'opération.

Tableau 6.3 – Paramètres clés pour tous les régimes d'opération détectés

Régime d'opération	Équipe responsable	Débit de pâte (tonne/jour)	Conductivité de la pâte (µmhos)	Profit (M\$/an)
1	C/D	654	3816	32
2	C/D	642	4045	30
3	C	661	4099	33
4	C/D	673	4006	35
5	C/D	681	4124	37
6	A/C	683	4363	37
7	D	677	4616	36
8	C/D	660	4626	33
9	D	668	4489	34
10	C/D	701	4436	40
11	C/D	721	4376	43
12	A/B	754	4389	49

À partir du Tableau 6.3, il est possible d'observer que la combinaison la plus intéressante est le régime d'opération 12. Il permet d'économiser 49 millions de dollars par an, et il est obtenu lorsque le procédé est opéré par les équipes A et B, que 754 tonnes/jour sont produites et que la conductivité de la pâte est de 4389 μmhos . Afin d'évaluer les raisons de la performance du régime d'opération 12, celui-ci est comparé au régime d'opération 2, le moins performant. Même s'il s'agit dans les deux cas de périodes de fonctionnement en régime permanent, lorsque le procédé produit plus de 15 lots de pâte par jour, c'est-à-dire en cas de fonctionnement optimal de l'usine, un profit de 19 millions de dollars différencie ces deux régimes d'opération. Par conséquent, une analyse de contribution de ces régimes d'opération a mis l'accent sur les variables qui varient (et impactent) le plus les régimes d'opération dans le département de lavage. L'identification de ces variables critiques aborde et explique les différences entre les régimes d'opération, et met en évidence la façon dont le procédé est opéré lorsque les performances idéales se produisent (Figure 6.12). La logique ou le rationnel derrière les variables contrôlées ayant un impact est expliquée grâce à la collaboration d'une personne experte de procédé.

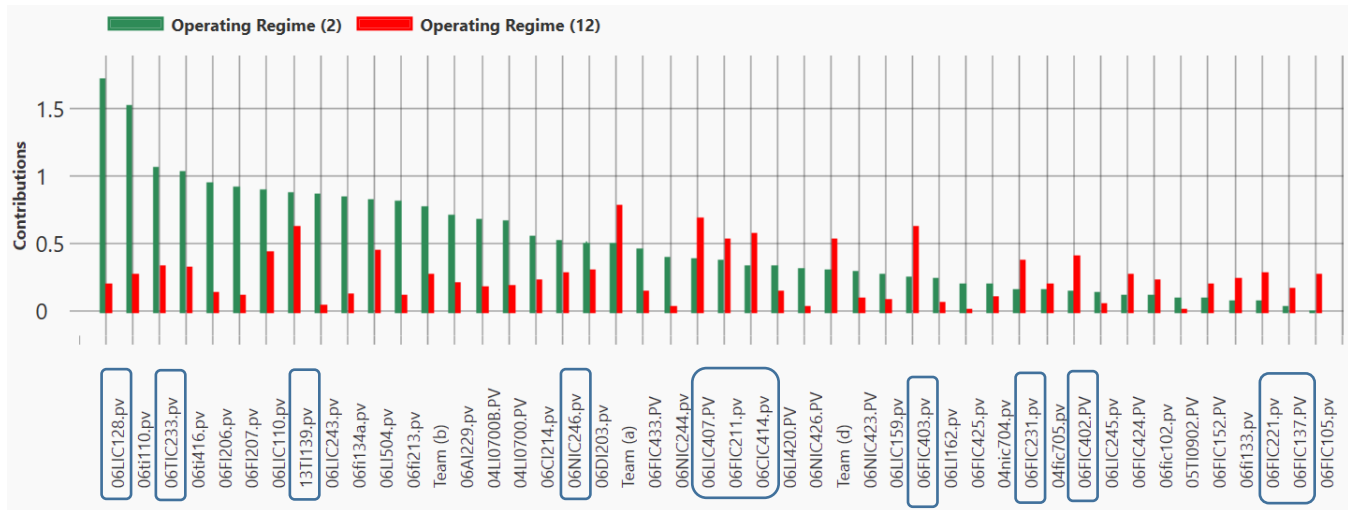


Figure 6.12 – Contribution des variables pour les régimes d'opération 2 et 12

La Figure 6.12 montre que la variabilité (ce qui différencie les régimes d'opération 2 et 12) est principalement expliquée par 06LIC128, qui est le niveau du laveur 1A. C'est une variable qui est contrôlée par les opératrices et opérateurs; ces personnes décident du niveau dans le laveur. Ce niveau peut influencer la façon dont l'usine performe, ce qui peut influencer le débit de pâte, et donc générer un profit. Le niveau est presque 50% plus élevé pendant le régime d'opération 2 que pendant le 12. Il est observé que les régimes d'opération 1 à 5 fonctionnent à un niveau plus élevé, tandis que les régimes 6

à 12 présentent un niveau du laveur 1A plus faible et produisent, en général, plus de pâte, ce qui réduit davantage les coûts.

En outre, 06NIC246, qui est la consistance du courant quittant la cuve de mélange, explique également la différence entre les deux régimes. Il est observé que l'opération autour de 2,150 conduit à de meilleures performances (régime d'opération 12) par rapport à 1,916 pour le régime d'opération 2.

Une troisième variable intéressante est 06LIC407, le niveau du réservoir de filtrat de l'épaississeur. Les opératrices et opérateurs modifient souvent sa valeur pour atteindre un débordement en pensant que cela va augmenter le débit de pâte. Cependant, le régime d'opération 12 produit le plus grand débit de pâte avec un niveau du réservoir de filtrat de l'épaississeur inférieur de 17% à celui du régime d'opération 2.

Une autre variable contrôlée intéressante est 06CIC414, la conductivité de la pâte. Contrairement à ce que les opératrices et opérateurs pensent généralement, un débit de pâte élevé ne rime pas avec une conductivité supérieure à 5000 μmhos , c'est-à-dire qu'il n'est pas nécessaire de pousser la conductivité à des valeurs supérieures à 5000 pour atteindre une quantité de pâte élevée. Dans les 12 régimes d'opération représentant de bonnes périodes de fonctionnement, la conductivité est systématiquement inférieure à 5000. Cela prouve qu'il est possible d'atteindre un débit de pâte élevé avec une faible conductivité.

Enfin, le débit d'eau des douches (06FIC402, 06FIC403, 06FIC137, 06FIC211, 06FIC221, 06FIC231) est contrôlé par les opératrices et opérateurs; ce sont des leviers importants. De meilleures performances sont obtenues lorsque plus d'eau est ajoutée par les douches. Il est observé que les six variables présentent des valeurs plus élevées dans le régime d'opération 12 par rapport au régime d'opération 2. En outre, la température de l'eau des douches (06TIC233) est contrôlée par les opératrices et opérateurs en ajoutant de l'eau chaude ou tiède au débit d'eau des douches. Cette variable est tout aussi importante que le débit d'eau des douches. De la même manière, la température de l'eau chaude (13TI139) a un impact important sur les performances de lavage. Une température plus élevée entraîne une meilleure performance de lavage.

Une autre variable qui n'est pas contrôlée, mais qui affecte grandement les performances de lavage est 06TI416 (température du réservoir de pâte écrue). Une température plus élevée correspond à des opérations de lavage plus faciles et à un débit de pâte plus élevé. Cet indicateur de température est influencé par la température de l'eau de la douche et la température de la pâte à l'entrée de l'épaississeur.

Par conséquent, en se concentrant particulièrement sur les variables contrôlées afin de comprendre ce que les opératrices et opérateurs ont fait différemment au régime d'opération 12, il est observé que les variables qui ont le plus de potentiel d'influencer positivement le procédé sont le niveau du laveur 1A, la consistance de la pâte, le niveau du réservoir de filtrat de l'épaississeur et le débit et la température de l'eau de la douche.

CHAPITRE 7 DISCUSSION GÉNÉRALE

7.1 Prises de décisions en industrie : régime permanent et régime d'opération

La caractérisation du traitement des données présentée au Tableau 6.1 révèle qu'au niveau de l'usine, le nettoyage des données consiste principalement en des tâches manuelles basées sur la visualisation des données. Bien que cela soit très différent des techniques présentées dans la littérature, en industrie, le nettoyage de données est effectué préalablement aux applications de prises de décision et varie d'une application à l'autre.

En ce qui concerne l'optimisation, la littérature (Abu-el-zeet et al., 2001; Delou et al., 2021; Reyes et al., 2015) suggère comme étapes de traitement des données de réduire le bruit des mesures, d'effectuer la détection des périodes en régime permanent et de réconcilier les données. L'industrie applique la vérification des valeurs limites et les filtres conditionnels et effectue le nettoyage des données pour éliminer les valeurs nulles et les erreurs. Le bruit de mesure est également pris en compte. L'industrie considère le fonctionnement normal pour les analyses d'optimisation. La détection des périodes en régime permanent est plutôt simpliste; le fonctionnement normal signifie nécessairement l'état d'équilibre pour l'industrie. Enfin, la réconciliation des données n'est pas mentionnée par les spécialistes en usines.

En ce qui concerne l'analyse des coûts, la littérature (Korbel, Milan, 2011; Korbel, M. et al., 2007; Korbel, M. & Stuart, 2011) propose de définir et d'identifier les régimes d'opération comme première étape. Il est ensuite suggéré d'effectuer le nettoyage des données de procédé, c'est-à-dire la correction des anomalies et le filtrage des données, puis la détection de l'état d'équilibre et enfin la réconciliation des données de procédé pour la détection des erreurs grossières. Les spécialistes en usines éliminent les valeurs aberrantes et nettoient les données; ces personnes établissent la moyenne des données et recherchent une opération normale qu'elles jugent représentative. Par conséquent, l'industrie a omis la première étape proposée par la littérature, à savoir la détection des régimes d'opération, ainsi que la réconciliation des données, mais son étape de traitement des données est conforme à la littérature. Enfin, la détection du régime permanent est effectuée de manière plus conforme à la littérature que dans le cas de l'optimisation, car les spécialistes en usine font la moyenne de bonnes données, nettoyées et représentatives.

Pour la détection et le diagnostic de fautes, la littérature (Kaneko & Funatsu, 2017; Rogers et al., 2021; Su et al., 2006) conseille le lissage et la détection de l'état d'équilibre comme première étape du traitement des données et la détection des régimes d'opération comme seconde étape. La littérature sur la surveillance des usines (Aldrich & van Deventer, 1993; Upadhyaya et al., 2014) met l'accent sur l'utilisation de données d'opération en régime permanent, exempt de bruit et erreur grossière. L'industrie utilise des données moyennées. Par conséquent, le bruit est réduit, les valeurs aberrantes sont assimilées à la valeur moyenne, et il y a la notion de la détection du régime permanent derrière ces moyennes à long terme. Il n'est pas question d'erreur grossière ou de détection des régimes d'opération.

Enfin, l'étape de traitement des données pour les simulations de procédé nécessite la réduction du bruit et l'élimination des valeurs aberrantes. Cette étape est réalisée correctement par les spécialistes en usine, qui ont éliminé les valeurs aberrantes et appliqué différents types de filtres pour supprimer le bruit et les données indésirables.

En outre, le Tableau 6.1 indique que le traitement des données a été effectué en fonction de l'application, c'est-à-dire dans un but précis, en utilisant une manière différente de traiter les données selon la nature de la prise de décision. Pour une application donnée, les usines traitent leurs données de manière assez similaire d'une personne à l'autre, il semble y avoir un consensus. Par conséquent, il est constaté que la décision ou l'analyse en cours a un impact sur la manière dont les données sont traitées.

Le Tableau 6.1 indique que les applications de prise de décision sont mises en œuvre à l'aide du concept de régime permanent, car elles sont toutes basées sur des données moyennées. Malgré que le calcul de la moyenne des données ne soit techniquement pas équivalent à la détection des périodes en régime permanent, lorsque le personnel de l'usine fait la moyenne des données de procédé, il s'assure que les données sélectionnées sont représentatives d'un bon fonctionnement en régime permanent. Cependant, ces personnes ne cherchent pas systématiquement à savoir si les données sont stables, aucune approche mathématique n'est employée. De ce fait, les expertes et experts effectuent une analyse en régime permanent sans utiliser d'algorithme, ces personnes connaissent leur procédé, elles sélectionnent un intervalle de temps spécifique qu'elles savent être en régime permanent.

La plupart des spécialistes en usines partent généralement du principe que le procédé fonctionne en permanence en régime permanent et toutes les décisions que ces personnes prennent s'inscrivent donc dans un paradigme de régime permanent. En effet, le personnel de l'usine a mentionné qu'il est plus facile

de comprendre le procédé dans un paradigme de régime permanent. Les données en régime permanent sont essentielles à la prise de décision, elles sont faciles à interpréter, pratiques et apportent une valeur substantielle.

En raison du caractère hautement dynamique des procédés de production de pâte et papier et des nombreuses interactions entre les départements (sections de l'usine), l'amélioration des performances opérationnelles en ligne constitue un véritable défi. Par conséquent, les analyses hors ligne sont préconisées. Les expertes et experts en usine considèrent que les analyses hors ligne sont d'emblée statiques et en régime permanent. Ces personnes les exécutent en utilisant des données historiques résumées (agrégées). Essentiellement, pour le personnel de l'usine, l'analyse des données hors ligne est équivalente à l'état d'équilibre. Cela est soit dû au fait que les spécialistes considèrent *de facto* que le procédé est constamment en régime permanent, soit au fait que dans une analyse hors ligne, une période de référence est choisie et une agrégation dans le temps est effectuée. En réalité, une analyse hors ligne peut être dynamique ou statique et peut considérer les périodes transitoires ou stables. Également, le procédé peut être en régime permanent, que les données soient en ligne ou hors ligne.

Les périodes de régime permanent peuvent être déclarées lorsque les opérations sont stables, ce qui signifie qu'il y a peu ou pas de variation dans les données. Il existe de nombreuses périodes de régime permanent distinctes dans un procédé d'opération. La plupart de ces périodes ne sont pas connues *a priori*, c'est pourquoi un algorithme de détection des états stables est nécessaire.

Interrogés sur l'analyse en régime permanent, les spécialistes mentionnent que le régime permanent implique que le procédé est stable depuis un certain temps; il n'a pas connu de démarrage ou d'arrêt au cours des dernières heures. Le régime permanent suppose que le procédé a atteint un "fonctionnement normal"; lorsque le procédé démarre, il faut attendre que les conditions d'opération normales soient atteintes. Il implique également que le procédé n'a pas connu de changement significatif du taux de production ou de grade au cours des dernières heures.

Dans bon nombre de leurs opérations, le personnel de l'usine part du principe que l'usine est en régime permanent presque tout le temps, ou du moins ce qu'il considère comme tel. En d'autres termes, s'il ne modifie pas le grade ou les cibles à atteindre, le procédé fonctionne à peu près en régime permanent. Le personnel mentionne que les équipements fonctionnent pendant des semaines sans perturbation et que le procédé est donc en régime permanent la plupart du temps. Quelques expertes et experts ont mentionné

que le régime permanent signifie que l'état actuel du procédé varie moins qu'un certain seuil (en pourcentage). D'autres mentionnent qu'à part l'appréciation visuelle des données (par exemple, un graphique de tendance), aucun critère spécifique n'est utilisé pour évaluer la stabilité des données, outre, pour certaines variables, un intervalle de deux fois l'écart-type autour des variables.

En général, ce sont les expertes et experts du procédé qui définissent les périodes de régime permanent. Pour leur analyse, ces personnes segmentent les données en fonction des opérations normales. Une variable dans le système PI d'AVEVA classe les données en tant que « fonctionnement normal » ou « anormal » (ce qui signifie qu'il y a un démarrage, un arrêt ou une perturbation) en utilisant des règles qu'elles ont écrites. Un exemple de règle est que lorsqu'elles démarrent une pompe, le procédé doit opérer pendant une heure à un débit supérieur à X gallons par minute pour être considéré en fonctionnement normal. Une fois ce critère rempli, le drapeau « fonctionnement normal » est activé. Tant que le procédé reste dans ce régime, les expertes et experts le considèrent comme un fonctionnement normal. Ces règles concernent les principales lignes de production et s'appliquent donc à l'ensemble des opérations.

Somme toute, les expertes et experts en usines supposent généralement que le procédé opère dans des conditions stables. Ces conditions sont considérées dans la plupart des décisions prises en usines, c'est-à-dire dans les simulations de procédé, les analyses de coûts et d'énergie, les modèles d'optimisation, les analyses comparatives, la surveillance des conditions et l'analyse des défaillances. Cependant, il existe de nombreuses périodes en régime permanent au sein des procédés qui peuvent être classées en régimes d'opération caractérisés par différents taux de production, produits, grades, saisons, recettes, matières premières, opératrices et opérateurs, etc.

Comme le montre le Tableau 6.1, le personnel des usines n'a pas explicitement tenu compte des régimes d'opération pour la prise de décision, sauf pour ce qui est de l'impact saisonnier sur l'analyse énergétique et les différents grades dans les simulations. En général, il y a peu ou pas de reconnaissance du fait qu'il existe de nombreuses périodes d'opération en régime permanent différentes au sein des opérations.

Cependant, les expertes et experts ont reconnu la valeur et le potentiel de la détection et l'identification des régimes d'opération. Cela peut conduire à de nombreuses décisions de procédé, et une analyse plus approfondie peut être entreprise en prenant en compte les régimes d'opération.

La détection des régimes d'opération offre la possibilité d'évaluer l'impact de chacun d'entre eux sur la consommation d'énergie, l'efficacité, la performance, le coût, le rendement et la qualité du produit et de

déterminer le régime préféré. Il est important que les simulations de procédés tiennent compte de ces différents régimes. De même, les régimes d'opération doivent être pris en compte dans les projets de conception ou les dépenses d'investissement. Pour l'analyse énergétique, il est essentiel de reconnaître les impacts saisonniers, qui peuvent être importants. Les régimes d'opération peuvent également être utilisés pour fixer des limites supérieures et inférieures dans les systèmes de contrôle. La planification de la production, le dépannage et l'optimisation sont également liés aux régimes d'opération.

La détection et l'identification des régimes d'opération réduit la variabilité, ce qui rend les modèles de procédé plus fiables et facilite la détection des déviations et évite les rejets inutiles. Lorsque les différents régimes d'opération ne sont pas identifiés, les écarts-types sont beaucoup plus importants et, par conséquent, lors de la détection des fautes, la défaillance des équipements, la surveillance ou le dépannage, il est beaucoup plus facile de passer à côté de quelque chose. Ainsi, en général, la détection des régimes d'opération aide à comprendre davantage le procédé.

En outre, le personnel de l'usine est vraiment intéressé à connaître le coût d'opération des différents régimes ainsi que leur empreinte carbone respective. Si les implications de chaque régime étaient connues, cela pourrait guider le personnel vers des décisions différentes. Ces expertes et experts veulent identifier la meilleure façon d'opérer; cela pourrait être basé sur une analyse des coûts. Les coûts de production sont inévitablement liés au régime d'opération.

Ainsi, le personnel de l'usine doit reconnaître que l'usine a différents régimes d'opération et utiliser leur définition pour toutes les applications mentionnées afin de prendre de bonnes (ou de meilleures) décisions. La détection des régimes d'opération aide à mieux comprendre le fonctionnement de l'usine et augmente la capacité d'interpréter les données de procédé de manière approfondie et profonde. Elle fournit des informations supplémentaires sur l'opération du procédé et offre un moyen applicable de définir le meilleur scénario pour obtenir un procédé plus efficace et plus rentable. Par conséquent, la reconnaissance des régimes d'opération ouvre la porte à de nombreuses opportunités.

En conclusion, les entrevues des expertes et experts ont confirmé la sous-hypothèse #1 selon laquelle la prise de décision dans l'industrie des pâtes et papier est généralement basée sur un fonctionnement en régime permanent pour la plupart des analyses de procédé, mais que toutefois, ces périodes stationnaires ne sont pas attribuées à des régimes d'opération spécifiques.

Également, la description des étapes de traitement des données pour les différentes applications et analyses a permis de souligner le fait que la littérature propose des méthodes de traitement très avancées, mais qui ne sont pas nécessairement alignées avec les préoccupations des industriels. En effet, les barrières et défis présentés à la section 1.1 expliquent pourquoi les données ne sont pas utilisées au maximum de leur potentiel. De ce fait, la section 7.2 discute d'une manière efficace et robuste de traiter les erreurs aléatoires.

7.2 Heuristiques basées sur les transformées en ondelettes pour un traitement plus efficace des données

La performance d'une méthode de traitement des signaux dépend fortement du choix de ses paramètres. Les paramètres des transformées en ondelettes sont théoriquement sélectionnés en fonction des caractéristiques du signal. La fonction ondelette doit correspondre à la forme du signal, les paramètres d'échelle et de seuillage doivent être choisis de manière à ne retenir que la tendance ou les composantes importantes du signal et à réduire les éléments de haute fréquence. Une décomposition excessive du signal peut entraîner un lissage excessif et comporter le risque d'éliminer des composantes importantes. D'autre part, un débruitage insuffisant (décomposition inadéquate) aurait pour conséquence de conserver une partie considérable du bruit dans le signal reconstruit. Par conséquent, le niveau de décomposition (échelle) doit être optimisé pour obtenir la meilleure qualité de signal filtré. Le niveau de décomposition est spécifique à chaque cas, il dépend notamment de la nature du signal et de l'objectif de l'étude. Finalement, le *vanishing moment* associé à la fonction ondelette aide à générer des caractéristiques plus lisses. Une valeur plus élevée indique que plus de *moments* sont capturés dans le signal.

Cependant, en pratique, il est fastidieux, exigeant et chronophage de sélectionner pour chaque signal individuel ses paramètres de transformée en ondelette. Par conséquent, pour réduire le temps consacré au traitement du signal, des heuristiques permettant de choisir les paramètres sont proposées.

De ce fait, 400 signaux sont analysés un par un avec différents paramètres d'ondelettes pour identifier les paramètres optimaux à l'aide d'un balayage paramétrique. Les paramètres optimaux de chaque combinaison de paramètres de classification des signaux sont saisis dans des heuristiques (Figure 6.8). Par conséquent, étant donné les caractéristiques de classification d'un signal, les signaux doivent être traités d'une manière spécifique. Enfin, la considération des caractéristiques du signal en plus des

paramètres de classification a permis d'obtenir des informations importantes et de mieux comprendre le comportement des signaux. Il est observé que les signaux ont des caractéristiques similaires selon leur classification. Cette analyse exhaustive a permis de mieux comprendre les signaux et il est maintenant possible de traiter tous les signaux de l'usine.

Les heuristiques proposées pour le traitement de différents types de signaux basés sur leurs caractéristiques sont validées en comparant la valeur de la fonction objectif (ligne 12 du pseudo-code, section 4.4) pour les paramètres de transformée en ondelettes obtenus avec le balayage paramétrique et ceux suggérés par les heuristiques. La fonction objectif minimise l'amplitude des pics des composantes dont la fréquence est supérieure à la fréquence d'analyse ainsi que l'erreur quadratique moyenne entre l'amplitude des pics des composantes dont la fréquence est inférieure à la fréquence d'intérêt pour le signal brut et le signal filtré. Un total de 7 signaux est considéré dans l'analyse de validation; un de chaque branche de la Figure 6.8 (un exemple est représenté en jaune à la Figure 6.8).

Tableau 7.1 – Validation des heuristiques sur sept signaux

Paramètres de classification du signal	Paramètres du balayage paramétrique	Fonction objectif du balayage paramétrique	Fonction objectif des heuristiques	Différence en %
Contrôleur – Consistance – Pâte	db11, niveau 7	0.77	0.86	10.11
Contrôleur – Débit – Liqueur noire faible	db13, niveau 6	719.15	755.31	4.79
Contrôleur – Niveau – Pâte	db14, niveau 7	38.90	39.22	0.81
Contrôleur – Température – Liqueur noire forte	db15, niveau 6	73.14	73.82	0.92
Indicateur – Débit – Liqueur blanche	db14, niveau 6	141.71	167.58	15.44
Indicateur – Niveau – Liqueur verte	db13, niveau 10	0.71	0.78	9.32
Indicateur – Température – Vapeur	db14, niveau 6	78.07	85.51	8.70

À partir du Tableau 7.1, il est possible d'observer que la plus grande différence entre la fonction objectif du balayage paramétrique et la fonction objectif de l'heuristique se produit dans le cas d'un indicateur de débit de liqueur blanche. La Figure 7.1 montre le résultat de l'utilisation de l'heuristique présentée à la Figure 6.8 pour le signal de ce capteur. Elle montre le signal original brut de ce capteur (en bleu), le signal filtré résultant de l'utilisation des paramètres d'ondelette donnés par le balayage paramétrique (en rouge), et le signal traité obtenu en utilisant les paramètres fournis par l'heuristique proposée (en jaune). Le balayage paramétrique propose de traiter les signaux de débit de liqueur blanche des indicateurs en utilisant db14, échelle 6, alors que l'heuristique suggère db13, niveau 6. Par conséquent, l'heuristique est à un *vanishing moment* près des résultats du balayage paramétrique. Ceci est acceptable pour l'utilisation prévue des signaux filtrés. Par conséquent, les heuristiques ont démontré une performance exceptionnelle

dans la réduction du bruit des signaux non stationnaires, et la qualité des signaux est jugée favorable pour la prédiction des tendances. Étant donné que tous les autres capteurs testés affichent des différences inférieures, cette validation démontre la robustesse des heuristiques générées et montre que la transformée en ondelettes peut réduire efficacement le bruit des signaux de procédés industriels. Ces heuristiques servent de base à l'extraction efficace d'informations utiles pour une analyse de prédiction.

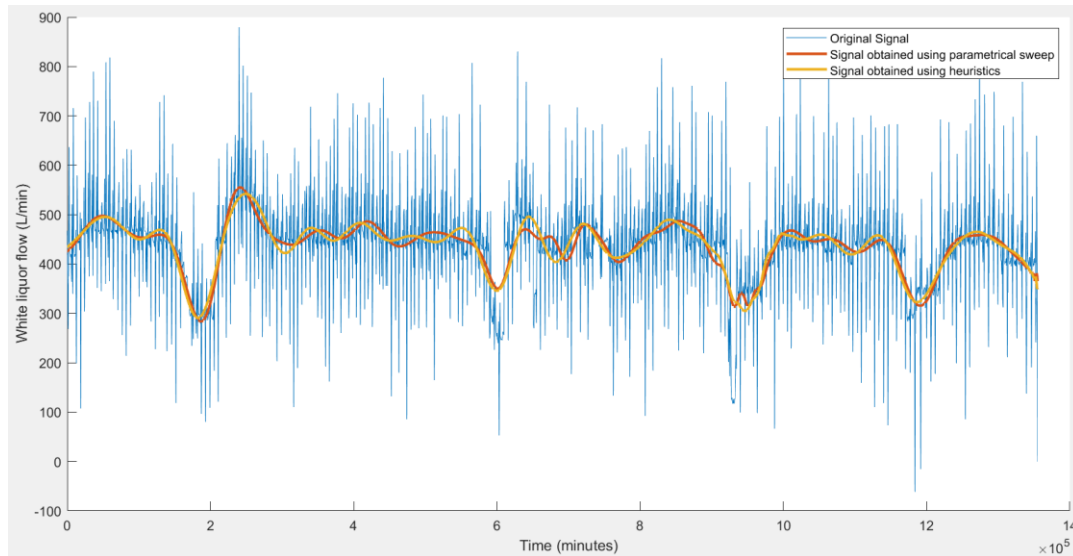


Figure 7.1 – Paramètres d'ondelettes optimaux appliqués à de nouveaux signaux pour évaluer leur applicabilité

Enfin, de la même manière que les signaux sont toujours traités en fonction d'une application ciblée, la fréquence de stockage dépend de l'objectif de l'analyse. La fréquence d'échantillonnage et le stockage doivent tenir compte de la dynamique du procédé (comment le procédé évolue dans le temps). Lorsque la dynamique d'un procédé est rapide, il convient d'enregistrer les données moyennes sur un pas de temps court, alors que pour une dynamique lente, un pas de temps plus important peut être autorisé; une moyenne de 10 minutes peut être courte pour certaines analyses et trop longue pour d'autres. Pour certains équipements de procédé, la vitesse d'un phénomène étudié (par exemple, une réaction) peut être élevée et la fréquence d'échantillonnage et de stockage doit donc être ajustée. Il en va de même si un phénomène est évalué sur une longue période de temps; il n'est pas nécessaire d'échantillonner toutes les secondes.

L'objectif du traitement du signal étant d'extraire la tendance quotidienne des signaux industriels dans un contexte de prédiction (afin de la prédire dans le futur), prendre la moyenne sur un intervalle de 10 minutes des variables qui sont échantillonnées plus rapidement ne pose pas de problème. Dans le cas des

variables échantillonnées à un intervalle inférieur à 10 minutes, l'interpolation est supprimée et l'analyse FFT permet d'identifier la bonne fréquence à partir de laquelle la tendance peut être observée. Cependant, les résultats de cette analyse sont spécifiques à l'intervalle de stockage de 10 minutes; il a un impact sur l'heuristique. L'impact de l'intervalle d'échantillonnage et de stockage sur les paramètres optimaux des ondelettes pourrait être évalué et élucidé dans le cadre d'une autre recherche.

Les signaux industriels traités pourront être utilisés pour construire un modèle de prédiction de la tendance pour la surveillance, l'optimisation et la prise de décision; les variations instantanées ne sont pas ciblées. En effet, un modèle peut être utilisé pour prédire une variable bien qu'elle soit mesurée. Ce dernier peut servir au contrôle *feed forward*, à prédire l'impact d'une modification sur le procédé et à mieux comprendre le procédé dans son ensemble. Il pourrait être pertinent dans de futurs travaux de confirmer si la prédiction est effectivement améliorée après le retrait du bruit, c'est-à-dire si la corrélation entre les variables est davantage au sein de la tendance que dans les variations instantanées. Également, outre le signal prédit (le Y dans un modèle de projection sur structures latentes) traité pour ne conserver que les éléments à la fréquence d'intérêt, les variables qui le prédisent (les X) pourraient être traitées avec comme objectif de maximiser la corrélation R^2 entre les X et la variable prédite.

En conclusion, la mise en évidence des relations entre les paramètres de classification des signaux, les caractéristiques des signaux et les paramètres de la transformée en ondelettes a permis de mieux comprendre le traitement des signaux industriels et a fourni des heuristiques. Les heuristiques conduisent à des actions utilisant un minimum de temps, de connaissances et de calculs; les défis de la complexité des calculs sont surmontés en utilisant les heuristiques. Elles sont utilisées pour des tâches difficiles et le traitement de milliers de signaux de procédés industriels ne fait pas exception. Dans certains cas, les heuristiques peuvent être considérées comme des lignes directrices de première approximation pour les tâches techniques fastidieuses. Les heuristiques tiennent compte des relations potentiellement inconnues ou inexplorées dans l'ensemble de données; elles mettent en évidence les relations dont les experts et experts de procédé pourraient ne pas avoir connaissance. En outre, avec les heuristiques, il est possible d'obtenir des bénéfices plus rapidement et de les maintenir avec un minimum d'efforts; elles sont utilisées pour économiser du temps et des ressources.

De ce fait, la comparaison du filtre de Kalman, de la moyenne mobile à pondération exponentielle, de la transformée de Fourier à court terme et de la transformée en ondelettes a confirmé que cette dernière est

performante et robuste pour retirer le bruit de signaux de procédés industriels. Également, il a été constaté qu'en traitant de nombreux signaux du même type et de même nature, des heuristiques permettant un traitement plus rapide et efficace des données pouvaient être formulées. En conclusion, la sous-hypothèse #2 selon laquelle l'exploitation de la relation entre les paramètres de la transformée en ondelettes et les caractéristiques des signaux permet de définir des heuristiques qui rendent possible le traitement de nombreux signaux de manière systématique et efficace à travers le procédé est confirmée.

Outre le retrait des erreurs aléatoires, d'autres étapes de traitement sont requises afin d'obtenir des données permettant une prise de décision fiable. De ce fait, la section 7.3 discute l'obtention d'un cadre méthodologique de traitement des données pour la prise de décision.

7.3 L'établissement d'un cadre méthodologique de traitement des données : régimes d'opération et coûts d'opération pour une interprétation plus concrète

À l'usine, les expertes et experts exploitent au maximum les ressources disponibles, c'est-à-dire en évitant autant que possible les modifications nécessitant un investissement en capital. Par exemple, pour optimiser et désengorger le département de lavage, ces personnes commenceraient par stabiliser les variables contrôlées du procédé. La variable la plus importante est la consistance à la sortie du réservoir de soufflage; cette consistance varie chaque fois qu'il y a une décharge du lessiveur. En stabilisant cette consistance, la pâte entrant dans le département de lavage est stable. Ensuite, étant donné que la pâte est déchargée dans un réservoir de soufflage qui n'a pas une capacité suffisante, l'usine pourrait utiliser un vieux réservoir qui n'est pas en fonction pour servir de tampon. Ce réservoir tampon permettrait également de réduire les variations de consistance. Cette stabilisation donne la possibilité d'augmenter la capacité de lavage. De plus, considérant que l'efficacité du lavage diminue avec la température puisque la lignine reste collée sur les fibres et ne se détache pas avec de l'eau froide, c'est-à-dire que le lavage de la pâte est plus facile avec une température d'eau de lavage plus élevée, il serait possible d'optimiser les opérations du département de lavage en installant un *steam heater* sur le courant d'eau de lavage pour les conditions hivernales. Ce faisant, l'écart de production entre les conditions d'été et d'hiver pourrait être réduit et le procédé pourrait rester plus stable tout au long de l'année. Par conséquent, un meilleur contrôle de la température de l'eau de traitement permettrait d'augmenter la capacité. Cependant, même si le fait d'avoir une eau plus chaude en hiver conduisait à une meilleure performance, les lessiveurs sont davantage un goulot d'étranglement en hiver que le sont les laveurs, et un *steam heater* de grande capacité

serait nécessaire pour chauffer l'eau d'hiver afin d'atteindre la même température qu'en été. Une analyse plus approfondie devrait donc évaluer les avantages par rapport aux efforts requis. Une dernière option, qui n'est pas sans coût d'investissement, consiste à remplacer les laveurs existants par des modèles plus efficaces.

Par conséquent, en considérant les étapes entreprises en pratique par l'usine, l'analyse des données et la détection et l'identification des régimes d'opération est proposée. Dans le cadre de cette recherche, le département de lavage de la pâte brune a été optimisé (désengorgé) en tenant compte de la surimposition des données d'opération traitées par le cadre méthodologique proposé et des informations sur les coûts. Les résultats sont présentés aux expertes et experts de l'usine comme un complément à ce qu'ils font déjà au niveau de l'usine. Cette analyse a pour but de fournir des solutions significatives (ayant un impact) et rentables qui peuvent être mises en œuvre de rapidement.

L'approche systématique et structurée (cadre méthodologique) permet d'analyser les signaux de procédé historiques afin d'aider à la prise de décisions de gestion de haut niveau. Ces décisions sont prises hors ligne. Le cadre méthodologique a été validé par des expertes et experts dans le domaine du traitement des données pour inclure les meilleures pratiques de l'industrie.

Le cadre méthodologique proposé met l'accent sur le traitement des données selon un objectif établi. La connaissance du procédé en prise en considération à chaque étape de la méthodologie proposée. Également, l'emphasis est mise sur la détection et l'identification des régimes d'opération. Dans un procédé, il y a plus d'une façon d'opérer compte tenu des différents produits, points de consigne, grades, recettes, saisons, etc., et ceci est dicté par les différents régimes d'opération. Leur détection et identification est basée sur la combinaison de la connaissance du domaine et d'une approche axée sur les données. Ainsi, le potentiel des données est exploité par la considération de ces régimes d'opération.

Le traitement et la manipulation des données sont liés à l'objectif de l'analyse, à l'utilisation des données; ils sont spécifiques à la finalité. En fonction de l'objectif de cette étude de cas, le cadre méthodologique proposé, et tout particulièrement les régimes d'opération, est nécessaire. En d'autres termes, le niveau de précision acquis par l'utilisation du cadre méthodologique est nécessaire; en considérant uniquement la moyenne estivale de toutes les variables, il n'aurait pas été possible d'identifier la manière préférée d'opérer. La distinction entre les différents régimes d'opération permet de mieux comprendre ce qui se

passé dans le procédé; les régimes d'opération sont utilisés pour répondre aux questions et résoudre les problèmes dans une usine.

La résolution de ce problème d'optimisation a montré la valeur et les avantages de l'utilisation du cadre méthodologique proposé; il ajoute une valeur significative à la compréhension du procédé. Le cadre méthodologique (en particulier l'identification des régimes d'opération) donne un aperçu de la variabilité du procédé, il aide à expliquer d'où viennent les variations du procédé, quand exactement elles se sont produites et quelles variables ont été affectées. En outre, les données traitées résultant du cadre méthodologique constituent un point de départ ou une base pour de nombreuses analyses ou prises de décision qui pourraient avoir lieu à l'usine. Le fait de disposer de données (traitées) de bonne qualité et de régimes d'opération, obtenues par la combinaison d'approches axées sur les données et d'interprétation des procédés, peut entraîner une augmentation des bénéfices de l'usine. Par exemple, si l'on connaît la capacité maximale de chaque opération unitaire du procédé ou de l'équipement d'un système, les données traitées pour chaque régime d'opération permettent de recommander des changements peu coûteux pour augmenter la capacité (désengorgement).

En outre, ensemble, le cadre méthodologique et le modèle de coût axé sur les opérations augmentent la transparence, la compréhension des causes et des effets et la granularité (détail) des résultats d'optimisation, et permettent l'identification des variables critiques au niveau des opérations. Ils permettent également de mieux comprendre les coûts de production, d'améliorer le contrôle des coûts grâce à l'analyse comparative des indicateurs clés de performance (variables de procédé dans le Tableau 6.3) et d'augmenter la rentabilité globale de l'usine.

Par conséquent, le cadre méthodologique permet de mieux comprendre les opérations d'un procédé et l'analyse des coûts axée sur les opérations favorise une prise de décision éclairée. Elle fournit des informations sur les différents objets de coûts et étudie leur rentabilité dans le cadre d'un processus décisionnel. La prise de décision basée sur les données issues du cadre méthodologique améliore l'efficacité et la rentabilité des opérations. La superposition des données traitées du cadre méthodologique aux données de l'analyse économique basée sur le concept d'*activity-based costing*, en tenant compte des opérations de l'usine, permet de prendre de meilleures décisions concernant les opérations. Cela fournit le coût marginal de chaque régime d'opération en attribuant les frais généraux aux objets de coût en réponse directe à leur consommation et identifie le régime d'opération qui crée le plus de valeur.

En résumé, l'attribution de coûts aux régimes d'opération permet de résoudre des problèmes d'opération et d'identifier les régimes qui sont les plus rentables. Ainsi, afin de répondre aux questions opérationnelles, la définition des différents régimes est essentielle. Les données économiques combinées aux régimes d'opération ont permis de découvrir des informations cachées sur les différents facteurs de rentabilité, de minimiser les coûts d'opération du département de lavage et d'identifier les actions à entreprendre pour augmenter les bénéfices.

Par conséquent, le fait de fonctionner dans les conditions préférées prescrites par le régime d'opération 12 (Tableau 6.3) conduit à une production de pâte lavée accrue ou plus élevée. Cette augmentation de la productivité rime avec haute performance; produit de haute qualité, réduction des coûts, augmentation des revenus et flexibilité opérationnelle. Ces conditions d'opération préférées augmentent la rentabilité. Par conséquent, l'analyse avancée des données combinée à son contexte (connaissance du procédé) donne plus d'informations et conduit à une compréhension plus profonde du procédé.

En conclusion, il a été démontré par l'élaboration d'un cadre méthodologique et son application à un problème d'optimisation (désengorgement) que la fusion des régimes d'opération définis par des conditions reconnues du procédé (recettes ou analyse "*top-down*") avec ceux détectés par l'analyse des données (analyse "*bottom-up*") fournit un ensemble de régimes d'opération qui permettent d'interpréter plus concrètement l'opération du procédé et donc de prendre de meilleures décisions liées à la conception et aux opérations concernant le procédé. Ainsi, la sous-hypothèse #3 est confirmée.

CHAPITRE 8 CONCLUSION

Les données de procédé ont une valeur limitée tant qu'elles ne sont pas transformées en informations. L'utilisation d'approches de traitement des données permet de tirer pleinement parti des données collectées. Les approches de traitement des données sont utilisées pour répondre à de nombreux contextes de prise de décision industrielle, y compris des facettes de l'excellence opérationnelle telles que le contrôle des procédés, la détection et le diagnostic de fautes, l'optimisation, la surveillance et la réduction des coûts; la gestion des actifs telle que la maintenance prédictive; les analyses environnementales (émissions, rejets), et l'analyse énergétique (consommation, efficacité, intensité, etc.).

Des entrevues effectuées avec des expertes et experts dans le domaine du traitement des données et personnel d'usine dans le secteur des pâtes et papier ont mis de l'avant que le traitement des données est une étape préalable essentielle, et selon l'application, la méthodologie de traitement des données varie. Également, il a été constaté que la plupart des applications de prise de décision pour l'amélioration des procédés réalisées par les spécialistes en usines supposent un paradigme de régime permanent. Finalement, la prise de décision éclairée dans les usines nécessite une connaissance approfondie des procédés, qui, dans une certaine mesure, peut être optimisée par l'identification des régimes d'opération à l'état d'équilibre. La compréhension de ces régimes peut donc potentiellement créer de la valeur en aidant à l'interprétation du procédé, en fournissant de meilleures données exploitables et en permettant une meilleure prise de décision. Bien que les spécialistes en usine soient capables d'envisager les avantages de la détection et l'identification des régimes d'opération, leur prise en compte est actuellement sous-exploitée dans l'industrie et représente une opportunité importante pour la prise de décision.

Dans le cadre du traitement des données, le filtrage est essentiel, car le bruit affecte directement la précision du système de contrôle, l'évaluation des performances du procédé, le diagnostic de fautes et la fiabilité des données. De plus, la réduction du bruit est nécessaire pour améliorer les performances, la précision et la robustesse des prédictions. Ainsi, elle doit être effectuée avant toute analyse, notamment l'optimisation, l'analyse des coûts et la simulation des procédés. La comparaison de différentes approches de traitement des erreurs aléatoires a permis d'identifier que la transformée en ondelettes est la plus performante. Elle s'est avérée robuste pour la réduction du bruit dans les signaux de procédés industriels et adéquate pour améliorer leur rapport signal/bruit et réduire l'erreur quadratique moyenne. En fait, la transformée en ondelettes a permis en moyenne une réduction de 14,14% de l'erreur quadratique

moyenne et une amélioration de 3,12% du rapport signal/bruit comparativement au filtre de Kalman, à la moyenne mobile à pondération exponentielle et à la transformée de Fourier à court terme. Il s'agit d'une bonne technique de traitement des signaux qui permet de réduire le bruit tout en conservant les informations utiles; elle conserve la variabilité du signal original, qui reflète les véritables changements et caractéristiques du procédé.

La sélection des paramètres optimaux des transformées des ondelettes étant fastidieuse et peu pratique, des heuristiques permettant leur sélection en fonction des caractéristiques et des paramètres de classification des signaux sont formulées à l'aide d'une analyse en composantes principales. L'analyse a révélé qu'il est possible de généraliser et d'extraire des heuristiques de traitement du signal en fonction du type de capteur, c'est-à-dire si le capteur est dans une boucle de contrôle ou non, du type de mesure et de la nature du fluide. Les paramètres optimaux sont obtenus par un balayage paramétrique.

Les heuristiques de traitement des signaux sont un moyen robuste et pratique d'accélérer le traitement de grands ensembles de données (milliers de signaux) provenant de procédés industriels. Elles permettent de traiter les signaux de manière structurée, systématique et efficace. Elles simplifient le travail des expertes et experts qui peuvent alors se concentrer sur des tâches à plus haute valeur ajoutée. Les heuristiques proposées sont validées; il est prouvé qu'elles sont efficaces pour réduire le bruit des signaux d'un procédé industriel. Les paramètres suggérés par les heuristiques sont applicables pour traiter tous les signaux de l'usine. Sur un autre site, les paramètres seraient très probablement différents, mais la méthodologie reste la même.

Ainsi, bien que la quantité de données de procédé disponibles ne cesse d'augmenter, leur qualité ne doit pas être compromise. De ce fait, des actions doivent être entreprises pour s'assurer que les données soient de bonne qualité. Le cadre méthodologique proposé met à profit les avantages de la connaissance du domaine ainsi que des approches axées sur les données afin de prendre des décisions de gestion pour des procédés complexes. Cette combinaison d'expertise en matière de procédé (connaissance du domaine) et d'approches axées sur les données commence par la définition du champ d'application, qui est l'objectif de l'analyse. Ensuite, les régimes d'opération "*top-down*" sont pris en compte. Une fois les données collectées, un prétraitement des données est effectué et les périodes en régime permanent sont détectées. L'algorithme de détection de ces périodes fournit un moyen fiable et efficace d'identifier le moment où chaque variable de référence est à l'état d'équilibre sur une fenêtre temporelle convenablement choisie.

Ensuite, une réconciliation des valeurs moyennes des variables sur la durée de chaque période en régime permanent est effectuée. Cette étape met en évidence les capteurs qui sont en faute de manière persistante.

La dernière étape du cadre méthodologique proposé est la détection et l'identification des régimes d'opération "*bottom-up*" par la synergie entre la connaissance du domaine et les techniques mathématiques. Dans les procédés industriels, il y a différentes recettes, différents produits ou grades sont obtenus, la matière première peut varier, les conditions météorologiques peuvent changer, différentes personnes opérant le procédé peuvent faire fonctionner le procédé de manière distincte, etc. Tous ces aspects peuvent conduire à des régimes d'opération distincts dans une usine. Certains des changements susmentionnés sont connus, mais d'autres peuvent être cachés dans les données du procédé. Par conséquent, la détection des régimes d'opération permet de mieux comprendre le fonctionnement de l'usine en fournissant des informations supplémentaires, d'approfondir l'analyse du procédé pour optimiser la gestion de l'usine et potentiellement amener le procédé à mieux opérer.

Somme toute, l'intégration de la détection des régimes d'opération, telle que décrite dans ce cadre, est un moyen d'atteindre le plein potentiel des données de procédé pour la prise de décision. Les données de procédé interprétées en fonction des régimes d'opération permettent de prendre des décisions de conception et d'opération plus éclairées. Finalement, la reconnaissance des différents régimes d'opération ouvre la porte à de nombreuses nouvelles opportunités.

Le présent travail porte sur l'amélioration de la prise de décision en matière de conception et d'opération dans des environnements industriels complexes à l'aide d'un cadre méthodologique de gestion des données en temps réel. Ce dernier permet de prendre des décisions plus éclairées et peut être utilisé pour obtenir une conception et une exploitation plus efficaces et plus respectueuses de l'environnement des installations industrielles.

Une application du cadre méthodologique à une usine de pâte à dissoudre a démontré que cet outil pratique de traitement des données est adapté pour être déployé dans les usines et qu'il peut être mis en œuvre dans les industries de transformation. En effet, bien qu'il fût appliqué au secteur des pâtes et papier, il est applicable à l'amélioration des processus industriels continus et semi-continus. Cependant, chaque étape du traitement des données doit être adaptée au type d'industrie ainsi qu'à l'usine analysée en particulier. Ainsi, le cadre peut être généralisé, mais les paramètres sont spécifiquement sélectionnés, car les mesures provenant des capteurs sont différentes selon l'industrie, l'usine et l'instrument.

La valeur et le potentiel du cadre méthodologique proposé sont illustrés dans une application d'optimisation des opérations cherchant à désengorger le département de lavage de la pâte. Les opérations du département de lavage d'une usine de pâte à dissoudre sont évaluées afin d'identifier le régime d'opération préféré en termes de débit de pâte, de qualité de pâte (conductivité) et de profits engendrés. Plus précisément, la fonction objectif d'optimisation de la performance de lavage considère la maximisation du débit de pâte et de la qualité du lavage de la pâte ainsi que la minimisation de la quantité d'antimousse et de produits chimiques de blanchiment, c'est-à-dire la maximisation du profit. L'analyse des coûts prend en compte la philosophie de la comptabilité par activité; le taux de consommation des ressources est lié à une activité, et l'activité est ensuite liée au produit.

Par conséquent, les données de procédé traitées obtenues à partir du cadre méthodologique proposé, combinées aux données économiques obtenues par l'analyse des coûts basée sur les activités, sont interprétées pour identifier les opérations rentables. Le régime d'opération le plus rentable est détecté et les raisons expliquant sa haute performance sont évaluées, c'est-à-dire que les variables contrôlées qui semblent influencer les performances du département de lavage sont identifiées, grâce à la combinaison de techniques axées sur les données et la connaissance du procédé. Une analyse plus poussée, reposant sur un plan d'expérience, pourrait isoler l'effet de chaque variable, évaluer leurs relations, prouver que ces variables ont un impact sur les performances, expliquer davantage comment elles conduisent à des performances plus élevées et confirmer le scénario d'opération préféré. Par conséquent, pour prendre des décisions informées basées sur un grand volume de données d'opération collectées au sein de procédés industriels complexes, les informations sur les coûts et les régimes d'opération doivent être pris en compte.

En conclusion, l'utilisation d'un cadre de gestion des données de procédé permet d'identifier les avantages potentiels que l'usine pourrait obtenir; il apporte donc une valeur considérable aux expertes et experts du procédé. La connaissance des régimes d'opération améliore le processus de prise de décision et la capacité à résoudre des problèmes d'opération et de conception. Ainsi, l'ensemble des travaux réalisés dans le cadre de cette recherche démontre qu'en utilisant un cadre méthodologique de traitement des données qui identifie les régimes d'opération en régime permanent et en considérant leurs coûts d'opération associés, les opérations et la conception des améliorations d'un procédé peuvent être optimisées, de sorte que des modifications rentables et des possibilités de désengorgement soient identifiées.

Quelques avenues de développement de cette recherche ont été identifiées. Premièrement, il serait intéressant de sonder et d'interroger des expertes et experts pour valider et justifier les paramètres identifiés. Par conséquent, une méthodologie hybride combinant les observations des relations identifiées dans l'ACP et les entrevues avec des expertes et experts est une avenue intéressante. Les méthodes de collecte de données d'entrevues donnent la possibilité d'obtenir des heuristiques dérivées empiriquement. Ainsi, les heuristiques qui peuvent être corroborées par des expertes et experts seraient potentiellement plus raffinées que celles basées uniquement sur des approches mathématiques.

De plus, toujours en lien avec les heuristiques formulées, des travaux futurs pourraient porter sur l'exploration d'autres ensembles de données, le test de diverses fréquences d'échantillonnage et l'extension de la méthodologie de formulation d'heuristiques dans le but d'inclure des capacités supplémentaires de détection des erreurs. L'application de cette méthodologie à d'autres usines ou à d'autres types de procédés (autre que celui des pâtes et papier) est également envisagée comme un domaine intéressant pour des activités de recherche supplémentaires à mener. Il est également pertinent d'étendre le travail actuel en approfondissant l'analyse de la configuration du procédé et du type de capteur.

Également, une recherche future pourrait analyser comment les caractéristiques des signaux entre différentes usines d'une même industrie varient. Il serait logique de croire que la même mesure prise dans différentes usines de pâte et papier serait différente considérant que les opératrices et opérateurs n'agissent pas de la même façon, les instruments de mesure ne sont pas installés de la même façon, etc. Enfin, différentes caractéristiques riment avec de nouvelles façons de traiter les données. Ainsi, la généralisation du traitement des signaux d'une autre usine pourrait être comparée avec les résultats obtenus dans le cadre de cette recherche.

En lien avec le cadre méthodologique proposé, les expertes et experts ont mentionné que la valeur du cadre pourrait augmenter considérablement s'il pouvait être utilisé en ligne, pour établir, par exemple, si les opérations dévient des opérations normales, et pour évaluer la cause sous-jacente. Des travaux futurs pourraient s'attarder à cette question; le cadre méthodologique conserverait *a priori* les mêmes étapes, il s'agirait essentiellement d'identifier des approches qui s'utilisent en ligne. En outre, les données traitées dans ce cas représentent des opérations en régime permanent; les états transitoires ne sont pas pris en compte, mais leur analyse représente un potentiel intéressant et peut faire l'objet de futurs travaux. Par

exemple, pour maximiser son profit, il pourrait être pertinent pour une usine d'évaluer le coût de la transition entre les régimes permanents et les régimes transitoires, et d'essayer de minimiser le temps de transition. En effet, afin de réduire le coût ou la durée de la transition d'un grade à l'autre, les périodes transitoires sont évaluées pour identifier les périodes de changement de grade (et non les périodes instables présentes au sein de la production d'un même grade) et déterminer où sont les transitions les plus longues et à partir de quel grade elles se produisent. Une meilleure planification des opérations permettrait de résoudre ce problème.

Pour l'analyse des périodes transitoires, le cadre méthodologique devra être modifié légèrement. Les changements requis au cadre méthodologique incluent la détection des régimes, la réconciliation et la détection des régimes d'opération. En effet, au lieu de ne pas prendre en compte les périodes qui ne sont pas stables, l'accent serait mis sur ces périodes transitoires. La détection du régime permanent serait toujours de mise afin de déterminer quelles sont les périodes transitoires. Par conséquent, suite à cette étape, des étapes parallèles à celles existantes pourraient envisager une réconciliation dynamique des données et une analyse dynamique en composante principale pour la détection et l'identification des régimes d'opération transitoires.

Enfin, au lieu de surimposer un modèle de coût aux données obtenues à partir du cadre proposé, une analyse environnementale aurait pu être effectuée pour mettre en évidence le régime le plus problématique et réduire l'impact environnemental.

RÉFÉRENCES

- Abbaszadeh, P., & Alipour, A. (2017). Development of a coupled wavelet transform and evolutionary Levenberg-Marquardt neural networks for hydrological process modeling. *Computational Intelligence*, 34, 1-25. <https://doi.org/10.1111/coin.12124>
- Abramovich. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika*, 85(1), 115-129. <https://doi.org/10.1093/biomet/85.1.115>
- Abu-el-zeet, Z. H., Becerra, V. M., & Roberts, P. D. (2001). Steady-State Optimisation Using Data Reconciliation and Bias Estimation. *IFAC Proceedings Volumes*, 34(8), 283-288. [https://doi.org/https://doi.org/10.1016/S1474-6670\(17\)40830-5](https://doi.org/https://doi.org/10.1016/S1474-6670(17)40830-5)
- AccountingTools. (2018). *Cost accounting definition*. <https://www.accountingtools.com/articles/what-is-cost-accounting.html>
- Aldrich, C., & van Deventer, J. S. J. (1993). Use of neural nets to detect systematic errors in process systems. *International Journal of Mineral Processing*, 39(3-4), 173-197.
- Alyasseri, Z. A. A., Khader, A. T., Al-Betar, M. A., Abasi, A. K., & Makhadmeh, S. N. (2020). EEG Signals Denoising Using Optimal Wavelet Transform Hybridized With Efficient Metaheuristic Methods. *IEEE Access*, 8, 10584-10605. <https://doi.org/10.1109/ACCESS.2019.2962658>
- Alyasseri, Z. A. A., Khader, A. T., Al-Betar, M. A., & Awadallah, M. A. (2018). Hybridizing β -hill climbing with wavelet transform for denoising ECG signals. *Information Sciences*, 429, 229-246. <https://doi.org/https://doi.org/10.1016/j.ins.2017.11.026>
- Arakawa, K. (1994). Fuzzy rule-based signal processing and its application to image restoration. *IEEE Journal on Selected Areas in Communications*, 12(9), 1495-1502. <https://doi.org/10.1109/49.339917>
- Averkamp, H. (2019). *What is cost accounting?* Accounting Coach. <https://www.accountingcoach.com/blog/what-is-cost-accounting>
- Bagajewicz, M. J. (2000). *Process Plant Instrumentation: Design and Upgrade*. CRC Press
- Bakshi, B. R. (1999). Multiscale analysis and modeling using wavelets. *Journal of Chemometrics*, 13(3-4), 415-434. [https://doi.org/10.1002/\(SICI\)1099-128X\(199905/08\)13:3<43.0.CO;2-8](https://doi.org/10.1002/(SICI)1099-128X(199905/08)13:3<43.0.CO;2-8)
- Bakshi, B. R., & Stephanopoulos, G. (1994). Representation of process trends—III. Multiscale extraction of trends from process data. *Computers & Chemical Engineering*, 18(4), 267-302. [https://doi.org/https://doi.org/10.1016/0098-1354\(94\)85028-3](https://doi.org/https://doi.org/10.1016/0098-1354(94)85028-3)
- Barton, M., & Lennox, B. (2022). Model stacking to improve prediction and variable importance robustness for soft sensor development. *Digital Chemical Engineering*, 3, 100034. <https://doi.org/https://doi.org/10.1016/j.dche.2022.100034>
- Bazin, C., Hodouin, D., Duchesne, C., Thibault, J., & Trusiak, A. R. (1998). Reconciliation of Mass and Energy Data Measurements: Application to a Rotary Dryer. *Canadian Metallurgical Quarterly*, 37(3-4), 333-342. <https://doi.org/10.1179/cmqr.1998.37.3-4.333>
- Beaver, S., Palazoglu, A., & Romagnoli, J. A. (2007). Cluster Analysis for Autocorrelated and Cyclic Chemical Process Data. *Industrial & Engineering Chemistry Research*, 46(11), 3610-3622. <https://doi.org/10.1021/ie060544v>

- Bellec, S. (2004). *Réconciliation des données en ligne*, Polytechnique].
- Bellec, S., Jiang, T., Kerr, B., Diamond, M., & Stuart, P. (2007). On-line processing and steady-state reconciliation of pulp and paper mill process data. *108*(6), 36-40.
- Ben Slama, A., Lentka, Ł., Mouelhi, A., Diouani, M. F., Sayadi, M., & Smulko, J. (2018). Application of statistical features and multilayer neural network to automatic diagnosis of arrhythmia by ECG signals. *Metrology and Measurement Systems*, 25. <https://doi.org/10.24425/118163>
- Canada, G. o. (2022). *EXPLORE*. <https://www.rncan.gc.ca/cartes-outils-et-publications/outils/outils-modelisation/explore/24825>
- Cantrell, J. (2019). *Brownstock washing cost optimization* TAPPI, St. Louis.
- Cao, Z., Jia, H., Zhao, T., Fu, Y., & Wang, Z. (2022). An Adaptive Self-Organizing Migration Algorithm for Parameter Optimization of Wavelet Transformation. *Mathematical Problems in Engineering*, 2022, 6289215. <https://doi.org/10.1155/2022/6289215>
- CBC. (2009). Pulp and paper woes reach far and wide. *CBC News*. <https://www.cbc.ca/news/business/pulp-and-paper-woes-reach-far-and-wide-1.830189>
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Comput. Surv.*, 41. <https://doi.org/10.1145/1541880.1541882>
- Chew, I. M. L., Foo, D. C. Y., Bonhivers, J.-C., Stuart, P., Alva-Argaez, A., & Savulescu, L. E. (2013). A model-based approach for simultaneous water and energy reduction in a pulp and paper mill. *Applied Thermal Engineering*, 51(1), 393-400. [https://doi.org/https://doi.org/10.1016/j.applthermaleng.2012.08.070](https://doi.org/10.1016/j.applthermaleng.2012.08.070)
- Chiang, L., Lu, B., & Castillo, I. (2017). Big Data Analytics in Chemical Engineering. *Annual review of chemical and biomolecular engineering*, 8. <https://doi.org/10.1146/annurev-chembioeng-060816-101555>
- Cooper, R., & Kaplan, R. S. (1988). How Cost Accounting Distorts Product Costs. *Management Accounting*, 69(10), 20.
- Cooper, R., & Kaplan, R. S. (1990). Measure Costs Right: Make The Right Decision. *The CPA Journal*, 60(2), 38.
- Cooper, R., & Kaplan, R. S. (1991). Profit Priorities from Activity-based Costing. *Harvard business review U6*, 69(3), 130-135.
- Dambros, J. W. V., Farenzena, M., & Trierweiler, J. O. (2018). Signal Preprocessing for Stiction Detection Methods. *Industrial & Engineering Chemistry Research*, 57(1), 302-315. <https://doi.org/10.1021/acs.iecr.7b04441>
- Daubechies, I. (1992). 1. The What, Why, and How of Wavelets. Dans *Ten Lectures on Wavelets* (p. 1-16).
- Delou, P. A., Ribeiro, L. D., Paiva, C. R., Niederberger, J., Gomes, M. V. C., & Secchi, A. R. (2021). A Real-Time Optimization Strategy for Small-Scale Facilities and Implementation in a Gas Processing Unit. *9*(7), 1179. <https://www.mdpi.com/2227-9717/9/7/1179>

- Desbiens, A., Nunez, E., Del Villar, R., Hodouin, D., & Poulin, E. (2008). Using process control to increase the energy efficiency of mineral and metal processing plants. *International Journal of Power and Energy Systems*, 28(2), 146-151.
- Dong, Y., Li, Y., Xiao, M., & Lai, M. (2009). Unscented Kalman filter for time varying spectral analysis of earthquake ground motions. *Applied Mathematical Modelling*, 33(1), 398-412. <https://doi.org/https://doi.org/10.1016/j.apm.2007.11.020>
- Donoho, D. L., Johnstone, I. M., Kerkycharian, G., & Picard, D. (1995). Wavelet Shrinkage: Asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2), 301-369. www.jstor.org/stable/2345967
- Dunn, K. (2022a). Box plots. Dans *Process improvement using data*. <https://learnche.org/pid/data-visualization/box-plots>
- Dunn, K. (2022b). Principal Component Analysis (PCA). Dans *Process improvement using data*.
- Dyason, W., Van Niekerk, T. I., Phillips, R., & Stopforth, R. (2017). *Performance evaluation and comparison of filters for real time embedded system applications*. 28th Annual Symposium of the Pattern Recognition Association of South Africa and the 10th Robotics and Mechatronics International Conference, PRASA-RobMech 2017, November 29, 2017 - December 1, 2017, Bloemfontein, South africa (vol. 2018-January, p. 242-248). <https://doi.org/10.1109/RoboMech.2017.8261155>
- El-Dahshan, E.-S. A. (2011). Genetic algorithm and wavelet hybrid scheme for ECG signal denoising. *Telecommunication Systems*, 46(3), 209-215. <https://doi.org/10.1007/s11235-010-9286-2>
- El-Halwagi, M. M., & Stuart, P. R. (2013). *Integrated biorefineries : design, analysis, and optimization*. CRC Press.
- Emblemsvåg, J. (2003). *Life–Cycle Costing : Using Activity–Based Costing and Monte Carlo Methods to Manage Future Costs and Risks*. John Wiley & Sons, Inc.
- Emblemsvåg, J., & Bras, B. (2000). *Activity-based cost and environmental management : a different approach to ISO 14000 compliance*. Kluwer Academic Publishers.
- Everitt, B. (1993). *Cluster analysis*. E. Arnold ; Halsted Press.
- Fan, H., Tariq, S., & Zayed, T. (2022). Acoustic leak detection approaches for water pipelines. *Automation in Construction*, 138, 104226. <https://doi.org/https://doi.org/10.1016/j.autcon.2022.104226>
- Fillingim, K. B., Nwaeri, R. O., Borja, F., Fu, K., & Paredis, C. J. J. (2020). Design Heuristics : Extraction and Classification Methods With Jet Propulsion Laboratory's Architecture Team. *Journal of Mechanical Design*, 142(8). <https://doi.org/10.1115/1.4044160>
- Fillingim, K. B., Shapiro, H., Paredis, C. J. J., & Fu, K. (2020). Process Heuristics : Extraction, Analysis, and Repository Considerations. *IEEE Systems Journal*, 14(4), 5148-5159. <https://doi.org/10.1109/JSYST.2019.2959538>
- Flehmg, F., Watzdorf, R. v., & Marquardt, W. (1998). Identification of trends in process measurements using the wavelet transform. *Computers & Chemical Engineering*, 22, S491-S496. [https://doi.org/https://doi.org/10.1016/S0098-1354\(98\)00092-1](https://doi.org/https://doi.org/10.1016/S0098-1354(98)00092-1)

- Fu, K., Yang, M., & Wood, K. (2015). *Design Principles: The Foundation of Design*.
- Fugal, L. D. (2009). *Conceptual Wavelets in Digital Signal Processing* (1^e éd.). Space & Signals Technical Publishing.
- Ganesan, R., Das, T. K., & Venkataraman, V. (2004). Wavelet-based multiscale statistical process monitoring : A literature review. *IIE Transactions*, 36(9), 787-806. <https://doi.org/10.1080/07408170490473060>
- Gao, Q., Duan, C., Fan, H., & Meng, Q. (2008). Rotating machine fault diagnosis using empirical mode decomposition. *Mechanical Systems and Signal Processing*, 22, 1072-1081. <https://doi.org/10.1016/j.ymssp.2007.10.003>
- García Plaza, E., Núñez López, P. J., & Beamud González, E. M. (2018). Multi-Sensor Data Fusion for Real-Time Surface Quality Control in Automated Machining Systems. *18*(12), 4381. <https://www.mdpi.com/1424-8220/18/12/4381>
- García Plaza, E., Núñez López, P. J., & Beamud González, E. M. (2019). Efficiency of vibration signal feature extraction for surface finish monitoring in CNC machining. *Journal of Manufacturing Processes*, 44, 145-157. <https://doi.org/https://doi.org/10.1016/j.jmapro.2019.05.046>
- Greenwood, T. G., & Reeve, J. M. (1992). Activity-based cost management for continuous improvement: A process design framework. *Journal of Cost Management the Manufacturing Industry* 5(4), 22-40. <https://maaw.info/ArticleSummaries/ArtSumGreenwoodReeve92.htm>
- Harashima, H., Odajima, K., Shishikui, Y., & Miyakawa, H. (1982). ϵ -separating nonlinear digital filter and its applications. 65(4), 11-19. <https://doi.org/https://doi.org/10.1002/ecja.4410650403>
- Harkat, S., Boukharouba, K., & Douaoui, A. E. K. (2016). Multi-site modeling and prediction of annual and monthly precipitation in the watershed of Cheliff (Algeria). *Desalination and Water Treatment*, 57(13), 5959-5970. <https://doi.org/10.1080/19443994.2014.956798>
- Harrison, R. P., & Stuart, P. (2007). Spatial and temporal resolution in data-driven process modeling of an integrated newsprint mill. *Chemical Product and Process Modeling*, 2(1). <https://doi.org/10.2202/1934-2659.1041>
- Harrison, R. P., & Stuart, P. R. (2011). *Data pre-processing techniques for multivariate analysis to treat industrial operating data for retrofit design*.
- Hasan, A. M., Samsudin, K., & Ramli, A. R. (2011). Intelligently tuned wavelet parameters for GPS/INS error estimation. *International Journal of Automation and Computing*, 8(4), 411-420. <https://doi.org/10.1007/s11633-011-0598-9>
- Heikkinen, M., Poutiainen, H., Liukkonen, M., Heikkinen, T., & Hiltunen, Y. (2011). Subtraction analysis based on self-organizing maps for an industrial wastewater treatment process. *Mathematics and Computers in Simulation*, 82(3), 450-459. <https://doi.org/https://doi.org/10.1016/j.matcom.2010.10.021>
- Hsieh, C. L., Tsai, W. H., & Chang, Y. C. (2020). Green Activity-Based Costing Production Decision Model for Recycled Paper. *Energies*, 13(10), 23. Article. <https://doi.org/10.3390/en13102413>
- Huang, G. (2021). Missing data filling method based on linear interpolation and lightgbm. *Journal of Physics : Conference Series*, 1754, 012187. <https://doi.org/10.1088/1742-6596/1754/1/012187>

- Hytönen, E., & Stuart, P. (2011). Operations driven cost impact evaluation of kraft process retrofit projects for capital appropriation : case of forest biorefinery. *Journal of Production Economics*.
- Ibrahim, R. K., Watson, S. J., Djurović, S., & Crabtree, C. J. (2018). An Effective Approach for Rotor Electrical Asymmetry Detection in Wind Turbine DFIGs. *IEEE Transactions on Industrial Electronics*, 65(11), 8872-8881. <https://doi.org/10.1109/TIE.2018.2811373>
- Jain, A. K. (2010). Data clustering : 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666. <https://doi.org/https://doi.org/10.1016/j.patrec.2009.09.011>
- Janssen, M., Naliwajka, P., & Stuart, P. (2012). Development of an operations-driven cost model for continuous processes part II: Retrofit process design application. 2, 43-54.
- Janssen, M., & Stuart, P. R. (2004). Business Model Framework for Using Real-Time Process Data for Design Decision Making.
- Jiang, T., Chen, B., & He, X. (2000). Industrial application of Wavelet Transform to the on-line prediction of side draw qualities of crude unit. *Computers & Chemical Engineering*, 24(2-7), 507-512.
- Jiang, T., Chen, B., He, X., & Stuart, P. (2003). Application of steady-state detection method based on wavelet transform. *Computers & Chemical Engineering*, 27(4), 569-578. [https://doi.org/10.1016/s0098-1354\(02\)00235-1](https://doi.org/10.1016/s0098-1354(02)00235-1)
- Jiang, T., Chen, B., Jasim, K., & Stuart, P. R. (2003). *Strategy for improving data quality for a kraft pulp mill recausticizing plant* Proceedings of the Computer-Aided Process Operations (FOCAPO) Conference, Coral Springs, Florida, USA.
- Joe Qin, S., Guo, S., Li, Z., Chiang, L. H., Castillo, I., Braun, B., & Wang, Z. (2021). Integration of process knowledge and statistical learning for the Dow data challenge problem. *Computers & Chemical Engineering*, 153, 107451. <https://doi.org/https://doi.org/10.1016/j.compchemeng.2021.107451>
- Kaneko, H., & Funatsu, K. (2017). Improvement of process state recognition performance by noise reduction with smoothing methods. *Journal of Chemical Engineering of Japan*, 50(6Special Issue), 422-429. <https://doi.org/10.1252/jcej.16we325>
- Katunin, A., & Przysławka, P. (2015). Automated wavelet-based damage identification in sandwich structures using modal curvatures. *Journal of Vibroengineering*, 17, 2977-2986.
- Kelly, J., & Hedengren, J. (2013). A steady-state detection (SSD) algorithm to detect non-stationary drifts in processes. *Journal of Process Control*, 23. <https://doi.org/10.1016/j.jprocont.2012.12.001>
- Kelly, J. D. (2023). *Industrial Modeling & Programming Language (IMPL) Manual*.
- Kopra, R., Kari, E., Harinen, M., Tirri, T., & Dahl, O. (2010). Optimization of wash water usage in brown stock washing. *2010 TAPPI PEERS Conference and 9th Research Forum on Recycling*, 1, 660-695. <https://doi.org/10.32964/TJ10.9.27>
- Kopra, R., Karjalainen, S., Tirri, T., & Dahl, O. (2011). Optimisation of pressure filter performance using refractometer measurements -Mill investigations. *Appita Journal*, 64.

- Kopra, R., Koivula, K., & Dahl, O. (2015). Optimization of a brownstock washing line. *Tappi Journal*, 14, 83-91. <https://doi.org/10.32964/TJ14.2.83>
- Korbel, M. (2011). *Ph. D. Thesis : On-line steady-state data reconciliation for advanced cost analysis in the pulp and paper industry*, Polytechnique].
- Korbel, M., Bellec, S., Jiang, T., & Stuart, P. (2014). Steady state identification for on-line data reconciliation based on wavelet transform and filtering. *Computers & Chemical Engineering*, 63, 206-218. <https://doi.org/10.1016/j.compchemeng.2014.02.003>
- Korbel, M., Garrigues, L., & Stuart, P. R. (2007). *Process and Business Data Reconciliation in the Pulp & Paper Industry*. Innovations in Engineering, Pulping and Environmental, Jacksonville, FL.
- Korbel, M., & Stuart, P. (2011). Assessment and interpretation of advanced cost data for process improvement in an integrated newsprint mill. *International Journal of Production Economics*.
- Korpunen, H., & Paltakari, J. (2013). Testing an activity-based costing model with a virtual paper mill. *Nordic Pulp and Paper Research Journal*, 28(1), 146-155.
- Korpunen, H., Virtanen, P., Dahl, O., Jylhä, P., & Uusitalo, J. (2012). An activity-based cost calculation for a kraft pulp mill. *Tappi Journal*, 11(9), 19-27.
- Krishnamurthi, R., Kumar, A., Gopinathan, D., Nayyar, A., & Qureshi, B. (2020). An Overview of IoT Sensor Data Processing, Fusion, and Analysis Techniques. *Sensors*, 20, 6076. <https://doi.org/10.3390/s20216076>
- Laflamme-Mayer, M., Janssen, M., & Stuart, P. (2011). Development of an operations-driven cost model for continuous processes Part 1: Framework for design and operations decision making. *J-For-Journal of Science & Technology for forest products and processes*, 1(1), 32-41.
- Laflamme-Mayer, M., Linkewich, J., & Stuart, P. R. (2008). Methodology for using real-time process data for cost modeling and supply chain decision support in the pulp and paper industry. 109(3), 44-51.
- Lee, B., Fillingim, K. B., Binder, W. R., Fu, K., & Paredis, C. J. J. (2017). *Design Heuristics: A Conceptual Framework and Preliminary Method for Extraction*. ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (vol. Volume 7: 29th International Conference on Design Theory and Methodology). <https://doi.org/10.1115/detc2017-67467>
- Li, H., Ma, Y., Liang, K., Tian, Y., & Wang, R. (2012). Parameters optimization for wavelet denoising based on normalized spectral angle and threshold constraint machine learning. *Journal of Applied Remote Sensing*, 6, 3579. <https://doi.org/10.1117/1.JRS.6.063579>
- Liu, H., Dong, H., Liu, Z., Ge, J., Bai, B., & Zhang, C. (2017). Noise characterization for the FID signal from proton precession magnetometer. *Journal of Instrumentation*, 12, P07019-P07019. <https://doi.org/10.1088/1748-0221/12/07/P07019>
- Liu, H., Shah, S., & Jiang, W. (2004). On-line outlier detection and data cleaning. *Computers & Chemical Engineering*, 28(9), 1635-1647. <https://doi.org/https://doi.org/10.1016/j.compchemeng.2004.01.009>

- Liu, W., Chang, Q. R., Guo, M., Xing, D. X., & Yuan, Y. S. (2011). Extraction of first derivative spectrum features of soil organic matter via wavelet de-noising. *Guang Pu Xue Yu Guang Pu Fen Xi*, 31(1), 100-104.
- Liukkonen, M., Heikkinen, M., Hiltunen, T., Hälikkä, E., Kuivalainen, R., & Hiltunen, Y. (2011). Artificial neural networks for analysis of process states in fluidized bed combustion. *Energy*, 36(1), 339-347. <https://doi.org/https://doi.org/10.1016/j.energy.2010.10.033>
- Liukkonen, M., Hiltunen, T., Hälikkä, E., & Hiltunen, Y. (2011). Modeling of the fluidized bed combustion process and NOx emissions using self-organizing maps: An application to the diagnosis of process states. *Environmental Modelling & Software*, 26(5), 605-614. <https://doi.org/https://doi.org/10.1016/j.envsoft.2010.12.002>
- Ma'arif, A., Iswanto, Nuryono, A., & Alfian, R. (2020). Kalman Filter for Noise Reducer on Sensor Readings. *Signal and Image Processing Letters*, 1, 11-22. <https://doi.org/10.31763/simple.v1i2.2>
- MacGregor, J. F., Kozub, D. J., Penlidis, A., & Hamielec, A. E. (1986). State Estimation for Polymerization Reactors. *IFAC Proceedings Volumes*, 19(15), 147-152. [https://doi.org/https://doi.org/10.1016/S1474-6670\(17\)59414-8](https://doi.org/https://doi.org/10.1016/S1474-6670(17)59414-8)
- Malghan, P. G., & Kumar Hota, M. (2022). Grasshopper optimization algorithm based improved variational mode decomposition technique for muscle artifact removal in ECG using dynamic time warping. *Biomedical Signal Processing and Control*, 73, 103437. <https://doi.org/https://doi.org/10.1016/j.bspc.2021.103437>
- Mallat, S., & Hwang, W. L. (1992). Singularity detection and processing with wavelets. *IEEE Transactions on Information Theory*, 38(2), 617-643. <https://doi.org/10.1109/18.119727>
- Mallat, S., & Zhong, S. (1992). Characterization of Signals From Multiscale Edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(7), 710-732. <https://doi.org/10.1109/34.142909>
- Mallat, S. G. (2009). *A wavelet tour of signal processing : the Sparse way* (3^e éd.). Elsevier /Academic Press.
- Marinova, M., Mateos-Espejel, E., Jemaa, N., & Paris, J. (2009). Addressing the increased energy demand of a Kraft mill biorefinery : The hemicellulose extraction case. *Chemical Engineering Research and Design*, 87(9), 1269-1275. <https://doi.org/https://doi.org/10.1016/j.cherd.2009.04.017>
- MathWorks. (2019). *Continuous and Discrete Wavelet Transforms*. MathWorks. <https://www.mathworks.com/help/wavelet/gs/continuous-and-discrete-wavelet-transforms.html>
- Matthews, P., Blessing, L., & Wallace, K. M. (2002). The introduction of a design heuristics extraction method. *Advanced Engineering Informatics*, 16, 3-19. [https://doi.org/10.1016/S1474-0346\(02\)00002-2](https://doi.org/10.1016/S1474-0346(02)00002-2)
- Mikhled, A., & Khaled, D. (2008). ECG Signal Denoising By Wavelet Transform Thresholding. *American Journal of Applied Sciences*, 5. <https://doi.org/10.3844/ajassp.2008.276.281>
- Moreno, R. d. P. (2010). *Steady State Detection, Data Reconciliation, and Gross Error Detection: Development for Industrial Processes*, University of New Brunswick].

- Moridani, M., & Pouladian, M. (2009). Detection Ischemic Episodes from Electrocardiogram Signal using Wavelet Transform. *Journal of biomedical science and engineering*, 02, 239-244. <https://doi.org/10.4236/jbise.2009.24037>
- Murphy, C. B. (2019). *What are the types of costs in cost accounting?* . <https://www.investopedia.com/ask/answers/041415/what-are-different-types-costs-cost-accounting.asp>
- Narasimhan, S., & Jordache, C. (2000). *Data Reconciliation and Gross Error Detection : An Intelligent Use of Process Data*. Gulf Professional Publishing.
- Narayan, K., Mahesh, B., & Andreas, S. (2013). *An Introduction to Kalman Filtering with MATLAB Examples*. Morgan & Claypool.
- Nounou, M. N., & Bakshi, B. R. (1999). On-line multiscale filtering of random and gross errors without process models. *AIChE Journal*, 45(5), 1041-1058. <https://doi.org/10.1002/aic.690450513>
- Nounou, M. N., & Bakshi, B. R. (2000). Chapter 5 - Multiscale Methods for Denoising and Compression. Dans B. Walczak (édit.), *Data Handling in Science and Technology* (vol. 22, p. 119-150). Elsevier.
- Nourani, V., Alami, M. T., & Aminfar, M. H. (2009). A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation. *Engineering Applications of Artificial Intelligence*, 22(3), 466-472. <https://doi.org/https://doi.org/10.1016/j.engappai.2008.09.003>
- Nourani, V., Kisi, Ö., & Komasi, M. (2011). Two hybrid Artificial Intelligence approaches for modeling rainfall-runoff process. *Journal of Hydrology*, 402(1), 41-59. <https://doi.org/https://doi.org/10.1016/j.jhydrol.2011.03.002>
- Nurminen, T., Korpunen, H., & Uusitalo, J. (2009). Applying the Activity-Based Costing to Cut-to-Length Timber Harvesting and Trucking. *Silva Fennica*, 43(5), 847-870. Article. <https://doi.org/10.14214/sf.177>
- Park, S.-M., Lee, T.-J., & Sim, K.-B. (2016). Heuristic feature extraction method for BCI with harmony search and discrete wavelet transform. *International Journal of Control, Automation and Systems*, 14(6), 1582-1587. <https://doi.org/10.1007/s12555-016-0031-9>
- Park, S., Gil, M.-S., Im, H., & Moon, Y.-S. (2019). Measurement Noise Recommendation for Efficient Kalman Filtering over a Large Amount of Sensor Data. *Sensors (Basel, Switzerland)*, 19(5), 1168. <https://doi.org/10.3390/s19051168>
- Patil, R. (2015). Noise Reduction using Wavelet Transform and Singular Vector Decomposition. *Procedia Computer Science*, 54, 849-853. <https://doi.org/10.1016/j.procs.2015.06.099>
- Perez, E., & Barros, J. (15-18 Aug. 2006 2006). *Voltage Event Detection and Characterization Methods: A Comparative Study*. 2006 IEEE/PES Transmission & Distribution Conference and Exposition: Latin America (p. 1-6). <https://doi.org/10.1109/TDCLA.2006.311552>
- Pérez, E., & Barros, J. (2012). Application of Advanced Digital Signal Processing Tools for Analysis of Voltage Events in Power Systems. *International Journal of Electrical Engineering & Education*, 46(3), 211-224. <https://doi.org/10.7227/ije.46.3.1>
- Perry, M. (2010). The Exponentially Weighted Moving Average. Dans.

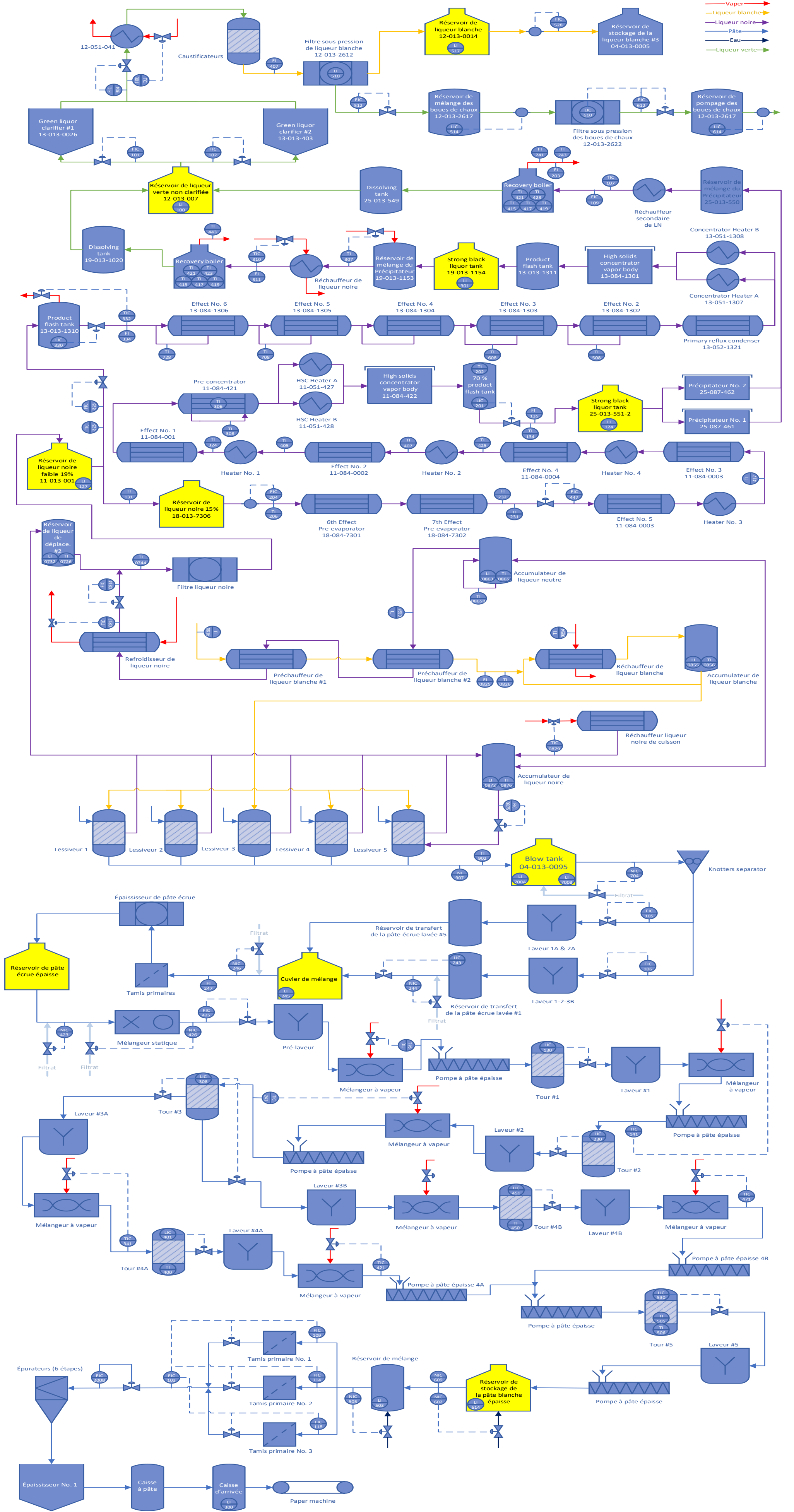
- Polikar, R. (NA). *The Wavelet Tutorial - The Engineer's Ultimate Guide to Wavelet Analysis*. <http://users.rowan.edu/~polikar/WTtutorial.html>
- Pomalaza-raez, C., & McGillem, C. D. (1984). An Adaptive, Nonlinear Edge-Preserving Filter. *IEEE Transactions on Signal Processing*, 571-576. <https://doi.org/10.1109/TASSP.1984.1164361>
- Quesada-Pineda, H. (2004). The ABCs of Cost Allocation in the Wood Products Industry: Applications in the Furniture Industry. *Virginia cooperative extension*, 420-147. <https://www.pubs.ext.vt.edu/420/420-147/420-147.html>
- Rajala, S., & Figueiredo, R. (1981). Adaptive nonlinear image restoration by a modified Kalman filtering approach. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(5), 1033-1042. <https://doi.org/10.1109/TASSP.1981.1163679>
- Ranchin, T., & Wald, L. (1992 1992). *Applications of wavelet transform in remote sensing data processing*. EARSeL Symposium 1992 "Remote sensing for monitoring the changing environment of Europe", Budapest, Hungary (p. 261-266). <https://hal-mines-paristech.archives-ouvertes.fr/hal-00466450>
- Ren, Y., Liu, P., Hu, L., Huang, J., Qiao, R., Chen, H., . . . Huang, S. (2021). Research on Noise Reduction Method of Pressure Pulsation Signal of Draft Tube of Hydropower Unit Based on ALIF-SVD. *Shock and Vibration*, 2021, 5580319. <https://doi.org/10.1155/2021/5580319>
- Reyes, J. D., Rodríguez, A. L., & Riascos, C. A. M. (2015). Data Analysis and Modelling of a Fluid Catalytic Cracking Unit (FCCU) for an Implementation of Real Time Optimization. Dans K. V. Gernaey, J. K. Huusom, & R. Gani (édit.), *Computer Aided Chemical Engineering* (vol. 37, p. 611-616). Elsevier.
- Rhinehart, R. R. (2013). *Automated steady and transient state identification in noisy processes*.
- Rivero, E. J., & Emblemståg, J. (2007). Activity-based life-cycle costing in long-range planning. *Review of Accounting and Finance*, 6, 370-390. <https://doi.org/10.1108/14757700710835041>
- Rogers, A., Guo, F., & Rasmussen, B. (2021). An Equilibrium Prediction Method for Control and Fault Detection of Energy Systems. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg*, 7(3). <https://doi.org/10.1115/1.4050039>
- Romagnoli, J., & Sanchez, M. (2000). *Data Processing and Reconciliation for Chemical Process Operations* (1^e éd., vol. 2). Academic Press
- Saari, J., & Odelius, J. (2018). Detecting operation regimes using unsupervised clustering with infected group labelling to improve machine diagnostics and prognostics. *Operations Research Perspectives*, 5, 232-244. <https://doi.org/https://doi.org/10.1016/j.orp.2018.08.002>
- Sangelkar, S., & McAdams, D. A. (2013). Mining Functional Model Graphs to Find Product Design Heuristics With Inclusive Design Illustration. *Journal of Computing and Information Science in Engineering*, 13(4). <https://doi.org/10.1115/1.4025469>
- Sawant, C., & Patil, H. T. (2014). ECG Signal De-noising using Discrete Wavelet Transform. *International Journal of Electronics Communication and Computer Engineering*, 5(4).

- Sawhney, R. S. (1991). An activity-based approach for evaluating strategic investments in manufacturing companies. *Journal of Manufacturing Systems*, 10(5), 353-367. [https://doi.org/https://doi.org/10.1016/0278-6125\(91\)90054-6](https://doi.org/https://doi.org/10.1016/0278-6125(91)90054-6)
- Seider, W. D., Seader, J. D., Lewin, D. R., & Widagdo, S. (2009). *Product and process design principles - Synthesis, analysis and evaluation* (3^e éd.). Wiley.
- Selesnick, I. (2009). The short-time Fourier transform and speech denoising. *Connexions*, <http://cnx.org/content/m32294/1.1/>.
- Shafri, H., & Yusof, R. (2009). Determination of optimal wavelet denoising parameters for red edge feature extraction from hyperspectral data. *Journal of Applied Remote Sensing*, 3. <https://doi.org/10.1117/1.3155804>
- Sigmafine. (2022). <https://www.sigmafine.com/>
- Simon, D. (2001). Kalman Filtering. *Embedded Systems Programming*.
- Smallwood, D. O., & Cap, J. S. (1998). *Salvaging pyrotechnic data with minor overloads and offsets*.
- Srinivasan, R., Wang, C., Ho, W. K., & Lim, K. W. (2004). Dynamic Principal Component Analysis Based Methodology for Clustering Process States in Agile Chemical Plants. *Industrial & Engineering Chemistry Research*, 43(9), 2123-2139. <https://doi.org/10.1021/ie034051r>
- Srinivasan, R., Wang, C., Ho, W. K., & Lim, K. W. (2005). Context-based recognition of process states using neural networks. *Chemical Engineering Science*, 60(4), 935-949. <https://doi.org/https://doi.org/10.1016/j.ces.2004.09.061>
- Su, H., Kim, Y.-M., & Chong, K. T. (2006). *A novel method for high-performance fault detection of induction machine* (vol. 6041). SPIE.
- Svensson, E. (2014). Flexibility to Seasonal Demand Variations in Pulp Mill Steam Production: The Effect of Steam Savings Leading to Off-design Heat Loads. *Applied Thermal Engineering*, 70. <https://doi.org/10.1016/j.applthermaleng.2014.04.059>
- Svensson, E., & Berntsson, T. J. C. e. t. (2013). Operational Flexibility in Pulp Mill Steam Production at Off-design Heat Loads. 35, 919-924.
- Svensson, E., Edland, R., Langner, C., & Harvey, S. (2020). Assessing the value of a diversified by-product portfolio to allow for increased production flexibility in pulp mills. 35(4), 533-558. <https://doi.org/doi:10.1515/npprj-2020-0034>
- Tan, F., Feng, X., Li, M., Wang, Z., Yang, L., Li, Y., . . . Nie, F. (2008). Construction and application of a novel library : Fourier transform infrared wavelet coefficients library. *Anal Chim Acta*, 629(1-2), 38-46. <https://doi.org/10.1016/j.aca.2008.09.017>
- Tanaka, Y. (1994). Video signal noise characteristics with regard to scanning method. *IEEE Transactions on Broadcasting*, 40(2), 68-74. <https://doi.org/10.1109/11.286286>
- Taspinar, A. (2018). *A guide for using the Wavelet Transform in Machine Learning*. <https://ataspinar.com/2018/12/21/a-guide-for-using-the-wavelet-transform-in-machine-learning/>

- Tatsiopoulos, I. P., & Panayiotou, N. (2000). The integration of activity based costing and enterprise modeling for reengineering purposes. *International Journal of Production Economics*, 66(1), 33-44. [https://doi.org/10.1016/S0925-5273\(99\)00080-8](https://doi.org/10.1016/S0925-5273(99)00080-8)
- Thibault, É., Chioua, M., McKay, M., Korbel, M., Patience, G., & Stuart, P. (2023). Experimental Methods in Chemical Engineering : State-of-the-Art Review of Industry Practice for Data Processing and Data Usage in Decision-Making. *Canadian Journal of chemical engineering*.
- Thibault, É., Kelly, J. D., Lebreux Désilets, F., Chioua, M., Poulin, B., & Stuart, P. (2023). An operating regime-based data-driven industrial framework for improved process decision-making - Part 1: Framework development *Industrial & Engineering Chemistry Research*.
- Thibault, É., Lebreux Désilets, F., Poulin, B., Chioua, M., & Stuart, P. (2023). Comparison of signal processing methods considering their optimal parameters using synthetic signals in a heat exchanger network simulation. *Computers & Chemical Engineering*.
- Thibault, É., Lebreux Désilets, F., Poulin, B., Chioua, M., & Stuart, P. (2023). Signal processing heuristics for the selection of wavelet transform parameters considering signal characteristics derived from a case study on a chemical pulp mill. *Computers & Chemical Engineering*.
- Tong, H., & Crowe, C. M. (1996). Detecting persistent gross errors by sequential analysis of principal components. *Computers & Chemical Engineering*, 20, S733-S738. [https://doi.org/https://doi.org/10.1016/0098-1354\(96\)00000-2](https://doi.org/https://doi.org/10.1016/0098-1354(96)00000-2)
- Towler, G., & Sinnott, R. (2008). *Chemical Engineering Design - Principles, Practice and Economics of Plant and Process Design* (1^e éd.). Elsevier.
- Turton, R., Bailie, R. C., Whiting, W. B., & Shaeiwitz, J. A. (2009). *Analysis, Synthesis, and Design of Chemical Processes* (3^e éd.). Prentice Hall.
- Upadhyaya, B. R., Mehta, C., & Bayram, D. (2014). Integration of Time Series Modeling and Wavelet Transform for Monitoring Nuclear Plant Sensors. *IEEE Transactions on Nuclear Science*, 61(5), 2628-2635. <https://doi.org/10.1109/TNS.2014.2341035>
- Wang, C., Shan, L., Zhou, L., & Zhang, G. (2008// 2008). *Study on the Application of ABC System in the Refinery Industry*. Global Design to Gain a Competitive Edge, London (p. 773-781).
- Wang, P., Yan, M., Zhang, L., & Yang, N. (2021). Shock Signal Trend Term Error Correction Method Based on Discrete Wavelet Transform and Low-Frequency Oscillator Combination. *Shock and Vibration*, 2021, 9939547. <https://doi.org/10.1155/2021/9939547>
- Wang, W., & Ding, J. (2003). Wavelet Network Model and Its Application to the Prediction of Hydrology. *Nat. Sci.*, 1.
- Wang, Y., & Veluvolu, K. C. (2017). Time-Frequency Analysis of Non-Stationary Biological Signals with Sparse Linear Regression Based Fourier Linear Combiner. *Sensors (Basel, Switzerland)*, 17(6), 1386. <https://doi.org/10.3390/s17061386>
- Welch, G., & Bishop, G. (2006). An Introduction to the Kalman Filter. *Proc. Siggraph Course*, 8.
- Yang, B., Xia, H., Annor-Nyarko, M., & Wang, Z. (2020). Application of total variation denoising in nuclear power plant signal pre-processing. *Annals of Nuclear Energy*, 135, 106981. <https://doi.org/https://doi.org/10.1016/j.anucene.2019.106981>

- Yinfeng, D., Yingmin, L., Mingkui, X., & Ming, L. (2008). Analysis of earthquake ground motions using an improved Hilbert–Huang transform. *Soil Dynamics and Earthquake Engineering*, 28(1), 7-19. <https://doi.org/https://doi.org/10.1016/j.soildyn.2007.05.002>
- Zhao, Z., Särkkä, S., & Rad, A. B. (17-20 Sept. 2018 2018). *Spectro-temporal ECG analysis for atrial fibrillation detection*. 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP) (p. 1-6). <https://doi.org/10.1109/MLSP.2018.8517085>
- Zhou, D., Wang, Q. J., Tian, Q. J., Lin, Q. Z., & Fu, W. X. (2009). Wavelet analysis and its application in denoising the spectrum of hyperspectral image. *Guang Pu Xue Yu Guang Pu Fen Xi*, 29(7), 1941-1945.
- Zuo-wei, P., Hong-wen, A., Shuang-yin, L., & Yi-bing, L. (31 May-2 June 2014 2014). *Amplitude envelope analysis for feature extraction of power plant blower bearing failure*. The 26th Chinese Control and Decision Conference (2014 CCDC) (p. 2821-2824). <https://doi.org/10.1109/CCDC.2014.6852653>

ANNEXE A CONFIGURATION DU PRODÉCÉ DE PÂTE DISSOUTE



ANNEXE B ARTICLE #1 EXPERIMENTAL METHODS IN CHEMICAL ENGINEERING : DATA PROCESSING AND DATA USAGE IN DECISION- MAKING

Émilie Thibault, Moncef Chioua, Maxime McKay, Milan Korbel, Gregory Patience, Paul Stuart
Submitted to the Canadian Journal of Chemical Engineering on 09/02/2023

Abstract

Industrial facilities collect large volumes of data, store them according to prescribed protocols, and then interpret them for process decision-making. Several sources and types of error contaminate these data for various reasons, but especially because they come from unreliable or unpredictable instrumentation. Data (or signal) processing corrects measurement errors to improve fidelity. Here, we highlight decision-making applications and signal processing methods. To fully appreciate the state-of-the-art, we interviewed plant data experts and software developers in the pulp and paper industry to examine how they apply signal processing methods in the context of decision-making, including the value of process data, how these data are used, and the major barriers that prevent plants from using data. Process experts clean data thoroughly with basic approaches compared to the advanced techniques available in the recent literature. The interviews demonstrate that decisions in industry are primarily based on steady-state process operating data. Challenges and barriers that prevent the use of process data to their full potential relate to resource limitations (people, time, and money), an entrenched culture, and access to recent technology. In practice, experts consider, implicitly or explicitly, data that represent the process operating under steady-state conditions. A plant model that represents steady-state operations is easier to interpret, is presented in a form that is usable by plant operators, and in this way, better enables decision-making.

Keywords : signal processing, industrial process decision-making, process operating regimes, pulp mill, machine learning

Introduction

This review of industry practice for data processing and data usage in decision-making is one in a series of articles dedicated to experimental methods in chemical engineering.

Industrial decision-making includes for example process optimization, process control, fault detection and diagnosis, cost or energy analysis. This process decision-making is based upon data from sensors, laboratory data, weather information, cost information, and other kinds of data. Process industries collect billions of sensor data points daily, however these data are corrupted by several sources and types of error (Alighardashi, 2017). Erroneous data may lead to deterioration in plant operations. For example, erroneous data can degrade control system performance, mask process changes, lead to asset health deterioration and failure, inhibit process optimization, or cause the process to operate in an uneconomical or unsafe regime (Alighardashi, 2017). In contrast, rectified data can enable the rapid identification of equipment maintenance needs, process leaks, faulty equipment, troubleshooting of operating problems, and improvement of process control and decision-making processes generally. To leverage industrial data in support of decision-making, a series of steps must be followed to rectify or “clean” data (Bellec et al., 2007; Korbel, Milan et al., 2014). This article discusses the state-of-the-art in industrial data rectification and signal processing for process decision-making.

Signal processing was initially a domain of electrical engineering focussing on analyzing, modeling, treating data to increase quality, and smoothing, modifying, and extracting features, patterns, trends, and useful information from time-series data (Ahmed, 2015; Romagnoli & Sanchez, 2000). The 1930s and 1940s saw the advent of radar, spread-spectrum communication systems, and military devices, which were mostly developed during World War II (Farina, 2006). Later on, the 1950s saw the

development of analog signal processing, and the first digital computers appeared (O'Regan, 2016). In the 1960s, signal processing was driven by large industries such as communications, oil exploration, and space programs because they possessed the type of computers that could perform such tasks. At that time, the discipline of digital signal processing began to form with the advent of microelectronics and with digital signals becoming more common. By the 1980s, the number of signal processing applications was tremendous including for example medical imaging, communications, and digital signal processors continued to be developed (ETHW, 2017). In the 2000s and beyond, signal processing became an important branch of engineering, and the range of applications exploded in many fields. Today, signal processing employs a sophisticated set of tools to manipulate data for many applications, including for example image compression and processing, audio and video processing, industrial process sensor data, speech recognition, and hearing aids (Kumar & Kumar, 2018).

With the proliferation of inexpensive sensors and instrumentation, the number of process data points available in chemical plants is increasing. Process engineers need data processing methods to exploit these data. Collected data are an essential source of knowledge for plant-wide applications and troubleshooting decisions (Rojek & Kusiak, 2012). Raw data are of limited value until they have been “rectified” by outlier detection, noise reduction, time alignment, delay estimation, steady-state detection, and data reconciliation; data cleaning is a critical step in the knowledge discovery process and improves overall data quality for further analysis and model/simulation building (Xu et al., 2015).

This paper compares signal processing techniques that are used in practice versus the state-of-the-art reported in the literature. Are there signal processing techniques reported in the literature that should be more present in industry? Do the methods reported in the literature take into consideration adequately the concerns of industry, and address real industrial issues? Interviews were conducted with process specialists and leading software developers to explore how data are used for decision-making applications in industry, and how process signals are cleaned and processed for these applications. The pulp and paper industry was specifically targeted to avoid disparate answers from various industrial sectors having different maturity levels, addressing unique process challenges, having different resources, etc.

This paper addresses the following objectives : (1) To present a review of the approaches used in signal processing through a literature analysis and a bibliographic map; (2) To assess the state-of-the-art in pulp and paper industry practice related to data usage and signal processing for decision-making; (3) To review approaches proposed by software developers working in the pulp and paper industry in terms of data processing for decision-making; (4) Compare the techniques (a) proposed in the literature with (b) the current state-of-the-art by software developers and with (c) what is used by practitioners in mills.

Theoretical Background

In this section, a bibliographic map illustrating data (and signal) processing is presented, and the most cited data applications and data processing techniques are highlighted. Papers reporting how data were processed in some steady-state applications are highlighted.

Data and Signal Processing in Chemical Engineering

In the past 10 years, signal or data processing appeared in 106,808 articles indexed by Web of Science (WoS). Electrical Engineering and Electronics was the first WoS Category with 46,193 articles, in second place Telecommunications (6,864), then Mechanical Engineering (6,657), and Computer Science Information Systems (6,353). Chemical Engineering was 54th in the ranking, with only 446 articles.

The green cluster, with 24 elements, is more oriented towards data-driven analysis and techniques. This cluster includes *artificial intelligence (AI)*, *ANN*, *big data*, *clustering*, *data mining*, *fuzzy logic*, and *Industry 4.0*. These elements focus more on statistical relationships, ie black-box data-driven approaches.

Lastly, the blue cluster has 19 items. The keywords in this cluster are mostly about *signal processing*, and more specifically about frequency and spectral analysis. *Short-time Fourier transform*, and *wavelet transform* are examples of frequency analysis approaches. On the other hand, this cluster also includes keywords such as *algorithm*, *machine learning* and *pattern recognition*. Hence, this cluster could be considered as hybrid, containing keywords that consider both data-driven and fundamental approaches.

When considering all keywords in Figure 1 without considering the various clusters, it is possible to observe bubbles associated with data or signal processing techniques like *machine learning*, *artificial neural network*, *filters*, *transforms*, *artificial intelligence*, *Kalman filter*, *EMWA filter*, *short-time Fourier transform*, and *wavelet transform*. Regarding keywords associated with applications where data/signal processing approaches are required in chemical engineering, *anomaly detection*, *condition monitoring*, *decision-making*, *design*, *diagnosis*, *maintenance*, *model/modeling*, *simulation*, *fault detection and diagnosis*, (process) *monitoring*, *optimization*, *prediction*, *predictive maintenance*, and *process control* are mentioned.

Focussing on the last five years only, it can be seen that AI and other ML algorithms are increasingly present when the database is queried for articles about data and signal processing in the chemical engineering field. Not surprisingly, this indicates that the literature is increasingly emphasizing the use of machine learning, an important part of signal processing. More specifically, between 1992 and 2017 (25 years), there were 26 papers mentioning these keywords, but between 2018 and 2022 (4 years), there were 37 papers on this topic.

Signal Processing in Practical Steady-State Industrial Applications

Data processing in the process industries provides many benefits such as better plant knowledge, reduced downtime (throughput increase), decrease in maintenance costs, increase in product quality, minimized energy consumption, and improved energy efficiency (Desbiens et al., 2008). There are several areas where data processing is directly being applied in the process industries. A non-exhaustive list of applications that may use steady-state data includes optimization, cost analysis, condition monitoring, benchmarking, failure analysis (fault detection and diagnosis), energy analysis, and process simulation. The following paragraphs describe how data were processed as found in the literature.

Focussing on papers that present the steps undertaken to process data before using them in optimization applications, Reyes et al. (2015) considered reducing measurement noise, steady-state identification, and gross error detection as the initial steps of a real-time optimization implementation. Signal denoising was performed both with the short-time Fourier transform and the wavelet transform, but the latter was preferred. Furthermore, the authors mentioned the importance of choosing the mother wavelet properly, and that, in this case, the most accurate wavelet family is Daubechies. Steady-state detection was carried out using the wavelet method proposed by Jiang, Taiwen, Chen, He, et al. (2003) (Reyes et al., 2015). Gross error detection was done by minimizing the generalized maximum likelihood objective function with Hampel's redescending M-estimator. Gross error detection is performed on 24 steady-state intervals of 8000 data points each.

Process optimization requires the following steps : (1) data acquisition, (2) data treatment, (3) steady-state detection, (4) data reconciliation, (5) optimization, and (6) solution validation (Delou et al., 2021). In the data acquisition stage, the user specifies whether the algorithm is going to run off-line, i.e., independently of the actual state of the plant, or on-line, i.e., reading the values directly from the process database. In the second step, each input variable may undergo a linear transformation to scale or to perform unit change, and any unwanted or out-of-limit values are excluded. The first two steps are repeated until all possible error has been removed. Steady-state detection is mandatory before the reconciliation stage to reliably estimate the optimization model parameters. The authors applied two statistical tests for steady-state detection. In the fourth step, data reconciliation is formulated as a nonlinear programming problem. Then the optimization problem can be solved, and the solution can be validated.

Abu-el-zeet et al. (2001) proposed to perform data reconciliation ahead of optimization because it can considerably reduce the inaccuracy of process data due to measurement errors . They state that doing this improves process knowledge and control

system performance. In fact, results show a marked improvement in optimization algorithm performance when data reconciliation and bias estimation techniques are used to process the data before the optimization stage. For the authors, data reconciliation consists in reducing or eliminating as much as possible the effect of random measurement error, improving confidence in the calculation of unmeasured variables, and lastly identifying process losses and faulty measurements. The data reconciliation problem was solved using sequential quadratic programming, but there was no mention of how steady-state periods were identified and whether the reconciliation is performed at every point in time, or on averaged data over a time period deemed representative of the process operations. The solutions of the steady-state optimisation algorithm are used as setpoints in low level control loops, as shown in Figure 2.

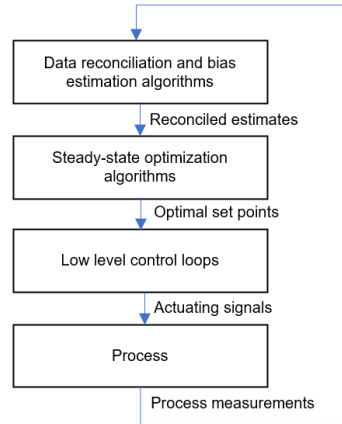


Figure 2 – Data processing algorithm for control and optimization (Abu-el-zeet et al., 2001)

Regarding cost analysis, Korbel, M. et Stuart (2011) stated that process knowledge must be integrated to assess the operating costs of each manufacturing regime. They did not use a data-based approach to detect operating regimes, but rather, employed good engineering judgment and process knowledge. By understanding the various operating regimes, decision-makers may avoid more costly ones, and use the analysis of operating regimes to better keep the process inside the most profitable operating regions for as much time as possible. Cost analysis was performed for steady-state periods as the authors assessed the running production costs of different operating regimes (they didn't consider the transient production cost).

The authors believed that to identify pseudo steady-state process operation, advanced data processing techniques must be used. Figure 3 summarizes the systematic methodology used, so that the raw data could be used for cost analysis. The reduction of random noise and elimination of abnormalities were carried out simultaneously. To obtain acceptable steady-state data, measured data must be processed to reduce signal noise before steady-state identification can be performed (Korbel, M. et al., 2007). The second step was data reconciliation, which further the process model representativeness. After steady-state identification of plant-wide operation, the data sets were ready to be used as input to operations-driven cost analysis.

With this approach, it was recognized that measured process data from sensors are a source of process knowledge, but are however inherently corrupted by different types of errors, and therefore data cleaning techniques are essential prior to data being used (Korbel, Milan, 2011).

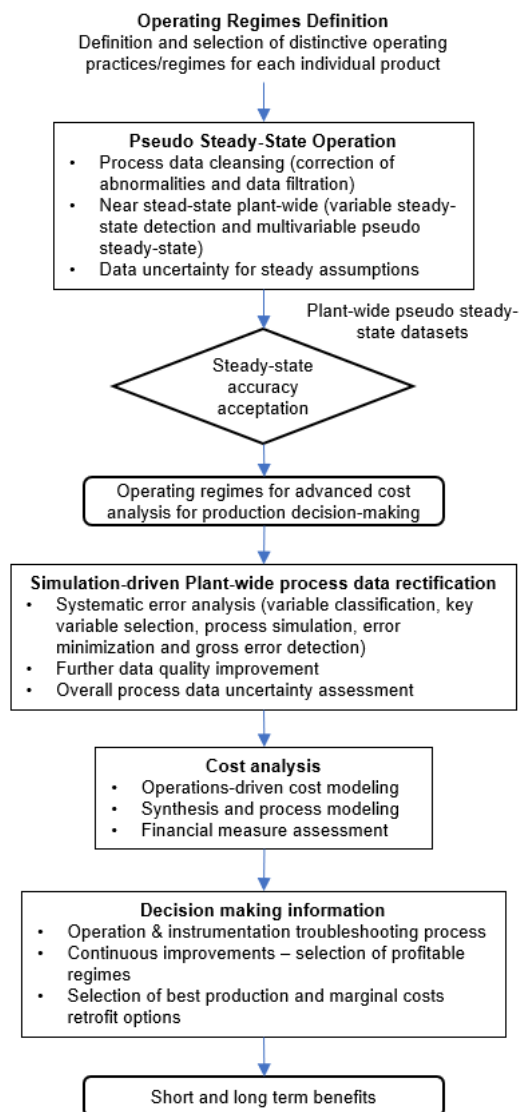


Figure 3 – Data processing algorithms for cost analysis (Korbel, Milan, 2011)

For fault detection and diagnostics in rotating equipment, the first step is to detect the steady-state segments of non-stationary time-series vibration signals. Su et al. (2006) used a Fourier-based signal processing method to account for these signals. Processing vibration signals before detecting faults ensures more accurate detection and fewer false alarms and missed faults. It was mentioned that to obtain high-performance fault indicators that are not influenced by time-varying equipment characteristics, the signal trend must be extracted from steady-state equipment vibration signals. Moreover, Rogers et al. (2021) stated that fault detection and diagnosis in industry typically require processing of transient data before applying fault detection and diagnosis methods. Kaneko et Funatsu (2017) confirmed that fault detection and identification accuracy are improved using smoothing methods (such as moving averages and Savitzky-Golay filtering), which also lead to enhanced state recognition performance. The authors found that measurement and sensor noise in operating data makes it difficult to recognize process states because it conceals important variations in process variables and decreases process state recognition performance significantly. Therefore, it is best to apply smoothing methods before steady-state and operating regime detection. Furthermore, noise makes it difficult to detect and isolate faults and diagnose faulty process variables. It was found that smoothing each process variable improved

fault detection and identification accuracy and also led to enhanced process state recognition performance. Hence, process data denoising is critical in state recognition, fault detection, isolation, and diagnosis (Kaneko & Funatsu, 2017).

Regarding plant monitoring, this application typically considers either steady-state or quasi steady-state operating data (Upadhyaya et al., 2014). Indeed, Aldrich et van Deventer (1993) mentioned that plant monitoring depends on reliable steady-state data sets. The acquisition of such data is fundamental for an interpretable understanding of the operating behavior of a plant, as well as the basis for modelling and optimization of the process. Measurement data are generally subject to random noise and/or gross errors that violate system process constraints (because of inadequate instrumentation, failure or miscalibration of measuring instruments). It is thus necessary to pre-process process data and account for systematic or gross errors in the data before performing plant monitoring. Aldrich et van Deventer (1993) accomplished this by using a back-propagation neural net. In another example, to build a simulation model of the TMP refiners in an integrated newsprint mill, Harrison et Stuart (2007) undertook data pre-processing in conjunction with multivariate analysis. This included outlier identification and removal with principal component analysis and filtering based on exponentially weighted moving average (EWMA) (Harrison & Stuart, 2007).

Often, in scientific papers discussing applications of process data such as optimization, cost analysis, monitoring, fault detection, and process simulation, data processing is more often than not summarized in two or three sentences. Hence, it can be difficult to elaborate on whether data processing was done appropriately. Nevertheless, for any decision-making application, data-driven analysis needs clean data to provide reliable detection and diagnostic results (Liu, Hancong et al., 2004). Depending on the application, the data processing approach may vary, but is a critical prerequisite needed to increase the realism and usefulness of further analysis such as multivariate and other analyses (Harrison & Stuart, 2011). The next section explores further the various techniques and approaches that can be used in the different steps of signal processing, for various applications.

Methodology to Grasp the Industrial Background

For this review paper, interviews with experts were conducted in collaboration with an AVEVA PI System provider (formerly OSIsoft) regarding the use and processing of process data for decision-making. The PI System is "a data management system built specifically to collect, store, and manage sensor and time-series operations data" (AVEVA, 2021). AVEVA PI Systems provide historical data, and provide the opportunity for users to visualize "actionable data". This data can be used for example for process troubleshooting, to assess how much resource and energy has been consumed, if the plant is operating productively, if the production assets are in good health, and to identify bottlenecks. PI software applies exception and compression to reduce noise in order to preserve data that are meaningful. Compression algorithms look for slopes in time series data and can interpolate data that don't vary significantly along that slope. For the exception algorithm, in order to store a value, the measurement must change by a specified amount compared to the previous measurements – typically about twice the sensitivity of the sensor.

Fundamental aspects were discussed with experts from industry both from process operations and from software development companies, and this information contrasted with that from the literature, to understand decision-making using process data. Figure 4 summarizes the different actors/sources and conceptually, their context versus time horizon.

Process operators focus on process issues that they are facing, and the troubleshooting required to resolve these. Process specialists rely upon software to help them in their day-to-day operations. Software serves process operators, providing a sophisticated ability to treat data and put it in a form suitable for decision-making.

On the other hand, software developers must consider the future to ensure their product is state-of-the-art; they must stay ahead of the technology curve, while respecting the pace of adoption by mills. Software development must be economically viable considering the market's ability to adopt new capabilities. Software developers will sometime strive to educate their customers on innovative data processing technology/techniques, but they also need to align with market trends, customer demands, and the competition. They must make the "right" decisions on development considering how to transfer tools into mills, and how to cope with the associated challenges.

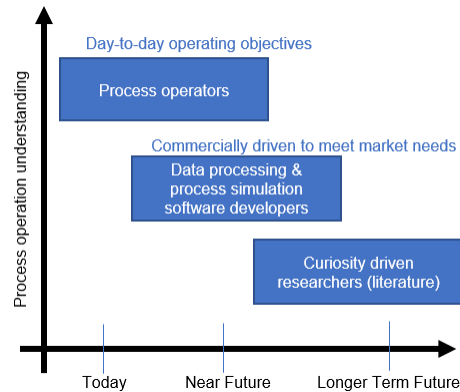


Figure 4 – Process operation understanding and time horizon for different actors/sources

The literature is typically further-reaching, and perhaps inspiring disruption to the status quo. Few techniques represented in the literature will be implemented in software development tools as proposed. Research influences software developers, who in turn interpret the practicality of the research result and explore this with industry.

The interviews aimed primarily to understand how organizations currently use their process data. The goal was to identify practical applications of process data and the applications of data processing tools used in practice. Several respondents worked for leading forest products companies, and for leading software development companies.

All the pulp and paper mill respondents used the AVEVA PI System. The individual targeted to complete the interview was either a person collecting data and trying to interpret it, a person responsible for process operation improvement based on data analysis, or someone who worked on the corporate side of a forest products company (Figure 5). The people interviewed in the plants were primarily chemical engineers. Their roles were enterprise-wide project expert, process specialist, and digital transformation project manager. The project managers were leading operational, digital transformation, and Industry 4.0 projects at the plant based on various strategies for digitizing the forest industry.

The pulp and paper industry respondents typically worked in advanced process control and were responsible for the plant's data and information technologies. In this capacity, they were designing and publishing production dashboards dealing with operation, quality, energy, cost, and utilities. Some oversaw mill information systems at the enterprise level. Enterprise-wide project experts oversaw technologies that might be applied at their sites. They drive best practices for monitoring mill performance, equipment health to avoid unplanned equipment failure, improving equipment reliability, and supporting mill staff when they have questions about PI and process data. These experts might lead advanced analytics teams to increase the speed of moving data to insights, developing talent in advance using analytic tools, and vetting new technology and tools for analytics.

People in engineering departments, in corporate development, and company presidents were respondents for software developers (Figure 5). The presidents were responsible for new feature development, deployment, training, support, marketing, and sales. Corporate development managers support sales through a consultative selling approach, and offer consulting services for advanced analytics across different process sectors. Their role is to expand their software offering either to consultants or plants. Lastly, senior research engineers are domain experts that offer data analysis support for users. A total of eight software developer companies were interviewed.

A critical analysis of the interview results was performed, and discussions were held with data processing consultants to validate the critical analysis.

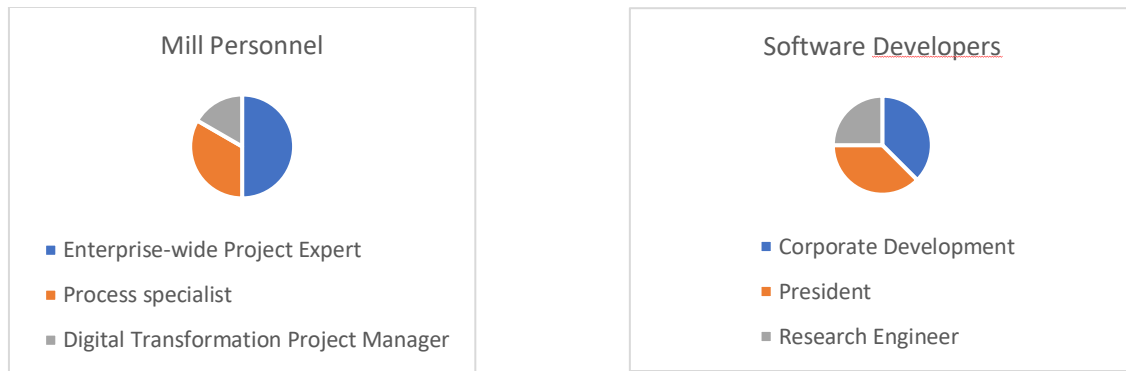


Figure 5 – Roles of the interview respondents

Industrial Perspective on Data Usage and Data Processing : Software Developers

Regarding the software developer group, the specific objectives of the interviews were (1) to identify the added value of process data, (2) to understand the barriers that prevent plants from using data more fully, (3) to explore the usage of machine learning in data processing (presented in the Future Perspectives section), and (4) to understand specific applications of their software and data processing procedures.

Added value of process data

Interviews with mill staff gave a limited and narrow perspective; their answers were solely based on their experience, and they could only testify to that. Their conclusions were case- and mill-specific, but they could narrow down into more detailed analytics. On the other hand, discussion with software developers gave a much broader perspective because they interact with many mill personnel over the years and generally have a better concept of the industry as a whole.

When asked about the most important value that they can obtain from process data, software developers mentioned that process knowledge is what matters the most. Until a data collection system is implemented, they cannot imagine what is going on. Process data provide tangible information that plants can make sense of. That is the big advantage. Mill workforces are making important discoveries about what is happening in the plant. It is also possible to analyze the relationship between process variables to understand the process better.

Furthermore, by having historical data that are representative of the process, it is possible to understand where the process is right now, and that could lead to better decisions for future operations. In fact, mill personnel cannot start predicting anything without knowing exactly where they currently are. During the discussions, it was pointed out that almost all the data that mills have been using for monitoring has been intended to enable managers to keep an eye on some part of the plant without going into the control room. Moreover, data make it possible to analyze process changes. Mill specialists can see the ways that different operators run the mill and the seasonal impacts as well.

In addition, using process data, post-mortem analysis can be undertaken. Process experts use operating time-series data to achieve good awareness of what is happening at the mill, what is happening now, and what happened in the past that led to an issue. Often a root-cause analysis of operating problems leads them to ask how they were operating and how they deviated from what they normally do to lead them to this problem. However, the most valuable achievement would be to figure out how to anticipate and prevent problems before they happen, to be able to use data to determine that a problem already experienced is happening again. Hence, there is tremendous value in identifying operating problems, diagnosing them to try to avoid them, and identifying new opportunities. Overall, this is a typical continuous improvement process, captured by the famous quote : “what gets measured gets managed”.

Plant performance can be assessed using key performance indicators such as process yield and product quality, and the availability of equipment can be assessed through a debottlenecking analysis. Conjointly, data open up the possibility of doing down-time management to understand the causes of stoppages or quality deviations to be able to correct them. With process data,

mill personnel can aim for operational excellence (efficiency) using optimization solvers, perform energy efficiency analysis, and assess resource consumption as well as compliance with environmental regulations. Likewise, process data may enable condition-based and predictive maintenance through analyses of equipment health. Lastly, process data can be integrated with other parts of the organization in an Enterprise Resource Planning like SAP to achieve better inventory management, to assess which product to order, for reporting, tracking, and more importantly for performing cost analysis. Suffice it to say that the greatest value is having actionable data from which real decisions can be made. Those are the things that mills expect from their data.

The value of data for software developers is data mining, which is about pulling data together to make useful decisions. The complete potential of decision-making based upon process data may only be obtained if it is linked to process knowledge. Almost all the aspects of data discussed above consider steady-state data. As well, it can be concluded that software developers mostly find value in analyzing historical data. Every added-value element raised up by software developers goes back to having actionable data from which real decisions can be made. More advanced applications such as digital twin, machine learning or artificial intelligence were not highlighted as part of this discussion.

Challenges and Barriers

In the eyes of software developers, the most important barrier is culture. Mill staff tend to be very stuck in the way they do things, they are running very lean, and they do not have enough people to do what they need to do. In particular, they do not have time to innovate, explore, and try to do things differently or to see things from a different perspective because their main focus is meeting production and quality goals. Hence, they do not have the culture that would enable them to look at data in a different way, to try new tools, and to try new methodologies. It is a substantial challenge for them to continue evolving in their use of process data. Mill personnel do what they are shown, but they are not going to do more than that. In connection with the culture comes the resistance to change. For some, changes are happening faster than they can absorb them.

According to the software developers, even if technology is present in a preponderant way in our lives, for many people, their first reflex is not to go towards technologies to use data for decision-making. People will avoid implementing technological solutions to problems. This is a question of knowledge, familiarity, and comfort with technology, fear of the unknown, and lack of trust. It is not what drives them. It is not intuitive for them to go towards technological solutions to solve problems. There are people for whom this is not a priority, not a source of motivation or inspiration.

Software developers mentioned that collecting everything makes it hard to know where to turn. Mill staff must ask themselves what the data will be used for and why they want to collect so many tags, and if they do it, they have to do it properly. Mills must avoid the black-box syndrome where they collect the maximum amount of data to be sure they can find the necessary information. They store data because they never know when they will need them, but if this was not done in a smart and systematic way, they will realize that they have stored poor-quality data. Mills record massive amounts of data, but not necessarily the critical tags. Therefore, another barrier mentioned by the software developers was that in the historian, the data are not organized correctly, they are not at the right resolution, and there are missing data, meaning that the data are available, but not usable per se. It is a question of data quality; quality is measured by having the least amount of error possible, the least number of missing data points, and the least amount of inconsistent data. The challenge is to transform these stored data, which were collected without knowing what they would be used for, into a usable data set. Most people go straight to analysis using raw data and do not do any data engineering (making data usable) beforehand to make sure that the data they put into their application algorithm have been conditioned for the application they want to do or the problem they want to solve.

Personnel at the mill don't necessarily know how to process data to make them usable. The problem of not having enough capacity at the mills in term of data processing is not unique to this industry. It happens to most, if not all process industries. There is definitely a gap between the people that are close to the process but do not have the statistical skills to understand what more can be done with data and the people that have those skills but that don't know much about the process. Another barrier is the fact that data processing is very labor-intensive, especially if the plants don't have the right tools. There is a lack of tools that are easy to use by people in industry, that would be software that are low code or no code for process engineers. These software need to be packaged up for the mills and be as automated as possible.

Other aspects that should be considered in order to use data effectively is time shifting the data to reflect the lag in the process given the fact that it takes 20 hours to go from the digester to the machine as well as the transient nature of mills. Understand the dynamics of the process could be a major challenge in regard to data cleaning.

Software developers mentioned that it is hard for mill people to know which tool to use for which application because they have no time to analyze that in detail. Mill specialists surely select the tool based on the application they are targeting, and the strength of the tool relative to that application, but they also consider which software was the first on the market, which will truly bring them where they need to go, which is the most user friendly, accessible, and easy to implement. Furthermore, technology companies that are strong on marketing might be more successful in selling their solutions as well as those how directly reach out to the users ultimately. The skepticism of the industry towards a solution itself – especially technologies that are not well understood and that are from small companies – is a barrier as well. It is not trivial to convince plants to deploy substantial efforts required to implement a new tool, despite the evident benefits. Generally speaking, they are questioning the pertinence of the solution, the technicalities, the security, the cost, the deployment and hosting, the customer satisfaction, the legal aspects, etc.

Software developers can help bringing out the significant potential of process data by providing the know-how and the expertise required to understand the data, how to use them, and how to process them. However, one problem that most software vendors have is that their clients are only leveraging 10% of their software capability. This is a daunting challenge. Mills must build, support, and sustain complex solutions comprised of multiple systems, acquired from multiple vendors, with diverging opinions and roadmaps, while interfacing with each other in a seamless way.

On the other hand, software developers must be aware of the mill's reality. For instance, the situation is not that mill staff are deliberately choosing to record data that do not have a specific application just yet; it is that staff do not know what the applications are. Mills will probably perform different applications with those data. They do not have the time to decide a priori what they are going to do, in which application data will be used, to justify the value of each measurement, and to risk missing something. Therefore, mills collect everything. It is a question of shared responsibility; whereas an IT department may be responsible for setting up the software and systems to collect and store data, it likely is not their role to use it. The engineer who needs access to data may not have been given the privileges and may not have the skills to correct the underlying software.

Another aspect of the mill's reality that should be considered by the software developers is the practical standpoint put forward in industry. They should discuss with the operators and try to understand the reality of being on the production floor.

Mills that have more resources, such as corporate resources that are not at the mill level to help with data analysis to create actionable data for future decision-making, are using state-of-the-art technologies and software. Therefore, the present situation perhaps has less to do with the cultural aspect of the industry, but rather more with a lack of resources.

Application and Processing of Industrial Data

This section describes some examples of software that are involved in data management leading to decision-making in plants to show the diversity of the companies that are involved in the pulp and paper sector. There are an enormous number of tools out there, only some of which are presented in Table 2. The table below states for each software, its objective, what kind of decisions are made using it, its ability to detect operating regime, or to consider the fact that there are several possible operating regimes, and lastly its targeted user. Softwares here fall into two categories : the more generic ones that propose data processing as the end purpose (there is no specific application, it is more like a mathematical toolbox) where users have extensive power to decide how they want to pre-process the data and the way they will process it; and secondly, software designed for a specific application.

The main objective and functionality of each these software environments are different. Therefore, the goal is to emphasize that they are all part of the solution to help bring value to industrial data. All software developers have their own perspectives on how data need to be processed and used in regard to process mode detection; the key here is to find balance between ease of use in an industrial context and the operating regime detection algorithms the tool contains.

Company	Software	Objective & Functionality of the Software	Nature of decision-making	Ability to detect Operating Regimes	Principal User
Aurel Systems Inc. (SME)	CADSIM Plus	Dynamic Process simulation – Energy, Heat and Mass Balance. Performing live data cleaning to create actionable data for decision-making. Identifying opportunities for process optimization and improvement. Providing soft sensors anywhere in the process.	Real-time optimization, test different scenarios, troubleshooting, debottlenecking, impact of capital expenditures projects, process design or existing plant retrofits/upgrades	Input manually. Seasons for energy balance, simulate different production rates.	Process Engineers
AVEVA (large corporation)	AVEVA PI System	Collects sensor data, stores and enhances them (context, analytics), and delivers them to users for applications in real time or analysis in PI Process Book or PI Vision.	Generates trend graphs for real time monitoring, root-cause analysis, and reporting	Cannot detect operating regimes but provides the ability to put a context on data (input the operating regime)	Operators Process Engineers
Braincube (SME)	Braincube	Turn data into competitive advantage with business intelligence apps : data visualization, condition monitoring, advanced analytics, process and quality control, predictive maintenance, modeling, prescription	Comply with process and quality standard, solve problems, optimize resources, improve operational performance, increase agility and efficiency, save time	No operating regime detection, but can identify the optimal conditions to operate at	Operators Process Engineers
Canmet ENERGY (Government)	EXPLORE	Data pre-processing to obtain usable data, perform data analytics, and create a monitoring interface to generate alarms	Fault detection and diagnosis, troubleshooting, prediction models, create soft sensors, analyze the links between variables and the difference between groups	Possible with PCA analysis, k-means clustering, validation with PLS	Operators Process Engineers
Industrial Algorithms Limited (SME)	IMPL	Off-line and on-line scheduling and planning optimization as well as real-time estimation such as data reconciliation for multi-unit continuous and batch industrial processes	Optimization, scheduling, planning, and estimation.	Uses optimization to schedule operating modes, but no operating regime detection	Process Engineers
Minitab (large corporation)	Minitab Statistical Software	Predict, visualize, analyze, and harness the power of data to solve business challenges	Discover trends, find and predict patterns, uncover hidden relationships between variables, create visualizations	Possible with Cluster analysis, Classification and Regression Trees	Data scientists
Optimistik (SME)	Oi Analytics	Make industrial data a daily performance lever and provide operational intelligence solution for industrial operations	Better understand the process behavior, find influent parameters, optimise operational conditions, improve productivity, compute new advanced indicators, investigate causes of non-performance	Operating regime detection is done by visual analytic	Operators Process Engineers
Pepite (SME)	DATAMAestro	Use no-code or low code advanced analytics for plant optimization (product yield, productivity, energetic efficiency, product quality), process performance monitoring and asset health.	Monitor, diagnose and optimize plant performance	Possible with clustering (e.g., K-Means), classification (e.g., Decision Tree, ANN), regression (e.g., PLS) and monitoring (e.g., ISHM) techniques	Operators Process Engineers
SAS (large corporation)	JMP	Data analysis tool used to leverage powerful statistical and analytical capabilities	Understand trends and patterns using statistical models, scenario (what-if) analysis, product performance, reliability analysis, address design vulnerabilities, enhance quality, uncover hidden relationships, understand cause and effect of variation	Regression, neural networks and decision trees	Data scientists
SEEQ Corporation (large corporation)	SEEQ	Self-serve advanced data analytics including statistics, linear regression, multivariate regression models, derivatives (for ex. to detect steady state), and linear interpolation	Root-cause analysis, reporting, monitoring, centerlining, signal relationship analysis, predict undesired events and prescribe better ways to operate.	Cannot detect operating regimes but provides the ability to put the data in a context that engineers want to use for analysis	Process Engineers

Table 2 – Examples of software objectives, data processing capacity, and operating regime detection.

Table 2 summarize the objectives and capability of some tools to detect operating regimes (how they could be used in that regard). It also presents the target audience of the software (operators, process engineers or data scientist) to assess the implementation complexity or ease of use of the algorithms implemented in these software in the context of operating regime detection. There are various algorithms discussed in the literature used to detect operating regimes such as unsupervised machine learning techniques like clustering (variational Bayesian Gaussian mixture (Saari & Odelius, 2018), Principal Component Analysis (Beaver et al., 2007; Srinivasan et al., 2004), K-mean (Bagherzade Ghazvini et al., 2021; Heikkinen et al., 2011; Liukkonen, Heikkinen, et al., 2011; Liukkonen, Hiltunen, et al., 2011)), neural network (Adebiyi & Corripio, 2002; Srinivasan et al., 2005) and classification (Lee, Y.-H. et al., 2006). Some of the software presented above provide more than one algorithm.

Performing Operating Regime Detection using a simulation software such as CAMSIM Plus is possible, but it is not trivial, it is done using logic and it must be configured by process expert. It requires physical model of the mill, process knowledge, maintenance, and it is time-consuming. Likewise, AVEVA PI System (AVEVA, 2022a) is not an ideal tool to detect operating regimes, there is no dedicated algorithm for this purpose. Similarly, Braincube (BRAINCUBE, 2022) doesn't have algorithm for operating regime detection. This Industrial Internet of Things Platform focuses on optimization data-driven techniques. For instance, Braincube may detect the optimal operating mode in which some KPI would be maximized or make recommendations in regard to an operating target. On the other hand, EXPLORE and DATAmaestro offers algorithms that can be use by process experts to detect process operation modes. Both of these software offer a maximum of functionality in terms of data processing for users that are not necessarily data scientist.

With IMPL, data are always pre-processed; then the process model is built, optimization is carried out, and lastly, data post-processing may occur. This tool offers a steady-state detection algorithm as well. The optimization objective function has three components : (1) maximize profitability (reduce feedstock cost, raise product prices) over some time horizon as the economic goal; (2) maximize performance, e.g., minimizing deviation from the target, maximizing efficiency; and (3) minimizing penalties like hitting infeasibility due to lack of feedstock. There are also constraints : (1) quantity (continuous) : flow, hold-ups; (2) logic/logistics (discrete) : start-up mode, shutdown, mode switching; and (3) quality (continuous) : consistency. This optimization algorithm gives rise to scheduling the proper operating modes, but those must be known a priori, there is no detection of operating regime. On the contrary, Minitab and JMP offer an array of data analysis tools, but these software are not very adapted to operators as they require a strong background in statistics. Using those software require an extensive background in statistics. Optimistik is adequate for industrial environment, but offers less advanced analytics functionality than Minitab and JMP and doesn't include an operating regime algorithm. However, like AVEVA PI System, DATAmaestro, Braincube, EXPLORE and Minitab, this tool proposes data visualization. Process experts may look at trends and visualize the process state in which they are at a given time period.

SEEQ is a tool for data connection, data management, and statistically based signal processing; it allows to prepare the data set for a predictive model. The cleaned data are saved and can be sent to a machine learning algorithm. In SEEQ, users can select the important data features they require. Depending on the analysis, the users may want to preserve the noise, but other analyses do not require it. With SEEQ, to put data into a context means defining the operating regime, to point out which grade the process is in – they also refer to this as capsule and condition and it is one of the first step experts must do. That data context, most of the time, is recorded in places other than the historical database or the historian. Mills probably have access to a manufacturing execution system (MES) that describes the various grades and SEEQ is able to read this information from the MES and use it to group the process data by grade. Seasonal data that come from weather stations can also provide context to analyze data based on weather.

In addition to Table 2, a non-exhaustive list of software and analytical tools used today in the pulp and paper industry for process monitoring and fault detection is presented here. A discussion with mill specialists led to the unanimous conclusion that "Excel is the Swiss Army knife of engineers everywhere". Almost all of any mill's process studies are done in Excel. It is their primary reporting and analysis tool that they consult on a regular basis. Any type of application usually starts with Excel; "it is the closest thing to a data preprocessor if you want to use your data for more advanced analytics". On the other hand, Power BI (Microsoft, 2022) is an extremely powerful data visualization and reporting tool. Some mills use it to create dashboards of their process with the goal of doing fault detection, monitoring, and troubleshooting and helping with decision-making. Mills compare strategic KPIs, calculate benchmarking metrics, and look for early failures.

In the same category as Excel and Power BI of tools that are generally applied in most mills, PI ProcessBook, PARCView, and PI Vision are widely used to display real-time data on screens, for process monitoring, detecting faults (putting rules or limits and setting alarms on mobile devices) and if there is something wrong, then troubleshooting. In addition, it is fairly common to see real-time process data integrating operational systems like SAP.

Ekhosoft is used for summary data and event triggering. In mills, some tags trigger an event in EKHO, and operators input the reason behind the event in this MES. For instance, when a detector notice that a breakage occurred on the paper machine, it triggers an event in EKHO and the operators will input the reason behind the breakage. Another example would be a production rate tag that drops below a certain threshold value that also triggers an event and the operators would enter the cause of this slowdown. EKHO can assess some statistics and build a report that gives in essence the pedigree of a production run (grade, quality parameters, production conditions, and parameters). This tool is used for quality tracking as well. Lastly, GE Proficy Plant Applications are used for environmental monitoring and analysis. This package is also used as an environmental compliance system and a product quality management system (end-product characteristics and properties). It combines laboratory test data with online sensor data. Although most mill personnel perform all their operations using the tools mentioned above, the more advanced users are moving towards machine-learning software such as Braincube and SIMCA. Braincube looks at all the AVEVA PI System data and performs a lag analysis; it accounts for tank volume, production rates, and rate changes and puts all the tags on the same timeline/stamp. Moreover, it is extremely good at cleaning data : filtering data, removing all outages, removing time when the production was less than a threshold value, removing every time it rains, etc. This data analytics tool takes out all the bad data values and does data conditioning and lagging. For some mills, Braincube is used for monitoring, troubleshooting, identifying the top 10 reasons something happened, filtering data, process control, and reporting. It may report the percent of compliance with a recipe, determine the variability of key output parameters, which are the top three parameters that impacted something, and assess the correlation between variables or parameters. Another use of Braincube is to set up statistical process control charts (setting up of control limits) for many of their process areas. This tool will identify when a tag is off-spec and assess why. Braincube is an adequate learning tool for young engineers who lack process knowledge. There are a lot of highly experienced mill workers that will retire in the upcoming years and this tool can help someone who does not have that experience. On the other hand, SIMCA (SARTORIUS, 2022) is used to develop a multivariate predictive statistical model based on historical data. If something is going wrong, the model can tell what is causing the issue, what mill personnel should do next, and what they could change to make sure that the process comes back where it should be. These prediction models are based on lag data; hence, in general, mills will use Braincube to set up and clean the data and then export them to SIMCA to build the model.

Another tool used by more advanced users is MATLAB. It offers generic machine learning and statistical as well as deep learning toolboxes. Some corporate resources are using MATLAB as a statistical tool to build custom applications that are specific to paper machine data and plot 3D graphs. Even if it is not specific to the pulp and paper industry, one mill has used MATLAB to do a lime kiln simulation. Lastly, Eigenvector offers a wide variety of MATLAB Toolbox (Eigenvector Research Incorporated, 2022).

Some software are dedicated to mass, energy and composition balance using first order conservation principles (linear equation model). This is the case for Sigmafine (Pimsoft Inc.). Its role consist in Reporting inventory and production numbers (production accounting), evaluating the measurement system performance using KPI, estimating performance of unit operations, production planning and scheduling, quality, and composition tracking, etc. Gross error detection (GED) is performed prior to running Sigmafine, and if process experts find patterns of errors, they can set up rules (based on variance, high and low limit value) to detect and correct them if a correction rule is available. Those gross error usually doesn't come from sensor data (if the sensors are working well), they come from manually collected data that is not coming from a sensor, such as lab data. GED is not the same as data reconciliation as gross error are obvious and can be fixed easily, which is not the case for data reconciliation.

Sigmafine has not been applied in the pulp and paper sector because the data reconciliation techniques they use do not apply to discrete data, and therefore wood chips being fed in batch to the digesters are a problem. Furthermore, measurement equipment for slurries and solid/liquid mixtures that have enough information redundancy to take advantage of these GED techniques are sparse. Instrumentation in general is lacking; pulp and paper is not a highly instrumented industry, and not an industry that is easy to instrument. In fact, to apply GED techniques, a certain level of information redundancy is needed; otherwise.

With Sigmafine, each tag has a tuning parameter (the design tolerance of the sensor given by the manufacturer). They are trying to implement an adaptive tuning parameter according to data quality to reduce variability, increase precision, and enhance model performance. Lastly, Sigmafine does not perform steady-state detection. The users must know the operations and how to segment the data to consider operating regimes and steady-state periods.

Industrial Perspective on Data Usage and Data Processing : Mills

The objectives of the interviews with mill personnel were to (1) assess how various organizations use real-time data (what are the current applications of real-time data) and identify some of the barriers that prevent using data to their full potential; (2) recognize how data are processed and cleaned for these applications; and (3) comprehend the major barriers that those organizations face in terms of data usage for decision-making.

Data acquisition

Every process generates a wide range of data types that are collected, gathered, and stored in the process database. These include process sensor data (pressure, flow, temperature, level), laboratory results, cost information/business data (accounting information, unit price, and stock market price), business plans, weather, orders, manufacturing schedules, equipment specifications, receipts, shipments, and transfers (pumping reports), scale data (weights), inventory data, production mode, product grade, and many others.

Mill personnel collect between 25,000 and 40,000 tags per site on average and store them for typically five to ten years. Every data point has its value, and provide historical perspective, but a few years from now, the mill might not be the same anymore. Therefore, the most recent data are more likely to be pertinent for decision-making.

Most mill specialists collect all possible and available data in case they need the data one day. When some tags are neglected, they could someday prove to be useful after all, and then they are not available when they are needed. If the tag is not created to store the information, the history is lost. They are aware that they may be storing some data for nothing, data that they will never look at, but they cannot predict the need. There are tags that they program thinking they will never need them, and eventually they do. The key message here is that they must not assume that the need is not there, but they should not decide the need for others. In conclusion, it is not possible to know what tags they might need eventually (to solve a problem or to understand something), and therefore they prefer to collect everything anyway in case someone needs it. Moreover, it is so inexpensive to get everything upfront and store it, and hence by default they get everything and assess what might be done with these data afterward.

One mill person mentioned that even when they take that approach, it is surprising to see how many times they fail to collect everything; sometimes they do miss sensors. They do not have a deterministic/systematic way of identifying additional sensors or calculations that have been created in the process control system and adding those to the list. Having a master data management team, with data owners responsible for all data and their maintenance was proven to be helpful.

Mill specialists are aware that they use a very small percentage of collected and available data for decision-making, KPI tracking, and troubleshooting. One respondent estimated that 10–15% of their collected data (tags) are really used to drive decisions, whereas another mentioned that 30% of their data were used in some form, but they did not have a formal way to confirm what is being consumed. Lastly, another mill estimated that, on a regular basis, probably only 20-30% of their data were used. However, data that are typically not needed for day-to-day operation might eventually be used for a process expansion study, or to debottleneck part of the plant. In conclusion, it is not easy to assess how many tags mills were using out of those that were collected.

The data sampling frequency varies depending on the type of measurement; if it is critical to the process, the sampling frequency is higher, and if the rate of change of the measurement is important (changing rapidly), the sampling frequency is higher again. Most real-time data are sampled in the 1–5 second time range, or even down to a quarter-second range in some cases. In some cases, the sampling frequency is set arbitrarily by the person configuring the AVEVA PI system tags.

When selecting the sampling frequency, mill personnel need to identify how each tag is going to be used; different applications require different sampling frequency (or the sampling frequency will depend on the application). For instance, for long-term

KPI, they do not need data every second. Generally, process control, fault detection, and diagnosis require understanding process dynamics and thus require high sampling frequency. In contrast, cost analysis, yield and quality monitoring, environmental and energy analysis, optimization, and debottlenecking are steady-state analyses and thus need lower sampling frequency.

Mills have working rules or rules of thumb that are far from the state of the art to select the sampling frequency of their tags. In general, mill staff prefer to have a high sampling frequency to avoid missing data, even if the compression and exception rules reject these values because they are similar to what was stored a few seconds ago. If they sampled at the higher frequency, they could miss or lose important events. Too low a sampling frequency could prohibit some analysis, whereas a high sampling frequency can always be dealt with by retaining only relevant information. Therefore, sampling frequency impacts the steps required to process the data.

Challenges and Barriers

Mill personnel mentioned that the biggest challenge that prevents the use of data to their full potential is a lack of resources, both human and temporal. There are few resources at the mill that have the time to work on data analytics. Mills do not have enough manpower in general, and hence if they are asked to learn a new tool or to use data analytics, even if there is substantial value in it, they just do not have time because they are too busy trying to keep production running as smoothly as possible. As well, it is time-consuming to retrieve large quantities of data. Data analytics is not a priority; they focus on day-to-day operation and deliver high-quality product.

In addition, there is a line of progression, or in other words, the people who work in the mill as part of the data analysis team move on to another position or get promoted into management. As well, with all the retirements, there are fewer and fewer people at the mill level, which exacerbates the situation. As a result, the expertise required to understand the data and what they mean is becoming an issue. The resources are lacking to retrain new personnel. To overcome this challenge, in the case of one company, the corporate side of the mill dedicated a special team to develop new process data applications as their full-time job. Furthermore, because there is no long-term stability in mills, there is an overall need for rapid onboarding.

Furthermore, mills have all too many tags that are overly compressed, meaning that significant events were missed, events that are not in their archive. Choosing the right compression rate for data is a challenge because they do not want to store too much information or to store noise.

All mills are using compression and exception to filter their tag data points. By using compression and exception, mills may maintain the quality of information without consuming too much space. There is a balance between information quality and performance. If data collection is in the hands of the IT department, data are probably compressed too much in the spirit of saving storage space, whereas in the hands of process engineers, data will not be compressed enough. Data must be compressed under specific criteria according to the application, for the reason that the data are being used. Mill management must make sure that the people who compress data and the people who use compressed data are one and the same.

One of the mills started by taking AVEVA's recommendation and cut their data compression rate in half to make the system compress data less and store twice as much. Then they set the right compression rates for each tag across the board by assessing them one at a time using their process knowledge. Recently, some mills brought a third-party tool called *Compression Insight* (Pattern Discovery Technologies, 2022) that analyze the incoming data and make recommendations on how to compress. It will tell them what the compression settings should be and whether their tags are over- or under-compressed; this is moving towards intelligent compression. As for the exception algorithm, unfortunately, in many places, the exception threshold is set to zero, and hence all changes are coming through the filter.

Data quality is indeed linked to data compression. Mills want to store all the data they can and not compress them at all. Data are generally not stored in the form in which they are going to be used. When mill personnel need them, they process the data according to the targeted application and store them under new variables so that others can also benefit from those data. In addition, the raw variables can be used for other applications, where they may be processed accordingly. On the downside, tag duplication for filtered data may create network bottlenecks, slowdowns, and unsatisfactory performance. Therefore, data could potentially be compressed and stored, and interpolation could then be used for some applications if required.

Even if storage memory in clouds environments is becoming cheaper as time goes by – pricing is often a function of tag count, therefore storage cost considerations are not totally out of the picture – data compression is still being done since it has to do with the performance of retrieval in online real-time (live) situation. Data compression have a huge performance impact on mills that are doing live data monitoring and processing and that have millions of data points that are stored on a continuous basis. They cannot wait 10 to 30 seconds for the data to come up if they need to have it in real-time and online. On the other hand, for offline analysis, waiting for the data to be extracted from the system is not an issue. Therefore, it depends on how the mills are consuming their data.

If process industries misuse data compression algorithms, they will end up with performance and data quality issues. They must select which variables (tags) should be compressed and how it should be done. Ideally, the compression algorithm should be specific to each variable, but mills tend to go more towards wide approaches since they have a huge amount of tags.

Moreover, the large amount of data makes it difficult to identify what to use. There is a huge volume of data, thousands of sensors in the mill, and it is a big job to look at all those data and to understand what they mean. There are so many opportunities that arise from the use of process data that it is tedious to decide exactly what to do and which data are significant. The whole process can become overwhelming. In this regard, it came to light that companies are beginning to use machine learning techniques with their data because their data are so voluminous, and this may change how much data they use and the value of data in the near term. This could help to overcome some of the barriers already described. Some mill specialists use Braincube, as mentioned earlier, to help them identify the most important variables to narrow down the critical sensors that need more investigation.

Furthermore, mill personnel mentioned another major challenge that prevents the use of data to their full potential : determining whether data are valid or not. Therefore, before using data in any algorithm, it is critical for mill specialists to see whether the data and their rates of change are within normal limits, i.e., valid. To address these challenges, some mill specialists are trying a new AI tool (APER10) that looks at all AVEVA PI System tags and gives a report on the condition of each. From this analysis, the tool can set the “quality” attribute of the AVEVA PI System tag : good, bad, etc. Consequently, if this tool changes the quality attribute of an AVEVA PI System tag, then any tool that uses that tag can benefit from knowing whether the incoming AVEVA PI System tag data are good or not. APER10 offers data integrity solutions powered by artificial intelligence that validate process data to improve its accuracy, enabling better and smarter decisions. This software provides a layer between the data source (e.g. AVEVA PI System) and the application. The tool reviews AVEVA PI System tag data in real time and checks their values using several data quality assessments : abrupt changes, out of range, correlation, not updating, noise, flat line, spikes, no data, etc. This could help greatly because we often hear how process data are not that great, but nobody has a concrete number representing how bad those data really are. For other mills, the tag condition analysis is done on Excel, but more and more new easy-to-use tools can provide better answers at low cost.

Therefore, having data quality issues automatically marked (but not necessarily fixed) and taken into consideration into a data infrastructure is the way to the future for them. It was further mentioned that data quality consider the following definitions that are based on ISO 8000-8 (ISO, 2015) :

- Syntactic quality : the degree to which data conforms to their specified syntax, i.e. when expecting a temperature within a given range, the sensor provides a valid reading.
- Semantic quality : the degree to which data corresponds with that which it represents, i.e. when a sensor measures 72°C, the actual temperature should also be 72 °C.
- Pragmatic quality : the degree to which data is appropriate and useful for a particular purpose, i.e. temperature sensor must be collecting data at the right place in the environment to serve a given purpose.

The first two are measured through a verification process; it complies with requirements and specifications. In contrast, the third one is measured through a validation process; it must meet the needs of the users.

The last barrier mentioned by mill specialists was the fact that process data are corrupted by different kind of errors and that they need to be treated. Following the data validation, if the data were confirmed to be invalid, that is a major roadblock for them, and they must therefore take the time to process these data according to the targeted application.

Applications of Industrial Data and their Processing

Without processing data beforehand, it is quite certain that engineers will not reach the right decision when using data in a specific application, and depending on the latter, data processing will be different, depending on the goals of the application. During the interviews, one mill specialist mentioned that without carefully planned data processing, important resources are being wasted.

Mill respondents were asked about the way they use process data and how they process them for each application. The targeted applications were benchmarking analysis, failure analysis (fault detection and fault diagnosis), condition monitoring, cost analysis, process optimization, energy analysis, and process simulation. When asked to rank applications by value, it was observed that mill personnel clearly see the value of failure analysis, which is coherent with the fact that they are mainly focussed on production. However, process simulation is not an application *per se*, but rather a tool; it brings tremendous value to the mill, especially if it includes cost analysis. The latter is the most important data application for plants; it is what keeps them in business. The following table presents the details of different applications and examples where process data are used and how they were processed.

Category	Application (Mill Example)	Data Processing	Consideration of Operating Regimes
Benchmark/ Failure analysis/ Condition monitoring	Identifying process changes from target and/or historical performance. Ensuring strategic process Key Performance Indicators are running within tolerance. Predicting future performance and/or failures to take proactive corrective actions.	Collecting process data, using the 90th percentile of the last 3 years to set the benchmark by grade, and once a day, comparing the actual averaged data to those benchmark values. Long-period averages are used to detect a change.	None
Cost analysis	Production and maintenance cost. Dashboards. Cost of chemicals according to the grade they produce. Compare the cost value with the budgeted value.	Manually done in Excel; remove outliers and upset conditions, then take 15-minute averages over 30 days (noise filtering). The average is calculated over good, cleaned data so that it may be deemed representative. The filters are bound checking, run flag (is the process running), is it a normal operation, conditional filters.	None
Energy analysis (efficiency)	Energy consumption and distribution. Yield in boilers. Track steam use over time in Power BI. Analysis around heat exchangers.	Manual download in Excel, filter out bad data, conditional filters. Use the filters in Asset Framework for bound checking, run flag (is the process running), is it a normal operation. Take 5 min average.	Identified operating regimes to make decisions for energy management (seasons).
Process optimization	Cost optimization during audits. Perform energy and mass balances for steam and electricity to optimize energy sources and to handle seasonal changes.	In Excel: apply filters, data sorting and cleaning, removing null values, errors in the data entry interface (timeout, bad input, error code, loss of connection, etc.). Done manually by an expert. The filters are bound checking, run flag (is the process running), is it a normal operation, conditional filters.	The cost of manufacturing the product is more challenging because there are different grades and specifications.
Process simulation	Mass and energy balance for decision-making. Monitor for differences between the model and the actual values and flag a failure. Provide virtual sensors.	Manually done in Excel; remove outliers and upset conditions, then take 15-minutes average over 30 days (noise filtering). Conditional filters in Asset Framework as well. The filters are bound checking, run flag (is the process running), is it a normal operation, conditional filters.	They consider the different grades and the different type of wood chips (hardwood or softwood), and they simulate both normal production rates and the minimum one.

Table 3 – Summary of how the industry uses their data for decision-making and how they are processed.

Benchmarking, failure analysis, or condition monitor include looking back over time (a long time frame), comparing operation to steady-state optimal performance values, making sure that equipment is running in its best operating range, and indicating meter drift using Excel, PI ProcessBook, and PARCview. These applications detect whether the averages over some time periods are shifted from where they have been historically, or in other words, whether they match the steady-state model. They use AVEVA PI System data to calculate daily key performance indicators and compared them to the one based on the best three consecutive months or on the 90th percentile over the last three years. Types of analysis that look across several years of data are typically done more in process audits rather than for continuous monitoring.

Furthermore, before the cost analysis, as was the case for process simulation, they perform a visual analysis, looking for what they consider a good run time using the *sort* feature and removing unwanted data points. The filters in Asset Framework are also used to create new AVEVA PI System tags with the filtered values. Hence, all subsequent analysis can transition from using the wrong values that came from the measurements to these filtered clean values where they are available. On the other hand, energy analysis is performed to make sure that heat exchangers are not fouling, or to assess whether they should be cleaned because they have started to foul up.

Before feeding data into their offline optimizer, mill personnel filter out unwanted data. They mainly use averaged data since they have little to no trust in real-time data and they are looking for the process data to be at the same frequency as the cost information which is not available in real-time. On the other hand, process simulations are decision support tools for process operation. Some mills simulate the whole mill in CADSIM Plus to assess operations, rejects, yield, and equipment performance. Most of the time, the simulation compares today's performance to a simulated performance to see whether something is going wrong. For process simulation, data averaging is done over good, cleaned data, so that the latter may be deemed representative. To do this, mill personnel look for a period of relative stability (relatively constant operation), for example, 24h of everything running stably with no outliers. Data are pre-processed through visual analysis, using the *sort* feature as well and removing bad data points. For the simulation, they pick a period (a rather long one), over which they do an average based on trends. Steady-state process simulations are also used for design and typically have at least two operating conditions : summer and winter. If the mill is doing a project in retrofit, the operating regime definition is critical to reduce the risks to the existing process and to obtain a better estimate of the return on investment.

In CADSIM Plus, noisy data are taken care of by dynamic data reconciliation technology and as for the outliers, it uses an average expected value, a boundary condition, and anything that goes beyond that will not be trusted. The further a measurement deviates from that typical value, the less importance that measurement has in terms of its weight; a weight can be added to data that is related to how much the measurement can be trusted and how it degrades over time. This is where the knowledge of the engineers becomes important, to set everything up correctly. They developed their own method based on first principles for data reconciliation. Steady-state detection is done manually, and operating regime detection is done using logic.

As an opening for further analysis, it is interesting that mills are mostly in a troubleshooting, problem-solving mode, and there is still very little mention by the mill respondent of prediction functionalities. Data are still perceived as the black box of an airplane; after a crash, they are used, but they are not used systematically to anticipate events or performance drifts. The value of data could be expanded to predictive and prescriptive functions. Anticipating that something not normal will happen and giving instructions to optimize operations could potentially have a huge economic impact on the process. To a large extent, the technology capabilities exist today, but have not been widely adopted across the pulp and paper industry. Advanced pattern recognition software such as AVEVA Predictive Analytics is aimed at addressing this specific scenario where multiple sensor data are compared over a period of steady-state normal operation and monitored for deviations, which can then be tied to specific anomalies and prescriptive corrective actions (AVEVA, 2022b).

Discussion

In this section, the techniques proposed in the literature and by software developers are compared with what mill specialists are doing with regard to 1) data cleaning, 2) steady-state detection, and 3) operating regime detection.

Data cleaning

The data processing characterization presented in Table 3 reveals that at the mill level, data cleaning consists mostly of manual tasks based on data visualization. This is quite different from the techniques presented in the literature and even from what the software developers are offering (Table 2). However, even if it is not state-of-the-art, the industry is doing data cleaning in all the applications described earlier.

Regarding optimization, the literature (Abu-el-zeet et al., 2001; Delou et al., 2021; Reyes et al., 2015) suggests reducing measurement noise, performing steady-state detection, and data reconciliation as data processing steps. The industry has applied bound checking and conditional filters and has performed data cleaning to remove null values and errors. Also, measurement noise has been addressed. It was noted that they consider normal operation for these analyses. Their steady-state detection is rather simplistic; normal operation does not necessarily mean steady state. Lastly, neither gross error detection nor data reconciliation was mentioned by mill specialists.

As for cost analysis, the literature (Korbel, Milan, 2011; Korbel, M. et al., 2007; Korbel, M. & Stuart, 2011) proposes to define and identify the operating regime as the first step. It is then suggested to perform process data cleaning, that is, correction of abnormalities and data filtering, followed by steady-state detection and lastly process data rectification for gross error detection. Mill specialists remove outliers and clean the data; they averaged the data and looked for a normal operation that they deemed representative. Therefore, the industry missed the first step proposed by the literature, which is operating regime detection, as well as data rectification, but their data processing step is consistent with the literature. Lastly, steady-state detection was performed more closely in accordance with the literature than was optimization because the mill specialists averaged good, cleaned, and representative data.

For fault detection and diagnostics, the literature (Kaneko & Funatsu, 2017; Rogers et al., 2021; Su et al., 2006) advises smoothing and steady-state detection as the first data processing step and operating regime detection as a second step. The literature on plant monitoring and data processing emphasizes the use of noiseless, gross error-free, steady-state operating data. The industry used averaged data. Therefore, noise was reduced, outliers were assimilated into the average value, and there was some idea of steady-state detection behind those long-term averages. There was no mention of gross error or of operating regime detection.

Lastly, the data processing step for process simulation requires noise and outlier removal. This step was done correctly by the mill specialists, who removed outliers and applied different kinds of filters to reduce noise and unwanted data.

Furthermore, Table 3 indicates that data processing was done according to the application, that is, for a specific purpose, using a different way to process the data according to the nature of the decision-making. For a given application, mills processed their data quite similarly from one user to another. Therefore, it was found that the decision does impact how data will be processed. For those five applications, five data processing algorithms were applied.

Steady-State Detection

Table 3 indicates that decision-making applications were implemented using some sort of steady-state analysis because they are all based on averaged data, but the last two applications made sure that the selected data were representative of good, steady-state operation. This means that mills process their data over periods of time by averaging them, but they do not look systematically at whether the data are stable. None of them are using mathematical methods.

Although averaging data is not technically the same as saying that the process is in steady state in the sense that if the process varies and data are averaged over a period, it brings the process back to steady state, this was not necessarily the case; all applications in Table 3 are steady-state analyses. Indeed, when mill personnel average their process data, they say that the process was in steady state. It is implicit that they are removing inappropriate data. They are doing steady-state analysis without using any algorithm. They know their process; they select a specific time interval that they know to be in steady state.

Given the above examples of the nature of the decisions mill specialists are making and the way they preprocess the data, it can be concluded that *de facto*, they see things in steady state, and they interpret steady-state data for process improvement.

Most mill specialists usually assume that the process is continuously running in steady state, and every decision they are making is under a steady-state paradigm. It is true that steady-state operations constitute the majority of operations.

Mill personnel mentioned that it is easier to understand the process in a steady-state paradigm for process troubleshooting. Steady-state data are critical in decision-making. They are easier to interpret than dynamic data, they are more usable, and they bring substantial value. The mill specialists cut out all the high frequencies in the signal because they are not interested in them. Hence, they see the process largely in steady state unless they are making a control- or fault detection-related decision.

Due to the highly dynamic character of pulp and paper processes and the numerous interactions between departments (mill sections), it is a real challenge to improve operating performance in real time. Therefore, most representations of pulp and paper processes usually assume steady-state conditions. These conditions are used in simulations, cost and energy analyses, optimization models and benchmarking, condition monitoring, and failure analysis. Therefore, today, and to comply with this assumption, data input into plant-wide applications are in the form of average values over a certain period, for example a 24-hour span, and hence no real-time insight is available.

During the interviews, the emphasis was on applications that are executed offline on a steady-state period in the process and that require data processing. The discussion with mill people highlighted the need to define a few terms, which may vary depending on the individual and according to their experience.

Online data analysis is performed while connected to the process. It uses data that are currently in the process, with little or no latency from their acquisition time. Data are measured as the process continues to operate. When asked what online data analysis means, mill respondents answered that online data analysis means the use of current data from the process at routine intervals. The models run all the time (every few minutes or every hour or every shift), and data flows to the model without human intervention. Online analysis means that input data are used on a continuous or nearly continuous basis with no human intervention, with results available immediately; it means that the process is analyzed in real time.

Hence, it seems that everyone agrees on the definition of online data analysis, but with a few subtle differences. The clarification of these definitions is critical because we need to make sure that mill personnel understand what we mean when we ask a question. What matters is not that they understand the formal definition, but to know what they do with their data in a practical way.

There seem to be a time component in their definitions, but the fact that the analysis is done online or offline has nothing to do with time. Indeed, an online data processing algorithm may either be continuous or batch. Continuous data processing would suggest that every time there a new data point is generated, it is incorporated into the algorithm and a new calculation is performed. On the other hand, batch data processing would also collect data every minute, for instance, but would only calculate something every hour (there is no refreshed value of each KPI each minute). Another element is the presence or absence of human intervention. This has to do with being automated or not, but it has, once again, nothing to do with online or offline processing.

For mill personnel, offline data analysis means that they capture data and run the model independently of the process with that set of data. Offline models for them are static, steady-state. They run these daily using summarized (aggregated) historical data. They stated that offline analysis means that input data must be reviewed and/or corrected by a human, and hence results are not available in the near term. With offline data analysis, the process is analyzed based on historical, *a posteriori* data sets.

Essentially, for mill personnel, offline data analysis is equivalent to steady state. This is either because *de facto*, mill personnel consider their processes to be at steady state, or because in an offline analysis, they perform an aggregation over time and choose a reference period. In reality, the process may be in steady state regardless of whether the data are online or offline.

Steady-state periods might be declared when operations are stable, meaning that there is little to no variation in process data. There are many distinct steady-state periods in an operating process. Most of those steady-state periods are not known *a priori*, which is why a steady-state detection algorithm is required. Between those periods are transient states. These periods can be classified into operating regimes, and they can be further analyzed to assess which period was optimal.

When questioned about steady-state analysis, they mentioned that steady state implies that the process has been stable for some time; it has not experienced a start-up or shutdown in the past X hours, where X varies from 15 minutes to 4 hours depending on the process. Steady state assumes that the process has achieved “normal operation”; when the process starts, it is necessary to wait until normal operating conditions have been reached. It also implies that the process has not experienced a significant production rate change or a significant grade change (which changes inputs and/or output targets) in the past X hours.

In many of their operations, mill personnel assume that the mill is in steady state almost all the time, or at least what they consider steady state. In other words, if they are not changing grades or the targets in the tanks, they are pretty much running at steady state. Moreover, they mentioned that their equipment runs for weeks at the time without an upset, and therefore they are in steady state most of the time. For some, steady state means that the current condition is varying less than a certain percent threshold. Typically, a process expert will define steady-state periods. They segregate the data for their analysis into normal operation. As an example, from the process being down to running, it takes one hour to reach steady state. Hence, they segment their analysis using those flags. Normal operation can be anything, like as soon as they start up a pump. They have a variable in the AVEVA PI System that classifies data as normal operation or not normal (meaning that they are starting up or shutting down equipment or dealing with a process upset) using rules that they wrote. An example of a rule would be that when they start a pump, if the equipment is running, the process must run for one hour at a flow rate greater than X gallons per min to be considered normal operation. Once that criterion has been met, then the “normal operation” flag should be set. As long as the process stays in that regime, they consider it normal operation. These rules refer to the major production lines and therefore apply to the whole unit of operation.

Lastly, for some mills, the process is always in steady state because they take a moving average over 15-minute periods. They do not really consider the dynamicity of process data. They analyze the data as if it were always a steady-state model, even if it is dynamic. For another process expert, apart from visual data appreciation (e.g., a trend graph), they do not generally use specific criteria to assess data stability, but for some variables, they consider the 2-sigma rule (twice the standard deviation). It is possible to conclude that each respondent had a personal definition of steady-state data and that it is a question of perception.

The last concept to define is real-time data, which are “real-time” with respect to the dynamics of the process under analysis. Real-time data means that the speed of data analysis approaches the real time at the process scale. If there is a process that only varies every hour and otherwise nothing happens, and a value is sampled every minute, the data are in real time because everything that is going to vary will be captured with this sample. If another process varies every hour, but a value is taken every 24 hours, these data are not in real time. “Real time” is always in relation to something. You need enough points to recognize all the events in the process to be in real time.

In conclusion, mill personnel all seem to understand what steady state is, they grasp the objective and criteria of steady-state analysis, and they do make the distinction between dynamic and steady-state analysis. However, the way steady-state periods are identified could be improved because the methods used by the mill specialist requires continual human attention, they are manual, and they are subject to human error in the recognition of steady state. Features that can compromise visual interpretation include noisy measurements, slow process changes, multiple dynamic trends, or change-of-shift timing. Furthermore, some of the mill specialists interviewed used visual examination based on process knowledge instead of steady-state detection assisted by an algorithm because, according to them, when they are not transitioning, the process is at steady state *de facto*. Therefore, there is no driver for them to do SSD. However, mill specialists do not consider or detect different operating regimes.

Operating Regime Detection

From the previous section, it is clear that mill personnel make most of their troubleshooting decisions based on steady-state periods. However, there are many steady-state regimes (operating regimes) in a process, there are shutdown and ramping-up periods, different grade runs and grade switching periods, seasons, and recipes.

As seen in Table 3, mill personnel did not explicitly consider operating regimes for decision-making except for the seasonal impact on energy analysis and the various grades in simulations, but they did recognize their value and potential. They do not recognize the particular operating regime where they are operating now and the fact that there are many different steady-state operating periods in their operations. In general, there is little to no acknowledgement of the fact that a process has different steady-state periods that are characterized by set-points, seasons, recipes, products, etc. In other words, for most of them, there is only one operating regime. This is consistent with the fact that among the software examples presented in Table 2, only one offered the opportunity to perform operating regime detection. Segmenting time-series data into different operating regimes may lead to many process decisions, and more analysis can be undertaken by considering them. The following section presents the benefits and the added value of operating regime detection in each case.

First of all, there should be a distinction regarding the operation of each equipment or process to identify whether it is functioning or stopped. A second level would be to identify the various operating modes of production (change in operating conditions), product grades (or different products), and feedstock changes (nature or quality of feedstock). Making different products requires numerous transitions and adds a huge amount of complexity to the process, especially regarding control, optimization, and monitoring. The transition between grades is an operating regime in itself. Much attention is given to these to assess how long they last and to optimize this transition, to get through it as fast as possible to minimize transition time (since they are losing money when they are transitioning grades or starting up and shutting down). Plants are looking at more carefully tracking and reporting what is happening during these transient periods, not only from a plant performance point of view, but also from an asset reliability point of view. There are a lot of failures that happen during transient states as well.

Furthermore, it is well known that different operators will operate the mill differently when producing the same product. That itself could result in different operating regimes. Another level could be the changing seasons. As a rule, mills rarely consider these in their analysis (except steam consumption) because they have no control over them, unlike product grade transitions. Hence, they are aware of the different grades that are produced, but not of seasonal impacts.

Operating regime detection opens up the possibility of assessing the impact of each on energy consumption, efficiency, performance, cost, yield, and product quality and to determine the optimal one. It is important that process simulations account for these different regimes. Likewise, operating regimes must be considered in design projects or capital expenditures. For energy analysis, it is essential to acknowledge seasonal impacts, which may be significant. Operating regimes may also be used to set upper and lower bounds in control systems depending on what grade the mill is producing. Scheduling and production planning, troubleshooting, and optimization are all about operating regimes as well.

Segmenting process data into operating regimes reduces variability, making process models more reliable and making it easier to detect process deviations and avoid unnecessary rejects. When the various operating regimes are not identified, the standard deviations are much wider, and hence when focussing on fault detection, equipment failure, monitoring, or troubleshooting, it is much easier to miss something. Hence, in general, defining operating regimes helps to understand what happened during different periods.

Furthermore, mill staff are really interested in knowing the cost of operating in different regimes and the carbon footprint based on the operating regime. If people knew the implications of each regime, this could guide them to make different decisions. They want to identify the best way to operate; this could be based on a cost analysis. Production costs are inevitably related to the operating regime, and mills need to know where they stand to assess these costs. In addition, mills want to know the cost of transitioning from one operating regime to another (the cost of the sequence of actions leading from one operating regime to another). They are especially interested in capturing the sequence of steps to take to reach that ideal operating regime transition, capture it, and make sure everyone follows it.

The closest thing that mills have had recently to operating regimes is what they call a *rate strategy*, according to which they are trying to operate the whole mill. Essentially, mills following this *rate strategy* have determined fast, medium, and slow process set-points for each area of the mill. Normally, mills are running fast, but if scheduled maintenance comes up or an unexpected area shutdown occurs, then the whole mill changes to either medium or slow. Staff in each area of the mill know all their associated set-points for each of the inputs based on the three production speeds.

If a mill does not use a *rate strategy*, then the production rate is always changing and never stabilizes because all the subprocesses are changing in response to each small production rate change that occurs, so that everything is moving all the time. A rate strategy aims to stabilize the process.

Most of the time, mills run fairly close to full speed, so they are approaching steady state, but when they are about to do maintenance, a typical mill will run flat out until they fill up inventory and then shut down, which is not ideal. It would be better to use a *rate strategy* so that they can run steadily for the next hours because they know how to run that rate.

In conclusion, mill personnel need to know the operating scenarios for all applications to make good (or better) decisions as they characterize process operations. It must be recognized that the plant has different operating regimes, and that people must work with them; process data must always be analyzed based on these regimes. An operating regime definition helps to better understand how the mill operates and increases the ability to interpret process data thoroughly and deeply. It provides additional information on process operation and offers an applicable way of defining the best mode for achieving a more efficient and cost-efficient process. Therefore, recognizing different operating regime opens the door for many new opportunities.

It is worth noting that in other process sectors, such as polymer processing, the type of data and the ability to relate data are quite different from the situation in the pulp and paper industry. To varying extends, they already do identify operating regimes and optimize these, especially for the objectives of grade sequencing in order to minimize product lost.

Future Perspectives

According to mill process specialists and software developers, of the time spent on data analysis, data preparation (collecting and cleaning) takes the greatest effort and has no apparent impact. For this and other reasons, it is sometimes not done rigorously. If there were systematic methodologies to make this first step easier and faster, it would help ensure that data would be better processed and interpreted. Machine learning techniques (ML) could potentially be used to automate and improve this time-consuming step of industrial process data analysis. Moreover when mill specialists were asked whether they believe that machine learning could improve data processing practices, they indicated that they believe there is tremendous potential value.

However while ML application is inevitable, survey respondents generally believed that ML will be difficult because pulp and paper mill process data are inherently noisy compared to data from other processes such as petrochemical plants, even with good maintenance and calibration. ML alone does not address data cleaning and processing needs rigorously because of the external factors that come from the process. They believed that the intervention of a process expert (subject matter expert) was critical to identify what good data look like versus bad data, and be implicated in “cleaning” process data. This is referred to as “supervised ML” and can be provided by some software such as AVEVA Insight Guided Analytics. However, it is tedious to instruct the machine about what normal data look like and what data make up a system that is correlated, so that ML can look for deviations. Some mill personnel hoped that the artificial intelligence could eventually accomplish this alone.

On the other hand, software developers mentioned that certain simple things can be done instead of implementing ML for data processing, for example by establishing simple data collection rules such as computing the variance of data points to check whether they are abnormal, and trying to understand why. Time-series analysis programs are available today that can be implemented to improve data quality in this manner, for example the AVEVA Insight Automated Analytics capability. However the value of simple, low-level anomaly detection at the individual tag level is insufficient for an increasing number of customers.

Software developers also mentioned that ML is complex to apply in that no flow sheet and no tangible model that reflects the process operations. All data processing steps are complex without a good understanding of the process (mass and energy balance, time constants, etc.). The data must be appropriately cleaned before the learning algorithm learn can provide consistently correct information.

Hence, it would be possible to use ML for data processing, but it would have to be assisted; it could not be used in a completely unsupervised and automated way. In a process, there are many elements that make it impossible to automate data processing completely. For example, someone must identify what variables to use to identify shutdowns. Algorithms can help, but human intervention is mandatory. Process experts can train an algorithm to do something useful for them.

Furthermore, processing data with machine learning presents many challenges. For instance, ML algorithms are generally dedicated to work on a steady-state time interval (in the same operating mode) that lasts for long periods (hours, days, weeks, months), not minutes. Therefore, if the process is characterized by short steady-state time periods with frequent fluctuations from one regime to another, e.g., different operating modes, it is more complex for the algorithm to detect trends.

Furthermore, processes are often complex and highly dynamic, and therefore it is more difficult for an ML algorithm to detect data problems by itself. Under these circumstances, in fact, often the ML algorithm may flag a multitude of outliers and anomalies, which become nuisances. To implement something interesting, a good understanding of how the process works is required. Hence, the association of which algorithm to run under which operating regime would be very pertinent. Throwing all the data into the algorithm and saying “find problems in the data” never works. People often think that ML will figure out everything by itself, but this is rarely the case in the process industry (when dealing with sensor data). In fact, sensors are always subject to failure; if they are not taken care of, they will get fouled up and need calibration, and data accuracy will degrade over time. The algorithm can find all sorts of things in the data, but not necessarily useful things. Moreover, the algorithm does not know the difference between normal and desirable data. How could an algorithm differentiate noise from real operation without the context and knowledge of what real operation looks like? This is a dangerous trap for any automated method. Therefore, process expert input and engineering knowledge are critical when considering the use of ML for data processing. Lastly, ML must not only detect errors in data, but it must also fix them; hence, a further challenge is to select or build a good machine learning-based data remediation algorithm. It is one thing to see a data problem, but another thing to solve it.

In conclusion, the expertise and process knowledge of the plant personnel need to be involved in the data processing, but there is no need to repeat the same task over and over again. This gives rise to another model called the hybrid (grey) model, where process knowledge (engineering) is combined with ML to have the best of both worlds. It was mentioned by a software developer that the best success cases of data processing is when an inductive ML model was used alongside an engineer deductive (first principles) model. It is the best compromise; the way to the future is when we can do these two things together.

ML has significant potential for improving data processing, and especially as processes are becoming larger, more integrated, and more complex. One could imagine an algorithm that would use an ML algorithm combined with a heuristics-based directive to perform data processing in an automated fashion. In this scenario, the automatic part would be done by the algorithm, and the technical process knowledge part would be incorporated by the heuristics.

Conclusions

Data processing methods are used to address many industrial decision-making contexts including facets of operational excellence such as process control, fault detection and diagnostics, optimization, monitoring, and cost reduction; asset management such as predictive maintenance; environment analysis such as emissions, rejects, operating safety, and compliance with regulatory requirements; and energy analysis (consumption, efficiency, intensity, etc.). A bibliographic map of data and signal processing papers over the last 10 years underlined the various data processing techniques and methods available, and the gamut of applications where they are used. A literature review of practical industrial signal processing applications emphasized that data processing is an essential prerequisite step, and depending on the application, the data processing methodology varies, generally classified in three categories : data cleaning, steady-state detection, and operating regime detection.

Interviews were conducted with software developers and process specialists in the pulp and paper sector to establish how data were processed for different decision-making objectives. The interviews with software developers highlighted that their software must have a data processing capability no matter what the objective or functionality; however, only a few provide

the opportunity to detect steady-state periods or operating regimes. At the same time, it was found that most decision-making applications for process improvement made by mill specialists assume a steady-state paradigm.

The data processing steps undertaken by mill process specialists were compared with what has been found in the literature for similar applications, and it was found that :

- Mill specialists are rigorous in their data cleaning approaches before decision-making, and the methods reported in the literature generally align with the methods used in industry.
- Steady state was considered in almost all plant applications, but near steady-state detection was performed in a rather rudimentary manner.

Proper decision-making in plants requires insightful process knowledge, which to some extent, can be illuminated through the identification of near steady state operating regimes. Understanding these operating regimes can thus potentially create value by helping with quantitative process interpretation, providing better actionable data, and enabling better decision-making. However, operating regime detection is not addressed systematically in the literature for applications such as process simulation, optimization, cost analysis, and many other purposes. Mill personnel performing “surrogate” steady-state detection by averaging data points, which is prone to significant error relative to the application/decision-making.

Although mill specialists are able to envisage the benefits of operating regimes’ detection, their consideration is currently under-exploited in industry and represents an important opportunity for data processing and process decision-making.

Acknowledgement

The authors gratefully acknowledge the financial support of the CRSNG. The authors further thank Philippe Mack from Pépite for fruitful discussions concerning data usage and processing, AVEVA for helping with the interviews, and all the mill personnel and software developers that took the time to discuss with us.

References

- Abu-el-zeet, Z. H., Becerra, V. M., & Roberts, P. D. (2001). Steady-State Optimisation Using Data Reconciliation and Bias Estimation. *IFAC Proceedings Volumes*, 34(8), 283-288. [https://doi.org/https://doi.org/10.1016/S1474-6670\(17\)40830-5](https://doi.org/https://doi.org/10.1016/S1474-6670(17)40830-5)
- Adebiyi, O. A., & Corripio, A. B. (2002). Operating regime decomposition using neural networks. *Chemical Engineering Communications*, 189(11), 1550-1568. <https://doi.org/10.1080/00986440214997>
- Ahmed, S. N. (2015). 8 - Signal processing. Dans S. N. Ahmed (édit.), *Physics and Engineering of Radiation Detection (Second Edition)* (p. 477-540). Elsevier.
- Aldrich, C., & van Deventer, J. S. J. (1993). Use of neural nets to detect systematic errors in process systems. *International Journal of Mineral Processing*, 39(3-4), 173-197.
- Alighardashi, H. (2017). *Simultaneous Gross Error Detection and Data Reconciliation Using Gaussian Mixture Distribution*, University of Alberta.
- AVEVA. (2022a). AVEVA PI System. <https://www.aveva.com/en/products/pi-system/>
- AVEVA. (2022b). AVEVA Predictive Analytics. <https://www.aveva.com/en/products/predictive-analytics/>
- Bagherzade Ghazvini, M., Sánchez-Marré, M., Bahilo, E., & Angulo, C. (2021). Operational Modes Detection in Industrial Gas Turbines Using an Ensemble of Clustering Methods. 21(23), 8047. <https://www.mdpi.com/1424-8220/21/23/8047>
- Beaver, S., Palazoglu, A., & Romagnoli, J. A. (2007). Cluster Analysis for Autocorrelated and Cyclic Chemical Process Data. *Industrial & Engineering Chemistry Research*, 46(11), 3610-3622. <https://doi.org/10.1021/ie060544v>
- Bellec, S., Jiang, T., Kerr, B., Diamond, M., & Stuart, P. (2007). On-line processing and steady-state reconciliation of pulp and paper mill process data. *108*(6), 36-40.

- BRAIN CUBE. (2022). *Solutions for manufacturers* <https://braincube.com/solutions/>
- Clarivate Analytics. (2022). *Web of Science™ Core Collection*. <http://apps.webofknowledge.com>
- Delou, P. A., Ribeiro, L. D., Paiva, C. R., Niederberger, J., Gomes, M. V. C., & Secchi, A. R. (2021). A Real-Time Optimization Strategy for Small-Scale Facilities and Implementation in a Gas Processing Unit. 9(7), 1179. <https://www.mdpi.com/2227-9717/9/7/1179>
- Desbiens, A., Nunez, E., Del Villar, R., Hodouin, D., & Poulin, E. (2008). Using process control to increase the energy efficiency of mineral and metal processing plants. *International Journal of Power and Energy Systems*, 28(2), 146-151.
- Eigenvector Research Incorporated. (2022). <https://eigenvector.com/>
- ETHW. (2017). *Digital Signal Processing*. https://ethw.org/Digital_Signal_Processing
- Farina, A. (2006). *Introduction to Radar Signal & Data Processing : The Opportunity*.
- Harrison, R. P., & Stuart, P. (2007). Spatial and temporal resolution in data-driven process modeling of an integrated newsprint mill. *Chemical Product and Process Modeling*, 2(1). <https://doi.org/10.2202/1934-2659.1041>
- Harrison, R. P., & Stuart, P. R. (2011). *Data pre-processing techniques for multivariate analysis to treat industrial operating data for retrofit design*.
- Heikkinen, M., Poutiainen, H., Liukkonen, M., Heikkinen, T., & Hiltunen, Y. (2011). Subtraction analysis based on self-organizing maps for an industrial wastewater treatment process. *Mathematics and Computers in Simulation*, 82(3), 450-459. <https://doi.org/https://doi.org/10.1016/j.matcom.2010.10.021>
- ISO. (2015). *Data quality — Part 8: Information and data quality: Concepts and measuring*. <https://www.iso.org/obp/ui/#iso:std:iso:8000:-8:ed-1:v1:en>
- Jiang, T., Chen, B., He, X., & Stuart, P. (2003). Application of steady-state detection method based on wavelet transform. *Computers & Chemical Engineering*, 27(4), 569-578. [https://doi.org/10.1016/s0098-1354\(02\)00235-1](https://doi.org/10.1016/s0098-1354(02)00235-1)
- Kaneko, H., & Funatsu, K. (2017). Improvement of process state recognition performance by noise reduction with smoothing methods. *Journal of Chemical Engineering of Japan*, 50(6Special Issue), 422-429. <https://doi.org/10.1252/jcej.16we325>
- Korbel, M. (2011). *Ph. D. Thesis : On-line steady-state data reconciliation for advanced cost analysis in the pulp and paper industry*, Polytechnique.
- Korbel, M., Bellec, S., Jiang, T., & Stuart, P. (2014). Steady state identification for on-line data reconciliation based on wavelet transform and filtering. *Computers & Chemical Engineering*, 63, 206-218. <https://doi.org/10.1016/j.compchemeng.2014.02.003>
- Korbel, M., Garrigues, L., & Stuart, P. R. (2007). *Process and Business Data Reconciliation in the Pulp & Paper Industry*. Innovations in Engineering, Pulping and Environmental, Jacksonville, FL.
- Korbel, M., & Stuart, P. (2011). Assessment and interpretation of advanced cost data for process improvement in an integrated newsprint mill. *International Journal of Production Economics*.
- Kumar, A., & Kumar, R. (2018). Role of Signal Processing, Modeling and Decision Making in the Diagnosis of Rolling Element Bearing Defect: A Review. *Journal of Nondestructive Evaluation*, 38(1), 5. <https://doi.org/10.1007/s10921-018-0543-8>
- Lee, Y.-H., Jin, H. D., & Han, C. (2006). On-Line Process State Classification for Adaptive Monitoring. *Industrial & Engineering Chemistry Research*, 45(9), 3095-3107. <https://doi.org/10.1021/ie048969+>
- Liu, H., Shah, S., & Jiang, W. (2004). On-line outlier detection and data cleaning. *Computers & Chemical Engineering*, 28(9), 1635-1647. <https://doi.org/https://doi.org/10.1016/j.compchemeng.2004.01.009>
- Liukkonen, M., Heikkinen, M., Hiltunen, T., Hälikkä, E., Kuivalainen, R., & Hiltunen, Y. (2011). Artificial neural networks for analysis of process states in fluidized bed combustion. *Energy*, 36(1), 339-347. <https://doi.org/https://doi.org/10.1016/j.energy.2010.10.033>
- Liukkonen, M., Hiltunen, T., Hälikkä, E., & Hiltunen, Y. (2011). Modeling of the fluidized bed combustion process and NOx emissions using self-organizing maps: An application to the diagnosis of process

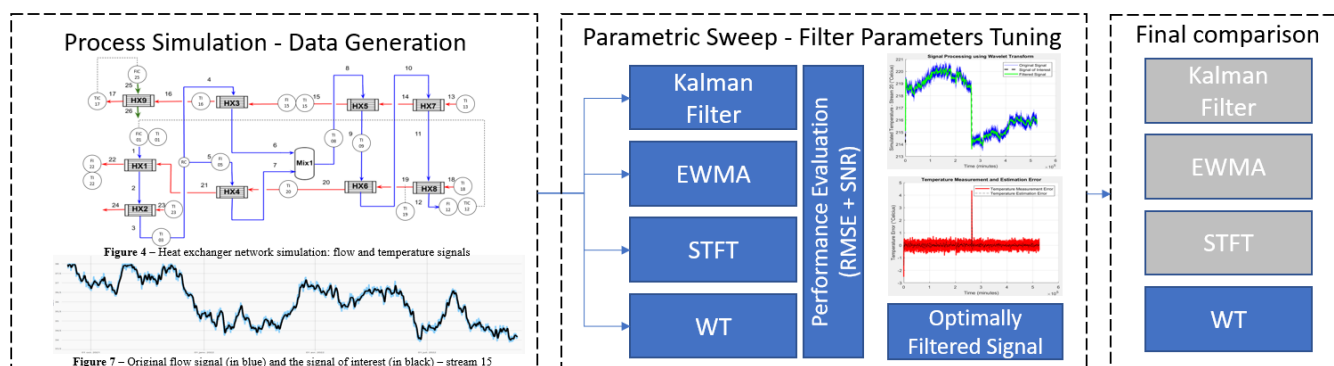
- states. *Environmental Modelling & Software*, 26(5), 605-614.
<https://doi.org/https://doi.org/10.1016/j.envsoft.2010.12.002>
- Microsoft. (2022). *Power BI*. <https://powerbi.microsoft.com/en>
- O'Regan, G. (2016). The First Digital Computers. Dans *Introduction to the History of Computing: A Computing History Primer* (p. 55-72). Springer International Publishing.
- Pattern Discovery Technologies. (2022). *Compression Insight*.
<https://www.patterndiscovery.com/products/compressioninsight/>
- Reyes, J. D., Rodríguez, A. L., & Riascos, C. A. M. (2015). Data Analysis and Modelling of a Fluid Catalytic Cracking Unit (FCCU) for an Implementation of Real Time Optimization. Dans K. V. Gernaey, J. K. Huusom, & R. Gani (édit.), *Computer Aided Chemical Engineering* (vol. 37, p. 611-616). Elsevier.
- Rogers, A., Guo, F., & Rasmussen, B. (2021). An Equilibrium Prediction Method for Control and Fault Detection of Energy Systems. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg*, 7(3).
<https://doi.org/10.1115/1.4050039>
- Rojek, G., & Kusiak, J. (2012// 2012). *Industrial Control System Based on Data Processing*. Artificial Intelligence and Soft Computing, Berlin, Heidelberg (p. 502-510).
- Romagnoli, J., & Sanchez, M. (2000). *Data Processing and Reconciliation for Chemical Process Operations* (1^e éd., vol. 2). Academic Press
- Saari, J., & Odelius, J. (2018). Detecting operation regimes using unsupervised clustering with infected group labelling to improve machine diagnostics and prognostics. *Operations Research Perspectives*, 5, 232-244. <https://doi.org/https://doi.org/10.1016/j.orp.2018.08.002>
- SARTORIUS. (2022). *SIMCA - Master your data*. <https://www.sartorius.com/en/products/process-analytical-technology/data-analytics-software/mvda-software/simca>
- Srinivasan, R., Wang, C., Ho, W. K., & Lim, K. W. (2004). Dynamic Principal Component Analysis Based Methodology for Clustering Process States in Agile Chemical Plants. *Industrial & Engineering Chemistry Research*, 43(9), 2123-2139. <https://doi.org/10.1021/ie034051r>
- Srinivasan, R., Wang, C., Ho, W. K., & Lim, K. W. (2005). Context-based recognition of process states using neural networks. *Chemical Engineering Science*, 60(4), 935-949.
<https://doi.org/https://doi.org/10.1016/j.ces.2004.09.061>
- Su, H., Kim, Y.-M., & Chong, K. T. (2006). *A novel method for high-performance fault detection of induction machine* (vol. 6041). SPIE.
- Upadhyaya, B. R., Mehta, C., & Bayram, D. (2014). Integration of Time Series Modeling and Wavelet Transform for Monitoring Nuclear Plant Sensors. *IEEE Transactions on Nuclear Science*, 61(5), 2628-2635. <https://doi.org/10.1109/TNS.2014.2341035>
- van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538. Article. <https://doi.org/10.1007/s11192-009-0146-3>
- Xu, S., Lu, B., Baldea, M., Edgar, T., Wojsznis, W., Blevins, T., & Nixon, M. (2015). Data cleaning in the process industries. *Reviews in Chemical Engineering*, 31, 453-490. <https://doi.org/10.1515/revce-2015-0022>

ANNEXE C ARTICLE #2 COMPARISON OF SIGNAL PROCESSING METHODS CONSIDERING THEIR OPTIMAL PARAMETERS USING SYNTHETIC SIGNALS IN A HEAT EXCHANGER NETWORK SIMULATION

Émilie Thibault, Francis Lebreux Désilets, Bruno Poulin, Moncef Chioua, Paul Stuart
Submitted to Computers and Chemical Engineering on 06/02/2023

Abstract

Plant sensor data contain errors that can hamper process analysis and decision-making. These data are not used to their full potential due to the complexity of their processing. This paper addresses these challenges by comparing popular data processing techniques based on their ability to process sensor data, all that while using optimal parameters. The latter are obtained for all approaches using an algorithm that performs a parametric sweep. The performance of Kalman filter, exponential weighted moving average filter, short-time Fourier transform, and wavelet transform to process synthetic flow and temperature signals from a heat exchanger network simulation is quantified given two criteria : signal-to-noise ratio (SNR) and root mean square error (RMSE). It is found that most of the time, wavelet transform showed the highest RMSE reduction and SNR improvement; the wavelet transform can effectively filter signals from distinct variables from a heat exchanger network simulation when optimal parameters are selected.



Keywords Signal processing, Performance analysis, EWMA, Kalman Filter, Short Time Fourier Transform, Wavelet transform

Introduction

In process industry, several variables are measured and recorded automatically by sensors for process control, process optimization and/or economic evaluation. Sensor data present generally several types and sources of error. This is unavoidable because data acquisition systems have inherent sources of error including sensor uncertainty, signal transmission losses, fluctuations in the electrical network, incorrect calibration of measuring instruments, and deficiencies in sensor quality/installation/maintenance. There are two main types of errors : random, and systematic (gross error or bias). Readers are referred to (Bagajewicz, M., 2000; Bhat & Saraf, 2004; Crowe, 1996; Mah & Tamhane, 1982; Martini et al., 2014; Narasimhan & Jordache, 2000; Sun et al., 2011) for a comprehensive survey of gross error detection. Some types of measurement error, especially bias, can lead to significant deterioration in a plant's performance, optimization and control which can cause unprofitable or unsafe operation. As part of this paper, methods for random error (noise) reduction are evaluated to obtain precise data for steady-state detection.

Industrial process data are generally corrupted with different forms of noise, and extracting the process state from the noisy data, i.e., reducing noise with maximum retention of the underlying process information, before performing any analysis is an essential preliminary step – often referred to as data cleaning, cleansing, filtering or denoising (Ganesan et al., 2004). Reducing random errors increases the quality and added-value of the data collected and enables detection of the steady state, which is required for many decision-making processes.

To extract relevant information from recorded signals, random measurement errors must be reduced by signal processing. Signal processing is used widely in industrial processes because the right data will enable quick identification of equipment maintenance needs, leakage in the process, or faulty equipment, troubleshooting to solve operational problems, and improvements in process control and decision-making. Therefore, rigorous signal processing is a critical step in performing these tasks properly.

A variety of data processing methods reducing unwanted high-frequency components, i.e., noise, from signals are available, such as finite-impulse response filters (moving average filter, polynomial filter) and infinite-impulse response filters (exponential filter, Kalman filter) (Narasimhan & Jordache, 2000). In addition to filters, time-frequency domain methods, e.g., the short-time Fourier transform, and the wavelet transform, can be used. All those approaches won't necessarily present the same efficiency, especially when it comes to filtering process sensor signals which presents a wide range of frequency components.

An important aspect to consider is the fact that the performance of these data processing techniques depends highly on the selected parameters. Therefore, when comparing the performance of said filtering techniques, it is critical to make sure that optimal parameters are being used. In addition to the filters performance, the way their parameters are selected must be accounted for, i.e., the ease of selecting the proper parameters for each approach, and especially considering the need to carry out large-scale manipulation of industrial data.

As part of this paper, four methods are considered : Kalman Filter (KF), Exponential Weighted Moving Average (EWMA) filter, short-time Fourier transform (STFT), and Wavelet transform (WT). These methods are selected because of their numerous applications as stated in the introduction, their industrial applicability, and their respective advantages. More particularly, the Kalman filter is theoretically an optimal estimator, can correct itself between the prediction and update phases, and provides high-quality-estimation with a minimum variance of the estimator error. The STFT is an excellent noise reduction tool for non-stationary signals, and the EWMA filter smooths data series readings and consumes significantly less memory and works faster than most techniques. Finally, the WT can carry out local analysis, is an efficient tool for short-time changes detection in signals, and can denoise particularly complex signals and then perfectly reconstruct them, giving many basis functions for decomposition, e.g. Haar, Morlet, Symlet, Daubechies.

The performance of the aforementioned signal processing methods has been compared in the literature to identify the most adequate technique for various applications (Dong et al., 2009; Dyason et al., 2017; Ibrahim et al., 2018; Perez & Barros, 2006; Pérez & Barros, 2012; Wang, Y. & Veluvolu, 2017; Yinfeng et al., 2008; Zhao et al., 2018). Among these applications, none specifically focuses on process data, i.e. data collected from process sensors (flow, temperature, level, etc.), but some are specific to the industry and engineering sector (Dyason et al., 2017; Ibrahim et al., 2018; Perez & Barros, 2006). Ibrahim et al. have compared the performance of Extended Kalman filter (EKF), continuous wavelet transform (CWT), and an iterative localized discrete Fourier transform in terms of their ability to track the magnitude of fault frequencies generated by a wind turbine (Ibrahim et al., 2018). In their work, various wind turbine operating conditions are tested, and the performance of all methods is evaluated using RMSE. Information about the parameters such as wavelet family and decomposition level could not be identified in the paper, and the noise reduction performance of each method is not under advisement (Ibrahim et al., 2018). Perez and Barros assessed the application of various advanced digital signal processing tools such as root mean square

magnitude, short time Fourier transform, Kalman filtering, and wavelet transform for detection and analysis of voltage events in power systems (Perez & Barros, 2006). These methods are studied for their potential in estimating the magnitude and duration of voltage events with various characteristics according to two criteria : maximum error in the estimation of magnitude, and maximum delay in detecting a voltage event. Although the voltage events had various magnitudes and durations and different start times, the characteristics of the various signal processing methods for detection and analysis are unchanged. The authors did not look at optimizing the parameters for specific voltage events. In addition, there is no explanation for the choice of signal processing parameters. Finally, they considered abnormal and sudden changes in signals from power systems but did not look at performance in reducing noise or other quality disturbances (Perez & Barros, 2006). The exponential weighted moving average (EWMA) filter is never compared with the wavelet transform, but the two methods are combined for noise reduction (Ge et al., 2015) and for fault detection (Harrou et al., 2018). However, the EWMA filter is compared with the Kalman filter (KF) to characterize the quality and usefulness of their outputs for real-time embedded system applications (Dyason et al., 2017). The relative computational efficiency and resource requirements of both filters are also studied. In this paper, the authors considered only normally distributed noise, but tested three signal-to-noise ratios. They used many performance metrics such as the sum of squared errors (SSE), the sum of differences squared (SoDS), %SSE, %SoDS, signal-to-noise ratio, and root mean squared.

Table 1 below present a summary of the papers that addressed the comparison of the different signal processing techniques aforementioned and considers as well how the tuning parameters are selected from a level of 1 to 3. A level 1 means that the selection is trivial or succinct (or that the parameters are just mentioned without any explanation), level 2 indicates that it is refined but not adapted to the process type, lastly level 3 suggests refined and adapted to the domain. Concomitantly, several papers discuss ways to select (manually or automatically) the values of the parameters and the ease of tuning them for Kalman Filter (Saho & Masugi, 2015; Yuen et al., 2007), EWMA Filter (Everett, 2011), Short-Time Fourier Transform and Wavelet Transform (Abbaszadeh & Alipour, 2017; Cao et al., 2022; El-Dahshan, 2011; Hasan et al., 2011; Katunin & Przysławka, 2015; Liu, W. et al., 2011; Shafri & Yusof, 2009; Zhou et al., 2009). However, studies that evaluate the ability or performance of these approaches to reduce noise in industrial process data while considering the selection of their optimal parameters are lacking in the literature. Table 1 showcases that papers comparing signal processing techniques don't always disclose the selected approach's parameters, and when they are indeed available, they are selected rather arbitrarily, or the selection method is not explicit.

Table 1 – Comparison of approaches considering parameters selection

Reference	Techniques compared				Parameters available	Level of sophistication regarding the parameter selection
	KF	EWMA	STFT	WT		
(Yinfeng et al., 2008)	X		X	X	Yes	1
(Zhao et al., 2018)	X		X	X	Yes	1
(Wang, Y. & Veluvolu, 2017)	X		X	X	Yes	1
(Dong et al., 2009)	X		X	X	Yes	1
(Pérez & Barros, 2012)	X		X	X	Yes	1
(Ibrahim et al., 2018)	X		X	X	No	NA
(Dyason et al., 2017)	X	X			Yes	2
(Perez & Barros, 2006)	X		X	X	Yes	1
This paper	X	X	X	X	Yes	3

Hence, the objective of this paper is to assess the optimal parameters of each signal processing method, to compare which approach shows the best performance towards synthetic process sensor data and to elaborate on how their parameters are generally established. In this regard, Section 2 of this paper presents the different parameters of various signal processing techniques and section 3 addresses the pros and cons of each. Section 4 elaborates on the methodology for comparing those techniques based on their performance in removing errors in process data considering their optimal parameters. Then, section 5 reports the comparison of the signal processing methods by firstly constructing and characterizing synthetic noisy signals, assessing their denoising performance and discussing their optimal parameters establishment. Section 6 gives a conclusion of the present work.

Signal Processing Methods : Parameters

The method-specific parameters of KF, EWMA, STFT and WT are described below. They are the input parameters, the information required to use the various signal processing techniques.

Kalman filter

The Kalman filter (KF) is a recursive linear estimation algorithm with the goal, given a sequence of noisy observations, of estimating the state of a system such that the estimation error is minimized. The KF is a very powerful tool for estimating the states of a linear system in the presence of noise. It minimizes the variance of the estimation error (the mean square error of the estimated states). This filter consists of a prediction phase and an update phase. Information about KF can be found in several references (Dash & Chilukuri, 2004; Dyason et al., 2017; Ibrahim et al., 2018; Kim & Bang, 2018; Labbe Jr, 2020; Narayan et al., 2013; Park, S. et al., 2019). A brief description of the KF parameters is provided here.

The system is described by the following state-space representation :

$$x_t = F_t x_{t-1} + B_t u_t + w_t \quad (1)$$

$$y_t = H_t x_t + v_t \quad (2)$$

where F_t is the state transition matrix, x_{t-1} is the previous state, B_t is the input matrix, u_t is the control vector, w_t is the process noise vector, Q_t is the process or system noise (co)variance matrix, y_t is an observation vector, H_t is the observation matrix, x_t is the state, v_t is an observation noise vector, and R_t is the measurement noise (co)variance matrix.

Prediction phase of the filter : In this first phase, a prediction (estimation) of the state $\hat{x}_{t|t-1}$ and the state estimate covariance matrix $P_{t|t-1}$ are calculated at time t as follows :

$$\hat{x}_{t|t-1} = F_t \hat{x}_{t-1|t-1} + B_t u_t \quad (3)$$

$$P_{t|t-1} = F_t P_{t-1|t-1} F_t^T + Q_t \quad (4)$$

The covariance matrix must be initialized at the beginning of the process.

Update phase of the filter : In the second phase, the measurement residual vector $\tilde{z}_t$ and the residual variance matrix S_t are calculated as follows :

$$\tilde{z}_t = y_t - H_t \hat{x}_{t|t-1} \quad (5)$$

$$S_t = H_t P_{t|t-1} H_t^T + R_t \quad (6)$$

These give the updated state estimation vector

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t \tilde{z}_t \quad (7)$$

and the updated estimate variance matrix

$$P_{t|t} = (I - K_t H_t) P_{t|t-1} \quad (8)$$

where K_t is the optimal Kalman gain that minimizes the mean square error $E[(x_t - \hat{x}_{t|t})^2]$:

$$K_t = P_{t|t-1} H_t^T S_t^{-1} \quad (9)$$

The Kalman filter parameters represent the distribution of noise in the system. The first one is the process noise covariance matrix. Process noise is associated with noise in the states. This value indicates, to some extent, the nonlinearity of the state-space model, and the larger it is, the more nonlinearity the model shows (Dong et al., 2009). Typically, process noise causes signal drift, and the result is related to the system dynamics. The covariance matrix Q expresses how much variance and covariance exist in the process. The matrix diagonal contains the variance of each state variable, and the off-diagonal elements contain the covariances between various state variables (Labbe Jr, 2020). If the state parameters are independent, then the matrix has only a diagonal. Setting the process noise to 0 tells the KF that the process model is perfect (Labbe Jr, 2020). The second parameter is the measurement noise covariance matrix. Measurement noise comes from a sensor and represents the noise characteristic of the measurement device. The measurement noise matrix R models the noise from sensors as a covariance matrix (Labbe Jr, 2020), which varies accordingly to sensor sensitivity. The covariance matrix contains the variance of each measurement and is calculated from the sensor accuracy. This is obtained through the standard deviation of the measured values using the signal of interest.

EWMA Filter

The exponentially weighted moving average filter (Evans, NA; Nounou & Bakshi, 2000; Perry, 2010; Smith, T. H. & Boning, 1997), a popular infinite impulse response (IIR) filter, is used to smooth time-series data by exponentially averaging incoming reading with all previous measurements. This technique is often employed to obtain an average of samples at the time they are “arriving” to the control system, i.e., it is used on-line. The EWMA is used for its noise filtering capabilities, it cancels the effects of noise at the output. The noise superimposed to the input is assumed Gaussian, with zero mean value and a known variance. This low pass filter reduces high frequency components. The EWMA filter makes it possible to specify the weight of the last sample acquired versus the previous filtered value using the alpha parameter. The average is “moving” in the way that it is computed each time a new sample is obtained. This method is widely used in chemical plants, consumes significantly less memory than others, and works faster. EWMA is mathematically defined as follows :

$$S_t = \alpha x(t) + (1 - \alpha) S_{t-1} \quad (10)$$

where the α coefficient is the smoothing factor and represents the weight, it defines the cut-off frequency above which features are eliminated; $x(t)$ is the current input value; and $S(t)$ is the filtered output at time step t . This recursive formula saves memory and makes fast calculation.

The smoothing factor, α , ranges from 0 to 1; the higher the value, the greater is the impact of the last reading. Hence, less smoothing is done, and more weight is put on this last entry. On the other hand, a small value of α means more weight on the average, meaning that more smoothing is done and the last entry has little impact; the filter is less responsive to new measurements (Evans, NA). Hence, a value of one results in no filtering (the output is only the current observation), and a value of zero does not take new measurements into account. The value of the smoothing parameter is typically correlated with measurement precision or accuracy.

Short-Time Fourier Transform

The Fourier transform (FT) gives the characteristics of a signal in the frequency domain. However, time information is lost because the transformed signal has a frequency on the x-axis and an amplitude on the y-axis. This would not be an issue if the signal under evaluation are stationary (constant frequency content). The short-time Fourier transform (STFT) is used to process time-variant frequency signals. STFT carries out the FT on a short period, accounting for the time component. This makes STFT adequate for non-stationary signals. Information about the short-time Fourier transform can be found in several references (Kehtarnavaz, 2008; Manhertz et al., 2016; Rahimi et al., 2007; Selesnick, 2009; Smith III, 2011; Zabidi et al., 2012).

The STFT is defined as

$$X_L[k] = \sum_{n=0}^{N-1} x[n] w[n - LH] e^{-j2\pi nk/N} \text{ for } L = 0, 1, \dots \quad (11)$$

where $x[n]$ is the time-discrete signal, $w[n]$ is the analysis window, n is the time-domain index, L is the frame number, and H is the hop (step) size, or the time separation between adjacent frames.

STFT is an effective signal processing technique that can decompose a time-domain signal into its time-frequency components. It divides a whole signal into shorter data frames of equal length by multiplying it by an analysis window function. Consequently, the frequency content is stored in time windows, where each of them is then analyzed using the fast Fourier transform. This procedure is called the overlap-add method (Bahoura, 2019). STFT reveals the frequency content of the signal at each time point (Zabidi et al., 2012).

STFT is performing iterations over L , hence skipping through time with L , and jumping to the adjacent window, i.e., going to the next time interval by hop size. The STFT is usually computed using overlapping windows, where the adjacent segments share some overlapped data instead of just being end-to-end. The purpose of using overlapping frames is to avoid spectral leakage, achieve better time locality, and obtain smoother results (Smith III, 2011). The output of the STFT is a sequence of spectra having the same size.

The first parameter used in STFT is the sampling rate; it corresponds to the number of data points per second. A higher sampling rate results in more samples and therefore more flexibility in choosing the window size, the FFT size, and the hop size. In the case of this paper, the discrete synthetic signals (Section 4.1) are sampled every minute, and consequently the sampling rate is a fixed value.

The second parameter is the Fast Fourier transform (FFT) size (or length); it is displayed as a number of bins in the analysis window, rather than as an absolute value. Hence, it defines the number of equal bins used to divide the window, which is typically a power of two. The number of bins is proportional to the number of samples in the window. It expresses the number of frequency bands into which the analysis window will be cut (Diatkine, 2011). A bin is a spectrum sample that defines the frequency resolution of the window (Diatkine, 2011). The FFT size is independent of window length (size), but it must be greater than or equal to the latter.

Third, the window function (Diatkine, 2011; Mitra, 2000; Smith, J. O., 2011) is a real function that is symmetric around the origin. Window functions are defined and distinguished by their main lobe (peak at the center) width and their highest side lobe (small lobe next to the main lobe), i.e., peak side-lobe, amplitude relative to the main lobe. The many windows available include Blackman, Chebyshev, Gaussian, Hamming, Hann, Rectangular, Triangular, and others. For each window function, there is a compromise with respect to these two parameters. The trade-off is that the wider the main lobe, the lower must be the peak side-lobe amplitude, and vice-versa. The

compromise is between spectral resolution and spectral leakage. A window with a smaller side lobe yields less spectral leakage, but has increased main-lobe width, which limits spectral resolution. Increasing the window length for a given accepted leakage level will increase spectral resolution, but decrease the time domain resolution. To minimize the impact of the latter side effect, adjacent windows are overlapped. The window length is an important parameter that specifies the frame length of the window analysis function. Its value must be greater than one, and it represents the number of samples.

The width of the main lobe is primarily determined by the window length, i.e., increasing its value will decrease the main-lobe width. The window length (Diatkine, 2011; Mitra, 2000; Smith, J. O., 2011) determines the resolution in both the time and frequency domains, i.e., the spectro-temporal resolution of the signal; a good time resolution requires a short window, whereas a good frequency resolution requires a long window. It is not possible to have at the same time both high time resolution and high frequency. This phenomenon is known as the Gabor limit, or the Heisenberg-Gabor limit, or Heisenberg's Uncertainty principle. In the same way that there is a trade-off in window choice, there is also one in window size.

Window functions can be described as having high resolution, but low dynamic range, or as having high dynamic range, but low resolution. The latter can distinguish features with different frequencies and amplitudes, whereas the former can distinguish components with similar amplitudes and frequencies. High-dynamic-range windows have a wide main lobe and very low amplitude peak side lobes and are used in wideband applications. The rectangular window – the simplest function – shows the narrowest main-lobe width and the highest-peak side lobe, resulting in what is described as high resolution. The Gaussian, Blackman-Harris, and Kaiser-Bessel windows have a relatively large dynamic range. On the dynamic range spectrum, between these window functions, there are moderate windows such as the Hamming and Hann windows.

Window functions are used in signal analysis, i.e., as an analysis window, as well as in signal reconstruction, i.e., as a synthesis window. The two do not have to be the same, but in this paper, they are.

Finally, the hop size (Diatkine, 2011; Mitra, 2000; Smith, J. O., 2011) corresponds to the number of samples between each time frame, i.e., the window step. The window is moved forward in time by the step size of H samples. In STFT, the windows are not juxtaposed successively, but overlap to improve the precision of the analysis. They are overlapped by a certain overlap length. Overlapping affects the calculation precision, improves the temporal resolution, but may increase the calculation time. Hence, the hop size is the difference between the window length and the overlap length of two consecutive frames (segments), as shown in Figure 1. Typical values of the overlap length are one-half, two-thirds, or three-quarters of window length.



Figure 1 – Hop size, inspired by MathWorks (2023b)

Wavelet Transform

The wavelet transform provides the characteristics of non-stationary signals in the time-frequency domain. The WT is defined as the projection of a signal $x(t)$ onto the scaled wavelet, i.e., the convolution of $x(t)$ and $\psi_{\tau,a}(t)$. The wavelet is an oscillatory window function of finite length, i.e., compactly supported, and defined by translation and dilation :

$$\psi_{\tau,a}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-\tau}{a}\right)$$

The continuous wavelet transform is expressed as follows :

$$W(\tau, a) = \langle x(t), \psi_{\tau,a}(t) \rangle = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t) \psi\left(\frac{t-\tau}{a}\right) dt$$

where τ is the shift parameter used to control the translation, a is the scale parameter, $\psi(t)$ is the mother wavelet function, and $*$ indicates complex conjugation. When the values of τ and a are continuous, the WT is called continuous, and when they are discrete, it is called discrete. DWT suppresses the redundancy of the continuous WT while conserving the reversibility property.

A wavelet transform decomposes a signal into several levels by using a set of filters, i.e., low-pass and high-pass filters, to produce the approximate and detailed coefficients respectively. These filters enable the signal to be represented by a sum of approximations (scaling functions $\Phi_{j,i}$) that form the smoothed signal and a sum of details (noise) represented by wavelet functions $\psi_{j,k}$ at different levels. The levels are used to display the characteristics of the signal over time and at various frequencies (Bakshi, Bhavik R., 1999) :

$$\begin{aligned} x(t) = x_0 &= \sum_{i \in I_0} c_{0,i} \Phi_{0,i} = \sum_{i \in I_1} c_{1,i} \Phi_{1,i} + \sum_{k \in K_1} d_{1,k} \psi_{1,k} \\ &= \sum_{i \in I_2} c_{2,i} \Phi_{2,i} + \left(\sum_{k \in K_1} d_{1,k} \psi_{1,k} + \sum_{k \in K_2} d_{2,k} \psi_{2,k} \right) \\ &= \sum_{i \in I_j} c_{j,i} \Phi_{j,i} + \left(\sum_{k \in K_j} d_{j,k} \psi_{j,k} + \sum_{L=1}^j \sum_{k \in K_L} d_{L,k} \psi_{L,k} \right) \\ &= \sum_{i \in I_j} c_{j,i} \Phi_{j,i} + \left(\sum_{j=1}^J \sum_{k \in K_L} d_{j,k} \psi_{j,k} \right) \end{aligned}$$

The first term, $\sum_{i \in I_j} c_{j,i} \Phi_{j,i}$, corresponds to the smooth signal that contains the low-frequency elements of the original signal present at level j (Figure 2, left). The second term, $\sum_{j=1}^J \sum_{k \in K_L} d_{j,k} \psi_{j,k}$, represents the detailed signal containing the high-frequency elements (noise) of the original signal at this level (Figure 2, right) (Jiang, Taiwan, Chen, He, et al., 2003; Korbel, Milan et al., 2014). Therefore, the difference between the smooth signal of level 1 ($s=1$) and the detailed signal of level 2 ($s=2$) gives the smooth signal at level 2, and so forth (Bakshi, B. R. & Stephanopoulos, 1994). The multi-level representation makes it possible to adequately identify signal characteristics located at different times and places, thus reducing noise in the data (Nounou & Bakshi, 1999). Information about wavelet transforms can be found in several references (Bakshi, Bhavik R., 1999; Daubechies, 1992a, 1992b; Doymaz et al., 2001; Fugal, 2009; Mallat, S. G., 2009; Rioul & Vetterli, 1991).

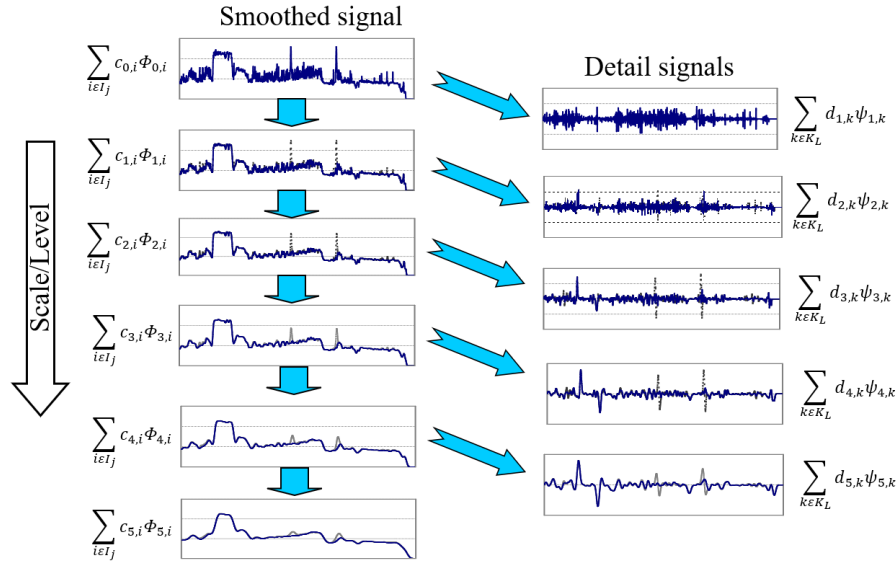


Figure 2 – Decomposition of a signal into smoothed and detailed signals, inspired by (Korbel, Milan et al., 2014)

The quality of WT denoising is strongly affected by several critical parameters, including the mother wavelet function type, the threshold rule, the decomposition level, and the threshold estimate. First, the mother wavelet function is usually selected from families that somewhat resemble the shape of the desired signal. Many types of wavelet families exist (supported, orthogonal, biorthogonal, etc.), and they contain wavelet functions of different orders: Daubechies (Haar, db2...db45), Biorthogonal (bior1.1...bior1.5, bior2.2...bior2.8, bior3.1...bior3.9), Coiflets (coif1...coif5), Symlets (sym2...sym45), Morlet, Meyer, Splines, Gaussian, etc.

Second, for the threshold rule, either soft or hard thresholding can be used (Fig. 3). For a given threshold λ (parameter #4) and value of the wavelet coefficient d , hard thresholding sets to zero those coefficients with absolute values lower than the selected threshold and maintains the same value for coefficients that exceed it :

$$D^H(d|\lambda) = \begin{cases} 0, & |d| \leq \lambda \\ d, & |d| > \lambda \end{cases}$$

Soft thresholding, which is an extension of hard thresholding, acts the same way for values lower than the threshold, but shrinks the coefficient values by the value of the threshold limit :

$$D^S(d|\lambda) = \begin{cases} 0, & |d| \leq \lambda \\ d - \lambda, & |d| > \lambda \\ d + \lambda, & |d| < -\lambda \end{cases}$$

Soft thresholding, also called wavelet shrinkage, takes both positive and negative coefficient values, and shrinks them towards zero, whereas hard thresholding either keeps or removes coefficients. The latter may result in larger variance in the signal after reconstruction but can better represent peaks and discontinuities. The former, on the other hand, has a larger bias, but gives good visual quality of filtering.

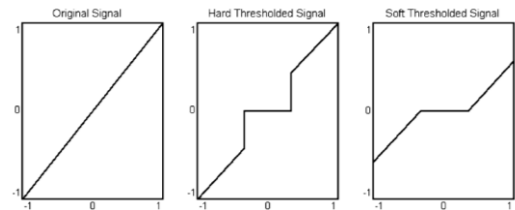


Figure 3 – Soft and Hard Thresholding

Third, the scale factor – the decomposition level, is related to the error frequency, i.e., the frequency correlates with the scale. Small scale values correspond to high-frequency information (compressed signals), i.e., rapid changes in

the detailed signal, whereas high values correspond to low-frequency components (stretched-out signals) associated with a non-detailed global view of the signal. Scaling means either dilatation or compression of the signal.

Finally, the fourth parameter is the thresholding selection rule – the threshold estimate. The threshold value is generally defined based on the standard deviation of the error (noise) amplitude (El-Dahshan, 2011). Selecting a proper threshold value is essential for effective denoising. Some thresholds are : Rigrsure – a threshold selected using the Stein’s Unbiased Risk Estimate (SURE) principle; Universal – a threshold equal to universal $\sigma_j \sqrt{2 \log N}$; and Minimax – a threshold equal to max MSE. The threshold is selected according to the nature of the noise. The Universal threshold reduces noise more efficiently, whereas the other two are more conservative and are more convenient when small details of the signal lie near the noise range (MathWorks, 2020).

Analysis of Signal Filtering Techniques

This section presents some advantages and drawback for each signal filtering technique and discusses how their parameters are selected.

First, the EWMA filter is widely used in process industries. A positive aspect of this approach is that it is easy to implement, and it requires low computational effort (reduced memory and fast calculation). On the downside, enough knowledge on the process data is required to set the alpha parameter – the weight of the last entering value in the data set. Some tests might be required to find the most suitable value according to the application.

Second, the Kalman Filter is theoretically an optimal estimator; it can correct itself between the prediction and update phase. It will accurately estimate the states affected by external disturbance. A major advantage of the Kalman Filter is its ability to provide the estimate with an indicator of the filter accuracy, i.e. the prediction error covariance (Harkat et al., 2016). Of all existing filters, Kalman is the only one that minimizes the covariance of the estimation error (Simon, 2001). On the other hand, this filter needs as an input the process noise variance as well as the observation noise variance. Selecting those values require great knowledge of the process and the sensors (sensor reaction time, sensor bandwidth).

Additionally, to use the KF, the state transition matrix needs to be established. To do so, a perfect model is rarely available, especially in process industries. Equations that compose the model (mass and energy balances) are never exact due to the many assumptions behind them. To the best of the authors knowledge, no one ever used the KF as a noise reducer on process sensor readings. What comes closest to this are publications related to the estimation of states that are not measurable or not measured – one of principal objective of the filter – for reactors based on a model (MacGregor et al., 1986).

In order to be fair with the three other processing technique discussed here, which have no clue about the process model (mass and energy balances) of the HEN, equations (1), (3) and (4) need to be adjusted. Firstly, there is no state transition ($F_t = 1$). Thus, it is not a matrix, it is a scalar value; the Kalman filter is used in a univariate way – it is processing one variable at the time. Also, the system does not have any input control vector (u_t), hence B_t , the control matrix, is null. The assumption here is that the only input in the system is the process noise. Therefore, the (co)variance matrix update (equation 4) depends greatly on Q_t ; the states vary due to the process noise variance matrix, the signals are filtered according to this. Hence, the implementation of the KF approach for reducing noise in simulated sensor readings does not consider the model, it only depends on the measurements quality. In other words, the part regarding the model dynamics is ignored, the algorithm focuses on the data and noise aspects. Otherwise, considering the model information in the KF technique would automatically introduce a bias since the other approaches don’t have access to this information.

Conventional linear filters are effective in removing random errors above or below a certain frequency (out-band). These are simple methods that are easy to use and implement. However, their action are limited for in-band signals since the data is represented for only one frequency level. In other words, these filters are single scale in nature as the basis functions have a fixed time-frequency localization (Nounou & Bakshi, 2000). Indeed, when they are applied to in-band signals, they also remove the signal of interest. Consequently, linear filters must find a balance between efficient removal of temporally global random error and accurate representation of temporally localized characteristics retained in the signal (Nounou & Bakshi, 2000). Sensor measurement data are well known to be non-stationary and multiscale given the variety of events occurring in a process, and thus are less suited for single scale analysis.

This problem is partially overcome with Fourier Transform (FT), which allows frequency analysis of data. On the other hand, it is inefficient in extracting temporal information from signal characteristics (Fugal, 2009). Moreover, few spectral methods rely on the implicit fundamental assumption of signals being periodic and stationary, which is not the case for signals from process sensors (Ganesan et al., 2004). Indeed, FT is not suitable for analyzing non-stationary signals and especially those containing sudden changes. For non-stationary signals, the FT takes the average of the frequency components over the whole duration of the signal (Ganesan et al., 2004). However, these downsides can be overcome with STFT that is adequate for non-stationary signal. STFT, in addition to a good frequency resolution, allows evaluating at what time a particular frequency event occurs in the signal.

One limitation of STFT is that the resolution remains constant over time, i.e., STFT uses a single, fixed analysis window, which means that the width of the time and frequency window is constant throughout the entire STFT process (Figure 4). Hence the time-frequency resolution is quite a limitation mainly due to using a fixed window size; once it is fixed, that window is the same for all frequencies. It is difficult to know which size would be adequate for the signal; a wide window size would have good frequency resolution but poor time resolution and vice versa for a too narrow window. The latter is not adequate for low frequency events and the former is not required for high-frequency components. Thus, the accuracy of the filtering method is greatly limited by the shape of the window. Therefore, STFT is not as well suited for signals with a wide range of frequencies and for locating singularities or short time events of high frequency (Fugal, 2009; Ganesan et al., 2004). This aspect makes it difficult to get proper characteristics from the signals (Park, S.-M. et al., 2016). Henceforth, a major disadvantage of the STFT is the resolution trade-off between time and frequency.

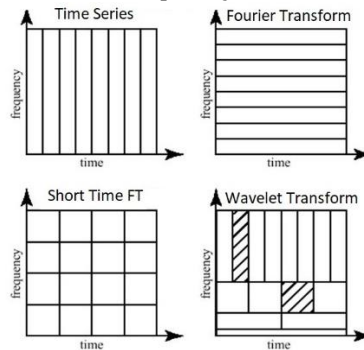


Figure 4 – STFT Fixed-Window (constant aspect ratio) and WT Variable-Window (Ganesan et al., 2004)

With WT, the notion of scale is introduced as an alternative to frequency, leading to a so-called *time-scale representation* – the scale factor is defined as the inverse of frequency. It is the equivalent of the time-frequency plane used in the STFT. In contrast to the STFT, which uses a single analysis window, WT uses windows of varying sizes adjusted to capture the signal important characteristics : narrow windows for high frequencies and broader windows for low frequencies (Figure 11). As a result, WT achieves an ideal balance of time-scale resolution (it has

more flexibility in time and scale resolution). Also, compared to STFT, it is better at revealing short-time changes or characteristics in signals occurring at high frequency (carrying out local analysis) and describing local features.

The wavelet Transform is known to be the most common, powerful, and successful technique for non-stationary signal denoising (Alyasseri, Z. A. A. et al., 2020). DWT is the most generally used time-frequency filtering method (Ben Slama et al., 2018). By looking at a signal at various scales and analyzing it with different resolutions (given its inherent multi-resolution capabilities), WT gets very efficient at detecting and representing elements in the signal (e.g., noise, abnormalities, features, and trends). This approach can reduce unwanted features that are superimposed on the signal of interest through frequency decomposition and the use of statistically based thresholding, without impacting it. In other words, the implementation of wavelet transform results in minimum distortion of the filtered data (WT is able to decompose complex signals and then reconstruct them without losing information relevant to the process) and helps to extract features that are associated only with sensor dynamics (Upadhyaya et al., 2014).

This method provides excellent time-frequency localized information and can deliver a multiresolution representation of a signal for multiscale events examination. It also provides an excellent means to analyze nonstationary and autocorrelated non-Gaussian data since it generates stationary and independently and identically distributed Gaussian wavelet coefficients irrespective of the input data distribution, even if it is non-stationary. Hence, the WT analysis gives the opportunity of applying statistical procedures that require Gaussian distributed data to the wavelet decomposed data.

Additionally, signals with sharp changes might be better analyzed with wavelet functions than with a smooth sinusoid provided by STFT. Wavelets are used in a broad range of application fields because it can process signals of various shapes thanks to the different families of wavelets. Indeed, STFT decomposes signals into sines and cosines whereas WT offers many different basis functions for decomposition that have various shapes (Figure 5). Hence, these basis functions can fit to various signal's characteristics, and it gives the WT the ability to denoise unusual signals.

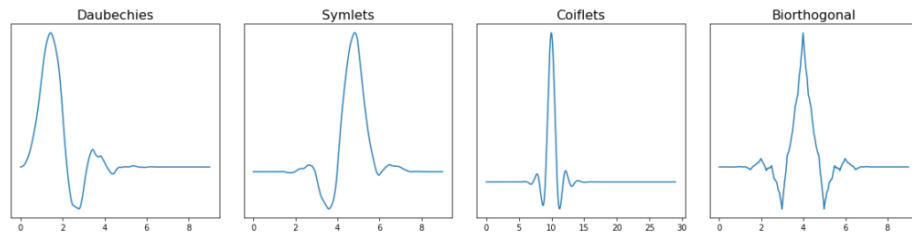


Figure 5 – Examples of Wavelets Functions, inspired by (Taspinar, 2018)

Methodology for comparing signal processing methods

This paper proposes to compare the denoising performance of four signal processing methods on synthetic signals (with known errors) from a heat exchanger network (HEN) simulation, i.e., assess their performance in adequately removing errors. The performance analysis is carried out using two criteria : the root-mean-square error and the signal-to-noise ratio between processed signal and the signal of interest. Furthermore, this paper considers the selection of the optimal parameters of each method. The latter are those that minimize the error between the signal filtered by the processing technique and the signal of interest (presented in Figure 7) using a parametric sweep (grid search). The different values of the parameters for each filtering technique (see Section 2) are presented in Table 2.

Table 2 – Ranges of commonly used method-specific denoising parameters for KF, EWMA, STFT and WT

Methods	Method-specific Parameters	Range or value
Kalman Filter	Measurement noise R	Standard deviation of the noise
	Process noise Q	R/1000; R/100, R/10, R (Ma'arif et al., 2020)
Exponential Weighted Moving Average	Smoothing factor α	0.05 to 0.95 (with a step of 0.05)
Short Time Fourier Transform	Window Type	Blackman; Chebyshev; Gaussian; Hamming; Hann; Rectangular; Triangular
	Window Size	Hour; Day; Week; Month
	FFT length	Twice the window size
	Sampling rate	One point per minute
	Overlap length	1%; 10%; 25; 50; 75 % overlap of the window size
Wavelet Transform	Wavelet function	Daubechies (db2-db15); Symlet (sym2-sym15); Coiflet (coif1-coif5); Biorthogonal (bior1.1-bior6.8)
	Scale (Decomposition level)	1 to10
	Thresholding function	Soft threshold; Hard threshold
	Threshold selection rule	Rigrsure; Universal; Minimax

Regarding the Kalman Filter parameters (a covariance matrix that reflects the uncertainty of the measurements and another one that reflects the uncertainty on the model), when experimenting with various values for the measurement noise and process noise, their ratio is what matter the most. In other words, when tuning a Kalman filter, rather than adjusting separately Q and R, it is their ratio that matters, it is related to the trust in the model versus the trust in the measurements. The measurement noise is assumed to be the measurement variance of the signal of interest. The different ratios for the process noise used in this work are taken from (Ma'arif et al., 2020).

Comparison of signal processing methods for error detection

First, the synthetic signals are obtained through a simulation (heat exchanger network in this case) to which fictional random errors are added. Next, the optimal parameters are identified for each processing technique using a parametric sweep that keeps the parameters that minimizes the error between the filtered signal and the signal of interest, i.e., the RMSE. Then, the optimal parameters are used to process the synthetic signals and appraise the performance of the Kalman filter, EWMA filter, short-time Fourier transform, and wavelet transform using quantitative criteria, i.e., RMSE and SNR. Lastly, the approaches are compared qualitatively in the discussion.

5.1 Simulation and synthetic signal

The synthetic (simulated) signals are obtained from the simulation of a heat exchanger network, inspired by a crude oil preheating train, developed in VBA Excel based on the number of transfer units (NTU) approach (Figure 6) (Bergman et al., 2011). Flow and temperature sensors are not available on all streams, as it would be the case in a real industrial process. This simulation provides six flow signals and 14 temperature signals. The blue streams are water, whereas the red streams have the specific heat capacity (Cp) of oil.

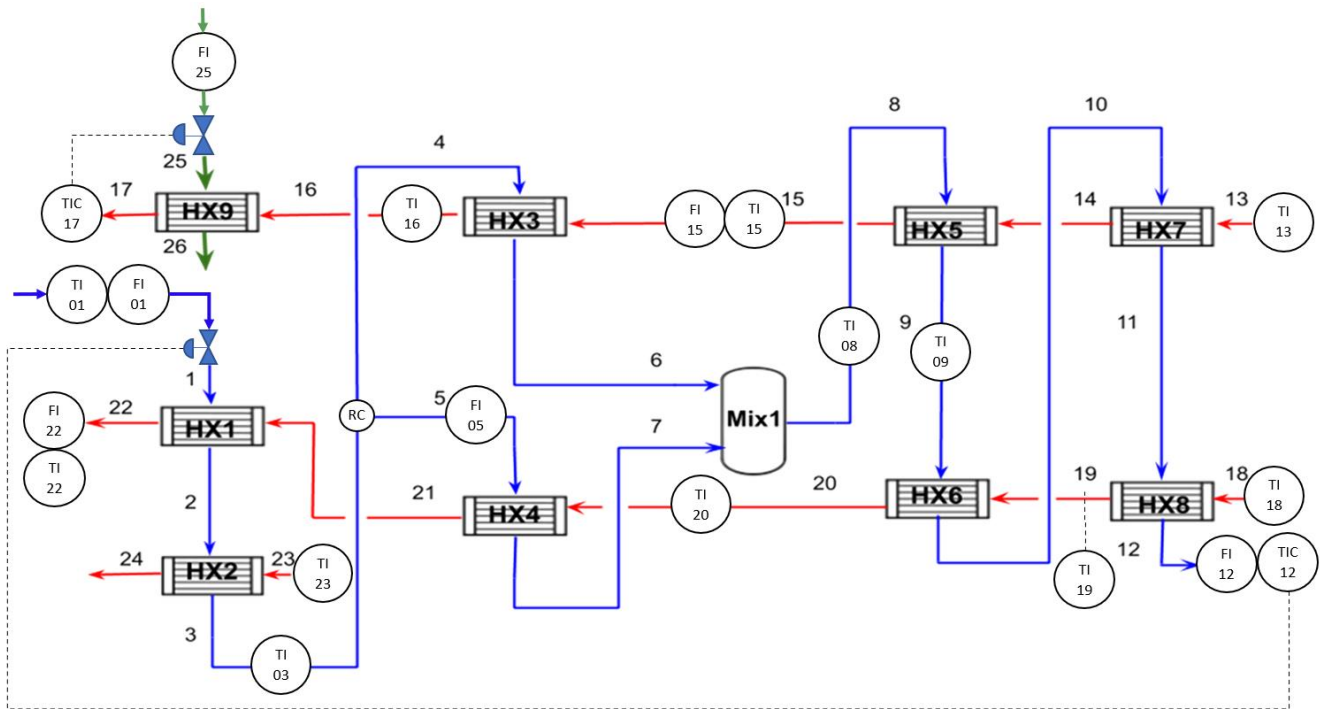


Figure 6 – Heat exchanger network simulation : flow and temperature signals

Out of the twenty (20) sensors, two (2) are part of a control loop and 18 are indicators. The temperature controller TIC12 adjusts the opening of a control valve on stream 1 according to a temperature set point on stream 12. Additionally, stream 25 is cooling stream 16, it has the Cp of air, and its flow is controlled according to the temperature of stream 17. Thus, TIC17 adjusts the opening of a control valve on stream 25 to control the temperature of stream 17. Lastly, the split ratio between stream 4 and 5 can also be controlled. Figure 6 shows the ratio controller (RC) at the split.

The simulation signals are generated considering the rate of change by using a superimposed high frequency random function over seasonal and operational variations (trend of a process). Small high frequency random values are used to demonstrate the fast variability that a signal can have. The latter are generated uniformly around the trend. Those small random values are normally distributed given the fact that the hypothesis stating that random errors are following a gaussian/normal distribution is generally accepted, it is a good approximation of many real-world environments (Park, S. et al., 2019). Random errors are added considering the amplitude of the measurement to provide realistic and interesting process signals.

The simulation provided a year of data sampled every minute. These simulated signals aim to be as close as possible to real signals without random errors; they are representative of true and realistic process signals. For that reason, sequential calculation method has been used to simulate the noise.

To obtain a good representation of time-series data from process sensors, understand the process dynamics and mimic real process signals, a setpoint change is applied to the stream 12 temperature controller (TIC12) (Figure 7). This allows to see how the other process variables react to this change and thus evaluate which filtering approach is robust enough to reduce errors notwithstanding the setpoint change; that is, to assess if the approaches are able to adjust to this change, reduce random error while estimating the signal's trend.

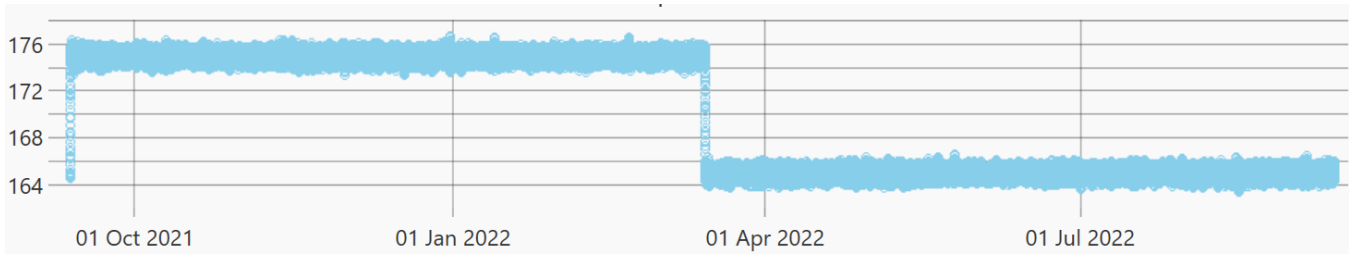


Figure 7 – Change in the final temperature setpoint (March 14th) – stream 12

Therefore, the HEN simulates normal continuous operations, i.e., (no tube leaks, fouling, sudden variation of the split ratio, heat exchanger by-pass), with small random variations as well as a setpoint change. In an actual industrial process system, there are, among other things, grade changes, adjustments in production and seasonal variation. Seasonal variation is considered in the simulation, as presented in Figure 8. The temperature of stream 1 is increasing during summer and goes back around 8 °C afterwards.

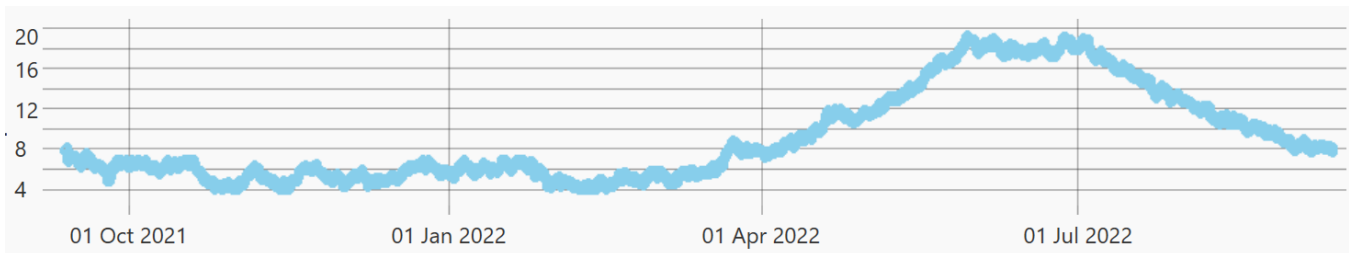


Figure 8 – Temperature on stream 1 exhibit seasonal variation

Lastly, Figure 9 presents an example of a raw flow measurement signal in blue and the useful signal (signal of interest) in black. The signal of interest aims to represent the signal trend as accurately as possible, reducing the effects of measurement and process errors, whereas the raw signal is the one containing noise.

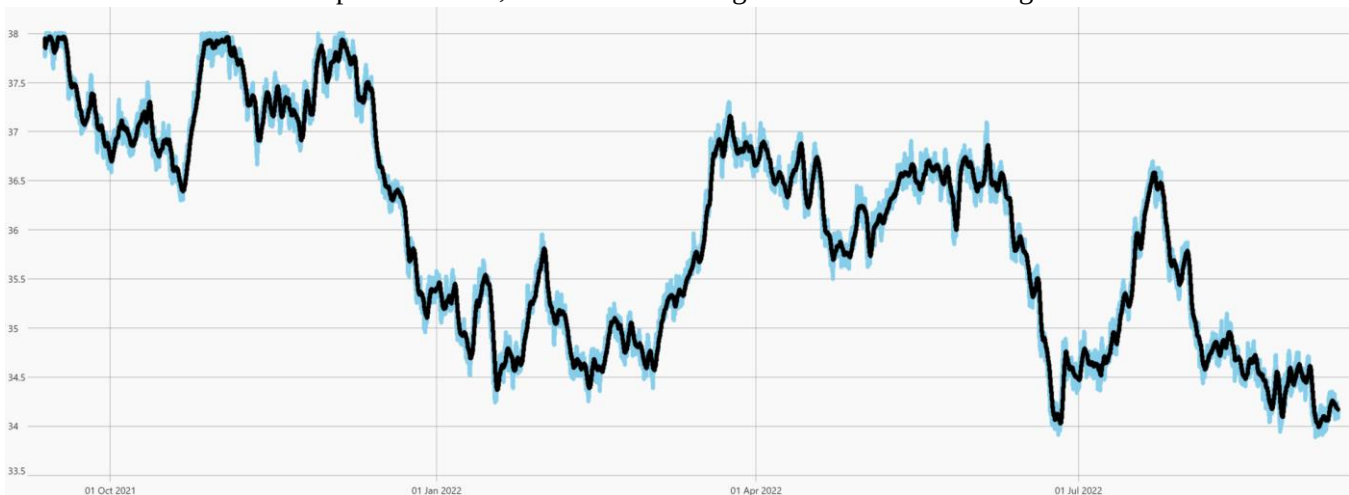


Figure 9 – Raw flow measurement signal (in blue) and the signal of interest (in black) – stream 15

5.2 Denoising Performance Evaluation – Synthetic signals from a heat exchanger network simulation

In this section, the performance of four signal processing methods (Kalman Filter, EWMA Filter, STFT and Wavelet Transform) are assessed and compared using the root mean square error and the signal-to-noise ratio, two easily interpretable criteria and the most used performance criteria in signal processing (Ben Slama et al., 2018; Muppalla

et al., 2017). These denoising performance metrics are evaluated systematically before and after signal processing for each method, and the difference is analyzed.

RMSE is the average distance (in absolute value), measured along a vertical line, between every pair of data points in two signals. It measures how close each processed signal is to its respective noiseless version, i.e., the quality of the filtered signal. The values are always non-negative, they have the same units as the signal (they are directly interpretable in terms of measurement units), and the closer they are to zero, the better, i.e., the smaller the RMSE, the more similar the processed signal is to the original noiseless data. The objective is to minimize the error between the processed signal and the original signal. The deviations between the real (original) signals and the measured noisy ones, as well as between the real signals the processed ones, are given respectively by the following equations:

$$RMSE_{before} = \sqrt{\frac{1}{N} \sum_{i=1}^N [x(i) - \bar{x}(i)]^2}; \quad RMSE_{result} = \sqrt{\frac{1}{N} \sum_{i=1}^N [x(i) - \hat{x}(i)]^2}$$

where $\bar{x}(i)$ is the synthetic noisy signal, $\hat{x}(i)$ is the denoised signal, $x(i)$ is the original (reference) signal, and N is the number of samples in the signal.

This criterion is used to assess the effectiveness of a processing method by grading the difference between $RMSE_{before}$ and $RMSE_{result}$. The former should be larger than the latter. Then this difference is divided by $RMSE_{before}$ to evaluate the $RMSE_{reduction}$. The larger the value, the greater is the reduction, and the more effective is the algorithm. A 100% reduction would mean that there is no more error and that the filtered signal is identical to the original one:

$$RMSE_{difference} = RMSE_{before} - RMSE_{result}$$

$$RMSE_{reduction}[\%] = \frac{RMSE_{difference}}{RMSE_{before}} \times 100$$

The noise reduction efficiency is evaluated using the signal-to-noise ratio. SNR is defined as the ratio of signal power to noise power, often expressed in decibels, as presented below :

$$SNR_{db} = 10 \log_{10} \left(\frac{P_{signal}}{P_{noise}} \right)$$

SNR compares the magnitude of a desired signal to the level of background noise. This characteristic reflects the dynamic of the signal. A ratio greater than 0 dB indicates more signal than noise, so the higher the SNR, i.e., the lower the noise content, the better is the signal quality.

The SNR before denoising is defined as

$$SNR_{before}(dB) = 10 \log_{10} \frac{\sum_{i=1}^N \bar{x}^2(i)}{\sum_{i=1}^N [\bar{x}(i) - x(i)]^2}$$

The SNR after denoising is defined as

$$SNR_{result}(dB) = 10 \log_{10} \frac{\sum_{i=1}^N \hat{x}^2(i)}{\sum_{i=1}^N [\hat{x}(i) - x(i)]^2}$$

SNR_{before} should be less than SNR_{result} to define the effectiveness of denoising techniques. $SNR_{improvement}$ is used to assess the noise reduction efficiency. The higher the value, the better is the result.

$$SNR_{difference}(dB) = SNR_{results}(dB) - SNR_{before}(dB)$$

$$SNR_{improvement}[\%] = \frac{SNR_{difference}}{SNR_{before}} \times 100$$

The results of denoising by Kalman filter, EWMA filter, STFT, and WT are compared by giving the RMSE and SNR parameters. The ideal signal processing technique would be the one with the highest value for both criteria, i.e., showing the greatest SNR improvement and the greatest RMSE reduction. Figures 10 and 11 present the RMSE and SNR results, denoted respectively as $RMSE_{reduction}$ and $SNR_{improvement}$ for all four filtering methods for all temperature and flow signals.

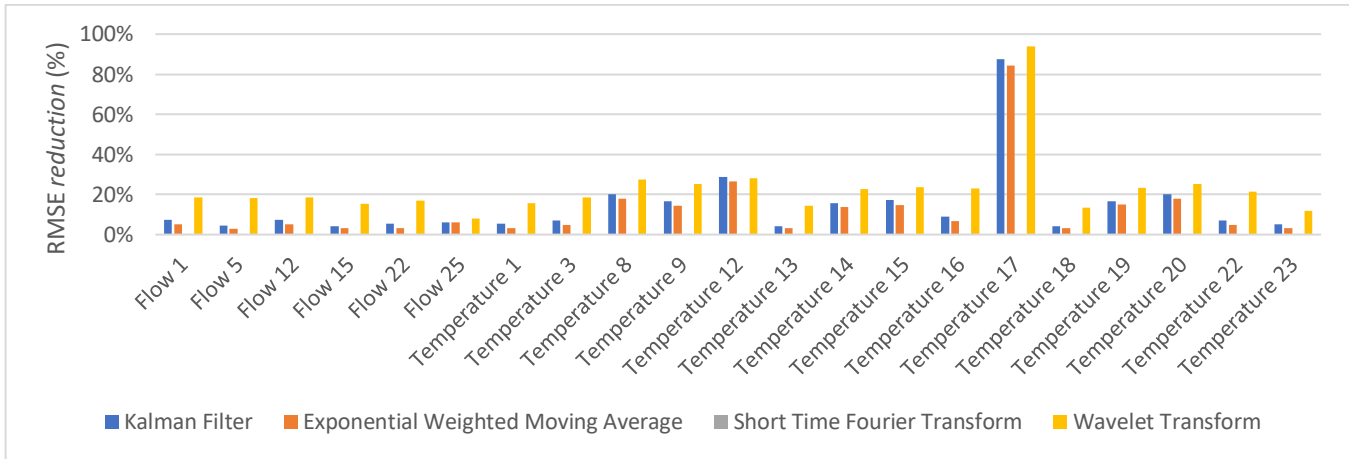


Figure 10 – RMSE reduction for synthetic signals for KF, EWMA, STFT and WT

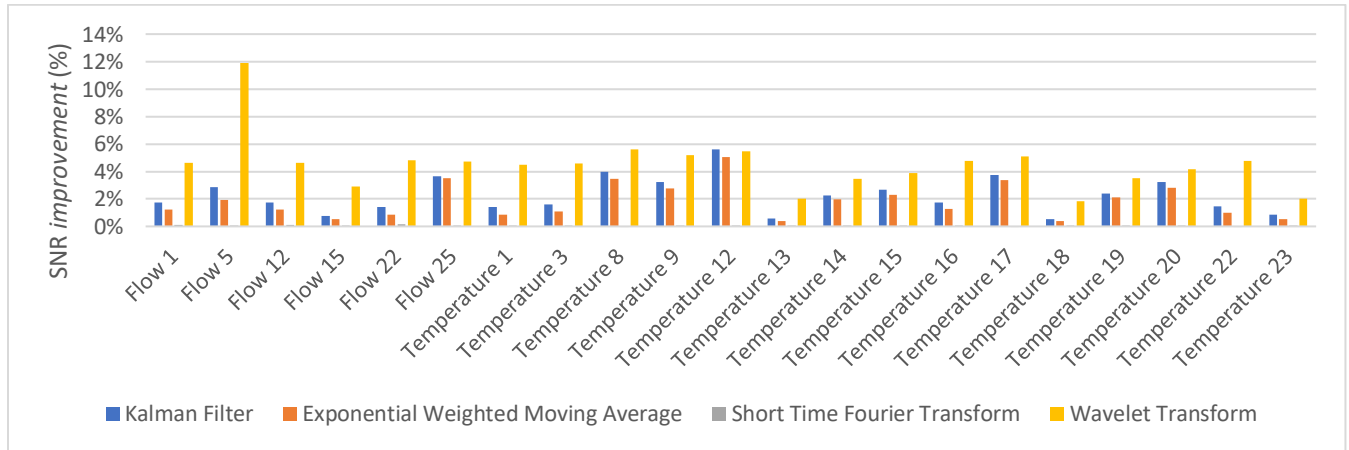


Figure 11 – SNR improvement for synthetic signals for KF, STFT, WT and EWMA

Figures 10, and 11 present respectively the comparison of different signal processing techniques in terms of RMSE reduction and SNR improvement in percent when processing synthetics temperature and flow signals from a HEN simulation. WT approach dominates in term of RMSE reduction, it is the best-performing method for most of the signals, except for few signals (temperature 12 and flow 25) where WT is as performant as the other techniques. In regard to the SNR improvement, WT is the preferred approach for all signals, except temperature 12 where the SNR

improvement is 5.60% for KF and 5.45% for WT. STFT appears to be systematically the least efficient techniques towards RMSE reduction as well as SNR improvement. An analysis is proposed in the following section.

RMSE reduction varied over a wide range, from null values for STFT to almost a 100% for WT. This means that signal processing technique performance varied widely, and the same is true for SNR improvement. Finally, both criteria led to the same conclusions : WT is the most efficient for Gaussian noise reduction for simulated HEN signals when compared with KF, EWMA, and STFT. Denoising with WT secured the lowest RMSE and the highest SNR most of the time, hence the largest difference from the original signal.

All processing techniques presented higher RMSE reduction than SNR improvement values, which means that they all reduced the difference between the original signals and those of interest, but the signals are still noisy. It is not possible to perfectly reconstruct a noise-free signal, but a proper filter will give the trend in the data rather than its instantaneous variation. Therefore, in this case, the filtered signals must resemble the signal of interest.

The two criteria are always in agreement; a signal that has a high value in RMSE reduction also has a high value in SNR improvement. WT showed its effectiveness by performing almost equally well for both metrics (RMSE and SNR) regardless of the signal; the RMSE reduction is about 23% and SNR improvement around 5%. Table 3 represents averaged values of RMSE reduction and SNR improvement, along with their standard deviations, for all signals (flows, temperatures, and both). The results unambiguously show that the WT are globally significantly better than other approaches according to the RMSE reduction and SNR improvement. In fact, WT shows on average 14.14% more RMSE reduction and 3.12% more SNR improvement when compared to KF, EWMA and STFT when considering all signals.

Table 3 – RMSE and SNR averages with their standard deviations for all approaches

	Sensors	KF	EWMA	STFT	WT
Average RMSE	Flow	5.79% $\pm$ 1.33%	4.31% $\pm$ 1.18%	0.21% $\pm$ 0.21%	15.88% $\pm$ 4.04%
	Temperature	17.62% $\pm$ 20.66%	15.55% $\pm$ 20.32%	0.01% $\pm$ 0.01%	25.88% $\pm$ 19.52%
	All	14.24% $\pm$ 18.15%	12.34% $\pm$ 17.79%	0.06% $\pm$ 0.14%	23.03% $\pm$ 17.09%
Average SNR	Flow	2.02% $\pm$ 1.05%	1.56% $\pm$ 1.07%	0.08% $\pm$ 0.06%	5.59% $\pm$ 3.18%
	Temperature	2.35% $\pm$ 1.42%	1.96% $\pm$ 1.34%	0.01% $\pm$ 0.01%	4.05% $\pm$ 1.26%
	All	2.26% $\pm$ 1.31%	1.84% $\pm$ 1.26%	0.03% $\pm$ 0.05%	4.94% $\pm$ 2.03%

Figure 13 exhibits, on the left side, the original TI20 signal (blue), its signal of interest (black), and the filtered signal (green) for all methods and on the right side, both the measurement error (red) and the estimation error (black) of the respective method. The former evaluates the difference or distance between the original signal and the signal of interest, and the latter is the difference between the original and filtered signal.

Figure 13 a) shows a greatly decreased level of noise and a smaller error than what is originally in the signal. The KF presents a considerable improvement – a RMSE reduction of 20.17% and a SNR improvement of 3.21%. Even if the results presented in Figure 13 b) are quite similar as those obtained with KF, the denoising is not as good as the one obtained using KF – EWMA presents a RMSE reduction of 17.82% and a SNR improvement of 2.80%. Figure 13 c) presents a poor execution of filtering with STFT; there is almost no difference from the error before filtering (RMSE reduction and SNR improvement are respectively at 0.04% and 0.03%. In Figure 13 d), almost no noise remains; the signal of interest is practically superimposed with the filtered signal. WT offers the lowest percentage of error relative to the signal of interest : RMSE reduction of 25.39% and a SNR improvement of 4.79%. The improvement of the estimation error is clearly visible and differs from all the other approaches.

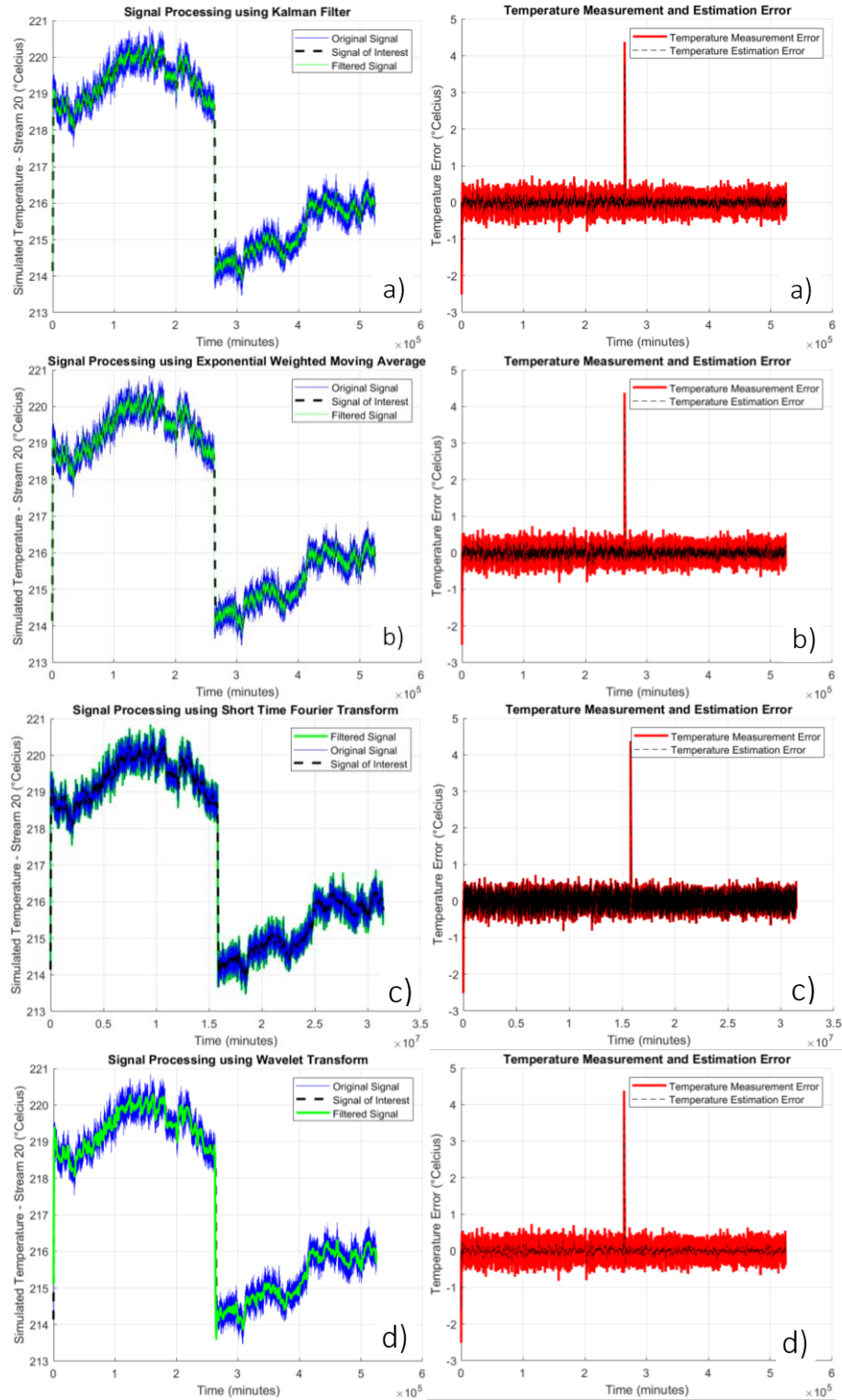


Figure 13 – Denoising results of synthetic signal temperature 20 using a) KF, b) EWMA, c) STFT, and d) WT

5.3 Discussion

Following the comparison analysis of four processing technique, the identification of optimal parameters – those that could minimize the RMSE – is presented in Table 5.

Table 5 – Method-specific parameters values for KF, STFT, WT, and EWMA filter for all signals

Signals	Signal processing approaches optimal parameters									
	Kalman Filter		Short Time Fourier Transform			Wavelet Transform				EWMA
	Measurement noise R*	Process noise Q	Window type	FFT length	Overlap length (% of the window size)	Wavelet Function	Decomp. level	Threshold selection rule	Thresholding function*	Smoothing factor α
F 1	1.69E-01	1.69E-04	rectwin	Week	10%	db4	10	UT	Soft	0.05
F 5	9.28E+00	9.28E-03	chebwin	Day	10%	sym2	10	UT	Soft	0.05
F 12	1.69E-01	1.69E-04	rectwin	Week	10%	db4	10	UT	Soft	0.05
F 15	5.55E-03	5.55E-06	chebwin	Month	25%	sym2	10	MM	Soft	0.05
F 22	3.98E-01	3.98E-04	chebwin	Week	10%	db4	10	MM	Soft	0.05
F 25	9.41E+00	9.41E-03	gausswin	Day	10%	sym4	10	UT	Soft	0.05
T 1	1.98E-02	1.98E-05	hamming	Hour	1%	sym2	10	UT	Soft	0.05
T 3	3.77E-02	3.77E-05	chebwin	Week	25%	sym3	10	MM	Soft	0.05
T 8	6.78E-02	6.78E-05	chebwin	Day	10%	db12	10	UT	Soft	0.05
T 9	9.81E-02	9.81E-05	triang	Week	25%	db13	10	UT	Soft	0.05
T 12	1.03E-01	1.03E-04	rectwin	Hour	75%	db12	10	UT	Soft	0.05
T 13	4.78E-03	4.78E-06	chebwin	Week	1%	sym3	10	MM	Soft	0.05
T 14	9.83E-03	9.83E-06	chebwin	Day	1%	db12	10	UT	Soft	0.05
T 15	2.11E-02	2.11E-05	rectwin	Hour	75%	db12	10	UT	Soft	0.05
T 16	7.36E-02	7.36E-05	hamming	Month	75%	db4	10	UT	Soft	0.05
T 17	3.53E-02	3.53E-05	hamming	Day	10%	db7	10	MM	Soft	0.05
T 18	5.37E-03	5.37E-06	gausswin	Hour	25%	sym3	10	MM	Soft	0.05
T 19	9.35E-03	9.35E-06	hamming	Hour	75%	db12	10	UT	Soft	0.05
T 20	2.37E-02	2.37E-05	rectwin	Hour	75%	db12	10	UT	Soft	0.05
T 22	2.88E-01	2.88E-04	gausswin	Day	10%	sym3	10	UT	Soft	0.05
T 23	5.91E-03	5.91E-06	chebwin	Hour	1%	sym2	10	MM	Soft	0.05

F= Flow ,T = Temperature. *°C for T sensors and ton/h for Flow sensors. UT = UniversalThreshold, MM = Minimax

For the KF, the ratio to obtain the optimal process noise value is identical for all experiments; it is R divided by 1000. The algorithm selected the smallest value as the optimal one. Moreover, a high process noise value implies a higher state uncertainty (Cortés et al., 2022). However, on the contrary, a lower value suggest a higher confidence on the process or state model, i.e., both are well known (Cortés et al., 2022). Nevertheless, the measurement noise values also have physical meaning as they represent the variance of the measurement.

With the short-time Fourier transform, when comparing various overlap lengths, FFT lengths and window type, the results showed that the optimal values differ from one signal to another. In general, more overlap – smaller hop size – will give more analysis points and hence smoother results over time, but the computational expense is proportionately greater. Moreover, to achieve a smooth signal, the frequency domain information does not need to be very specific. Hence, when looking for a window length that results in the loss of spectral resolution, a narrow window is preferred. However, a longer window enables better trend identification. Therefore, for the window size (FFT length), a compromise must be found between the time and the frequency resolution.

For the wavelet transforms, different mother wavelet functions are tested. Because orthogonal wavelets such as the Coiflet, Daubechies and Symlet families are the best for denoising signals, these wavelet functions are tested for various orders (coif1-coif5, sym2 to 15 and db2 to db15). The biorthogonal wavelet function (bior1.1-bior6.8) is also tested. Decomposition levels (scale) from 1 to 10 are considered. The latter must not be too low, in which case high frequency components might hide the trend, but also, it must not over-smooth the signal (hence, not too high of a scale). The optimal values for the wavelet function and the scale are obtained for all signal by minimizing the

RMSE. It is found that Daubechies and Symlet are the two selected function and decomposition level 10 is chosen for all signals. This value is coherent with the fact that the signals are corrupted with high frequency components that needed to be reduced. Therefore, the highest scale is required to reach the signal of interest. The optimal thresholding selection rule is either minimax or universal threshold. The former is more conservative and is more convenient when small details of the signal lie near the noise range whereas the latter is known to reduce the noise more effectively (MathWorks, 2020). Finally, the soft thresholding function led to the best results for all signals. This may be explained by the fact that hard thresholding results in larger variance after reconstruction and can very well represent peaks in the signal, but since the signal's trend is what matter here (signal of interest), this is not applicable.

Lastly, for the exponential weighted moving average, the smoothing factor is all the time at its smallest value (0.05). A value near zero means that the filter is less responsive to new incoming measurements, hence there is a greater smoothing effect, more weight is put on the average.

The success of a signal processing method depends on the optimal configuration of its control parameters. Working with synthetic simulated signals with known errors makes possible the identification of optimal parameters values. However, in a real-life context, the selection of these parameters is much more complex. In the case of KF, setting the state equations requires a good knowledge of the process, and this is hard because processes are complex. Additionally, the process noise is unknown, and the measurement noise is hard to determine, it varies according to sensor sensitivity and accuracy. Therefore, for a real process, obtaining optimal result would be much harder because the process is more complex and assessing the state equation and both matrices is tedious. In the case of EWMA, the value of the smoothing parameter is typically correlated with measurement precision or accuracy. This parameter is not set according to the signal's appearance or characteristics but rather on the confidence in data.

However, because measurement signals are generally characterized by noise at several frequencies, these filters force a compromise between the quality of improvement over a long period (signal distortion) and of local signal characteristics. Linear filters are therefore not effective in filtering signals containing several characteristics located at different positions in the time and/or frequency domains (Bakshi, B. R. & Stephanopoulos, 1994; Nounou & Bakshi, 1999). Problems with linear filter efficiency have led to the development of multi-stage filtration methods, such as WT. Moreover, WT makes it possible to shift the wavelet function to align it with events to reduce noise efficiently.

Wavelet transforms parameters are selected according to the signal's characteristics. The mother wavelet function must match the shape of the signal, the scale and thresholding parameters must be chosen in order to retain only the signal's trend or important component and reduce high frequency elements. STFT parameters are also selected according to the frequency content of the signal, however this approach did not perform very well on either criteria, there is less error in the signals, but noise is still very much present and Figures 10, 11-12 and 13 c) corroborates that. Therefore, it would be interesting to assess the relation between signals characteristics and the WT parameters used to process them in specific applications (as signals are always processed according to a targeted application).

Conclusion

This article proposes a comparison of four signal processing approaches. Synthetic signals containing random errors are created and processed one at the time with KF, EWMA, STFT, and WT. Because the errors are known, it is possible to assess their optimal parameters and compare the methods to find the best-performing one. The performance of these denoising algorithms is evaluated using two criteria : RMSE and SNR.

As a result, single scale filtering methods (EWMA and KF) are much less efficient for simultaneously filter non-stationary measured signals containing several characteristics/multiple features located at different positions in time-frequency domain and accurately represent these characteristics and features. Additionally, short time Fourier transform is found to be inadequate to reduce error from heat exchanger network simulated signals.

Compared to other signal processing approaches, WT is found to be robust towards noise reduction in industrial process signals and adequate for improving their SNR and reducing the RMSE. In fact, WT enabled on average 14.14% more RMSE reduction and 3.12% more SNR improvement than KF, EWMA and STFT when considering all signals. It is a good signal processing technique that can reduce errors from signals while retaining useful information; it retains the variability of the original signal, which reflects the true changes and characteristics of the process. WT stand out for its signal processing performance for simulated industrial signals, and for the way its parameters (mother function, scale, threshold, etc.) are selected. The idea of assessing the relation between the signal's characteristics and the signal processing technique's parameters to observe how beneficial it could be for processing a wide range of signals in a systematic and efficient way could be further addressed in a following paper. Additionally, further research could be conducted on the 4 approaches by adding different types of faults in the signals. As part of this paper, the HEN simulated normal continuous operations only, however, tube leaks, fouling, sudden variation of the split ratio, heat exchanger by-pass could be considered.

Rectified data will serve as the basis for many subsequent decisions and enable better control of plant operations and decision-making. Signal processing increases the efficiency of tasks related to the operation of chemical processes : configuration of control systems, detection and diagnosis of faults directly related to process operations (leaks or deposits), and planning and scheduling of process operations and maintenance. Because process data contain errors that may be detrimental, it is critical to reduce them before making any important decisions.

Acknowledgements

The authors gratefully acknowledge the financial support of the CRSNG and further thank CanmetENERGY for the technical support using the EXPLORE software.

References

- Abbaszadeh, P., & Alipour, A. (2017). Development of a coupled wavelet transform and evolutionary Levenberg-Marquardt neural networks for hydrological process modeling. *Computational Intelligence*, 34, 1-25. <https://doi.org/10.1111/coin.12124>
- Alyasseri, Z. A. A., Khader, A. T., Al-Betar, M. A., Abasi, A. K., & Makhadmeh, S. N. (2020). EEG Signals Denoising Using Optimal Wavelet Transform Hybridized With Efficient Metaheuristic Methods. *IEEE Access*, 8, 10584-10605. <https://doi.org/10.1109/ACCESS.2019.2962658>
- Bagajewicz, M. (2000). A brief review of recent developments in data reconciliation and gross error detection/estimation. *Latin American Applied Research*, 30(4), 335-342.
- Bahoura, M. (2019). Efficient FPGA-Based Architecture of the Overlap-Add Method for Short-Time Fourier Analysis/Synthesis. 8(12), 1533. <https://www.mdpi.com/2079-9292/8/12/1533>
- Bakshi, B. R. (1999). Multiscale analysis and modeling using wavelets. *Journal of Chemometrics*, 13(3-4), 415-434. [https://doi.org/10.1002/\(SICI\)1099-128X\(199905/08\)13:3/43.0.CO;2-8](https://doi.org/10.1002/(SICI)1099-128X(199905/08)13:3/43.0.CO;2-8)
- Bakshi, B. R., & Stephanopoulos, G. (1994). Representation of process trends—III. Multiscale extraction of trends from process data. *Computers & Chemical Engineering*, 18(4), 267-302. [https://doi.org/https://doi.org/10.1016/0098-1354\(94\)85028-3](https://doi.org/https://doi.org/10.1016/0098-1354(94)85028-3)
- Ben Slama, A., Lentka, Ł., Mouelhi, A., Diouani, M. F., Sayadi, M., & Smulko, J. (2018). Application of statistical features and multilayer neural network to automatic diagnosis of arrhythmia by ECG signals. *Metrology and Measurement Systems*, 25. <https://doi.org/10.24425/118163>

- Bergman, T. L., Lavine, A., Incropera, F. P., & DeWitt, D. P. (2011). *Introduction to heat transfer*. Wiley.
- Bhat, S. A., & Saraf, D. N. (2004). Steady-State Identification, Gross Error Detection, and Data Reconciliation for Industrial Process Units. *Industrial & Engineering Chemistry Research*, 43(15), 4323-4336. <https://doi.org/10.1021/ie030563u>
- Cao, Z., Jia, H., Zhao, T., Fu, Y., & Wang, Z. (2022). An Adaptive Self-Organizing Migration Algorithm for Parameter Optimization of Wavelet Transformation. *Mathematical Problems in Engineering*, 2022, 6289215. <https://doi.org/10.1155/2022/6289215>
- Cortés, I., van der Merwe, J. R., Lohan, E. S., Nurmi, J., & Felber, W. (2022). Performance Evaluation of Adaptive Tracking Techniques with Direct-State Kalman Filter. *Sensors (Basel)*, 22(2). <https://doi.org/10.3390/s22020420>
- Crowe, C. M. (1996). Data reconciliation — Progress and challenges. *Journal of Process Control*, 6(2), 89-98. [https://doi.org/10.1016/0959-1524\(96\)00012-1](https://doi.org/10.1016/0959-1524(96)00012-1)
- Dash, P. K., & Chilukuri, M. V. (2004). Hybrid S-transform and Kalman filtering approach for detection and measurement of short duration disturbances in power networks. *IEEE Transactions on Instrumentation and Measurement*, 53(2), 588-596. <https://doi.org/10.1109/TIM.2003.820486>
- Daubechies, I. (1992a). 1. The What, Why, and How of Wavelets. Dans *Ten Lectures on Wavelets* (p. 1-16).
- Daubechies, I. (1992b). 3. Discrete Wavelet Transforms: Frames. Dans *Ten Lectures on Wavelets* (p. 53-105).
- Diatkine, C. (2011). *Signal analysis introduction - FFT Parameters*. Support IRCAM. <http://support.ircam.fr/docs/AudioSculpt/3.0/co/FFT%20Params.html>
- Dong, Y., Li, Y., Xiao, M., & Lai, M. (2009). Unscented Kalman filter for time varying spectral analysis of earthquake ground motions. *Applied Mathematical Modelling*, 33(1), 398-412. <https://doi.org/https://doi.org/10.1016/j.apm.2007.11.020>
- Doymaz, F., Bakhtazad, A., Romagnoli, J. A., & Palazoglu, A. (2001). Wavelet-based robust filtering of process data. *Computers & Chemical Engineering*, 25(11), 1549-1559. [https://doi.org/https://doi.org/10.1016/S0098-1354\(01\)00718-9](https://doi.org/https://doi.org/10.1016/S0098-1354(01)00718-9)
- Dyason, W., Van Niekerk, T. I., Phillips, R., & Stopforth, R. (2017). *Performance evaluation and comparison of filters for real time embedded system applications*. 28th Annual Symposium of the Pattern Recognition Association of South Africa and the 10th Robotics and Mechatronics International Conference, PRASA-RobMech 2017, November 29, 2017 - December 1, 2017, Bloemfontein, South africa (vol. 2018-January, p. 242-248). <https://doi.org/10.1109/RoboMech.2017.8261155>
- El-Dahshan, E.-S. A. (2011). Genetic algorithm and wavelet hybrid scheme for ECG signal denoising. *Telecommunication Systems*, 46(3), 209-215. <https://doi.org/10.1007/s11235-010-9286-2>
- Evans, W. C. (NA). *Exponentially-Weighted Moving Average*. GeoGebra. <https://www.geogebra.org/m/tb88mqrm>
- Everett, J. (2011). *The exponentially weighted moving average applied to the control and monitoring of varying sample sizes* (vol. 51).
- Fugal, L. D. (2009). *Conceptual Wavelets in Digital Signal Processing* (1^e éd.). Space & Signals Technical Publishing.
- Ganesan, R., Das, T. K., & Venkataraman, V. (2004). Wavelet-based multiscale statistical process monitoring: A literature review. *IIE Transactions*, 36(9), 787-806. <https://doi.org/10.1080/07408170490473060>
- Ge, X., Fan, Y., Li, J., Wang, Y., & Deng, S. (2015). Noise reduction of nuclear magnetic resonance (NMR) transversal data using improved wavelet transform and exponentially weighted moving average (EWMA). *Journal of Magnetic Resonance*, 251, 71-83. <https://doi.org/10.1016/j.jmr.2014.11.018>
- Harkat, S., Boukharouba, K., & Douaoui, A. E. K. (2016). Multi-site modeling and prediction of annual and monthly precipitation in the watershed of Cheliff (Algeria). *Desalination and Water Treatment*, 57(13), 5959-5970. <https://doi.org/10.1080/19443994.2014.956798>
- Harrou, F., Taghezouit, B., & Sun, Y. (2018). *A robust monitoring technique for fault detection in grid-connected PV plants*. 7th International IEEE Conference on Renewable Energy Research and Applications, ICRERA 2018, October 14, 2018 - October 17, 2018, Paris, France (p. 594-598). <https://doi.org/10.1109/ICRERA.2018.8566768>

- Hasan, A. M., Samsudin, K., & Ramli, A. R. (2011). Intelligently tuned wavelet parameters for GPS/INS error estimation. *International Journal of Automation and Computing*, 8(4), 411-420. <https://doi.org/10.1007/s11633-011-0598-9>
- Ibrahim, R. K., Watson, S. J., Djurović, S., & Crabtree, C. J. (2018). An Effective Approach for Rotor Electrical Asymmetry Detection in Wind Turbine DFIGs. *IEEE Transactions on Industrial Electronics*, 65(11), 8872-8881. <https://doi.org/10.1109/TIE.2018.2811373>
- Jiang, T., Chen, B., He, X., & Stuart, P. (2003). Application of steady-state detection method based on wavelet transform. *Computers & Chemical Engineering*, 27(4), 569-578. [https://doi.org/10.1016/s0098-1354\(02\)00235-1](https://doi.org/10.1016/s0098-1354(02)00235-1)
- Katunin, A., & Przystalka, P. (2015). Automated wavelet-based damage identification in sandwich structures using modal curvatures. *Journal of Vibroengineering*, 17, 2977-2986.
- Kehtarnavaz, N. (2008). Frequency Domain Processing. Dans N. Kehtarnavaz (édit.), *Digital Signal Processing System Design* (2^e éd., p. 175-196). Academic Press.
- Kim, Y., & Bang, H. (2018). Introduction to Kalman Filter and Its Applications. Dans *Kalman Filter*. InTechOpen.
- Korbel, M., Bellec, S., Jiang, T., & Stuart, P. (2014). Steady state identification for on-line data reconciliation based on wavelet transform and filtering. *Computers & Chemical Engineering*, 63, 206-218. <https://doi.org/10.1016/j.compchemeng.2014.02.003>
- Labbe Jr, R. R. (2020). *Kalman and Bayesian Filters in Python*.
- Liu, W., Chang, Q. R., Guo, M., Xing, D. X., & Yuan, Y. S. (2011). Extraction of first derivative spectrum features of soil organic matter via wavelet de-noising. *Guang Pu Xue Yu Guang Pu Fen Xi*, 31(1), 100-104.
- Ma'arif, A., Iswanto, Nuryono, A., & Alfian, R. (2020). Kalman Filter for Noise Reducer on Sensor Readings. *Signal and Image Processing Letters*, 1, 11-22. <https://doi.org/10.31763/simple.v1i2.2>
- MacGregor, J. F., Kozub, D. J., Penlidis, A., & Hamielec, A. E. (1986). State Estimation for Polymerization Reactors. *IFAC Proceedings Volumes*, 19(15), 147-152. [https://doi.org/https://doi.org/10.1016/S1474-6670\(17\)59414-8](https://doi.org/https://doi.org/10.1016/S1474-6670(17)59414-8)
- Mah, R. S. H., & Tamhane, A. C. (1982). Detection of gross errors in process data. *AIChE Journal*, 28(5), 828-830. <https://doi.org/10.1002/aic.690280519>
- Mallat, S. G. (2009). *A wavelet tour of signal processing: the Sparse way* (3^e éd.). Elsevier /Academic Press.
- Manhertz, G., Modok, D., & Á, B. (25-28 Sept. 2016 2016). *Evaluation of short-time fourier-transformation spectrograms derived from the vibration measurement of internal-combustion engines*. 2016 IEEE International Power Electronics and Motion Control Conference (PEMC) (p. 812-817). <https://doi.org/10.1109/EPEPEMC.2016.7752098>
- Martini, A., Coco, D., Sorce, A., Traverso, A., & Levorato, P. (2014). *Gross Error Detection Based on Serial Elimination: Applications to an Industrial Gas Turbine Volume 3A: Coal, Biomass and Alternative Fuels; Cycle Innovations; Electric Power; Industrial and Cogeneration*. <https://doi.org/10.1115/gt2014-26746>
- MathWorks. (2020). *Threshold selection for denoising*. MathWorks. <https://www.mathworks.com/help/wavelet/ref/thselect.html>
- MathWorks. (2023). *Short-Time Fourier Transform*. MathWorks. <https://www.mathworks.com/help/dsp/ref/dsp.stft.html>
- Mitra, S. K. K. (2000). *Digital Signal Processing: A Computer-Based Approach*. McGraw-Hill Higher Education.
- Muppalla, V., Suraj, N. S. S. K., Reddy, V. Y. S., & Suman, D. (15-17 Dec. 2017 2017). *Performance Evaluation of Different Denoising Techniques for Physiological Signals*. 2017 14th IEEE India Council International Conference (INDICON) (p. 1-6). <https://doi.org/10.1109/INDICON.2017.8487739>
- Narasimhan, S., & Jordache, C. (2000). *Data Reconciliation and Gross Error Detection: An Intelligent Use of Process Data*. Gulf Professional Publishing.
- Narayan, K., Mahesh, B., & Andreas, S. (2013). *An Introduction to Kalman Filtering with MATLAB Examples*. Morgan & Claypool.
- Nounou, M. N., & Bakshi, B. R. (1999). On-line multiscale filtering of random and gross errors without process models. *AIChE Journal*, 45(5), 1041-1058. <https://doi.org/10.1002/aic.690450513>

- Nounou, M. N., & Bakshi, B. R. (2000). Chapter 5 - Multiscale Methods for Denoising and Compression. Dans B. Walczak (édit.), *Data Handling in Science and Technology* (vol. 22, p. 119-150). Elsevier.
- Park, S.-M., Lee, T.-J., & Sim, K.-B. (2016). Heuristic feature extraction method for BCI with harmony search and discrete wavelet transform. *International Journal of Control, Automation and Systems*, 14(6), 1582-1587. <https://doi.org/10.1007/s12555-016-0031-9>
- Park, S., Gil, M.-S., Im, H., & Moon, Y.-S. (2019). Measurement Noise Recommendation for Efficient Kalman Filtering over a Large Amount of Sensor Data. *Sensors (Basel, Switzerland)*, 19(5), 1168. <https://doi.org/10.3390/s19051168>
- Perez, E., & Barros, J. (15-18 Aug. 2006 2006). *Voltage Event Detection and Characterization Methods: A Comparative Study*. 2006 IEEE/PES Transmission & Distribution Conference and Exposition: Latin America (p. 1-6). <https://doi.org/10.1109/TDCLA.2006.311552>
- Pérez, E., & Barros, J. (2012). Application of Advanced Digital Signal Processing Tools for Analysis of Voltage Events in Power Systems. *International Journal of Electrical Engineering & Education*, 46(3), 211-224. <https://doi.org/10.7227/ijeee.46.3.1>
- Perry, M. (2010). The Exponentially Weighted Moving Average. Dans.
- Rahimi, F., Bustami, A., Hanif, M., Md Saad, M. H., Jailani, M., Mohd Nor, J., . . . Inayatullah, O. (2007). The Application Of Short Time Fourier Transform And Image Processing Techniques To Detect Human Heart Abnormalities.
- Rioul, O., & Vetterli, M. (1991). Wavelets and signal processing. *IEEE Signal Processing Magazine*, 8(4), 14-38. <https://doi.org/10.1109/79.91217>
- Saho, K., & Masugi, M. (2015). Automatic Parameter Setting Method for an Accurate Kalman Filter Tracker Using an Analytical Steady-State Performance Index. *IEEE Access*, 3, 1-1. <https://doi.org/10.1109/ACCESS.2015.2486766>
- Selesnick, I. (2009). The short-time Fourier transform and speech denoising. *Connexions*, <http://cnx.org/content/m32294/1.1/>.
- Shafri, H., & Yusof, R. (2009). Determination of optimal wavelet denoising parameters for red edge feature extraction from hyperspectral data. *Journal of Applied Remote Sensing*, 3. <https://doi.org/10.1117/1.3155804>
- Simon, D. (2001). Kalman Filtering. *Embedded Systems Programming*.
- Smith III, J. O. (2011). *Spectral Audio Signal Processing*. Stanford University.
- Smith, J. O. (2011). *Spectral Audio Signal Processing*. Center for Computer Research in Music and Acoustics (CCRMA). <https://ccrma.stanford.edu/~jos/sasp/>
- Smith, T. H., & Boning, D. S. (1997). A self-tuning EWMA controller utilizing artificial neural network function approximation techniques. *IEEE Transactions on Components, Packaging, and Manufacturing Technology: Part C*, 20(2), 121-132. <https://doi.org/10.1109/3476.622882>
- Sun, S., Huang, D., & Gong, Y. (2011). Gross Error Detection and Data Reconciliation using Historical Data. *Procedia Engineering*, 15, 55-59. <https://doi.org/10.1016/j.proeng.2011.08.012>
- Taspinar, A. (2018). *A guide for using the Wavelet Transform in Machine Learning*. <https://ataspinar.com/2018/12/21/a-guide-for-using-the-wavelet-transform-in-machine-learning/>
- Upadhyaya, B. R., Mehta, C., & Bayram, D. (2014). Integration of Time Series Modeling and Wavelet Transform for Monitoring Nuclear Plant Sensors. *IEEE Transactions on Nuclear Science*, 61(5), 2628-2635. <https://doi.org/10.1109/TNS.2014.2341035>
- Wang, Y., & Veluvolu, K. C. (2017). Time-Frequency Analysis of Non-Stationary Biological Signals with Sparse Linear Regression Based Fourier Linear Combiner. *Sensors (Basel, Switzerland)*, 17(6), 1386. <https://doi.org/10.3390/s17061386>
- Yinfeng, D., Yingmin, L., Mingkui, X., & Ming, L. (2008). Analysis of earthquake ground motions using an improved Hilbert–Huang transform. *Soil Dynamics and Earthquake Engineering*, 28(1), 7-19. <https://doi.org/https://doi.org/10.1016/j.soildyn.2007.05.002>

- Yuen, K.-V., Hoi, K.-I., & Mok, K.-M. (2007). Selection of noise parameters for Kalman filter. *Earthquake Engineering and Engineering Vibration*, 6(1), 49. <https://doi.org/10.1007/s11803-007-0659-9>
- Zabidi, A., Mansor, W., Lee, Y. K., & Fadzal, C. W. N. F. C. W. (11-12 Sept. 2012 2012). *Short-time Fourier Transform analysis of EEG signal generated during imagined writing*. 2012 International Conference on System Engineering and Technology (ICSET) (p. 1-4). <https://doi.org/10.1109/ICSEngT.2012.6339284>
- Zhao, Z., Särkkä, S., & Rad, A. B. (17-20 Sept. 2018 2018). *Spectro-temporal ECG analysis for atrial fibrillation detection*. 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP) (p. 1-6). <https://doi.org/10.1109/MLSP.2018.8517085>
- Zhou, D., Wang, Q. J., Tian, Q. J., Lin, Q. Z., & Fu, W. X. (2009). Wavelet analysis and its application in denoising the spectrum of hyperspectral image. *Guang Pu Xue Yu Guang Pu Fen Xi*, 29(7), 1941-1945.

ANNEXE D ARTICLE #3 SIGNAL PROCESSING HEURISTICS FOR THE SELECTION OF WAVELET TRANSFORM PARAMETERS CONSIDERING SIGNAL CHARACTERISTICS DERIVED FROM A CASE STUDY ON A CHEMICAL PULP MILL

Émilie Thibault, Francis Lebreux Désilets, Bruno Poulin, Moncef Chioua, Paul Stuart
Submitted to Computers and Chemical Engineering on 06/02/2023

Abstract

Clean and noiseless data is necessary to improve process plants operation and management decision-making. However, tuning filters parameters for processing industrial signals can be tedious, time-consuming and requires expertise in the field of signal processing and in the plant operation. To alleviate the downside of data processing, it must be approached systematically with scalable solutions. Using a parametric sweep, optimal wavelet parameters, i.e., those that reduce unwanted frequencies, i.e. frequencies higher than the one of interest, while retaining useful information in the signals, are identified for hundred of signals. Those parameters are combined with the signal's characteristics (measurement type, fluid nature, sensor type and process configuration) and features (variance, rate of change, data distribution statistics, coefficient of variation, and frequency) in a Principal Component Analysis to assess their correlation and relationship. As a result, a set of explicit heuristics to select signal processing tuning parameters for denoising industrial process signals is proposed.

Keywords Signal Processing, Noise, Wavelet transforms, Heuristics, Process data, Pulp Mill

Introduction

Real-time data available from sensors in today's pulp mills are abundant and contain the knowledge that can help plants with their decision-making processes, monitoring, and optimization. Pulp mills are complex industrial processes that generate multiscale signals due to the variety of events that occur at different times and frequencies. These process signals are affected by the quality of the sensors, the process environment and are bound to be mixed with high-frequency noise (Yang et al., 2020). Noise in process sensor signal may be assimilated to any unwanted interference that degrades the signal; its reduction is one of the most difficult challenges in signal processing (McSharry et al., 2003). This noise directly affects the accuracy of the control system, performance evaluation, fault diagnosis and the reliability of extracted data used in detecting operating state. Ultimately, data sets denoising is necessary to improve performance, accuracy and prediction robustness (Chiang et al., 2017).

To improve decision-making based on information from plant sensors, and for process data to reflect characteristics of the original data properly, they must be cleaned and then interpreted. However, given that there are thousands of signals in a mill, tuning filters parameters to process industrial signals may be tedious and time-consuming as it is often done manually by data scientists or process experts. Since it is practically infeasible to tune them one at the time, data processing must be approached in such a way to be able to deal with large volume of data. In addition, even if the process of reducing undesirable components is critical, it is not a step that brings tremendous value by itself. It is a pre-requisite step, something to consider before any further added-value analysis such as optimization, cost analysis, process simulation, etc. Therefore, to expedite the signal processing procedure, there is a need to systematically select the values of signal processing parameters. Heuristics could tune the filter parameters for industrial signals processing to save time and reduce process expert intervention.

The objective of this paper is to propose a structured way to process large-scale industrial data by defining heuristics based on wavelet transform parameters and signal characteristics that will standardize this process. Those heuristics provide an efficient and practical way to obtain plant-wide processed data that could be used for a wide variety of application and decision-making process. In order to obtain those heuristics, many signals are processed using the Wavelet transforms (WT) as it is shown to be quite performant for industrial signal processing (Thibault et al., 2023). Then, the signal's WT filtering parameters as well as the data that reflect the signal's characteristics are incorporated in a multivariate correlation analysis – Principal Component Analysis (PCA). This qualitative analysis is performed to identify correlations between signals characteristics and WT parameters.

This paper is organized as follows. Section 2 reviews WT parameters selection methods proposed in the literature while section 3 focuses on papers on signal characterization and features extraction. The proposed methodology is detailed in Section 4. The systematic methodology applied in this paper for the analysis of process signals lead to formulation of heuristics for the automatic treatment of errors contained in measurements data from sensors. Section 5 presents a case study based on a dissolving pulp mill. It includes the selection and classification of sensors according to the measurement type (temperature, level, flow, consistency) and nature (vapor, pulp, white liquor, weak black liquor, and strong black liquor), the description of the process configuration around the sensor and the raw signals features extraction. Then, section 5.4 presents the parametric sweep used for the identification of optimal denoising wavelet parameters. Section 5.5 explains the Principal Component Analysis results and section 5.6 presents the derived heuristics. Section 6 refers to the discussion and results validation. Section 7 concludes the paper.

State-of-the-art on the selection of wavelet transforms parameters

Sensor measurements representing an industrial process are generally tainted with noise which may mask features and/or numerous other events. In most applications, the multiple events in an industrial process occur at various frequencies (scales) and are localized at different points in time. In other words, inherent process variations are often multiscale with different features localized in time and frequency. This would imply that it is essential that the signal trend be extracted from noisy data according to the application before going forward. Among other approaches, Wavelet transforms are known to provide an appropriate mean to examine multiscale events as it has good resolution both in time and frequency. This approach breaks down a signal into its low and high frequency components to analyze them. It reduces the high-frequency components that represents the signal noise and reconstructs the low frequency components that reflects the basic trend of the signal.

When performing signal processing using wavelet transform, users must select various parameters including the type of wavelet function and its order (vanishing moment), level of decomposition, denoising method and threshold rule as they strongly affects the denoising performance. To the authors best knowledge, there is no universal rule for the selection of these parameters since they are problem-dependent (Katunin & Przyszałka, 2015). Although the efficiency of noise reduction greatly depends on the selection of the filtering parameters, determining WT parameters may be ambiguous and not very practical. In fact, given the number of parameters, WT are more difficult to implement than other filters, heavy to tune and to maintain. This is particularly true for industrial signals as they are not periodic, they are prone to outliers, shutdowns, and abnormal events. Analyzing the current state of the art, it can be noted that there is no recommendation of which wavelet function should be used to analyze industrial process signals, to which level of decomposition this analysis should be carried out and which thresholding rule should be applied. The following paragraphs elaborates on how the WT parameters are tuned in various applications.

All fields combined, not only related to chemical engineering and industrial processes, few papers used algorithms (mostly the Genetic Algorithm) to select optimal wavelet denoising parameters to maximize filtering performance

(Abbaszadeh & Alipour, 2017; Cao et al., 2022; El-Dahshan, 2011; Hasan et al., 2011; Katunin & Przyszałka, 2015), others rely on empirical methods to select the wavelet parameters (Liu, W. et al., 2011; Shafri & Yusof, 2009; Zhou et al., 2009), whereas others use a trial and error approach (Tan et al., 2008) or personal experience (Smallwood & Cap, 1998). In Wang, P. et al. (2021), the authors determined the optimal WT parameters by comparing the Pearson correlation coefficient between the corrected shock response spectrum and a low-frequency spectrum baseline at different decomposition levels and vanishing moments. It is found that Db7 with seven-level decomposition is the optimal combination to process shock test signals as the correlation coefficient reached 0.9995.

Hasan et al. (2011) presents an algorithm for denoising global positioning system and inertial navigation system data. The selection of wavelet parameters is vital in de-noising this kind of data. Hasan et al. (2011) tested a range of wavelet parameters in their optimization algorithm: the types of wavelet function – Daubechies (Db2-Db10), Coiflets (Coif1-Coif5), Symlets (Sym2-Sym8) and Biorthogonal (Bior1.3-Bior6.8), the threshold selection rule – minimax, heursure, rigrsure, universal, and the level of decomposition – 1 to 10. By implementing the proposed algorithm, the following parameters are identified as the optimum parameters for the position in x, y and z axes: Wavelet Function Db3, Db4, Sym4, and Bior4.4, threshold universal and minimax and 5-6 as the level of decomposition.

Wavelet parameters selection directly impact the denoising performance, therefore some papers investigated what are the ideal parameters in specific applications. In the case of hyperspectral applications, Zhou et al. (2009) identified that the optimal WT denoising parameters for vegetation (lawn grass) spectrum in the United States Geological Survey spectrum library is the sym6 wavelet function, the heursure threshold selection rule, and seven-level decomposition. Liu, W. et al. (2011) showed that the sym8 wavelet function, the single level noise rescaling method, the heursure threshold selection rule and three-level decomposition produce the finest denoising performance for organic matter (soil) spectrum. Using synthetic signals, Shafri et Yusof (2009) indicated that the most suitable WT parameters to denoise palm oil reflectance signal are sym8 or Db9 wavelet function (after testing Db4-30, Sym4-15 and Coif1-5) and single level noise threshold rescaling factor when using the Universal threshold value, soft thresholding and five-level of decomposition. Li et al. (2012) looked for the parameters (threshold selection rule, threshold rescaling, level of decomposition and wavelet type) that would perform well for any single spectrum in the library using a normalized spectral angle criterion.

In applications of WT to Machine Learning (ML) based hydrological process modeling, the identification of an optimal set of wavelet parameters such as wavelet function type, vanishing moment, and maximum decomposition level is mostly performed using a trial-and-error approach (Abbaszadeh & Alipour, 2017). The downside to this process is that it is very time-consuming and doesn't guarantee whether the optimal wavelet parameters are selected. Abbaszadeh et Alipour (2017) developed a framework that helps choosing the most appropriate WT parameters (wavelet function type, vanishing moment and maximum decomposition level) specifically for hydrological process modeling. The parameters are chosen with the objective of enhancing the efficiency of the developed model (Abbaszadeh & Alipour, 2017). The authors also provided a meaningful relationship between the appropriate WT parameters and watershed data (rainfall and runoff) statistic characteristics, i.e. minimum, maximum, mean, variance. In this study, both daily and monthly data are assessed.

In the context of hydrological process modeling, the appropriate selection of parameters is capital as it leads to a well-calibrated ML-model. Although WT has been successfully used by researchers to pre-process hydrological time-series repeatedly, none of them presented how to identify these parameters altogether (Abbaszadeh & Alipour, 2017). The inappropriate choice of WT parameters may lead to a more computationally demanding ML-model and may decrease its efficiency (Abbaszadeh & Alipour, 2017). Therefore, to circumvent this problem, some researchers have used an equation to find the optimum decomposition level (Nourani et al., 2009; Nourani et al., 2011; Wang,

W. & Ding, 2003), however its usage is restricted to signals represented by autoregressive models for which the maximum decomposition level is only determined based on the signal length (Nourani et al., 2011). Furthermore, Abbaszadeh et al. mentioned the absence of specific research done to address the appropriate selection of wavelet function, another important factor that affects the efficiency of WT pre-processing hydrological series. The similarity in shape between the wavelet function and the raw signal has often provided a guideline in choosing a reliable wavelet function. Lastly, the wavelet function vanishing moment has received little attention by the researchers in developing hydrological models (Abbaszadeh & Alipour, 2017).

Abbaszadeh et al. considered four types of wavelet function – Daubechies, Coiflet, Haar, Symlet – with different vanishing moments and decomposition level. Based on their work, the most appropriate wavelet parameters should be determined using the following guidelines to capture the hydrologic characteristics :

1. Daubechies wavelet is the most suitable wavelet function to extract informative features of hydrological time series due to its similarity in shape to the original signals.
 2. The vanishing moment of this wavelet function should be selected based on the watershed behavior. For the chaotic watershed in general, db4 is recommended, while for the mild watershed, db1 is preferred.
 3. The decomposition level, on the other hand, should be determined based on the time resolution of hydrological time series. The daily time scale require level 8 of decomposition to converge to an optimal result, while the monthly time scale need level 4 of decomposition to achieve a desirable estimation.
- Using those levels, the periodicity of hydrologic series can be properly captured.

Alyasseri, Zaid Abdi Alkareem et al. (2018) stated that electrocardiogram signals, given their non-stationarity (as it is the case for industrial signals), frequently encounter noise and a low signal-to-noise ratio (SNR). In their paper, Wavelet transform is used for denoising, but the selection of this approach best combination of parameters is a challenging task because they are not determined by applying a given technique but rather by referring to empirical evidence or experience (Alyasseri, Zaid Abdi Alkareem et al., 2018). Therefore, using signals from a database that are widely used as benchmark for comparing the performance of different noise reduction techniques, the authors applied an algorithm to identify the optimal wavelet parameters that minimizes the mean square error between the original and denoised ECG signals. The algorithm's performance is evaluated by using percentage root mean square difference (PRMSD) and signal-to-noise ratio as criteria. The proposed method reduced white gaussian noise in non-stationary signals efficiently since the output signal obtain a high SNR (which correspond to a low noise level, a smooth signal) and low PRMSD (which indicates an efficient denoising of the original signal). The authors found that the optimal wavelet parameters to denoise electrocardiogram signals are sym7 wavelet function, decomposition level 2, soft thresholding function, rigrsure thresholding selection rule and single level rescaling approach.

Similarly, Sawant et Patil (2014) tested several wavelet parameters including db2 to db10 as a wavelet function, soft and hard thresholding function and Universal, Minimax, Heursure and Rigrsure as threshold selection rules, to find the best ones that can obtain a denoised electrocardiogram signal with a high SNR. They conducted their experiments on synthetic data and identified db5, level 4, soft thresholding with rigrsure threshold as the optimal parameters. Lastly, Mikhled et Khaled (2008) recommended using db4, level 5 and soft thresholding as denoising parameters to minimize the error between the original and noisy signals.

Therefore, for better performance, the choice of wavelet parameters should be tailored based on the different kind of signals being processed. However, optimal parameters are challenging to determine as they often require empirical knowledge about the system or process. Some publications identified the optimal combination of wavelet denoising parameters for GPS, geological, hydrological and electrocardiogram signals, but no equivalent work for industrial process sensor data is found in the literature (Table 1). Furthermore, the approaches employed as cited in the literature did not explicitly mention reducing high frequency components while making sure to retain pertinent

information according to a targeted application. Lastly, only Abbaszadeh et al. address the correlation between signals characteristics and WT parameters.

Hence, the objective of this paper is to assesses the correlation of WT tuning parameters with process signal characteristics and features to speed up signal processing of chemical process signals.

Table 1 – Census of papers that consider optimal WT parameters selection

Reference	Domain of application	Analyzed WT parameters	Approach employed to select the WT parameters	Correlation between signals characteristics and WT parameters
(Katunin & Przyszałka, 2015)	Composite structures	Wavelet function and its order	Multi-objective meta-optimization algorithm	No
(Hasan et al., 2011)	Global positioning system	Wavelet function, thresholding selection rule, level of decomposition	Genetic algorithm	No
(Cao et al., 2022)	Image processing	Number of layers, threshold, type of wavelet	Adaptive self-organizing migration algorithm	No
(El-Dahshan, 2011)	Electrocardiogram signals	Wavelet function, decomposition level	Genetic algorithm	No
(Abbaszadeh & Alipour, 2017)	Hydrological process modeling	Wavelet function, vanishing moment, decomposition level	Selected with the objective of enhancing the efficiency of the developed model	Yes
(Shafri & Yusof, 2009)	Hyperspectral reflectance data	Thresholding rules, wavelet function	Selected using synthetic data	No
(Tan et al., 2008)	Infrared spectroscopic analysis	Decomposition levels, wavelet function	Trial and error approach	No
(Wang, P. et al., 2021)	Shock response spectrum	Decomposition layers, vanishing moments	Selected using Pearson correlation coefficient	No
(Li et al., 2012)	Hyperspectral applications	Threshold selection rule, thresholding rules, multiplicative threshold rescaling, level of decomposition, wavelet type	Selected using the normalized spectral angle	No
(Alyasseri, Zaid Abdi Alkareem et al., 2018)	Electrocardiogram signals	Wavelet function, decomposition level, thresholding methods, thresholding selection rule, threshold rescaling methods	Selected by using minimum MSE between original and denoised ECG signals	No
(Sawant & Patil, 2014)	Electrocardiogram signals	Wavelet function, thresholding function, threshold selection rules	Selected using signal-to-noise ratio	No
This paper	Chemical process	Wavelet function, scale, thresholding function, threshold selection rule	Parametrical sweep	Yes

State-of-the-art signal characterization and features extraction

The parameters selection of pre-processing method parameters depends on the characteristics of the disturbance (noise), and on the signal itself (Dambros et al., 2018). Process signal characterization is addressed in some papers, particularly for vibration signals. Vibration signals are analyzed in the time domain using standard deviation and Kurtosis value as features and in the frequency and cepstrum domain using visual inspection (Zuo-wei et al., 2014). The amplitude and frequency of a vibration signal of a faulty power generator in a thermal-electric plant are analyzed using signal decomposition methods to extract the fault characteristics (Gao et al., 2008). In a unit of a hydropower station, a vibration signal is characterized by its frequency, amplitude, and its trend using waveform diagram and spectrogram (Ren et al., 2021). A water level signal is characterized by assessing the distribution, the power spectrum and the probability density function of the noise from the water level sensor whereas, vibration signals are characterized by the frequency distribution (Yang et al., 2020).

In the literature, signal characteristics and feature such as mean (García Plaza et al., 2018, 2019), standard deviation (dispersion of the data) (Fan et al., 2022; García Plaza et al., 2018, 2019; Malghan & Kumar Hota, 2022), average amplitude (strength of the signal) (Fan et al., 2022), Peak to peak amplitude (Fan et al., 2022; García Plaza et al., 2018, 2019), Energy (Fan et al., 2022; García Plaza et al., 2018), Shannon Entropy (Fan et al., 2022; García Plaza et al., 2018, 2019), Root Mean Square (Fan et al., 2022), Kurtosis (Fan et al., 2022; García Plaza et al., 2018, 2019), Skewness (Fan et al., 2022; García Plaza et al., 2018, 2019), Mean of Absolute Value (Fan et al., 2022), (Root) Mean Square Error (Malghan & Kumar Hota, 2022) and Signal-to-Noise Ratio (Malghan & Kumar Hota, 2022) are discussed in some papers. The SNR is often defined to investigate the noise characteristics associated with a signal (Liu, Huan et al., 2017; Tanaka, 1994).

Given the signals features, an assessment of the correlation between them and the WT parameters would highlight some rule of thumbs (guideline or heuristics) for a systematic, efficient, and structured tuning of the WT parameters. To the best of the authors knowledge, there is no mention of research that assessed the correlation of industrial process sensor signal characteristics or features with wavelet transform's parameters. However, the idea of a relation between signal characteristics and a processing technique parameters is not new. To filter nonstationary signals effectively, filter's parameters should be selected depending on the local characteristics of the signals (Harashima et al., 1982; Pomalaza-raez & McGillem, 1984; Rajala & Figueiredo, 1981). Arakawa (1994) proposed a signal processing technique based on fuzzy rules for estimating nonstationary signals contaminated with additive random noises. For this filtering technique, fuzzy rules concerning the relationship between signal characteristics and filter design are applied to set the filter parameters, taking the local characteristics of the signal into consideration (Arakawa, 1994).

Nevertheless, filters parameters deduced from rules or guidelines cannot be employed unquestionably if the rules describing how to set them are expressed ambiguously. For instance, it is common knowledge that the window size of a filter aiming to smooth random noises must change adaptively depending on the frequency component of the raw input signal. However, it is difficult to distinguish a large window size from a small one by a single threshold. Another guideline, specifically for WT, states that the selected wavelet function should be similar to the input signal. These rules are not trivial to implement and take into consideration. Therefore, there is a need for more specific rules for filter tuning, and more specifically, the wavelet transform in this case.

Methodology : Signal processing through Heuristics Framework

In the present work, the identification of optimal WT parameters for noise reduction⁶ of industrial sensor signal is obtained by a parametrical sweep. The parametrical sweep test each combination of parameters and keep the combination that minimized an objective function minimizing the loss of pertinent information, retain information below the frequency of interest, and reduce unwanted high frequency component. The obtained wavelet transform parameters, combined with the signal's characteristics, are used to generate heuristics. These heuristics will select the denoising parameters and allow to process signals based on their inherent characteristics in an efficient and structured way. The steps required to complete this are presented in Figure 1.

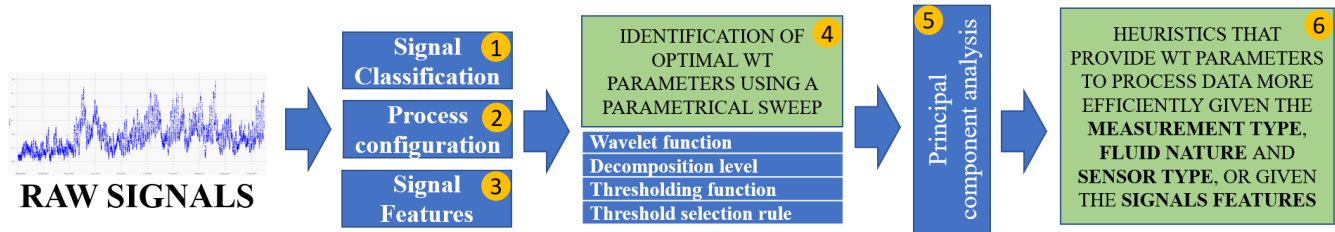


Figure 1 –Methodology employed to formulate heuristics for processing signals based on their features

STEP 1: Selection and classification of signals

The signals under study originate from sensors located in several departments of a dissolving pulp mill. Different measurement types such as temperature, level, consistency, and flow as well as different fluid nature such as vapor, pulp, weak and strong black liquor (BL), white liquor, etc. are considered. A total of 400 signals coming from 100 sensors are denoised using wavelet transform.

STEP 2: Description of process equipment configuration around each sensor

The physical environment of each sensor, i.e., pieces of equipment before and after the sensor, is considered. Succinct flow diagrams around the sensor related to the signal are drawn. They allow to see where the measurements are located in the process and therefore to better understand the impact of their locations. In addition to the process configuration, the sensor type is also considered, that is, whether the sensor is part of a control loop or solely an indicator.

STEP 3: Features extraction of industrial process signals

Process data are likely to be affected by the process configuration (pieces of equipment) around the sensor, the type of signal that is being measured and the nature of the fluid. Therefore, once signals are classified according to the type of measurement and fluid nature, features are extracted from these signals to evaluate their correlation. Four data sets of approximatively one month are used for each sensor. For each signal, stable operations from one operating regime (specific recipe), are targeted to isolate the effects of noise. Based on the literature review, the features considered in this analysis include the variance, rate of change, trend frequency, skewness, kurtosis and

⁶ The terms denoising, noise reduction, filtering and processing are used throughout the paper to designate the same thing.

coefficient of variation. The trend frequency refers to a spectral analysis of the signals performed by the Fast Fourier Transform (FFT) algorithm (Gao et al., 2008). It provides the most occurring frequencies, including the trend.

Variance (S^2) is a measure of dispersion, meaning that it measures how far values (x_i) are spread out from their average value ($\bar{x}$).

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Rate of change is a feature that captures locally the process dynamics. It shows how fast the process variable (x) changes with respect to time (t).

$$m = \frac{x(t_2) - x(t_1)}{t_2 - t_1} \quad (1)$$

As the rate of change is computed for each data point, the average value is considered to characterize the whole signal.

Measurements originating from a process system like the dissolving pulp mill under study need to be centered (around the mean μ) and scaled (given the standard deviation σ) to bring them on an equal basis. A common technique is standardization.

$$Z = \frac{x - \mu}{\sigma} \quad (2)$$

Trend frequency is obtained by computing a Fast Fourier Transform. The latter transforms signals from the time domain to the frequency domain, it measures the spectral (or frequency) content of the signals. Therefore, numerous signal characteristics can be investigated to a much greater extent than when limiting the inspection to the time domain. In the frequency domain, the process signal characteristics are described by independent frequency components. The output of a FFT analysis is a frequency spectra. A spectra indicated how much of each frequency is present in the signal. Frequency is plotted along the x-axis and amplitude is plotted along the y-axis. From those, it is possible to assess which frequencies are strongly present in signals, the dominant frequencies, that is those that are represented by a vertical peak of high amplitude. The location of the peak along the horizontal axis indicates the frequency whereas the height of the spike gives the amplitude, it indicates how much amplitude is located at each frequency. Complex signals have multiple peaks present on the FFT graph. Low frequency peaks appear at the left end of the graph; the left-most peak represents the fundamental trend.

The FFT analysis show which frequencies are present but lacks the information about when an event occurred within the signal. Therefore, to make a meaningful trend frequency analysis, the samples must be chosen carefully, i.e., the signals should remain reasonably constant throughout the entire sample. Consequently, the sample in the time-domain is chosen according to the fact that it is not changing much (the signal present relatively stable operations, with no step – change of set point, but still very much representative of the operations) while making it as long as possible since longer intervals improves the frequency resolution. The longer the selected interval, the more likely it is that the values will change during the selection.

Skewness is a measure of the distortion from a normal distribution, it characterises the asymmetry in a set of data. When a distribution lacks symmetry, it is said to be skewed. Skewness gives the direction of the outliers, whether they are more present right or left side of the distribution, however, it doesn't give any indication about the number of outliers.

Kurtosis, like skewness, is a statistical measure that is used to describe a distribution. More specifically, it describes the shape of a distribution's tail in relation to its overall shape, thus it assesses the "tailedness" by identifying whether the tails contain extreme value. In other words, the kurtosis defines how heavily the tails of a distribution

differ from those of a normal distribution. A distribution with a kurtosis value above five standard deviations from the mean will exhibit a tail that exceed the one of a normal distribution.

Coefficient of variation (CV) is defined as the ratio of the measured signal standard deviation to its mean: $CV = \sigma/\mu$. The result is dimensionless and thus, comparable between signals of different units. CV shows the extent of variability in relation to the mean of the sample, higher the value, greater the dispersion.

Having signals that are classified according to the nature and type of the measurement and the process configuration around the sensor and that their features are extracted, the following section address their processing, the identification of the optimal wavelet parameters, the correlation analysis, a discussion, and the results validation.

STEP 4: Identifying optimal wavelet transform parameters that provide the signal's trend

Optimal WT parameters are used to denoise 400 industrial process signals in order to make a predictive model of the trend of these variables. The best denoising performance are obtained when the shape of the wavelet function is similar to the shape of the signals, however manually assessing this is tedious and may be ambiguous. In this study, four types of wavelet function families including Daubechies (db), Symlet (sym), Coiflet (coif) and Biorthogonal (bior), with different vanishing moments and decomposition level, are analyzed in a parametrical sweep. The parametrical sweep is part of the proposed approach to formulate heuristics as shown in Figure 1.

Decomposition levels determine to what level the approximation (smooth component of the signal) and detail (high frequency component) coefficients are extracted. For instance, 4-level decomposition yields 5 sub signals: one approximation and details at level 1, 2, 3 and 4. Excessive decomposition of the signal may result in over smoothing and involve the risk of eliminating important components. On the other hand, insufficient denoising (inadequate decomposition) would result in a considerable portion of the noise to be retained in the reconstructed signal. Hence, the level of decomposition (scale) needs to be optimized to have the best quality of filtered signal. The level of decomposition is case-specific, it depends on the nature of the signal, the signal length, and the objective of the study, among other things.

Vanishing moments helps in generating smoother *features* of a signal. A higher value indicates that more moments are captures from the signal. Db1, which is known as the Haar Wavelet, is similar to a step function and is not well adapted to approximate and denoise smooth signals. Therefore, the Daubechies wavelet considered here are limited to db2 to db15 – that is Daubechies wavelet with 2 to 15 vanishing moments. This family of orthogonal functions are among the most popular ones and are a good choice for denoising signals (MathWorks, 2023a). The same can be said about the Symlet wavelet, which is a modified version of the Daubechies with increased symmetry. Another orthogonal wavelet, this one near symmetric, the Coiflet, is designed to have scaling function and wavelets with the same number of vanishing moments. All existing functions are studied here : coif1-coif5.

Different thresholding functions and selection rules are also compared as indicated in Table 2.

Table 2 – Ranges of commonly used wavelet denoising parameters

Wavelet Denoising Parameters	Range
Wavelet function	Daubechies (db2-db15); Symlet (sym2-sym15); Coiflet (coif1-coif5); Biorthogonal (bior1.1-bior6.8)
Decomposition level	1-10
Thresholding function	Soft threshold; Hard threshold
Threshold selection rule	Rigsure; Heursure; Universal; Minimax

It can be seen from Table 1 that there are 3 840 possible combinations, as all these parameters are independent from each other. In this regard, a parametric sweep coded in MATLAB is developed to select the best combination of parameters listed in Table 2 that provides the daily trend in the data. This data will then be used to make a predictive model for plant management decisions for which a granularity of one day is sufficient. Therefore, information every second would be considered as noise. The latter is a subjective concept that is application-dependent, it is unwanted high frequency component that do not contain any pertinent information for a specific application, i.e., noise consist of a faster variation than a studied phenomenon. Hence, solely information at the daily-level are kept – data are processed in order to be synchronized with what is of interest in this case, i.e., to match the frequency of what is being analyzed considering their initial frequency content. It is the daily information that is interesting, it is the reference, thus all the variations that are faster than a day are not needed.

All raw signals do not have the same frequency content, therefore the optimal parameters that allow retaining pertinent information, i.e., information related to targeted frequency (the frequency of interest), are obtained using a parametrical sweep based on the rejection of components at frequencies higher than one day (unwanted information). However, while smoothing the data to highlight the daily trend is important, it is critical to ensure that the signal is not oversmoothed and that useful information won't be lost. A rigorous criterion that allows adequate filtering is needed. Therefore, based on the frequency components of the raw and filtered signals issued by FFT analysis (line 3 in the pseudo-code), the optimal parameters are determined by applying an objective function (line 12) that minimizes the peaks amplitude of components whose frequency is higher than the analysis frequency (frequency of interest) (line 10) as well as the root-mean-square error (RMSE) between the peaks amplitude of components whose frequency is lower than the frequency of interest for the raw signal and the filtered signal (line 11). Figure 2 shows an example of a flow measurement processed according to different levels (left side) alongside the spectrum of each signal as well as the analysis frequency. In this case, the parametrical sweep selected level 5 which reduce the high frequency components while retaining the trend information (analysis frequency). On the other hand, higher level starts to loose critical information and show a shift according to the original signal.

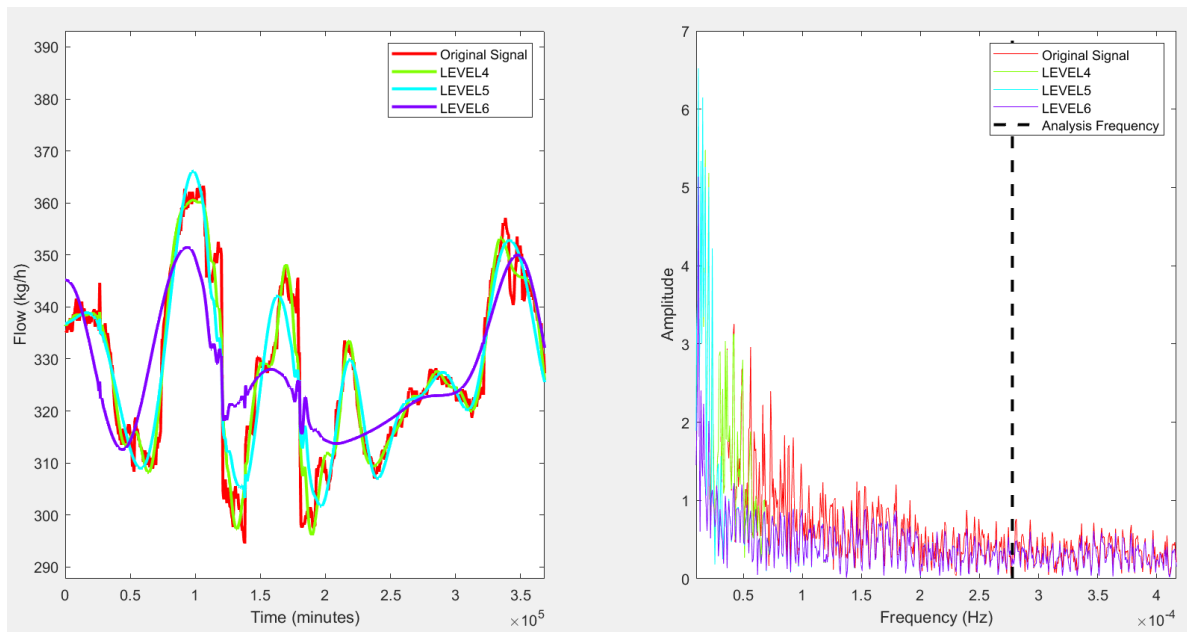


Figure 2 – Signals filtered with the different levels (left) with their FFT analysis (right)

The optimal wavelet transform parameters obtain as part of this parametrical sweep are used to filter signals for this specific application, i.e., prediction. Data processing is always performed according to a specific objective (it is objective-driven), which, in this case, is to obtain the daily trend in industrial data in order to predict it in the future. Industrial signals are used to build a trend prediction model for monitoring, optimization and decision-making.

Pseudo-code for the identification of optimal wavelet parameters for a trend prediction analysis

```

1 : for all signals  $i$ 
2 :     initialize the objective function value to infinite (stopping criteria)
3 :     compute the FFT of raw (input) signal  $i$ 
4 :     for all wavelet functions  $j$ 
5 :         for all level of decomposition  $k$ 
6 :             for all denoising method  $l$ 
7 :                 for all thresholding rule  $m$ 
8 :                     signal denoising with Wavelet Transform
9 :                     compute the FFT of filtered signal
10 :                     evaluate the sum of the amplitude for frequency > analysis frequency
11 :                     evaluate the RMSE of amplitude for the frequencies < analysis frequency
12 :                     evaluate the objective function value :  $OBJ = SAHF + RMSE$ 
13 :                     keep denoising parameters if  $OBJ_{jklm} < OBJ_{min}$ 
14 :                      $OBJ_{min} = OBJ_{jklm}$ 
15 :     return optimal wavelet parameters

```

STEP 5: Principal Component Analysis

The relations between signals classification parameters, features and identified wavelet transform parameters (Table 3) are assessed using a multivariate analysis (MVA). The multivariate analysis is a way to contextualize a large volume of data by exploiting the naturally occurring correlation structures. In other words, it allows for extracting interactions and useful correlations from dense data sets (Chiang et al., 2017). MVA can be applied in process monitoring to detect faults, it may predict important variables using available process data. Among the multivariate analysis approaches, projection-based methods such as Principal Component Analysis (PCA) is successful because of their effectiveness characterizing large, complex data sets (Qin, 2012). In addition, the PCA results are easy to understand, interpret and explain. PCA is used in the present work as it serves the objective of this paper, i.e., to establish heuristics to denoise signals based on the relations of signals' classification parameters, features and their optimal wavelet parameters.

PCA identifies and explain correlations and variability within the data and highlight the most influent variables. It is common to have correlated variables and redundant information in a data set. These variables may be combined to create new variables, fewer in number, thus reducing the initial size of the system, representing the data in a more compact form. These variables, called principal components, are independent of each other. They are linear combinations of the original variables. Therefore, beyond simple dimension reduction, PCA synthesize the information contained in all the variables; it reduces the size of large data matrices while retaining most of the variance (information). The 1st principal component is identified as the vector that can describe the greatest variance within the data set. The 2nd principal component is a vector that is orthogonal to the 1st, chosen so that it projects the greatest remaining variability, so on so forth. These components are ordered in descending magnitude of explained variance.

This descriptive analysis (PCA) identifies clusters of variables based on their similarities and differences. This can be assessed visually on the score and loading charts. The former shows the observations (data points of each

variable) and help in the determination of the various observation groups, whereas the latter shows the variables and reveals the correlations and importance of the variables. For instance, an important variable has a high loading value and variables that are link/correlated to it will have high values as well (in absolute value). In other words, variables that are grouped together are correlated – the reason for this correlation should be investigated and validated by process knowledge. Principal component analysis also gives the opportunity to assess the variable's contributions in a selected group of data.

Table 3 – Variables used in the Principal Component Analysis

Signal Classification	Signal's Features	Wavelet Parameters
Measurement type	Variance	Wavelet function
Measurement nature	Coefficient of variation	Vanishing Moments
Process configuration	Kurtosis	Decomposition level
Sensor type	Skewness	Thresholding function
	Trend Frequency	Threshold selection rule
	Rate of change	

STEP 6: Formulate heuristics based on principal component analysis

Heuristics enabled by the correlation between process signals classification parameters, features and optimal Wavelet Transform's Parameters derived from a parametrical sweep are formulated. The heuristics would reduce noise in process signals, in a more systematic, and structured way, before using them for a prediction analysis. Those heuristics select the right wavelet denoising parameters based on the signal's characteristics. Hence, the PCA highlight the optimal filter parameters based on the signal's characteristics as shown in Figure 1. The proposed heuristics that provide optimal denoising parameters are tested/validated.

Case study : Signal processing by heuristics extraction for different types of process signals in a dissolving Pulp Mill

In this case study, industrial signals from a dissolving pulp mill are examined to identify which wavelet transform parameters are optimal. Then, heuristics are formulated in order to minimize engineering time for data processing and make filter tuning easier in the future, when processing any new signals from the mill. A total of 100 sensors measuring levels, flows, temperatures and consistency all over the mill are analyzed. Sensor data used in this case study are stored in a database on a 10-minutes average basis.

As part of this analysis, anomalous observations are not under advisement. Therefore, data are pre-processed to reflect noisy data in normal circumstances, that is, by removing low production periods, spikes (outliers), start-ups, and shutdowns in the signals.

The following sub-sections present the results of the 6-steps methodology described in section 4.

5.1 Selection and classification of sensors

The 400 signals from a dissolving pulp mill – more specifically from the digester, washing, bleaching and machine department – are classified according to their type : temperature (188), flow (112), consistency (36) and level (64) and their nature : pulp, weak and strong black liquor (BL), white and green liquor and vapor.

5.2 Description of process equipment configuration around each sensor

In addition to the type and nature of signals, the process configuration, i.e., the pieces of equipment around each measurement instrument are assessed as well as the type of sensor, i.e., whether the variable is controlled or not (only given by an indicator). The following types of equipment are considered : pump, heat exchanger, tank, filter, boiler, vapor collector, desuperheater, paper machine, dryer, washer, knoter, static mixer, primary sieve, purifier, bleaching tower, mixer and convey. The process equipment configuration includes the equipment located right before and after the sensor. Figure 3 presents a part of the flow diagram around a sensor (NIC423). In this case, prior to the sensor is a pump and a static mixer comes right after.

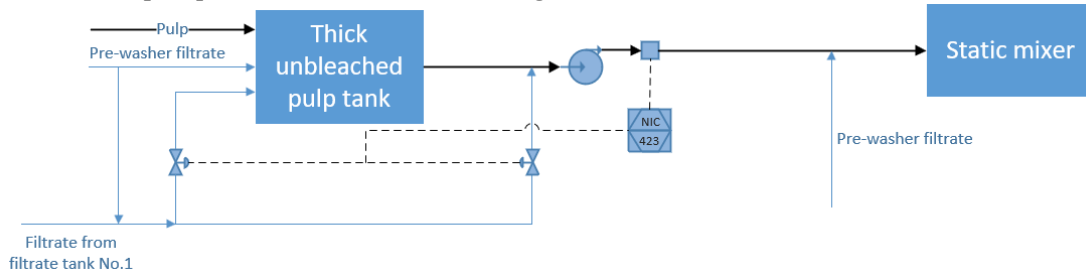


Figure 3 – Process configuration around a pulp consistency sensor in the washing department

Process equipment configuration around each sensor is relevant because the process sensor data are likely to be affected by the process environment. For instance, the vibration of a pump near a sensor could cause additional noise in a signal whereas a tank right before one might smooth or attenuate the variations in a signal. Regarding the equipment after a sensor, it might affect the way a signal looks in the way that depending on which equipment it is, it may require a more precise control. For instance, a tank will absorb variations whereas a paper machine needs a precise input flow to always have a sheet that is conform and identical from time to time. Therefore, this might impact the control loop, hence, the way the signal look.

The process configuration is not considered for signals coming from sensors that are placed directly on tanks – temperature and level.

5.3 Features extraction of industrial dissolving pulp mill signals

The features of the 400 signals – variance, rate of change, trend frequency, skewness, kurtosis, coefficient of variation – are computed as described in Section 4 (step 3). Comments on how these features are informative for this case study are displayed in sub-section 5.5, the principal component analysis results.

5.4 Identification of optimal denoising wavelet parameters of industrial dissolving pulp mill signals

The identification of optimal WT parameters values to obtain the daily trend for each signal is the first step towards the establishment of heuristics that could manage the treatment of a large amount of data. Table 4 shows a sample of the optimal wavelet parameters selected using the proposed parametrical sweep (Section 4, step 4).

Table 4 – Optimal Wavelet denoising parameters obtained by the parametrical sweep

Fluid Nature	Measurement Type	Type of sensor	Process conf. : Before	Process conf. : After	Wavelet Function	Decomp. Level	Thresh. Rule	Thresh. Function
Pulp	Consistency	Controller	Pump	Primary Sieve	db14	7	Universal Threshold	Soft
Pulp	Consistency	Controller	Static Mixer	Mixer	db15	8	Universal Threshold	Soft
Pulp	Consistency	Controller	Pump	Tank	db13	7	Universal Threshold	Soft
Pulp	Flow	Controller	Knotter	Washer	db15	6	Universal Threshold	Soft
Weak BL	Flow	Controller	Pump	Tank	db15	6	Universal Threshold	Soft
White Liquor	Flow	Indicator	Pump	Filter	sym9	6	Universal Threshold	Soft
Pulp	Level	Indicator	NA*	NA*	db12	7	Universal Threshold	Soft
Strong BL	Level	Indicator	NA*	NA*	db12	8	Universal Threshold	Soft
Green Liquor	Level	Indicator	NA*	NA*	db12	8	Universal Threshold	Soft
Pulp	Temperature	Indicator	Pump	Tank	db15	6	Universal Threshold	Soft
345 kPa Vapor	Temperature	Controller	Desuperheater	Heat Exchanger	db12	6	Universal Threshold	Soft
3 100 kPa Vapor	Temperature	Controller	Desuperheater	Tank	db15	6	Universal Threshold	Soft

*See Section 5.2

Soft thresholding function is selected 390 times out of 400. This may be explained by the fact that hard thresholding results in larger variance after reconstruction and can very well represent peaks in the signal, but since the signal's trend is what matter here, this is not applicable. Additionally, although hard thresholding is a simpler approach, it can produce discontinuities in the signals (El-Dahshan, 2011). Soft thresholding has a larger bias; however it gives good visual quality of filtering. Therefore, the objective function used in the algorithm seemed to have prioritized soft thresholding.

For most signals (398 out of 400), the optimal threshold selection rule is the Universal threshold as it reduces noise more efficiently, whereas the others are more conservative and are more convenient when small high frequency content of the signal lie near the noise range (MathWorks, 2020). Minimax is selected 2 times.

For the majority of the signals (51%, 204/400), the decomposition level is 6, whereas 32% are at 7, 5% at 8, 2% at 9 and the signals remaining have a decomposition level of 10. High decomposition level values can be explained by the fact that only the daily trend is targeted in the case of this prediction analysis and the decomposition level is defined according to the noise level – noise being all unwanted high frequency component. On the other hand, for comparison purposes, when aiming for the hourly trend, the parametrical sweep prioritized decomposition levels between 2 and 4.

Coiflet and biorthogonal wavelet functions are not selected by the parametrical sweep. Symlet are chosen close to 5% of the time, in contrast to 95% for Daubechies. This is coherent with the fact that orthogonal wavelets are a good choice for denoising signals and that Daubechies and Symlet can facilitate a perfect yet simple reconstruction of the original signal (Donoho, D. L., 1995; Mallat, S. G., 2009).

5.5 PCA : Signal classification parameters, features and optimal wavelet parameters

The optimal WT parameters, as well as the signal's classification parameters and features are studied in a principal component analysis to extract relevant correlations between them. The results of the multivariate analysis are discussed in the next section.

The principal component analysis starts by analyzing the signal's classification parameters against one feature (variance, rate of change, trend frequency, skewness, kurtosis, coefficient of variation) at a time to identify and explain its variability as well as variables that influence the feature under study. The following observations are made :

1. Temperature signals show low coefficient of variation values whereas level and flow signals show high values (Figure 4).
2. Flow and consistency signals present high variance values.
3. Flow signals appear to have high kurtosis values.
4. Level signals present high skewness values, while temperature signals present low skewness values.
5. Nothing significant can be observed related to the trend frequency and the rate of change.

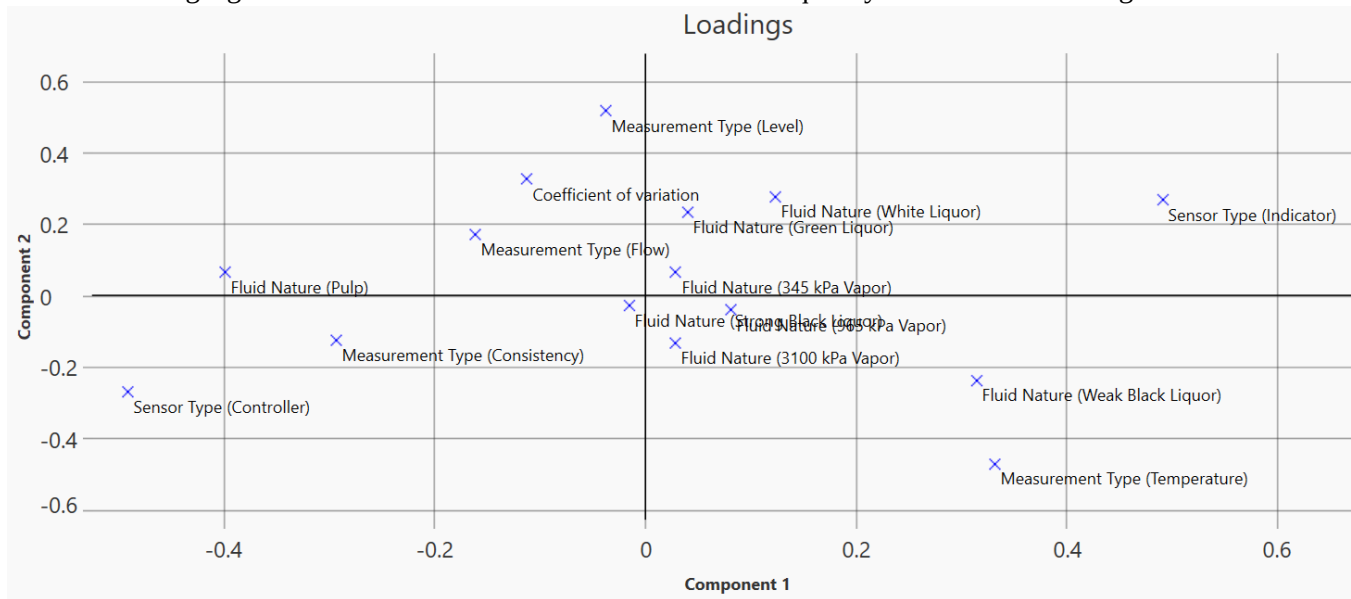


Figure 4 – Loading chart for the first and second component

The loading chart of the first and second component in the case of the coefficient of variation is presented in Figure 4. Variables that are close to each other in the loading chart show similarities (they are correlated with one another), variables that are in the same location in the opposite quadrant are inversely proportional and the rest are not related. Figure 4 displays that the level and flow measurement types are in its near vicinity, meaning that they vary proportionally. Oppositely, the temperature measurement type is inversely proportional. Therefore, the temperature's signals mean is more important that its standard deviation. This is explained by thermal inertia; temperature signal presents slower variation. Then, in the second statement flow and consistency signals are found to be consisted of values that are spread out far from their average values, that is, compared to level and temperature. This is justified by the fact that they show more variations – consistency measurements usually present less variation than flow measurements. The high kurtosis values of flow signals imply a higher probability of presence of extreme values – more extreme than the usual $3\sigma \pm \mu$ that is predicted by normal distribution. In the opposite manner, distributions with low kurtosis exhibit generally less extreme tail data than those of a normal distribution. Flow signals are highly affected and responsive to events happening in the process, and therefore are more prone to outliers. Lastly, high positive skewness values of level signals indicates that the extreme data values are larger, which is characterize by an important mean value for the dataset and that leads to a left-skewed distribution. On the other hand, temperature signals low (negative) skewness value indicates that extreme data values are smaller, the mean decrease and the distribution have a tail on its left side, therefore, it is a right-skewed distribution.

Therefore, the signals have distinct characteristics depending on their measurement type. However, the fact that the variance is not unitless might skewed the results; the flow units are not comparable to the level units, and if the units change, the results might change as well. As using standardized signals to assess the variance is pointless, unitless

features or features on the same basis should be preconized, i.e., rate of change, trend frequency (Hz), skewness, kurtosis, coefficient of variation.

Considering signal's features and WT parameters, the following observations are made :

1. High coefficient of variation are often processed using db9, scale 8 whereas high rate of change rely often on sym7.
2. Signals having a high trend frequency are related to db15 and sym15, and those having a high kurtosis are using db12, scale 9.
3. High variance values are correlated with db15.
4. Nothing conclusive can be observed in regard to the skewness.

From these observations, it is possible to conclude that signals presenting important variations, changes or major events are characterize by higher vanishing moments and scale. To assess further the relations with the wavelet transform parameters, the signal classification and wavelet transform parameters relationships are investigated. The score chart of these variables is considered to observe the presence of clusters (Figure 5). Furthermore, the variable' contributions are explored (Figure 6); variables that present a high contribution value (absolute value) explain the variability among the variables the most.

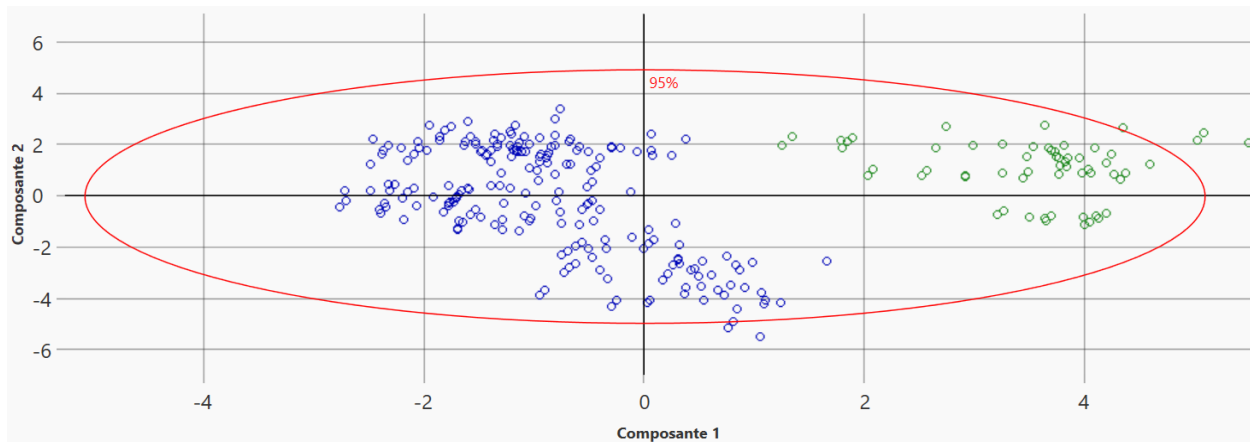


Figure 5 – Scores chart for all signal classification and wavelet transform parameters

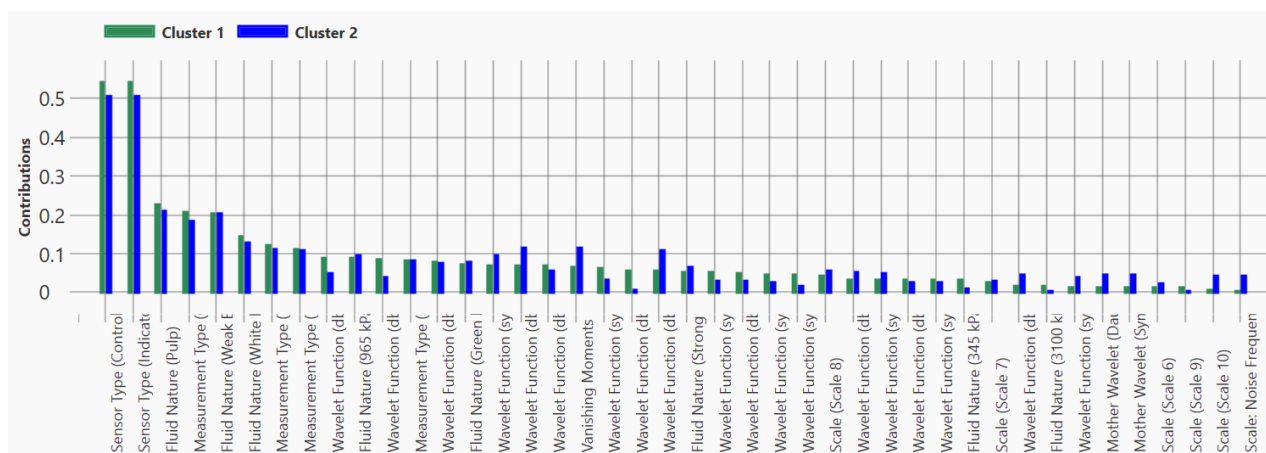


Figure 6 – Contributions of the clusters present in the scores chart

The score chart of the signal classification and wavelet transform parameters reveals the presences of two distinct clusters (Figure 5) that can be explained by Figure 6. A contribution analysis (Figure 6) showed that whether a sensor is part of a control loop or solely an indicator explain largely this phenomenon. It is observed that the variance is higher when the sensor is an indicator. Therefore, a sensor that is not controlled will provide a signal with more important variation – that is to be expected since a controller is used to reduce the variation. In other words, signal variations are highly influenced by whether the sensor is used as an indicator or in a control loop. The sensor's precision, reliability and accuracy could be an interesting addition to this analysis. As the characteristics of the measurement instrument would influence the high frequency component of the signal, if the data are not stored on a 10-minutes average basis, it could have been possible to assess the impact of those characteristics on the optimal wavelet parameters selection. This paper considers the type of sensor, that is indicator and controller, but for the latter, the type of controller, i.e., PID or PI could probably influence the signals dynamics as well. Also, whether the signal represents a controlled or manipulated variable might be of interest.

Now, focusing only on sensors in control loops, it is possible to detect that the score graph presents clusters (Figure 7) that are mostly explain by the measurement type (flow, temperature, consistency and level). That could be explained by the fact that different measurement type have different impact on the noise level which affect the signal as their dynamic is different. For instance, because of thermal inertia, temperature signal presents slower variation (they have lower frequency values) than flow signals; flows vary faster and present more high frequency components (noise) than temperatures. Then, considering solely the flow measurements, conclusions can be drawn onto which wavelet parameters is used for each fluid nature. The same can be archive for all measurement type. Those conclusions are presented in sub-section 5.6.

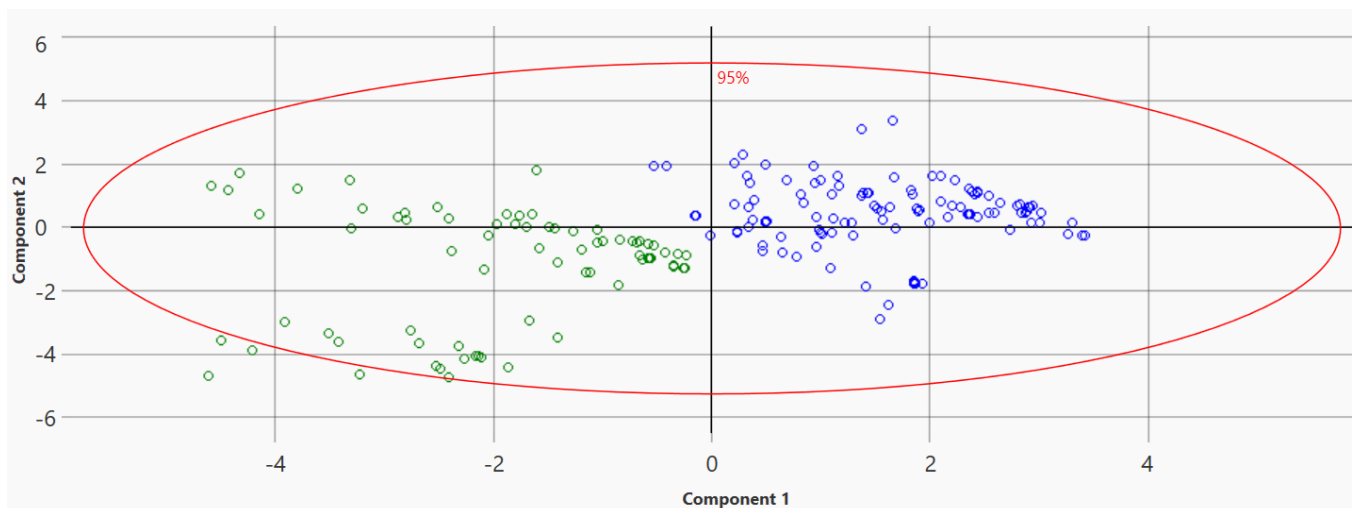


Figure 7 - Scores chart considering only measurements from sensors in control loops

In addition to the whether the sensor is an indicator or used in a control loop, sensor type, measurement type and fluid nature, the process configuration is considered. The purpose of the process configuration analysis is to identify if it has an influence on the signal's features and/or on the wavelet parameters. In this analysis, high level process configuration is accounted for whereas small process elements including valves, elbow, pipe restriction (change in diameter), splitter, etc. are not considered. Those elements might as well influence the signal's response and should be assessed in future analysis. The distance between pieces of equipment and the sensor might also be something of interest. All previously mentioned aspects are more difficult to assess as they are not visible in P & IDs, which is the only accessible information for the authors. Another paper could work on isolating completely each

component and assess its impact on each signal. Furthermore, there might be other aspect that could influence the way a signal look like, this paper does not present an exhaustive list in that matter.

Ultimately, it was found that isolating the impact of the equipment surrounding sensors is not trivial and requires redundancy, i.e. many signals with identical classification parameters for which only one element of the process configuration varied. Considering the lack of redundancy in the 400 signals identified through the principal component analysis, process configuration could not be assessed properly, and further analysis should be undertaken to evaluate this aspect. Hence, the equipment before and after the sensor won't be a part of the heuristics formulated in this paper.

Sensor type, measurement type and fluid nature are identified as the key factors that influence the way the signals are processed (Figures 6 and 7). It is found that signals from a specified sensor type such as controller, of specific measurement type, such as temperature, and measuring a specific fluid, pulp for example, had common characteristics. The next section presents the heuristics formulated based on the sensor type, measurement type and fluid nature.

5.6 Derived heuristics

The principal component analysis highlighted the fact that the sensor type, measurement type and the fluid nature are critical to the processing parameters; since the signals are characterized differently based on that, they must be process accordingly. The results of the case study are generalized and presented in the following heuristics. These heuristics capture the correlations between the signal's classification characteristics and the filtering parameters.

To emphasize the importance of context when extracting heuristics, the context from which they came from is well established; heuristics are presented by pairing a context in which they are applicable to an action that may be taken, a context-action form. In this case, the action is denoising signals. Outside this context, the heuristic might not be applicable, there might be some physical phenomena or constraints that prevent the generalization and extension of those heuristics to other mills, but the methodology is generic and is applicable elsewhere.

Heuristics derived as part of this research would guide process expert in the signal processing procedure; they could be relayed to mill personnel to be used as a starting point in their analysis, to improve the training processes, and shorten the learning curve regarding signal processing. Ultimately, heuristics are a means for fast, inexpensive guidance towards promising and valuable actions; they reduce the complexity of tasks.

Heuristics are extracted to include the context in which they are applicable followed by a suggested action to take. Therefore, in the context of industrial process data from a dissolving pulp mill,

1. the soft thresholding function should be considered by default for noise reduction.
2. the universal threshold selection rule should be considered by default for noise reduction.
3. the classification tree in Figure 7 should be used to select the wavelet function and decomposition level.

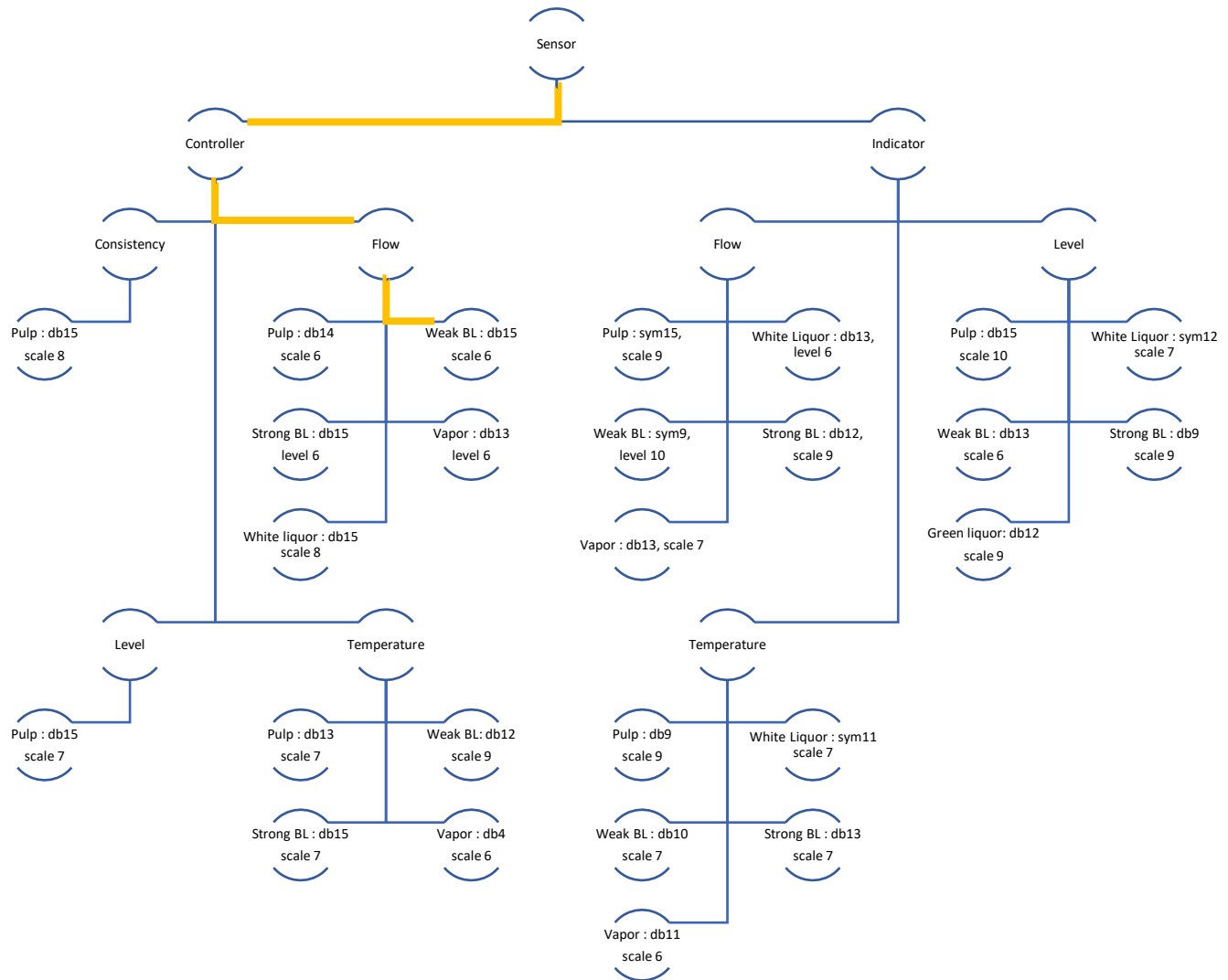


Figure 8 – Optimal Wavelet Parameters according to the signal's classification characteristics

Figure 8 summarizes which are the optimal wavelet parameters according to the combination of attributes sensor type/measurement type/fluid nature. Sub-section 5.5 emphasized that the measurement type, fluid nature and sensor type impacted the choice of wavelet parameters. For instance, considering the level indicators for the various fluid nature, the fact that the scale (decomposition level) is higher for the pulp then for the other fluids can be explained by the fact that, in general in dissolving pulp processes, pulp tank levels vary greatly, whereas green, white, and weak black levels are steadier.

The three heuristics formulated as part of this case study provide a quick approach to determine optimal WT parameters that could be used at a large scale, and therefore avoid tedious manual tuning of thousands of industrial signals. These signal processing heuristics are formulated for a specific set of data (they exist within a particular context); they are case-specific heuristics, which is, in this case, a dissolving pulp mill. Therefore, the relationships identified here have only been verified for a specific dataset (for a specific mill). Therefore, the same analysis should be done with data collected from another dissolving pulp mill to compare the conclusions and assess whether or not they are all case-specific. In other terms, going from an instance of an action choice (the choice of a filter's parameters for this dissolving pulp mill) to a heuristic that could be used in different plants require generalization.

Even if the heuristics might be case-specific, since a parametrical sweep is used to assess the optimal wavelet parameters, they are not user specific. In this work, heuristics are articulated and deduced from observation; a systematic approach is developed and employed.

Discussion and results validation

The 400 signals are analyzed one at the time given different wavelet parameters to identify the optimal ones using a parametrical sweep. The optimal WT parameters for each combination of signal's classification parameter is captured in heuristics. Therefore, given a signal's classification characteristics, the signals should be processed in a specific way. Lastly, the consideration of the signal's features alongside the classification parameters provided insight, important information and a better understanding of the signals' behavior.

The proposed heuristics are validated by comparing the objective function value (line 12 of the pseudo-code) for the WT parameters obtained with the parametrical sweep and those suggested by the heuristics. The objective function minimizes the peaks amplitude of components whose frequency is higher than the analysis frequency as well as the RMSE between the peaks amplitude of components whose frequency is lower than the frequency of interest for the raw signal and the filtered signal. A total of 7 signals are considered in the validation analysis; one from each branch of Figure 7 (an example is represented in yellow on Figure 7).

Table 5 – Heuristics validation on seven signals

Signal's classification parameters	Parametrical sweep parameters	Parametrical sweep objective function	Heuristics objective function	% difference
Controller – Consistency – Pulp	db11, scale 7	0.77	0.86	10.11
Controller – Flow – Weak BL	db13, scale 6	719.15	755.31	4.79
Controller – Level – Pulp	db14, scale 7	38.90	39.22	0.81
Controller – Temperature – Strong BL	db15, scale 6	73.14	73.82	0.92
Indicator – Flow – White liquor	db14, scale 6	141.71	167.58	15.44
Indicator – Level – Green liquor	db13, scale 10	0.71	0.78	9.32
Indicator – Temperature – Vapor	db14, scale 6	78.07	85.51	8.70

From Table 5, it is possible to observe that the highest difference between the parametrical sweep objective function and the heuristics objective function occurs in the case of a white liquor flow indicator. Figure 9 shows the result of the usage of the heuristics presented in the previous section this sensor's signal. It shows the raw original signal from this sensor (blue), the filtered signal resulting from using the WT parameters given by the parametrical sweep (red), and the processed signal obtained using the parameters provided by the proposed heuristics (yellow). The parametrical sweep proposed to process white liquor flow signals from indicators using db14, scale 6, whereas the heuristics suggested db13, level 6. Therefore, the heuristics are one scale away from the parametrical sweep results. This is acceptable for the planned use of the filtered signals. Hence, the heuristics demonstrated an outstanding performance in reducing noise in non-stationary signals, and the quality of the output signal is deemed favorable for trend prediction. Given the fact that all other tested sensor displayed lower differences, this validation demonstrated the robustness of the generated heuristics and showed that the wavelet transform can effectively reduce noise in industrial process signals. Those heuristics laid some initial groundwork for efficient extraction of useful information for prediction analysis.

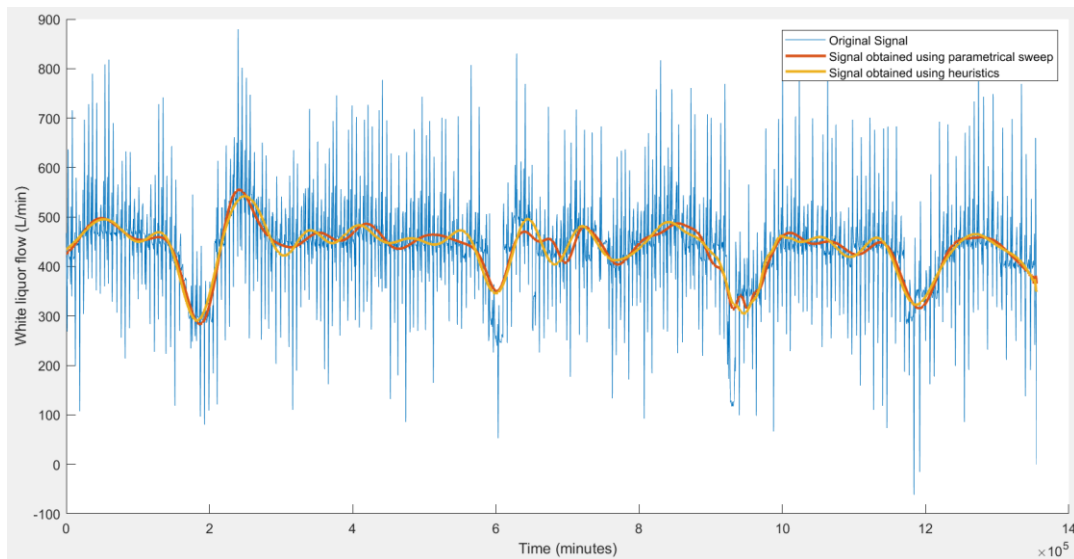


Figure 9 – Optimal wavelet parameters applied on new signals to assess their applicability

At last, in the same way that signal processing is objective-oriented, storage frequency depends on the objective of the analysis. The sampling frequency and storage should consider the process dynamic (how a process change in time). When a process dynamic is fast, averaged data over a short time step should be logged, whereas for a slow dynamic, a larger time step can be afforded; a 10-minutes average basis may be short for some analysis and too long for others. In some process equipment, the speed of a studied phenomenon (ex. a reaction) might be high and therefore the sampling and storage frequency would require adjustment. The same is true if a phenomenon is assessed over a long period of time; there is no need to sample every second.

As the objective of signal processing is to extract industrial signals' trend in a prediction context, taking the average over a 10-minutes interval of variables that are sampled at a frequency higher than 10 minutes is not problematic. In the case of variables that are sampled at a frequency lower than 10 minutes, the interpolation where removed and the FFT analysis helped identify the right frequency from which the trend could be seen. However, the results of this analysis are specific to the 10-minutes average storage frequency; it impacts the heuristics. The impact of the sampling frequency and the storage frequency on the optimal wavelet parameters could be gauged and elucidated in another paper.

Lastly, it would be interesting to survey and interview experts to validate and justify the identified parameters. Therefore, a hybrid methodology combining observations from the relations identified in the PCA and interviews with experts could address these limitations. Survey data collection methods gives the opportunity to archive empirically derived heuristics. Consequently, heuristics that may be corroborated by experts would be more refined than those based solely on mathematic approaches. Ultimately, heuristics may be considered more valuable if they could be implemented in a variety of contexts.

In conclusion, highlighting the relationships between the signal classification parameter, the signals features and the wavelet transform parameters provided a greater understanding of industrial signal processing and provided heuristics. Heuristics lead to actions using a minimum amount of time, knowledge, and computation; computational complexity challenges are overcome by using heuristics. They are used for challenging tasks and performing signal processing on thousands of industrial process signal is no exception. In certain cases, heuristics can be seen as guidelines for first approximation in tedious engineering tasks. Heuristics account for potentially unknown or unexplored relationships in the dataset; they highlight relationships from which process expert might be unaware

of. Additionally, with heuristics, it is possible to achieve benefits quicker and sustain these benefits with minimum effort; they are used to save time and resources.

Conclusion

The data storage capacities of modern process automation systems have grown rapidly. Nowadays, the systems can frequently collect thousands of measurements in parallel and store them in databases. However, these data are still usually not fully exploited in the analysis of processes because, among other things, they contain noise. Filtering is critical because noise directly affects the accuracy of the control system, performance evaluation, fault diagnosis and the reliability of extracted data used in detecting operating state. Therefore, noise reduction is necessary to improve performance, accuracy and prediction robustness and it should be performed before any analysis including optimization, cost analysis and process simulation. Because processing one signal at the time is tedious and impractical, heuristics to select data processing tuning parameters are extracted by using a principal component analysis. Signal processing heuristics are used to speed up the processing of large datasets of industrial process signals.

To reduce noise in industrial process signals denoising effect in the context of prediction, wavelet transform method is used. To achieve the best denoising effect, a parametrical sweep identified the optimal wavelet denoising parameters out of four kinds of wavelet functions including Daubechies (db), Symlets (sym), Coiflets (coif), and Biorthogonal (bior), decomposition levels 1 to 10, soft and hard threshold as well as various threshold selection rule (Rigsure; Heursure; Universal; Minimax) for signals representing various measurement type and fluid nature.

The identified optimal wavelet transform parameters are then used alongside signal's characteristics in a principal component analysis to assess their relations. It is demonstrated that the signal's classification parameters (measurement type, fluid nature, sensor type and process configuration) impacted the wavelet transform parameters and signal's features (frequency, coefficient of variation, variance, rate of change, skewness, and kurtosis). The principal component analysis enabled the formulation of heuristics and conclusions by coupling the optimal wavelet transform signal processing parameters and the signal's features and classification parameters so that random errors can be processed in a structured, systematic and efficient way. The multivariate analysis reveal that it is possible to generalize and extract signal processing heuristics base on the sensor type, i.e., whether the sensor is in a control loop or not, the measurement type, and the fluid nature.

These heuristics provide a robust and practical way to treat thousands of industrial process signals. Heuristics is a mean for faster data processing, they simplify the work of process experts who may then focus on more added-value tasks. The proposed heuristics are validated ; they are proven to be efficient at reducing noise in industrial process signals. The set of parameters suggested by the heuristics are applicable to process all the signals for this one mill application. At a different site, the parameters would most likely be different, but the methodology remains.

The authors have identified several directions for extending this research. Future work can focus on validating the heuristics with an expert, exploring other datasets, testing various sampling frequency, and expanding the methodology with the aim of including additional errors detection capabilities. The application of this methodology to other mills or other types of processes is also envisioned as an interesting area for additional research activities to be carried out. The authors also propose to extend the current work by going deeper in the analysis of the process configuration and sensor type. In summary, the heuristics are a key to achieving intelligent signal processing for process industry that could further be applied to automate portions of the signal pre-processing process.

Acknowledgements

The authors gratefully acknowledge the financial support of the CRSNG.

References

- Abbaszadeh, P., & Alipour, A. (2017). Development of a coupled wavelet transform and evolutionary Levenberg-Marquardt neural networks for hydrological process modeling. *Computational Intelligence*, 34, 1-25. <https://doi.org/10.1111/coin.12124>
- Alyasseri, Z. A. A., Khader, A. T., Al-Betar, M. A., & Awadallah, M. A. (2018). Hybridizing β -hill climbing with wavelet transform for denoising ECG signals. *Information Sciences*, 429, 229-246. <https://doi.org/https://doi.org/10.1016/j.ins.2017.11.026>
- Arakawa, K. (1994). Fuzzy rule-based signal processing and its application to image restoration. *IEEE Journal on Selected Areas in Communications*, 12(9), 1495-1502. <https://doi.org/10.1109/49.339917>
- Cao, Z., Jia, H., Zhao, T., Fu, Y., & Wang, Z. (2022). An Adaptive Self-Organizing Migration Algorithm for Parameter Optimization of Wavelet Transformation. *Mathematical Problems in Engineering*, 2022, 6289215. <https://doi.org/10.1155/2022/6289215>
- Chiang, L., Lu, B., & Castillo, I. (2017). Big Data Analytics in Chemical Engineering. *Annual review of chemical and biomolecular engineering*, 8. <https://doi.org/10.1146/annurev-chembioeng-060816-101555>
- Dambros, J. W. V., Farenzena, M., & Trierweiler, J. O. (2018). Signal Preprocessing for Stiction Detection Methods. *Industrial & Engineering Chemistry Research*, 57(1), 302-315. <https://doi.org/10.1021/acs.iecr.7b04441>
- Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3), 613-627. <https://doi.org/10.1109/18.382009>
- El-Dahshan, E.-S. A. (2011). Genetic algorithm and wavelet hybrid scheme for ECG signal denoising. *Telecommunication Systems*, 46(3), 209-215. <https://doi.org/10.1007/s11235-010-9286-2>
- Fan, H., Tariq, S., & Zayed, T. (2022). Acoustic leak detection approaches for water pipelines. *Automation in Construction*, 138, 104226. <https://doi.org/https://doi.org/10.1016/j.autcon.2022.104226>
- Gao, Q., Duan, C., Fan, H., & Meng, Q. (2008). Rotating machine fault diagnosis using empirical mode decomposition. *Mechanical Systems and Signal Processing*, 22, 1072-1081. <https://doi.org/10.1016/j.ymssp.2007.10.003>
- García Plaza, E., Núñez López, P. J., & Beamud González, E. M. (2018). Multi-Sensor Data Fusion for Real-Time Surface Quality Control in Automated Machining Systems. 18(12), 4381. <https://www.mdpi.com/1424-8220/18/12/4381>
- García Plaza, E., Núñez López, P. J., & Beamud González, E. M. (2019). Efficiency of vibration signal feature extraction for surface finish monitoring in CNC machining. *Journal of Manufacturing Processes*, 44, 145-157. <https://doi.org/https://doi.org/10.1016/j.jmapro.2019.05.046>
- Harashima, H., Odajima, K., Shishikui, Y., & Miyakawa, H. (1982). ϵ -separating nonlinear digital filter and its applications. 65(4), 11-19. <https://doi.org/https://doi.org/10.1002/ecja.4410650403>
- Hasan, A. M., Samsudin, K., & Ramli, A. R. (2011). Intelligently tuned wavelet parameters for GPS/INS error estimation. *International Journal of Automation and Computing*, 8(4), 411-420. <https://doi.org/10.1007/s11633-011-0598-9>
- Katunin, A., & Przystalka, P. (2015). Automated wavelet-based damage identification in sandwich structures using modal curvatures. *Journal of Vibroengineering*, 17, 2977-2986.
- Li, H., Ma, Y., Liang, K., Tian, Y., & Wang, R. (2012). Parameters optimization for wavelet denoising based on normalized spectral angle and threshold constraint machine learning. *Journal of Applied Remote Sensing*, 6, 3579. <https://doi.org/10.1117/1.JRS.6.063579>
- Liu, H., Dong, H., Liu, Z., Ge, J., Bai, B., & Zhang, C. (2017). Noise characterization for the FID signal from proton precession magnetometer. *Journal of Instrumentation*, 12, P07019-P07019. <https://doi.org/10.1088/1748-0221/12/07/P07019>

- Liu, W., Chang, Q. R., Guo, M., Xing, D. X., & Yuan, Y. S. (2011). Extraction of first derivative spectrum features of soil organic matter via wavelet de-noising. *Guang Pu Xue Yu Guang Pu Fen Xi*, 31(1), 100-104.
- Malghan, P. G., & Kumar Hota, M. (2022). Grasshopper optimization algorithm based improved variational mode decomposition technique for muscle artifact removal in ECG using dynamic time warping. *Biomedical Signal Processing and Control*, 73, 103437. <https://doi.org/10.1016/j.bspc.2021.103437>
- Mallat, S. G. (2009). *A wavelet tour of signal processing: the Sparse way* (3^e éd.). Elsevier /Academic Press.
- MathWorks. (2020). *Threshold selection for denoising*. MathWorks. <https://www.mathworks.com/help/wavelet/ref/thselect.html>
- MathWorks. (2023). *Choose a Wavelet*. <https://www.mathworks.com/help/wavelet/gs/choose-a-wavelet.html>
- McSharry, P. E., Clifford, G. D., Tarassenko, L., & Smith, L. A. (2003). A dynamical model for generating synthetic electrocardiogram signals. *IEEE Transactions on Biomedical Engineering*, 50(3), 289-294. <https://doi.org/10.1109/TBME.2003.808805>
- Mikhled, A., & Khaled, D. (2008). ECG Signal Denoising By Wavelet Transform Thresholding. *American Journal of Applied Sciences*, 5. <https://doi.org/10.3844/ajassp.2008.276.281>
- Nourani, V., Alami, M. T., & Aminfar, M. H. (2009). A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation. *Engineering Applications of Artificial Intelligence*, 22(3), 466-472. <https://doi.org/10.1016/j.engappai.2008.09.003>
- Nourani, V., Kisi, Ö., & Komasi, M. (2011). Two hybrid Artificial Intelligence approaches for modeling rainfall-runoff process. *Journal of Hydrology*, 402(1), 41-59. <https://doi.org/10.1016/j.jhydrol.2011.03.002>
- Pomalaza-raez, C., & McGillem, C. D. (1984). An Adaptive, Nonlinear Edge-Preserving Filter. *IEEE Transactions on Signal Processing*, 32(6), 571-576. <https://doi.org/10.1109/TASSP.1984.1164361>
- Qin, S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control*, 36(2), 220-234. <https://doi.org/10.1016/j.arcontrol.2012.09.004>
- Rajala, S., & Figueiredo, R. (1981). Adaptive nonlinear image restoration by a modified Kalman filtering approach. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(5), 1033-1042. <https://doi.org/10.1109/TASSP.1981.1163679>
- Ren, Y., Liu, P., Hu, L., Huang, J., Qiao, R., Chen, H., . . . Huang, S. (2021). Research on Noise Reduction Method of Pressure Pulsation Signal of Draft Tube of Hydropower Unit Based on ALIF-SVD. *Shock and Vibration*, 2021, 5580319. <https://doi.org/10.1155/2021/5580319>
- Sawant, C., & Patil, H. T. (2014). ECG Signal De-noising using Discrete Wavelet Transform. *International Journal of Electronics Communication and Computer Engineering*, 5(4).
- Shafri, H., & Yusof, R. (2009). Determination of optimal wavelet denoising parameters for red edge feature extraction from hyperspectral data. *Journal of Applied Remote Sensing*, 3. <https://doi.org/10.1117/1.3155804>
- Smallwood, D. O., & Cap, J. S. (1998). *Salvaging pyrotechnic data with minor overloads and offsets*.
- Tan, F., Feng, X., Li, M., Wang, Z., Yang, L., Li, Y., . . . Nie, F. (2008). Construction and application of a novel library: Fourier transform infrared wavelet coefficients library. *Anal Chim Acta*, 629(1-2), 38-46. <https://doi.org/10.1016/j.aca.2008.09.017>
- Tanaka, Y. (1994). Video signal noise characteristics with regard to scanning method. *IEEE Transactions on Broadcasting*, 40(2), 68-74. <https://doi.org/10.1109/11.286286>
- Wang, P., Yan, M., Zhang, L., & Yang, N. (2021). Shock Signal Trend Term Error Correction Method Based on Discrete Wavelet Transform and Low-Frequency Oscillator Combination. *Shock and Vibration*, 2021, 9939547. <https://doi.org/10.1155/2021/9939547>
- Wang, W., & Ding, J. (2003). Wavelet Network Model and Its Application to the Prediction of Hydrology. *Nat. Sci.*, 1.
- Yang, B., Xia, H., Annor-Nyarko, M., & Wang, Z. (2020). Application of total variation denoising in nuclear power plant signal pre-processing. *Annals of Nuclear Energy*, 135, 106981. <https://doi.org/10.1016/j.anucene.2019.106981>

- Zhou, D., Wang, Q. J., Tian, Q. J., Lin, Q. Z., & Fu, W. X. (2009). Wavelet analysis and its application in denoising the spectrum of hyperspectral image. *Guang Pu Xue Yu Guang Pu Fen Xi*, 29(7), 1941-1945.
- Zuo-wei, P., Hong-wen, A., Shuang-yin, L., & Yi-bing, L. (31 May-2 June 2014 2014). *Amplitude envelope analysis for feature extraction of power plant blower bearing failure*. The 26th Chinese Control and Decision Conference (2014 CCDC) (p. 2821-2824). <https://doi.org/10.1109/CCDC.2014.6852653>

ANNEXE E ARTICLE #4 AN OPERATING REGIME-BASED DATA-DRIVEN INDUSTRIAL FRAMEWORK FOR IMPROVED PROCESS DECISION-MAKING – PART 1: FRAMEWORK DEVELOPMENT

Émilie Thibault, Jeffrey Dean Kelly, Francis Lebreux Désilets, Moncef Chioua, Bruno Poulin, Paul Stuart
Submitted to the Industrial & Engineering Chemistry Research Journal on 08/02/2023

Abstract

Data management systems are increasingly used in industrial processes. However, data collected as part of industrial process operations, such as sensor or measurement instruments data, contain various sources of errors that can hamper process analysis and decision-making. Therefore, in order to take full advantage of collected and stored data and to increase data quality, an operating regime-based data-processing framework for industrial process decision-making is proposed in this paper. This systematic and structured approach includes the following stages : (1) scope of the analysis, (2) data management and (3) operating regime detection and identification. All steps are based on the combination of process knowledge and data-driven approaches.

The scope definition starts with the analysis objective definition since data must always be processed according to a specific objective. Then, the data management step includes data synchronization, lag correction, shutdown and start up periods, outlier removal and noise reduction. The detection of steady-state operating periods and data reconciliation for gross error detection, identification and removal are accounted for in this same step. Lastly, given that operating regimes detection and identification using a combination of process knowledge and data-driven approaches are lacking in the literature, this paper attempts to address this by setting-up an operating-regime based data processing framework in which regimes are considered and detected using both domain-specific knowledge and clustering analysis (principal component analysis and *k*-means algorithm).

The proposed framework is applied to data from a brownstock washing department of a dissolving pulp mill. All steps of the framework are applied in order to identify how many operating regimes can be detected when the process is running smoothly (in steady-state). The framework identified 20 steady-state spans that lasted for more than 12 hours and the top four sensors that are persistently showing gross errors.

Lastly, the framework is employed in a case study presented in Part II of this publication. The case study presents an improved way to operate the brownstock washing by incorporating an activity-based costing analysis to the process data obtained from the framework application.

Keywords Industrial data, data processing, steady-state detection, data reconciliation, operating regime

Introduction

Data available in today's industrial processes are abundant and contain the knowledge that can help plants with their decision-making processes. However, big data doesn't always include good data only; raw process data alone have no refined value and are of limited use until they are converted or transformed into information. Measured process data from sensors and instruments are the source of process data, but are however inherently corrupted and distorted by different types of errors and therefore data cleaning techniques must be applied before incorporating

these data in any analysis Korbel, Milan (2011). In other words, plants need to pre-process their big data to obtain useful good data of reproducible and reliable quality.

Measurements taken in industrial processes are generally impacted by several types of errors that can be grouped under three main categories : random and gross errors as well as abnormalities. The presence of these errors can be evidenced by noise in process data and the inconsistency between the measured values and the material, energy and momentum balances related to the process. This makes the direct use of raw data inefficient. In order to overcome these issues, data filtering and reconciliation have become established industrial practices (Bagajewicz, M., 2000). Data manipulation approaches allow us to take full advantage of the information collected and provide many benefits such as better plant knowledge, reduced downtime (hence, throughput increase), decrease in maintenance costs, reduce over-specified feed quality, increase in product quality, minimized energy consumption and improved energy efficiency (Desbiens et al., 2008).

An adequate elimination of the errors present in the measurements enables us to increase the effectiveness and efficiency of high-level decisions related to the operation of chemical processes. The latter include the configuration of the control systems, detection and diagnosis of the faults related to the operation of the process, planning and scheduling and coordinating and real-time optimization of the process operations and cost analysis.

The operation of a process can exhibit different regimes that may be characterized by feedstock changes (nature or quality of feedstock), diverse operating or processing conditions, different products runs and product switching periods, seasons, maintenance periods, shutdown and ramping-up periods, and recipes. Operational decisions related to a process could benefit greatly from the detection and identification of its operating regimes.

Segmenting industrial process sensor data (time-series data) into different operating regimes opens the possibility of assessing energy (or resources) consumption, detecting ineffectiveness and inefficiency, monitoring performance, cost, yield, and product quality for each regime. This analysis should help dictate and direct the best way to operate a plant. The identification of the most economic, effective and efficient operating regimes also determines the least interesting ones. The detection and identification of these regimes may also help the debottlenecking of the process and design projects with low or no capital expenditures which could be undertaken. Operating regime detection and identification helps to understand what happened during different periods. Therefore, it may impact the production coordination, scheduling and planning of the plant. The extent of what is presented here will be confirmed through a case study in part II.

High-level decisions related to the management of chemical processes, i.e., decisions made offline by the process management team, are directly dependent on an adequate interpretation of the process variables trends. However, the escalating complexity and specificity of chemical processes require the consideration of a priori process knowledge when analyzing industrial data for decision-making in order to extract and analyse these trends and behaviours.

Therefore, this paper proposes an operating regime-based data management framework that combines the interpretation of the process operation (understanding of the process) with data-driven mathematical approaches for design, production, process and/or operation related decision-making. This systematic and structured way to analyze historical process signals for making offline high-level management decisions puts a specific emphasis or focus on operating regime detection and identification (ORDI). This practical framework will be applied to the data of a brownstock washing department of a dissolving pulp mill.

To pull together this architecture, where the major contribution is on the operating regime detection and identification, section 2 presents the existing data processing frameworks that are established for industrial decision-making. Section 3 reviews existing methods for industrial process operating regime detection proposed in the

literature. Then, section 4 describes the steps employed to obtain the proposed methodology (operating-regime based data processing framework) and section 5 presents the results (the framework based on the combination of domain knowledge and data-driven approach). The next section demonstrates the applicability of the framework on data from a dissolving pulp mill. Section 7 gives a conclusion of the present work.

Review of data processing frameworks for industrial decision-making

Several industrial decisions made in plants require the application of data management steps. The data manipulation and treatment steps required vary from one application (decision) to another. In the sequel, data processing frameworks employed for industrial decision-making, focusing on offline process diagnostics, troubleshooting, optimization and cost analysis, will be reviewed.

In 2003, Jiang and al. proposed a systematic strategy for improving data quality (Figure 1) (Jiang, Taiwen, Chen, Jasim, et al., 2003). Starting with process data, the data processing step reduced errors and extracted key process trends. To do so, the authors performed a multi-scale data processing using wavelet transform. They discarded random noise and abnormalities and extracted process trends by using an approach proposed in 2000 by Jiang (Jiang, Taiwen et al., 2000). Secondly, a method based on wavelet transform presented by Jiang (Jiang, Taiwen et al., 2000) is used for the steady-state detection. Lastly, the data reconciliation step is performed using the Sigmafine software (Sigmafine, 2022).

This data processing framework is proposed to improve and validate a steady-state simulation model of a recaustizing plant in a pulp mill. The processed and reconciled data showed that the simulation is reasonable for practical application such as process sensitivity analysis, investigation of pulp production increase and providing insight related to capital effective options for plant upgrade alternatives.

Data must be processed for a targeted specific application. In this case, it is for the validation and calibration of a process simulation – which, itself, could be used for a variety of applications. Therefore, it is critical to highlight the scope of the analysis and take into consideration the period that is under study here. Lastly, operating regimes should be considered in order to account for the mill or plant variability.

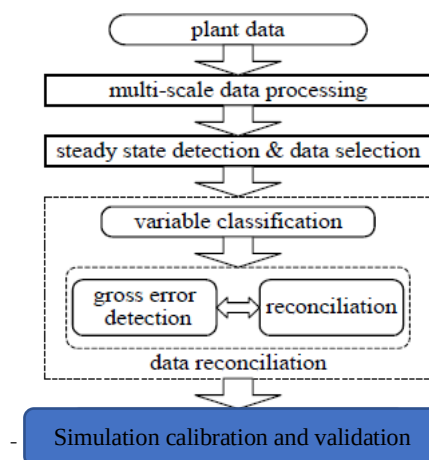


Figure 1 – Strategy for improving data quality for simulation calibration and validation, adapted from (Jiang, Taiwen, Chen, Jasim, et al., 2003)

Based on Jiang's work, Bellec et al. (Bellec et al., 2007) suggested a sequence of steps to increase value by improving the quality of real-time process measurements (Figure 2). Jiang's framework aims to improve the overall accuracy and precision of process data. The selected data treatment techniques are suitable for online applications.

Bellec et al. (2007) data processing framework helps to identify various sources of errors present in real-time measurements, including noise, abnormal data points or outliers and inconsistencies in the material and energy balances to provide better quality data. These data can then assist process engineers in quickly identifying equipment maintenance requirements, troubleshooting and diagnosing operating problems and improving process related decision-making. Concerning the latter, examples of process management decisions that are generally based on steady-state datasets include planning, scheduling and coordinating.

The first step of Bellec et al. (2007) presented in Figure 2 is the detection and correction of abnormal measurements; this is accomplished by employing a method based on wavelet transforms (Jiang, Taiwen, Chen, He, et al., 2003). Then, a data filtration technique (also based on wavelet transform (Jiang, Taiwen, Chen, He, et al., 2003)) is used to correct random errors. Thereafter, in order to reconcile measurements, steady-state operating periods are carefully identified based on the approach mentioned in (Jiang, Taiwen, Chen, He, et al., 2003). Lastly, data reconciliation is undertaken to correct gross errors using a weighted least squares method.

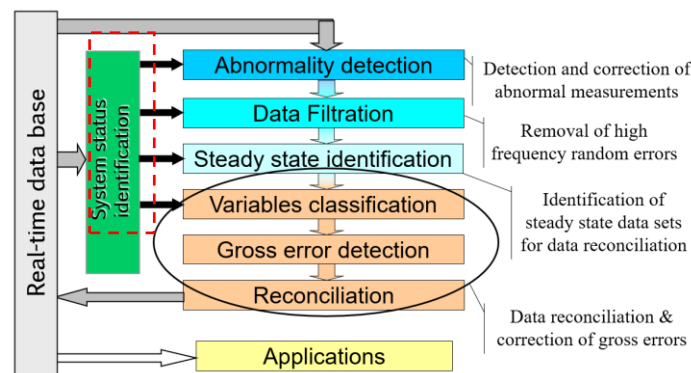


Figure 2 – Methodology to improve data quality proposed by Bellec et al., adapted from (Bellec et al., 2007)

In Bellec et al. (2007) framework, the authors included a system status identification where the main purpose is to identify the conditions of operation for the system under study and select the data processing parameters associated with the identified process status (operating regime). The selection is based on an offline investigation of all steady-state operating conditions and the parameters are set to achieve optimal results. Bellec et al. (2007) are vague regarding the optimality of the results.

During the validation of the framework presented by Bellec et al. on historical data from a paper machine, the authors included the notion of operating regimes also known as modes or grades. They underlined that a process may have many different steady-state operating regimes. Those are identified for the system under study and then the characteristics associated with the different operating regimes are compiled and stored in a database. For instance, in their example, 27 different operating regimes are merged into 4 groups, namely production of paper grade A, production of paper grade B, maintenance stops, and the occurrences of paper breaks. The probability of occurrence, the duration of each regime and the delay occurring between those are assessed. Additionally, the mean and variance of process measurements within each operating regime are also compiled and stocked in a database. These characteristics, associated with each process measurement within an operating regime, are useful to properly identify the operating and processing conditions of the studied process (Bellec et al., 2007).

Furthermore, the authors ensure that the lags in the process are considered (i.e., the data are synchronized) by retaining only steady-state periods longer than a threshold value, i.e., the system delay plus a safety factor. Therefore, they targeted steady-state periods long enough for them to compensate for the fact that the various departments in a process are not synchronized.

The authors mentioned that improving data quality and reliability (accuracy and precision) justify the effort and expense of implementing their framework. Furthermore, by compiling steady-state reconciled data in different operating regimes, plants can build a database of accurate process data. The latter could lead to benefits related to process operation and troubleshooting such as identification of process leaks and instruments failures for improved process yield and conversion calculation. The understanding of the process is improved from knowing the different process operating conditions. Lastly, plant-wide or multi-unit-wide optimization could be done by coupling operational data and operating cost data for specific products in particular operating regimes.

Therefore, compared to Jiang, Taiwen, Chen, Jasim, et al. (2003) framework, Bellec et al. (2007) considered operating regimes. To the best of our knowledge, the operating regimes are manually identified by experts in the process. Additionally, the lags in the process are considered but not systematically and rigorously accounted for; to eliminate them, Bellec et al. (2007) retained only steady-state period that lasted a certain amount of time. Lastly, there is no mention of an identification of steady-state periods for a whole system, i.e., a unit-wide or plant-wide steady-state detection.

In 2011, Korbél proposed a framework that uses plant-wide rectified process data in cost analysis for identifying short- and long-term benefits (Korbél, Milan, 2011). The initial step consists of defining operating regimes, i.e., characterizing the operating differences that occur while producing each specific product. According to Korbél, Milan (2011), operating regimes are established when changes occur in process design, when different equipment are used, when changes occur in production target or in setpoint control strategy as well as feedstock diversity. The second step is the identification of steady-state operation within the operating regime under analysis. Noise and abnormalities are extracted from the process measurement trends by applying optimal wavelet transform parameters which are adjusted iteratively until optimal data pre-processing and accurate steady-state detection are perceived to be achieved. Then, the de-noised signals, series or trends are analyzed for steady-state occurrences using a methodology based on wavelet transform presented in Korbél, Milan et al. (2014). The detection of steady-state periods is done in two steps, (1) univariately and (2) in a multivariate way by comparing the whole set of key variable states over time. Then, plant-wide data reconciliation assesses the data validity by comparing them to the values of the underlying process model i.e., a simulation that accounts for the material and energy balances and/or other laws of conservation. The details behind the steps performed for data rectification are presented in (Korbél, Milan, 2011).

The reconciled data obtained following this last step is used for advanced cost analysis for production decision-making in (Korbél, Milan, 2011). This operations-driven cost analysis is performed through an activity-based cost modelling approach for a given operating regime. Ultimately, with Korbél, Milan (2011) framework, the operating regimes are defined based on process knowledge and sound engineering judgement.

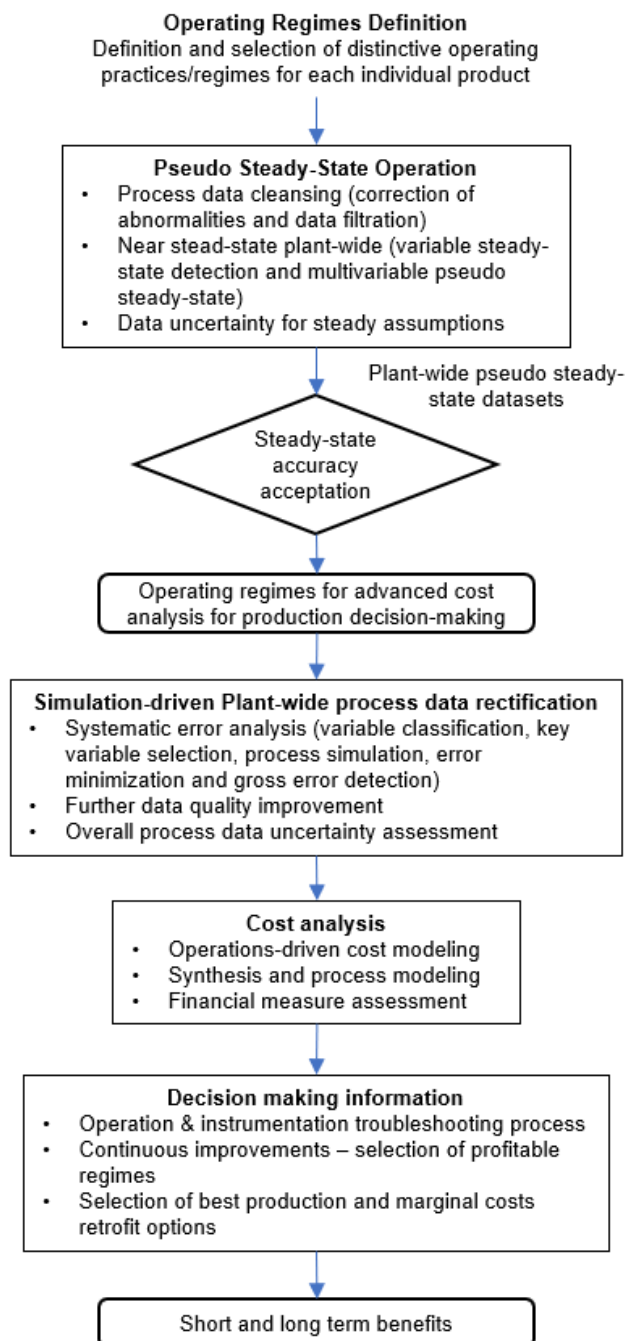


Figure 3 – Data processing framework for cost analysis, adapted from (Korbel, Milan, 2011)

Lastly, other papers proposed a sequence of data processing steps that lead to process optimization. For instance, Delou et al. (Delou et al., 2021) put forward that process optimization must start with a data acquisition step where a user specifies whether the analysis will be performed offline, independently of the actual state of the plant, or online, reading the values directly from the process database. Then, in the data manipulation step, unwanted or out-of-limit values are excluded from the input variables. These two first steps are re-iterated until all possible error have been removed. Next, to reliably estimate the model parameters, steady-state detection must be done before the data reconciliation step (Delou et al., 2021). Prior to solving the optimization problem and validating its solution, a nonlinear programming problem (data reconciliation) must be solved. Their approach does not consider operating regimes detection.

Ultimately, it is found in the literature that most of the data management frameworks for offline decision-making considered data processing steps (noise reduction, outlier detection and data reconciliation), but their scope is not always explicit, and operating regime detection is not systematically considered. This paper proposes a systematic and structured way to process data inspired by the frameworks presented in section 2, that considers a scope definition, data processing, steady-state detection, data reconciliation and operating regime detection. The next section discusses industrial process operating regime detection.

Industrial Process Operating Regime Detection

Operating regimes are different operating strategies. For instance, some examples of those are different feedstocks, product recipes, grades, production rates, seasons, equipment modes, etc. However, changes in processing conditions, such as long-term evolution and transition from one regime to another, have often been ignored when analyzing industrial process data (Liukkonen, Hiltunen, et al., 2011). Modern automated data acquisition systems, improved data storage capacity and retrieval, combined to the proliferation of inexpensive sensors and instrumentation increased massively availability of data. This massive amount of data is not easy to interpret by conventional means. To exploit and take advantage of this data, chemical engineers must use data science tools, such as those discussed in this paper, and especially in section 3. Industrial (chemical) plants commonly operate in a finite set of operating regimes, which may be classified as modes where process variables vary within a relatively narrow band, i.e., are at steady-state (Srinivasan et al., 2005). On the other hand, transition or sequencing periods are characterized by large changes in one or more variables (Srinivasan et al., 2005). The latter represent the state of the process between two modes, e.g., a change in setpoint, the opening of a valve, and/or a change in equipment configuration. Research has shown that the behavior of a process may differ greatly between modes (Liukkonen, Hiltunen, et al., 2011).

Archived process data can be used to cluster or characterize process states (Heikkinen et al., 2011). Heikkinen et al. grouped variables by the *k*-means algorithm into clusters that represent different process states (Heikkinen et al., 2011). Therefore, the state of the process is determined through data analysis. The clusters are analyzed to indicate differences in process states by examining their properties and by using process knowledge. A cluster representing periods of plant shutdown is easily identified based on very low volumetric flow. Their findings showed that seasonality is one of the clearest distinguishing factors between clusters. One cluster is prevalent in the summertime, another in the wintertime, a third represented periods when the whole process is either out of control or shut down, and a fourth represented periods when the process is unstable or unsteady. Various identified process states gave information about its process behavior. Furthermore, this information is used to determine optimum set-points for key process variables within a process state. Process knowledge is needed for variable selection and for associating and allocating process states to clusters (Figure 4). Lastly, there is always measurement noise that affects accuracy; therefore, the variables are filtered by a moving average filter (Figure 4).

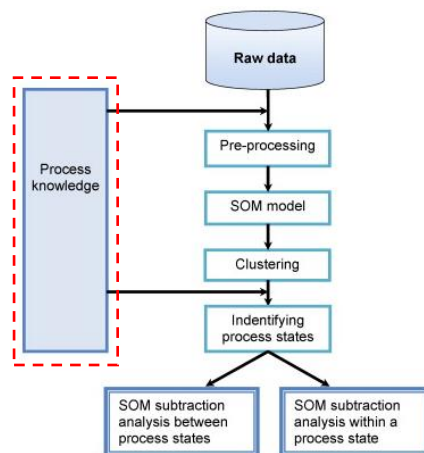


Figure 4 – Data processing chain diagram for self-organizing maps (SOM), adapted from (Heikkinen et al., 2011)

The idea of using k -means clustering to create subsets representing process states is taken up by Liukkonen et al. as well (Liukkonen, Heikkinen, et al., 2011; Liukkonen, Hiltunen, et al., 2011). One of the most commonly used unsupervised clustering method is the k -means algorithm (Jain, 2010). Perhaps the biggest issue when trying to implement this technique is the requirement to set the number of clusters in advance; this is rarely known for complex processes operating under unknown conditions or in a changing environment (Saari & Odelius, 2018). Therefore, Saari et Odelius (2018) used another unsupervised clustering technique to detect operating regimes in condition monitoring data, e.g., vibration signals, called the variational Bayesian Gaussian mixture.

Traditional clustering techniques are computationally expensive and generally perform poorly on time-series data because they are intended for more time-independent observations, Srinivasan et al. (2004) proposed a two-step clustering method based on principal component analysis (PCA). Chemical plants normally operate in a number of operating states (e.g., plant start-up, grade change, shutdown), including steady state, and frequently switch between them. Therefore, in the first step, process states are classified into modes, i.e., the process unit operates in steady-state and transitions, using a novel multivariate algorithm to segment historical data. A PCA algorithm is then used in the second step to compare the different modes and transitions to cluster them based on their similarities. Process control, simulation, fault detection, and alarm management are examples of applications where the controller parameters must be adjusted to fit the current process state (Srinivasan et al., 2004).

In practice, translating the obtained clusters to operating regimes periods requires adequate knowledge of process history, but this is often unavailable. Hence, unknown operating regimes may exist. Therefore, Beaver et al. (2007) used cluster analysis (Everitt, 1993) as a framework to simultaneously identify process states, detect transition periods between states, and label times of occurrence using historical time-series data. As part of the cluster interpretation, the most desired operating regimes are identified on the basis of knowledge of the process history.

Even if the notion of process knowledge is considered (Heikkinen et al., 2011) or mentioned (Beaver et al., 2007), studies, approaches or frameworks that identify operating regimes using a combination of process knowledge and data-driven approaches are lacking in the literature. Therefore, this paper attempts to address this by setting-up an operating-regime based data processing framework in which regimes are considered and detected using both domain-specific knowledge and clustering analysis (principal component analysis and k -means).

Methodology : Operating-Regime Based Data Processing Framework

This paper presents a systematic and structured approach (framework) to analyze historical process signals to assist in the making high-level management decisions. These decisions would be made offline. This framework puts a specific focus or emphasis on operating regime detection and identification. The latter is based on the combination of domain knowledge and a data-driven approach.

The activities undertaken as part of this study, to arrive at our framework, are presented here.

To obtain this architecture, where the major contribution is on the operating regime detection and identification, a literature review of (1) existing data processing frameworks, and (2) operating regime detection algorithms is accomplished. Then, a way to consider both the process knowledge as well as the information contained in industrial process data for ORDI is described. As part of this research, process knowledge consideration falls into what we call “top-down” analysis whereas the data-driven approach is referred to a “bottom-up” analysis. Operating regimes classified as top-down are obvious and easily detectable; process experts are aware of those. Grades, seasons, production rate, recipes, setpoints, lines closing, equipment shutting down, pumps status, etc. are examples of these. On the other hand, operating regimes resulting from the bottom-up analysis are identified through a clustering analysis. These operating regimes are found through principal component analysis (PCA) and *k*-means algorithm, along side process knowledge. The latter is required in the bottom-up analysis as well in order to give meaning and interpretation of the PCA results. Therefore, from the process data, it is possible to learn more about how the plant operates.

The top-down analysis may be assimilated into a white-box approach since it puts forward how a plant is supposed to operate, how it is designed, what grades are produced, what are the setpoints for those grades, what are the standard operating procedure (SOP) says, etc. On the other hand, the bottom-up can be perceived as the black-box approach; process data, through clustering analysis, may uncover unknown elements about the process.

In addition to the detection and identification of the process operating regimes, the analysis scope must be defined, the data are processed, the steady-state periods are detected, and the data are reconciliated. The last three steps are assimilated to the data management stage. Finally, the information obtained from the framework is used for decision-making. Figure 5 captures the overall architecture of the framework.

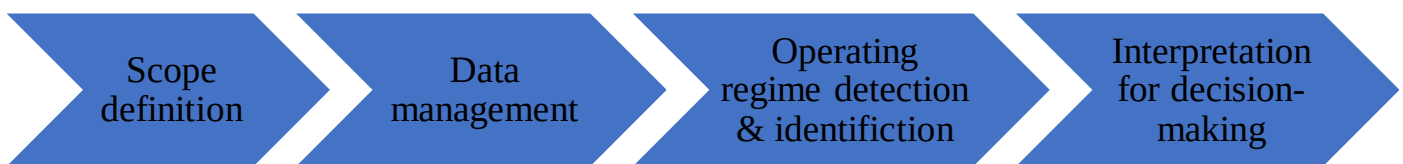


Figure 5 – Framework presentation

Scope definition

The first step, in any data processing analysis, consists of defining the objectives of the analysis, e.g. optimizing the production sequence, reducing the operating cost in a plant’s department, identifying unprofitable operating regimes. Then, the operating regimes are defined using process knowledge. Lastly, the data are collected over a horizon that is representative for the analysis.

Data management

As part of the data management stage, the data are processed, meaning that the random errors are reduced, the steady-state periods are detected, that would be the periods where the whole system under analysis is relatively stable and the last step is a unit-wide data reconciliation to detect, identify and eliminate gross errors.

Operating regime detection and identification

The ORDI is done through a cluster analysis based on a combination of principal component analysis, *k*-means algorithm, and domain knowledge. The reconciled steady-state data are classified into operating regimes and, lastly, some of those regimes are selected according to the scope to solve a high-level management problem (related to process design or operations).

Following the framework development, an application is demonstrated on data from the brownstock washing department of a dissolving pulp mill in order to put this architecture together.

Results and Discussion

The following sections present the development of a framework for which all steps consider the combination of domain-knowledge and data-driven approaches.

5.1 Framework Development

Measurements taken in industry contains several types of errors that can be grouped under three main categories : abnormalities, random errors, and gross errors. These errors are corrected in a series of steps. First, given the fact that data are always processed according to a specific and targeted goal, the objective of the analysis (the scope) is defined, and data are collected. Then, data are processed, i.e., data undergo a synchronization step, a lag correction, the abnormal operations are removed, and data are filtered to reduce unwanted high frequency components. The third step is to detect the steady-state data series that will allow the data reconciliation step. Lastly, the various operating regimes are detected and identified. This framework is limited to historical operating data (as opposed to real-time online process data) as they are an important source of process insight.

The proposed framework focuses on data that are representative of normal and desirable operating conditions. Abnormal operations, fault detection and diagnosis are critical, but are not the subject under advisement in this paper. Therefore, shutdown and startup periods as well as outliers are removed, and steady-state periods are detected.

Figure 6 depicts the proposed methodological framework. While some of the reasoning is already explained in the previous section, the key elements of this framework are detailed.

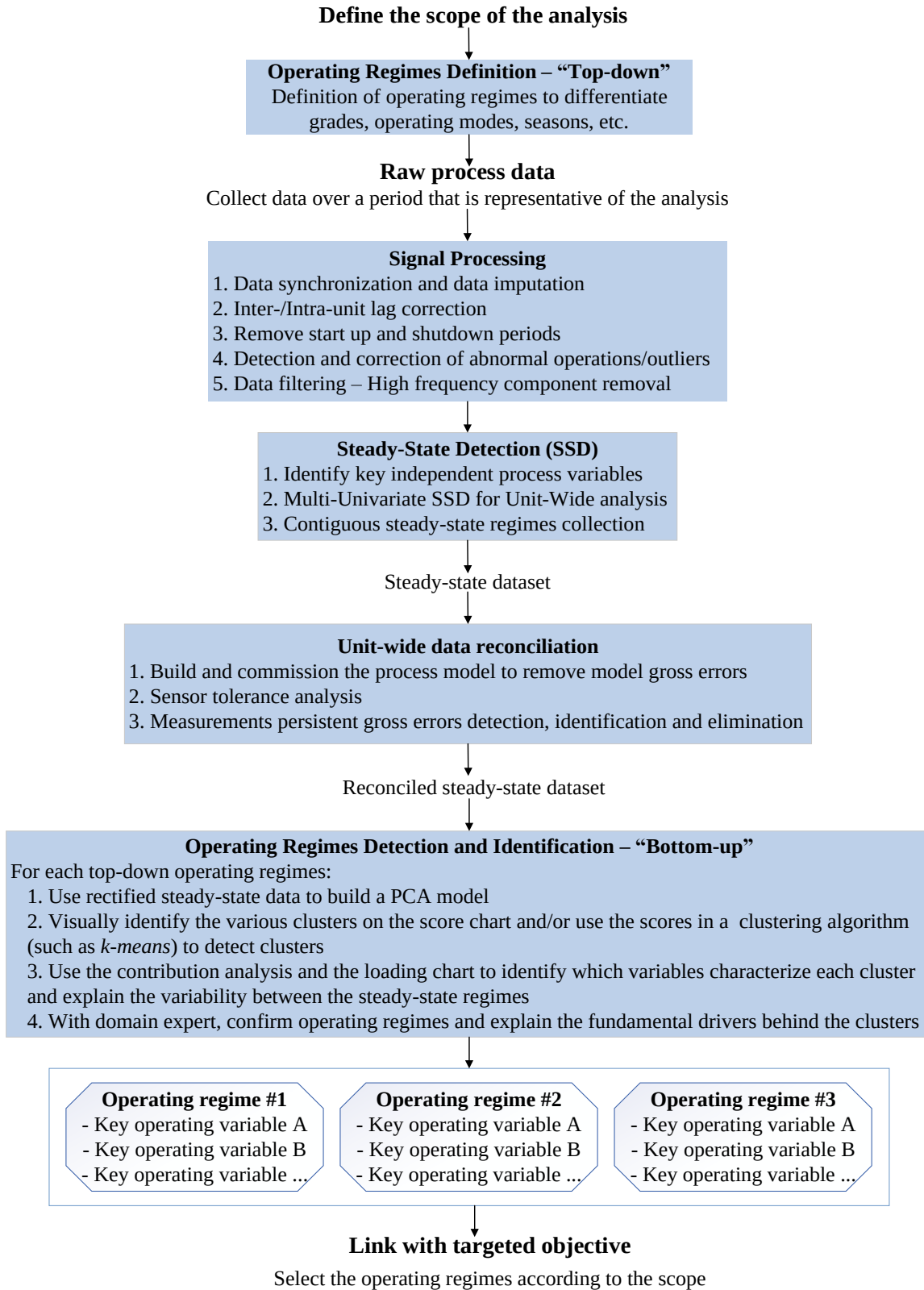


Figure 6 – Framework combining process knowledge and data-driven approaches

The data processing methodology developed in this paper is enabled by the work of Korbel (Korbel, Milan, 2011), Bellec (Bellec et al., 2007) and Jiang (Jiang, Taiwen, Chen, Jasim, et al., 2003). This framework is putting forward the combination of process expertise and data-driven approaches to obtain enriched knowledge about the process and use it to make informed offline management decisions. The mathematical data analysis techniques are employed to learn, discover some operating regimes, and extract (hidden) knowledge (that perhaps is not present in the domain knowledge) from data which can help to make quick and smart decision in complex situations whereas the domain knowledge contains information which is not accessible in the datasets.

In summary, the proposed methodology exploits the advantages of data-driven approaches and uses process expertise as complementarity information sources. This is a generic approach that can be applied for a wide range of industrial processes. Using process operation knowledge in all steps of data analytics is also addressed in Joe Qin et al. (2021).

Lastly, all steps have their own internal feedback loop. For instance, in the data reconciliation step, the moment one gross error is eliminated, another may be found; a practical way to perform data reconciliation is to remove gross errors one at the time.

Scope definition

Any data processing analysis starts with the definition of the objectives. It is critical to put forward why this analysis is performed, what are the data used for and what is being studied. Therefore, the operating regimes known and recognized by the process experts, i.e., the different recipes as expressed in standard operating procedure that are related to the analysis objective, should be acknowledged. Operating regimes are used to answer questions, solve problematics and make decisions in industrial processes. Depending on the question, the problematic or the process decision to be made, specific operating regimes will be required.

“Top-down” operating regimes are driven by changes in the control setpoint strategy caused by different operating conditions (seasons, feedstock changes, maintenance, the usage of different production line, etc.). The fact to the matter is that those setpoints does not account of everything happening in industrial processes. In fact, even if operating regimes are defined by standard operating procedures (SOP), setpoints are changed by operators since a plant is not a 100% automated. The process can be operated in automatic or manual mode; some variables are left in automatic control all the time while others are not. Furthermore, it could be possible to observe differences in some variables when an equipment needs maintenance, hence before and after maintenance. Moreover, equipment performance changes all the time because of fouling, deterioration, therefore the recipe should be adjusted accordingly. Hence, operators are acting when something changes in the process; there could be a problem in the process, variation in raw material, hence the process needs adjustments. Also, two different operators won’t necessarily take the same action. These considerations are made later on with the “bottom-up” operating regime detection through clustering analysis and the drivers of these operating regimes are assessed (process knowledge is necessary to understand why variables act as they do). In the scope definition, only the “top-down” operating regimes are considered; the idea is to go as far as possible using the recipes (SOP) considering what is required related to the scope, then those regimes will be sub-categorized by using clusters.

Lastly, raw process data are collected over a period that is representative for the targeted analysis. Every process generates a wide range of data types that are collected, gathered, and stored in the process database/ historian systems. These include process sensor data (pressure, flow, temperature, level), lab results, manufacturing schedule,

production mode, product grade, and many others. Therefore, pertinent data must be gathered considering the objective of the analysis.

Data processing

Data must be synchronized to the same time step since they are often stored on the change of value, which is not a regular frequency, and variables used in an analysis won't necessarily be sampled at the same rate (for instance hourly for some lab data versus by the second for some sensors). Compression is used to save data storage space, therefore sensor data are often only logged when the values differ significantly from the previously stored value (change-of-value basis), whereas lab data might be available once every hour or every day. Therefore, to perform any analysis, all data should be synchronized, and preferably, to the frequency of interest. Thus, if the sampling frequency of a variable is lower than the frequency of interest for an analysis, data imputation must be performed (linear interpolation of data, repeat the last valid value) (Huang, 2021). On the other hand, if the sampling frequency is higher than the frequency of interest, then the mean, the closest value, the minimum value or the maximum value over a sampling interval equal to the desired time step must be considered, depending on the analysis.

For most analyses, it is not mandatory to have all the sampled data points values. For instance, to determine a trend, a few missing values in a dataset are not a problem (Gitzel, 2016). Larger segments without data are, however, unusable in most analysis. Therefore, missing half of the values is a problem in the case of continuously missing data that results in large empty stretches, but, depending on the analytics algorithm, it might be less of an issue if every other value is missing (Gitzel, 2016).

Once the synchronization process to set data on a regular time interval is done, the lag is accounted for. For an analysis that is taken place inside one department (system or unit) only, intra-unit lag correction is performed. On the other hand, a plant-wide analysis considers inter-unit lag as well – that is, the lag from one unit to another. For instance, when the production stops in the digester units of a pulp and paper mill, it takes a few hours for the repercussions to be visible at the paper machine. The lag account for the time required for the operations, the retention time, the transport time, etc. Lags vary according to the production rate. By ignoring the lags, some correlations may be lost. Furthermore, most of the subsequent steps involved in this framework employed reference variables (independent variables that describe the process) to perform the analysis, therefore all variables must be aligned in time beforehand. The lag correction analysis is achieved based on Fourier analysis (software like EXPLORE (Canada, 2022) offers this functionality).

As part of the data processing phase, the shutdown and start up periods should be removed. To do so, variable of reference are selected to detect those periods. Even when the process is stopped, some pieces of equipment are still functioning, hence some sensors are still collecting data and therefore couldn't be used as reference. Using reference variables, specific periods are detected visually and deleted from the dataset by numerical validation, i.e., all values under a certain limit that are not representative of the normal and desirable operating conditions. The threshold values are specified by a process expert. The removal of process shutdown and start up events from the reference variables are mirrored onto the other variables used as part of the analysis (values at certain time stamps will be eliminate for all variables). Lastly, additional data points beyond the specified period which are impacted by a process shutdown (a time delay/interval or buffer around these periods) are removed to account for the time required to shut the system and to reach normal operations again. In other words, periods of downtime are extended to include transient periods. The time intervals removed around downtime is chosen based on operating practice and process expert judgement.

Alongside shutdown removal, an important step in signal processing is to detect and remove outliers (Liu, Hancong et al., 2004). Therefore, once downtime periods are eliminated from the dataset, outliers, i.e., observations that

deviates from the expected behavior of a process variable – values that are out of range, are removed. These values are often of short duration and present abrupt changes in the signal. The fact that top-down operating regimes are defined earlier on helps in the identification of outliers as segmenting process data reduces variability (and the standard deviation), and thus makes it easier to detect process deviations and avoid unnecessary rejects. Similarly as what is done for the shutdown and start up periods, variables that represent the system under analysis (key process variables) are used to detect outliers. The quartiles analysis (Dunn, 2022a) is used to identify and remove outliers univariately in the key process variables. This removal is reflected on all the other variables. The quartiles analysis is based on the median and is therefore independent from the distribution and more robust (and efficient) towards outliers detection.

Additionally, a spatial correlation-based outlier detection analysis is carried out using principal component analysis (PCA) to validate the results of the quartile analysis. The PCA identify the residual values by extracting the principal components of the dataset, and these are evaluated through detection mechanism such as Hotelling T^2 score and squared prediction error to find important discordant data points – observations falling outside the (95% or 99%) confidence level of the model (Krishnamurthi et al., 2020). The Hotelling's 95 or 99-percentile appears as an ellipse on the score chart, showing how each observation fits within the model relative to all the others. Data points around the outliers might also be removed to ensure that the period of disruption, i.e., the period of abrupt change in the signal, is covered. Once identified and removed, abnormal data points may be corrected using imputation. This value can be predicted using previous data, interpolation or by using the mean value. Lastly, abnormal data can also be labeled as missing data or non-naturally occurring number (-99999). References that considered PCA for outlier removal include (Barton & Lennox, 2022); Joe Qin et al. (2021)

Afterwards, in order to reduce noise that affect a signal by making it less representative of the true process state (random errors) and, therefore, distinguish the process trend, signals are filtered. Various sources of noise add a random component to the signal fluctuations. To achieve that, the wavelet transform are used. Some approaches reported in the literature are compared in Thibault, Émilie et al. (2023). It is highlighted that low-pass filters do not have the ability to reduce noise present at different frequencies and that lead to the development of multi-level analysis. In this category, the use of wavelet transform is preferred to that of short time Fourier transform because of the weaknesses towards temporal localization of the latter. However, the quality of wavelet transform denoising depends on the tuning of its control parameters : signal cut-off level (scale), wavelet function and threshold parameters. Thibault, Émilie, Lebreux Désilets, Francis, et al. (2023) proposed an approach to identify optimal wavelet parameters for industrial process in order to retain only the trend associated with the operation of the process. At last, to ensure that the filtering does not induce unwanted lag, a second lag analysis might be performed.

Similarly as for the outlier detection, the consideration of the inherent variability of the data from the top-down operating regimes definition before noise reduction can benefit this data processing step as the process noise amplitude can fluctuate with operating conditions, and therefore the filter parameters must be adapted (Rhinehart, 2013).

Unit-wide steady-state detection

Industrial process signals describing a process cover both transient and steady-state operations. The former happen between steady-state operating periods and occur because of environmental changes, upsets, changes in setpoints, etc. Plants are tracking and reporting what is happening during these transient periods; the analysis of these periods includes how long they lasted, how to minimize the transition time and the asset reliability (equipment failures happening during transient states). However, this framework focuses on steady-state periods to provides a better

understanding of how the process operates as well as significant information for offline management decision-making to improve the process analytics. Therefore, transient periods are not considered.

In order to detect when a process is operating under PSS conditions, the first step consist of the identification of critical variables that will be used as the basis for the steady-state detection (SSD). This identification is done using process knowledge. The variables should be those that best characterize the stability of the system and should additionally be independent, i.e., controlled by different setpoints or variables, and not correlated, i.e., don't vary together at the same time. When all the key process variables are simultaneously in steady-state, the system is said to have reached PSS conditions. An alternate solution would be to claim steady-state if some percentage of the critical variables are at steady-state. This percentage value would vary depending on the complexity of the process.

The SSD algorithm employed as part of this framework is inspired by the one presented by Kelly and Hedengren in 2013 (Kelly, J. & Hedengren, 2013). This window-based algorithm utilizes the Student-t test to determine if the difference between the process signal value and its mean is above or below the standard deviation times its statistical critical value – below would mean that the time instant is steady whereas above would indicate that it is unsteady.

This approach offers a practical, efficient and robust way to detect PSS conditions, the implementation is computationally simple (easy to apply/implement or easily implemented), the criteria are generalizable, the results are reproducible, easily interpretable and reliable, and lastly, process knowledge can and must be combined with the approach – there are adjustable parameters.

Furthermore, this method can accommodate noisy data, multivariable analysis and is scale independent. The computational load imposed by the approach is minimal as it only involves calculation of a mean, standard deviation, and slope. Lastly, this effective method gained industrial acceptance as it is simple both in concept and implementation. It has been extensively applied to commercial/industrial-scale multivariate processes/processing units. The details of this algorithm are now presented.

It is assumed that any process signal may be represented as

$$x_t = mt + \mu + a_t \quad (1)$$

Where m is a non-zero slope, t the relative time within a window (the cycle count), mt is the deterministic drift component, μ is the sample mean or arithmetic average over the time-window with zero slope and a_t is the independent and identically distributed random error series or white-noise sequence with zero mean and standard deviation σ_a . The t index indicates the cycle at which the sample is collected.

By taking the discrete difference of x_t and its immediate past time-shifted value x_{t-1} , the following equation is obtained

$$x_t - x_{t-1} = m + a_t - a_{t-1} \quad (2)$$

Where $a_t - a_{t-1}$ has an expected value of zero with a standard deviation of $2\sigma_a$ by definition. Therefore, the drift slope m can be estimated as

$$m = \frac{1}{n-1} \sum_{t=2}^n x_t - x_{t-1} \quad (3)$$

with n sampled values in the window which are equally spaced in time.

The intercept or mean μ is obtained through equation (4)

$$\mu = \frac{1}{n} \left(\sum_{t=1}^n x_t - m \sum_{t=1}^n t \right) \quad (4)$$

Next, the standard deviation of the noise is estimated as

$$\sigma_a = \sqrt{\frac{1}{n-2} \sum_{t=1}^n (x_t - mt - \mu)^2} \quad (5)$$

The latter could also be provided to the algorithm if it is known by the user, and thus equation (5) is ignored. The standard deviation obtained from this equation is rather liberal (looser); it will most likely be bigger than an externally supplied standard deviation that could be more conservative (tighter/smaller).

Both the mean and standard-deviation are corrected for the drift component.

Lastly, considering a specified Student-t critical or threshold value at a particular significance level α and degrees-of-freedom n (the DOF is the number of sample in the time-window), the null hypothesis that the process signal is steady may be tested. Therefore, if

$$|x_t - \mu| \leq t_{critical} \sigma_a \text{ then } y_t = 1 \text{ else } y_t = 0 \quad (6)$$

The null hypothesis is rejected (the window is flagged as unsteady) when $y_t = 0$ whereas a value of one means that the null hypothesis is accepted, all of the points in the time-window are deemed to be at steady-state. Therefore, if the measurement minus the estimated mean divided by the standard deviation is greater than the $t_{critical}$, then one can be $(1 - \alpha)\%$ confident that it is unsteady. The null hypothesis is presuming that there is no autocorrelation in the time-series process data. If there is autocorrelation, the variance will be inflated. Therefore, to ensure that the approach will be insensitive and immune to any autocorrelation in data fluctuations, providing the standard deviation is recommended.

After performing the Student-t test individually on all data points of a variable (tag), the algorithm returns the frequency of $y_t = 1$. The sum of y_t divided by n (the non-missing data amount number in the time-window) represents the probability of the null hypothesis to be true; it is the percentage of time within the window that the process variable is deemed to be at steady-state. For instance, a value of 95% would indicate that 5% of the points are not at steady-state.

The algorithm calls this routine (equations 1 to 6) for each time-window with an increment of one time-interval. In other words, all equations are computed over the time-window continuously (for all samples in the monitoring horizon) given a sliding window. At every time points, the algorithm is looking back at the time-window minus one data point as it includes itself. In the end, every data point has a steady-state flag.

In this algorithm, the samples with time index lower than the time-window (1 to window-1) always have a returned probability of 0 to be in steady-state as there is not enough past data, they are always declared to be unsteady; the algorithm is using a sliding-window that is looking back at the window size. Therefore, after the sample (window), there is enough points so that the algorithm is able to evaluate the probability.

This methodology consider domain-knowledge; it is a combination of process expertise and statistics. The algorithm is subjective as there are a lot of parameters that may impact whether or not a data point accept or reject the null hypothesis, i.e., the process is at steady-state : the number of samples in the time-window, the alpha value (for the threshold, the Student-t critical value) and the standard deviation. In the algorithm (Kelly, J. & Hedengren, 2013), the cut-off probability (the sum of y_t over n) is also tunable, it could be a setting of the algorithm, however, in this

paper, the control limit for the probability is assumed to be $1 - \alpha$. Therefore, for a given time-window, a probability of being steady is assign by counting the number of time-points that exceed the 90%, 95% or 99% confidence-interval defined by the Student-t statistic. If the probability is greater than some upper limit such as 90%, 95% or 99%, then the time-samples are declared steady, else unsteady.

The significance level α represents the probability of a type I error, or the probability of rejecting the null hypothesis when it is true (false positive/false alarm). Therefore, a 95% confidence interval means that when the null hypothesis is actually true, it will be falsely identified as being unsteady (rejected) 5% of the time. Therefore, alpha is related to the importance of a process value being truly at steady-state. There is also the type II error (significance level β), or the probability of accepting the null hypothesis when it is false (false negative/missing alarm). Theoretically, if steady-state is critical for a specific application, the SSD must ensure a low probability of a type II error, however, this algorithm doesn't consider it.

As part of this algorithm, the incoming process signals must contain contiguous data points. Therefore, missing data points are given a non-naturally occurring number (NNON) of -99999. These NNON are always considered/defaulted as unsteady. Furthermore, all the equations presented earlier are handling missing data as they are assessed considering only non-missing data points; there must be at least the number of sample in the time-window of non-missing data for the SSD algorithm to proceed, else it returns a probability of 0 for unsteady.

The SSD algorithm is performed on all the reference variables, univariately, but always considers the presence of the others. To manage multiple process signal that collectively determine whether a system is steady, i.e., multi-univariate SSD, the individual significance level α_i is evaluated based on equation (7). The validity of this equation is predicated on independence of all key process variables. The sole difference between strict univariate and multi-univariate SSD (where there are two or more key independent process variables considered) is the threshold value (more specifically the alpha value). The latter is reduced from the overall process alpha value following the Sidak significance level adjustment equation (Kelly, J. & Hedengren, 2013)

$$\alpha'_i = 1 - \sqrt[k]{1 - \alpha_{process}} \quad (7)$$

where k is the number of key independent variables, $\alpha_{process}$ is the desired overall level of significance, α_i is the required level of significance for each individual variable. Therefore, using α'_i in the Student-t statistic instead of $\alpha_{process}$ results in a larger critical value for the same DOF for an individual signal. In a nutshell, the exact same calculation are used for each variable, however, when looking at more than one, α_i is used instead in the Student-t statistic. In a multivariable analysis, we are more cautious about rejecting steady-state conditions and more accepting towards the acceptance of steady-state conditions.

Cross correlation between variables is handled by including key process variables that are not correlated, therefore strategically selecting the variables. Additionally, since cross correlation affects the statistical level of significance, the Sidak adjustment is considered in the SSD algorithm. Otherwise, to properly handle cross-correlation across multiple tags, a multivariate statistic like the Hotelling should be used.

Following the detection of PSS periods, a threshold (criteria) for the minimal period of time for the steady-state operation to be deemed representative is considered. The minimal PSS period is established based on process knowledge; this criteria is often set as the average time required to reach steady-state in the process. This is a generally accepted justification for the threshold as it is generalizable, reproducible and robust.

As part of this steady-state detection algorithm, a start, stop and span range are calculated for every steady-state periods (Figure 7). The algorithm is able to keep track of the date where steady-state periods detected using a time

index vector. The start range returns the starting index when y_t vector first becomes steady for all found spans, then the stop range returns the time index when the last contiguous $y_t = 1$ is obtained, i.e., when the y_t vector switches from 1 to 0, and the span range returns the number of data point or time index in the interval (the length of these spans in terms of the number of indexes when it is contiguously steady). The span range considers additionally the number of data point in the time-window. Finally, SSD algorithm return some summary metrics based on the number of indexes per span such as the maximum and minimum number of data points in a span, the span mean, mode, median and variance. All data windows or span deemed acceptable (how reached a minimal period of time) are then reconciled. This idea is consistent with the fact that in process operations, there are sporadic and persistent steady-state periods. When looking for something that is persistently steady, a minimal period for SS operation should be considered.

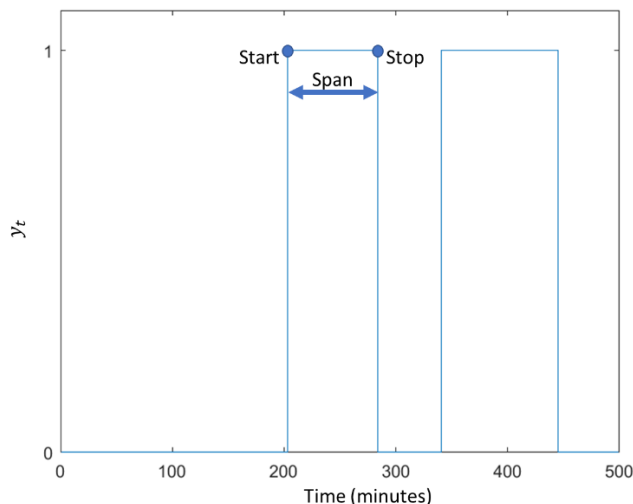


Figure 7 – Start, stop and span representation

Unit-wide data reconciliation

Following SSD, the remaining time-series data are reconciled. Data reconciliation (DR) consist of identifying measurement systematic (gross) error – inconsistencies with respect to known conservation laws – and more specifically persistent gross error. This is done by fitting the data to the theoretical material and energy balances. This analysis does indicate the presence of systematic error, but doesn't specify which instruments (sensors) are wrong – and whether or not there are more than one sensor in fault.

The primary purpose of applying DR is to improve industrial plant data considering the inherent uncertainty and complexity in the process measurement systems. This is done by making sure they satisfy all material, energy and momentum equality constraints or balances and any other inequality constraints or bounds that may be justifiably included. Therefore, DR leads to the identification/diagnose of defective, faulty and/or inconsistent measuring instrument. Furthermore, DR provide estimates of unmeasured quantities and qualities to be used, for instance, in daily, weekly or monthly process, production and/or yield accounting reports. When substituting the raw measured values by the reconciled and estimated (regressed) unmeasured variables, all balances equal zero.

The process of data reconciliation aims to align measured process data with the principles of conservation of matter. By solely relying on these conservation laws, gross errors can be identified. However, if these laws do not reveal any errors, it does not mean that the data are necessarily free of gross errors. More complex models, incorporating transport phenomena, fluid mechanics, and reaction kinetics details, may uncover additional errors. Material, energy

and momentum balances are typically just a subset of the available constraints, and adding more constraints may uncover errors that were not detected using only these three (Kelly, Jeffrey Dean, 2004).

As part of the framework proposed here, the DR is performed using IMPL-DATA (Kelly, Jeffrey Dean, 2023). The latter is a data reconciliation modeling and solving platform, implemented in Microsoft Excel, to perform any type of unit-wide and plant-wide nonlinear data reconciliation application in off-line and on-line environments. IMPL-DATA is specifically streamlined and suited to the interactive workflow of DR in terms of its gross error detection, identification and elimination capabilities. Additionally, it contains a smart model building capability to configure small to large nonlinear data reconciliation problems. The straightforward formulation of these models is based on *units* such as mixers, splitters, processes, etc. interconnected via *streams* where quantity and quality meters (flows, holdups, densities, etc.) may be assigned or mapped.

Prior to running the DR algorithm, the process variables are categorized as measured, unmeasured or fixed (constant in the process). After running the solver, measured variables can be indicated as redundant or not whereas unmeasured variables could be marked as observable or not. This classification highlights which variables can be reconciled (redundant variables), which ones can be estimated (observables) and, finally, those whose reconciliation is not possible and whose accuracy remains unchanged (non-redundant). A measurement is redundant when, if it is removed (marked as unmeasured), that value/variable is still observable. On the other hand, an observable variable is one that is uniquely calculatable from the model and measurements. Even if data reconciliation can only take place if there is redundancy⁷ in the system, i.e., there exists more constraints or model equations than unmeasured variables, observability is more important. Among unmeasured variables, defining which may be calculated (which one are solvable) is required. Indeed, unobservable variables are non-unique, not reliable and their values are arbitrary in the sense that they only satisfy the model constraints numerically.

If there are unobservable variables in the system, there are three approaches to remove the unobservabilities : (1) add new sensors which requires capital investments, (2) add more equations if possible and (3) reconfigure and simplify the model to remove the unobservable variables, i.e., the model perhaps can be too granular.

Following variables classification, a data reconciliation model based on physical theoretical models (thermodynamic, material and energy balance models) is built to identify inconsistencies. This model represent the flowsheet of the plant or sub-plant, it is a subset of the equations used in a process simulation. This step requires a profound comprehension of the process. DR model relates to the fundamental laws of material, energy and momentum conservation, and it is also dependent of the processing operations (operating modes, grades, equipment set-ups, start-ups, shutdowns and switchovers as well as recycle/recirculation loops via bypassing, repetitive cleaning and purging, etc.). From a DR perspective, these variations in the processing operations may be mathematically expressed using the notion of temporary stream variables (versus permanent) with zero/one, open/closed, on/off or active/inactive stream-switches. Temporary streams can be switched on or off depending on state of the process. Usually, industrial plant operations know a priori when mode, grade and/or any other type of processing logic changes will occur via their planning, scheduling and coordinating department. As such, before, during or after the changeover occurs, manual and/or automatic indicators such as valve actuator positions, key process conditions, pump/compressor starts/stops, etc. will be available to aid in the determination of when and which stream-switches changed. The DR model represent all possible streams, both permanent and temporary. Most of the streams are permanent in industrial processes – these are represented as solid lines in Figure 7. Nevertheless,

⁷ The data reconciliation algorithm still works even if the DOF is null, however all the sensors become non-redundant.

when the process conditions changes (going into another operating mode), there are streams switches or temporary streams (represented in dotted lines in Figure 8). Thus, when a process switches to a different operating mode, some flows are no longer required whereas other may be opened just for the time being. Therefore, handling those stream-switch configurations can reliably indicate a certain mode, grade or changeover that occurred and thus reducing the model gross errors possibility, plausibility and probability of occurrence.

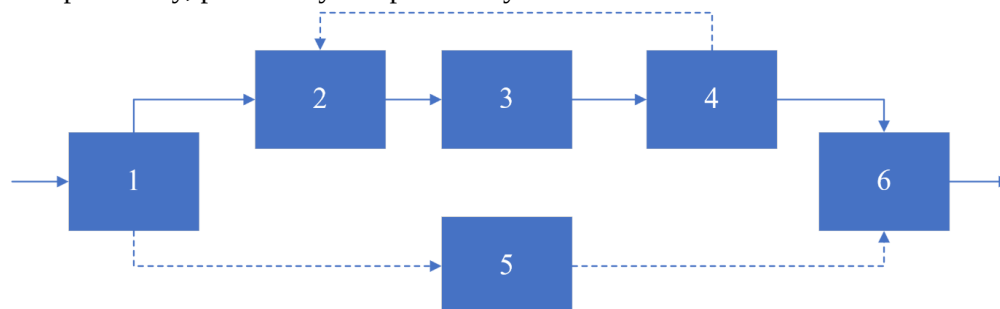


Figure 8 – Example of a block flow diagram to present permanent and temporary streams

The next step is to validate the model, i.e., making sure that there are no model errors. Moreover, it should be validated for each “top-down” operating regime. Indeed, in order to properly validate the model, the operating regimes should be acknowledged since for every operating mode, the operations change, there might be different streams (temporary streams). Therefore, if the operating mode changes, then the model could be wrong if it is not changed properly. Hence, the temporary streams due to operating-mode or grade switches are properly accounted for in the model structure.

Ultimately, before validating the model, both model and measurement gross error are to be expected in the process data sets. However, once it is done, data are only left with measurement gross error; there shouldn’t be any more modelling errors. Moreover, when the process is unsteady, there will be model gross error because the mass balance equation won’t equal zero. Therefore, in order to only identify measurements gross error, unsteady periods are excluded from the analysis, i.e., this is one of the reason SSD is required.

As part of the model validation concept, a preliminary validation stage could be performed if a simulation of the plant process (or some process units) is available. The simulated values would be regressed against the DR model. If no gross errors are present, then a proper (or the actual) validation is performed based on real process data. Afterwards, the gross error detection (GED) might take place.

All three concepts (pre-validation, validation and GED) can use initial values or starting guesses from the variables from a simulation base case. These may be used every time the DR is run, as they “prime” the DR solver with default results. The initial values are used for a warm start, they can influence the solution since it is a non-convex problem (i.e., multiple local optima exists), convergence to the global optimum cannot be guaranteed. Hence, the starting guesses may find a different local solution. Lastly, these values are only used as default results and are updated during the reconciliation solving.

Once the process model is validated, it is used to analyze typically averaged or aggregated steady-state industrial process data. All the time-windows that are found to be statistically steady are reconciled. The steady-state data reconciliation (SSDR) is computed based on the span average which is the average over the multiple time-steps declared to be steady in the span, whereby the SSDR is executed for the number of spans detected.

Data reconciliation consider the impact of process measurement instrumentation by using a tolerance (weight) metric, i.e., the raw sensor or measurement variances are supplied. This metric involve the sensor precision,

accuracy and reliability. It considers the type and brand of instrumentation and the measurement data quality. The measurement error tolerance is two times the standard-deviation. Hence the variance is equal to the square of half the tolerance.

Data reconciliation is based on statistics. A well-established method to detect one or several gross errors is the objective function or global test (Narasimhan & Jordache, 2000). This analysis is a multivariate test. The null hypothesis for DR is there is no gross error. This hypothesis is rejected when the objective function is greater than the Chi-squared statistic (χ^2). On the other hand, if its value is less than or equal to its Chi-squared upper control limit, there is no statistically detectable gross error, H_0 is accepted – that means that statistically, all of the measurements are consistent with the model. For the Chi-squared statistic, the degree-of-freedom λ equal to the number of balance equations minus the number independent unmeasured variables. It can be assumed that all unmeasured variables are independent, which means that they are all observable. Coupled with a specified level-of-significance α , yields its critical value χ^2 and provides an upper bound or threshold limit for the hypothesis testing.

When DR is performed, there are an infinite number of feasible solutions that satisfies the material, energy and momentum balance. Every feasible solution is consistent, meaning that it satisfies all of the balances and constraints. The algorithm seeks the one that minimizes an objective function that reads a weighted sum of squares of residuals:

$$\text{Objective function} = \sum \frac{(\text{raw measurement} - \text{reconciled value})^2}{\text{raw variance}} \leq \chi_{\alpha, \lambda}^2 \quad (8)$$

In Equation 8, the adjustments of measured variables are weighted, squared, and minimized. However, unmeasured variables do not contribute to the objective function since their weight is zero (0.0). On the other hand, one of the most important information that comes out of the data reconciliation solving process is the reconciled variances (calculated via the propagation of errors) of all measured variables. The latter provide the gross error detection statistics presented below.

If the null hypothesis H_0 is rejected, the first hypothesis alternative, H_1 , is that there is only one gross error whereas the second alternative, H_2 , is stating that two or more gross errors exist. Therefore, employing the maximum power measurement test statistic (MPMT) give the opportunity to assess where the GE are in the data set (Narasimhan & Jordache, 2000). This is achieved by comparing the MPMT values to a critical value (the Student-t distribution) for statistical significance for each raw variable. The MPMT is calculated as follow :

$$\text{MPMT} = \sqrt{\frac{(\text{raw measurement} - \text{reconciled value})^2}{\text{reconciled variance}}} \leq \text{Student}_{\alpha', \nu} \quad (9)$$

Where α' is the Sidak significance level adjustment and ν is the number of measurements (sensors).

Hence, there is the overall or global hypothesis that there are no gross errors, then there are the individual hypothesis tests for each sensor; the hypothesis test is whether or not an individual sensor is in gross error. Following the calculation of MPMT for every measurement (sensor), the square of the MPMT maximum absolute value is subtracted from the DR objective function value (weighted sum of squares of residuals).

$$\text{Objective function} - \max(|\text{MPMT}^2|) \leq \chi_{\alpha, \lambda-1}^2 \quad (10)$$

If that difference is less than or equal the Chi-squared critical value (with one less DOF), H_1 is accepted, and if it is greater, then H_2 is accepted. When H_1 is accepted, the statistical power of the MPMT can isolate the most-likely bad sensor, it is a reliable identifier of the sensor with the gross error. If H_2 is accepted, then a combinatorial search,

also known as a subset-selection enumeration search, is necessary to identify multiple gross errors although serial elimination (remove or delete one measurement at the time) may or may not be effective (Kelly, Jeffrey D. & Zyngier, 2008). In other words, there is no easy way to find multiple gross errors, it becomes a combinatorial problem, and it is even more difficult as the number of expected gross errors needs to be known a priori. Nevertheless, this statistical test is useful when there is a single gross error in the system. In this case, the MPMT statistic can reliably indicate the problematic sensor or instrument – it will be flagged as being in gross error. However, if multiple gross errors exist, then maximum power is not guaranteed, and a more sophisticated analysis is necessary.

Serial elimination can be used to identify two or more gross errors (Bagajewicz, M. et al., 2010); the first gross error is identified using the maximum absolute MPMT statistics, its corresponding measurement is eliminated and the DR is repeated.

Consequently, limiting the DR to a scope where one or even two gross errors are expected is the most tractable approach to accurately and precisely detecting, identifying and eliminating true gross errors in industrial plant data. In that regard, the DR routine is intended to be executed on a regular basis (say on an hourly average frequency) on-line in real-time at a unit-wide level for volume-only and/or volume-density-mass reconciliation. It is also possible to collect past measurement data, and run the DR algorithm off-line (in the past) on hourly averages. A daily plant-wide DR would highlight a lot of gross errors and closing the balance would be nearly impossible. Therefore, by running the GED more often, i.e., on a shorter time interval (every hour instead of every day), and on a smaller scope/scale (unit-wide instead of plant-wide), the ability to diagnose gross errors easily and accurately increases.

As part of the DR algorithm, a distinction is made between sporadic (transient) and sustained (persistent) gross error (Tong & Crowe, 1996) by pairing the temporal context to the DR problem. At first, when a gross error is detected, whether it is sporadic or sustained is unknown. In order to distinguish between the two, the number of times a sensor shows up as being inconsistent with the material and energy balances over a monitoring horizon (the probability) is computed; the occurrence is logged and recorded. This metric identifies the most sustained gross errors by ranking the probability of occurrence for the sensors. IMPL-DATA is detecting which of the sensors are persistently in gross error versus sporadically. Generally speaking, the plant personnel will only act on the sustained ones. In other words, data reconciliation focuses on diagnosing persistently or sustained defective, faulty or bad measurements. Data reconciliation is used to screen the data for persistent gross error and for plant personnel to fix them over time by recalibrating the sensor for instance. And as mentioned, unless the gross errors are eliminated, the only recourse is to use the raw values with regard to further analysis and reporting since the reconciled values are unreliable.

Data reconciliation is primarily focused on the detection, identification and elimination of gross errors. There seems to be a wide-spread misconception in industry that after performing data reconciliation on process data, the reconciled values are better than their raw measurements, and that reconciled data should be used for subsequent analysis. In the absence of model gross errors, the reconciled values are by definition consistent with respect to its sensor certainty and model constraints and may be considered as conforming, to some degree, to their true values. Yet, if the dataset contain gross error, all the reconciled numbers may in fact be worse than the raw readings given that all (weighted) least squares methods systematically smear or spread the effects of gross errors throughout the model and ultimately corrupt and distort the accuracy of the reconciled data. Even if data reconciliation provides reconciled values, they are rarely employed; the end game of data reconciliation is never to get reconciled values. This is explained by the fact that if there is one gross error in the system (most of the time there are more than one), all the reconciled values are wrong – if one measurement is bad, they are all bad, the whole recipe is wrong. When there is gross errors in a dataset, the latter will propagate into the whole system. Additionally, when there are more

than one gross error in the system, it is practically impossible to determine which one are bad. It may sound counter intuitive, but it is better to use the raw, unreconciled measurements for all subsequent analysis such as optimization.

However, if it turns out that after DR, there are no statistically detectable gross error (all the sensors are consistent and all of their errors are random), then the raw values and the reconciled values can both be used. The only difference between the raw and the reconciled numbers, is that the reconciled values completely satisfy statistically all of the constraints, they close all of the material, energy and momentum balances perfectly, whereas the raw values, technically do not.

Nevertheless, the objective function (global test) could either be the sum of squares of residuals (L2- or Euclidean-norm), i.e., raw measurements minus reconciled, or it may be the sum of weighted absolute deviations (L1- or Manhattan-norm), i.e., the absolute value of the residual divided by the standard deviation (Zhan et al., 2012). When performing DR with the 1-norm, the reconciled values are less impacted and sensitive to gross errors (biases or outliers). Hence, if one really wishes to use the reconciled values, while keeping in mind that it is not possible to be a 100% gross error-free, the 1-norm is an interesting alternative and may be considered instead of the 2-norm whereby the reconciled values may exhibit less corruption and distortion.

As a final consideration, the data reconciliation stage will avoid removing sensors and instruments. Instead, the sensors declared as persistently in gross error (over a long period) following the DR are flagged and are only ignored temporarily in the DR problem. Specifically, gross error measurement are set to unmeasured variable in the reconciliation, which is identical to eliminating the sensor value. In order to achieve that, the measurement raw variance is set to be very large, hence its weight becomes null in the objective function. Additionally to the probability of occurrence, there is also the notion of severity, i.e., the degree to which the sensors exceed their statistical threshold tolerances. Therefore, based on this quantitative analysis that consider both metrics, justified and appropriate action may be taken to eliminate the most likely gross errors by sensor recalibration, reconfiguration, repair, replacement or modification of its tolerance or uncertainty.

In summary, gross error detection is applied to improve the accuracy in measured data and to identify instrumentation problems that require special maintenance and correction. Moreover, detection of persistent gross errors can reduce maintenance costs and provide smoother plant operation. Eliminating gross error allows to close the material balances to within statistical error every hour, shift, day, etc.

Operating regime detection and identification

The last step of the proposed framework is to translate processed steady-state data into operating regime. This gives a snapshot of the plant and, therefore, the idea is to perform a multivariate data analysis to detect what is the operating regime when the plant is operating that specific way. In order to classify the snapshots, i.e., to identify the operating regime, independent key process variables (variables of interests that are important in a process unit) that distinguish the operating regimes are identified by a process experts.

Starting from a “top-down” perspective, the steady-state processed data are used to build a principal component analysis model. PCA exploits historical process data and discovers hidden phenomena that may be useful for detecting unexpected/novel or unknown operating conditions. The principal component analysis allows to represent process variability and identify and understand significant correlations (between variables) that are inherent to the plant operation. It groups together variables with similar characteristics into clusters unmindful of the link between

consecutive time-steps. Standard PCA ignores dynamicity (it is not a time-series technique), it is treating all data points (time increments) equivalently, regardless of how far apart they are timewise (Harrison & Stuart, 2011).

Then, the principal components scores are used as input variables in a *k*-means clustering algorithm. Only the components that explain an important proportion (90% for example) of the variation among the operation variables are selected. Trial and error is used to input the number of clusters detected by the *k*-means algorithm.

These clustered data are assessed by domain experts to highlight what makes a cluster different from another and assess whether they all represent distinct operating regimes. The cluster analysis gives process insight to make the process more effective, economical, efficient, helps understanding the process, phenomenon, deviation, and improves decision-making.

Additionally, the contribution analysis and the loading chart (Dunn, 2022b) are employed to identify which variables characterize each cluster and explain the variability between steady-state operating regimes; hence, which variables are important for a particular regime, i.e., which variables are affected by the different clusters. Lastly, the experts confirm the operating regimes and explain the fundamental drivers behind them, i.e., the variable of interest helps identifying the clusters.

These “bottom-up” operating regimes based on a data-driven approach as well as domain-knowledge offer the opportunity to interpret and understand the process in more depth. Those regimes could be linked with the way the plant is run by operators since they might operate the plant differently when producing the same product.

The identified operating regimes should be mutually exclusive. Hence, some variables specific values must uniquely be attributable to one operating regime. In other words, there is a unique set of parameters values to distinguish each regime. All operating regimes are unique, and they cover all the possibilities for characterization variables (attributes).

Operating regime detection and identification allows improving process knowledge as analyzing each cluster may lead to discover unknown operation mode. On the other hand, domain knowledge is required to interpret the patterns extracted from the principal component analysis and to achieve an accurate representation of the operating regimes in complex processes. The PCA results interpretation must be performed by process experts, and they should be cautious when doing so as the PCA blindly finds correlations, however the physical reality of the process plays no part in generating the statistical outputs (Harrison & Stuart, 2011). Therefore, PCA results should be interpreted based on an understanding of the process fundamentals. This is a task for process experts; the “bottom-up” operating regimes are detected and identified by merging data-driven approach and process knowledge.

As a final observation, performing all the previously mentioned data management steps on the raw process data – including abnormal operations removal and noise reduction – maximizes the usefulness and veracity of multivariate analysis as a statistical analysis is only as good as the raw data and pre-treated data represent the process more accurately (Harrison & Stuart, 2011). Multivariate analysis (such as PCA) are entirely data-driven techniques, and thus highly susceptible to the issue of “garbage in, garbage out”. They are sensitive to outliers and instrument drift; the latter can appear as a long-term trend to which multivariate algorithm could blindly ascribes statistical significance (Harrison & Stuart, 2011). Lastly, noise in process data represent a different problem. As each data point (one-hour period for instance) is treated as a separate observation which bear little resemblance to each other before bringing them all to the frequency analysis, some kind of smoothing is therefore required; this filtering improves the multivariate analysis ability to find correlations between variables (Harrison & Stuart, 2011). Therefore, a direct use of the raw data would yield meaningless multivariate analysis results, since the algorithm

could, for instance, blindly attribute most of the correlation to the start/stop phenomenon and not to actual changes in the process (Harrison & Stuart, 2011).

5.2 Application : brownstock washing department in a dissolving pulp mill

An application is demonstrated on data from the brownstock washing of a dissolving pulp mill in order to illustrate the benefits of the proposed framework. This unit is isolated by buffer tanks. Those tanks absorb process fluctuations, so events in one system will not impact downstream systems. Therefore, the unit can be marked as independent from the rest of the process. This system is chosen as it is the one showing the highest redundancy (allows data reconciliation analysis); the pulp and paper mills are famously known for their lack of redundancy. The results from a simulation of this department built by Brucel et al. (2023) are used to increase redundancy.

The implication of each step of the framework are briefly summarized here. This section is divided in accordance with the main parts of the framework; scope definition, data processing, unit-wide steady-state detection, unit-wide data reconciliation and operating regime detection and identification. Each step plays a critical role in the whole framework, and together they allow the use of rectified and reconciled (clean) segmented steady-state data for design decision-making.

Scope definition

The intended use of the data is to assess how many bottom-up regimes can be detected and identified in the brownstock washing (BSW) through the use of the framework when the process is running smoothly (in steady-state). The top-down regime considered is the summertime when a specific grade is being produced. Brownstock washing data that fit this context is collected.

Data processing

In this case, data synchronization and imputation are not required as all data points are sampled at the same rate, i.e., a sampling interval or time-step of 10 minutes (process experts confirmed that a 10-minute sampling interval is adequate for this SSD). A sampling interval that reduces autocorrelation in the time-series process data is used. Additionally, as the pulp crosses the whole unit in a few minutes (the overall residence time of the system is around 10 minutes), there is no need for intra-unit lag correction. This is confirmed by using the lag correction analysis offered in EXPLORE; a straight line means that there is no lag detected (Figure 9). If a lag was present, the graphs would have shown some waves and a cycle.

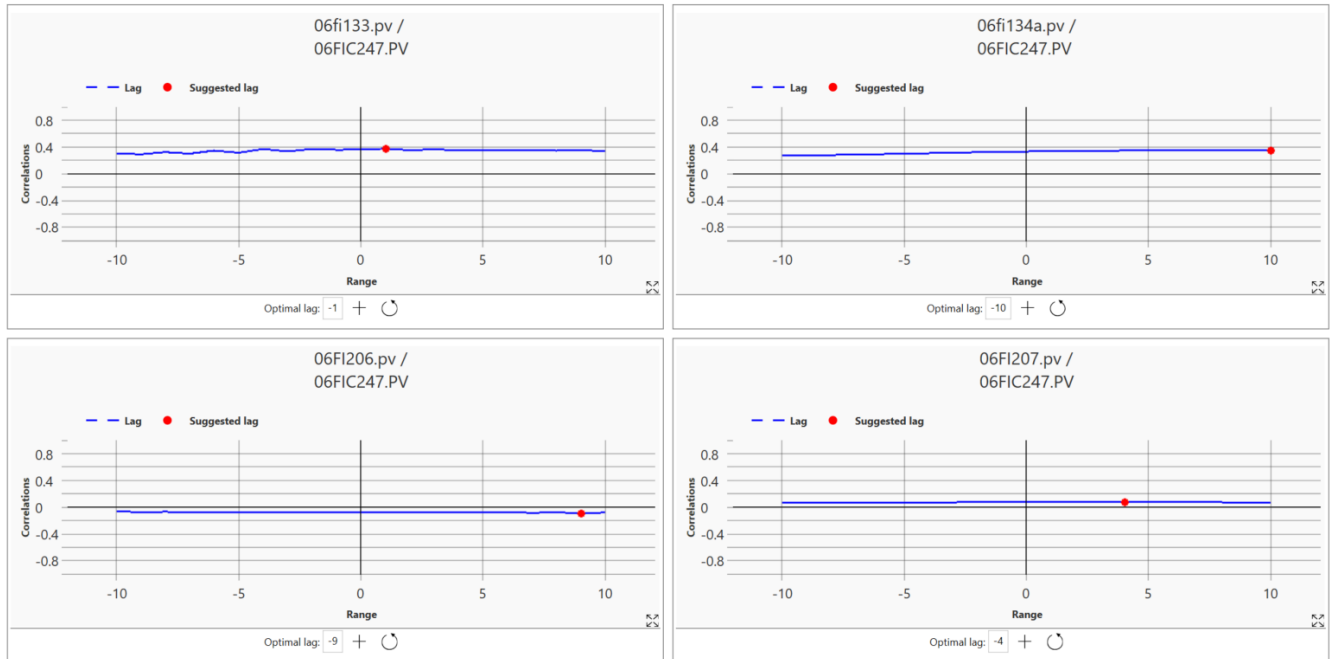


Figure 9 – Lag analysis where X is the number of lags and Y the correlation

From there, start up and shutdown periods are removed. To do so, three independent flow variables that describe the process are used as reference to detect those periods. The latter are removed by using threshold values specified by a process expert. A two- and four-hour time interval are removed respectively before and after each upset periods to account for the time required to go into shutdown and restart the system. These time interval are based on the time required for the thickener consistency measurement to restabilize itself after an upset. Lastly, the removal of process shutdown and start up from reference variables is mirrored onto the other variables used as part of the analysis.

Then outliers are removed using the quartile analysis as well as the principal component analysis (Figure 10). Similarly as what is done for the shutdown and start up periods, key process flow variables are used to detect outliers. This removal is reflected on all the other variables. Having almost 20,000 data points, outliers are not corrected, but rather attributed a NNON value.

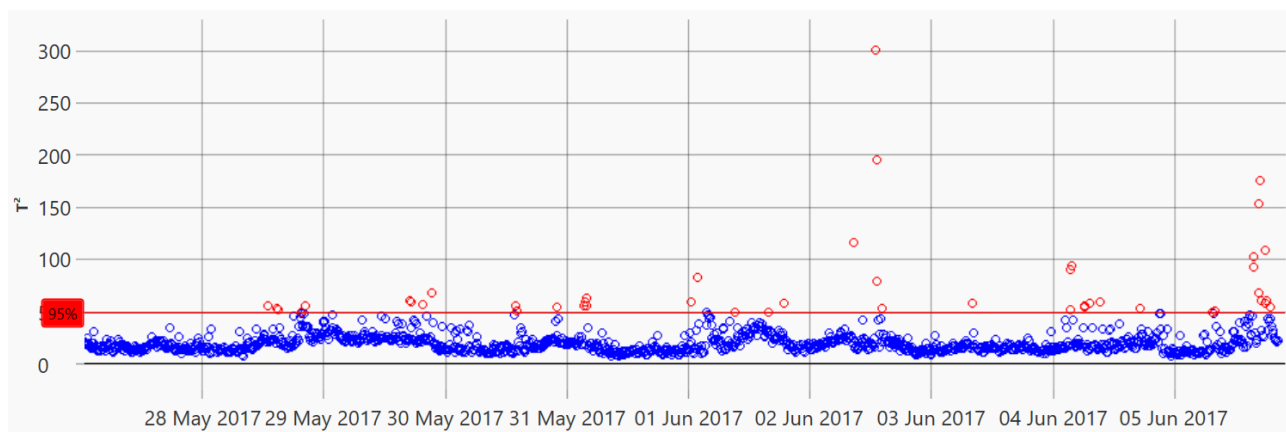


Figure 10 – Outlier removal using principal component analysis

Last of all, the signals are filtered using wavelet transform. Given that the success of wavelet transform denoising relies on the optimal configuration of its control parameters, the selection is done through the use of heuristics presented in Thibault, Émilie, Lebreux Désilets, Francis, et al. (2023).

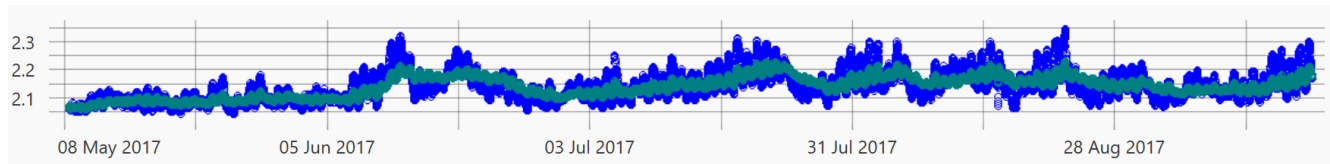


Figure 11 – Consistency signal filtered using wavelet transform with db15, scale 8

Unit-wide steady-state detection

This section describes an industrial implementation of the proposed SSD algorithm. The latter requires a data-vector of time-series data discretized into a time-ordered sequence of uniformly distributed timestep of equal durations with the oldest data point referenced as timestep one (1). The SSD algorithm also necessitates the tuning of the time-window, threshold, standard deviation of the noise (if desired) and cut-off probability; these tuning parameters are part of the configuration of the algorithm. This implies that some prior knowledge and understanding of the process is necessary.

It is not uncommon for experts at the mill (or in plants in general) to be overoptimistic on how well the process unit performs and mentioned how the process runs for many consecutive days in steady-state, possibly overestimating the stability of the plant. There are mainly two factors that may explain this. It might come from the fact that their steady-state analysis is most of the time qualitative. This is what they aim, expect, and wish for, although without performing a quantitative data-driven analysis, there is no way to confirm that the process is indeed at steady-state. On the other hand, detection of steady-state periods might also be a visually subjective process expert decision. However, this visual steady-state recognition approach requires continual human attention, and it is subject to human error. More specifically, noisy measurements, slow process changes, multiple dynamic trends, and change-of-shift timing are features that may compromise the visual interpretation of data. The statistically-based SSD standardize the procedure.

In the context of this application, the monitoring horizon is five months in the summertime whereas the time-window is four hours. The latter should have a number of samples equivalent to three (3) to five (5) times the time constant of the overall process divided by the sampling interval; it should be long enough for the variable to reach steady-state or equilibrium. Here, experts stated that it corresponds to the average time required to reach steady-state in the process. Since the time to reach steady-state varies with operating conditions (Rhinehart, 2013), an average provided by a process expert is used. Too short of a time-window will not give sufficient time to the process to reach some level of stability, thus the steady-state probability will always be low. On the other hand, a too long time-window may lead to the false conclusion that the signal is steady when in fact it is not. Additionally, long windows are not well suited for detecting unsteady behavior with short duration; it is harder to detect. Therefore, the time-window long enough to account for the system dynamics, but short enough to detect undesirable changes in the process value that has short duration.

Additionally, since different variables present different dynamics a slow process should consider a longer time-window, while if the dynamics are fast, the time-window duration should be shorter. Therefore, the time-window is the parameter that can account for both the sampling interval as well as the process dynamics. Furthermore, it is possible for every KPV to have its own time-window size (Moreno, 2010).

The threshold is set at 95% confidence interval (Student-t score). This parameter is related to the importance of a period to be labeled properly; for a critical application, a higher α should be used. Therefore, if the steady-state detection is not extremely critical for particular application, a 5% probability of incorrectly rejecting steady-state is accepted.

Then, even if the standard deviation of the noise could be assessed by equation (5), the latter is provided for all key process variables (KPV) targeted to detect steady-state periods. As the standard deviation of the noise value increases, the maximum steady-state span increases as well – more data are deemed at steady-state, the spans get longer.

The KPV selected by a process expert represent well the process and the operations undertaken by the mill, they have a significant impact on the latter and they should have the least autocorrelation as possible. Experts decide what is the minimum number of KPV required to detect whether the process is at steady-state, and then set up the procedure on that minimal set. For this application, a process expert at the mill mentioned that the best indicator to know if the process is stable is when three distinct pulp flows meet the steady-state conditions at the same time.

The SSD algorithm returns a probability that the data-vector is steady with a probability near one (1) and is unsteady with a probability near zero (0). The determination of the probability limit which indicates a steady or unsteady signal is the responsibility of the user. A suitable cut-off value of whether the process is deemed to be at steady-state depends on the application. In this case, the cut-off probability is set to 95% (probability of the time-window being steady) for each KPV. Once the tuning parameters are established, the SSD function is called for every time-window until the end of the monitoring horizon.

This approach, which uses a combination of good judgment, knowledge and interpretation of the process operations, and statistics, have two different modes. The first one runs every time-step (sampling interval) continuously (every 10 minutes in this case) – it incrementally move over the smallest time step. Therefore, every 10 minutes, the algorithm is looking back over the time-window duration, and it declares if the process is steady or unsteady for the past 4 hours. The second mode is the batch one where the algorithm runs every 4 hours, still looking back, declare the whole time-window steady or unsteady, and then move forward to another time-window to conclude about its steadiness. However, in both cases, the time-window duration is the minimal threshold of time for which the process can be deemed steady or not, and for which the steady-state duration is judged representative; the algorithm is not going to find anything less than the time-window. In other words, every time the process is declared to be in SS, it lasted for the past time-window. As part of this proposed framework, the first mode is preferred.

In order to label a period as steady, all the key process variables must be steady at the same time, i.e., all KPV have a probability higher than 95%. It would also be possible to label a period as being in steady-state if a specified fraction of the KPV are in steady-state at the same time. In this case, since there are only three of those, all three must meet the probability threshold.

As a result, the algorithm evaluated that out of the five (5) months monitoring horizon, the process is in steady-state 32 % of the time. However, a major challenge in this case is the constant starting and stopping of production lines and pieces of equipment. Consequently, knowing that there are sporadic and persistent steady-state periods, a run length of a minimum of 12 hours is targeted to deem the period as persistently steady; when the process runs reasonably steady for a while, these operating data might be used as a basis for decision-making. The run length is the amount of time or samples required to be confident in declaring that the process is persistently steady. In the algorithm, a second routine identifies the contiguous sets of steady data. This routine is a data function that determines the statistics of the contiguous span of when the process is steady; a start, stop and span range are

assessed as well as the maximum and minimum number of data points in a span, the span mean, mode, median and variance. The following table presents information about the contiguous steady-state regimes collection.

Table 1 – Information on the contiguous steady-state regimes collection

	Value
Minimum	24
Maximum	182
Mean	31
Median	26
Mode	24
Variance	214

However, considering only when the process unit is contiguously steady for longer than the run length, a total of 20 spans are found. Once all steady-state periods are detected, they all individually and separately go through the process of data reconciliation to determine if there are measurement gross errors – each one of the contiguous steady-state set respecting the run length becomes a dataset for data reconciliation.

Unit-wide data reconciliation for various datasets

As part of this framework, the data reconciliation problem is modeled from a mathematical programming perspective in opposition to a matrix algebra perspective. From this standpoint, every stream variable has a reconciled value and an adjustment (revision) value. In order to perform data reconciliation, a process model is built and validated. This model consists of a set of equations⁸. There are sensor constraints equations, i.e., adjustment + reconciled = measurement, and model constraints, which are all the laws of conservation of material, energy, and momentum, i.e., mixer, splitter, process, and density equations. The process recycling loops are included in the DR model. These generates unbound variables which lead to an infinite number of solutions. Hence, the simulation results are used to evaluate the split fraction and create a hard constraint.

The equations are a mixture of linear and non-linear constraints. This makes the DR problem more difficult to solve, but it is important to represent the process in its most fundamental way to reduce model gross errors. Having non-linear equations makes the problem non-convex; the problem is subject to local solutions, there are local optimum. Thus, the solver is runed a few times and each time, it randomizes the initial values (starting value generation). This also helps with convergence issues. IMPL-DATA can solve non-linear data reconciliation problem.

Generally, adding constraints to any kind of mathematical minimization optimization problem will either increase the value of the objective function or it will stay the same if the information is redundant. The only way adding a constraint could reduce the objective function value is if the solution is a local optimal (it converges to a different local optimum).

The model is validated when it is consistent with the simulation data (given some tolerance). Hence, the process simulation is used for the validation of the data reconciliation model. When the DR model is validated, it is possible to say in confidence that from now on, only measurement GE will be identified.

In industry, not every stream has a flow and/or consistency measurement. This application is no exception; measurements are very sparse. There are many unmeasured and unobservable variables and not enough

⁸ See Appendix for the model

measurements to identify gross errors. Thus, there is not enough redundancy in the mill; there is redundancy overall, but not around some equipment. Therefore, redundancy is created using the simulation results (this is not a common practice in DR). Hence, a sensor is assigned to all streams; the simulation results are considered as measurements – the simulation approximates the real process. By doing so, observability is increased. There is no level of acceptable observability, but IMPL-DATA have a pre-solve observability functionality that can be applied to better determine the observability, to improve the numerical robustness of the observability detection (Kelly, Jeffrey D. & Zyngier, 2008). The pre-solve algorithm goes through the sparse unmeasured variables incidence matrix and excludes variables that are strongly or guaranteed to be observable in the singleton constraints (Kelly, Jeffrey D. & Zyngier, 2008). Declaring these as observable shrinks the matrix. A large matrix has a large condition number, therefore making the matrix smaller inertly means a smaller condition number, which makes it more numerically reliable. Therefore, the pre-solve provides a smaller matrix that has less constraints and variables to go ahead with the observability detection analysis.

Additionally, to build a DR model, weights are assigned to measurements. These weights reflect the reliability, precision, and accuracy of measurements. For instance, temperature measurements tend to be more accurate than flow measurements, and more specifically, steam flows are more problematic as they are harder to calibrate and need to be corrected for temperature and pressure to get the flows right. On the other hand, liquid flows are generally fairly good. Then, temperature measurements from thermocouple are generally pretty decent, they don't drift much. Next, level measurements from tanks are, most of the time, not even reconciled as plants probably have those values right since the operators wouldn't let tanks overflow or underflow. Lastly, pressure signals are reasonably good and consistency data have a certain range where the sensor seem to work reasonably well.

In this application, only flow and consistency measurements are being reconciled. Measurements are given a tolerance or precision interval which is translated into a raw variance where its inverse is used as an objective function weight in equation (8). As consistency measurements present less variation than flow measurements, and are well controlled, process experts decided to assign a tolerance value of 0.5% to consistency measurements and a tolerance value of 2% to flow measurements. These tolerances are for hard sensors (direct measurements); a tighter tolerance is put when the stream has a physical sensor. However, for values assumed from the simulation (soft sensors), a larger tolerance is considered as there is more uncertainty. Therefore, a suitable tolerance for indirect (soft sensor) consistency measurements is 4% whereas 5% is assumed for flows. Lastly, fixed variables are given a tolerance of 0%.

Following the tolerance analysis, for every variable on all streams, there is something indicating whether it is measured (1) or unmeasured (0). This is the sensor switch. When there are a lot of zeros in the sensor switch, the results start showing unobservability, and some flows present negative values (this is a sign of GE). Therefore, to maximize observability, there is a tradeoff that must be found in labeling variables as unmeasured and getting negative flows. The sensor switch helps in turning off the wrong sensors, either hard or soft; to keep the sensor, a value of one (1) is used. Therefore, even if a sensor is assigned to all streams, a value is assumed only for those having their sensor switch value to one (1).

In addition to the sensor switches, which can indicate whether or not a sensor is good, whether it may be trusted or not, stream switches manage permanent and temporary streams. The latter must be considered to make sure that the reconciliation results make sense. Since processes have different operating regimes, the DR model changes and must therefore allow streams to be turned on and off i.e., it is parameterized in the sense that streams may be active or inactive. A logic is required to manage whether the streams are active or not. For instance, when a stream has a flow sensor, the latter could be used as a stream switch by setting the sensor value to zero when required.

When plants perform data reconciliation, it is mostly based on mass balance. However, most of the flows in plants are volumetric. Therefore, to convert those to mass, density measurements are required. The latter can be assumed, or calculated given an ad-hoc formula, density measurements can come from a simulator, or they could be taken every week or every day at the lab. However, densities are not as good of a measurement as flow or consistency because they are not continuous, they are sampled intermittently. Ultimately, the reliability of the density data is poor primarily because there is little, if any, statistically-driven measurement quality feedback being transferred or relayed back to the engineering, operations, instrumentation, and maintenance departments. Additionally, they could be wrong because they are based on assumptions about the process and operating modes that don't necessarily apply. Given the inherent sparsity in the density measuring system, these densities may be biased (wrong, not accurate), especially if these densities are operating-condition or -mode dependent. Hence, many problems could come from inferring the densities. Performing DR can help remove that bias provided that bad densities are detected, identified, and removed from the data reconciliation problem.

Once all variables are accounted for and the model equations are set, a degree-of-freedom analysis is assessed. A negative value of the degree-of-freedom (number of variables – number of equations or constraints – number of measurements or fixed variables) is expected as the simulation values were used as measurements substitution. This non-linear data reconciliation problem reconciles both volume flow and density simultaneously, involving volume, density, consistency, and mass balances. In the present application, there are 53 sub-units, 107 streams, and 19 flow sensors as well as 4 consistency sensors. The sub-units in the brownstock washing are mixers (18), splitters (9) and processes (26). For mixers, both volume (28) and mass (18) balances or equations are required, splitters count 46 equations, and lastly, process equations have a total of 38 mass balances and 30 volumetric equations. In addition to these, the BSW process unit accounts for 107 density equations. Then, as each stream has a flow, consistency and density value, there are 321 variables. Therefore, the number of variables, equations and measurements (including those assumed from the simulation) yield a DOF of $321 \text{ variables} - 267 \text{ equations} - 90 \text{ measurements} = -36$.

In this analysis, the qualities/intensive (*compositions* – fractions, consistency, *properties* – density, molecular weight and *conditions* – temperature, pressure, velocity) and quantities/extensive (flow, mass inventory, moles, volume, energy and momentum) are kept distinct to have different constraints on the individual variables : volumetric flow, consistency and volumetric flow*density*consistency (to obtain the mass flow). Both variable categories are reconciled simultaneously.

This DR problem only considers material balance (flow in – flow out = 0). However, if there is a hold-up unit in the model (tank, drum, vessel), the hold-up balance (flow in – flow out + opening – closing = 0) could be performed considering measurements of the level (hold-up, inventory). To do so, the opening hold-up is considered as a fixed variable, meaning that it has 0 uncertainty, or its weight is infinity, and then only the closing value is reconciled. Another way to do this would be to reconcile the difference between the opening and the closing.

In summary, when performing data reconciliation, the first step is to find unobservable unmeasured variables. DR would still run with these, but it would calculate numbers that are completely meaningless since they are non-unique. Hence, from a mathematical perspective, there is no issue, however, from an engineering perspective, unobservable unmeasured variables may be problematic. If required, all unmeasured variables can be made observable either by adding assumed values or shrinking the data reconciliation model. Then, after running the algorithm, one must make sure that there is no negative flows greater than constraint and convergence tolerances, because they mean that there are measurement gross errors. In fact, negative flows could mean that their model directions should be reversed. Once all the negative flows have been resolved, then the gross error detection may begin.

For this application, data reconciliation is run offline on averaged steady-state data. However, data reconciliation is generally run online on hourly average steady-state data in order to detect the most persistent or sustained sensor with GE. The execution frequency of DR should be based on how many gross error could happen within that time frame. The goal is to run it when there is zero or no more than one. As a matter of fact, as there are sporadic and persistent GE, a monitoring report is available when reaching the end of the reporting horizon (moment when conclusions about DR is made), such as a shift, a week or a month, to inform process experts about the most problematic sensors. Hence, only persistent GE are recorded or added to the ongoing list; they may as well be ranked. The ratio of the number of H_0 acceptances over the duration of the reporting horizon gives a probability of occurrence and indicates what percentage of the time a sensor is persistently faulty. Therefore, deploying online unit-wide data reconciliation may continuously improve the reliability of process data, assure that the sensor networks is functioning with consistency and integrity, and provide the level of assurance required for descriptive, predictive and prescriptive analytics.

Nevertheless, in this case, DR is run on the averaged steady-state spans of arbitrary duration, but always longer than the run length. Table 2 presents the objective function value (Equation 8) for all spans. According to the total number of DOF and the 95% confidence interval, the Chi-squared statistic is 85.965. The objective function values are all greater than the chi-squared statistic, hence, out of the 20 steady-state span found from SSD, all of them contain gross errors. Column 3 of Table 2 gives the worst (biggest) maximum power measurement test values (Equation 9) across all variables in each span. Lastly, the next column provides the values of Equation 10. Since all values are higher than the Chi-squared critical value (with one less DOF), 84.821, there are more than one gross error in each span. Therefore, it is difficult to isolate the most-likely bad sensors or to reliably identifier of the sensors with the gross error. Considering all the measurements in each span with a MPMT value close to the worst one, it was concluded that the top four persistently problematic sensors across the steady-state spans that would need to be verified are three flow meters (06FIC137, 06FIC152, 06FIC433), and one consistency sensor (06NIC423). This information is transmitted to process experts for them to take a look at the problematic sensors. Knowing which sensors are faulty is important because experts base their analysis (such as optimization) on these data, and until these sensors are fixed, gross errors are present in the datasets.

Spans	Objective function	Worst MPMT value	Obj – MPMT
1	2069.364	30.383	1146.250
2	1960.541	29.919	1065.403
3	1899.865	28.611	1081.287
4	1944.359	28.350	1140.662
5	1870.016	27.920	1090.486
6	1910.728	28.635	1090.786
7	1914.603	29.442	1047.766
8	2144.591	30.767	1198.001
9	2077.743	31.400	1091.786
10	2128.247	30.996	1167.487
11	2106.752	30.351	1185.586
12	2041.512	29.570	1167.146
13	1922.061	29.515	1050.945
14	2033.123	29.566	1158.979
15	1861.946	29.297	1003.613

16	1928.288	29.459	1060.441
17	1685.618	27.045	954.180
18	1652.962	26.277	962.464
19	1706.588	25.994	1030.917
20	1832.242	26.900	1108.613

In Part II, some of the 20 steady-state spans will be further assessed.

Operating regimes detection and identification

In general, in the pulp and paper industry, there is little to no acknowledgement of the fact that a process has many different steady-state operating periods. In other words, operating regimes are not explicitly considered for decision-making (Thibault et al, 2022). However, the value and potential of operating regime detection and identification is well recognized. In the present study, a model based on principal component analysis as well as *k*-means algorithm is used to identify the operating regimes of the brownstock washing department of a dissolving pulp mill. In this application, five months of data are used to detect and identify the process operating regimes.

The clusters apparent on the score chart of the PCA (Figure 12) where confirmed with a *k*-means clustering analysis a detailed in section 5.1. The first and second components separated the observations that are different and gathered the identical observations. The variables that influence these clusters can be observed through a contribution analysis (Figure 13). Analyzing the results with process experts, it is found that the main drivers for the “bottom-up” operating regimes are the pulp level in tanks, its density, and the shower wash water flow rate. The clusters represent changes in the operating conditions. Part II of this publication will analyze further what explain the changes in operation between the different steady-state spans, and how does they affect the operating cost.

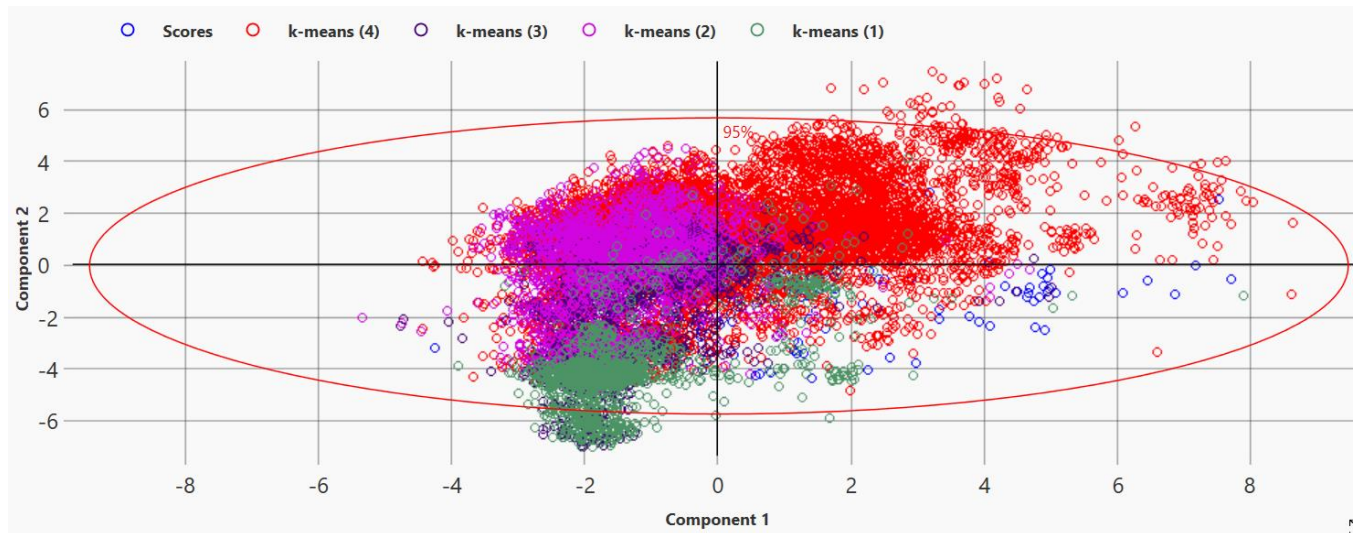


Figure 12 – Score chart showing the clusters detected by *k*-means clustering algorithm

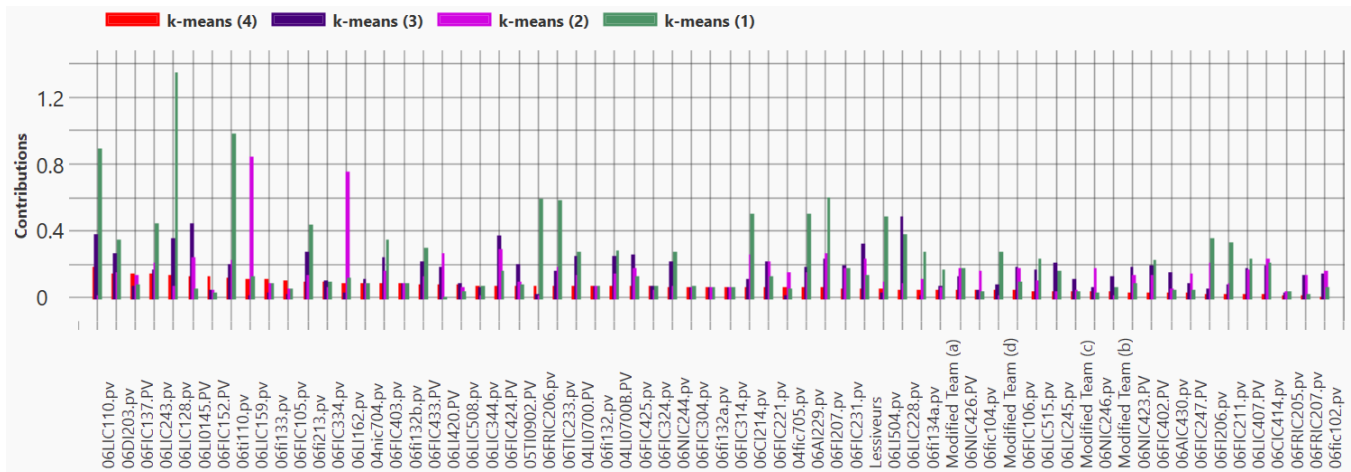


Figure 13 – Contribution analysis of four detected clusters through PCA and *k*-means

Lastly, the loading chart of the first and second components is showed in Figure 14. This chart identified which variables characterized each cluster and explain the variability between the different regimes. Variables close to the center of the chart don't have a lot of importance for component 1 and 2 (in this case) whereas variables away from the center and close to either component will be of great importance – further from the center, more influence they have. Lastly, those located diagonally are influenced by both components. Figure 14 shows for instance that the pulp consistency explain a lot of variability in the first component while the pulp density explains most the variability of the second component.

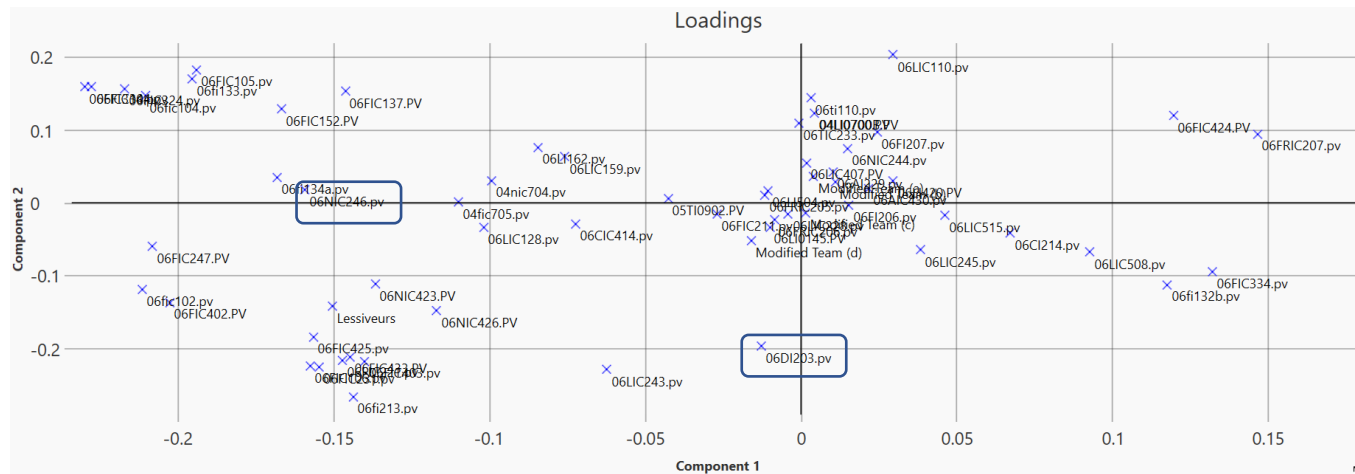


Figure 14 – Loading chart of the first and second components to identify variables of significant influence

Conclusion

With the quantity of process data being collected increasing year over year, actions should be undertaken to put them to good use, and to do that, one must make sure that the data is of good quality. Several authors proposed methodologies to manage, treat and analyze process sensor data and to use them for decision-making (Bellec et al., 2007; Delou et al., 2021; Jiang, Taiwen, Chen, Jasim, et al., 2003; Korbel, Milan, 2011). However, these data processing frameworks do not explicitly consider operating regimes nor mention the concept of unit- or plant-wide analysis, even though industrial applications of operating regime detection are extensive. This paper proposes an

operating regime-based data-driven industrial framework for improved process decision-making. In this first part, the latter is established. Its added-value will be further explored in an upcoming paper (Part II).

The proposed framework exploits advantages from both domain knowledge and data-driven approaches. The former are a critical complementary information source. On the other hand, principal component analysis help to understand the process further and can extract and unlock hidden knowledge that is not present in the domain knowledge. It allows enrichment and updating of existing domain knowledge to take accurate management decisions for complex processes.

This combination of process expertise (domain knowledge) and data-driven approaches starts with the scope definition which is the objective of the analysis. Then, given that the means and variances of the data signals change with operating conditions, top-down operating regimes must be considered early on in the framework.

Once the data are collected, data pre-processing is performed and steady-state periods, intervals or windows are detected. The steady-state detection algorithm provides a reliable and effective way to identify when each key sensor reading is stationary over a suitably chosen time-window. Given that most management decisions, as well as design and diagnostics, are based on the steady-state representation, it is easier to understand and is more practical. Then, unit-wide steady-state data reconciliation is performed. The averaged variable values over the duration of each steady-state span are used for all individual data reconciliation runs. This step highlights persistently faulty sensors in the process and allows ranking them according to their probability of occurrence. Also, in opposition to common belief, if there are gross errors in the measurements, then the reconciled values are in fact less reliable than the original raw data as these outliers can unpredictably distort and bias all related or co-incident process variables involved in the model.

The last step of the proposed framework is the “bottom-up” detection and identification of operating regime. In industrial processes, there are often various component material recipes used, different products or grades made, the operating conditions may change, the feedstock may vary (slightly or not), the weather may change, different operators may operate the process differently, etc. All of these aspects may lead to different operation regimes in a plant. Some of the aforementioned changes are known, however some may be hidden in process data. Therefore, operating regime detection can yield knowledge otherwise not available, provide insights for management decision-making and thus potentially create opportunities for better process operations. Putting forward the synergy between domain experts and mathematical techniques help detecting operating regimes and thus better understand how the processes, production and plant works, and can lead to substantially improved process analytics. Incorporating operating regime detection as described in this framework is an enabler to reach the full potential of process data for decision-making.

An application of the developed framework in a brownstock washing unit was demonstrated. We showed that this practical decision support tool can be implemented in the process industries, and is suitable to be deployed in mills, plants and other types of processing facilities.

The insights generated by the framework will be demonstrated for the optimization of the brownstock washing department operations in Part II of this paper. As future work, identified operating regimes will be used for enhanced process decision-making. This will enable the analysis of cost-efficiency for various operating regimes. More specifically, to manifest further the value of the framework, Part II will address the incorporation of an activity-based costing analysis. Production costs are inevitably related to the operating regime and plants need to know where they stand to assess these costs.

As a final observation, to understand process data, and to make sense of them, they must be processed (cleaned) and their context should be acknowledged (operating regimes) by coupling mathematical methods with process expertise. Then proper and improved decisions and/or actions may be taken.

Acknowledgements

The authors gratefully acknowledge the financial support of the CRSNG and further thank CanmetENERGY for the technical support using the EXPLORE software.

References

- Bagajewicz, M. (2000). A brief review of recent developments in data reconciliation and gross error detection/estimation. *Latin American Applied Research*, 30(4), 335-342.
- Bagajewicz, M., Jiang, Q., & Sanchez, M. (2010). Performance evaluation of PCA tests in serial elimination strategies for gross error identification. *Chemical Engineering Communications*, 183(1), 119-139. <https://doi.org/10.1080/00986440008960505>
- Barton, M., & Lennox, B. (2022). Model stacking to improve prediction and variable importance robustness for soft sensor development. *Digital Chemical Engineering*, 3, 100034. <https://doi.org/https://doi.org/10.1016/j.dche.2022.100034>
- Beaver, S., Palazoglu, A., & Romagnoli, J. A. (2007). Cluster Analysis for Autocorrelated and Cyclic Chemical Process Data. *Industrial & Engineering Chemistry Research*, 46(11), 3610-3622. <https://doi.org/10.1021/ie060544v>
- Bellec, S., Jiang, T., Kerr, B., Diamond, M., & Stuart, P. (2007). On-line processing and steady-state reconciliation of pulp and paper mill process data. *108*(6), 36-40.
- Brucel, C., Thibault, É., & Stuart, P. (2023). Real-time data triage and validation of a steady-state process simulation of brownstock washing. *TAPPI JOURNAL*.
- Canada, G. o. (2022). *EXPLORE*. <https://www.rncan.gc.ca/cartes-outils-et-publications/outils/outils-modelisation/explore/24825>
- Delou, P. A., Ribeiro, L. D., Paiva, C. R., Niederberger, J., Gomes, M. V. C., & Secchi, A. R. (2021). A Real-Time Optimization Strategy for Small-Scale Facilities and Implementation in a Gas Processing Unit. 9(7), 1179. <https://www.mdpi.com/2227-9717/9/7/1179>
- Desbiens, A., Nunez, E., Del Villar, R., Hodouin, D., & Poulin, E. (2008). Using process control to increase the energy efficiency of mineral and metal processing plants. *International Journal of Power and Energy Systems*, 28(2), 146-151.
- Dunn, K. (2022a). Box plots. Dans *Process improvement using data*. <https://learnche.org/pid/data-visualization/box-plots>
- Dunn, K. (2022b). Principal Component Analysis (PCA). Dans *Process improvement using data*.
- Everitt, B. (1993). *Cluster analysis*. E. Arnold ; Halsted Press.
- Gitzel, R. (2016). *Data Quality in Time Series Data: An Experience Report*. CBI.
- Harrison, R. P., & Stuart, P. R. (2011). *Data pre-processing techniques for multivariate analysis to treat industrial operating data for retrofit design*.
- Heikkinen, M., Poutiainen, H., Liukkonen, M., Heikkinen, T., & Hiltunen, Y. (2011). Subtraction analysis based on self-organizing maps for an industrial wastewater treatment process. *Mathematics and Computers in Simulation*, 82(3), 450-459. <https://doi.org/https://doi.org/10.1016/j.matcom.2010.10.021>
- Huang, G. (2021). Missing data filling method based on linear interpolation and lightgbm. *Journal of Physics: Conference Series*, 1754, 012187. <https://doi.org/10.1088/1742-6596/1754/1/012187>
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666. <https://doi.org/https://doi.org/10.1016/j.patrec.2009.09.011>
- Jiang, T., Chen, B., & He, X. (2000). Industrial application of Wavelet Transform to the on-line prediction of side draw qualities of crude unit. *Computers & Chemical Engineering*, 24(2-7), 507-512.
- Jiang, T., Chen, B., He, X., & Stuart, P. (2003). Application of steady-state detection method based on wavelet transform. *Computers & Chemical Engineering*, 27(4), 569-578. [https://doi.org/10.1016/s0098-1354\(02\)00235-1](https://doi.org/10.1016/s0098-1354(02)00235-1)
- Jiang, T., Chen, B., Jasim, K., & Stuart, P. R. (2003). *Strategy for improving data quality for a kraft pulp mill recausticizing plant* Proceedings of the Computer-Aided Process Operations (FOCAPO) Conference, Coral Springs, Florida, USA.

- Joe Qin, S., Guo, S., Li, Z., Chiang, L. H., Castillo, I., Braun, B., & Wang, Z. (2021). Integration of process knowledge and statistical learning for the Dow data challenge problem. *Computers & Chemical Engineering*, 153, 107451. <https://doi.org/10.1016/j.compchemeng.2021.107451>
- Kelly, J., & Hedengren, J. (2013). A steady-state detection (SSD) algorithm to detect non-stationary drifts in processes. *Journal of Process Control*, 23. <https://doi.org/10.1016/j.jprocont.2012.12.001>
- Kelly, J. D. (2023). *Industrial Modeling & Programming Language (IMPL) Manual*.
- Kelly, J. D., & Zyngier, D. (2008). A new and improved MILP formulation to optimize observability, redundancy and precision for sensor network problems. 54(5), 1282-1291. <https://doi.org/10.1002/aic.11475>
- Korbel, M. (2011). *Ph. D. Thesis : On-line steady-state data reconciliation for advanced cost analysis in the pulp and paper industry*, Polytechnique].
- Korbel, M., Bellec, S., Jiang, T., & Stuart, P. (2014). Steady state identification for on-line data reconciliation based on wavelet transform and filtering. *Computers & Chemical Engineering*, 63, 206-218. <https://doi.org/10.1016/j.compchemeng.2014.02.003>
- Krishnamurthi, R., Kumar, A., Gopinathan, D., Nayyar, A., & Qureshi, B. (2020). An Overview of IoT Sensor Data Processing, Fusion, and Analysis Techniques. *Sensors*, 20, 6076. <https://doi.org/10.3390/s20216076>
- Liu, H., Shah, S., & Jiang, W. (2004). On-line outlier detection and data cleaning. *Computers & Chemical Engineering*, 28(9), 1635-1647. <https://doi.org/10.1016/j.compchemeng.2004.01.009>
- Liukkonen, M., Heikkinen, M., Hiltunen, T., Hälikkä, E., Kuivalainen, R., & Hiltunen, Y. (2011). Artificial neural networks for analysis of process states in fluidized bed combustion. *Energy*, 36(1), 339-347. <https://doi.org/10.1016/j.energy.2010.10.033>
- Liukkonen, M., Hiltunen, T., Hälikkä, E., & Hiltunen, Y. (2011). Modeling of the fluidized bed combustion process and NOx emissions using self-organizing maps: An application to the diagnosis of process states. *Environmental Modelling & Software*, 26(5), 605-614. <https://doi.org/10.1016/j.envsoft.2010.12.002>
- Moreno, R. d. P. (2010). *Steady State Detection, Data Reconciliation, and Gross Error Detection: Development for Industrial Processes*, University of New Brunswick].
- Narasimhan, S., & Jordache, C. (2000). *Data Reconciliation and Gross Error Detection: An Intelligent Use of Process Data*. Gulf Professional Publishing.
- Rhinehart, R. R. (2013). *Automated steady and transient state identification in noisy processes*.
- Saari, J., & Odelius, J. (2018). Detecting operation regimes using unsupervised clustering with infected group labelling to improve machine diagnostics and prognostics. *Operations Research Perspectives*, 5, 232-244. <https://doi.org/10.1016/j.orp.2018.08.002>
- Sigmafine. (2022). <https://www.sigmafine.com/>
- Srinivasan, R., Wang, C., Ho, W. K., & Lim, K. W. (2004). Dynamic Principal Component Analysis Based Methodology for Clustering Process States in Agile Chemical Plants. *Industrial & Engineering Chemistry Research*, 43(9), 2123-2139. <https://doi.org/10.1021/ie034051r>
- Srinivasan, R., Wang, C., Ho, W. K., & Lim, K. W. (2005). Context-based recognition of process states using neural networks. *Chemical Engineering Science*, 60(4), 935-949. <https://doi.org/10.1016/j.ces.2004.09.061>
- Thibault, É., Lebreux Désilets, F., Poulin, B., Chioua, M., & Stuart, P. (2023). Comparison of signal processing methods considering their optimal parameters using synthetic signals in a heat exchanger network simulation. *Computers & Chemical Engineering*.
- Thibault, É., Lebreux Désilets, F., Poulin, B., Chioua, M., & Stuart, P. (2023). Signal processing heuristics for the selection of wavelet transform parameters considering signal characteristics derived from a case study on a chemical pulp mill. *Computers & Chemical Engineering*.
- Tong, H., & Crowe, C. M. (1996). Detecting persistent gross errors by sequential analysis of principal components. *Computers & Chemical Engineering*, 20, S733-S738. [https://doi.org/10.1016/0098-1354\(96\)00000-2](https://doi.org/10.1016/0098-1354(96)00000-2)

Zhan, H.-r., Miao, Y., & Wang, W. (2012). Extended support vector regression based data reconciliation and its application to plant-wide mass balance. *Internation Journal of Innovative Computing, Information and Control*, 8(6), 4111-4122.

Appendix

I M P L (c)

Copyright and Property of i n d u s t r i @ l g o r i t h m s

NMONIKERS = 1000

! Dilution factor parameters / unmeasured variables.

REPEAT(1;6;1;df\$ = X10000 + \$)
 REPEAT(1;107;1;f\$ = X1000 + \$)
 REPEAT(1;107;1;fa\$ = X2000 + \$)
 REPEAT(1;107;1;x\$ = X3000 + \$)
 REPEAT(1;107;1;xa\$ = X4000 + \$)
 REPEAT(1;107;1;d\$ = X5000 + \$)
 REPEAT(1;107;1;D\$ = F5000 + \$)

MA1 = F1101
 MB1 = F1102
 MC1 = F1103
 MD1 = F1104
 ME1 = F1105
 MF1 = F1106
 MG1 = F1107
 MH1 = F1108
 MI1 = F1109
 MJ1 = F1110
 MK1 = F1111
 ML1 = F1112
 MM1 = F1113
 MN1 = F1114
 MO1 = F1115
 MX1 = F1116
 MY1 = F1117
 MZ1 = F1118

MA2 = F1201
 MB2 = F1202
 MC2 = F1203
 MD2 = F1204
 ME2 = F1205
 MF2 = F1206
 MG2 = F1207
 MH2 = F1208
 MI2 = F1209
 MJ2 = F1210
 MK2 = F1211
 ML2 = F1212
 MM2 = F1213

MN2 = F1214
MO2 = F1215
MX2 = F1216
MY2 = F1217
MZ2 = F1218

MA3 = F1301
MB3 = F1302
MC3 = F1303
MD3 = F1304
ME3 = F1305
MF3 = F1306
MG3 = F1307
MH3 = F1308
MI3 = F1309
MJ3 = F1310

SP1 = F2101
SQ1 = F2102
SR1 = F2103
SS1 = F2104
ST1 = F2105
SU1 = F2106
SV1 = F2107
SW1 = F2108
SZ1 = F2109

SP201 = F22101
SP202 = F22102
SP203 = F22103

SQ201 = F22201
SQ202 = F22202
SQ203 = F22203

SR201 = F22301
SR202 = F22302

SS201 = F22401
SS202 = F22402

ST201 = F22501
ST202 = F22502
ST203 = F22503
ST204 = F22504
ST205 = F22505
ST206 = F22506
ST207 = F22507
ST208 = F22508

ST209 = F22509
ST210 = F22510
ST211 = F22511
ST212 = F22512
ST301 = F23501
ST302 = F23502
ST303 = F23503
ST304 = F23504
ST305 = F23505
ST306 = F23506

SU201 = F22601
SU202 = F22602

SV201 = F22701
SV202 = F22702

SW201 = F22801
SW202 = F22802

SZ201 = F22901
SZ202 = F22902
SZ203 = F22903

P011 = F3101
P021 = F3102
P031 = F3103
P041 = F3104
P051 = F3105
P061 = F3106
P071 = F3107
P081 = F3108
P091 = F3109
P101 = F3110
P111 = F3111
P121 = F3112
P131 = F3113
P141 = F3114
P151 = F3115
P161 = F3116
P171 = F3117
P191 = F3119
P201 = F3120
P221 = F3122
P231 = F3123
P251 = F3125
P271 = F3127
P281 = F3128
P291 = F3129

P301 = F3130
 P012 = F3201
 P022 = F3202
 P032 = F3203
 P042 = F3204
 P052 = F3205
 P062 = F3206
 P072 = F3207
 P082 = F3208
 P092 = F3209
 P102 = F3210
 P112 = F3211
 P122 = F3212
 P132 = F3213
 P142 = F3214
 P152 = F3215
 P162 = F3216
 P172 = F3217
 P192 = F3219
 P202 = F3220
 P222 = F3222
 P232 = F3223
 P252 = F3225
 P272 = F3227
 P282 = F3228
 P292 = F3229
 P302 = F3230
 P023 = F3302
 P043 = F3304
 P063 = F3306
 P083 = F3308
 P103 = F3310
 P153 = F3315
 P163 = F3316
 P203 = F3320
 P223 = F3322
 P233 = F3323
 P253 = F3325
 P273 = F3327
 P274 = F3427
 P293 = F3329
 P294 = F3429

Objective

REPEAT(1;107;1;fw[\$] fa\$^)
 REPEAT(1;107;1;xw[\$] xa\$^)

Constraints

! Meter balances or equations i.e., reconciled + revision = raw.

```
REPEAT(1;107;1;F40000 + $| f$ + fa$ = fm[$])
REPEAT(1;107;1;F50000 + $| x$ + xa$ = xm[$])
```

! Mixer equations.

```
MA1| f001 + f004 - f002 = 0
MB1| f002 + f005 - f003 = 0
MC1| f007 + f010 - f094 = 0
MD1| f016 + f017 - f018 = 0
ME1| f021 + f022 - f023 = 0
MF1| f030 + f034 - f031 = 0
MG1| f031 + f035 - f032 = 0
MH1| f032 + f036 - f033 = 0
MI1| f008 + f056 - f057 = 0
MJ1| f060 + f063 - f061 = 0
MK1| f084 + f078 - f073 = 0
ML1| f075 + f079 - f081 = 0
MM1| f074 + f076 - f075 = 0
MN1| f048 + f049 - f050 = 0
MO1| f044 + f047 - f049 = 0
MX1| f038 + f098 - f099 = 0
MY1| f077 + f099 - f096 = 0
MZ1| f101 + f102 - f103 = 0
```

! Splitter equations.

```
SP1| f095 - f007 - f008 = 0
SQ1| f087 - f009 - f010 - f011 = 0
SR1| f088 - f014 - f016 = 0
SS1| f091 - f020 - f021 = 0
ST1| f093 - f027 - f028 - f034 - f035 - f039 - f066 - f067 - f078 - f079 - f097 - f098 - f100 = 0
SU1| f089 - f055 - f056 = 0
SV1| f092 - f004 - f059 = 0
SW1| f090 - f062 - f063 = 0
SZ1| f103 - f104 - f105 = 0
SP201| x095 - x007 = 0
SP202| x095 - x008 = 0
SQ201| x087 - x009 = 0
SQ202| x087 - x010 = 0
SQ203| x087 - x011 = 0
SR201| x088 - x014 = 0
SR202| x088 - x016 = 0
SS201| x091 - x020 = 0
SS202| x091 - x021 = 0
ST201| x093 - x027 = 0
ST202| x093 - x028 = 0
ST203| x093 - x034 = 0
ST204| x093 - x035 = 0
ST205| x093 - x039 = 0
ST206| x093 - x066 = 0
```

ST207| $x_{093} - x_{067} = 0$
 ST208| $x_{093} - x_{078} = 0$
 ST209| $x_{093} - x_{079} = 0$
 ST210| $x_{093} - x_{097} = 0$
 ST211| $x_{093} - x_{098} = 0$
 ST212| $x_{093} - x_{100} = 0$
 SU201| $x_{089} - x_{055} = 0$
 SU202| $x_{089} - x_{056} = 0$
 SV201| $x_{092} - x_{004} = 0$
 SV202| $x_{092} - x_{059} = 0$
 SW201| $x_{090} - x_{062} = 0$
 SW202| $x_{090} - x_{063} = 0$
 SZ201| $x_{103} - x_{104} = 0$
 SZ202| $x_{103} - x_{105} = 0$

! Process equations.

P011| $f_{003} + f_{009} - f_{006} - f_{095} = 0$
 P021| $f_{015} + f_{014} + f_{094} - f_{017} - f_{012} = 0$
 P031| $f_{012} + f_{104} - f_{013} - f_{087} = 0$
 P041| $f_{018} + f_{020} - f_{019} - f_{022} = 0$
 P051| $f_{019} - f_{088} - f_{106} = 0$
 P061| $f_{023} + f_{026} + f_{027} - f_{024} - f_{025} = 0$
 P071| $f_{024} - f_{091} - f_{101} = 0$
 P081| $f_{057} + f_{062} - f_{058} - f_{060} = 0$
 P091| $f_{058} + f_{013} + f_{105} - f_{089} - f_{092} = 0$
 P101| $f_{061} + f_{066} + f_{086} - f_{064} - f_{065} = 0$
 P111| $f_{064} - f_{090} - f_{102} = 0$
 P121| $f_{025} + f_{028} - f_{029} = 0$
 P131| $f_{065} + f_{067} - f_{068} = 0$
 P141| $f_{029} + f_{068} + f_{070} - f_{030} = 0$
 P151| $f_{033} + f_{039} - f_{037} - f_{038} = 0$
 P161| $f_{037} + f_{041} + f_{042} + f_{085} - f_{040} - f_{045} = 0$
 P171| $f_{040} + f_{043} - f_{044} = 0$
 P191| $f_{006} + f_{054} + f_{055} - f_{051} = 0$
 P201| $f_{051} - f_{052} - f_{005} = 0$
 P221| $f_{096} + f_{100} - f_{036} - f_{084} = 0$
 P231| $f_{073} + f_{097} - f_{072} - f_{074} = 0$
 P251| $f_{072} - f_{054} - f_{070} = 0$
 P271| $f_{081} - f_{077} - f_{080} = 0$
 P281| $f_{080} - f_{083} = 0$
 P291| $f_{083} - f_{082} - f_{076} = 0$
 P301| $f_{045} - f_{093} - f_{107} = 0$

! Mixer equations.

MA2# $d_{001} * x_{001} * f_{001} + d_{004} * x_{004} * f_{004} - d_{002} * x_{002} * f_{002} = 0$
 MB2# $d_{002} * x_{002} * f_{002} + d_{005} * x_{005} * f_{005} - d_{003} * x_{003} * f_{003} = 0$
 MC2# $d_{007} * x_{007} * f_{007} + d_{010} * x_{010} * f_{010} - d_{094} * x_{094} * f_{094} = 0$
 MD2# $d_{016} * x_{016} * f_{016} + d_{017} * x_{017} * f_{017} - d_{018} * x_{018} * f_{018} = 0$

ME2# $d021 \times d021 \times f021 + d022 \times d022 \times f022 - d023 \times d023 \times f023 = 0$
 MF2# $d030 \times d030 \times f030 + d034 \times d034 \times f034 - d031 \times d031 \times f031 = 0$
 MG2# $d031 \times d031 \times f031 + d035 \times d035 \times f035 - d032 \times d032 \times f032 = 0$
 MH2# $d032 \times d032 \times f032 + d036 \times d036 \times f036 - d033 \times d033 \times f033 = 0$
 MI2# $d008 \times d008 \times f008 + d056 \times d056 \times f056 - d057 \times d057 \times f057 = 0$
 MJ2# $d060 \times d060 \times f060 + d063 \times d063 \times f063 - d061 \times d061 \times f061 = 0$
 MK2# $d084 \times d084 \times f084 + d078 \times d078 \times f078 - d073 \times d073 \times f073 = 0$
 ML2# $d075 \times d075 \times f075 + d079 \times d079 \times f079 - d081 \times d081 \times f081 = 0$
 MM2# $d074 \times d074 \times f074 + d076 \times d076 \times f076 - d075 \times d075 \times f075 = 0$
 MN2# $d048 \times d048 \times f048 + d049 \times d049 \times f049 - d050 \times d050 \times f050 = 0$
 MO2# $d044 \times d044 \times f044 + d047 \times d047 \times f047 - d049 \times d049 \times f049 = 0$
 MX2# $d038 \times d038 \times f038 + d098 \times d098 \times f098 - d099 \times d099 \times f099 = 0$
 MY2# $d077 \times d077 \times f077 + d099 \times d099 \times f099 - d096 \times d096 \times f096 = 0$
 MZ2# $d101 \times d101 \times f101 + d102 \times d102 \times f102 - d103 \times d103 \times f103 = 0$

! Process equations.

P012# $d003 \times d003 \times f003 + d009 \times d009 \times f009 - d006 \times d006 \times f006 - d095 \times d095 \times f095 = 0$
 P022# $d015 \times d015 \times f015 + d014 \times d014 \times f014 + d094 \times d094 \times f094 - d017 \times d017 \times f017 - d012 \times d012 \times f012 = 0$
 P023# $(1 - x014) \times f014 \times d014 - f017 \times d017 + (1 - df1) \times (f017 \times d017 \times x017) = 0$
 P032# $d012 \times d012 \times f012 + d104 \times d104 \times f104 - d013 \times d013 \times f013 - d087 \times d087 \times f087 = 0$
 P042# $d018 \times d018 \times f018 + d020 \times d020 \times f020 - d019 \times d019 \times f019 - d022 \times d022 \times f022 = 0$
 P043# $(1 - x020) \times f020 \times d020 - f022 \times d022 + (1 - df2) \times (f022 \times d022 \times x022) = 0$
 P052# $d019 \times d019 \times f019 - d088 \times d088 \times f088 - d106 \times d106 \times f106 = 0$

P062# $d023 \times d023 \times f023 + d026 \times d026 \times f026 + d027 \times d027 \times f027 - d024 \times d024 \times f024 - d025 \times d025 \times f025 = 0$
 P063# $(1 - x027) \times f027 \times d027 + f026 \times d026 - f025 \times d025 + (1 - df3) \times (f025 \times d025 \times x025) = 0$
 P072# $d024 \times d024 \times f024 - d091 \times d091 \times f091 - d101 \times d101 \times f101 = 0$
 P082# $d057 \times d057 \times f057 + d062 \times d062 \times f062 - d058 \times d058 \times f058 - d060 \times d060 \times f060 = 0$
 P083# $(1 - x062) \times f062 \times d062 - f060 \times d060 + (1 - df4) \times (f060 \times d060 \times x060) = 0$
 P092# $d058 \times d058 \times f058 + d013 \times d013 \times f013 + d105 \times d105 \times f105 - d089 \times d089 \times f089 - d092 \times d092 \times f092 = 0$
 P102# $d061 \times d061 \times f061 + d066 \times d066 \times f066 + d086 \times d086 \times f086 - d064 \times d064 \times f064 - d065 \times d065 \times f065 = 0$
 P103# $(1 - x066) \times f066 \times d066 + f086 \times d086 - f065 \times d065 + (1 - df5) \times (f065 \times d065 \times x065) = 0$
 P112# $d064 \times d064 \times f064 - d090 \times d090 \times f090 - d102 \times d102 \times f102 = 0$
 P122# $d025 \times d025 \times f025 + d028 \times d028 \times f028 - d029 \times d029 \times f029 = 0$
 P132# $d065 \times d065 \times f065 + d067 \times d067 \times f067 - d068 \times d068 \times f068 = 0$
 P142# $d029 \times d029 \times f029 + d068 \times d068 \times f068 + d070 \times d070 \times f070 - d030 \times d030 \times f030 = 0$
 P152# $d033 \times d033 \times f033 + d039 \times d039 \times f039 - d037 \times d037 \times f037 - d038 \times d038 \times f038 = 0$
 P162# $d037 \times d037 \times f037 + d041 \times d041 \times f041 + d042 \times d042 \times f042 + d085 \times d085 \times f085 - d040 \times d040 \times f040 - d045 \times d045 \times f045 = 0$
 P163# $(1 - x042) \times f042 \times d042 + f041 \times d041 + f085 \times d085 - f040 \times d040 + (1 - df6) \times (f040 \times d040 \times x040) = 0$
 P172# $d040 \times d040 \times f040 + d043 \times d043 \times f043 - d044 \times d044 \times f044 = 0$
 P192# $d006 \times d006 \times f006 + d054 \times d054 \times f054 + d055 \times d055 \times f055 - d051 \times d051 \times f051 = 0$
 P202# $d051 \times d051 \times f051 - d052 \times d052 \times f052 - d005 \times d005 \times f005 = 0$
 P203# $(1 - x005) \times d005 \times f005 - 0.85 \times (1 - x051) \times d051 \times f051 = 0$
 P222# $d096 \times d096 \times f096 + d100 \times d100 \times f100 - d036 \times d036 \times f036 - d084 \times d084 \times f084 = 0$
 P232# $d073 \times d073 \times f073 + d097 \times d097 \times f097 - d072 \times d072 \times f072 - d074 \times d074 \times f074 = 0$
 P252# $d072 \times d072 \times f072 - d054 \times d054 \times f054 - d070 \times d070 \times f070 = 0$
 P272# $d081 \times d081 \times f081 - d077 \times d077 \times f077 - d080 \times d080 \times f080 = 0$
 P273# $(x080 \times f080 \times d080) - b \times (x081 \times f081 \times d081) = 0$

P282# $d080 * x080 * f080 - d083 * x083 * f083 = 0$
 P292# $d083 * x083 * f083 - d082 * x082 * f082 - d076 * x076 * f076 = 0$
 P293# $(x082 * f082 * d082) - c * (x083 * f083 * d083) = 0$
 P302# $d045 * x045 * f045 - d093 * x093 * f093 - d107 * x107 * f107 = 0$

! Density equation as function of consistency.

REPEAT(1;107;1;D\$# $0.9894 + 0.2327 * x\$^{\{2\}} + 0.3061 * x\$ - d\$ = 0$)

Bounds

REPEAT(1;6;1;-LARGENUMBER <= df\$ <= LARGENUMBER)
 REPEAT(1;107;1;-LARGENUMBER <= f\$ <= LARGENUMBER)
 REPEAT(1;107;1;-LARGENUMBER <= fa\$ <= LARGENUMBER)
 REPEAT(1;107;1;-LARGENUMBER <= x\$ <= LARGENUMBER)
 REPEAT(1;107;1;-LARGENUMBER <= xa\$ <= LARGENUMBER)
 REPEAT(1;107;1;-LARGENUMBER <= d\$ <= LARGENUMBER)

Initials

REPEAT(1;6;1;df\$ = DF\$)
 REPEAT(1;107;1;f\$ = fm[\$] + NOT{fm[\$]})
 REPEAT(1;107;1;x\$ = xm[\$] + NOT{xm[\$]})
 REPEAT(1;107;1;d\$ = $0.9894 + 0.2327 * \{xm[\$]\}^{\{2\}} + 0.3061 * \{xm[\$]\}$)

End

ANNEXE F ARTICLE #5 AN OPERATING REGIME-BASED DATA-DRIVEN INDUSTRIAL FRAMEWORK FOR IMPROVED PROCESS DECISION-MAKING – PART 2: FRAMEWORK APPLICATION CONSIDERING ACTIVITY-BASED COSTING CONCEPTS

Émilie Thibault, Christian Ledoux, Paul Stuart

Submitted to the Industrial & Engineering Chemistry Research Journal on 10/02/2023

Abstract

Operating data collected and stored in historian must be managed to extract their full potential. Part I of this paper proposed a structured way to process industrial data. This framework considers the analysis scope definition, data management steps and operating regimes detection and identification. These regimes are detected and identified by known process conditions and data driven analysis (principal component analysis and *k*-means algorithm). Furthermore, all steps proposed in the framework are based on the combination of process knowledge and data-driven approaches.

The value of this proposed framework is demonstrated in Part II through the use of cost accounting for operational problem-solving. With the use of operations-driven cost modeling and processed data corresponding to steady-state operation, incremental profit can be assigned to each operating regime in order to identify the most cost-efficient one.

The overall objective of this paper is to convert processed industrial steady-state data and cost information into knowledge that can be used to optimize the washing department of a chemical pulp mill. More specifically, the different operating regimes are assessed, and the most suitable operating strategy is defined.

The washing process performance is assessed for 12 steady-state operating conditions, and an optimization analysis explored and compared these scenarios to highlight an improved way to operate (preferred combination of set point values), and recognized which controlled variables are responsible for this high performance. Therefore, the optimization analysis is identifying which operating regimes are presenting an improved performance, i.e., when the process is showing the highest pulp throughput, minimal pulp conductivity (higher washing performance) and minimal cost.

A contribution analysis highlighted which variables are impacted (in which a difference can be observed) by the highest washing performance. This is assessed by comparing operating regimes 2 and 12 that respectively showed the lower and higher profit (and overall washing performance). It is possible to conclude that when the process is operated this way, significant savings are achieved : 1A washer level, level of the filtrate tank of the thickener, pulp conductivity kept low, and consistency of the stream leaving the mixing tank, showers water flow, showers water temperature kept high. A concise analysis assessed if the different operating conditions of the controlled variables that impact the operations matched the operators schedule.

The analysis performed as part of this paper proved the potential of the framework and exhibited why operating regimes were required to solve this specific operating problem. For proper data-driven decisions or actions, clean and processed data are required, and especially the detection of operating regimes.

Keywords Activity-based costing, optimization analysis, operating regimes, framework, pulp mill

Introduction

Every plant exploits some level of cost-control strategies in order to ensure its financial performance and to maintain a competitive position in the marketplace. Over the past decade, the strategy employed by North American pulp and paper mills has been to lay off staff, downsize operations and sometimes close down (CBC, 2009). However, in the shorter term, some mills are adopting a strategy of minimizing production costs. For this purpose, financial and cost accounting are exploited. Financial accounting is required to establish the company's financial statements, prepared according to standardized guidelines. Cost accounting, on the other hand, is used internally by management to study plant operations and make decisions to maximize the plant's profits; it provides valuable insight to decision makers. The results of this analysis are confidential and are never made public, unlike its counterpart, financial accounting (El-Halwagi & Stuart, 2013; Hytönen & Stuart, 2011). In addition to decision making, cost accounting is used to solve operation problems and maximize product margins (AccountingTools, 2018; Averkamp, 2019; Murphy, 2019). To conclude with the strategies put forward by mills in the paper industry, some of them tend, in addition to minimizing their production cost, towards diversification of their product portfolio, i.e., biorefinery.

Nevertheless, the relationship between industrial process data and cost information is underexploited in the pulp and paper industry. Interviews conducted with process experts in the sector highlighted that the production costs per product and per operating regime are often unknown and that it would be interesting to know the cost of operating in different regimes (Thibault, Émilie, Chioua, Moncef, et al., 2023).

Furthermore, a cost analysis can only be as good as the process data provided to perform the analysis. In this perspective, a data management framework was proposed in Part I of this paper (Thibault, Émilie, Kelly, Jeffrey Dean, et al., 2023). This framework considers, among the other steps, the reduction of noise. Different data processing techniques to reduce those high frequency components are compared in Thibault, Émilie et al. (2023). The wavelet transforms were found to be robust and were used to establish heuristics to process a wide range of signal in a timely efficient manner (Thibault, Émilie, Lebreux Désilets, Francis, et al., 2023). Furthermore, the framework helps explaining the variations in the process and highlight which variables were responsible for the variations by considering operating regimes. Operating regime detection helps to better understand how a mill operates and increases the ability to interpret process data thoroughly and deeply. It provides additional information on process operation and offers an applicable way of defining the best mode for achieving a more cost-efficient process. Therefore, recognizing different operating regime opens the door for many new opportunities.

Part II of this paper proposes a concrete application of the framework. It highlight an example where it is worthwhile and justifiable to use the framework to solve an operational problem. The operating regimes are considered to refine the operating cost estimates in a pulp mill.

The objective of this paper is to make a juxtaposition of the data management framework with an activity-based costing analysis to solve an optimization problem in the washing department of a pulp mill and therefore assess how the process can operate better. More precisely, the activity-based costing assessment would provide the incremental cost and value generated by a product in each operating regime, and knowing these regimes as well as their associated cost, an optimization analysis could identify the preferred conditions. Therefore, the value of the framework is demonstrated in an industrial context by solving a process optimization problem that benefits from the recognition of steady-state operating regimes and associated cost.

This paper starts with an overview of operating cost assessment in industry (section 2) as well as optimization analysis in pulp and paper sector presented in the literature (section 3). In section 4, the proposed approach to merge

the data management framework exhibited in Part 1 with a cost analysis, focusing particularly on operating regimes is described using real-world data from a pulp mill as an illustrative example. A short summary of the case study background is stated in section 5, whereas the following section exhibit the optimization problem definition. Section 5.2 explain the results obtained as part of this analysis and the paper discusses these results and the added-value sought by the framework in section 6. Lastly, a conclusion is done in section 7.

Cost accounting historical background

Cost accounting counts three families of costs : raw materials, labor and overhead (El-Halwagi & Stuart, 2013). The first two are classified as variable costs that fluctuate according to the quantity of products manufactured (null if there is no production), while the third can either be variable or fixed, i.e., that do not vary according to the level of production – they are always recurring. The overhead costs refer to the expenses associated with the proper functioning of the company; expenses required to maintain the business.

In addition to being fixed or variable, costs can be classified as direct, i.e. including everything directly related to the production of a product, or indirect, i.e. expenses that are difficult to attribute to a single product; not related to the production of a product (Murphy, 2019).

Cost accounting consists of classifying, analyzing and allocating the different types of costs to the products that the plant produces, and then taking action to control and optimize these costs. Specifically, it involves identifying which products or regimes are the most profitable and which are the least profitable for the plant. Also, this method makes it possible to study the efficiency of resource consumption, optimize procedures, identify process bottlenecks and study their impact on production and reduce operating costs (AccountingTools, 2018; Averkamp, 2019; Murphy, 2019).

The accuracy with which these three cost elements – raw materials, labor and overhead – are evaluated will impact the reliability of the decision-making process. At the beginning of the 20th century, their evaluation was relatively simple since variable costs represented the majority of manufacturing costs and the range of products manufactured was quite limited (Cooper & Kaplan, 1988; El-Halwagi & Stuart, 2013). In fact, the proportion of overhead costs was negligible compared to the costs of labor and raw materials (Korbel, Milan, 2011). However, times are changing and production that was heavily labor-based is now shifting to a largely automation-based environment (Cooper & Kaplan, 1990; El-Halwagi & Stuart, 2013). Thus, allocating correctly overhead and indirect costs to products is more complex, and misallocation of these costs creates a distortion in the evaluation of product production costs (El-Halwagi & Stuart, 2013; Korbel, Milan, 2011).

There are different methods of cost accounting. The evaluation of costs differs according to the technique used. A first one is the cost-volume-profit analysis, it is widely used in the paper industry (El-Halwagi & Stuart, 2013). The total fixed and variable costs evaluated for the total quantity of products manufactured are used to calculate the level of production at which the company begins to make a profit (AccountingTools, 2018). A second approach, which is also very common in the pulp and paper industry, is standard cost accounting. The latter uses ratios that compare the labor and raw materials currently used to produce a product to what is required under so-called standard conditions (AccountingTools, 2018). Both of these methods were developed nearly 100 years ago, at the beginning of the industrial era, during which manufacturing costs were dominated by variable elements (Korbel, Milan, 2011). In fact, these approaches consider all types of costs to vary with regard to production volume, and thus create inaccuracy in fixed costs. Furthermore, overhead costs are allocated without looking for a cause-and-effect relationship. Instead, they are allocated arbitrarily, making traditional methods highly inappropriate a multiproduct

environment (Korbel, Milan, 2011). Furthermore, traditional methods do not allow for the calculation of margins for individual products (El-Halwagi & Stuart, 2013; Korbel, Milan, 2011).

It was found that the increase in overhead costs compared to direct costs is obvious in the manufacturing sector (Tatsiopoulos & Panayiotou, 2000). Hence, the allocation of these costs must be more equitable with respect to the different products (Tatsiopoulos & Panayiotou, 2000). Therefore, the distortion of the cost of production (accounting for the overhead and other indirect costs based on simple volume-based measures including labor and machine-hours) increases further over the years as overhead and other indirect costs become more and more important (AccountingTools, 2018; Cooper & Kaplan, 1988, 1990; El-Halwagi & Stuart, 2013). These methods were (and still are) employed because of their simplicity, they are widely known and understood, and the models employing them are inexpensive to maintain. Yet, they have not evolved even though labor has long since ceased to dominate production costs. Poor cost allocation can lead to poor decision making.

A more recent method based on business functions and activities add more levels of detail to the cost structure : activity-based costing (ABC) method. This approach, as its name indicates, identifies the activities incurred by industrial plants, allocates to them the cost of the resources they consume and then attributes these activities to products through activity cost drivers. Thus, the final products consume activities which themselves consume resources. The activities consequently represent the intermediate stage, and the cost of the resources is no longer assigned directly to the products according to their production volume (AccountingTools, 2018; Cooper & Kaplan, 1988, 1991; Murphy, 2019).

ABC method emerged in the 1980s in response to overhead costs allocation discrepancies (Cooper & Kaplan, 1988). Indeed, labor and raw materials are quite easily traceable to the different products, but it is more difficult to allocate indirect costs and overheads directly to products. ABC allows overhead costs to be associated with certain products in a rational manner. It allows a clear representation of the revenues generated and the consumption of resources (Cooper & Kaplan, 1991). This method uses a very specific lexicon presented in the table below (Table 1).

Table 1. Lexicon used in the ABC approach (Cooper & Kaplan, 1990; Korbel, Milan, 2011)

Word	Definition
<i>Activity</i>	Basic operation that incurs costs; characterized as the consumer of resources.
<i>(Activity) Cost driver</i>	Measure the use of an activity by cost objects; factor that creates or drives the cost of an activity; link activities to cost objects, e.g. machine operating hours, number of products, number of employees, hours worked, etc.
<i>Cost center (cost pool)</i>	Units that represent parts of the process (departments) to which costs are allocated; identification of indirect costs assigned to a cost object. Activity cost drivers must be determined for each cost center.
<i>Cost object</i>	Something to which a cost is assigned, e.g. production line, department, product, customer, service, etc. Consumes individual activities.
<i>Process work center (PWC)</i>	Represents the different processes of the plant (Figure 1)
<i>Resource driver</i>	What links resources to activities, determines the amount of resources consumed by each activity e.g. work/labor hours, kWh, number of deliveries

Figure 1 shows the PWCs. These are divided into two categories : those related to process support and those related to production (Janssen, M. et al., 2012).

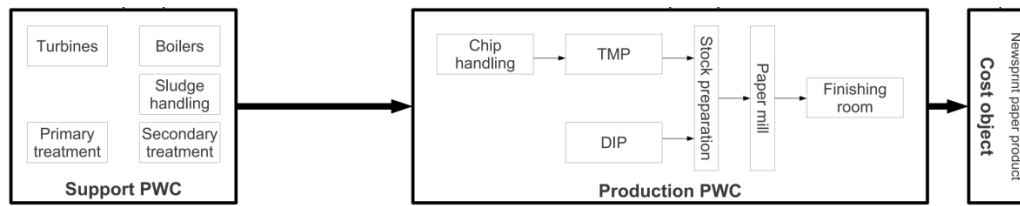


Figure 1. PWCs division into support and production, adapted from Janssen, M. et al. (2012)

Figure 2 shows a simplified representation of the steps to be followed in applying the ABC method. With this approach, resources are linked to activities by *resource drivers*, and similarly activities are linked to cost objects by *cost drivers*. It can be seen that the ABC method is implemented through two logical steps, while traditional accounting has only one. This is because the latter does not consider the causal relationship between what causes the cost of an activity and the activity itself. Instead, costs are allocated in an ad-hoc manner or on the basis of production volume (Emblemsvåg, J. & Bras, 2000).

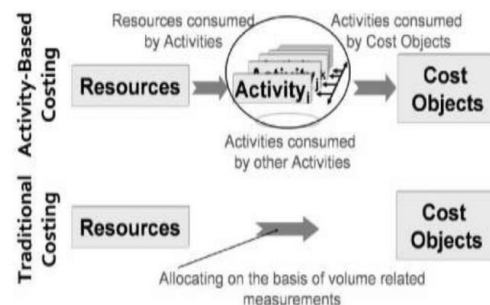


Figure 2. ABC method vs. Traditional cost analysis method (Korbel, Milan, 2011)

The process-oriented character of ABC and the causal relationships between cost drivers and activities increase the accuracy of the cost data, help management understand their actual costs and potentially improve their profits. On the other hand, traditional costing methods cannot encompass this critical linkage between actual causes (activities) and associated costs.

ABC considers different levels of activities and their expenses – resources consumed (Table 2). The consumption of resources at the *unit-level* varies according to the number of units produced, for the *batch-level* it varies according to the number of batches produced and for the *product-level*, it varies according to the number of different types of products. Only the expenses of these first three levels should be allocated to products. Thus, expenses related to the *facility-level* that support the operations of the entire plant – i.e. utilities, insurance, security, building maintenance – are never attributed to cost objects, they are kept at the plant level (Cooper & Kaplan, 1991). As shown in Table 2, the cost drivers at the *unit-level* are direct labor and raw materials and are directly assigned to the units produced. At the *batch-level*, the cost drivers affect an entire production batch, e.g. the number of machine installations, number of deliveries to customers, number of batches, etc. At the *product-level*, the production activities may be placing orders, product engineering and order scheduling (Quesada-Pineda, 2004).

Table 2. Hierarchy of operating expenses (inspired by (Cooper & Kaplan, 1991) and (Tekathen, 2019))

	Explanation	Activities	Cost Drivers
Unit-level	Performed for each unit produced	Labor-related; Materials Machine-related; Energy	Direct labor hours; Machine hours
Batch-level	Performed for each batch produced	Equipment set-up; Purchase ordering; Material handling; Inspection	Nb. of set-ups or set-up time; Nb. of purchase orders; Nb. of materials moves; Nb. of inspections or inspection time;
Product-level	Performed for support of an entire product line	Production design; Product specifications; Engineering changes;	Number of product designs; Number of changes (such as product enhancement)
Facility-level	Required to support an entire production process	Plant management salaries; Plant depreciation; Property (building and grounds) taxes; Utilities (heating and lighting)	Number of employees managed; Square meters

ABC allows for long-term decision making with respect to operations and design (Greenwood & Reeve, 1992). ABC makes it possible to differentiate the expenses incurred for the production of individual units from those for a set of products, regardless of the number of units sold. It only depends on the level of activity and thus on the consumption of resources. Traditional cost analysis methods, on the other hand, consider that all activities are carried out only at the unit-level.

Furthermore, traditional accounting systems tend to aggregate costs over a manufacturing period while considering solely a single standard recipe. However, the cost information might incorporate various changes in process operation (due to disturbances, raw material variation, operating regimes transition, etc.). Hence, recognizing that within the manufacturing period, the cost will most likely differ from one product to another as well as within the production process for the same product is critical and can be done through activity-based costing.

There are several industrial applications and/or development of process-driven accounting frameworks based on the activity-based philosophy. Some are discussed in the following section.

Operating cost assessment in pulp and paper industry

ABC method is process and operation oriented : it establishes what unites *cost objects* and *activities* and is suitable for analyzing the operating costs of a continuous production. There are many applications of this process-oriented cost model. First, an approach that integrates environmental metrics to ABC principles for analytical economic and environmental assessment for decision-making activities has been developed; environmentally related opportunities were ranked with the aim of improving competitiveness while reducing environmental impact (Emblemsvåg, J. & Bras, 2000). Second, an oil refinery used ABC to evaluate the costs of by-products as well as the main product. Indeed, this method improves the evaluation of the costs of each product when several products are manufactured simultaneously (Wang, C. et al., 2008). In addition, the ABC approach has been used to evaluate the performance of investments in manufacturing companies (Sawhney, 1991). ABC has also been used to perform a comparative life-cycle cost analysis of different plant budget scenarios to identify different key factors (Emblemsvåg, Jan, 2003; Rivero & Emblemsvåg, 2007). Lastly, a template was proposed to assist decision-making surrounding operations to enable continuous improvement using the ABC method (Greenwood & Reeve, 1992).

Now focusing on the pulp and paper sector, the activity-based costing method was employed for cost modelling and analysis in the case of a virtual paper mill (Korpunen & Paltakari, 2013). Following a very extensive analysis considering both general and process specific cost drivers (chosen carefully), the results shown that the ABC method is a pragmatic cost calculation approach, and with the parameters used, it was estimated that papermaking can still be profitable in Nordic conditions. Given that ABC delivers accurate information for strategic and tactical decisions (Korpunen & Paltakari, 2013), processes that generate higher costs than added-value were identified by mill managers. Korpunen and Paltakari mentioned that the traditional costing methods do not necessarily provide process-based information, and that some marginal products or raw materials go sometime unnoticed, and that may therefore lead to false determination of prices (Korpunen & Paltakari, 2013). The authors showed that a slight change in the market prices may be critical for profitability and that a few errors in cost estimations may eventually lead any business to a detrimental result.

Other similar studies – activity-based costing for the pulp and paper industry sector – were performed (Hsieh et al., 2020; Korpunen et al., 2012; Nurminen et al., 2009). For instance, Janssen presented a bottom-up approach to business modeling by exploiting the potential of real-time process data, ABC cost data, and environmental data for process decision-making (Janssen, Matty & Stuart, 2004). The business model concept was illustrated through an application in an integrated newsprint mill.

Also, a sophisticated approach proposes to integrate cost and process information to provide an operations-driven costing approach (ODCA) (Laflamme-Mayer et al., 2011; Laflamme-Mayer et al., 2008). This allows a thorough understanding and interpretation of resource consumption. This method links the costs to the process operations using ABC. The latter was developed by accountants as well as process engineers in the pulp and paper industry as it requires in-depth understanding of the process operations (Korbel, Milan, 2011). ODCA seeks to minimize production costs. It identifies which operating regimes are unprofitable, and hence should be avoided, as well as the margin for each product; it can be used as an operational decision support tool. The knowledge of individual product margins allows to identify which products should be emphasized more (El-Halwagi & Stuart, 2013). Finally, ODCA has been applied to assess production costs for different production campaign runs; the results show that the manufacturing cost of the same product varies significantly from one campaign (regime) to another (Laflamme-Mayer et al., 2011). It has also been used to analyze and optimize the supply chain of a forest industry mill (Laflamme-Mayer et al., 2008).

Each product from the pulp and paper mills is manufactured according to a specific recipe. However, each of these recipes can be produced according to different operating strategies, i.e. operating regimes. These are characterized by operating conditions and by choices in the process operation. Different set points for a parameter, the use of one or two production lines, the opening and closing of a recirculation loop are examples of operating regimes. With this in mind, Korbel et al. proposed to use a ODCA to evaluate the production cost of different products under different operating regimes (Figure 3) (Korbel, M. et al., 2007).

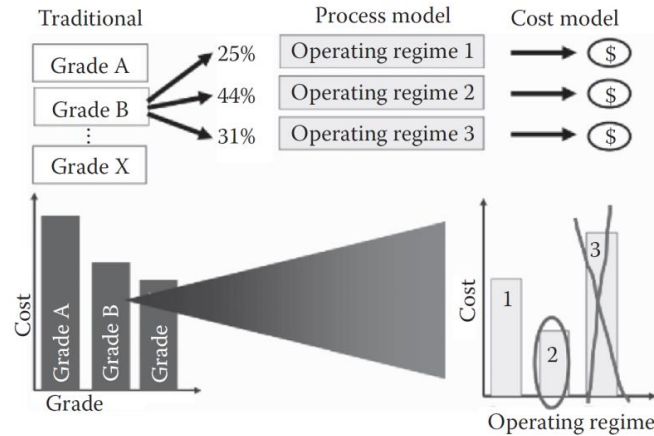


Figure 3. Presentation of the methodology presented by Korbel, M. et al. (2007)

The more traditional cost analysis methods analyze the different products (grades) margins in an *ad-hoc* manner (El-Halwagi & Stuart, 2013). However, ODCA coupled with the operation regimes is based on a process engineering perspective and thus provides a better understanding of costs and the product margins for each campaign. Indeed, this technique relies on the use of a process model that presents the operations in pseudo steady-state regime. Together, the cost and process models allow to evaluate the probability of each operation regime occurring and their individual and respective profitability (El-Halwagi & Stuart, 2013). Thus, this information is used to make decisions at the plant operations level. This approach operates in a *bottom-up* manner, i.e., from the identification of the different operation regimes to the process model, to the collection of cost data, and then to the calculation of operation costs. Figure 4 shows the overall information exchange that takes place with ODCA.

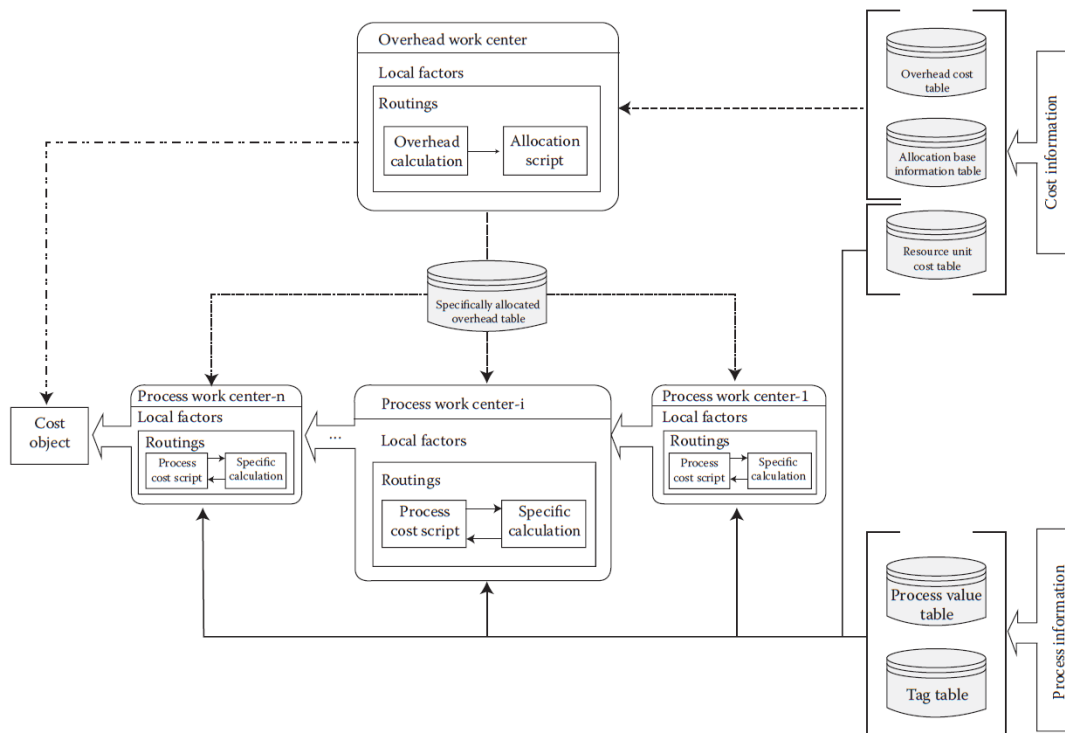


Figure 4. Information exchange between the process model and economic data (Laflamme-Mayer et al., 2011)

In addition, ODCA allows the economic evaluation of continuous processes generating a range of products. It offers great insight for decision-making, permits the identification of the optimal regimes by allowing the visualization of how costs vary with each regime, and gives the production cost and margin of different products for each operating regime. While it helps to identify operating regimes that are not cost effective, some of these are inherent to the process and are unavoidable. Figure 5 shows the contrast between traditional costing methods and operation-driven costing approach. The latter starts with the identification of the operating regimes.

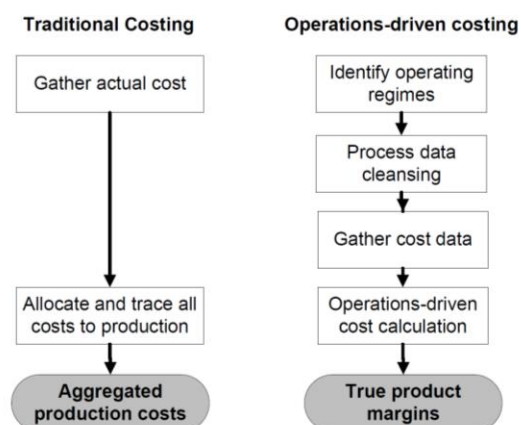


Figure 5. Comparison of cost analysis methodologies (Korbel, M. & Stuart, 2011)

Therefore, operations-driven costing approaches combine the cost and the operating regimes, allowing to assess the cost of each regime. A cost analysis is going to be applied in combination with the data processing framework proposed by Thibault, Émilie, Kelly, Jeffrey Dean, et al. (2023) to solve an optimization problem in the pulp and paper sector. Therefore, the following section discusses cases of optimization in this industry in more detail, with a particular focus on the consideration of operating regime in the identification of the solution.

Optimization analysis in pulp and paper sector

There are few paper addressing optimization analysis related to the pulp and paper industry such as one on energy optimization (Marinova et al., 2009). Another aim to improve washing efficiency and reduce water consumption by determining the optimum wash water usage level in a brown stock washing operation – level at which washing losses are satisfactory (Kopra et al., 2010). Additionally, a study that aim to improve the efficiency of a chemical pulp mill's brownstock washing (BSW) line was carry out (Kopra et al., 2015). The results indicate that the amount of brownstock washing loses may be reduced by controlling the dilution factor. Furthermore, another paper intent to review the cost parameters in brownstock washing and explore their sensitivities as well as the equipment performance, all that to present a cost optimization (Cantrell, 2019). Lastly, Chew et al. presents an optimization model aiming for simultaneous water and energy saving for BSW system in a pulp and paper mill as well as a sensitivity analysis to explore the relationship between washing efficiency and utilities, i.e., water and energy, consumption (Chew et al., 2013). All of these studies had no consideration for operating regimes.

Nevertheless, a paper tackled the optimization of washer performance by calculating the displacement ratio (Kopra et al., 2011). In this case, the results indicate that optimal operation of the pressure filter in the brown stock washing was sought by changing the processing experimental conditions and control parameters. Therefore, the type of wood pulp varies (pine or birch), as well as the production rate, the feed pulp consistency, the dilution factor, and the purity of the wash filtrate (kappa number). Hence, there is this idea of considering different recipes or operating conditions, and identifying the optimal one in the context of the operation of pressure filters.

In another paper, the concept of operating regimes was put forward through the consideration of varying seasonal conditions (heat load variations). In a chemical pulp mill, a multi-period optimization model for design decisions that considers the operational limits of the steam production units as well as the heat load variations over the year was proposed (Svensson & Berntsson, 2013). It was mentioned that a conventional approach that considers fixed annual averages (single-period) of process parameters would risk leading to sub-optimal solutions in the context of reducing the process heating demand because of neglecting the variations in heat demand and the operational limits (Svensson & Berntsson, 2013). Therefore, the approach suggested considered modelling the varying heat demand. As a result of the comparison of the aforementioned approach and the conventional approach, the latter shows significant errors and arrives at an inadequate estimation of the optimal solution, which provides poor guidance regarding financial decision-making. The results of the optimization analysis can only be correctly determined by properly including heat load variations data. The exact same idea was also addressed in (Svensson, 2014). These studies considered only the seasonal variations in steam demand. However, the different production rate, product campaigns and ambient temperatures are all additional factors for process steam demand variations. Furthermore, electricity and raw material (biomass) prices also vary over time, and may significantly impact the optimal operation of the system.

Therefore, a model for design optimization of a kraft pulp mill steam utility systems, subject to variations in energy (electricity) prices and operating conditions including steam production and demands, is presented in (Svensson et al., 2020). The latter is used to study how these variations affect investment decisions. The model included more sources of variations and tried to overcome the challenges that comes with the modelling tasks such as those related to data reliability, uncertainties in cost data, lack of documentation and technical specifications, the poor correspondence between literature models and measurement data as well as the wide range of sources of variations for which correlations are often difficult to identify (Svensson et al., 2020). The optimal investment obtained when accounting for the varying conditions was compared with the one found when only a single operating and price scenario is considered, and it was shown that the former may identify solutions with more flexibility than those identified with a single-period model.

As has been demonstrated, there are no optimization analysis that considers operating regimes in all of their complexity to highlight the preferred way to operate in in terms of cost, throughput volume and pulp quality in the washing department. The following section details the methodology employed to do so.

Using activity-based costing philosophy or operation-driven cost approach provides an estimation of how much resources are consumed in each operating regime; it allows a deeper understanding of operating costs. Rectified steady-state data integrated to a cost model are used as the basis for management decision-making. The interpretation of the results of the data-driven costing method employed in this paper lead to the optimization of the operation in the washing department. While as part of this paper an optimization problem is solved, there will not be any optimization algorithm per se. It is a case study were the operations are improved through the identification of the most suitable operating regime to produce in – the one that minimize the production cost, maximize the throughput and product quality.

Methodology : Operating regimes-driven cost analysis for problem resolution

Continuous processes, such as chemical pulp mills, are often characterized by a multiple products environment and numerous operating regimes. Therefore, it is crucial to evaluate individual product margins in each regime to make sure they are all profitable. Enhancing the insights and understanding of the cost variances between different operating regimes (production recipes) in the washing department of a chemical pulp mill is archive by :

- (1) ruling on a problem to solve – identifying a process optimization case study;

- (2) applying the framework proposed in Part I to obtain reconciled steady-state data in each operating-regime;
- (3) evaluating the cost for each operating regime using activity-based costing;
- (4) solving the optimization problem, and
- (5) propose an improved way to operate the washing department.

The optimization problem is detailed in section 6.1; the goal is to improve the operations. The process data required to solve it are processed based on the steps presented in Thibault et al., 2022. Namely, in order to perform the activity-based costing analysis, the process information are acquired using signal processing, steady-state detection and data reconciliation techniques for each operating regime (detected as part of the aforementioned framework).

As for the third step, the cost analysis is inspired by the operations-driven cost model framework proposed by Laflamme-Mayer et al. (Laflamme-Mayer et al., 2011). The first stage is to characterize the process operations based on process data that were dissected to describe the multiple operating regimes – identify the different resources and activities of the department. The latter must capture and represent the production chain of the system (washing department). Next, the consumption of resources is assigned to the different activities, and then the activities to the cost objects. Thus, the activities serve as a link between the consumption of resources and the cost objects (Cooper & Kaplan, 1991). In a nutshell, the cost information is extracted, and cost drivers are defined and organized into matrices that correspond to the underlying fundamental mass and energy equations. The production cost per ton of product (cost object) is evaluated.

Then, concerning the optimization problem, the most suitable operating regimes are identified by allocating the manufacturing costs to each regime. The result of this analysis will support informed decision-making, provide insights, highlight potential areas for improvement, and lead to the identification of an improved way to operate the washing department. Hence, by analysing and interpreting the outcomes of the operation-driven cost model, an improved way to operate will be identified. Further analysis to understand the causes of good and bad performances of operating regimes is performed with a principal component analysis.

All analysis as part of this paper are based on historical data. Retrieving these past data at the granularity of sub-hours (every minute for instance) over a long period of time (multiple months) allows the classification of whether the process units were steady or unsteady. The spans identified as steady may then be used in process optimization. Therefore, this could be called “optimizing the past”.

Case study : brownstock washing department in a dissolving pulp mill

The pulp, which has just been cooked, is composed of fibers (cellulose) and impurities suspended in black liquor. The latter contains dissolved wood components including lignin (the Kappa number is an indicator of the lignin content in the pulp), and the chemicals initially present in the cooking liquor. The principal objective of brownstock washing (BSW) is to separate the dissolved organics and spent cooking chemicals from the pulp so that it may then be bleached, and the liquor sent to the evaporators. In other words, the goal is to recover spent pulping chemicals as efficiently as possible and to minimize black liquor solids carryover to the bleach plant to reduce excessive bleaching chemical usage.

If the pulp is poorly washed, some dissolved components will remain and will impact the bleaching process that would hence require more chemicals. On the other hands, an over-washed pulp would mean a too diluted black liquor which would increase the steam demand of the evaporators; the amount of water to evaporate rises and therefore the energy demand increases proportionally.

The brownstock washing principles (for drum washers) are the dilution, extraction and displacement. First, the pulp is diluted in the washer vat, decreasing the concentration of dissolved solids (DS) in the black liquor. The dilution factor (DF) represents the mass of added water (wash liquor of the shower flow) minus the mass of liquid leaving with the pulp (liquor of the pulp stream), all that divided by the pulp mass. Then removing liquid from the pulp/liquor suspension reduces the total amount of DS, however the concentration stays the same. Lastly, the shower liquid flows through the pulp mat; replacing the liquid already present in the pulp (which is removed through the vacuum of the drum). The displacement ratio is a measure of washing efficiency.

Other indicators of the washing performance are the viscosity and brightness that are indirect measures of the fraction of cellulose in the pulp. The drum rotation velocity can affect the thickness of the sheet forming on the drum; to obtain the optimum sheet thickness, the best range is 2-3.5 rotations/min. Furthermore, the wash water temperature can impact the washing efficiency as a higher temperature will increase the DS removal. Lastly, the feed and discharge consistencies affect the energy consumption of the equipment, and therefore may affect the performance as well.

As part of this case study, the BSW department uses three different pieces of equipment to remove different components from the pulp (Figure 6). First off is the two-knotters cascade system used to remove uncooked wood fibers and knots. Then, the actual washing is done in two parallel lines of rotary washers (dissolved cooking components) arranged in a classic counter-current setup. The two different washing lines are not always working simultaneously. This happens when there is an operation problem happening either upstream, in the digesters, or downstream, in the beaching department. Finally is the three-screens cascade system and cleaners utilized to remove other debris (plastic, shives, sand, etc.)

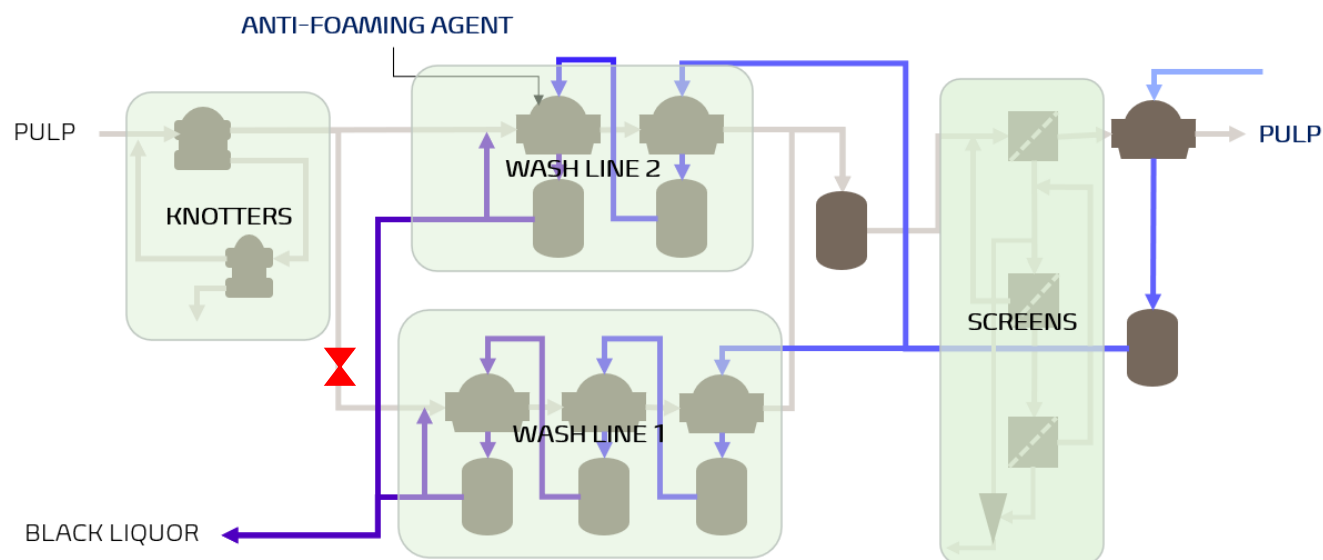


Figure 6. Brownstock washing department

This department presents several operating challenges; it is one of the most variable areas of the process, with significant fluctuations. The washing department, which operates continuously, is quite unstable given the pulping department that precedes it which operates in batches (discharge every 90 minutes). The blowing tank at the entrance of the department is not sufficient to counteract the fluctuations – these fluctuations are reflected on the tank level. Also, the low capacity of the washing department means that in order to wash all the pulp coming out of the digesters, the washers are operated at high speed, which makes them more difficult to control – the speed at which

the digesters are running influences the speed of operation of the washers; the latter must not slow down the digesters. This makes the wash department a major bottleneck. However, operating the washers at their maximum capacity could result in a loss of control over the operation and thus reduce the quality of the pulp wash, cause too much lignin to remain in the pulp and end up in the bleaching department, and result in much higher bleaching agent costs.

Moreover, the operators have little room for manoeuvre in this department; in order to wash the pulp correctly, they must put in just enough water so as not to have too much filtrate (and therefore black liquor, which will impact the other departments), put in the right amount of defoamer and try to keep up with the speed of the digesters (find the right operating speed).

Thus, one strategy used to incrementally optimize or debottleneck the washing department operations is to identify opportunities requiring little or no capital investment, based on data analysis performed using the framework presented in Thibault, Émilie, Kelly, Jeffrey Dean, et al. (2023), which considered the operating regimes, and production costs. Improved washing operating conditions (regimes) will be identified as describe in the following section.

Optimization problem definition : maximizing the washing performance

Optimizing the washing department must be in relation with its main objective : removing unwanted components from the pulp. In particular, this should be done in the most economical way possible by maximizing the profit for a maximum of dissolved solids removed from the pulp.

The objective of the optimization problem is to identify the preferred operating regime and unprofitable (or less profitable) ones; in other words, ascertain what conditions (operating parameters) would meet the following criteria:

- (1) maximize the overall yield/capacity of the washing department (pulp throughput quantity);
- (2) maximize washing quality/performance by minimizing the pulp conductivity;
- (3) maximize the profit, which consider the quantity of defoamer and the bleaching chemical usage.

The first criterion must account for the physical constraints of the department, i.e., the maximum pulp production capacities of each washer. The goal is to identify the regime that shows the highest amount of pulp delivered to the thick unbleached pulp tank. However, there is a trade-off between the capacity and the pulp quality (the washing performance).

The second criterion try to minimize a pulp cleanliness indicator, the conductivity. Quality of pulp washing is indirectly measured by its conductivity at the end of the department; the lower its value, the higher the quality. The conductivity is caused by the black liquor (lignin) that remains in the pulp. A high conductivity value means that the pulp contains lignin (component that the washing department is trying to remove as much as possible – this is what consumes the bleaching chemicals and causes the pulp to be brown). Hence, the cost of bleaching is related to the lignin content, a high lignin content left in the pulp leads to higher bleaching chemical costs. Therefore, the ideal production rate respects the washer load capacity and considers that overloading leads to poorer washing efficiency.

For the third criterion, the cost of operating under the different regimes are assessed. Foam formation is one of the main difficulties which negatively affects brownstock washing efficiency. Foam forms when wood chips contain high concentrations of extractives or when air entrainment occurs in the system. Foaming can cause production loss and increase chemical consumption. Defoamer (anti-foaming agent) is used to remove foam which prevents proper

washing; a sufficient anti-foaming agent usage increases washer capacity. The required anti-foaming agent quantity increases proportionally with the pulp production; the units are therefore kg per ton of pulp.

Another paper could also consider the shortfall in revenue when the brownstock washing department is the mill bottleneck and assess how much does it cost to not be available to the full capacity of the digesters. This would require a bottleneck analysis to highlight when the BSW was the bottleneck and how much more pulp could have been produced if that would have been optimized.

Lastly, the cost of water is not considered since its unit cost is insignificant. Something more critical would be the quantity (amount) of water sent to the evaporators – more water will overload the evaporators, and could slow down the digesters. Overloading the evaporators is very costly and will slow down the washers. Hence, the cost of evaporation per ton of pulp should be assessed, but this analysis stops at the bleaching department, and doesn't consider the evaporation department. Ultimately, minimizing the cost of water evaporation indirectly minimizes the amount of water (shower flow) used at the same time. However, there is a trade-off; the operators should consume as little water as possible, but without compromising the washing performance.

The result of this optimization analysis is the identification of the best (ideal) way to operate the washing department. The solution is not in the form of a design modification generating capital investment. As a matter of fact, paper mills are currently trying to do the most with less capital investment as possible, therefore, solutions requiring little to no investment are prioritized. The various operating regimes are going to be classified based on the added-value they bring and their cost, considering their feasibility. The activity-based costing analysis will provide the marginal cost of incremental production in each operating regime. The amount of time the process can produce in the different operating regime is also going to be ascertained.

Washing department debottlenecking and cost optimization

In order to concretize the systematic approach proposed in Thibault, Émilie, Kelly, Jeffrey Dean, et al. (2023), and its value, the operating costs of each regime are considered. These are not evaluated in an absolute way, but rather in relation to the average production of dissolving pulp, which is the reference or baseline. More specifically, knowing that the mill's objective is to produce as much pulp as possible, the economic analysis focuses on the profit generated for the additional (above average) tons of pulp produced. Also, it has been demonstrated by the mill that the revenues generated from the sale of the pulp produced always cover the costs of additional consumption of defoamer and bleaching chemicals. Every ton of pulp produced above average brings an incremental profit. Therefore, the goal is identifying the best way to operate to make the most money; finding the operating regimes that are the most profitable.

Each different way of operating is assigned a profit obtained with respect to a baseline, which is estimated at 475 tons/day. Resources considered in this operation-driven cost analysis are the average amount of defoamer, and bleaching chemicals consumed during the different operating regimes; the washing department is not energy intensive, therefore energy cost was not considered. Activities that consume these resources are the usage of the washers and the reactions happening in the bleaching towers. The cost object is the washing department.

By considering the absolute production costs per ton of pulp produced and their breakdown into direct and indirect costs, it is possible to identify that for each additional ton of pulp produced compared to the average, the mill makes a gain. Therefore, considering the amount of pulp produced above average and the amount of raw material consumed, by subtracting the costs from the gain, it is possible to obtain the profit for each operating regime.

This marginal cost analysis provides the savings per operating regime for a year of operation (355 days). However, the feasibility must be accounted for; identify how many days a year it is possible to operate this way and reduce the operating costs. These costs may be saved in specific conditions when the washing is in fact the mill's bottleneck.

The targeted application or analysis dictate which operating regimes are required and how they should be defined (the level of data aggregation or granularity). A possible avenue would be to focus on identifying operating regimes that are less cost-effective and assess whether they are essential and/or avoidable. However, the operating regimes identified for this case study are those in the summertime (where the mill operates at its fullest), and especially when the process is running smoothly (producing more than 475 ton of pulp per day). Therefore, days where less than 15 batches were produced were not considered; less than 15 batches indicates a problem in the operations, these data are not representative of good operations.

During summertime, the plant operates at a higher production rate and the washing department is the limiting factor. Thus, the maximum capacity of the equipment in this department is often reached and this results in lower washing performance. In winter, production rates are lower, which brings more stability to the operation – the wash flow is easier to control; the washing department is not the bottleneck. Thus, having less pulp to process causes the washing department to perform better in winter. However, having a lower production is never economically desirable. Thus, in general within the mill, seasons play a critical role in the operations of the process. This impact the number of digesters used (batches produced) each day (in winter, fewer digesters are used) – the number of digesters is directly related to the pulp production rate, and the process water temperature is colder in winter than in summer.

The steady-state operating regimes identified using the proposed Framework (Thibault, Émilie, Kelly, Jeffrey Dean, et al., 2023) should last at least 12 hours. It is the minimal time for the steady-state to be representative and considering that shift rotations happen twice a day. During each shift, the working team is responsible to achieve the department's objectives. Given the fact that some operating regimes lasted longer than 12 hours, these regimes show an overlap between the teams. The operating regimes are compared based on which team of operators was working during the operating regime duration; those are assimilated to different production scenarios as different operator teams won't necessarily take the same action when reacting to an event happening at the mill. The operating regime are also distinguished by key parameters including their averaged pulp production and quality (conductivity) values, and their profit relatively to the averaged production.

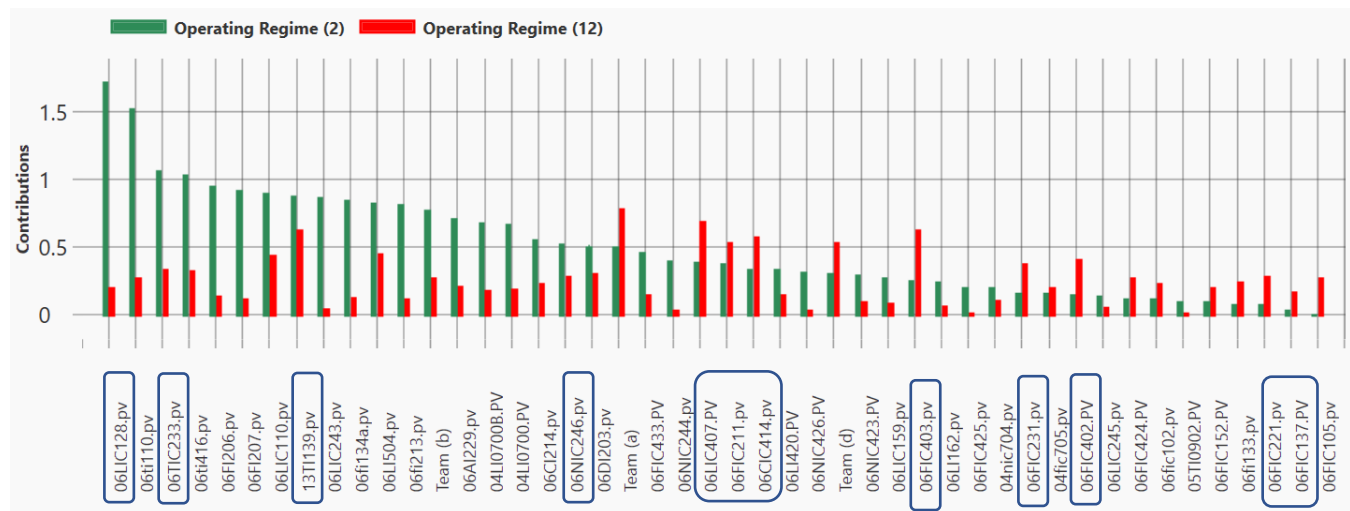
Table 3 presents the key parameters values averaged over the run length (duration of each operating regime). The last column provide the profit in millions of dollar per year for each operating regime. Given the different operating regime identified based on the framework presented in Part I, the most favorable ones can be recognized according to the parameters values that uniquely defines each one of the operating regime.

Table 3 – Key parameters for all detected operating regimes

Operating regime	Team responsible	Pulp throughput (ton/day)	Pulp conductivity (µmhos)	Profit (M\$/year)
1	C/D	654	3816	32
2	C/D	642	4045	30
3	C	661	4099	33
4	C/D	673	4006	35
5	C/D	681	4124	37
6	A/C	683	4363	37
7	D	677	4616	36
8	C/D	660	4626	33

9	D	668	4489	34
10	C/D	701	4436	40
11	C/D	721	4376	43
12	A/B	754	4389	49

From Table 3, it is possible to observe that the preferred combination is operating regime 12. It saves 49 M\$ per year, and it is obtained when the process is operated by teams A and B, that 754 ton/day are produced, and that the conductivity of the pulp is 4389 μ mhos. In order to assess the reasons behind the performance of operating regime 12, it was compared with the least performing one, operating regime 2. Even if both are steady-state operating periods happening when the process is producing more than 15 batches of pulp a day, i.e. improved mill operations, a profit of 19 M\$ differentiate these two operating regimes. Therefore, a contribution analysis of these operating regimes emphasised on which variables varied (and impacted) the most the operating regimes in the washing department. The identification of these critical variables addresses and explain the differences between operating regimes, and highlight how the process was operated when the improved performances happened (Figure 7). The logic or the rational behind the identified impactful controlled variables is ascertained with the collaboration of a process expert.



A third interesting variable is 06LIC407, level of the filtrate tank of the thickener. The operators modify its value to reach an overflow thinking that it would increase the pulp throughput. However, operating regime 12 produce the most amount of pulp throughput with its level of the filtrate tank of the thickener 17% lower than the one exhibited by operating regime 2.

Another interesting controlled variable is 06CIC414, the pulp conductivity. Unlike what the operators commonly believe, high pulp throughput does not rime with conductivity higher than 5000 μmhos , i.e., it is not required to push conductivity to values higher than 5000 to reach high pulp quantity. In all 12 operating regimes representing good operating periods, the conductivity is systematically below 5000. This proves that it is possible to achieve high pulp throughput with low conductivity.

Lastly, the showers water flow, which include 06FIC402, 06FIC403, 06FIC137, 06FIC211, 06FIC221, 06FIC231, are controlled by operators; these are major leverages. Better performance is obtained when more water is added through the showers. It was found that all six variables showed higher values in operating regime 12 compared to operating regime 2. Furthermore, showers water temperature (06TIC233) is controlled by operators by adding hot or warm water to the showers water flow. This variable is equally as important as the showers water flow. In the same way, warm water temperature (13TI139) greatly impact the washing performance. Higher temperature leads to higher washing performance. Even if this variable is not a control loop, it would be interesting to assess how warmer water can be obtained on a continuous basis.

Another variable that is not controlled but greatly affect the washing performance is **06TI416** (temperature of the unbleached pulp tank). Higher temperature rimes with easier washing operations, and higher pulp throughput. This temperature indicator is impacted by the shower water temperature and the pulp temperature at the entrance of the thickener.

Therefore, focusing especially on controlled variables in order to understand what operators did differently in operating regime 12, it was found that the variables that have the most potential to influence the process for better results are 1A washer level, pulp consistency, level of the filtrate tank of the thickener and shower water flow and temperature.

Finally, a succinct analysis of the impact of the working team is undertaken. Since Teams A and B, and C and D work together, Figure 8 shows the variables contribution to each pair of teams.

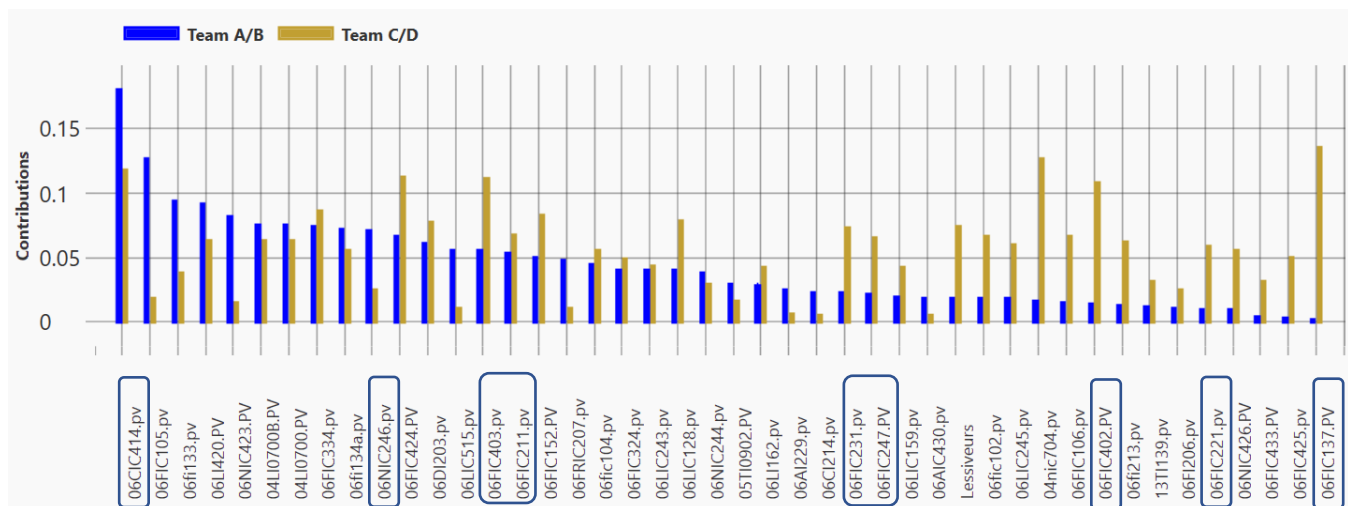


Figure 8 – Variables contribution for Teams A/B and C/D

Figure 8 illustrates their different way of operating the process; the impact of working team is reflected on several variables including the pulp conductivity (06CIC414), the pulp consistency (06NIC246), pulp throughput (06FIC247), showers water flow (06FIC402, 06FIC403, 06FIC137, 06FIC211, 06FIC221, 06FIC231). Team A/B showed higher conductivity values, they used less water, and they produced slightly less pulp, and its conductivity was lower. Globally, further analysis should be completed to assess which team performs better than the others.

Results and Discussion : Framework's added-value concretization, advantages, limitations and risks

At the mill, experts are trying to do the most with what they have without capital-investment modifications. For instance, in order to optimize and debottleneck the washing department, they would start by stabilizing the process-controlled variables. The most important variable is the consistency at the outlet of the blow tank; the consistency varies each time there is a digester discharge. By stabilizing this consistency, the pulp going into the washing department is stable. Then, considering that the pulp is discharged into a blow tank that does not have enough capacity, the mill could use an old tank that is not in function to act as a buffer. This buffer tank would help reduce the consistency variations as well. This stabilization gives the possibility to increase the washing capacity. Moreover, considering that the washing efficiency decrease with the temperature since lignin remains stuck on the fibers and won't come off with cold water, i.e., pulp washing is easier with higher wash water temperature, it would be possible to optimize the operations of the washing department by installing a steam heater on the wash water flow for the winter conditions. By doing so, the production gap between summer and winter conditions could be reduced and the process could remain more stable throughout the year. Therefore, controlling the process water temperature better would provide more capacity. However, even if having warmer water in winter would lead to better performance, digesters are more a bottleneck during winter than the washers, and a large steam heater would be required to heat winter water to reach the same temperature as in the summer. Hence, further analysis should assess the benefits versus the efforts required. A last option, which is not without investment cost, is to replace existing washers by more efficient ones.

Therefore, considering the steps undertaken in practice by the mills, data analysis and operating regime detection and identification is proposed. In this paper, the brownstock washing department was optimized (debottlenecked) given the surimposition of processed operating data and cost information. The results are presented to the mill expert as a complement of what they are already doing at the mill level. This analysis is intended to provide meaningful (impactful) solutions that can be implemented in a cost-effective and timely manner.

Data processing and manipulation is relative to the objective of the analysis, the data usage; it is purpose-specific. Based on the purpose of this case study, the framework, i.e., the operating regimes, was required. In other terms, the level of precision acquired by using the framework was necessary; by considering solely the summertime average for all variables, it would not be possible to identify the preferred way to operate. Distinguishing between the different operating regimes lead to a deeper understanding of what is happening in the process; operating regimes are used to answer questions and solve problematics in a plant.

Solving this optimization problem showed the value and benefits of using the proposed framework; it adds significant value for process understanding and troubleshooting. The framework (especially the identification of operating regimes) give insight about process variability, it helps explain where the variations in the process come from, when exactly they happened and which variables were impacted. Furthermore, processed data resulting from the framework is a starting point or the basis for numerous analysis or decision-making that could take place at the mill. Having good quality (processed) data and operating regimes obtained through the combination of data-driven

approaches and process interpretation may lead to an increase in the mill's profit. For instance, knowing the maximum capacity of each unit operation in the process or equipment in a system, processed data for each operating regime help to make recommendations for low-cost changes to increase capacity (debottleneck).

Furthermore, together, the framework and the operation-driven cost model increase transparency, understanding of cause and effect and granularity (detail) of optimization results, and allows the identification of critical variables at the operation level. They also provide improved insights into production costs, enhanced cost-control via benchmarking key performance indicators (process variables in Table 3), and helped to increase the overall corporate profitability.

Therefore, operation-driven cost analysis supports informed decision-making. It provides information on the various cost objects and studies their profitability as part of a decision-making process. Decision-making based on the data from the framework improves the efficiency and profitability of operations. Overlaying processed data from the framework onto ABC data by considering the plant operations helps make better decisions regarding operations. It provides the marginal cost of each operating regime by allocating overhead to cost objects in direct response to their consumption and identifies which operating regime creates the most value.

In summary, assigning costs to the operating regimes allow to solve operation problems and identify regimes that are the most profitable. Thus, in order to answer operational questions, the definition of the various regimes is essential. Economic data combined with operating regimes uncovered hidden insights of various drivers of profitability, lead to the minimization the operating costs of the washing department as well as the identification of the actions to take to increase profits.

Therefore, operating in the preferred conditions prescribe by operating regime 12 lead to an increased or higher throughput. This productivity increase rimes with high performance; high quality product, profit, revenue increase and operational flexibility. These improved operating conditions increase profitability by maximizing the production profit. Hence, advanced data analytics combined with its context (process knowledge) gives more insight and lead to a deeper understating of the process.

However, experts mentioned that the value of the framework could increase tremendously if it could be used online, in real-time, to establish, for instance, if the operations are deviating from normal operations, and to assess the underling cause. Furthermore, processed data in this case represents steady-state operations; transient state is not considered. To maximize its profit, it could be interesting for a mill to assess the cost of the transition between steady-sate regime, and try to minimize the transition time.

Lastly, instead for surimposing a cost model to the data obtained from the proposed framework, an environmental analysis could have been performed to highlight the most problematic regime and reduce the environmental impact.

Conclusion

The data storage capacities of modern process automation systems have grown rapidly. Nowadays, the systems are able to store hundreds of measurements in databases. However, these data have limited value until they are transformed into information. Using data manipulation approaches allow to take full advantage of the data collected.

Knowing the implications of each operating regime in an industrial process guide the management in their decision-making process. Plants have different operating regimes and process data must be interpreted based on these regimes in order to make more informed design and operating decisions. Operating regime detection and identification helps to better understand how the mill operates and increases the ability to interpret process data thoroughly and deeply. It provides additional information on process operation and offers an applicable way of defining the best mode for

achieving a more efficient and cost-efficient process. Therefore, recognizing different operating regime opens the door for many new opportunities.

In paper presents a case study where operating regimes are required. The operations of the washing department of a chemical pulp mill are assessed in order to identify the preferred operating regime in terms of pulp throughput, pulp quality (conductivity) and profit. More specifically, the washing performance optimization objective function considers the maximization of the pulp throughput quantity, the washing quality as well as the minimization of the anti-foaming agent quantity and bleaching chemical, i.e., profit maximization. The cost analysis considered the activity-based costing philosophy; the rate of resource spending is traced to an activity, and the activity is then traced to the product.

Therefore, processed data obtained from the proposed framework combined with cost information are interpreted to identify the profitable operations. The most cost-efficient operating regime is detected and the reasons explaining its high performance are assessed. This is achieved by using data-driven techniques in conjunction with process knowledge. Hence, to make smart decisions based on huge volume of operating data collected from complex industrial processes, cost information and operating regimes should be considered.

In conclusion, using a process data management framework allowed the identification of potential benefits that could be attained by the mill; hence it brings tremendous value the process expert. Knowing operating regimes improves the ability to solve operating problems and identify the most cost-efficient regime. Controlled variables that appear to influence the performance of the washing department are identified. Further analysis, starting with a design of experiment, could isolate the effect of each variable, assess their relations, prove that these variables impact the performance, explain further how they lead to a higher performance, and confirm the preferred operating scenario.

Acknowledgements

The authors gratefully acknowledge the financial support of the CRSNG.

References

- AccountingTools. (2018). *Cost accounting definition*. <https://www.accountingtools.com/articles/what-is-cost-accounting.html>
- Averkamp, H. (2019). *What is cost accounting?* Accounting Coach. <https://www.accountingcoach.com/blog/what-is-cost-accounting>
- Cantrell, J. (2019). *Brownstock washing cost optimization* TAPPI, St. Louis.
- CBC. (2009). *Pulp and paper woes reach far and wide*. CBC News. <https://www.cbc.ca/news/business/pulp-and-paper-woes-reach-far-and-wide-1.830189>
- Chew, I. M. L., Foo, D. C. Y., Bonhivers, J.-C., Stuart, P., Alva-Argaez, A., & Savulescu, L. E. (2013). A model-based approach for simultaneous water and energy reduction in a pulp and paper mill. *Applied Thermal Engineering*, 51(1), 393-400. <https://doi.org/https://doi.org/10.1016/j.applthermaleng.2012.08.070>
- Cooper, R., & Kaplan, R. S. (1988). How Cost Accounting Distorts Product Costs. *Management Accounting*, 69(10), 20.

- Cooper, R., & Kaplan, R. S. (1990). Measure Costs Right: Make The Right Decision. *The CPA Journal*, 60(2), 38.
- Cooper, R., & Kaplan, R. S. (1991). Profit Priorities from Activity-based Costing. *Harvard business review U6*, 69(3), 130-135.
- El-Halwagi, M. M., & Stuart, P. R. (2013). *Integrated biorefineries: design, analysis, and optimization*. CRC Press.
- Emblemsvåg, J. (2003). *Life-Cycle Costing: Using Activity-Based Costing and Monte Carlo Methods to Manage Future Costs and Risks*. John Wiley & Sons, Inc.
- Emblemsvåg, J., & Bras, B. (2000). *Activity-based cost and environmental management : a different approach to ISO 14000 compliance*. Kluwer Academic Publishers.
- Greenwood, T. G., & Reeve, J. M. (1992). Activity-based cost management for continuous improvement: A process design framework. *Journal of Cost Management the Manufacturing Industry* 5(4), 22-40. <https://maaw.info/ArticleSummaries/ArtSumGreenwoodReeve92.htm>
- Hsieh, C. L., Tsai, W. H., & Chang, Y. C. (2020). Green Activity-Based Costing Production Decision Model for Recycled Paper. *Energies*, 13(10), 23. Article. <https://doi.org/10.3390/en13102413>
- Hytönen, E., & Stuart, P. (2011). Operations driven cost impact evaluation of kraft process retrofit projects for capital appropriation : case of forest biorefinery. *Journal of Production Economics*.
- Janssen, M., Naliwajka, P., & Stuart, P. (2012). Development of an operations-driven cost model for continuous processes part II: Retrofit process design application. 2, 43-54.
- Janssen, M., & Stuart, P. R. (2004). Business Model Framework for Using Real-Time Process Data for Design Decision Making.
- Kopra, R., Kari, E., Harinen, M., Tirri, T., & Dahl, O. (2010). Optimization of wash water usage in brown stock washing. *2010 TAPPI PEERS Conference and 9th Research Forum on Recycling*, 1, 660-695. <https://doi.org/10.32964/TJ10.9.27>
- Kopra, R., Karjalainen, S., Tirri, T., & Dahl, O. (2011). Optimisation of pressure filter performance using refractometer measurements -Mill investigations. *Appita Journal*, 64.
- Kopra, R., Koivula, K., & Dahl, O. (2015). Optimization of a brownstock washing line. *Tappi Journal*, 14, 83-91. <https://doi.org/10.32964/TJ14.2.83>
- Korbel, M. (2011). *Ph. D. Thesis : On-line steady-state data reconciliation for advanced cost analysis in the pulp and paper industry*, Polytechnique].
- Korbel, M., Garrigues, L., & Stuart, P. R. (2007). *Process and Business Data Reconciliation in the Pulp & Paper Industry*. Innovations in Engineering, Pulping and Environmental, Jacksonville, FL.
- Korbel, M., & Stuart, P. (2011). Assessment and interpretation of advanced cost data for process improvement in an integrated newsprint mill. *International Journal of Production Economics*.
- Korpunen, H., & Paltakari, J. (2013). Testing an activity-based costing model with a virtual paper mill. *Nordic Pulp and Paper Research Journal*, 28(1), 146-155.

- Korpunen, H., Virtanen, P., Dahl, O., Jylhä, P., & Uusitalo, J. (2012). An activity-based cost calculation for a kraft pulp mill. *Tappi Journal*, 11(9), 19-27.
- Laflamme-Mayer, M., Janssen, M., & Stuart, P. (2011). Development of an operations-driven cost model for continuous processes Part 1: Framework for design and operations decision making. *J-Forest Journal of Science & Technology for forest products and processes*, 1(1), 32-41.
- Laflamme-Mayer, M., Linkewich, J., & Stuart, P. R. (2008). Methodology for using real-time process data for cost modeling and supply chain decision support in the pulp and paper industry. 109(3), 44-51.
- Marinova, M., Mateos-Espejel, E., Jemaa, N., & Paris, J. (2009). Addressing the increased energy demand of a Kraft mill biorefinery: The hemicellulose extraction case. *Chemical Engineering Research and Design*, 87(9), 1269-1275. <https://doi.org/10.1016/j.cherd.2009.04.017>
- Murphy, C. B. (2019). What are the types of costs in cost accounting? . <https://www.investopedia.com/ask/answers/041415/what-are-different-types-costs-cost-accounting.asp>
- Nurminen, T., Korpunen, H., & Uusitalo, J. (2009). Applying the Activity-Based Costing to Cut-to-Length Timber Harvesting and Trucking. *Silva Fennica*, 43(5), 847-870. Article. <https://doi.org/10.14214/sf.177>
- Quesada-Pineda, H. (2004). The ABCs of Cost Allocation in the Wood Products Industry: Applications in the Furniture Industry. *Virginia cooperative extension*, 420-147. <https://www.pubs.ext.vt.edu/420/420-147/420-147.html>
- Rivero, E. J., & Emblemståg, J. (2007). Activity-based life-cycle costing in long-range planning. *Review of Accounting and Finance*, 6, 370-390. <https://doi.org/10.1108/14757700710835041>
- Sawhney, R. S. (1991). An activity-based approach for evaluating strategic investments in manufacturing companies. *Journal of Manufacturing Systems*, 10(5), 353-367. [https://doi.org/10.1016/0278-6125\(91\)90054-6](https://doi.org/10.1016/0278-6125(91)90054-6)
- Svensson, E. (2014). Flexibility to Seasonal Demand Variations in Pulp Mill Steam Production: The Effect of Steam Savings Leading to Off-design Heat Loads. *Applied Thermal Engineering*, 70. <https://doi.org/10.1016/j.applthermaleng.2014.04.059>
- Svensson, E., & Berntsson, T. J. C. e. t. (2013). Operational Flexibility in Pulp Mill Steam Production at Off-design Heat Loads. 35, 919-924.
- Svensson, E., Edland, R., Langner, C., & Harvey, S. (2020). Assessing the value of a diversified by-product portfolio to allow for increased production flexibility in pulp mills. 35(4), 533-558. <https://doi.org/10.1515/npprj-2020-0034>
- Tatsiopoulos, I. P., & Panayiotou, N. (2000). The integration of activity based costing and enterprise modeling for reengineering purposes. *International Journal of Production Economics*, 66(1), 33-44. [https://doi.org/10.1016/S0925-5273\(99\)00080-8](https://doi.org/10.1016/S0925-5273(99)00080-8)
- Thibault, É., Chioua, M., McKay, M., Korbel, M., Patience, G., & Stuart, P. (2023). Experimental Methods in Chemical Engineering: State-of-the-Art Review of Industry Practice for Data Processing and Data Usage in Decision-Making. *Canadian Journal of chemical engineering*.

- Thibault, É., Kelly, J. D., Lebreux Désilets, F., Chioua, M., Poulin, B., & Stuart, P. (2023). An operating regime-based data-driven industrial framework for improved process decision-making - Part 1: Framework development *Industrial & Engineering Chemistry Research*.
- Thibault, É., Lebreux Désilets, F., Poulin, B., Chioua, M., & Stuart, P. (2023). Comparison of signal processing methods considering their optimal parameters using synthetic signals in a heat exchanger network simulation. *Computers & Chemical Engineering*.
- Thibault, É., Lebreux Désilets, F., Poulin, B., Chioua, M., & Stuart, P. (2023). Signal processing heuristics for the selection of wavelet transform parameters considering signal characteristics derived from a case study on a chemical pulp mill. *Computers & Chemical Engineering*.
- Wang, C., Shan, L., Zhou, L., & Zhang, G. (2008// 2008). *Study on the Application of ABC System in the Refinery Industry*. Global Design to Gain a Competitive Edge, London (p. 773-781).