UQAC
Université du Québec
à Chicoutimi

AI-DRIVEN ANOMALY DETECTION FOR HEALTHCARE SUPPLY CHAINS: A PIPELINE INTEGRATING PRE-PROCESSING, HYPERPARAMETER TUNING, AND POST-PROCESSING FILTERS.

BY COLIN BOUCHARD

MASTER'S THESIS PRESENTED TO L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE (M.SC.) IN COMPUTER SCIENCE

RESEARCH IN ARTIFICIAL INTELLIGENCE

QUÉBEC, CANADA

RÉSUMÉ

La gestion des chaînes d'approvisionnement dans le réseau de la santé et des services sociaux (RSSS) mobilise des processus complexes où des anomalies (variations de prix, retards de livraison, retours inhabituels) peuvent nuire à l'efficacité, à la transparence et à la sécurité des soins. Ce mémoire propose une approche intégrée de détection d'anomalies basée sur l'Intélligence Artificielle (IA), à partir de données réelles issues du système SAP du Centre intégré universitaire de santé et services sociaux du Saguenay–Lac-Saint-Jean(CIUSSS-SLSJ). En raison de contraintes de confidentialité institutionnelle, les jeux de données SAP et le code ne peuvent pas être rendus publics; seuls des résultats agrégés et les détails méthodologiques sont présentés. Le travail est structuré en deux manuscrits de type article (préprints) destinés à une soumission ultérieure. Le premier article décrit un pipeline dans un environnement pandas combinant des algorithmes non supervisés [Isolation Forest, One-Class Support Vector Machine, Local Outlier Factor, Robust Covariance utilisée dans une Envelope Elliptique (EE)], une optimisation bayésienne des hyperparamètres et des filtres de post-traitement adaptés au domaine. Les résultats suggèrent que l'association de méthodes IA et de filtres spécifiques au domaine améliore la précision et la pertinence opérationnelle de la détection d'anomalies dans les données d'approvisionnement du CIUSSS-SLSJ. Le second article traite de la portabilité vers Apache Spark en réimplantant le pipeline et en évaluant la reproductibilité des résultats en environnement distribué. Les performances demeurent robustes, tout en mettant en évidence des défis techniques liés aux algorithmes et au partitionnement des données. Globalement, le système vise à améliorer la qualité des alertes, notamment en réduisant les faux positifs afin d'identifier des écarts plus actionnables, et à préparer une intégration future dans des environnements ERP centralisés. Les résultats soutiennent qu'un pipeline piloté par l'IA peut rendre les alertes plus exploitables, notamment grâce à l'intégration des filtres de post-traitement comme éléments de conception clés. Le travail contribue à la littérature sur la détection d'anomalies en santé et propose des pistes pour renforcer l'efficacité des activités d'approvisionnement du CIUSSS-SLSJ, avec un potentiel de transfert vers d'autres organisations du RSSS sous réserve de validations multi-établissements. Cette conclusion demeure limitée par le caractère mono-institutionnel de l'étude et par un cadre d'évaluation semi-supervisé. Les métriques reposent sur des étiquettes binaires fournies par un seul spécialiste en procédés administratifs du CIUSSS-SLSJ, ce qui réduit la généralisabilité des scores. Pour les travaux futurs, une validation multi-experts et une mesure d'accord inter-juges sont recommandées pour renforcer la robustesse des conclusions.

Mots-clés : intelligence artificielle, détection d'anomalies, apprentissage non supervisé, algorithmes, optimisation bayésienne, hyperparamètres, pipeline de données, traitement distribué, Apache Spark, Python, Pandas, filtre de post-traitement, transformation numérique en santé, performance opérationnelle, indicateurs clés de performance, Santé Québec, Robust Covariance (RC) in Elliptic Envelope (EE), Isolation Forest (IF), One-Class SVM (OC-SVM), Local Outlier Factor (LOF).

ABSTRACT

Supply chain management in Health and Social Services Network / Réseau de la santé et des services sociaux (RSSS) involves complex processes where anomalies (price variations, delivery delays, abnormal returns) can undermine efficiency, transparency, and care safety. This thesis proposes an integrated Artificial Intelligence (AI)-based anomaly detection approach using real SAP data from the Integrated University Health and Social Services Center of Saguenay-Lac-Saint-Jean / Centre intégré universitaire de santé et de services sociaux du Saguenay-Lac-Saint-Jean (CIUSSS-SLSJ). Due to institutional confidentiality constraints, the underlying SAP datasets and implementation code cannot be publicly released; only aggregated results and methodological details are reported. The work is structured around two article-style preprints intended for future submission. The first presents a pipeline in a pandas environment that combines unsupervised Machine Learning (ML) algorithms [Isolation Forest (IF), One-Class Support Vector Machine (OC-SVM), Local Outlier Factor (LOF), Robust Covariance (RC) in Elliptic Envelope (EE)], Bayesian hyperparameter optimization, and domain-adapted post-processing filters. The results suggest that pairing AI-based methods with domain-adapted filters can improve both accuracy and operational relevance for anomaly detection in CIUSSS-SLSJ supply chain data. The second article focuses on portability by re-implementing the pipeline in Apache Spark and examining the reproducibility of the results of the first article in a distributed environment. Overall, the proposed system aims to improve alert quality, particularly by reducing false positives to better identify actionable outliers, and to support future integration into centralized Enterprise Resource Planning (ERP) environments. The findings indicate that an AI-driven pipeline can make alerts more actionable when domain-specific post-processing filters are treated as first-class design elements alongside hyperparameter tuning. The work contributes to the healthcare anomaly detection literature and outlines practical pathways to strengthen CIUSSS-SLSJ procurement operations, with potential relevance to other RSSS institutions subject to further multi-site validation. The quantitative findings are promising for this case study, but technical challenges related to algorithms and data partitioning are highlighted. This conclusion is limited by the single-institution scope and a semi-supervised evaluation setting. Performance metrics were derived from binary labels provided by a single Administrative Process Specialist from CIUSSS-SLSJ, reflecting alignment with that expert's interpretation of risk rather than an objective institutional reference. Future work should include multi-expert labelling and inter-rater agreement analysis to estimate label uncertainty and reduce dependence on a single expert's mental model.

Keywords: artificial intelligence, anomaly detection, unsupervised learning, algorithms, Bayesian optimization, hyperparameters, data pipeline, distributed processing, Apache Spark, Python, Pandas, post-processing filter, digital transformation in healthcare, operational performance, key performance indicators, Santé Québec, Robust Covariance (RC) in Elliptic Envelope (EE), Isolation Forest (IF), One-Class SVM (OC-SVM), Local Outlier Factor (LOF).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADR Adverse Drug Reaction

AI Artificial Intelligence

ANQ National Assembly of Québec / Assemblée nationale du Québec

ANOVA Analysis of Variance

AUC Area Under the Curve

AUC-ROC Area Under the Receiver Operating Characteristic Curve

AUC-PR Area Under the Precision-Recall Curve

AutoML Automated Machine Learning

API Application Programming Interface

BI Business Intelligence

BO Back Order

CER Research Ethics Committee / Comité d'éthique de la recherche

CISSS Integrated Health and Social Services Center / Centre intégré de santé et de services sociaux

CIUSSS Integrated University Health and Social Services Center / Centre intégré universitaire de santé et de services sociaux

CIUSSS-SLSJ Integrated University Health and Social Services Center of Saguenay-Lac-Saint-Jean / Centre intégré universitaire de santé et de services sociaux du Saguenay-Lac-Saint-Jean

CPI Provincial Indicators Committee / Comité Provincial des Indicateurs

CRAR Reference Framework for Responsible Procurement / Cadre de référence en approvisionnement responsable

CNN Convolutional Neural Network

DRFA Finance and Procurement Resources Department / Direction des ressources financières et de l'approvisionnement

DRI Information Resources Department / Direction des ressources informationnelles

DW Data Warehouse

DbOD Distance-based Outlier Detection

EDA Exploratory Data Analysis

EE Elliptic Envelope

ERP Enterprise Resource Planning

ETL Extract, Transform, Load / Extraire, transformer, charger

FP False Positive

FPR False Positive Rate

FN False Negative

FNR False Negative Rate

IF Isolation Forest

KNN k-nearest neighbors

GAN Generative Adversarial Network

GMM Gaussian Mixture Model

GRM Global Resource Management

ROC Receiver Operating Characteristic

RNN Recurrent Neural Network

HMM Hidden Markov Model

HP Hyperparameter

HPO Hyperparameter Optimization

IEO Improved Equilibrium Optimizer

IoT Internet of Things

KPI Key Performance Indicators

KDE Kernel Density Estimation

LCAG Act respecting the Government Acquisitions Center / Loi sur le Centre d'acquisitions gouvernementales

LCOP Act respecting contracting by public bodies / Loi sur les contrats des organismes publics

LGSSSS Act respecting the governance of the health and social services system / Loi sur la gouvernance du système de santé et de services sociaux

LMRSSS Act to Modify the Organization and Governance of the Health and Social Services Network / Loi modifiant l'organisation et la gouvernance du réseau de la santé et des services sociaux

LOF Local Outlier Factor

LSSS Act respecting health services and social services / Loi sur les services de santé et les services sociaux

LSTM Long Short-Term Memory

ML Machine Learning

M2M Machine-to-machine

MSSQL Microsoft SQL Server

MSSS Ministry of Health and Social Services (Québec) / Ministère de la Santé et des Services sociaux du Québec

OC-SVM One-Class Support Vector Machine

ODBC Open Database Connectivity

PR Precision–Recall

PABSGC Policy on the Acquisition of Goods and/or Services and Contract Management / Politique sur l'acquisition de biens et/ou de services et la gestion des contrats

PCA Principal Component Analysis

PPE Personal Protective Equipment

RBF Radial Basis Function

RC Robust Covariance

RDD Resilient Distributed Dataset

RF Random Forest

RSSS Health and Social Services Network / Réseau de la santé et des services sociaux

SAP Systems, Applications and Products

SD Standard Deviation

SQL Structured Query Language

SSSS Health and Social Services System / Système de services de santé et sociaux

SSIS SQL Server Integration Services

SSAS SQL Server Analysis Services

SSDQ SQL Server Data Quality

SVM Support Vector Machine

SMAC Sequential Model-Based Algorithm Configuration

TP True Positive

TPR True Positive Rate

TN True Negative

TNR True Negative Rate

UQAC University of Québec at Chicoutimi / Université du Québec à Chicoutimi

DEDICATION

On a personal and philosophical note. To all those who dare to challenge the limits of their own reality tunnels and imagine new possibilities. Everything is possible, for what we emit is what we receive. You don't need to change the outer world. Simply change the inner one, and the outer world will rearrange itself. In this future world shaped by artificial intelligence, it is not the machine that dictates our reality, but our consciousness which, amplified by these tools, creates new dimensions of existence. Be careful what you desire, for artificial intelligence tends to amplify the intentions and data we feed into it, whether they lead toward light or toward shadow.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to all the people who, directly or indirectly, contributed to the completion of this master's thesis.

I am grateful to my thesis supervisors: Bruno Bouchard, for his methodological guidance, and Julien Maitre, for his technical expertise in AI.

My thanks also go to the CIUSSS-SLSJ, and more particularly to the management, which made this project possible. A special thanks to my superiors: Martial Fradet, for his inspiring ideas that nourished my research topic, and Manon Tremblay, project director of finance and procurement, for her support with senior management of CIUSSS-SLSJ. I would also like to thank the Finance and Procurement Resources Department / Direction des ressources financières et de l'approvisionnement (DRFA), in particular Pierre-Olivier Pedneault, director, for approving and supporting the project by providing the necessary resources. Thanks as well to Marc Ouellet and Luc Boutin, Administrative Process Specialists in procurement and finance respectively, for their collaboration and explanations related to legal issues (Act respecting contracting by public bodies / Loi sur les contrats des organismes publics, etc.), which strongly impacted the labeling of the datasets. My gratitude also extends to the Information Resources Department / Direction des ressources informationnelles (DRI) of Integrated University Health and Social Services Center of Saguenay-Lac-Saint-Jean / Centre intégré universitaire de santé et de services sociaux du Saguenay-Lac-Saint-Jean. Thank you to Philippe Perron, Head of Cybersecurity, for his collaboration in ensuring compliance with cybersecurity standards and laws, as well as to Jean-Rock Duchesne and Luc Tremblay, IT service managers, for approving the institutional fit on the IT side. Finally, a special thank-you to Martin Villeneuve, Director of Information Resources, for facilitating access to data, information, and resource persons essential to this work. I also thank the Research Ethics Committee / Comité d'éthique de la recherche (CER) of Integrated University Health and Social Services Center of Saguenay-Lac-Saint-Jean / Centre intégré universitaire de santé et de services sociaux du Saguenay-Lac-Saint-Jean for their support and for greatly facilitating all the procedures with the various departments involved in this thesis. Thanks to Andrée Hardy, Administrative Technician, who helped me with the multitude of required forms. A note of recognition also goes to Elise Duchesne, Chair of the Integrated University Health and Social Services Center of Saguenay-Lac-Saint-Jean / Centre intégré universitaire de santé et de services sociaux du Saguenay-Lac-Saint-Jean Research Ethics Committee / Comité d'éthique de la recherche, researcher, and professor at University of Québec at Chicoutimi / Université du Québec à Chicoutimi (UQAC), without whom this project could not have been approved.

I also wish to thank the Provincial Indicators Committee / Comité Provincial des Indicateurs (CPI), previously attached to the Ministry of Health and Social Services (Québec) / Ministère de la Santé et des Services sociaux du Québec (MSSS) and now to Santé Québec, for its strategic guidance regarding the use of Key Performance Indicators (KPI). Their work on

supplier performance evaluation and on the prevention of supply disruptions, notably through the monitoring of delivery delays, had a decisive influence on the direction of this project. A special thank-you to Martin Michaud, formerly of the Ministry of Health and Social Services (Québec) / Ministère de la Santé et des Services sociaux du Québec and now at Santé Québec, for his expertise, as well as to Stefan Sawoszczuk, chair of the committee, for his strategic orientation of the committee, which influenced this thesis. Thanks also to all the other members who contributed, directly or indirectly, to the committee's work; your collaboration has been deeply inspiring.

Finally, I thank my wife, Marie Dubreuil, whose daily support, attentive listening, and shared experience were invaluable in facing the challenges of this project. My gratitude also goes to my mother, Diane Charbonneau, who introduced me to management concepts and supported me throughout this academic journey, as well as to my father, Luc Bouchard, for his unwavering support and enlightening explanations in plumbing, which I was able to fruitfully transpose to the field of data pipelines. I also thank my extended family and friends for their constant encouragement throughout this journey. A special nod to my dog, Soya, who reminded me, even in the most difficult moments, of the importance of getting fresh air and disconnecting. And an emotional final note to my unborn child, who accompanied me, in their own way, through the last stages of this *mémoire*.

To all of you, I say thank you.

PREFACE

In a context where the efficiency and reliability of healthcare supply systems directly determine the quality of care, this thesis is part of an innovation-driven approach aimed at exploring the contribution of AI to improving decision-making. Conducted in collaboration with the CIUSSS-SLSJ, it builds on the KPI established by the CPI of the MSSS, as well as on real data from the SAP system covering the period 2018–2023. While this work fundamentally relies on the KPI framework established by the CPI of the MSSS, the detailed, specific operational metrics used to define anomalies are considered confidential internal documents; consequently, the research was guided by the strategic orientation and expertise provided by Committee members rather than a formally citable public document to structure the anomaly criteria. The project adopts a retrospective observational approach to assess the ability of AI models to objectively and reliably detect anomalies in procurement processes. The objective is twofold: first, to compare the performance of these models with traditional monitoring methods, and second, to explore their potential to generate early alerts enabling more proactive and transparent resource management. This thesis is based on two complementary scientific works. The first provides a proof of concept developed in a Python/pandas environment, illustrating the potential relevance of combining various anomaly detection and post-processing techniques. The second expands this approach by migrating the architecture to Apache Spark, in order to meet large-scale data processing needs and to prepare for future integration into a forthcoming centralized ERP system. Beyond its academic contribution, this research aims to provide Québec's healthcare network managers with practical tools to anticipate critical anomalies and strengthen system resilience. It is thus part of the broader movement of digital transformation, at the crossroads of scientific rigor and operational usefulness. Although this thesis is longer than typical, its scope is justified by the inclusion of two scientific articles and a detailed Exploratory Data Analysis (EDA) appendix, both of which were necessary to presenting the research with clarity and rigor.

INTRODUCTION

0.1 RESEARCH CONTEXT

The Health and Social Services System / Système de services de santé et sociaux (SSSS) is a vast network of organizations working together to deliver healthcare and social services to the population of Québec [4]. The SSSS in Québec is a public system, where the State acts as the main insurer and administrator. It was established in 1971 following the adoption of the first Act respecting health services and social services / Loi sur les services de santé et les services sociaux (LSSS) [5] by the National Assembly of Québec / Assemblée nationale du Québec (ANQ) [6].

The system is made up of three levels of management: the MSSS, Santé Québec, and health and social service institutions. The regional health and social service agencies, which represented the regional tier, were abolished in 2015 after the implementation of the Act to Modify the Organization and Governance of the Health and Social Services Network / Loi modifiant l'organisation et la gouvernance du réseau de la santé et des services sociaux (LMRSSS) [7]. More recently, the Government of Québec passed the Act respecting the governance of the health and social services system / Loi sur la gouvernance du système de santé et de services sociaux (LGSSSS) [8], which introduced a new management tier (Santé Québec) between the MSSS and the institutions [9]. The addition of Santé Québec as the single operator of the network strengthens centralization and the ability to impose common standards [10].

Health and social service institutions include Integrated Health and Social Services Center / Centre intégré de santé et de services sociaux (CISSS) and Integrated University Health and Social Services Center / Centre intégré universitaire de santé et de services sociaux (CIUSSS) [11]. The CISSS are the result of the merger of public institutions within the same region and, where applicable, the regional health and social services agency [11]. The CIUSSS are

created on the same model but are located in regions that host a university offering a complete undergraduate medical program or that operate a center designated as a university institute in the social field [11]. The mission of the CISSS and CIUSSS is to provide integrated health and social services to the population [11]. These responsibilities are tailored to the needs of their populations and territorial realities, which illustrates the complexity of the SSSS megastructure [4].

Supply chain management represents a major challenge for health and social service institutions [12]. Indeed, outdated and archaic systems, as well as differences in software solutions across the RSSS, make it difficult to measure and standardize working methods [13].

The Framework for Responsible Procurement [12] of the MSSS seeks to mobilize stakeholders in the RSSS around responsible procurement. However, there is no one-size-fits-all solution for all institutions, which can lead to differences in work methods and processes [12]. The Policy on the Acquisition of Goods and/or Services and Contract Management / Politique sur l'acquisition de biens et/ou de services et la gestion des contrats (PABSGC) [14] establishes the specific regulations applicable to the SSSS context. Finally, the Act respecting contracting by public bodies / Loi sur les contrats des organismes publics (LCOP) [15] provides the legislative framework for procurement across all public bodies, including the RSSS.

Since the onset of COVID-19, supply chains in the healthcare sector have faced unprecedented pressures [16, 17, 18]. Disruptions such as these can lead to shortages of essential products, including Personal Protective Equipment (PPE), medicines, and vaccines [19, 20, 21]. This, in turn, may compromise the ability of health systems to meet patient needs effectively, while posing risks to the health and safety of healthcare personnel [22]. Moreover, changes in supply chains can also increase costs, which may affect accessibility and equity in healthcare [23, 24].

It is therefore important to implement robust strategies to manage risks associated with supply chain disruptions, particularly by strengthening resilience, diversifying sources, and improving transparency and traceability [25, 26]. This need for resilience and diversification is complemented by the adoption of anomaly detection methods [27, 28], which play a key role in real-time monitoring and data analysis [29, 30]. Such methods enable rapid intervention in the supply chain by identifying unusual variations, such as delivery delays or sudden increases in demand, thereby potentially enabling earlier identification of emerging issues, optimizing inventory management, and improving transparency by tracking real-time performance and highlighting areas requiring attention [31, 32]. More specifically, these approaches can enhance the fluidity and monitoring of processes [31].

In summary, anomaly detection can help reduce the costs associated with supply chain disruptions. However, based on the information available to the CIUSSS-SLSJ and the literature reviewed for this thesis, we did not find publicly available evidence that such anomaly detection methods have been deployed at large scale within the RSSS. Specifically, we performed targeted searches across academic and industry literature, government documentation, and publicly and non-publicly accessible organizational materials (including internal documentation and Santé Québec intranet/SharePoint resources) using keywords related to healthcare supply chains, anomaly detection, and real-time deployment. The implementation of these technologies represents a major challenge, as the administrative and legal processes involved are often complex [33].

In the context of procurement for public bodies such as the RSSS, it is essential that both requesters and suppliers comply with regulations and laws [33]. However, this framework sometimes complicates the management of calls for tenders and contractual commitments [15]. Detecting anomalies in these processes is therefore a major challenge [34, 35]. For instance, some departments occasionally place orders for goods and services not directly related to their

activities. Such orders represent anomalies and may lead to additional costs and procurement delays [33, 34, 36, 37]. In other cases, suppliers fail to comply with contract prices or the quantities requested in orders [37], which can generate additional costs [38]. Likewise, certain stakeholders do not always respect the processes specified by laws and regulations [15], and such process breakdowns can prove costly in terms of both time and money [38].

Furthermore, in the context of healthcare labor shortages, fulfilling complex commitments by suppliers presents challenges for this department [35]. Anomaly detection would therefore be an asset for reducing costs and improving management capacities [39, 40]. To some extent, procurement also encompasses inventory management [41], where irregularities may arise in product orders [42]. For example, during staff substitutions, orders may omit essential materials. Such shortages can sometimes disrupt organizational operations. Detecting anomalies in orders could therefore support managers in addressing these issues proactively [43].

Another important dimension of supply chain management is the detection of anomalies related to shortages, also known as *Back Order (BO)* [44]. Several studies have documented BO-related issues in healthcare supply chains, suggesting that this is a recurring challenge in many systems [44, 45, 46]. BO occur when orders cannot be completed due to supplier stockouts [44, 45]. This situation can cause delays in the delivery of essential products, thereby compromising patient care [47]. In this context, anomaly detection can potentially play an important role in supporting a steady supply flow and facilitating more effective responses to patient needs [39].

In conclusion, the RSSS, despite its complex structure and the continuous evolution of its governance frameworks, faces supply chain challenges. Disruptions caused by events such as the COVID-19 pandemic have highlighted supply chain vulnerabilities and the importance of resilience and efficiency in this field [48]. The adoption of anomaly detection technologies

appears to be a promising approach for anticipating and managing these challenges, by enabling more rapid and targeted interventions when irregularities arise, provided that data quality, organizational capacity, and integration constraints are adequately addressed.

0.1.1 ANOMALIES

An anomaly refers to a deviation or irregularity from what is considered normal, standard, or expected in a specific context [49]. It can manifest in various fields, such as biology, medicine, statistics, the environment, or within technical and social systems. Generally, an anomaly indicates a difference from a given set. In the field of data science, for example, an anomaly often means that a data point differs from others, suggesting either an error or an unusual event [50]. Detecting and analyzing anomalies can, in some cases, provide useful insights for understanding and addressing problems across different domains, potentially helping to reveal underlying issues or opportunities for improvement [51].

ANOMALIES IN HEALTHCARE PROCUREMENT

An anomaly in the context of healthcare procurement systems refers to any deviation or irregularity that disrupts the efficient and secure flow of medical supplies, medications, vaccines, or services necessary to ensure quality care [52, 53, 54]. This issue is of substantial importance, as such anomalies can negatively affect the quality of care delivered, the accessibility of treatments, and, in some cases, patient safety [55]. These anomalies may take various forms, including shortages of vital supplies, errors in inventory management, delivery delays, or product quality issues [56]. In a field where precision and reliability are essential, the rapid identification and correction of these anomalies are vital to maintaining the integrity

and efficiency of the procurement system [32], and by extension, to ensuring the continuity and quality of healthcare services [57].

0.1.2 ANOMALY DETECTION

Anomaly detection, also known as deviation detection or outlier detection, is a process used in data analysis and data science to identify observations that do not conform to an expected pattern or typical behavior [39]. These anomalies may indicate errors, fraud, defects, or rare but significant events [39].

Anomaly detection is important in various fields such as network monitoring [58], fraud prevention [59], healthcare (e.g., identifying anomalies in medical test results) [60], predictive maintenance of machines [61, 62], and environmental monitoring [63]. Techniques for anomaly detection range from simple statistical methods, such as standard deviations or interquartile ranges, to more advanced machine learning models like neural networks and random forests. These approaches can be categorized into three groups: supervised, semi-supervised, and unsupervised methods [39].

The goal of using such algorithms is to identify anomalies quickly and accurately, so that corrective action can be taken, either by fixing errors or by conducting further analysis to understand the cause of these deviations [64].

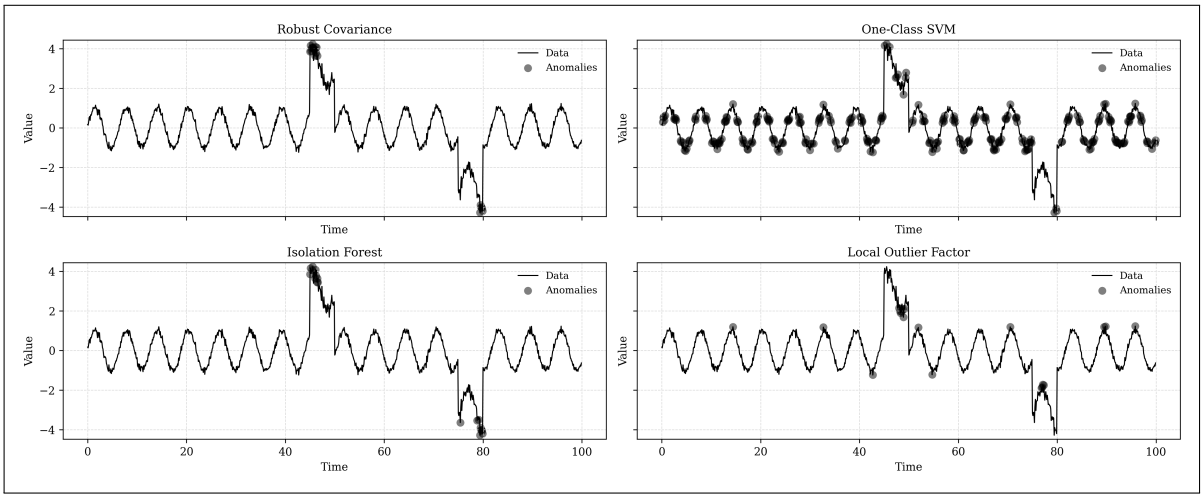
Due to the diversity, complexity, and ever-growing volume of data, anomaly detection is a constantly evolving field, regularly integrating new techniques and approaches to improve accuracy and efficiency [39]. This continuous evolution is driven by the need to adapt to increasingly varied and complex data types, as well as the growing scale of data, which places new demands on the performance and scalability of detection techniques [51].

For example, in cybersecurity, anomaly detection plays a key role in identifying suspicious behavior that may indicate a cyberattack or system intrusion [65]. Advanced algorithms are needed here to distinguish between normal activity and potentially malicious behavior hidden within large volumes of data. A similar analogy can be applied to procurement, where violations of laws or regulations in bidding processes or contractual commitments could be flagged, ultimately reducing costs by mitigating fraud risks.

Furthermore, with the advent of the *Internet of Things (IoT)* and *Industry 4.0*, anomaly detection has become essential for monitoring the condition of connected devices and industrial processes [61, 66]. In such contexts, anomalies may signal imminent failures or inefficiencies, allowing for preventive measures that avoid costly downtime or catastrophic breakdowns [61, 62]. This application is especially relevant to supply chain monitoring, where detecting delivery delays could prevent potentially life-threatening shortages for patients.

Anomaly detection is also critical in finance to identify suspicious transactions, such as money laundering or credit card fraud [59]. *Machine learning* [67] and *data mining* [68] techniques are particularly useful in this field for analyzing patterns in large transaction datasets and identifying behaviors that deviate from the norm [69, 70]. Similarly, these algorithms could potentially be applied to detect procurement fraud attempts.

Figure 0.1 illustrates some examples of machine learning algorithms applied to nonlinear data using the *Python Scikit-Learn* library [1]. These approaches are especially useful in identifying geospatial anomalies. For instance, in the transport of goods or people, it would be possible to detect deviations from established routes, thereby reducing the risk of delivery delays.



© Colin Bouchard

Figure 0.1 : Examples of various anomaly detection algorithms applied to nonlinear simulated data in the *Python Scikit-Learn* library [1].

ANOMALY DETECTION IN HEALTHCARE PROCUREMENT

According to McKinsey [71], in healthcare procurement, effective resource management is vital to ensure the fair and adequate distribution of medical services and health products. However, this process is complex, as it involves planning, acquisition, distribution, and management of medical supplies, equipment, drugs, and other essential goods, each step having its own critical level of importance. A high-performing supply chain must not only meet current patient and healthcare professional needs but also anticipate future demands, while remaining flexible enough to adapt to emergencies or sudden changes, such as those observed during pandemics [72].

A major challenge lies in balancing operational efficiency with maintaining sufficient inventory levels to prevent shortages, while minimizing costs and waste [73]. This requires a detailed understanding of consumption trends [73] and disease/clinical patterns [60], as well as predictive analysis to forecast future needs [68, 74]. In this context, anomaly detection becomes a

key component, enabling the rapid identification of irregularities in consumption, unexpected supplies, or sudden changes in demand patterns [32]. Health information systems that integrate anomaly detection tools provide real-time data to support informed and responsive decision-making [29, 75].

At the same time, compliance with regulatory requirements, such as the LCOP [15], is essential to ensure that products and services meet high standards of quality and safety, while preventing collusion and corruption. Supply chain management for medicines, medical devices, and equipment must guarantee their availability, safety, effectiveness, and regulatory compliance [75]. Anomalies stemming from non-compliance with such laws and regulations can be detected through a combination of simple and advanced algorithms.

The COVID-19 pandemic highlighted the importance of a resilient healthcare supply chain, underscoring the need for proactive planning, rapid responsiveness to changing demands, and close collaboration among key stakeholders [76]. Innovations such as blockchain for traceability [77], artificial intelligence for demand forecasting [74], and robotics for automating storage and distribution processes [78] have become increasingly important. These tools, when combined with anomaly detection mechanisms, contribute to making healthcare supply chains more robust, transparent, and efficient [32].

SELECTED ANOMALY DETECTION MODELS

As illustrated in Figure 0.2, each anomaly detection algorithm relies on a fundamentally distinct mathematical principle to define and identify outliers. The *Isolation Forest* (A) adopts an *isolation-based* strategy, where anomalies are detected through recursive random partitioning of the feature space [79]. At each node, a feature and split value are randomly selected, generating axis-aligned partitions that progressively isolate observations; anomalous points, being rare and distinct, require fewer splits and therefore exhibit shorter average path lengths across the ensemble. These path lengths are subsequently normalized to compute an anomaly score, with shorter normalized paths indicating a higher likelihood of anomaly [79]. In contrast, the *Robust Covariance* in *Elliptic Envelope* (B), based on *robust covariance estimation* [80], assumes that inliers follow an approximately *Gaussian* (elliptical) distribution and constructs an *ellipsoidal boundary* using the *mean vector* and *covariance matrix*; outliers are determined via a threshold on the *Mahalanobis distance*, with points lying sufficiently far from the center classified as anomalous [80]. Although the *One-Class SVM* (C) may appear visually similar, it is conceptually different: instead of relying on a parametric distributional assumption, it maps the data into a high-dimensional feature space using a kernel function and learns a maximum-margin boundary that separates the data from the origin, effectively estimating the support of the data distribution [81, 82]. This enables the detection of anomalies as points falling outside this learned support, making the method particularly suitable for non-linear and non-elliptical structures. Finally, the *Local Outlier Factor* (D) is a density-based approach that measures the local deviation of a point with respect to its *k-nearest neighbors* [83]; it computes a local reachability density and assigns an anomaly score based on the ratio between the average local density of the neighbors and that of the point itself, with significantly lower-density points identified as outliers [83].

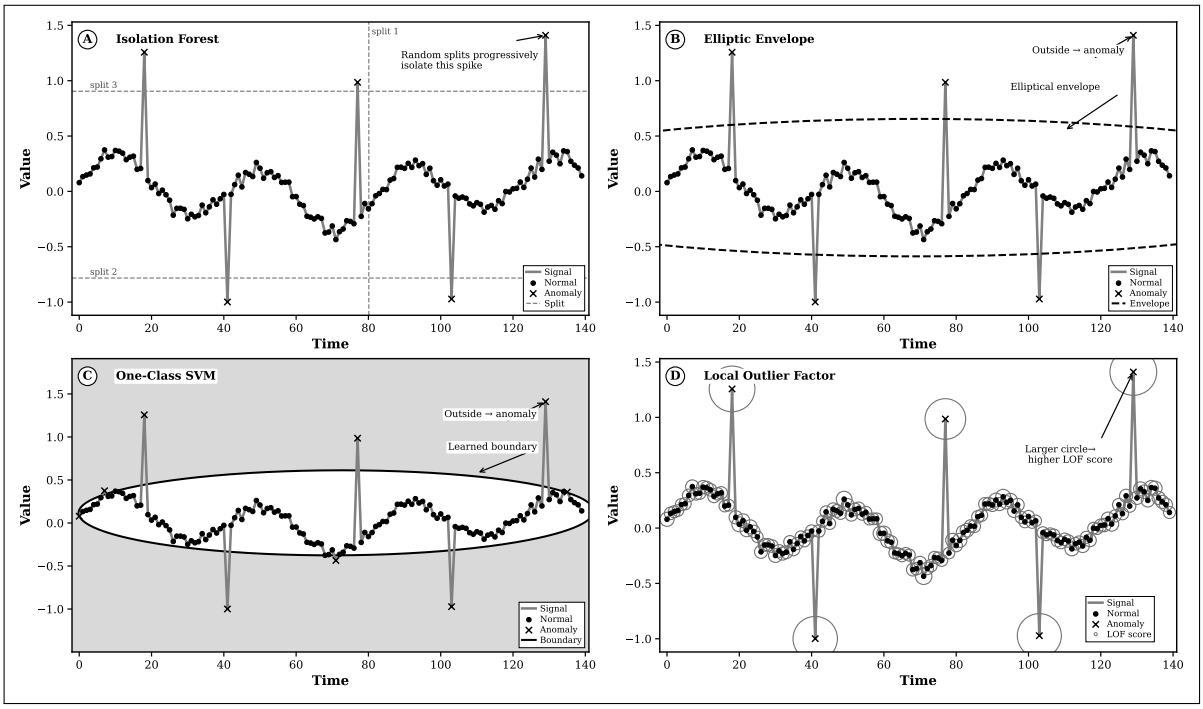


Figure 0.2 : Illustration of *Isolation Forest* (A), *Elliptic Envelope* (B), *One-Class SVM* (C), and *Local Outlier Factor* (D) for anomaly detection on a synthetic time-series dataset with injected spikes.

This side-by-side layout highlights how global assumptions (*Gaussian* versus *kernel-based support estimation*) and *local* or *isolation-based* criteria lead to markedly different geometric interpretations of what constitutes an outlier, thereby motivating the selection of these complementary algorithms.

0.1.3 ANOMALY CRITERIA OF THE STUDY

Having introduced anomaly detection broadly, the next section defines the specific anomaly types relevant to our healthcare procurement context. Healthcare procurement is exposed to a range of anomalies that, if not detected in time, compromise operational efficiency, the quality of services delivered to patients, and the overall performance of the system. The selected criteria cover the strategic, tactical, and operational dimensions of resource management. In this section, we review various anomaly criteria in healthcare procurement.

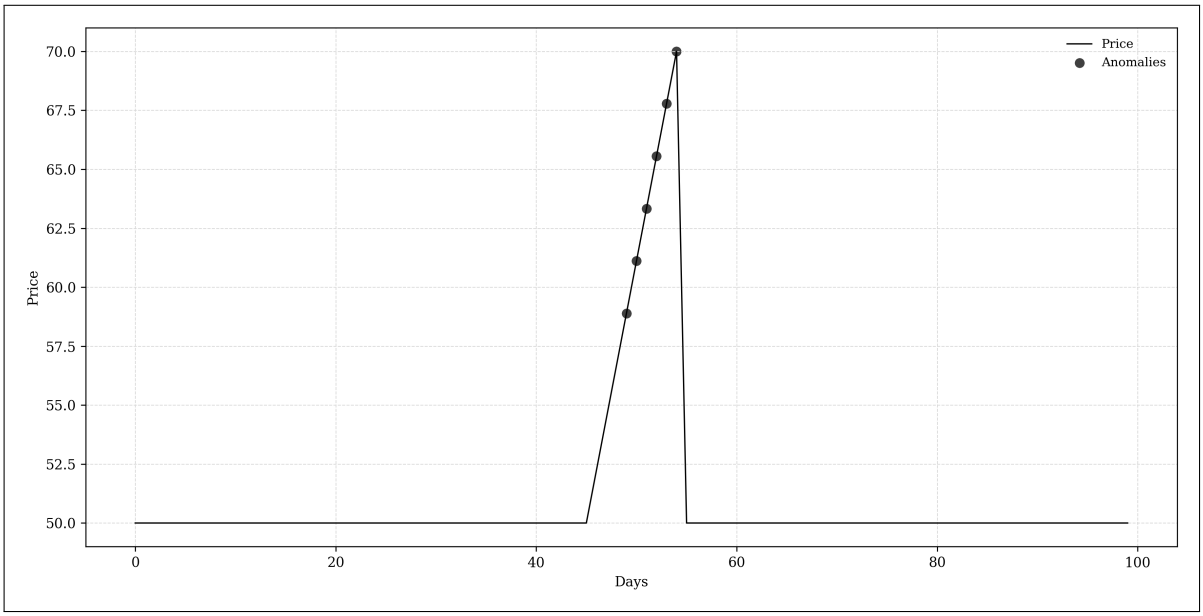
A one-way *Analysis of Variance*(*ANOVA*) is used as a descriptive test of whether the mean of a variable differs between two pre-labelled groups (in our case: anomalies and normal values). In practice, labels are assigned by a single domain expert (Administrative Process Specialist) following operational *CPI–KPI* rules and review criteria. The *ANOVA* is then applied using these expert-provided labels; when possible, we apply it to variables not directly used in the expert labelling criteria to limit circularity, and we interpret results cautiously otherwise. With two groups, one-way *ANOVA* is equivalent to a two-sample t-test under the same null hypothesis [84]. The resulting *F-statistic* is evaluated with between- and within-group degrees of freedom to obtain a p-value. We use a significance level of $\alpha = 0.05$ ($p \leq 0.05$); a significant result indicates a group-level mean difference (not significance of individual observations). We emphasize that *ANOVA* in this setting acts only as a proxy sanity check against a single-expert reference and is limited by its assumptions (e.g., independence, homoscedasticity, and approximately normal residuals) and by the quality and subjectivity of the initial labelling.

NON-COMPLIANCE WITH REGULATIONS, LAWS, AND CONTRACTUAL COMMITMENTS

Compliance with regulations and contractual commitments is a strategic foundation of public healthcare procurement. In Quebec, the legislative framework is primarily defined by the LCOP, which governs calls for tenders, direct contracts, and conditions for contractual amendments. The purpose of this law is to ensure transparency, fairness among suppliers, and responsible management of public funds.

The LCOP [15] is complemented by the PABSGC [14] and the Reference Framework for Responsible Procurement / Cadre de référence en approvisionnement responsable (CRAR) [12], which specify expected practices to guarantee efficiency, sustainability, and integrity in procurement processes. Finally, the Act respecting the Government Acquisitions Center / Loi sur le Centre d'acquisitions gouvernementales (LCAG) [85] centralizes and harmonizes purchasing processes to ensure better coherence in contract management across the healthcare and social services network.

Within this normative context, several anomaly criteria are particularly relevant. The first is **significant price variation**, as contracts specify exact prices that must be respected. Abnormal fluctuations may indicate overbilling, administrative errors, or lack of proper control. Figure 0.3 illustrates such a case, where a normally stable price shows a statistically significant variation detected by an algorithm. This criterion is strategic, as it protects both the financial integrity of the network and trust in public processes.



© Colin Bouchard

Figure 0.3 : Example of anomaly detection (Distance-based Outlier Detection) on a fictitious stable price (F-statistic = 735.80, p-value < 0.001).

A second criterion is **non-compliance with contractual deadlines**. While this is also related to shortage prevention (Section 1.6.2), it is primarily a contractual issue. When a supplier fails to meet deadlines specified under the LCOP [15], PABSGC [14], or CRAR [12], it violates its legal obligations, which can directly impact continuity of care. This criterion is therefore both strategic (legal compliance) and operational (timely delivery).

Finally, **unusual orders** represent another form of contractual anomaly. For example, when an administrative department, such as Human Resources, places an order for medications normally reserved for clinical or pharmaceutical services, this constitutes a deviation from allocation and responsibility rules. Such anomalies, linked to internal governance, are considered tactical criteria: they enable quick detection of organizational irregularities or fraud attempts and support the correction of order validation processes.

SHORTAGE PREVENTION

Shortage prevention is primarily an operational concern, as it relates to the immediate availability of medical products. Two main criteria are directly associated with this dimension. First, **unusual delivery delays** are an early warning sign of stockouts, as they may result in the absence of essential drugs or medical devices in facilities. Figure 0.4 illustrates this type of anomaly, showing significant variations in observed delivery times.

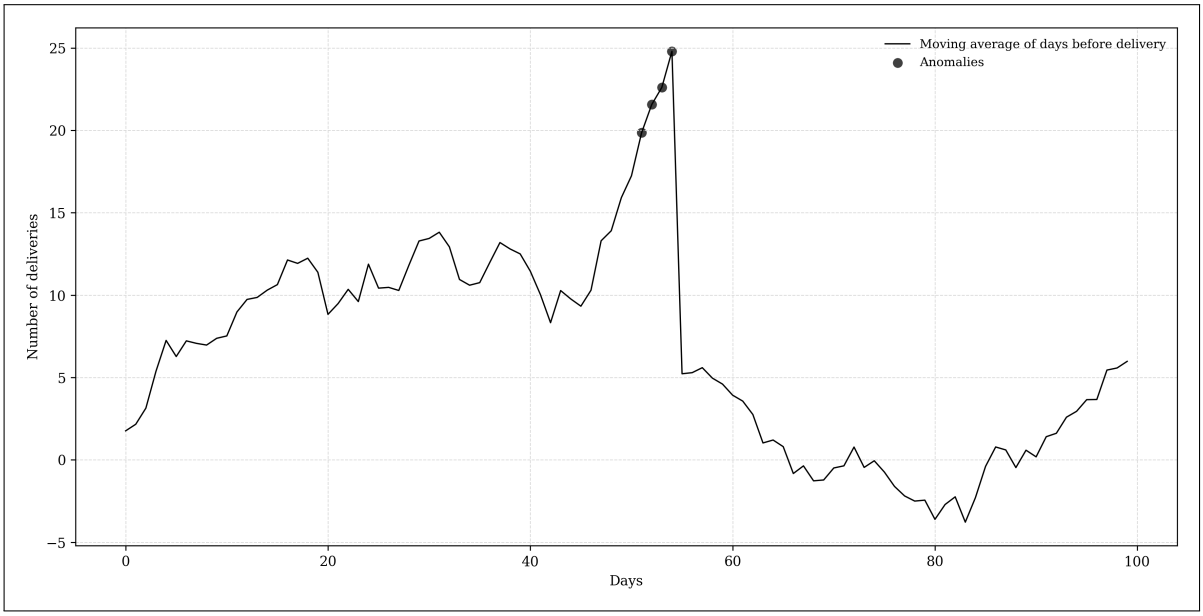


Figure 0.4 : Example of anomaly detection (*Z-score* [2]) on fictitious variable delivery days (F-statistic = 35.31, p-value < 0.001).

The second criterion is **unusual variability in received quantities**. When delivery volumes deviate significantly from expectations, the risk of understocking or overstocking increases. This operational criterion is important, as it helps maintain stable inventories while avoiding costs related to emergencies or waste.

Abnormal demand fluctuations are also included, as they indicate instability that may stem from sudden epidemics, organizational changes, or irregular orders. From a tactical perspective, this criterion enables managers to react quickly by adjusting procurement. Finally, **increased unusual returns**, as shown in Figure 0.5, constitute a clear warning of recurring non-compliance. This reinforces prevention by allowing interventions on quality issues before shortages occur.

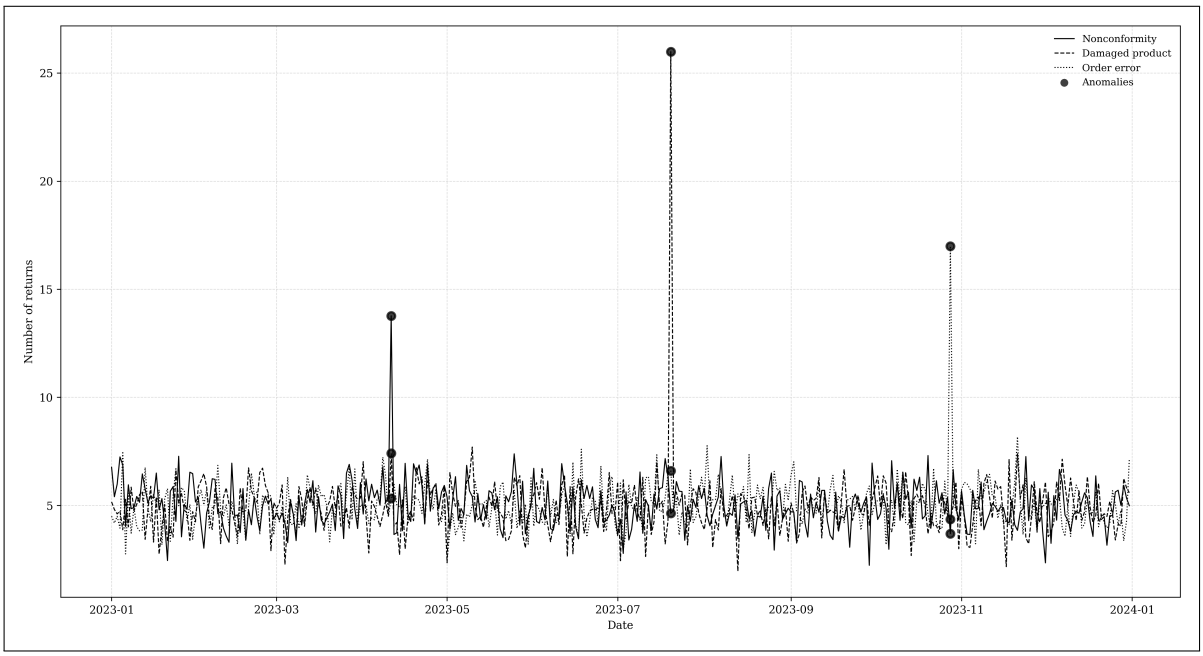


Figure 0.5 : Example of anomaly detection (*Grubbs Test* [3]) on fictitious return types (*F-statistic* = 562.61, p-value < 0.001).

SIGNIFICANT DEVIATIONS FROM STANDARDS

Significant deviations from established standards constitute anomalies detectable through statistical approaches and are both tactical and operational in nature. They relate to quality, compliance, and performance in procurement processes.

A central criterion is the **number of unusual returns**. When a delivered product is frequently returned, it signals recurring non-compliance, whether due to manufacturing defects, specification issues, or failure to meet expected standards. Figure 0.5 illustrates such a case, where the number of returns exceeds normal thresholds, revealing a significant anomaly. This criterion may be strategically relevant, as it can influence both the safety and the quality of care.

Sudden deviations in KPIs such as average delivery days, percentage of perfect orders, or the number of active contracts also constitute important anomalies. Since these indicators are used by the MSSS to evaluate network performance, any unexpected variation signals systemic dysfunction. This criterion is therefore strategic, as it provides decision-makers with a reliable and synthetic view of performance.

Finally, **significant variations in delivery times and prices**, although already mentioned in relation to contractual commitments and shortage prevention, are here confirmed as measurable deviations from standards. Their inclusion in multiple categories demonstrates their transversal relevance for overall procurement evaluation.

SELECTED CRITERIA

As shown in Table 1, a set of anomaly criteria has been selected for this study. These are used to implement the detection system.

In the context of this thesis, these anomaly criteria were selected because they align with the thesis objective of detecting early signals of risk and non-compliance in procurement and supply operations using data that organizations already track routinely.

Table 1 : Selected anomaly criteria

Anomaly Criterion	**Definition**
Significant price variation [86]	Significant change in price that may indicate an attempt at fraud.
Significant delivery time variation [87]	Significant change in delivery times (lead times) that may indicate non-compliance with contractual deadlines, with potential shortage risks.
Unusual number of returns (possible product non-compliance) [88]	Significant change in the number of returns that may indicate recurring non-conformity.
Abnormal demand fluctuation [89]	Significant change in demand for a given product that may signal abnormal or questionable overuse.
Unusual orders [90]	Orders outside the normal scope of a department (e.g., Human Resources ordering medications).
Deviation or sudden change in KPIs [91]	Significant changes in KPI dashboard indicators (e.g., average delivery days, deliveries per item, deliveries per order, percentage of perfect orders, number of orders, number of contracts, delays of $\pm$30 days, etc.) .
Unusual delivery delays [92]	Unusual delays that may indicate early signs of shortages .
Unusual variability in received quantities [93]	Variability in received quantities (in-full / service level) that may indicate early signs of shortages.

0.1.4 EVALUATION OF ANOMALY DETECTION MODELS

Once anomalies are detected, we need ways to evaluate detection performance. Therefore, we review common evaluation metrics.

The evaluation of anomaly detection models is essential to ensure their effectiveness and reliability in identifying atypical behaviors, data, or events. These models, widely used in anomaly detection [94], require specific evaluation metrics to assess their performance [95, 96]. Among the most common indicators are the false positive rate (the number of normal instances incorrectly classified as anomalies), the false negative rate (the number of anomalies missed), *precision*, *recall*, and the *F1-score*, which is the harmonic mean of *precision* and *recall* [95, 96].

The use of well-designed test datasets, including both known anomalies and normal behaviors, is essential to realistically evaluate a model's ability to generalize from unseen data [97]. In addition, techniques such as *ROC* and *PR* analyses are commonly used when a detector provides a continuous decision value. In this thesis, *ROC* and *PR* curves are computed by sweeping a decision threshold over the continuous anomaly scores $s(x)$ returned by each detector (after orienting scores so that larger values indicate more anomalous behavior). Threshold-dependent metrics (*accuracy, precision, recall, specificity, and F1-score*) are computed at a chosen operating point by thresholding $s(x)$ into binary anomaly flags $\hat{y} \in \{0,1\}$ (consistent with the output of `predict` after mapping $\{-1,+1\} \to \{1,0\}$).

DETERMINATION OF TRUE POSITIVES, FALSE POSITIVES, TRUE NEGATIVES, AND FALSE NEGATIVES

In the context of this study, an anomaly is considered a *true positive* (TP) when the model's binary anomaly decision matches the expert-provided label (treated as the study's reference *expert reference labels*), i.e., the model flags an observation as anomalous and the expert label is 1. Conversely, a *false positive* (FP) occurs when the model flags an observation as anomalous but the expert label is 0. The corresponding *true negatives* (TN) and *false negatives* (FN) are defined analogously. We note, however, that these metrics reflect agreement with an expert risk model rather than an objective, error-free expert reference labels; as such, they may inherit subjective or institutional biases present in the labeling process.

Here, *Positive* is synonymous with anomaly, and *Negative* with normal.

- The *True Positive Rate* (TPR), also called sensitivity or recall, corresponds to the proportion of anomalies correctly detected among all present anomalies.
- The *False Positive Rate* (FPR) indicates the proportion of normal data wrongly classified as anomalies.
- The *True Negative Rate* (TNR), or specificity, measures the proportion of normal data correctly identified as normal.
- The *False Negative Rate* (FNR) reflects the proportion of anomalies not detected and incorrectly classified as normal.

EVALUATION METRICS

Precision, *sensitivity*, *specificity*, *F1-score*, *AUC-ROC* and *AUC-PR* are key measures in statistics and machine learning to evaluate model performance [95, 96, 98].

Precision Precision measures the proportion of correct predictions among all positive predictions, while sensitivity (also called recall) evaluates the model's ability to correctly identify true positives [95, 96].

Sensitivity (recall) Sensitivity measures the proportion of actual positive examples that the model detects. A higher sensitivity means the model can effectively capture true positives, which is particularly important in fields where minimizing false negatives is critical, such as anomaly detection [95, 96].

Specificity Specificity measures the proportion of actual negative examples that the model correctly identifies. A high specificity means the model minimizes false positives, i.e., it does not wrongly classify normal examples as anomalies [95, 96].

F1-score In artificial intelligence, the *F1-score* (also called *F-measure* or *F-beta*) combines both *precision* and *recall* to evaluate the performance of a binary classification model. It is a measure of overall classification quality, taking into account *false positives* (FP), *false negatives* (FN), and *true positives* (TP). The *F1-score* provides a balanced measure of performance, considering both the ability to correctly classify positive examples (*precision*) and to detect all real positives (*recall*). It is especially useful when the dataset has imbalanced classes, as in anomaly detection [95, 96].

AUC-ROC The *Area Under the Receiver Operating Characteristic Curve (AUC-ROC)* evaluates a detector's ability to rank anomalies above normal instances by measuring the area under the *Receiver Operating Characteristic (ROC)* curve obtained by sweeping a decision threshold over continuous anomaly scores (e.g., decision values) rather than hard labels. It summarizes the trade-off between true positive rate (sensitivity) and false positive rate across thresholds. An *Area Under the Curve (AUC)* close to 1 indicates strong separability, while a value close to 0.5 reflects performance equivalent to random guessing [99, 100, 101].

AUC-PR The *Area Under the Precision-Recall Curve (AUC-PR)* measures the (*trapezoidal*) area under the precision–recall curve obtained by sweeping a decision threshold over continuous anomaly scores. In highly imbalanced settings such as anomaly detection, *AUC-PR* is often more informative than *AUC-ROC* because it focuses on performance on the rare positive class (anomalies). *Precision* is the proportion of flagged instances that are true anomalies, while recall reflects the proportion of anomalies that are detected. A high *AUC-PR* therefore indicates that the detector maintains high precision while recovering anomalies across thresholds [102, 103].

METRICS CALCULATION

Table 2 summarizes the evaluation measures used to assess and compare anomaly detection methods and provides the corresponding formulas expressed with the standard confusion-matrix terms: *true positives* (TP), *true negatives* (TN), *false positives* (FP), and *false negatives* (FN). The table first reports *Accuracy*, which captures overall correctness by counting both correctly detected anomalies and correctly identified normal instances, and *Precision*, which evaluates how trustworthy anomaly predictions are by quantifying the proportion of predicted anomalies that are truly anomalous [95, 96]. It then includes *Sensitivity* (*Recall*), which measures the ability to detect actual anomalies (i.e., the fraction of anomalies recovered), and *Specificity*, which measures the ability to correctly recognize normal behavior and therefore penalizes false alarms [95, 96]. To balance the trade-off between missing anomalies and generating *false positives*, the *F1-score* is provided as the harmonic mean of *Precision* and *Recall*, yielding a single value that is sensitive to both types of error [95, 96]. Beyond single-threshold measures, the table also reports two threshold-sweeping, ranking-based criteria computed from a continuous anomaly score $s(x)$ (higher values indicating more anomalous observations): *AUC-ROC*, which approximates the area under the *ROC* curve using trapezoidal integration across successive points defined by the *true positive rate* TPR_i and *false positive rate* FPR_i [99, 100], and *AUC-PR*, which similarly integrates the *Precision–Recall* curve over varying thresholds and is often informative when anomalies are rare [102, 103].

In this thesis, *ROC* and *PR* curves are computed by sweeping over continuous anomaly scores; confusion-matrix metrics are computed after thresholding scores into binary flags. For curve-based metrics, we use an anomaly score $s(x)$ such that larger values indicate greater outlierness (e.g., $s(x) = -f(x)$ when a library returns an inlier score $f(x)$).

Table 2 : Evaluation metrics measures across anomaly detection methods

Measure	Formula
Accuracy [95, 96]	$\text{Accuracy} = \dfrac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$
Precision [95, 96]	$\text{Precision} = \dfrac{\text{TP}}{\text{TP}+\text{FP}}$
Sensitivity (Recall) [95, 96]	$\text{Recall} = \dfrac{\text{TP}}{\text{TP}+\text{FN}}$
Specificity [95, 96]	$\text{Specificity} = \dfrac{\text{TN}}{\text{TN}+\text{FP}}$
F1-Score [95, 96]	$\text{F1} = \dfrac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$
AUC-ROC [99, 100]	$\text{AUC-ROC} = \sum_{i=1}^{m-1} \left(\text{FPR}_{i+1} - \text{FPR}_i\right) \dfrac{\text{TPR}_{i+1}+\text{TPR}_i}{2},$ $\text{TPR}_i = \dfrac{\text{TP}_i}{\text{TP}_i+\text{FN}_i}, \qquad \text{FPR}_i = \dfrac{\text{FP}_i}{\text{FP}_i+\text{TN}_i}.$ *Computed by sweeping thresholds on the continuous anomaly score $s(x)$ (higher = more anomalous).*
AUC-PR [102, 103]	$\text{AUC-PR} = \sum_{i=1}^{m-1} \left(\text{Recall}_{i+1} - \text{Recall}_i\right) \dfrac{\text{Precision}_{i+1}+\text{Precision}_i}{2}.$ *Computed by sweeping thresholds on the continuous anomaly score $s(x)$ (higher = more anomalous).*

AUC CONVENTION

We define the positive class as expert-labeled anomalies ($y = 1$) and normal samples as $y = 0$. Both *AUC-ROC* and *AUC-PR* are computed from the *same* continuous anomaly score $s(x)$, where larger values indicate more anomalous samples. For detectors that return an inlier score (e.g., `decision_function` in scikit-learn), we use $s(x) = -f(x)$ to keep a consistent direction across models. *AUC-PR* is computed as the *trapezoidal* area under the precision–recall curve obtained by sweeping the decision threshold on $s(x)$. Given anomaly prevalence π, the no-skill baseline for *precision–recall* is π; therefore *AUC-PR* should be interpreted relative to π in highly imbalanced settings. We verified that the same $(y, s(x))$ convention is used for both *AUC-ROC* and *AUC-PR* in all experiments and result tables.

0.1.5 HYPERPARAMETERS

The management of hyperparameters is an important aspect of data modeling, particularly in machine learning and artificial intelligence. Hyperparameters are configuration parameters set before training begins. Unlike model parameters, which are learned during training, hyperparameters must be defined in advance [104]. The performance and efficiency of machine learning models largely depend on an appropriate selection of these hyperparameters [105].

Hyperparameters directly influence both the structure of machine learning models (e.g., the number of layers in a neural network) and the learning process itself (e.g., the *learning rate* in *gradient-descent algorithms*). A well-chosen configuration can improve a model's ability to learn complex patterns and generalize to unseen data. *Hyperparameter tuning* remains a major challenge, especially for complex models such as deep neural networks, where the number

of hyperparameters can be very large, and the search for an optimal configuration is often time-consuming and computationally expensive.

Recent advances such as *automated hyperparameter tuning* (AutoML) [106] and *multi-objective optimization* approaches [107] aim to simplify and streamline the tuning process. In this study, *Bayesian optimization* is retained because it explores the hyperparameter space more efficiently than exhaustive or random search and typically converges faster toward near-optimal configurations [108].

HYPERPARAMETERS FOR THE ANOMALY DETECTION ALGORITHMS USED IN THIS THESIS

Table 3 summarizes the principal hyperparameters that govern the behavior of four anomaly detection methods. For *Robust Covariance* in *Elliptic Envelope*, the *support fraction* determines the proportion of observations treated as inliers when estimating a *covariance matrix* that is resistant to outliers. For *One-Class SVM*, $\nu \in (0, 1]$ controls model sparsity and contamination by specifying an upper bound on the expected anomaly rate and a lower bound on the fraction of support vectors, while the *kernel* selects the feature-space mapping used to separate normal data from anomalies; for the RBF kernel, γ sets the influence radius of training examples, thereby controlling decision boundary flexibility. For *Isolation Forest*, the *number of trees* defines the ensemble size and thus the stability of the anomaly score, and the *sample size* specifies how many points are subsampled to build each tree, affecting both computational cost and the granularity of isolation. Finally, for *Local Outlier Factor* (LOF), the *number of neighbors* fixes the locality scale used to estimate density around each observation, and the chosen *distance metric* (e.g., *Euclidean, Manhattan, Minkowski, cosine*)

defines how proximity is computed, which can substantially influence local density estimates and the resulting outlier scores.

Table 3 : Key hyperparameters for anomaly detection algorithms of this thesis

Method	Hyperparameter	Definition
Robust Covariance (RC) [80] in Elliptic Envelope (EE) [109, 110]	*Support fraction*	Proportion of observations retained as inliers when estimating the *robust covariance*.
One-Class SVM (OC-SVM) [81]	ν (nu)	Parameter in $(0, 1]$ that sets an upper bound on the fraction of anomalies and a lower bound on the fraction of *support vectors*.
	Kernel	Kernel function mapping data into a higher-dimensional feature space.
	γ (*RBF kernel*)	*Coefficient* controlling the influence radius of a training example in the feature space.
Isolation Forest (IF) [79]	*Number of trees*	Total number of *isolation trees* in the ensemble.
	Sample size	Number of points randomly sampled to build each tree.
Local Outlier Factor (LOF) [83]	*Number of neighbors*	Number of neighbors used to estimate *local density*.
	Distance metric	Proximity function between observations (e.g., *Euclidean, Manhattan, Minkowski, cosine*).

0.1.6 POST-PROCESSING

Most anomaly detection systems produce a raw model output (anomaly flag) or alert stream derived from a model (for example, via residuals, likelihoods, or deviation measures) rather than immediately yielding final labelled events [39, 111]. The post-processing stage refines the raw model outputs (anomaly flags). It applies *smoothing*, *filtering*, *temporal* or *spatial aggregation*, *thresholding*, and *business-logic rules*.

These operations transform raw outputs into actionable alarms. As noted by G. P. Spathoulas and S. K. Katsikas in the intrusion detection domain, the aim of post-processing research is to improve alert-set quality by *false positives* reduction, *alerts' correlation* and visualisation [112]. This underscores that post-processing is a distinct and important component of the detection pipeline.

A concrete example from a non-supply-chain domain can illustrate how post-processing adds value. In the work by Nag et al., the module for *Post-Processing Temporal Action Detection* introduces a refinement stage that adjusts start and end times of detected actions by modelling the boundary distributions and applying a *Taylor-series* approximation to correct for quantisation error after down-sampling [113]. Despite the primary detection model remaining unchanged, this post-processing step yielded measurable improvements in *temporal precision of action alerts*. In a different domain (hyperspectral anomaly detection), Wu et al. demonstrate how the application of a morphological operator after a baseline detector helps suppress isolated high-score pixels (false alarms), thereby improving overall anomaly accuracy [114]. In the context of operational anomaly detection pipelines, post-filtering/post-processing is widely used as a downstream refinement stage, well established in signal processing and detection systems (e.g., microphone-array post-filters [115] and imaging post-processing [116, 117]) to suppress noise and improve decision usefulness, while modern anomaly detection additionally benefits from adaptive, online thresholding under distribution shift [118] and, where appropriate, robust covariance/scatter estimation to mitigate contamination effects [119].

Translating these ideas into the supply-chain context, one may think of the raw models output (anomaly flags) as stemming from sources such as demand-forecast residuals, supplier-performance metrics, logistics-delay indicators or sensor readings in manufacturing. Post-processing could apply a *temporal smoothing* of the anomaly score time series to suppress

transient spikes caused by one-off events or measurement noise. It might also merge contiguous anomaly-flagged periods into single supply-chain incident events (e.g., a sustained supplier under-performance rather than a momentary blip). *Spatial or network-wide filters* could combine anomalies across multiple nodes (for example, cascading delays through a supplier network) and apply *domain-specific business-logic filters*, such as excluding known maintenance windows or expected seasonal shifts. The *post-processing layer* is configured to trade off earlier detection (lower latency) against reducing false alarms. This trade-off is particularly relevant in supply-chain settings, where *false positives* can lead to unnecessary operational follow-ups.

Further, because the post-processing stage often encodes domain heuristics and business rules, it tends to influence interpretability and trust: analysts and supply-chain managers must be able to interpret how the raw outputs were transformed into an alarm, what filters were applied, and why a given event was flagged. In many real-world deployments the design of the post-processing layer has as much impact on operational performance (false alarm rate, event quality, latency) as the core detection model itself. In summary, while a large body of research emphasizes the detection model (*feature extraction*, scoring algorithm, neural architecture), the *post-processing layer*, including *smoothing, filtering, thresholding, merging, correlation and domain logic*, is a critical and often under-emphasised component of end-to-end anomaly detection systems [112, 120].

In the context of this thesis, we implemented contextualized, business-oriented post-processing filters. These filters, such as *Excess* (e.g., a sudden price spike may be the most meaningful anomaly for a domain expert and is therefore prioritized when the signal reflects an upward deviation), *Lack* (e.g., an unexpected drop in received quantities/service level can flag shortage risk and is prioritized when the signal reflects a downward deviation), and signal reversal (e.g., when the model's anomaly polarity is misaligned with expert labels, reversing the signal can

restore agreement with domain expectations), are contextualized because they explicitly align anomaly direction with procurement meaning and actionable risk. In this setting, *Excess* and *Lack* are not just technical filters, but semantic labels that interpret the anomaly direction as a concrete change in a specific anomaly criteria or feature (too high vs. too low), thereby giving the signal clear business meaning.

0.2 THESIS OBJECTIVE

In line with the work of Painuly et al. [27], this thesis aims to assess whether and under which conditions artificial intelligence can improve decision-making within healthcare supply chains. More specifically, the research framework seeks to develop and train AI models designed to detect anomalies in procurement processes, focusing on issues such as delivery delays, unexpected increases in demand, order irregularities, and back orders, using techniques similar to those employed by Bauder et al. [121], Westerski et al. [122], Patel et al. [123] and Usman et al. [94].

The project also aims to enhance the models through hyperparameterization, as demonstrated by Feurer and Hutter [124]. In addition, the study intends to address administrative and legal issues associated with the implementation of these technologies, by developing strategies to manage tenders, contractual commitments, and regulations using approaches described by Lau et al. [125].

This approach is consistent with the latest advances in artificial intelligence for supply chains in the context of *Industry 4.0*, as highlighted by Jagadeesan et al. [31]. The thesis ultimately aims to lay the groundwork for a real-time alerting system for abnormal situations, thereby supporting rapid interventions to resolve procurement issues. All experiments use real data from the SAP system of CIUSSS-SLSJ, with prior authorization from the institution's CER,

and rely on advanced analytical techniques and machine learning algorithms. Because all experiments are based on SAP data from CIUSSS-SLSJ, this thesis should be read as a single-institution case study, and its quantitative results should not be generalized to other contexts without further empirical validation.

This thesis investigates whether an AI-driven anomaly detection pipeline, combining pre-processing of SAP healthcare procurement data, *Bayesian hyperparameter optimization* of unsupervised models, and domain-specific post-processing filters, can improve the accuracy and operational relevance of anomaly detection in the CIUSSS-SLSJ context, compared with traditional manual handling of the data. It also explores a first migration of this logic to *Apache Spark*, not to demonstrate full large-scale scalability, but to examine whether comparable detection behaviour can be reproduced in a distributed environment and to identify the technical and organizational constraints that currently limit broader deployment in Québec's public health network. These aims are examined through the following research questions:

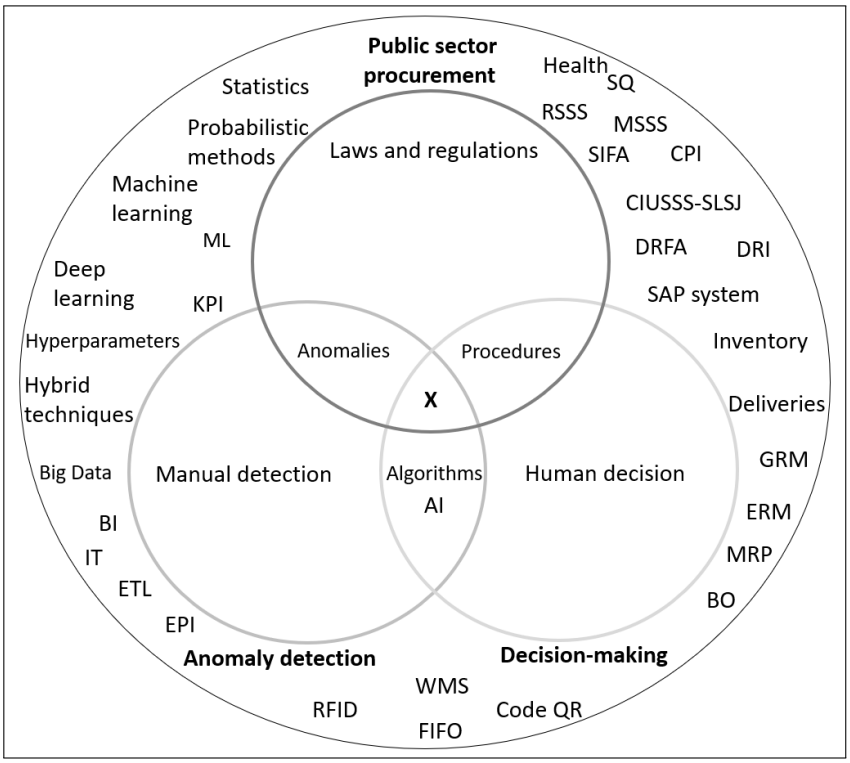
RQ1. How effective is an unsupervised AI-based anomaly detection pipeline, enriched with *Bayesian hyperparameter optimization* and *tailored post-processing filters*, at identifying operationally relevant anomalies in healthcare procurement data?

RQ2. To what extent do domain-specific *post-processing filters* improve precision and reduce *false positives* compared to relying on raw model outputs (anomaly flags) alone?

RQ3. Can the *Pandas*-based anomaly detection pipeline be reproduced in *Apache Spark* with comparable results, and what qualitative insights does this migration offer about potential scalability and performance limitations on the current infrastructure?

0.3 CONTRIBUTION OF THE THESIS

As illustrated in Figure 0.6, this thesis lies at the intersection of three important themes: public sector procurement, anomaly detection, and decision-making. This study explores the points of convergence between these domains, thereby revealing unique insights into their interdependence.



© Colin Bouchard

Figure 0.6 : Intersection between the thesis themes

At the core of public sector procurement are laws and regulations such as the LCOP (C-65.1) [15]. When these are not respected, anomalies emerge, which directly link procurement to the theme of anomaly detection. Public procurement also shares procedural elements that are closely tied to decision-making. In parallel, the field of anomaly detection encompasses topics such as manual anomaly detection, while also sharing concepts such as the use of

algorithms and artificial intelligence with decision-making. Within decision-making itself, we find both human decision-making and procedures shared with public procurement, as well as algorithmic and AI-based approaches that overlap with anomaly detection.

This study builds on a wide range of related topics to enrich its analysis. Among them: the MSSS, the CPI, the CIUSSS, as well as systems such as the SAP system. Concepts such as inventory management, shortages, BO, as well as ERP and Global Resource Management (GRM) are also discussed.

In addition, technical aspects such as Hyperparameter, KPIs, statistics, probabilistic models, machine learning, deep learning, and hybrid techniques are explored for their relevance to this research. Finally, at the intersection of the three themes, the *X* in Figure 0.6 symbolizes the main contribution of this thesis:

- A reproducible end-to-end procurement anomaly detection pipeline is introduced, combining preprocessing, unsupervised detectors, Bayesian hyperparameter optimization, and domain-informed post-filtering to prioritize actionable alerts.
- An empirical evaluation compares Isolation Forest (IF), Local Outlier Factor (LOF), One-Class SVM (OC-SVM), and Robust Covariance (RC) in Elliptic Envelope (EE) across multiple procurement anomaly criteria using real-world SAP data.
- A domain-tailored filtering suite is designed and shown to reduce false positives while improving the operational actionability of the generated alerts.
- The approach is re-implemented in *Apache Spark*, and deployment feasibility is assessed through scalability analysis and an explicit discussion of distributed processing constraints and trade-offs.

- Limitations, including label uncertainty and institutional specificity, are analyzed, and concrete avenues to improve external validity and generalizability are articulated.

In summary, this thesis offers an in-depth exploration of how these themes intersect, shaping the efficiency and effectiveness of processes in the public sector. It aims to contribute to a better understanding of the complex dynamics at play in public sector procurement, anomaly detection, and decision-making, by providing a fresh and multidimensional perspective.

0.4 MATERIALS AND METHODS

0.4.1 MATERIALS

The project takes place entirely within the secure network environment of CIUSSS-SLSJ. The computer used for testing is performant (Windows 10 Enterprise, 12th Gen Intel® Core™ i7-1165G7 4.70 GHz, 32 GB RAM, NVIDIA RTX 3080ti, 1 TB SSD), ensuring efficient and fast data processing.

The data is stored and protected on a local Microsoft SQL Server (MSSQL) 2019. This server is equipped with critical components such as Microsoft SQL SQL Server Integration Services (SSIS), SQL Server Analysis Services (SSAS), and SQL Server Data Quality (SSDQ). These tools allow Extract, Transform, Load / Extraire, transformer, charger (ETL) operations into the Data Warehouse (DW), while ensuring high data quality.

Preprocessing and data analysis is performed using Python, with a wide range of specialized libraries: `Numpy`, `Pandas`, `Scikit-Learn`, `pySpark`, `XGBoost`, `Statsmodels`, `SciPy`, `SymPy`, `CleanFill`, `Outlier-Utils`, `PyOD`, `ADTK`, `FBProphet`, `Matplotlib`, and `Seaborn`. This multidimensional approach ensures a comprehensive and robust analysis.

Docker and *Apache Spark* were used for Chapter 3 to ensure a reproducible computing environment and to support distributed data processing. Multiple compute nodes were containerized in CIUSSS-SLSJ cloud environment to replicate a distributed cluster setup for the experiments.

Additionally, PowerBI Enterprise is used to develop a sophisticated Business Intelligence (BI) data model, enabling multiple visualizations and analyses. This facilitate the understanding of trends, anomaly detection, and evidence-based decision-making.

0.4.2 METHODS

The methodology begins with a preliminary stage aimed at acquiring the necessary knowledge for the project. This is followed by a seven-phase procedure:

1. Define anomaly criteria.
2. Collect data.
3. Preprocess data.
4. Perform testing with classical statistical and machine-learning–based anomaly detection algorithms (*Robust Covariance* in *Elliptic Envelope*, *Local Outlier Factor*, *Isolation Forest*, *One-Class SVM*). *Deep learning* and *hybrid techniques* are identified as promising future directions but are not implemented in this thesis.
5. Optimize models using *hyperparameter tuning* and *apply post-processing filter*.
6. Evaluate results.
7. Conduct discussion and analysis of findings.

The experimental procedure involves the collection, preprocessing, and analysis of CIUSSS-SLSJ procurement data for the specified period, as shown in Figure 0.7. Data is extracted from the SAP system, preprocessed to ensure quality and anonymity, then submitted to artificial intelligence algorithms to detect anomalies.

Figure 0.7 illustrates the overall experimental procedure designed for anomaly detection. The process begins with the clear definition of detection criteria, identifying what types of irregularities, such as unusual demand, delivery delays, or price fluctuations, should be considered anomalies. Once these criteria are established, relevant data are collected from various sources to provide a solid foundation for analysis. An exploratory study then follows, combining descriptive statistics, visualization, and a series of statistical tests to better understand data structure, variability, and potential inconsistencies. The next step involves data preprocessing, where missing values are handled, variables are normalized or categorized, and overall data quality is ensured to make the dataset ready for modeling.

Subsequently, several analytical methods are applied: classical statistical approaches and unsupervised machine-learning algorithms are compared to identify the most effective strategy for anomaly detection in CIUSSS-SLSJ data. Deep-learning and hybrid models are discussed as potential extensions but remain outside the scope of the present experiments. Model performance is further enhanced through hyperparameter optimization, allowing fine-tuning of the algorithms to improve predictive accuracy. Finally, the models are rigorously evaluated using performance metrics such as precision, recall, and ROC curves. The procedure concludes with a synthesis of results, discussion of insights, and recommendations for practical implementation and future improvements. Although the anomaly detection algorithms are trained in a fully unsupervised manner (they do not use labels during fitting), their outputs are evaluated against binary anomaly labels provided by an Administrative Process Specialist from the CIUSSS-SLSJ procurement team. In practice, this means that the learning phase is

unsupervised, but model selection and benchmarking rely on a supervised comparison with this expert-defined expert-provided proxy labels; throughout the thesis, we therefore refer to this as a supervised evaluation of unsupervised detectors.

Figure 0.7 summarizes the anomaly-detection workflow: (a) define anomaly criteria (e.g., regulatory non-compliance, price/demand/KPI deviations, unusual orders/returns, delivery delays, abnormal received quantities), (b) collect the data, (c) run exploratory analysis (descriptive stats/visuals plus correlation and distribution/structure tests, including temporal dependence and stationarity), (d) preprocess (missing values, normalization, categorization/encoding, quality checks), (e) apply detection methods—(e1) statistical (*distance-based, Z-score, Grubbs*), (e2) ML (*Robust Covariance in Elliptic Envelope, Isolation Forest, Local Outlier Factor, One-Class SVM*), then (f) optimize via (f1) post-filters (*default (none), excess, lack, inverse*) then (f2) *Bayesian hyperparameter tuning* and (g) evaluate with *accuracy, precision, sensitivity (recall), specificity, F1-score, AUC-ROC, AUC-PR*, and (h) conclude with recommendations.

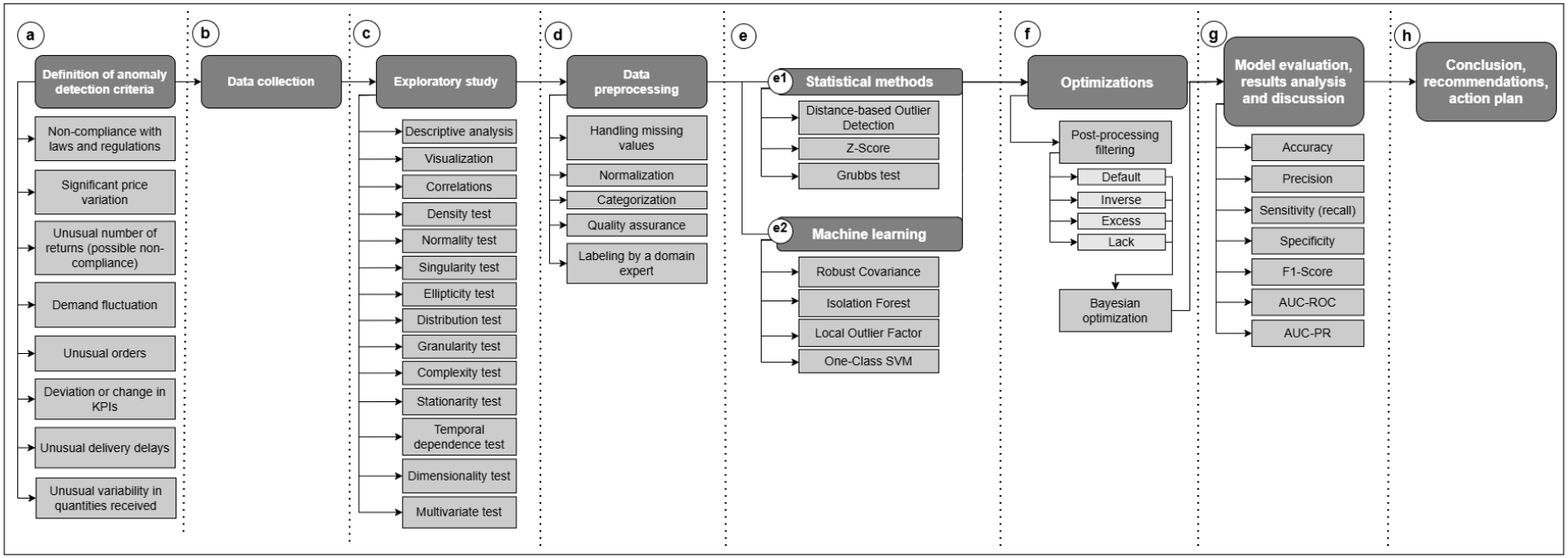


Figure 0.7 : Experimental procedure

DATA COLLECTION

Relevant data is collected from the SAP system of CIUSSS-SLSJ and other internal and external sources in a business intelligence perspective, as proposed by Fantini [126]. This includes orders, deliveries, contracts, and delays. The data extraction, preprocessing, and integration steps follow established guidance for preparing collected data for mining [127].

Table 4 : Selected SAP reports

Report	Transaction Code	Field
Articles	SQ01	ARTICLES_LIST
Orders	SQ01	BC_CODEP_DATEL
Valid Contracts	SQ01	CONTRATSVALIDES
Invoices	SQ01	FACTURATION
Receipts	MB51	CODE_MOUVEMENT == 101
Returns	MB51	CODE_MOUVEMENT == 122
Suppliers	SQ01	FOURNISSEURS
Divisions	MB51	DIVISION

Table 4 (Selected SAP reports) lists the SAP data sources used in this work by pairing each business object with its extraction mechanism and the corresponding selection field or filter applied. Specifically, the reports for *Articles*, *Orders*, *Valid Contracts*, *Invoices*, and *Suppliers* are retrieved via the custom query transaction `SQ01` using the fields `ARTICLES_LIST`, `BC_CODEP_DATEL`, `CONTRATSVALIDES`, `FACTURATION`, and `FOURNISSEURS`, respectively. Operational inventory movements are obtained from transaction `MB51`, where *Receipts* and *Returns* are isolated by filtering on the movement type (`CODE_MOUVEMENT == 101` for goods receipts and `CODE_MOUVEMENT == 122` for returns), and organizational attribution is captured using the `DIVISION` field (also from `MB51`) to represent *Divisions*. Data is extracted as XLS files using

transaction codes and their fields. Due to the large volume (Big Data), extractions are done in batches, cleaned, and merged using the algorithm outlined below.

Algorithm 0.1 parallel-processes each SAP report folder by cleaning every .XLS file (dropping empty rows/irrelevant columns and standardizing headers), concatenating the cleaned files into one DataFrame per folder, and exporting the merged result as an .XLSX dataset while freeing memory. Once cleaned, the .XLS files can safely be imported into an MSSQL relational database.

Then a series of exploratory and diagnostic analyses (Appendices A–I) were conducted prior to modeling. Descriptive statistics and visualizations highlighted heavy-tailed distributions and strong *skewness* for key monetary and delay variables, while density plots showed marked mass near zero with long right tails. Temporal dependence analyses revealed significant *autocorrelation* for several KPIs, and normality and ellipticity tests confirmed that *multivariate Gaussian* assumptions were not appropriate for most features. *Singularity* and *multicollinearity* checks informed the selection of variables and the use of *dimensionality reduction* (PCA), and *complexity and stationarity tests* suggested that simple *linear models* would be insufficient. These findings motivated the use of robust, *unsupervised anomaly detection algorithms* capable of handling *non-Gaussian*, correlated, and *non-stationary data*[50, 128].

Data: Excel files extracted from SAP
Result: Merged and cleaned dataset in XLSX format

```
foreach folder in {ITEMS, ORDERS, VALID_CONTRACTS, DIVISIONS, INVOICES,
  SUPPLIERS, RECEIPTS, RETURNS} do
    process in parallel with process_folder (folder);
end
Function clean_excel_file(file):
    read Excel file, skipping empty rows;
    remove irrelevant columns ("Unnamed", NaN);
    trim column names;
    return cleaned file;
Function merge_all_xls_files(folder):
    DataFrame ← empty;
    foreach file in folder do
        if file ends with .XLS then
            cleaned ← clean_excel_file(file);
            concatenate into DataFrame;
        end
    end
    return DataFrame;
Function process_folder(folder):
    print "Processing folder";
    merged ← merge_all_xls_files(folder);
    save merged dataset in .XLSX format;
    free memory with garbage collection;
```

Algorithm 0.1: Tool for Cleaning and Merging SAP Reports

GROUND TRUTH LABELLING BY DOMAIN EXPERT

In this project, the anomaly labels used for model tuning and evaluation were designed by an Administrative Process Specialist from the CIUSSS-SLSJ procurement team. This specialist is responsible for monitoring compliance with public procurement rules and internal procedures, and has extensive experience with SAP data and local purchasing practices. For each anomaly criterion and time period, the specialist examined the relevant SAP records and labelled observations as anomalous or normal according to their professional judgment, informed by applicable regulations (e.g. LCOP), internal guidelines, and operational knowledge of the supply chain. The domain expert carried out the labeling independently and was not exposed to any of the automatic anomaly scores or labels produced by the statistical and machine-learning models. The precise decision rules and thresholds used in this process are part of the institution's internal practice and were not fully documented for the author. In this thesis, we therefore treat these expert-provided labels a proxy label: they approximate the events that a knowledgeable administrator would consider worthy of attention, but they may include some subjectivity and label noise. This dependence on a single expert means the models may be biased toward the specialist's mental model of risk. No formal inter-rater agreement was computed, and thus the interpretation of borderline cases reflects only this expert's interpretation of CIUSSS-SLSJ policies. This limitation is discussed in Section 4.5. Applying classical, rule-based monitoring at CIUSSS-SLSJ is difficult because experts must interpret a complex, evolving regulatory framework (e.g., LCOP, PABSGC), making manual reviews slow, labor-intensive, and inconsistent. Moreover, EDA shows all variables deviate from normality ($p<0.01$), undermining Gaussian-based baselines and simple global thresholds on heterogeneous procurement values, thereby motivating robust, unsupervised AI methods with domain-adapted post-processing to produce more consistent and operationally relevant alerts.

0.5 ORGANIZATION OF THE THESIS

This thesis is structured into five chapters, each representing a milestone of the research methodology:

- **Introduction** presents the context, objectives, contributions of the thesis, materials, methods, and the overall organization of the document.
- **Chapter 1** provides a literature review, clarifying the state of research on public procurement, anomaly detection, algorithmic decision-making, and *hyperparameter tuning*.
- **Chapter 2** presents the first article, *"Anomaly detection using AI in healthcare procurement data across multiple criteria with tailored post-processing filters"* [129].
- **Chapter 3** presents the second article, *"AI-based anomaly detection in healthcare procurement using Apache Spark: a post-processing filters approach"* [130].
- **Chapter 4** discusses the results, highlighting methodological advances, the role of post-processing filters, and the challenges of scaling anomaly detection in healthcare procurement.
- **Chapter 5** concludes the thesis by summarizing the main findings, formulating recommendations for practice, and outlining directions for future research.

CHAPTER I

STATE OF THE ART

1.1 APPROACHES IN ANOMALY DETECTION

Research on anomaly detection has expanded rapidly, spanning a substantial body of literature, reflecting the field's diversity and complexity [39, 51]. Approaches vary widely and include statistical methods, probabilistic models, machine learning techniques, deep learning, and hybrid methods (Table 1.1).

Table 1.1 : Types of approaches, models, and algorithms for anomaly detection with examples and seminal references

Approach type	**Example methods/models)**
Statistical methods	Distance-based Outlier Detection (DbOD) [131]; Z-Score [132]; Grubbs' Test [3]
Probabilistic models	Gaussian mixture models (GMM) [133]; Hidden Markov models (HMM) [134]
Machine learning	Robust covariance (RC) [80] in Elliptic Envelope (EE) [109, 110]; One-class SVM (OC-SVM) [81]; Isolation Forest (IF) [79]; Local Outlier Factor (LOF) [83]
Deep learning	Autoencoders [135]; Convolutional Neural Networks (CNN) [136]; LSTM networks [137]
Hybrid techniques	RF–KMeans; KNN–KMeans [138]

Statistical and probabilistic foundations. Classical approaches model normality using rigorous statistical principles and flag significant deviations as anomalies. Widely used techniques include distance-based detection [131], *Z-score* [132], *Grubbs' test* [3], *extreme value theory* [139, 140], and regression-oriented treatments of extremes and outliers [141]. Probabilistic models such as HMMs and GMMs are also standard tools [133, 134]. These

techniques offer confidence measures and handle uncertainty, but may struggle with high-dimensional or highly non-linear data [142, 143].

Machine learning. ML-based methods learn patterns of normality from historical data and can better capture complex or non-linear structure. Common approaches include Robust Covariance [80] in Elliptic Envelope [109, 110], Isolation Forest [79], one-class SVM [81], and Local Outlier Factor [83]. These methods work well when normal behavior is complex or slowly evolving and can adapt over time.

Deep learning. Deep learning has advanced anomaly detection by learning rich, hierarchical representations from raw data [144, 145]. *Autoencoders* [135], *CNNs* [136] and *RNNs* such as *LSTMs* [137] are effective for images, audio, and time series. Benefits include modeling non-linear dependencies and high-dimensional interactions [146, 147]; challenges include computational cost and opacity [148].

Hybrid methods. Hybrid approaches combine strengths across paradigms to overcome individual limitations. For example, Lok et al. [138] propose *RF–KMeans* and *KNN–KMeans* hybrids, demonstrating improved flexibility and performance. This demonstrates how multiple anomaly detection methods can be combined to achieve results.

Applied impact and cross-domain trends. Applications in supply chains, inventory management, and decision support demonstrate practical and economic value [32, 82, 149]. The field continues to evolve toward higher accuracy, faster detection, and scalability to large, heterogeneous data [150, 151]. At the same time, interpretability and accountability are increasingly central, especially when decisions have significant real-world consequences [152, 153].

Ecosystem and data considerations. Advances in AI and analytics enable progress on complex, heterogeneous, high-dimensional data [154]. The proliferation of IoT and cyber-

physical systems has expanded real-time monitoring and anomaly detection across domains such as health [57], network and critical infrastructure security [58, 65, 66], and predictive maintenance [61, 62]. Addressing the challenges of massive data volumes and privacy constraints motivates robust, scalable techniques and privacy-preserving learning settings (e.g., federated learning) [155, 156].

Looking ahead, anomaly detection in healthcare will likely become more interdisciplinary and integral to innovative strategies for strengthening healthcare supply chains, alleviating workforce pressures, providing effective decision support, and improving the quality of life for workers and patients. This multidimensional trajectory aims to better surface and manage anomalies, supporting more stable supply chains and safer, more efficient care delivery environments.

1.2 APPLICATION DOMAINS OF ANOMALY DETECTION

Anomaly detection plays a critical role across many domains by revealing rare or unexpected patterns that may signal system malfunctions, fraud, or emerging risks. In cybersecurity and network monitoring, machine learning–based anomaly detection is used to identify intrusions and malicious activities in dynamic network environments [65, 66]. In finance, anomaly detection methods flag suspicious transactions and money-laundering activity among large volumes of transactional data [59, 121]. In industrial systems and the Internet of Things (IoT), anomaly detection is leveraged for predictive maintenance and sensor fault detection to preempt failures [61, 62, 67]. In healthcare, anomaly detection algorithms help identify abnormal patient vital signs, atypical imaging findings, or diagnostic outliers [58, 60, 75]. In public administration and procurement, data-driven anomaly detection is increasingly used to uncover irregularities in government spending, bid rigging, and procurement fraud [157, 158, 159]. Each domain brings unique challenges such as heterogeneous data, regulatory

constraints, data sparsity, and the need for interpretability, which drive the development of robust, context-aware, and scalable methods.

1.2.1 ANOMALY DETECTION IN THE PUBLIC SECTOR

Anomaly detection has gained momentum in public-sector applications, particularly in procurement, public spending, and regulatory oversight, where identifying irregular patterns can prevent corruption, fraud, and misuse of public funds. In public procurement, researchers have applied machine learning techniques and statistical models to detect collusion, favoritism, and bid rigging in government contracts [157, 158]. For instance, using open contracting data, methods have been developed to flag deviations from expected bidding behavior or contract award norms [157]. Graph-based techniques and pattern mining have also been leveraged to capture relational anomalies, such as unusual subgraph patterns among bidders that may indicate fraud-prone networks [158, 160]. In real-world deployments, explainability has emerged as a key requirement: explainable anomaly detection frameworks have been proposed to make results interpretable to auditors and public administrators [64, 150]. Challenges abound: public-sector data is often incomplete or distributed across heterogeneous systems, anomalies are rare and diverse, and labeled ground truth is scarce, making supervised learning difficult [30]. These constraints underscore the need for tailored, robust, and scalable anomaly detection approaches in this domain.

1.2.2 ANOMALY DETECTION IN PROCUREMENT AND SUPPLY CHAINS

In the domain of procurement and supply chains, anomaly detection is increasingly used to uncover irregularities such as fraud, collusion, invoice manipulation, and vendor opportunism across transactional flows. For example, researchers have developed methods

to detect collusion in public procurement auctions by applying and comparing a variety of machine learning algorithms over historical bid data from multiple countries [157, 158]. In procurement settings, explainable indicator-based frameworks have been proposed to generate interpretable *fraud scores* per transaction by combining multiple statistical signals [75, 150]. In supply chains, anomaly detection methods often combine forecasting and outlier detection; for instance, *LSTM*-based models integrated with *autoencoders* have been used to detect deviations in multivariate time series data tracking orders and shipments [74, 161]. Hybrid models using gradient boosting (e.g., *XGBoost*) have also been applied to flag suspicious supply chain behaviors and detect fraud in manufacturing contexts [75, 77]. To deal with scarce labeled instances of fraud, recent work employs semi-supervised learning: a two-phase framework that first isolates potential anomalies using an unsupervised detector (e.g., *Isolation Forest*) and then refines predictions using self-training support vector machines, showing promising results on real supply chain datasets [96]. *Isolation Forest* is cited here only for its original formulation [79]. However, challenges remain: high class imbalance, data heterogeneity across participants, dynamic concept drift, and the need for interpretability in business processes all complicate the deployment of such systems in real operational settings.

1.3 HYPERPARAMETER TUNING

Hyperparameter tuning is often considered as much an art as a science [162], due to the vast range of options and the absence of fixed rules applicable to all scenarios.

1.3.1 IMPORTANCE OF PARAMETERIZATION FOR PERFORMANCE

In the domain of anomaly detection, careful hyperparameter tuning is critical because small changes in configuration can drastically affect the sensitivity and specificity of detectors: for *Robust Covariance* (e.g. in *Elliptic Envelope* or *Minimum Covariance Determinant* meth-

ods) the support (or *support fraction* / contamination) parameter controls how many outliers the model tolerates before fitting the majority cluster, which determines how strictly the *Mahalanobis-distance* threshold is set, as per robust statistics theory [109]; in *One-Class SVM*, the choice of ν (upper bound on fraction of outliers), the kernel type (*linear*, *RBF*, *polynomial*), and *gamma* (*kernel bandwidth*) jointly govern the boundary tightness and generality of the normal region (and must often be cross-validated for good performance) [163]; for *Isolation Forest*, the number of trees and the subsample (or sample) size determine the randomness and stability of isolation paths, influencing the variance and bias tradeoff in anomaly scoring [79]; and in Local Outlier Factor, the number of neighbors and the distance metric (or weighting of distances) control the local density estimation sensitivity, making LOF's anomaly flags heavily dependent on neighborhood scale tuning (and often requiring heuristic or validation-based methods to calibrate) [83].

1.3.2 CLASSICAL TECHNIQUES

Grid Search performs exhaustive evaluation on a predetermined discretized grid, Random Search samples hyperparameters uniformly at random (and often more efficiently in high-dimensional spaces) [164], *Gradient-Based Search* uses hypergradient or bilevel methods to adjust hyperparameters via differentiable optimization [165], *Bayesian Optimization* builds a probabilistic surrogate (e.g. *Gaussian process*) to balance exploration and exploitation for global search [166], and *Evolutionary Algorithms* maintain a population of candidate hyperparameter settings and evolve them through mutation and crossover operators over generations [167].

GRID SEARCH

In many practical settings, this exhaustive method tests a predefined range of values for each hyperparameter. While simple to understand and implement, it becomes impractical as the number of hyperparameters increases due to the combinatorial explosion [168]. In practice, grid search is still widely used (e.g. in scikit-learn) for low-dimensional parameter spaces [169]. Its limitations are well documented: with d hyperparameters each with n candidate values, it requires n^d evaluations, which becomes infeasible even for moderate d [170].

RANDOM SEARCH

Proposed and studied by Bergstra and Bengio [164], random search samples hyperparameter combinations at random rather than exhaustively. They showed empirically and theoretically that random search often finds good configurations more efficiently than grid search, especially when only a subset of hyperparameters strongly influences performance [164]. In many cases, a modest number of random trials can outperform exhaustive grid-sampling [164].

GRADIENT-BASED SEARCH

Gradient-based methods (also called hypergradient methods) compute derivatives of the validation loss with respect to continuous hyperparameters and perform updates accordingly. These approaches are suitable in settings where the hyperparameter-to-loss mapping is differentiable (or can be approximated), and they can adapt continuously rather than via discrete sampling [165]. They are less commonly used in general-purpose Hyperparameter Optimization (HPO) compared to search-based methods, but have seen growing interest in deep-learning research (see, e.g., the survey by Chen et al. on *gradient-based bilevel optimization*) [165].

BAYESIAN OPTIMIZATION

Bayesian optimization builds a probabilistic surrogate model of the validation loss (e.g. *Gaussian processes or Tree-structured Parzen estimators*) and uses acquisition functions to propose promising hyperparameter values [171]. This approach is more sample-efficient (i.e. requires fewer evaluations) than random or grid search in many cases, especially for expensive-to-evaluate models [172]. However, the sequential nature of most *Bayesian* methods limits parallelism and complicates scaling to very high-dimensional hyperparameter spaces [173].

EVOLUTIONARY ALGORITHMS

Evolutionary (or population-based) algorithms leverage principles inspired by biological evolution, such as selection, mutation, and recombination, to evolve candidate hyperparameter configurations. They are effective in exploring complex, multimodal search spaces and do not require gradient information [174, 175]. For instance, Differential Evolution has been shown to outperform SMAC (a *Bayesian optimizer*) for many datasets [175]. More recently, hybrid and surrogate-assisted evolutionary approaches [e.g. Improved Equilibrium Optimizer (IEO)] have been developed to accelerate convergence in hyperparameter tuning [176]. The trade-offs are computational cost and sensitivity to the choice of genetic operators and control parameters [177].

1.3.3 CURRENT LIMITATIONS AND IMPACT ON ROBUSTNESS

A key limitation identified in the state of the art is the strong dependency of anomaly detection algorithms on carefully tuned hyperparameters, since small changes in values such as

contamination levels, kernel bandwidths, or neighborhood sizes can drastically alter sensitivity and specificity [178]. Scalability, interpretability, and stability further complicate deployment in real-world, large-scale temporal data such as healthcare procurement. In the first article of the current thesis, *Bayesian Optimization* was adopted as the most relevant strategy for hyperparameter tuning because of its sample-efficiency and ability to balance exploration and exploitation when only a few hyperparameters drive most of the performance variation [179]. By leveraging a probabilistic surrogate to guide the search, *Bayesian Optimization* quickly identified high-performing configurations on smaller-scale datasets, suggesting robust baselines without incurring excessive evaluation costs. In the second article of this master thesis (*Apache Spark*), our experiments indicated that these optimized hyperparameters transferred without re-optimization, which suggests their effectiveness was driven more by dataset characteristics than by computational scale. thereby saving significant distributed resources while preserving detection quality. This choice was particularly important because Bayesian Optimization presents challenges in distributed contexts: its inherently sequential nature contrasts with *Spark*'s parallel paradigm, stochastic variance across nodes introduces noise into performance estimates, Gaussian-process surrogates scale poorly with higher-dimensional parameter spaces [179], and repeated evaluations incur heavy communication and scheduling overhead across the cluster. Consequently, the two-stage strategy, optimizing hyperparameters once on smaller samples and then deploying them at scale, appeared to be both efficient and practically necessary in this case study, overcoming both methodological and infrastructural constraints while highlighting the importance of aligning hyperparameter optimization techniques with the realities of large-scale, distributed anomaly detection.

1.4 POST-PROCESSING IN ANOMALY DETECTION

After generating raw model outputs (anomaly flags) or tentative anomaly flags from a detection model, a necessary stage is post-processing, whose goal is to refine, filter, and consolidate the results so as to reduce false positives, improve interpretability, and produce actionable anomaly outputs. In many pipelines, post-processing bridges the gap between model outputs and final decisions (alerts, intervals, ranked anomalies).

One common post-processing operation is thresholding: selecting a cutoff such that flagged data above the threshold are declared anomalous and those below are considered normal. However, a static, global threshold often yields sub-optimal trade-offs between recall and precision, especially in non-stationary data streams subject to distribution shift [118]. Some works therefore adopt adaptive or dynamic thresholding, or even multi-level thresholds, to respond to distribution shifts [113]. In multivariate time-series anomaly detection, it has been shown that the choice of scoring or thresholding function often matters more than the base model itself, and dynamic scoring functions generally outperform fixed ones across diverse datasets [180].

Beyond simple thresholding, more advanced post-processing includes merging or smoothing of detected anomaly points into coherent intervals. For instance, isolated "spikes" of anomaly flags may be smoothed out to avoid labelling transient noise; conversely, gaps between flagged points may be bridged if they fall within a temporal tolerance. In visual or spatial anomaly detection, morphological or connected-component merging operations are often used to merge adjacent anomalous pixels or patches [114]. In video-oriented temporal action detection pipelines, post-processing modules aggregate per-frame detections into contiguous anomaly episodes [113].

Another key function of post-processing is false positive filtering and ranking. Especially in high-recall detectors, many candidate alarms may be spurious. Some systems introduce a secondary classifier or scoring step over proposed anomalies to prune false alarms; for example, in industrial image-level anomaly detection, a false-alarm classifier is applied on object-level features atop a baseline detector [181]. In network-intrusion or alert-monitoring contexts, alert-post-processing also involves correlation and aggregation of flagged flows to reduce operator burden [112].

Hybrid schemes also exist: for example, large-scale deployment frameworks decouple the anomaly-detection and alerting components, where the alert-filtering layer applies business logic, prioritisation or suppression rules on raw detection outputs; this filtering layer is effectively a form of post-processing that boosts precision in real-world operation [112].

Finally, some recent methods embed domain knowledge or pattern constraints into post-processing stages. For instance, in industrial defect-detection settings the post-processing module applies smoothing or filtering of the anomaly map to reduce isolated false positives and better localise meaningful defects [114]. While some contemporary models (e.g., end-to-end frameworks that jointly learn localisation and decision boundary) attempt to absorb or eliminate the need for ad-hoc post-processing, many legacy or industrial systems still rely on external post-processing modules for filtering, smoothing and event consolidation.

In contrast, drawing on literature that treats post-filtering / post-processing as an explicit refinement stage in detection pipelines, common in signal processing and imaging applications [115, 116, 117] and also discussed in anomaly detection under distribution shift and operational constraints [113, 118], the framework presented in the current thesis begins with a *default filter* that deliberately preserves raw outputs as a baseline, providing a clear reference point for assessing the incremental effect of subsequent filtering strategies. The *inverse filter* is

used as a robustness stress-test by systematically flipping anomaly classifications to probe boundary strictness; while this idea is related to robustness and sensitivity analyses, it appears less commonly described as an explicit post-processing step in the sources reviewed (which often assume the anomaly/normal direction is correct). The *excess filter* goes beyond a purely global percentile rule by retaining anomalies based on deviations above a moving average with a fixed 30-day horizon, chosen to align with the CIUSSS-SLSJ monthly reporting cycle; this coupling to an operational cadence can improve domain interpretability compared with purely distribution-based thresholds. Finally, the *lack filter* focuses on unusually low values (e.g., drops in sensor readings or economic activity), a class of anomaly that may receive less attention in approaches that primarily emphasize spikes and surges.

Taken together, these filters provide bidirectional sensitivity to both excesses and deficiencies, while embedding organisational and application-level constraints into the filtering logic. In summary, this contribution goes beyond noise-reduction: it reframes post-processing as a targeted, context-aware refinement stage that complements adaptive thresholding and related anomaly-scoring approaches, thereby enhancing robustness, interpretability, and practical relevance of anomaly detection outputs.

1.5 POSITIONING OF THE THESIS

This thesis is positioned at the intersection of public-sector procurement, anomaly detection, and decision support, with a concrete case in the Québec healthcare network (CIUSSS-SLSJ). Despite the rich temporal properties confirmed by exploratory data analysis, *deep learning* models (such as *Long Short-Term Memory (LSTM)s* and *Autoencoders*) were explicitly excluded from the current study in favor of more transparent machine learning methods, reflecting the acute need for interpretability, explainability, and resource efficiency required by public sector governance frameworks (e.g., LCOP) where decisions must be

auditable and accountable. It responds to a practical need for early detection of procurement irregularities, such as delivery delays, atypical price variations, demand fluctuations, or back-orders, using real SAP data and within the governance and regulatory context of the health system. Methodological choices have been aligned with operational realities rather than purely academic idealisations.

Relative to the state of the art, the contribution shifts emphasis from exclusively improving model internals (for example via extensive hyperparameter optimization [166, 171]) toward domain-aware post-processing that translates raw model outputs (anomaly flags) into actionable, manager-ready signals. While many works focus on scoring functions or architecture choices in anomaly detection, fewer seem to explicitly investigate the transformation of those flags into domain-tailored alerts suitable for decision support in procurement contexts. The choice of filtering logic thus becomes a core dimension rather than a secondary "cleanup" step.

Empirically, the thesis is structured around two original articles. The first, *Anomaly Detection Using AI in Healthcare Procurement Data Across Multiple Criteria with Tailored Post-Processing Filters* [129], demonstrates that unsupervised algorithms (IF, LOF, OC-SVM, RC in EE) combined with careful preprocessing [*normalisation*, *Principal Component Analysis (PCA)*] and the proposed filters yield operationally meaningful detections on CIUSSS-SLSJ procurement data. This first study primarily addresses RQ1 and RQ2 by designing and evaluating an unsupervised anomaly detection pipeline on CIUSSS-SLSJ procurement data. More specifically, it examines how combinations of algorithms (IF, LOF, OC-SVM, RC), *Bayesian hyperparameter optimization*, and tailored post-processing filters influence both detection performance and the operational relevance of the generated alerts. The second article, *AI-based anomaly detection in healthcare procurement using Apache Spark: a post-processing filters approach* [130], migrates the pipeline to *Apache Spark* to examine distributed execution,

identifying where the logic transfers seamlessly and where algorithmic constraints limit native scaling (notably for *distance-* and *kernel-based* models). This second study primarily addresses RQ3 by re-implementing the anomaly detection pipeline in *Apache Spark* and comparing its performance and outputs against the *Pandas* implementation. It focuses on scalability, execution constraints, and the preservation (or degradation) of anomaly flags and filtered alerts when the pipeline is deployed in a distributed environment.

Conceptually, this re-frames the optimization levers in anomaly detection: whereas Bayesian hyperparameter optimization remains valuable [166, 171], the experiments reveal that well-designed, domain-specific post-processing can supply comparable or superior gains at much lower computational cost, an especially pertinent trade-off for public organisations with limited compute resources. In this sense, post-processing is elevated from a generic clean-up stage to a core component that contextualises anomalies for governance and decision-making. This thesis is positioned by arguing, on the basis of the case study, that in resource-constrained, regulation-heavy environments, optimizing domain-specific post-processing filters may offer a promising operational trade-off compared to pursuing marginal detection gains via complex, expensive hyperparameter optimization.

Methodologically, the thesis contributes a pragmatic pipeline that is both interpretable and transferable. In the experiments, Isolation Forest (IF) combined with the 'excess' filter appeared to be among the more robust and stable configurations across settings, supporting clearer alerting with fewer tuning burdens for this particular case study. At scale, the *Spark* migration confirms the durability of the filter logic, while transparently documenting limitations (for example, broadcast-workarounds that hinder full scalability for Local Outlier Factor and One-Class SVM). This candour about constraints helps delineate the boundary between proof-of-concept and operational readiness.

Situated within the literature, the thesis contributes along three axes. Scientifically, it redirects attention toward post-processing as a complementary optimization lever in anomaly detection applied to public procurement. Methodologically, it advances a hybrid view that couples detection algorithms with business-rule filtering, improving interpretability and uptake by public decision-makers. Practically, it offers a replicable pathway, validated on real hospital procurement data, toward integrating AI-based anomaly detection into ERP environments (for example SAP today), while surfacing concrete research avenues (*Spark*-native methods, distributed HPO, federated learning, multi-institutional validation).

Based on the literature reviewed for this thesis, we did not find prior studies that simultaneously (i) integrate Bayesian hyperparameter optimization and domain-specific post-processing filters for multi-criteria anomaly detection in healthcare procurement data and (ii) examine scalability in a distributed *Apache Spark* environment.

In sum, the thesis occupies a pragmatic niche between academic prototypes and operational deployment: it retains rigorous anomaly-detection methodology yet re-orients design choices toward organisational usefulness, interpretability, and scalability in a public-health procurement context.

The following chapters (II and III) consist of the two core studies of this thesis. Chapter II presents the first study (as submitted in manuscript form), and Chapter III the second study.

CHAPTER II

ANOMALY DETECTION USING AI IN HEALTHCARE PROCUREMENT DATA ACROSS MULTIPLE CRITERIA WITH TAILORED POST-PROCESSING FILTERS

Colin Bouchard (Université du Québec à Chicoutimi)

Julien Maitre (Université du Québec à Chicoutimi)

Bruno Bouchard (Université du Québec à Chicoutimi)

Realised for

University of Québec at Chicoutimi / Université du Québec à Chicoutimi

In collaboration with the

Integrated University Health and Social Services Center of Saguenay-Lac-Saint-Jean / Centre intégré universitaire de santé et de services sociaux du Saguenay-Lac-Saint-Jean

December 15, 2025

2.1 ABSTRACT

This study evaluates unsupervised machine learning methods for anomaly detection in healthcare public procurement data from the CIUSSS-SLSJ SAP system. Four algorithms are assessed: Robust Covariance (RC) in Elliptic Envelope (EE), Isolation Forest (IF), One-Class SVM (OC-SVM), and Local Outlier Factor (LOF). Preprocessing pipelines include standardization or normalization and PCA, with all transforms fitted on the training split only to prevent information leakage. Outputs are optionally refined using rule-based post-processing filters (excess, lack, inverse, default), and hyperparameters are tuned via Bayesian optimization. Test sets were expert-labeled to identify anomalies related to price, value, purchased and received quantities, delivery delays, and atypical return, demand, and order patterns. Performance is reported using standard metrics (accuracy, precision, recall, specificity, F1-score, AUC-ROC, AUC-PR) and averaged across three independent expert-labeled hold-out datasets per criterion. In these limited experiments, Isolation Forest (IF) achieved the strongest overall performance, though conclusions are constrained by small sample sizes. The results support the feasibility of AI-based anomaly detection in SAP procurement workflows while highlighting practical barriers in Quebec's healthcare supply chain, including data warehousing requirements, storage and integration constraints, and the need for more centralized ERP standardization across the Santé Quebec network.

Keywords: public procurements, healthcare supply chain, anomaly detection, artificial intelligence, AI, unsupervised, machine learning, Robust Covariance (RC) in Elliptic Envelope (EE), Isolation Forest (IF), One-Class SVM (OC-SVM), Local Outlier Factor (LOF).

2.2 INTRODUCTION

In the evolving landscape of healthcare logistics and management, the integration of artificial intelligence (AI) into supply chain processes promises a potential cost reduction [28] and an increase in operational efficiency [182]. This paper aims to enhance our understanding of the challenges and future trends identified by Painuly et al. [27]. This research paper seeks to explore the application of AI within the supply chains of the Integrated University Health and Social Services Center of Saguenay-Lac-Saint-Jean / Centre intégré universitaire de santé et de services sociaux du Saguenay-Lac-Saint-Jean, utilizing key procurement data from SAP software systems. Our study is inspired by the methodologies of Ten et al. [65], Genge et al. [66], Shi et al. [67], Bauder et al. [121], and Patel et al. [123], which have previously demonstrated the potential of AI for anomaly detection in various domains.

This work emphasizes the integration of preprocessing techniques, such as normalization, standardization [183], and dimensionality reduction (e.g., PCA) [184] along with post-processing steps, understood broadly as operations applied after the initial output to enhance or refine results [116], include tailored output filters (e.g., *excess*, *lack*, *inverse*, and *default*) to refine anomaly detection outcomes. These filters rely on heuristic, domain-informed choices (e.g., a 30-day trailing moving-average window and expert-defined thresholds). They are not claimed to be optimal, and would need to be recalibrated when transferring the methodology to other institutions or KPI definitions. We hypothesize that by applying tailored filtering strategies, we can enhance detection performance metrics by reducing noise. To enhance algorithmic performance, *Bayesian optimization* is employed, following the approach outlined by Feurer and Hutter [124].

Beyond technical advancements, this study addresses practical considerations in adopting AI technologies within a complex procurement ecosystem. Including navigating public pro-

curement rules and the governance constraints that shape public-sector algorithm deployment [12, 15].

Ultimately, the research seeks to establish a robust, real-time alert system powered by AI, enabling the CIUSSS-SLSJ to proactively respond to anomalies. By leveraging live Enterprise Resource Planning (ERP) data, this system aims to enhance operational resilience and foster a culture of accountability and efficiency in public procurement management, in alignment with modern Industry 4.0 practices, as previously demonstrated by Jagadeesan et al. [31].

2.3 CONTRIBUTIONS

1. An empirical comparison of four unsupervised anomaly detection models on CIUSSS-SLSJ SAP procurement data across multiple anomaly criteria.

2. A evaluation pipeline with preprocessing, Bayesian hyperparameter tuning, and expert-labeled hold-out testing using standard performance metrics.

3. Tailored post-processing filters (default, inverse, excess, lack) to refine anomaly flags.

2.4 OBJECTIVE OF THIS STUDY

The primary objective of this study is to design and implement AI-driven methods to detect and address anomalies in healthcare procurement data. By leveraging specific unsupervised machine learning algorithms, the research aims to identify irregularities that may indicate inefficiencies, fraud, or compliance issues within the procurement process.

Since we hypothesize that applying tailored filtering strategies may enhance detection performance metrics by reducing noise, our secondary objective is to demonstrate the potential of these specific filters (default, inverse, excess, and lack).

2.5 RESEARCH PROTOCOL AND METHODOLOGY

This study follows a rigorous eight-step protocol to ensure reliable results. First, the anomaly criteria were defined and the relevant variables were selected. The data are split into an unlabeled training set used to fit the unsupervised models and three distinct expert-labeled hold-out datasets per anomaly criterion (triplicate) used only for supervised evaluation; all reported metrics are computed on each hold-out dataset and then averaged. No expert labels are used to fit the models. All expert labels were produced by a single administrative process specialist from CIUSSS-SLSJ, and no formal inter-rater agreement was computed. As a result, the ground truth reflects this expert's interpretation of the CIUSSS-SLSJ policies and may differ from how other specialists would categorize borderline cases. An exhaustive exploratory data analysis (EDA) was conducted to gain insights, followed by data preprocessing to handle missing values, apply normalization/standardization, encode categorical variables, and ensure quality. Statistical models (e.g., *Z-Score*, *Grubbs' Test*, and *Distance-Based Anomaly Detection*) and unsupervised anomaly detection algorithms (e.g., *Local Outlier Factor*, *One-Class SVM*, *Isolation Forest*, and *Robust Covariance* in *Elliptic Envelope*) were applied. The processed output was filtered using contextual filters (e.g., *excess*, *lack*, and *inverse* filters) and compared to a default, unfiltered scenario. We then optimized hyperparameters through *Bayesian optimization*. When hyperparameter tuning is performed, the selection role is rotated across the three labeled hold-out datasets (triplicate). Finally, the results were evaluated using metrics such as *accuracy, precision, recall, specificity, F1-score*, *AUC-ROC*, *AUC-PR*, and a confusion matrix, leading to actionable recommendations and conclusions based on the findings. Detailed descriptive statistics, distributional analyses, and assumption checks for all variables are provided in Appendices A–I; these diagnostics guided variable selection, scaling decisions, and the choice of anomaly detection algorithms.

2.5.1 DATA-ENGINEERING SUMMARY

Nine SAP export types were processed: `PRODUCTS`, `ORDERS`, `VALID_CONTRACTS`, `DIVISIONS`, `BILLS`, `SUPPLIERS`, `RECEIPT_OF_GOODS`, `RETURNS`, `PURCHASE_REQUESTS`.

A uniform five-step ETL routine was applied to each export:

1. **Ingest**: load all `.XLS` files, skipping the SAP header rows (3 or 6 lines).
2. **Rename**: apply hand-crafted French-to-English column mappings and drop unnamed or empty fields.
3. **Cast**: standardize types: IDs to strings (`Codify`); amounts to floats (`Floatify`); counters to integers (`Intify`); dates parsed with error coercion.
4. **Prune**: remove SAP-specific noise columns (e.g. `Lot`, `Username`) and de-duplicate on primary keys.
5. **Load**: write the cleaned data frames to the staging SQL database (`AD_Staging`), streaming one row per chunk to avoid ODBC row-size limits.

The script executes the full pipeline in $\sim$22 min on a 16-core workstation and yields nine analytics tables.

2.5.2 DATASETS AND VALIDATION SCHEME

Each anomaly criterion (Table 1) is populated by three independent datasets, together covering fiscal years 2018–2023.

Each dataset is built by joining one or more of the nine analytics tables produced in the previous step.

- Mean rows per dataset (after quality screening and before labeling): 7.42 M
- Mean expert-labeled rows per dataset : 3910
- Average anomaly prevalence of labeled dataset: $28.38\% \pm \sigma_p$
- σ_p is the standard deviation (21.64%) of prevalence across the held-out datasets.
- Chronological 70 / 30 train–test split; no temporal leakage
- LOF was run in novelty mode (novelty=True) and fitted on train only.

All artifacts (confusion matrices, curves, logs) are archived in the project repository; due to CIUSSS-SLSJ confidentiality constraints, the repository is not publicly accessible, but the full configuration schema (hyperparameters, preprocessing choices, random seeds, and software versions) is provided to enable auditability and internal reproducibility. Confusion matrices are computed by comparing binary anomaly flags $\hat{y}$ (returned by `predict()`) to the expert labels, whereas *AUC-ROC* and *AUC-PR* are computed from continuous anomaly scores $s(x)$ (higher $s(x)$ indicates more anomalous; scores are sign-inverted where necessary). For each anomaly criterion and dataset, threshold-dependent metrics are computed by comparing the binary anomaly flags $\hat{y}$ to the expert labels, whereas curve-based metrics (*AUC-ROC* and *AUC-PR*) are computed by comparing the continuous anomaly scores $s(x)$ to the expert labels.

(see Section 0.4.2). The algorithms themselves are fitted without labels, but their performance and operating points are assessed on this labelled subset. Methodologically, this corresponds to unsupervised anomaly detectors evaluated in a supervised manner, i.e., a semi-supervised evaluation scenario rather than a fully unsupervised setting.

2.5.3 BAYESIAN HYPERPARAMETER OPTIMIZATION

We use `scikit-optimize` with 40 evaluations per dataset; the objective is to maximise validation *F1-Score* from the binary anomaly flags $\hat{y} \in \{0,1\}$ returned by `predict()` on an internal validation fold. The internal validation fold is obtained via a chronological split within the training period, so that validation always uses data from later months than the corresponding training subset. This preserves the temporal structure and avoids leakage from the future into hyperparameter tuning. Each algorithm has its own search space (Table 2.1).

Table 2.1 : Hyperparameter search space

Algorithm	**Parameter**	**Range / distribution**
Robust Covariance (RC) in Elliptic Envelope (EE)	`support_fraction`	0.60–1.00 (continuous)
	`contamination`	0.10–0.30 (continuous)
Isolation Forest (IF)	`n_estimators`	1–150 (integer)
	`max_samples`	0.10–1.00 (continuous)
One-Class SVM (OC-SVM)	ν (nu)	0.01–0.40 (continuous)
	γ (gamma)	{`scale`, `auto`} (categorical)
Local Outlier Factor (LOF)	`n_neighbors`	2–50 (integer)
	`leaf_size`	1–40 (integer)

2.6 ANOMALY DETECTION ALGORITHMS FOR THIS STUDY

Table 2.2 : Theoretical formulas of selected anomaly detection algorithms

Anomaly detection algorithms	Simplified theoretical formula
Robust Covariance (RC) [80] in EllipticEnvelope [109, 110]	$(\hat{\mu}, \hat{\Sigma}) = \mathrm{MCD}_\gamma(X), \quad \gamma = \texttt{support\_fraction},$ $d^2(x) = (x-\hat{\mu})^\top \hat{\Sigma}^{-1}(x-\hat{\mu}),$ $s(x) = d^2(x) \quad (\text{higher} = \text{more anomalous}),$ $f(x) = \texttt{offset} - d^2(x), \quad \hat{y}(x) = \mathbb{I}[f(x) < 0],$ $\texttt{offset}$ set using $\varepsilon = \texttt{contamination},$ (sklearn) $s(x) = -\texttt{decision\_function}(x).$
Isolation Forest (IF) [79, 185]	$E(h(x)) = \frac{1}{t}\sum_{j=1}^{t} h_j(x),$ $s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (\text{higher} = \text{more anomalous}),$ $c(n) = 2H_{n-1} - \frac{2(n-1)}{n} \quad (n > 2),$ $H_{n-1} = \sum_{k=1}^{n-1} \frac{1}{k} \approx \ln(n-1) + \gamma,$ (sklearn) $s(x) = -\texttt{decision\_function}(x).$
One-Class SVM (OC-SVM) [81]	$\min_{w,\rho,\xi} \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n} \xi_i - \rho,$ $\text{s.t. } w^\top \phi(x_i) \geq \rho - \xi_i,\ \xi_i \geq 0,\ i = 1,\ldots,n,$ $g(x) = w^\top \phi(x) - \rho = \sum_{i=1}^{n} \alpha_i K(x_i, x) - \rho,$ $\hat{y}(x) = \mathbb{I}[g(x) < 0],$ $s(x) = -g(x) \quad (\text{higher} = \text{more anomalous}),$ $K(x,z) = \langle \phi(x), \phi(z) \rangle.$
Local Outlier Factor (LOF) [83]	$\text{reach\_dist}_k(x,o) = \max(\text{k-dist}(o), d(x,o)),$ $\text{lrd}_k(x) = \left(\frac{1}{\lvert N_k(x)\rvert} \sum_{o \in N_k(x)} \text{reach\_dist}_k(x,o) \right)^{-1},$ $\text{LOF}_k(x) = \frac{1}{\lvert N_k(x)\rvert} \sum_{o \in N_k(x)} \frac{\text{lrd}_k(o)}{\text{lrd}_k(x)},$ $s(x) = \text{LOF}_k(x) \quad (\text{higher} = \text{more anomalous}),$ (sklearn) $s(x) = -\texttt{negative\_outlier\_factor\_}(x).$

The Table 2.2 summarizes four widely used anomaly detection methods by pairing each algorithm with a compact "scoring rule" that ranks points by how atypical they are, plus (when applicable) how common software implementations convert that score into an inlier/outlier label: *Robust Covariance* (RC) in *Elliptic Envelope* (EE) via the *Minimum Covariance Determinant* (MCD) estimates a robust mean $\hat{\mu}$ and covariance $\hat{\Sigma}$ from a high-support subset (controlled by $\gamma = \texttt{support\_fraction}$), then measures atypicality using the squared *Mahalanobis distance* $d^2(x) = (x-\hat{\mu})^\top \hat{\Sigma}^{-1}(x-\hat{\mu})$, so larger $d^2(x)$ implies a more extreme point under an ellipsoidal *Gaussian*-like model; in practice, a threshold (the `offset`) is calibrated using an assumed outlier proportion $\varepsilon = \texttt{contamination}$ so that $f(x) = \texttt{offset} - d^2(x)$ becomes negative for predicted anomalies [109, 110]. Isolation Forest (IF) instead defines anomalies as points that become isolated quickly in random partition trees: it averages the path length $h_j(x)$ over t trees to obtain $E(h(x))$, then converts it to an anomaly score $s(x,n) = 2^{-E(h(x))/c(n)}$ where $c(n)$ is the expected path length in a random binary search tree (using harmonic numbers H_{n-1}), yielding higher scores for shorter-than-expected paths (more isolatable points) [79, 185]. One-Class SVM (OC-SVM) frames anomaly detection as learning a boundary in feature space that encloses most training data: it minimizes $\frac{1}{2}\|w\|^2$ while allowing slack ξ_i and enforcing $w^\top \phi(x_i) \geq \rho - \xi_i$, with ν controlling the trade-off (and implicitly the fraction of outliers/support vectors); prediction uses the signed decision function $g(x) = w^\top \phi(x) - \rho = \sum_i \alpha_i K(x_i, x) - \rho$, labeling points with $g(x) < 0$ as outliers and taking $s(x) = -g(x)$ as an *anomalousness* score [81]. *Local Outlier Factor* (LOF) is explicitly local-density-based: it defines a reachability distance $\text{reach\_dist}_k(x,o)$ that smooths distances by each neighbor's k-distance, computes the local reachability density $\text{lrd}_k(x)$ as the inverse of average reachability distance to k neighbors, and then sets $\text{LOF}_k(x)$ to the average ratio of neighbors' densities to the point's density, so $\text{LOF}_k(x) > 1$ indicates x lies in a region sparser than its neighbors (a local outlier), with larger values indicating stronger local deviation [83]; finally, the table notes that some `scikit-learn` APIs flip signs (e.g., return a higher

`decision_function` for inliers), so the displayed $s(x)$ is often implemented as the negative of those returned quantities to keep the interpretation *higher = more anomalous* consistent across methods [110, 185].

2.7 ANOMALY CRITERIA SELECTED FOR THE STUDY

As shown in Table 2.3, a set of anomaly criteria has been selected for this study. These criteria are used to implement the anomaly detection system. Each criteria consists of a different dataset and is evaluated individually.

The criteria selected for the anomaly detection study, as detailed in Table 2.3, encompass a range of operational and transactional parameters critical for identifying deviations that could indicate fraud, non-compliance, or operational inefficiency. Significant price variation, for example, captures changes in pricing that deviate markedly from historical norms, potentially signaling manipulative behaviors or market anomalies. Similarly, significant delivery delays are included as a criterion due to their ability to highlight disruptions in supply chain processes, which could forewarn of deeper issues such as logistical failures or breaches in contract compliance. The unusual number of returns is monitored as it may reflect underlying quality or satisfaction issues, directly impacting public trust and political reputation. Abnormal demand fluctuation is another key indicator, pointing to potentially unsustainable or irregular purchasing patterns that could suggest speculative behavior or stockpiling. Lastly, sudden changes in key performance indicators (KPIs) and unusual orders, such as those out of scope for specific departments, are scrutinized to ensure that all organizational activities align with established operational standards and ethical practices. By integrating these criteria into the anomaly detection system, the study aims to provide a robust framework for early detection and mitigation of activities that could adversely affect the organization. As shown in Table 2.3,

Table 2.3 : Anomaly criteria and associated variables

Anomaly Criterion	Associated Variables
Significant price variation	AverageProductPrice, AverageOrderedValue, AverageOrderedQuantity
Significant delivery time variation	AverageDaysDelayLessThan30Days, AverageDaysDelayGreaterThan30Days
Unusual number of returns	ReturnCounts
Abnormal demand fluctuation	DemandesCounts
Sudden deviation or change in KPIs	PercentagePerfectOrders, PercentagePerfectOrders_36M_RollingAvg, PercentagePerfectOrders_12M_RollingAvg, PercentagePerfectOrders_3M_RollingAvg, PercentagePerfectInvoices, PercentagePerfectInvoices_36M_-RollingAvg, PercentagePerfectInvoices_12M_-RollingAvg, PercentagePerfectInvoices_3M_RollingAvg
Unusual delivery delays	NumberDelaysLessThan30Days, NumberDelaysGreaterThan30Days, TotalNumberofDelaysPerPeriod, NumberOfDelayedItems
Unusual variability in received quantities	AverageReceivedQuantity

each of these anomaly criteria has been associated with the corresponding variable in each data set. Anomalies are specifically detected on these variables.

2.7.1 DEFINING NORMS

Variation in prices, delays in deliveries, or product conformity can impact healthcare supply chain operations. In this subsection, we will define what constitutes a norm and discuss deviations that could be indicative of anomalies [132]. Price variations, whether due to unintentional errors or attempts at overcharge, can lead to financial losses. Anomaly

detection algorithms can identify unusual price fluctuations by product. These detections allow managers to intervene with suppliers [186]. Ensuring delivery deadlines are met is important for maintaining the integrity of healthcare supply chains. Delays can directly affect the availability of critical medical supplies, impacting patient care and safety [68, 187, 188, 189]. We hypothesize that an analysis of return reasons from deliveries can effectively detect anomalies related to product non-conformity. This method focuses on why items are returned by procurement departments using traditional classification methods such as decision trees, random forests and other methods, revealing issues such as quality deficiencies or specification mismatches [190]. Preventing shortages is critical for the efficient management of healthcare supply chains. This involves detecting anomalies related to unusual delays in deliveries and variability in received quantities. Advanced monitoring systems and predictive analytics play a key role in identifying these anomalies early, allowing organizations to take proactive measures to avoid disruptions [68, 187, 188, 189]. Detecting unusual delays in deliveries is essential to prevent potential shortages. Close monitoring of delivery schedules and swift identification of deviations from the norm help in maintaining continuous supply chain operations. This not only prevents operational disruptions but also ensures patient safety by avoiding critical supply shortages [68, 187, 188, 189]. Monitoring the consistency of received quantities is vital. Any significant deviation from expected levels can signal potential supply issues, allowing procurement teams to act quickly to address potential shortages. This proactive approach helps maintain consistent supply levels and operational stability [191].

2.8 TAILORED POST-PROCESSING OUTPUT FILTERS

Anomaly detection systems often generate raw outputs that require further refinement to extract meaningful insights. In the context of the current article, post-processing filters helped tailor anomaly classification to specific application needs, ensuring that only the most relevant

deviations are considered. In some cases, not all detected anomalies are equally important. In fact, certain anomaly criteria require focusing exclusively on high-magnitude output, while others prioritize unusually low values. Additionally, what is traditionally classified as an anomaly may, under specific conditions, represent normal behavior, whereas its inverse could signify an actual anomaly. We hypothesize that tailored filtering strategies may reduce noise in the anomaly stream; any resulting effect on detection performance remains to be validated empirically. By refining anomaly classification, these filters might help focus on the most relevant deviations, reducing false flags and enhancing the overall reliability of the detection system.

DEFAULT (NO FILTER)

In the context of this study, the *default* filter applies no modifications, preserving the raw output for direct analysis. It serves as a baseline to evaluate the effectiveness of other filtering techniques. By comparing results against this unfiltered reference, one can assess how different filters influence anomaly detection and overall system performance.

INVERSE FILTER

This filter swaps anomaly classifications: observations labeled as anomalous under the default filter are reclassified as normal, and vice versa. By reversing the interpretation, this filter provides an alternative perspective on the dataset, helping to evaluate the sensitivity of the detection model. It is particularly useful for testing robustness and identifying cases where classification boundaries may be too strict or lenient.

EXCESS FILTER

The *Excess* filter retains only values that exceed a moving-average threshold; in our experiments we used a trailing 30-day window because it matches the CIUSSS-SLSJ monthly reporting cycle. Other horizons can be substituted to fit different operational cadences. This ensures that only high-magnitude deviations are retained while filtering out minor fluctuations. It is well-suited for applications where extreme outliers, such as sudden spikes in financial transactions, temperature anomalies, or performance surges, are of primary concern.

LACK FILTER

Conversely, the *Lack* filter identifies anomalies that fall below the 30-day moving average threshold, emphasizing unusually low values. This filtering approach is particularly beneficial in scenarios such as detecting unexpected drops in sensor readings, system failures, or economic slowdowns. By isolating low-value anomalies, it helps uncover patterns that might otherwise be overlooked in broader anomaly detection frameworks.

2.9 EVALUATION OF THE MODELS

The evaluation of anomaly detection models is essential to ensure their effectiveness and reliability across various applications, including cybersecurity [192], industrial monitoring [66], healthcare diagnostics [193], and fraud detection [59]. The justness of these models is assessed using various performance metrics, each derived from fundamental classification outcomes: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) [96, 194]. In this study, TP, FP, TN, and FN values are determined based on the labeling of the test dataset by an Administrative Process Specialist from the CIUSSS-SLSJ,

ensuring that the ground truth is treated as reference labels for evaluating anomaly detection performance. TP represents correctly identified anomalies, FP accounts for normal instances misclassified as anomalies, TN corresponds to normal observations correctly identified, and FN represents anomalies that the model failed to detect [96, 194].

Accuracy is the most used general measure of correctness [194]. It quantifies the proportion of correctly classified instances but may be misleading in imbalanced datasets where anomalies are rare. Precision, or positive predictive value, evaluates the proportion of detected anomalies that are genuinely anomalous [96], making it important in scenarios where reducing false alarms is essential, such as fraud detection [59]. Recall (sensitivity), on the other hand, assesses the model's ability to detect actual anomalies, ensuring that critical events are not missed, which is particularly important in applications like medical diagnostics [193]. Specificity measures the proportion of correctly classified normal instances, complementing recall by minimizing false positives [96, 194], which is valuable in automated monitoring systems where excessive alerts can cause inefficiencies [66]. The F1-score, a harmonic mean of precision and recall, is particularly useful when a balance between false positives and false negatives is necessary [96, 194]. *Area Under the Receiver Operating Characteristic Curve* evaluates a model's ability to discriminate between normal and anomalous instances across varying thresholds by plotting the true positive rate (TPR) against the false positive rate (FPR) [99]. Meanwhile, *Area Under the Precision-Recall Curve* provides a more insightful performance metric in highly imbalanced datasets, where precision-recall trade-offs better reflect model capabilities compared to ROC analysis [102]. The mathematical formulations for these metrics are provided in Table 2 [96, 194]. By analyzing these performance measures comprehensively, we gain deeper insights into the strengths and limitations of different anomaly detection models.

2.10 RESULTS

In this section, we analyze the performance of the anomaly-detection models. Each figure and analysis reports the mean over three independent held-out test datasets (one run per dataset) for each anomaly criterion; error bars show a heuristic aggregate variability indicator defined as $E_g = \sum_{k \in \mathscr{K}} \bar{\sigma}_{g,k}$, where $\bar{\sigma}_{g,k}$ is the configuration-averaged standard deviation of metric k across the three runs (see equations 2.1, 2.2, 2.3 and 2.4).

Aggregation and variability notation. Let $r \in \{1,2,3\}$ index the triplicate runs, and let $k \in \mathscr{K}$ index the evaluation metrics (e.g., *Accuracy*, *Precision*, *Recall*, *Specificity*, *F1*, *AUC-ROC*, *AUC-PR*). For a configuration c, we denote the run-level score by $s_{c,r,k}$ and define the per-metric mean and standard deviation across runs as:

$$\mu_{c,k} = \frac{1}{3} \sum_{r=1}^{3} s_{c,r,k}, \qquad \sigma_{c,k} = \sqrt{\frac{1}{2} \sum_{r=1}^{3} \left(s_{c,r,k} - \mu_{c,k}\right)^2}. \tag{2.1}$$

For any grouping g (e.g., model, feature, filter), with configuration set C_g, we aggregate by averaging over configurations:

$$\bar{\mu}_{g,k} = \frac{1}{|C_g|} \sum_{c \in C_g} \mu_{c,k}, \qquad \bar{\sigma}_{g,k} = \frac{1}{|C_g|} \sum_{c \in C_g} \sigma_{c,k}. \tag{2.2}$$

In stacked-bar figures, the total bar height is the sum of aggregated metric means,

$$H_g = \sum_{k \in \mathscr{K}} \bar{\mu}_{g,k}, \tag{2.3}$$

and the plotted error bar is an aggregate variability indicator computed as

$$E_g = \sum_{k \in \mathscr{K}} \bar{\sigma}_{g,k}. \tag{2.4}$$

2.10.1 OVERALL COMPARISON

For each model m, we aggregate over all configurations using that model (see equation (2.5)), i.e., $C_m = \{c : \text{Model}(c) = m\}$. The stacked bar segments correspond to $\bar{\mu}_{m,k}$ for each metric k, with total height and aggregate variability indicator:

$$H_m = \sum_{k \in \mathcal{K}} \bar{\mu}_{m,k}, \qquad E_m = \sum_{k \in \mathcal{K}} \bar{\sigma}_{m,k}. \tag{2.5}$$

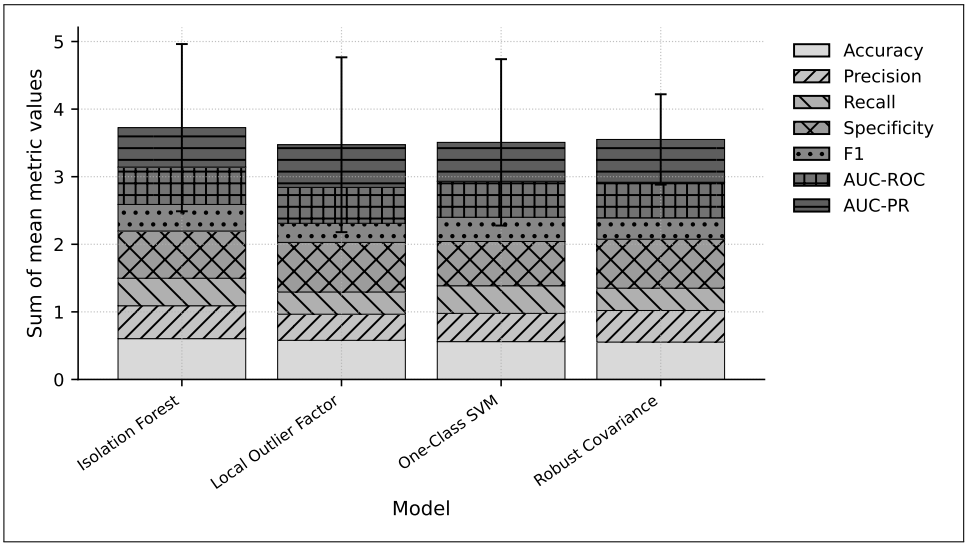


Figure 2.1 : Overall model performance

Across the three held-out test splits, Isolation Forest tended to achieve the highest aggregate performance metrics (Figure 2.1); however, the aggregate variability indicators overlap with those of *One-Class SVM* and *Local Outlier Factor* on several metrics. Figure 2.1 shows that the bar representing *Isolation Forest* is generally taller and its error bars (summed standard deviations) are comparatively moderate, indicating both high performance and relative stability. Models like *One-Class SVM* and *Local Outlier Factor* follow closely in certain metrics (e.g., *Recall*, *Specificity*), but they do not consistently outperform Isolation Forest across the entire

metric set. *Robust Covariance* in *Elliptic Envelope* trails slightly, yet remains a viable choice in specific parameter regimes.

2.10.2 COMPARISON PER ANOMALY CRITERION

For each anomaly criterion (feature) f, we aggregate over all configurations using that feature (see equation 2.6), i.e., $C_f = \{c : \text{Feature}(c) = f\}$. The stacked bar segments correspond to $\bar{\mu}_{f,k}$ for each metric k, and we plot:

$$H_f = \sum_{k \in \mathcal{K}} \bar{\mu}_{f,k}, \qquad E_f = \sum_{k \in \mathcal{K}} \bar{\sigma}_{f,k}. \tag{2.6}$$

We further broke down the analysis by individual anomaly criteria (or *features*), such as AverageProductPrice, PerfectOrderRate, and NumberDelaysGreaterThan30Days. Notably, NumberOfRequests (purchase request) emerges as the most consistent feature overall: it yields the best combination of average performance metrics (e.g., highest mean *F1*, *AUC-ROC*, *AUC-PR* and *Accuracy*) with respect to the expert reference labels, and the smallest aggregate variability indicator across triplicates. This stability suggests that NumberOfRequests (purchase requests) may be more predictable for anomalies in its own dataset. Although different features cater to different types of anomalies, results with NumberOfRequests (purchase requests) suggest this criterion can reliably enhance detection performance, regardless of the chosen model. (see Figure 2.2)

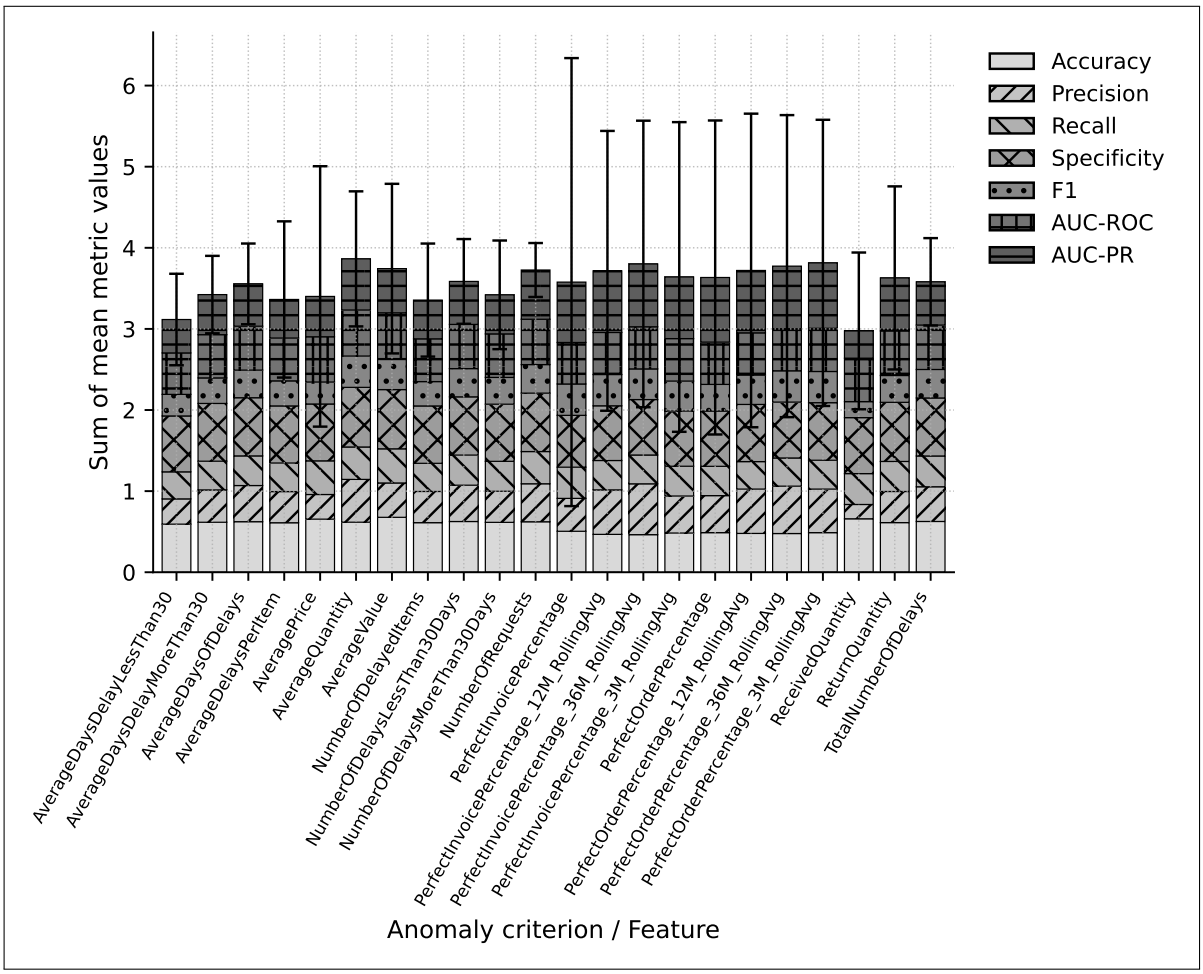


Figure 2.2 : Comparison of average performance with summed standard deviation per anomaly criterion

2.10.3 INFLUENCE OF POST-PROCESSING FILTERS

For each filter setting z and model m, we aggregate over configurations that match both (see equation 2.7), i.e., $C_{z,m} = \{c : \text{Filter}(c) = z \wedge \text{Model}(c) = m\}$. Within each filter panel, the stacked bar segments are $\bar{\mu}_{z,m,k}$ and we plot:

$$H_{z,m} = \sum_{k \in \mathscr{K}} \bar{\mu}_{z,m,k}, \qquad E_{z,m} = \sum_{k \in \mathscr{K}} \bar{\sigma}_{z,m,k}. \tag{2.7}$$

Figure 2.3 explores the effect of various post-processing filters (*default*, *excess*, *lack*, *inverse*) on the model's binary anomaly flags $\hat{y}$ and the resulting performance metrics computed against expert reference labels. Consistent with the overall metrics:

- **Excess Filter**: Consistently yields superior average performance metrics compared to the other filters and the no-filter (*default*) case. When the objective is to emphasize stronger outlier evidence (e.g., very high or low values), applying the *excess* filter can markedly improve metrics such as *F1*, *AUC-ROC*, *AUC-PR* and Accuracy.
- **Default Filter (No Filter)**: Often remains a close second, yet typically posts slightly lower performance than the excess filter.
- **Inverse** and **Lack Filters**: Each favor models like One-Class SVM, improving Recall, but they are less effective at maintaining balanced Precision and Specificity.

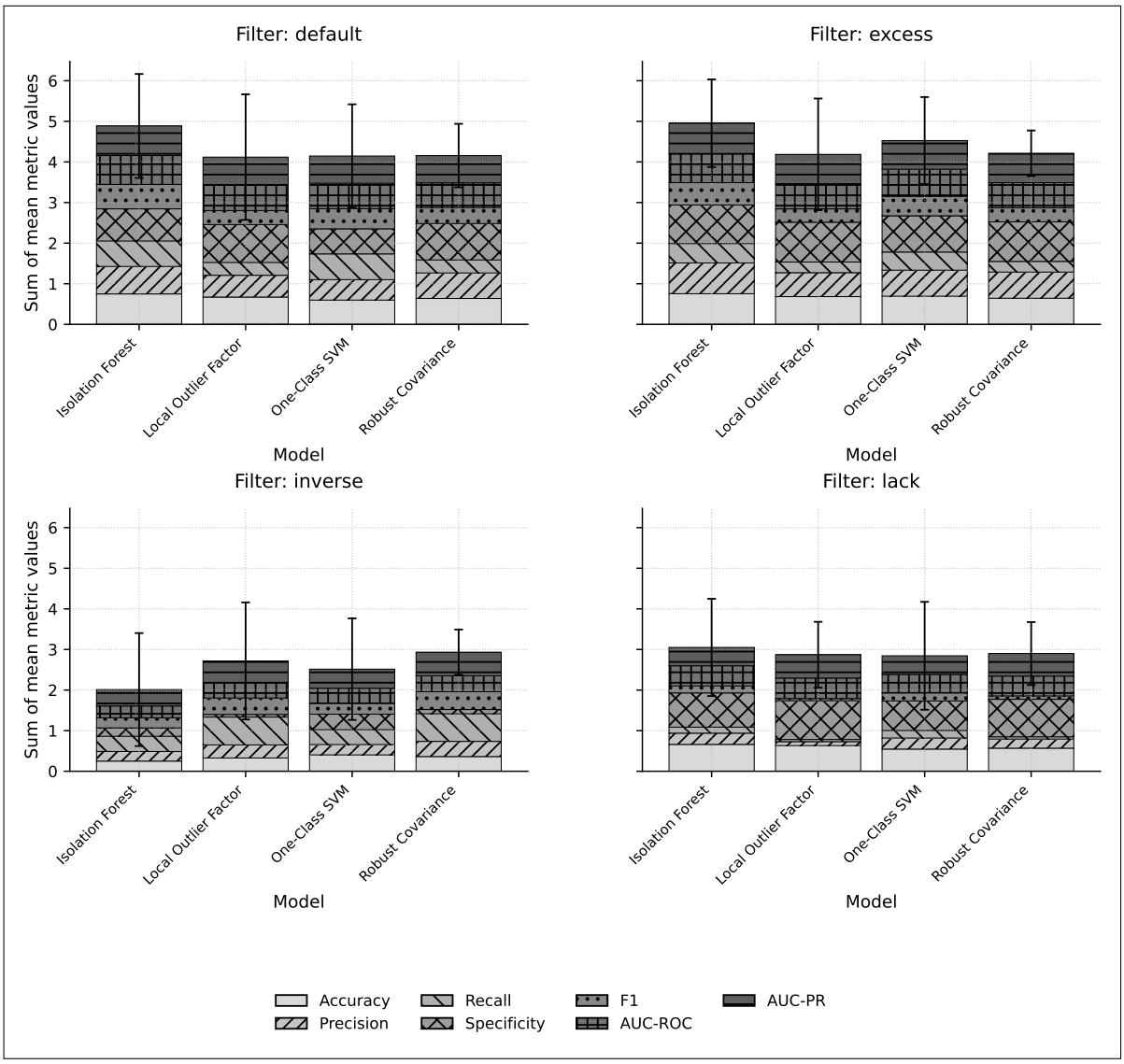


Figure 2.3 : Influence of post-processing filters on global performance

If the cost of false positives is tolerable and capturing strong anomalies is a priority then the excess filter appears to be the top choice overall.

2.10.4 MODELS PERFORMANCE IN SPECIFIC CONDITIONS

Equation 2.8 illustrates the ranking of individual configurations c. It's defined as global score of the average of the metric means:

$$G_c = \frac{1}{|\mathscr{K}|} \sum_{k \in \mathscr{K}} \mu_{c,k}. \tag{2.8}$$

For the two heatmap visualizations, we aggregate mean performance metrics over either *(configuration, model)* pairs or *(feature, model)* pairs. For the configuration heatmap, we group results by (g,m) with $C_{g,m} = \{c : \text{Config}(c) = g \wedge \text{Model}(c) = m\}$ and display the summed mean score:

$$I_{g,m} = \sum_{k \in \mathscr{K}} \bar{\mu}_{g,m,k}. \tag{2.9}$$

For the feature heatmap, we group results by (f,m) with $C_{f,m} = \{c : \text{Feature}(c) = f \wedge \text{Model}(c) = m\}$ and compute:

$$I_{f,m} = \sum_{k \in \mathscr{K}} \bar{\mu}_{f,m,k}. \tag{2.10}$$

Beyond global analyses, using the expert reference labels, we identified the top parameter configurations to pinpoint scenario-specific practices. We first added a *Performance_Global* score to each row (averaging *F1_mean*, *AUC_ROC_mean*, *AUC_PR_mean*, *Accuracy_mean*, *Precision_mean*, *Recall_mean*, and *Specificity_mean*) and then selected the highest-scoring rows. (see Figure 2.4 and 2.5)

A summary of best configurations highlights several recurring patterns:

- **Isolation Forest (IF)**: The strongest scores cluster under the *excess* filter, with a clear boost when Bayesian optimization is activated; the choice between PCA on/off and Standard vs Min–max preprocessing has a comparatively smaller effect.
- **Local Outlier Factor (LOF)**: Competitive performance is observed mainly with the *excess* filter (and slightly lower with *default*), while *inverse* and *lack* configurations trail behind; here, Bayesian optimization and PCA are secondary compared with the filter choice.
- **One-Class SVM (OC-SVM)**: Top configurations are again associated with the *excess* filter, frequently paired with Min–max preprocessing; Bayesian optimization often captures the peak, whereas PCA on/off plays a lesser role.
- **Robust Covariance (RC) in Elliptic Envelope (EE)** : High scores occur primarily when Bayesian optimization is enabled, most often with the *excess* filter (with *default* sometimes close); disabling Bayesian optimization generally leads to a noticeable drop, with PCA on/off exerting a smaller influence.
- **Feature-side signal**: Across models, quantity/value features (notably *AverageQuantity*, often also *AverageValue*) and rolling invoice/perfect-order rates tend to rank among the most informative; Local Outlier Factor's strongest feature rankings lean more toward delay-related variables and longer-window invoice rollups.

The feature–model heatmap (Figure 2.5) shows that *Isolation Forest* performs best on quantity/value anomalies and rolling quality KPIs, *Local Outlier Factor* is strongest on delay-related criteria, and *Robust Covariance* in *Elliptic Envelope* peaks on aggregate intensity signals (e.g., request volume/total delays). Read with the configuration heatmap (see Figure 2.4 and 2.5), these strengths most often occur with the *excess* filter and *Bayesian optimization* enabled, while PCA and preprocessing choices mainly fine-tune results.

These results confirm that there is no universally dominant parameter set for all contexts. Instead, practitioners should select model–filter–feature combinations based on the specific anomaly patterns, cost of false positives/negatives, and domain constraints.

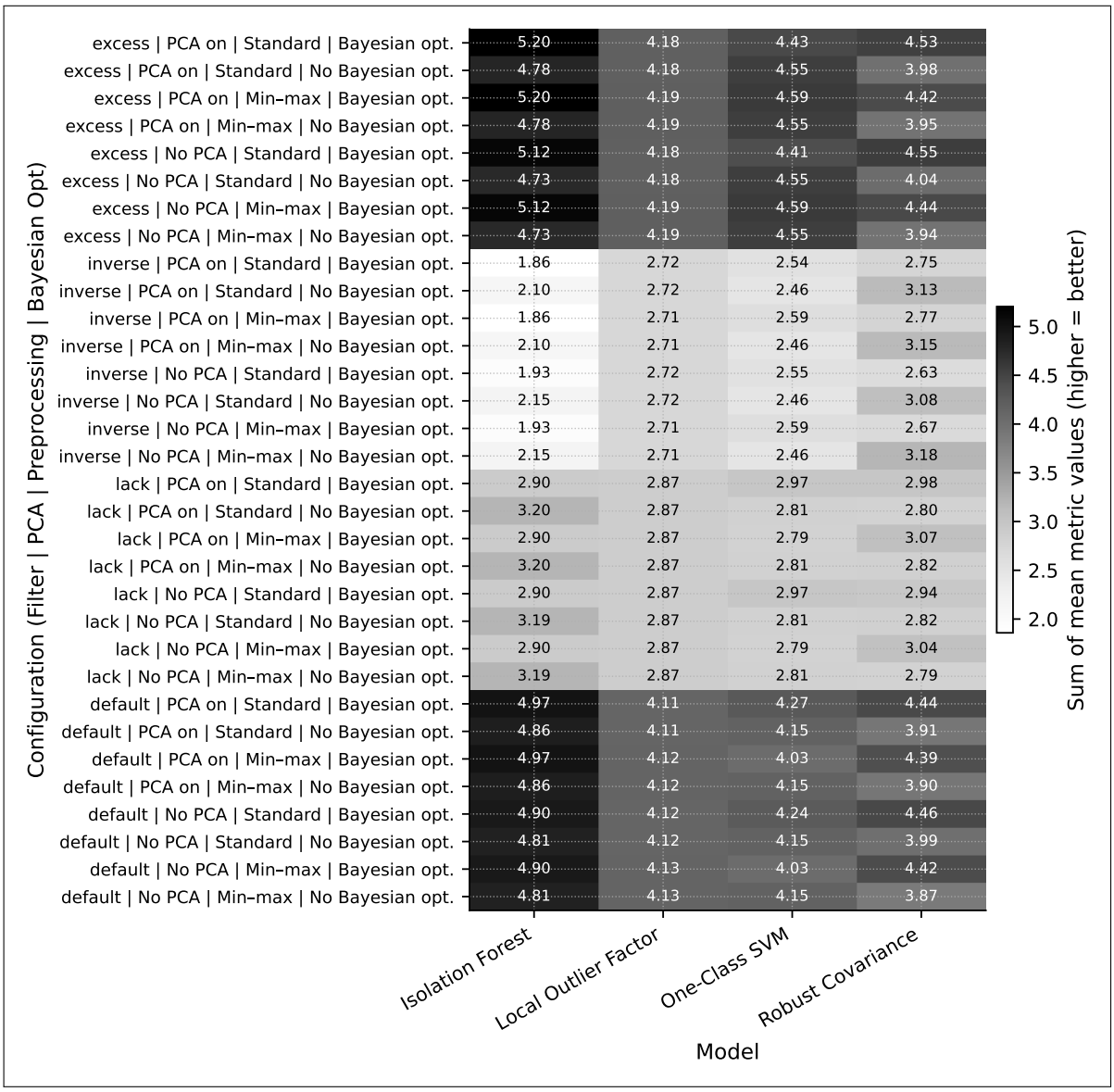


Figure 2.4 : Sum of averaged performance heatmap (configuration, model)

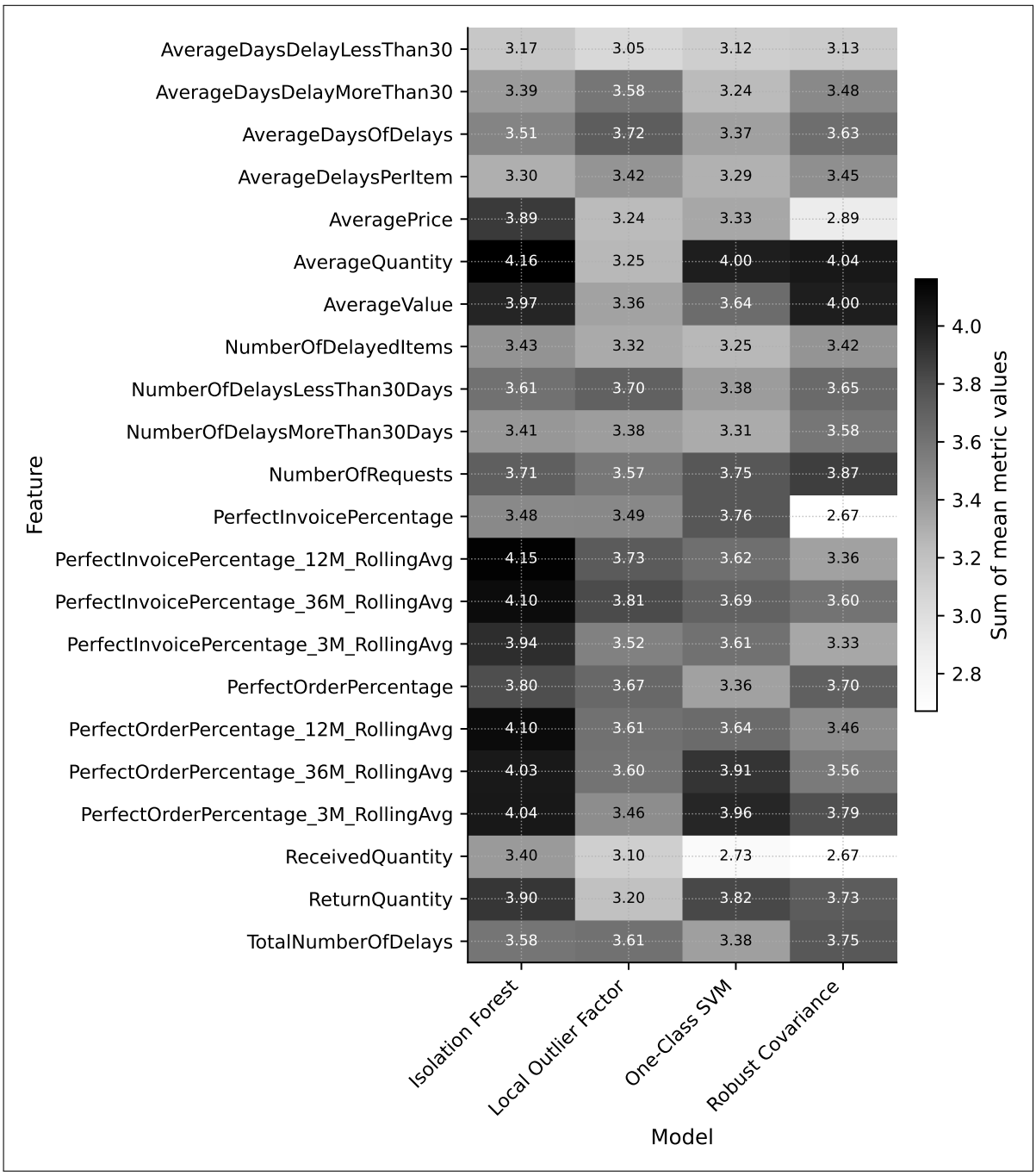
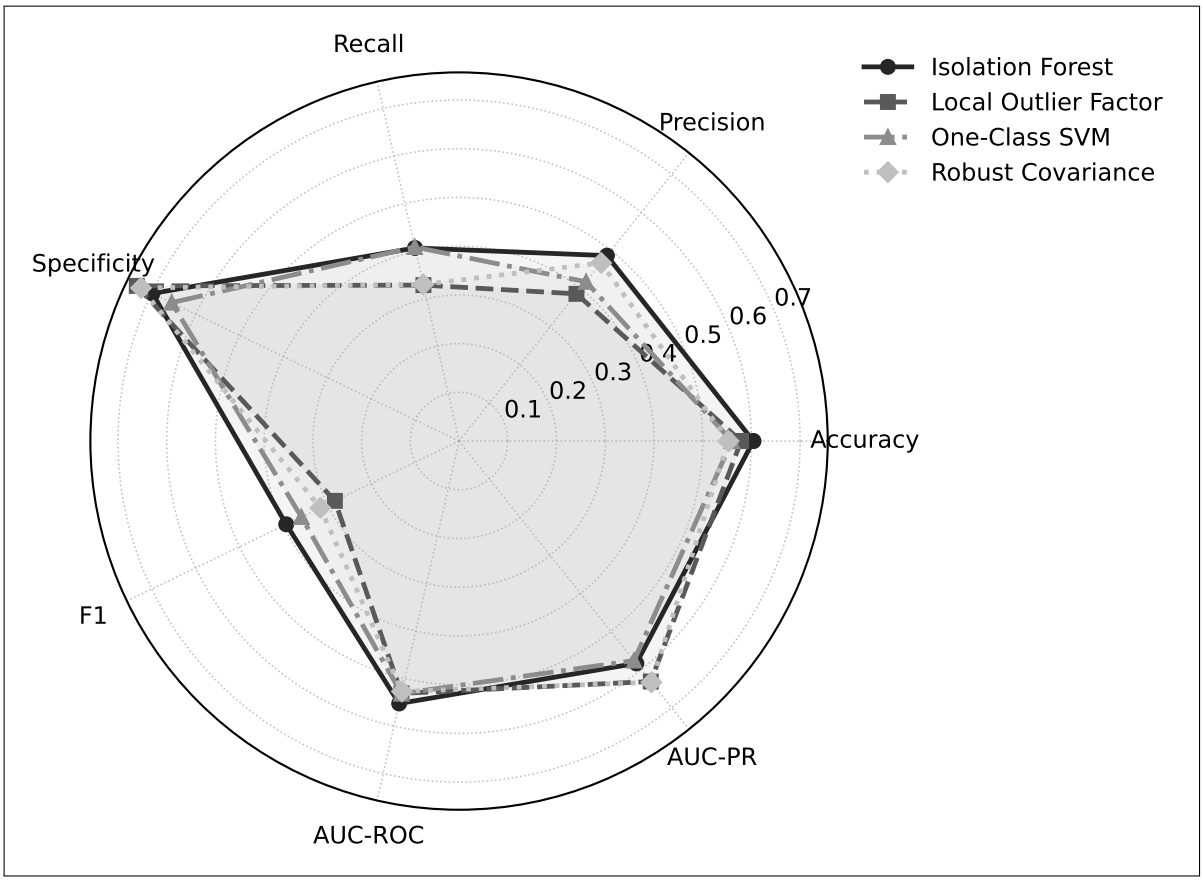


Figure 2.5 : Sum of averaged performance heatmap (feature, model)

2.10.5 THE IMPORTANCE OF THE MODEL

Equation (2.11) illustrates that for each model m, we compute the mean performance profile across all configurations using that model, $C_m = \{c : \text{Model}(c) = m\}$, and plot the vector of aggregated metric means:

$$R_{m,k} = \bar{\mu}_{m,k} = \frac{1}{|C_m|} \sum_{c \in C_m} \mu_{c,k}, \qquad \forall k \in \mathscr{K}. \tag{2.11}$$



© Colin Bouchard

Figure 2.6 : Average performance per model radar

Finally, a radar plot (Figure 2.6) visually contrasts the mean performance of all four models across *Accuracy, Precision, Recall, Specificity, F1*, *AUC-ROC*, and *AUC-PR*. As observed throughout:

- **Isolation Forest** covers a consistently large area in the radar, validating its top performance across most conditions.
- **Local Outlier Factor** and **One-Class SVM** can surpass *Isolation Forest* in narrowly targeted configurations, especially when the filter or feature selection heavily emphasizes certain anomaly traits (e.g., capturing extreme outliers or boosting Recall).
- **Robust Covariance in Elliptic Envelope** typically covers a smaller radar area, yet certain parameter synergies narrow the gap.

In conclusion, Isolation Forest (IF) emerges as a strong all-around model in this study, but the expert reference labels model–filter combination can vary across different features, filtering strategies, and optimization parameters. The choice of anomaly criterion (e.g., *NumberOfRequests* (purchase requests)) and post-processing filters (e.g., *excess*) has a decisive impact on overall accuracy and reliability, underscoring the importance of systematic experimentation and parameter tuning in anomaly-detection pipelines.

2.11 FUTURE DIRECTIONS SHOULD INCLUDE

- **Developing a *Spark*-based** framework that combines unsupervised machine learning with domain-informed post-processing filters, demonstrating reproducibility and indicating feasibility on moderate-scale ERP data.
- **Conducting a systematic error analysis of failures** observed during distributed execution to identify when *Spark*-native algorithms or alternative partitioning strategies are necessary.
- **Establishing best practices** for designing anomaly detection systems that remain robust under the constraints of distributed, high-dimensional data environments.

2.12 CONCLUSION

In the specific case studied here, artificial intelligence appeared to be a promising tool for anomaly detection in healthcare procurement data, suggesting that machine learning algorithms may complement existing practices and, under suitable conditions, contribute to improved operational efficiency and governance. These potential benefits must be viewed within the constraints of the semi-supervised evaluation framework used, where metrics reflect alignment with the judgment of a single domain expert. While Kumar et al. [28] explore the role of AI in healthcare supply chains before, their findings only suggested a possible benefit in spotting inefficiencies and suggest cost savings, warranting further study. Our findings align with prior research [182], highlighting that AI-driven anomaly detection can support data-driven decision-making and resource optimization in public sector procurement.

Given that all input variables significantly deviate from normality ($p<0.01$), making traditional parametric statistical baselines unreliable, the machine learning algorithms, particularly One-Class SVM and Local Outlier Factor, showed promising capabilities in identifying

anomalies related to pricing, delivery delays, and unusual purchasing behaviors by relying on non-parametric approaches. This aligns with [66], which demonstrated the advantages of unsupervised learning techniques in complex and dynamic environments. Furthermore, these performance metrics must be understood as quantifying the model's alignment with the single expert's interpretation of procurement risk within CIUSSS-SLSJ policies, acknowledging the inherent limitations imposed by this proxy ground truth on broader quantitative generalization.

Our study also underscores the importance of tailored post-processing filters in refining anomaly detection outputs. As outlined by [67], the integration of preprocessing and post-processing techniques can enhance anomaly detection accuracy, particularly in scenarios where subtle deviations may otherwise go unnoticed. The observed impact of Bayesian optimization on model performance further corroborates findings by [124], reaffirming the role of hyperparameter tuning in improving anomaly detection capabilities.

Moreover, this research has practical implications for public sector governance and policy compliance. Given the regulatory landscape defined by [4] and [5], the integration of AI-driven anomaly detection in healthcare procurement aligns with institutional objectives of transparency, accountability, and risk mitigation. As noted in [9], emerging legislative efforts aim to enhance efficiency in public administration, further reinforcing the relevance of AI applications in procurement oversight.

Future research should explore the scalability of these methods across broader datasets, integrating additional anomaly detection frameworks such as deep learning approaches [60] to further enhance predictive capabilities. Additionally, ethical considerations regarding AI adoption in public sector decision-making, as raised by [195], warrant further investigation to ensure equitable and responsible implementation.

In the specific context studied, AI-based anomaly detection appears capable of complementing manual validation by increasing speed and helping handle larger data volumes, but it does not replace the need for human oversight. Compared with purely manual methods, which are often slow, labour-intensive, and vulnerable to fatigue or inconsistency, AI systems can systematically scan large datasets and flag patterns that might be difficult for humans to detect. At the same time, these systems introduce their own risks (for example, dependence on data quality, model assumptions, or thresholds), and may reflect or amplify existing biases. As a result, AI should be viewed as a decision-support tool that can reduce manual workload and highlight candidate irregularities, while final judgments and responsibility remain with human experts. Although similar approaches are increasingly used in domains such as cybersecurity, finance, and manufacturing, the conclusions drawn in this thesis are limited to the healthcare procurement setting examined here and would need to be validated before being generalized to other industries.

Ultimately, this study contributes to the growing body of research advocating for AI-driven anomaly detection in healthcare supply chains. By leveraging advanced machine learning techniques and post-processing strategies, institutions could potentially improve procurement monitoring, mitigate financial risks, and uphold the integrity of public resource management, in line with contemporary healthcare governance initiatives [6].

CHAPTER III

AI-BASED ANOMALY DETECTION IN HEALTHCARE PROCUREMENT USING APACHE SPARK: A POST-PROCESSING FILTERS APPROACH

Colin Bouchard (Université du Québec à Chicoutimi)

Julien Maitre (Université du Québec à Chicoutimi)

Bruno Bouchard (Université du Québec à Chicoutimi)

Realised for

University of Québec at Chicoutimi / Université du Québec à Chicoutimi

In collaboration with the

Integrated University Health and Social Services Center of Saguenay-Lac-Saint-Jean / Centre intégré universitaire de santé et de services sociaux du Saguenay-Lac-Saint-Jean

December 15, 2025

3.1 ABSTRACT

Anomaly detection in public healthcare procurement can help identify inefficiencies, fraud, and operational risks. This paper re-engineers a prior *Pandas*-based, unsupervised anomaly detection pipeline in *Apache Spark* to support future large-scale deployments in centralized ERP settings across Quebec's healthcare system. Using SAP-derived procurement data labeled by CIUSSS-SLSJ domain experts, we evaluate four unsupervised algorithms: Robust Covariance (RC) in Elliptic Envelope (EE), Isolation Forest (IF), One-Class SVM (OC-SVM), and Local Outlier Factor (LOF). The workflow integrates preprocessing (normalization, standardization, PCA) and post-processing filters (default, inverse, excess, lack). Each modeling path is ported via Spark's `mapInPandas()` API, enabling partition-level execution of the original *Pandas* code without algorithmic changes. The migration highlights difficulties in distributing algorithms that rely on global computations, including partition sparsity effects, serialization overhead in Spark's Python API, and challenges in visualizing distributed outputs. Post-processing filters preserve their effects: the excess filter improves precision for high-magnitude anomalies, the lack filter helps detect demand or supply drops, and the inverse filter reveals overfitting in some models, notably Local Outlier Factor. Overall, results suggest that Spark can support AI-driven anomaly detection for moderate data sizes and suitable algorithms, but distance- and kernel-based methods (Local Outlier Factor and One-Class SVM) remain hard to scale and may require Spark-native or redesigned distributed implementations. In our setting, full dataset broadcasting was necessary to stabilize these global methods, indicating that true linear scalability on very large datasets remains an open challenge.

Keywords: Spark, anomaly detection, healthcare, public procurement, unsupervised learning, Isolation Forest (IF), One-Class SVM (OC-SVM), Robust Covariance (RC) in Elliptic Envelope (EE), Local Outlier Factor (LOF).

3.2 INTRODUCTION

Public healthcare systems are among the most complex and resource-intensive enterprises in modern economies [196]. They rely on intricate supply chains to ensure the timely delivery of critical goods and services, ranging from pharmaceuticals and medical devices to infrastructure and administrative materials [197]. Effective procurement management is essential not only for cost control but also for safeguarding patient safety and maintaining public trust [55]. However, these systems are inherently vulnerable to inefficiencies, fraud, waste, and abuse due to their scale, regulatory complexity, and the diverse ecosystem of suppliers and intermediaries involved [198]. As noted by prior studies, anomalies in procurement data can serve as early warning signals for these issues, enabling organizations to intervene proactively [122].

In recent years, AI and ML have emerged as powerful tools for enhancing anomaly detection in large datasets [199, 200]. Unsupervised ML algorithms, in particular, offer advantages where labeled data are scarce, which is often the case in public sector procurement audits [201]. Algorithms such as Isolation Forest, One-Class Support Vector Machine, Local Outlier Factor, and Robust Covariance (RC) in Elliptic Envelope (EE) have demonstrated success in identifying patterns and deviations that traditional rule-based systems may overlook [1, 202, 203]. Despite these promising developments, translating AI-driven anomaly detection pipelines from research prototypes to production-scale systems in public healthcare remains challenging [204, 205].

Our prior work presented a *Pandas*-based anomaly detection pipeline applied to SAP procurement exports from CIUSSS-SLSJ [129]. That study incorporated preprocessing (normalization, standardization, PCA), Bayesian hyperparameter optimization, and domain-informed post-processing filters (*default*, *inverse*, *excess*, *lack*). Experimental results on nine anomaly

criteria (price variation, delivery delays, demand fluctuations, etc.) showed that Isolation Forest (IF) with the *excess* filter best balanced precision, recall, and specificity, highlighting the potential for AI-enhanced monitoring in public procurement.

While those findings were encouraging, the single-node *Pandas* environment limited scalability. To explore portability and to assess practical scaling constraints, we re-engineer the pipeline within *Apache Spark* to evaluate whether a distributed architecture can preserve detection accuracy, interpretability, and robustness observed in the single-node prototype.

Deploying unsupervised anomaly detection algorithms in *Spark* presents non-trivial technical hurdles [206, 207]. Algorithms such as Local Outlier Factor [208, 209] and One-Class SVM [210] are inherently global: they rely on full distance matrices or kernel computations that assume access to the entire dataset [208, 209, 210]. In a distributed setting, partitions may not contain sufficient data to compute meaningful neighborhood statistics or support vector boundaries, leading to errors or degraded performance [211]. Even Robust Covariance in Elliptic Envelope can fail on partitions with limited cardinality, producing singular covariance matrices [212]. These issues are exacerbated by hyperparameter search, which multiplies model fittings across partitions [213]. These shortcomings have motivated distributed-friendly outlier detection designs, e.g., distributed Local Outlier Factor [209].

Contributions

1. A *Spark*-based framework that integrates unsupervised ML with domain-informed post-processing filters, illustrating potential reproducibility and suggesting probable feasibility on moderate-scale ERP data.

2. A systematic error analysis of failures encountered during distributed execution, surfacing when *Spark*-native algorithms or alternative partitioning strategies are required.

3. Best practices for designing anomaly detection systems robust to constraints of distributed, high-dimensional data environments.

3.3 OBJECTIVE OF THIS STUDY

This study aims to operationalize AI-based anomaly detection in healthcare procurement by porting a *Pandas*-based pipeline to *Apache Spark* to evaluate backend reproducibility and to characterize scaling constraints. Building on prior work using Isolation Forest (IF), Robust Covariance in Elliptic Envelope, One-Class SVM , and Local Outlier Factor on CIUSSS-SLSJ data [129], we address portability, reproducibility, and operational constraints in distributed environments. The primary objective is to assess whether *Spark*'s architecture can support these algorithms without sacrificing performance or interpretability. Secondary goals include integrating domain-informed post-processing filters (*default*, *inverse*, *excess*, *lack*) and documenting implementation challenges. This contributes theoretical insights on portability/reproducibility and on scaling constraints of classical models in *Spark* and a prototype framework for near-real-time monitoring in the CIUSSS-SLSJ context (and similar settings), subject to infrastructure and integration constraints.

Importantly, while we investigate operationalization within *Spark*, the underlying models remain fundamentally single-node in computational requirements. Results therefore reflect datasets amenable to in-memory processing and do not demonstrate *Spark*'s full distributed capabilities for truly large-scale data.

3.4 RESEARCH PROTOCOL AND METHODOLOGY

To enable rigorous comparison with earlier work [129], we retained the same dataset structure, labeling method, and preprocessing logic. Datasets include six SAP-derived procure-

ment dimensions managed by CIUSSS-SLSJ: monthly demand volume, supplier performance (e.g., average delivery delay), product return ratios, pricing inconsistencies, and received quantity anomalies. Each dataset contains timestamps and binary anomaly labels provided by a single Administrative Process Specialist.

Unlike the *Pandas* workflow, the current implementation uses *Spark*'s distributed capabilities via *mapInPandas()*. This executes arbitrary *Pandas* logic within partitions, keeping algorithmic flexibility while enabling parallelism. Each job is defined by a configuration tuple: (feature, model, preprocessing, PCA usage, filter type). Configurations are executed in batches across partitions, producing structured, reproducible outputs.

For each configuration, we apply preprocessing (standardization, normalization, or none), optional PCA (retaining 90% variance), unsupervised model fitting, prediction, and post-processing (default, excess, lack, inverse filters). We did not perform new hyperparameter tuning; we reused previously identified optimal values [129] to reduce cost and streamline execution.

Outputs include labeled predictions, *confusion matrices*, *AUC-ROC* and *PR curves*, *F1-scores*, and time-series plots, with file structure and logging aligned to the original study for comparability.

3.5 ANOMALY DETECTION ALGORITHMS IN *APACHE SPARK* FOR THIS STUDY

We evaluate four models: *Robust Covariance* in *Elliptic Envelope*, *Isolation Forest*, *One-Class SVM*, and *Local Outlier Factor*, across combinations of feature scaling, PCA, and filtering strategies. To preserve configuration independence and enable parallel evaluation, we run experiments in a distributed manner using *Spark* partitions.

Spark introduced constraints for LOF and OC-SVM, which depend on full distance matrices or support vectors [207]. To stabilize performance, we broadcasted the dataset and parallelized configurations [209, 214]. Retaining these models exposes the limits of *Spark* compatibility for traditional unsupervised algorithms.

3.6 ANOMALY CRITERIA FOR THIS STUDY

We adopt the same criteria used in the original article *Anomaly detection using AI in healthcare procurement data across multiple criteria with tailored post-processing filters* [129], ensuring methodological continuity and comparability of results across both works.

3.7 TAILORED POST-PROCESSING OUTPUT FILTERS

We reuse the four post-processing filters introduced previously [129]. *Default* preserves raw outputs. **Inverse** flips anomaly classifications to probe sensitivity to the decision threshold. *Excess* retains anomalies above a 30-day trailing moving-average threshold (emphasizing spikes). *Lack* retains anomalies substantially below the moving average (emphasizing drops). These filters reduce false positives and focus attention on context-relevant deviations.

3.8 EVALUATION OF THE MODELS

As in the *Pandas*-based study [129], the *Spark* experiments train the anomaly detection algorithms in an unsupervised way but evaluate their outputs in a supervised fashion against the expert-provided anomaly labels, so that the overall setting corresponds to unsupervised detectors with semi-supervised evaluation.

We compute threshold-dependent metrics (*accuracy, precision, recall, specificity, and F1-score*) from binary flags $\hat{y}$, and ranking metrics (*AUC-ROC* and *AUC-PR*) from continuous

anomaly scores $s(x)$. Confusion matrices and detection timelines support qualitative assessment [96, 194]. Confusion matrices and detection timelines support qualitative assessment. *Spark* adds parallel execution of configurations, automatic artifact storage in a structured filesystem, and embedded metadata for configuration-level comparisons. Its partitioned execution enables hundreds of configurations per dataset (beyond single-machine feasibility), broadening coverage of parameter/filter interactions while preserving comparability to the original setup.

3.9 RESULTS

We compare the traditional *Pandas* implementation [129] against the *Spark*-based implementation. Results are presented as overall metrics, per-feature performance, filter effects, parameter-specific behaviors, and model-wise summaries. All figures are generated using the same implementation as in prior work [129].

3.9.1 OVERALL COMPARISON

Table 3.1 summarizes the average performance across features and models; Threshold-dependent metrics (*accuracy, precision, recall, specificity, and F1-score*) are computed from the binary anomaly flags $\hat{y}$ produced by `predict()`, whereas *AUC-ROC* and *AUC-PR* are computed from the continuous anomaly scores $s(x)$ (higher $s(x)$ indicates more anomalous). Without re-running *Bayesian optimization* (reusing the previously optimized hyperparameters), performance differences are modest: accuracy increases from 0.579 to 0.583, precision from 0.429 to 0.442, while recall slightly decreases from 0.371 to 0.359, with F1 remaining effectively unchanged ($0.3358 \rightarrow 0.3361$). Overall, this case study indicates that the *Spark* implementation achieves aggregate performance comparable to the *Pandas* version even without additional optimization.

Table 3.1 : Overall metrics comparison between *Pandas* and *Spark*.

Metric	Pandas	Spark	% Diff.	±5%
Accuracy	0.580	0.584	+0.7	Yes
Precision	0.430	0.443	+3.0	Yes
Recall	0.371	0.359	-3.3	Yes
F1-Score	0.336	0.336	+0.1	Yes
Specificity	0.700	0.706	+0.9	Yes
AUC-ROC	0.536	0.533	-0.6	Yes
AUC-PR	0.594	0.583	-2.0	Yes

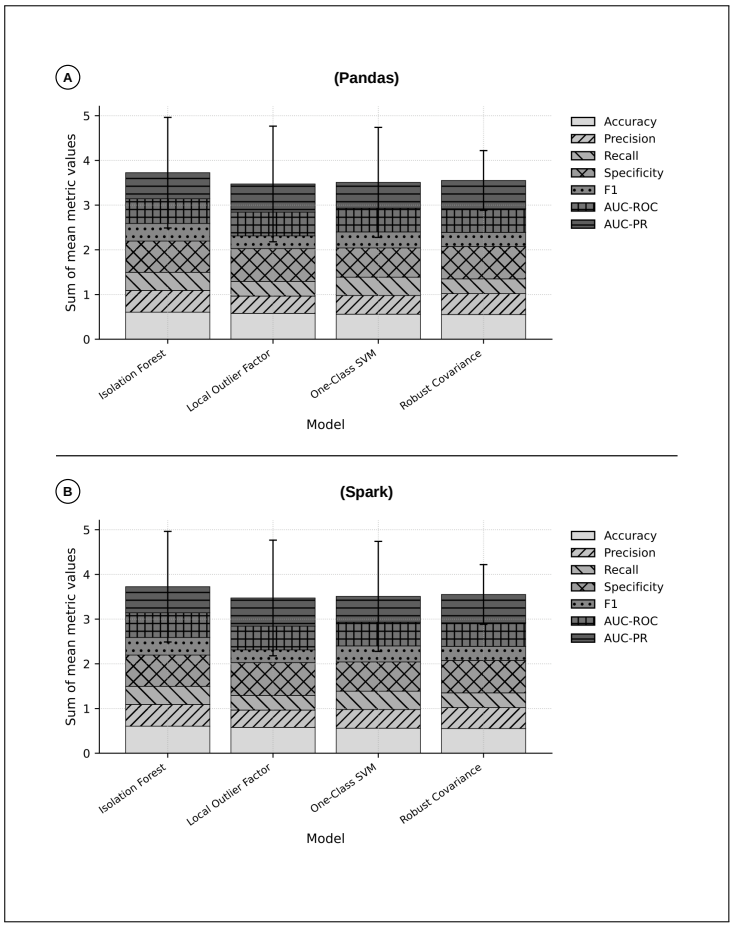


Figure 3.1 : Overall metrics comparison between (top [A]: *Pandas*, bottom [B]: *Spark*)

The stacked-bar comparison across seven metrics (top [A]: *Pandas*, bottom [B]: *Spark*), with SD error bars, indicates marginally higher aggregate performance for *Pandas*, particularly for *F1*, *AUC-ROC*, and *AUC-PR*, while *Spark* exhibits reduced variability. The relative ordering of methods is consistent across implementations: Isolation Forest outperforms Local Outlier Factor and One-Class SVM, which in turn exceed Robust Covariance in Elliptic Envelope.

3.9.2 COMPARISON PER ANOMALY CRITERION

Figure 3.2 (top [A]: *Pandas*, bottom [B]: *Spark*) shows average performance per feature. *NumberOfRequests (purchase requests)* remains strongest under *Spark* (highest *F1*, *AUC-ROC*, stability), reinforcing [129]. Transitioning from *Pandas* to *Spark* shows criterion-specific shifts that look even more nuanced when considering *dataset broadcasting* and the lack of *Bayesian optimization* in the *Spark* setup. *AverageProductPrice* records a strong precision gain (+50.90 %) with improved accuracy (+4.93%) and F1 (+8.91%), but recall drops (-9.12%), suggesting *Spark* may be operating with more conservative or non-re-optimised thresholds for price anomalies. *AverageOrderedQuantity* is the most uniformly improved signal, rising in precision (+13.67%), recall (+6.98%), and F1 (+17.10%), which may indicate this criterion is less sensitive to tuning gaps and benefits more directly from *Spark*'s scalable aggregation. By contrast, *AverageDaysLate* declines in precision (-12.26%) and F1 (-8.41%) with only a small recall change (-2.91%), implying that broader lateness aggregates might be more affected by distributed execution details or suboptimal parameter transfer from the *Pandas* pipeline. The windowed lateness criteria remain comparatively stable: *AverageLateLess30D* shows a mild precision increase (+5.65%) offset by recall reduction (-3.25%), while *AverageLateMore30D* improves precision (+16.07%) and F1 (+7.83%) with only a minor recall dip (-1.46%). Overall, broadcasting likely helped preserve consistency in join or baseline-dependent features, but the

absence of Bayesian optimization in *Spark* may have limited recall and hurt the more diffuse *AverageDaysLate* signal, even as most other criteria benefit from improved precision.

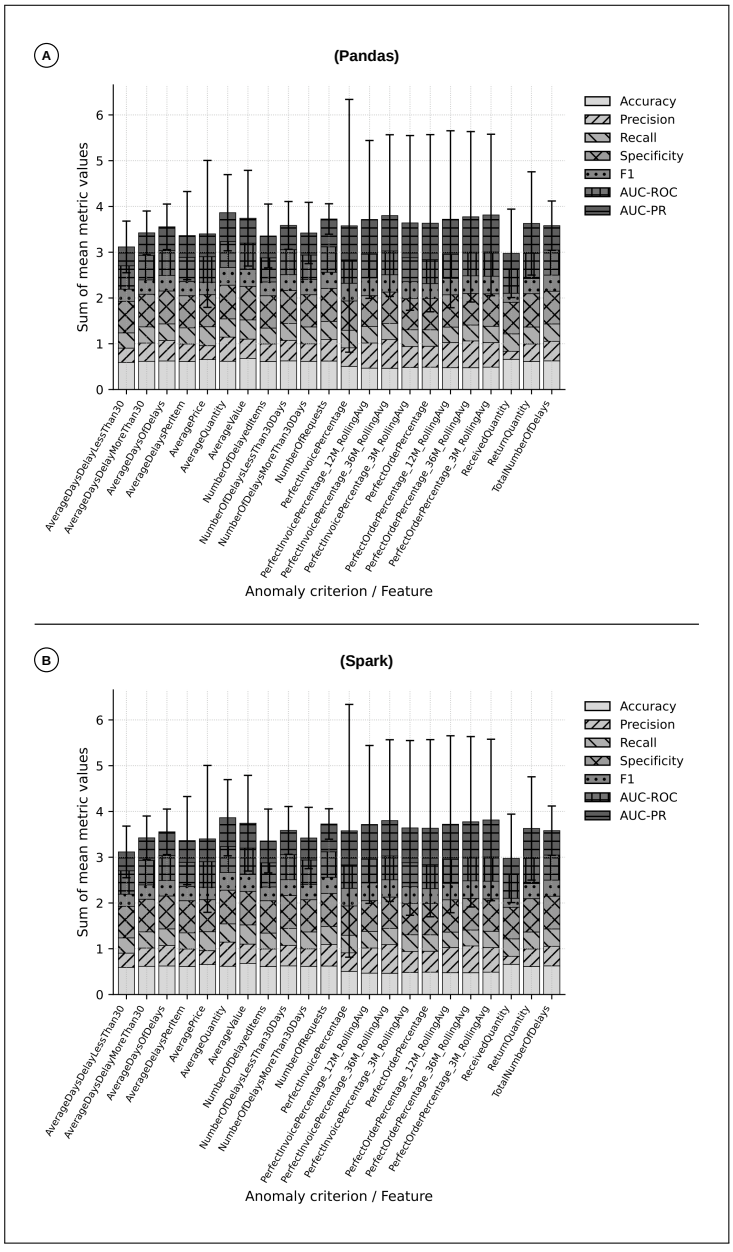


Figure 3.2 : Comparison per anomaly criterion between (top [A]: *Pandas*, bottom [B]: *Spark*)

3.9.3 INFLUENCE OF POST-PROCESSING FILTERS

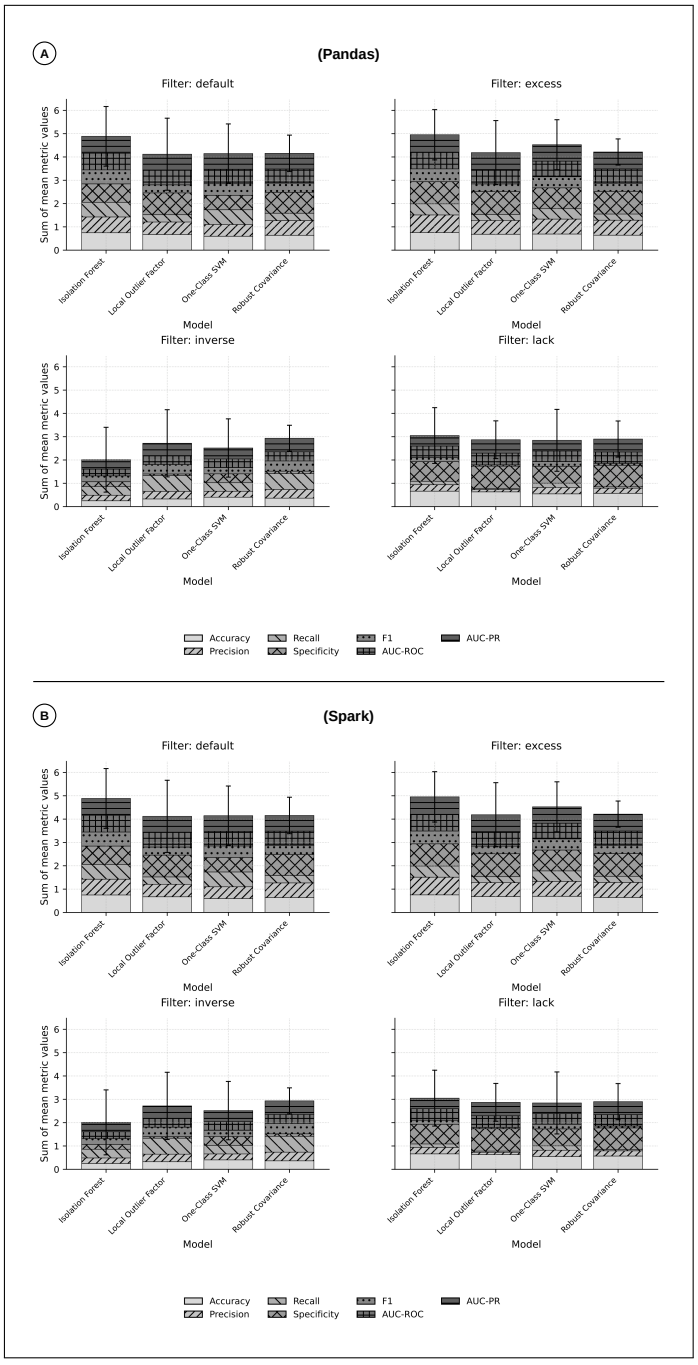


Figure 3.3 : Influence of post-processing filters between (top [A]: *Pandas*, bottom [B]: *Spark*)

Figure 3.3 (top [A]: *Pandas*, bottom [B]: *Spark*) shows that, in our experiments, *Excess* filter tended to increase F1, *AUC-ROC*, and *AUC-PR*, often $> 10\%$ over default (especially for *AverageProductPrice*). *Inverse* and *Lack* are less stable without tuning (higher false positives).

3.9.4 MODELS PERFORMANCE IN SPECIFIC CONDITIONS

Among the top 25 configurations, *Spark* favors moderate PCA (3–8 components), standard or min–max scaling, and the *Excess* filter, matching [129]. Even without new Bayesian optimization, Isolation Forest often achieves perfect scores (1.00) for *NumberOfRequests (purchase requests)* and *AverageOrderedQuantity* under *Excess*.

3.9.5 THE IMPORTANCE OF THE MODEL

Figure 3.4 (top [A]: *Pandas*, bottom [B]: *Spark*) summarizes model-wise performance (radar). *Isolation Forest* generally performed best across the evaluated metrics and configurations, balancing precision/recall, particularly under *default* and *excess*. *One-Class SVM* and *Local Outlier Factor* retain recall strengths (notably with *Lack*) but drop precision without tuning. *Robust Covariance* in *Elliptic Envelope* remains least consistent, with minor gains.

The radar plots suggest that the Figure 3.4 [A] *Pandas* (top) and [B] *Apache Spark* (bottom) pipelines yield broadly consistent performance profiles across the four models. In both cases, specificity and *AUC-ROC* appear comparatively strong while *F1* and *recall* remain modest, indicating a similar trade-off between capturing anomalies and avoiding false positives. The *Spark* chart looks like a close reproduction of the *Pandas* pattern with only minor shifts in the mid-range metrics (e.g., small changes in *precision–recall* and *AUC-PR*), implying that the relative model behavior is largely stable across execution engines rather than being sensitive to the pipeline technology.

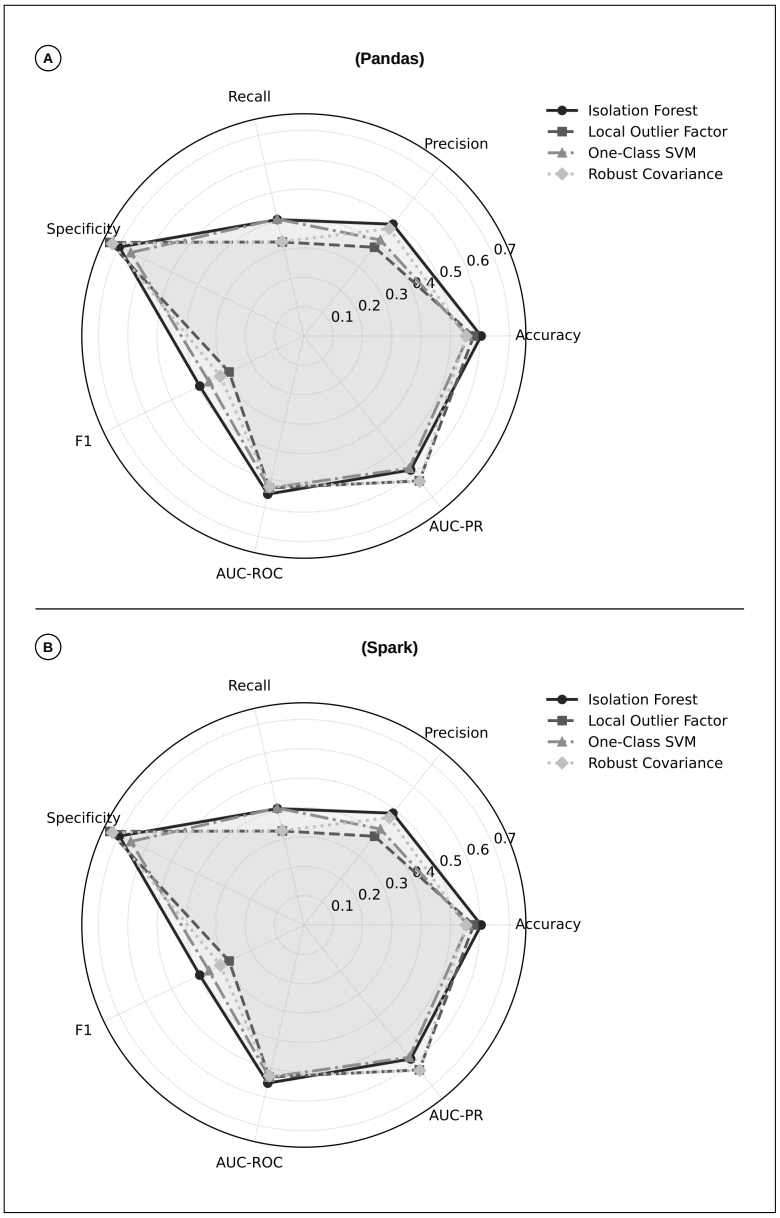


Figure 3.4 : Importance of the models across metrics between (top [A]: *Pandas*, bottom [B]: *Spark*)

A more model-wise read of the radar plots shows strong consistency between the Figure 3.4 [A] *Pandas* (top) and [B] *Spark* (bottom) pipelines. *Isolation Forest* appears to retain one of the most balanced profiles across both engines, with relatively solid accuracy,

precision, and *AUC-ROC*, while still sharing the overall pattern of high specificity and modest *recall* and *F1*. *Local Outlier Factor* looks very close to *Isolation Forest* in both pipelines, with only slight visual shifts in precision and *AUC-PR*, suggesting minimal sensitivity to the execution framework. *One-Class SVM* seems to be the most *recall* and *F1*-challenged in both charts (a smaller extension on the *F1* axis in particular), even though its specificity remains strong, indicating a more conservative anomaly flagging behavior regardless of pipeline. *Robust Covariance* in *Elliptic Envelope* also tracks similarly across *Pandas* and *Spark*, often matching or slightly edging others on precision and *AUC-PR* while preserving the same high-specificity, mid-*AUC* shape, which reinforces that the relative strengths and weaknesses of each algorithm are largely stable across the two implementations.

3.10 IMPLEMENTATION CHALLENGES

Partition-level application of *Local Outlier Factor* and *One-Class SVM* caused errors (e.g., *singular matrices, insufficient neighbors*) and inconsistent results due to limited local context. Serialization/pickling of models between workers complicated debugging (scattered worker logs).

3.10.1 DATASET BROADCASTING

To stabilize, we broadcasted the full dataset so each worker operated while parallelizing across configurations. This replicates *Pandas* results functionally but limits scalability and diverges from *Spark*'s intended partitioned model, reinforcing the need for partition-aware or *Spark*-native algorithms.

3.11 LIMITATIONS

Table 3.2 summarizes key limitations: reliance on broadcasting, single-threaded intra-model computation, lack of distributed hyperparameter optimization, and validation only on moderate-sized datasets.

Table 3.2 : Key limitations of the current approach.

Limitation	Description
Full dataset broadcasting	Reduces scalability and increases memory usage; prevents true distributed processing.
No distributed hyperparameter tuning	Limits optimization and model performance at scale.
Intra-model computation is not parallelized	Only orchestration is parallel; model fitting remains single-threaded.
Tested only on moderate data sizes	Scalability to very large datasets is not demonstrated.

3.12 FUTURE RESEARCH DIRECTIONS SHOULD INCLUDE

- Designing data ingestion pipelines that exploit partition-aware transformations without requiring full data replication.
- Adding distributed hyperparameter optimization comparable to *Bayesian* strategies used in prior work [129], and integrating *Spark*-compatible HPO frameworks.
- Exploring fine-grained intra-model parallelization to better utilize cluster resources during training and inference.
- Evaluating the approach on datasets that are orders of magnitude larger to validate scalability and robustness.

- Investigating algorithms that natively support partitioned data and *Spark*-native distributed execution to realize *Spark*'s full potential for real-time, scalable anomaly detection in high-volume healthcare procurement.

3.13 CONCLUSION

We reproduced the anomaly detection framework of the previous study [129] within *Apache Spark*, broadly comparable results with negligible deviations (non-deterministic Floating-Point arithmetic; no Bayesian tuning) [see Table 3.1 and Figures 3.1, 3.2, 3.3 and 3.4]. However, broadcasting the full dataset prevented exploitation of partitioning and shuffle-aware processing. Thus, while functional equivalence was achieved, claims about *Spark* scalability versus the single-node pipeline cannot be inferred.

Moreover, parallelism was limited to orchestration across independent models; intra-model computation remained effectively single-threaded. Performance gains were bounded by the number of concurrent configurations, not cluster scale, limiting generalizability to truly large, distributed scenarios.

From a reproducibility standpoint, the methodology is robust across computational backends. From a systems perspective, naive broadcasting and lack of intra-model parallelism contradict *Spark*'s distributed model when datasets exceed single-node memory.

Overall, the *Spark* experiments provide potential evidence for the reproducibility of the *Pandas* with *scikit-learn* results on the available infrastructure, but do not, in their current form, establish clear scalability gains for very large datasets. Therefore, while the *Spark*-based prototype confirms that the anomaly detection logic can be technically ported to a distributed framework, it does not yet demonstrate linear scalability for very large datasets, particularly due to reliance on dataset broadcasting and limitations of some classical algorithms.

CHAPTER IV

DISCUSSION

The present research, structured around the articles *Anomaly detection using AI in healthcare procurement data across multiple criteria with tailored post-processing filters* [129] and *AI-based anomaly detection in healthcare procurement using Apache Spark: a post-processing filters approach* [130], highlights a practical methodological advance in applying artificial intelligence to anomaly detection in healthcare procurement data, with a focus on domain-aware post-processing and scalable implementation.

4.1 EXPLORATORY DATA ANALYSIS

The research began with an in-depth exploratory analysis of the data, an essential step to evaluate their structure, distribution, and main statistical characteristics. This initial phase aimed to better understand the nature of the variables, verify their consistency, and identify possible peculiarities that could influence subsequent analytical methods. Detailed results of this analysis are presented in the appendix, along with tables and visualizations.

Initial descriptive statistics revealed wide variability among the studied indicators. Maximum values in some variables, such as average price or purchase value, indicated extreme cases, while others, such as the number of returns, showed limited dispersion. These discrepancies underline the heterogeneity of the measured phenomena and justify particular caution in the subsequent anomaly detection stage.

Distribution tests confirmed deviations from normality. The *Shapiro-Wilk* test showed that all variables significantly departed from a normal distribution (p-values < 0.01). This pattern may indicate that it would be prudent to consider analytical approaches that are less sensitive

to departures from normality [215]. Several commonly used classical procedures, including Z-scores and *Grubbs' test*, are derived under an assumption of at least approximate normality, and when that assumption is weak, unverifiable, or plausibly violated, their calibration can degrade in ways that may affect false-positive rates and the stability of the resulting inferences [216, 217, 218].

The exploratory study also revealed correlation and redundancy phenomena. *Singularity tests (matrix rank analysis)* showed that delay-related variables presented a singular structure, suggesting strong interdependence. This raises the risk of *multicollinearity*, affecting regression analyses or other sensitive methods [219]. Multivariate analysis confirmed this with perfect determination coefficients (R^2) for several delay variables, reflecting near-exact correlation. These findings require cautious interpretation.

Further, *ellipticity tests* distinguished two profiles: operational variables (price, purchase value, number of requests) with moderate ellipticity risk, and delay variables with much higher risk, reflecting structural complexity, asymmetry, and *kurtosis*.

Finally, temporal analysis confirmed the data's relevance for dynamic analyses. *Autocorrelation functions* revealed regularities, enabling predictive models based on temporal dependence [220]. *Augmented Dickey-Fuller tests* showed all variables were stationary, a major advantage for forecasting tools [221, 222]. Although not explored in this thesis, the observed *stationarity* could be leveraged in future forecasting-oriented studies.

Overall, the exploratory study identified both strengths and limitations of the data, clarifying their distribution, interdependencies, and temporal behavior. These findings guided the choice of analytical methods and formed the basis for anomaly detection techniques in the next stage.

4.2 STATISTICAL ANOMALIES AND AUTOMATIC LABELING

Before manual labeling by an Administrative Process Specialist from CIUSSS-SLSJ, a preliminary step of *statistical anomaly detection* was performed to ensure data coherence and reliability. This preparatory phase aimed to identify atypical values, either from entry/administrative errors or rare but valid cases, that could undermine model robustness. The goal was to provide an objective baseline from *quantitative methods* and test the feasibility of automatic labeling before expert validation; to avoid anchoring effects, the expert assessment was conducted without access to automatic labels.

To achieve this, classical anomaly detection methods were applied. *Distance-based outlier detection* highlighted isolated observations. *Z-Score*, though poorly adapted due to non-normality [2], allowed identification of extreme values by standard deviation. *Grubbs' test*, despite the same limitation [3], confirmed outliers within assumed normal distributions. The joint application of these techniques provided a first quality filter essential for later stages. As stated before, the *Shapiro-Wilk* results provide strong evidence of non-normality across all input variables ($p<0.01$), which makes parametric baseline detectors that are calibrated under normality assumptions, such as *Z-score* thresholds and *Grubbs' test*, less defensible in this setting [2, 3, 216, 217, 218]. In response, this study prioritizes the non-parametric ML methods applied here (IF, LOF, OC-SVM, RC), which do not require a Gaussian data model, while noting that their reliability still depends on data structure and appropriate tuning, thereby offering a more principled methodological step beyond conventional monitoring approaches.

Anomalies detected this way generated an initial set of automatic labels, but these labels were used strictly for exploratory testing and were not involved in the training, validation, or evaluation of the learning models presented in Chapters 2 and 3. Despite their limitations under non-normality, *Z-score*, *Grubbs' test*, and *distance-based anomaly detection* helped to

provided a rough indication of how classical statistical methods behave on the CIUSSS-SLSJ data. In this thesis, however, these automatically generated labels are treated purely as a non-diagnostic tool and all reported performance results rely exclusively on expert-provided labels. Also, as previously mentioned, expert did not see these automatic labels neither before or after their own labeling. Consequently, the automatic labels should be interpreted with caution and viewed as part of an exploratory phase, not as a substitute for expert annotation or as a basis for the conclusions drawn in the two articles. From a learning-theory perspective, the core anomaly detection algorithms therefore remain unsupervised, but they are evaluated and calibrated against this expert-derived expert-provided proxy labels. In other words, the studies reported in Chapters 2 and 3 operate in a semi-supervised evaluation framework: the detectors do not use labels to learn, yet their success is quantified by how well they match the expert notion of an anomaly.

The four retained models, *One-Class SVM*, *Isolation Forest*, *Robust Covariance* in *Elliptic Envelope*, and *Local Outlier Factor*, showed promising performances even with fully automated labels. This validated the methodological approach and justified moving to expert manual labeling, consolidating data quality for subsequent analyses.

This human validation led to the two articles [129, 130] mentioned, demonstrating both practical and academic value.

4.3 FROM HYPERPARAMETERIZATION TO POST-PROCESSING FILTERS

The initial study aimed to test Bayesian optimization of hyperparameters for improving unsupervised models. Results confirmed its benefits but also showed high computational costs and methodological complexity, limiting scalability.

Unexpectedly, post-processing filters, especially the *Excess* filter, often outperformed *hyperparameter optimization* gains. This shifted the research towards leveraging contextualized business-inspired filtering rules to increase operational value.

This methodological shift does not invalidate *hyperparameter optimization*; it refines how we position it. In our context, optimization is well-suited to pushing peak performance, but it can also introduce additional effort and operational complexity when moving toward production. In contrast, post-processing filters are parsimonious, interpretable, robust, and better suited for continuous monitoring systems.

4.4 TRANSITION TO A DISTRIBUTED FRAMEWORK

The follow-up study tested this approach in a distributed *Apache Spark* environment, evaluating feasibility at scale in centralized *ERP* systems. *Bayesian optimization* was abandoned in favor of reusing previously identified optimal parameters. This reduced computation load while testing filter robustness in a different technical setting.

Results showed no degradation, metrics like F1-Score, AUC-ROC, and AUC-PR remained stable, with reduced variance between partitions. Filters' stability compensated for the absence of dynamic optimization. However, some algorithms (LOF, OC-SVM) showed scalability limits, requiring workarounds like dataset broadcasting.

4.5 ROBUSTNESS AND MODEL LIMITATIONS

Across the two studies, Isolation Forest appeared to be the most stable and competitive algorithm in the tested single-node (*Pandas* with *scikit-learn*) and distributed (*Spark*) configurations for CIUSSS-SLSJ data. *Local Outlier Factor* and *One-Class SVM*, while theoretically

interesting, proved sensitive to technical constraints. *Robust Covariance* in *Elliptic Envelope* lagged except when combined with strong dimensionality reduction.

PRACTICAL SCOPE AND IMPLICATIONS

Together, the two articles illustrate the progression from proof-of-concept to pragmatic operational deployment. The first provided evidence for the feasibility of post-processing filters in controlled settings, while the second suggested that these filters can remain effective in distributed environments, which may help inform future real-world deployments.

Nonetheless, *Spark* migration still relied on methodological compromises like dataset broadcasting, limiting scalability gains. The studies underline the need for *Spark*-native algorithms adapted to distributed systems.

LIMITATIONS AND PERSPECTIVES

The research has limits, notably the narrow data context (SAP from CIUSSS-SLSJ only) and limited expert labeling, which restrict generalizability. The focus on four classical unsupervised algorithms also excluded deep learning approaches such as autoencoders or GANs. Because model selection and performance metrics are based on labels from a single Administrative Process Specialist, the reported scores primarily reflect alignment with this expert's judgment rather than comprehensive coverage of all possible anomalies. This semi-supervised evaluation may bias the models toward the specialist's mental model of risk and should be considered when extrapolating results to other institutions or regulatory environments. Given the unavoidable subjectivity in this expert-provided proxy labels, the study shifted away from purely technical optimization. Instead of pursuing marginal, hard-to-explain performance gains, which initially motivated computationally expensive Bayesian hyperpa-

rameter optimization (HPO), the central objective became maximizing the interpretability and operational relevance of alerts. This labeling context supports the thesis preference for transparent unsupervised methods and positions domain-specific post-processing filters as the primary lever for tuning the pipeline, helping ensure alerts remain auditable and suitable for use within the public sector rigorous governance framework.

Spark migration exposed key constraints. Some algorithms do not partition well, requiring broadcasting and limiting scalability. Distributed numerical variability can also reduce reproducibility. Reusing fixed hyperparameters may be stable but risky in fast-changing data. For researchers migrating anomaly detection pipelines to *Spark*, this thesis suggests a practical trade-off: reproducibility seems achievable, while linear scalability is harder to prove. The *Spark* version reproduced core detection logic, outputs, and filter effectiveness (RQ3), with metrics broadly comparable to the *Pandas* prototype, without costly distributed hyperparameter tuning. However, classical models that rely on global structure were less suited to partitioned execution. *LOF* and *OC-SVM* required broadcasting the full dataset to each worker to stabilize results and avoid partition-level failures (e.g., singular matrices or too few neighbors). As a result, the work supports pipeline portability in *Spark* but offers limited evidence of linear scalability for very large datasets. Most parallelism came from running independent model configurations in parallel, not clear intra-model speedups.

Conceptually, the experiments indicate that post-processing filters may potentially be effective in this context, although they rely on domain-specific heuristics, which limits generalization. Filters and hyperparameter optimization should be seen as complementary levers for stronger pipelines.

Finally, hospital deployment raises organizational, infrastructural, and regulatory challenges, especially around interoperability, cybersecurity, and real-time monitoring. Future directions

include *Spark*-native algorithms, hybrid models, unsupervised deep learning, federated learning for privacy-preserving multi-institutional training, and real-time anomaly detection in healthcare supply chains.

4.6 ANSWERS TO RESEARCH QUESTIONS (RQ1-RQ3)

Regarding RQ1, the experiments on CIUSSS-SLSJ procurement data suggest that multi-criteria AI-based anomaly detection using classical unsupervised methods (RC, IF, OC-SVM, and LOF) can identify patterns that are operationally meaningful for this specific institution, especially when anomalies are defined with respect to several distinct criteria (e.g., unusual prices, quantities, or delays). Across the different criteria and train/test splits, no single algorithm dominated universally, but IF and LOF tended to offer a good balance of detection quality on the labeled subsets, while RC was more sensitive to distributional assumptions and feature selection, and OC-SVM sometimes required careful tuning to avoid overly conservative or overly inclusive decision boundaries. Taken together, these results indicate that, in this case study, multi-criteria anomaly detection with unsupervised models is feasible and can complement existing monitoring practices, but its usefulness remains dependent on careful preprocessing, hyperparameter tuning, and ongoing expert validation, and the findings should not be generalized beyond the CIUSSS-SLSJ context without further testing.

For RQ2, the comparison between the default (unfiltered) models outputs (anomaly flags) and the different post-processing filters (in particular the *excess* and *lack* filters) indicates that contextual filtering can substantially change precision, recall, and the volume of alerts that must be reviewed, without modifying the underlying model scores themselves. In our results, the excess filter typically reduced the number of flagged cases and tended to increase precision and specificity, at the cost of lower recall, which made the alert list more manageable for practitioners but also increased the risk of missing borderline anomalies; the lack filter,

in contrast, highlighted atypical absences or under-activity and produced a different profile of anomalies, with more variable precision and recall depending on the criterion. Within the limits of the labeled test sets used here, these patterns suggest that post-processing filters are not a minor cosmetic step but a central design choice that trades off false positives against false negatives, and that their configuration should be treated as a policy decision informed by local risk tolerance rather than as a purely technical detail.

Finally, for RQ3, migrating the pipeline from a *Pandas*-based implementation to *Apache Spark* did not materially change the models outputs (anomaly flags) or evaluation metrics for the tested configurations. This suggests that, under the chosen implementation strategy (notably the use of broadcasting), the *Spark* version reproduced the main behavioural patterns of the original methods, producing broadly comparable results and showing no obvious numerical instability. At the same time, the available infrastructure and data volume were moderate, so the potential scalability advantages of *Spark* were only partially realised: runtime improvements were modest, and the current design did not fully exploit data partitioning or cluster-level parallelism in a way that would clearly demonstrate performance gains at much larger scales. As a result, these experiments support a cautious conclusion that *Spark* can host the proposed anomaly detection logic without degrading robustness on CIUSSS-SLSJ data, while offering only qualitative insight into scalability and performance limitations on the current infrastructure; more extensive tests on larger datasets, alternative cluster configurations, and more *Spark*-native implementations would be needed before making stronger claims about large-scale scalability.

CONCLUSION

CHAPTER V

CONCLUSION

5.1 CONCLUSION

This thesis examined the use of AI-based anomaly detection in healthcare procurement through a single-institution case study at CIUSSS-SLSJ, using operational SAP data and *KPI* definitions aligned with the CPI of the MSSS. Within the scope of the datasets, anomaly criteria, and evaluation protocol adopted, the results are consistent with the possibility that unsupervised anomaly detection, when complemented by domain-informed post-processing filters, can produce alerts that are more aligned with the assessed reference than the baseline approaches evaluated. These observations should not be interpreted as proof of effectiveness beyond the studied context, nor as evidence of general applicability to other institutions or domains.

The first article examined anomaly detection in CIUSSS-SLSJ procurement data and suggests that, in this specific setting, unsupervised AI methods can plausibly complement KPI-based monitoring and manual review by surfacing candidate irregularities across multiple criteria. The reported performance should be interpreted strictly within the semi-supervised evaluation design, since the metrics primarily quantify agreement with the judgments of a single domain expert rather than an objective institutional ground truth. A further contribution is methodological, showing that contextual post-processing filters and, to a lesser extent, hyperparameter tuning can refine raw model outputs and potentially improve their operational interpretability, without implying that such improvements would necessarily transfer to other institutions or datasets.

The second article explored portability of the same pipeline to *Apache Spark* and found that the results could be reproduced with only negligible deviations on the available infrastructure. However, the implementation relied on broadcasting the dataset and did not introduce meaningful intra-model distributed computation, which limits the extent to which scalability claims can be supported. As a result, the study provides evidence of technical transferability and backend-level reproducibility, while also clarifying that demonstrating true large-scale benefits would require distributed-native methods and implementations that avoid full-data broadcasting and better exploit partition-aware computation.

A central outcome of the work is methodological rather than definitive. The experiments suggest that practical improvements in alert relevance can be obtained not only through model selection and tuning, but also through structured post-processing rules that encode contextual knowledge. In this thesis, that lever appeared comparatively robust under the constraints considered, and it offers a pragmatic pathway for applied settings where computational resources, governance requirements, and operational integration constraints limit the feasibility of intensive optimization.

The work also highlights that scalability is not solely a question of computing capacity. The distributed implementation explored here indicates that some approaches can be reproduced in a *Spark* environment within acceptable variance for the reported experiments, while also revealing that certain algorithms and evaluation practices do not translate cleanly to distributed settings without methodological compromises. As a result, any claim of readiness for large-scale deployment would require additional engineering, monitoring, and validation work under real operating conditions.

The findings must be interpreted cautiously in light of the study's evaluation setting. The anomaly labels used for assessment reflect the judgments of a single domain expert, whose

decision criteria were not fully formalized and may involve subjectivity or label noise. Consequently, the reported performance should be read primarily as measured agreement with that reference, rather than as an objective measure of institutional ground truth. More broadly, the conclusions remain specific to the procurement data, anomaly definitions, and constraints studied here, and they require external replication and multi-expert validation before any broader generalization is warranted.

In summary, this thesis contributes a case-study-based pipeline and evidence that domain-informed post-processing can be a meaningful lever for improving the operational usefulness of anomaly detection outputs in healthcare procurement. It also clarifies key limitations that must be addressed before translating such approaches into durable decision-support tools, including validation design, distributed-native modeling choices, and governance structures aligned with public-sector realities.

5.2 RECOMMENDATIONS

In light of the results and insights developed throughout this thesis, the following considerations may be relevant for policy-makers, health-network managers, and researchers interested in anomaly detection in supply chains.

First, one possible approach is to explore a gradual integration of AI-based anomaly detection techniques into existing procurement management systems such as SAP or, over time, a more centralized ERP. A progressive implementation, initially limited to targeted use cases such as monitoring delivery delays or atypical price variations, may help limit operational and change management risks while providing early evidence about potential added value. This incremental approach may also support organizational acceptance and user trust, which are often discussed as important factors in adopting emerging technologies. Under the

operating thresholds explored in this thesis, the system produced an alert volume that appeared potentially compatible with manual review by a small team; however, actual manageability would depend on staffing, workflows, and local practices. From a data-governance perspective, this integration is typically better anchored in a data warehouse layer because anomaly detection for governance purposes benefits from stable, conformed definitions of measures, products, suppliers, and time, along with curated data quality checks and auditable lineage. A data warehouse can enforce schemas, harmonize *KPI* calculations, and provide consistent historical snapshots, which helps keep model inputs reproducible and alerts interpretable for oversight. By contrast, a data lake often concentrates heterogeneous, less curated, and more frequently evolving raw extracts, which can increase ambiguity, schema drift, and noise unless a strongly governed semantic layer is added. For this reason, even if data ingestion begins in a lake, the operational anomaly detection pipeline can be more defensible when executed on a curated data warehouse or equivalent governed lakehouse layer.

Second, pairing algorithmic models with post-processing filters that were effective in this case study (notably the *Excess* filter and, in certain contexts, the *Lack* filter) may help translate anomaly scores into more actionable outputs by reducing noise and clarifying anomaly types. Their performance should nonetheless be periodically reassessed as data distributions, procurement practices, and organizational constraints evolve, implying the need for monitoring, review, and controlled adjustment of rules and thresholds.

Third, the results highlight infrastructure considerations for scaling. The *Apache Spark* experiments suggest that distributed deployments can be explored, while also indicating limits for certain classical algorithms under distributed constraints. In our experiments, reproducing the pipeline for some models required broadcasting the dataset to all nodes; this choice supported experimentation but reduced the degree of true scalability. Organizations and public authorities may therefore consider prioritizing approaches that are compatible with

natively distributed computation, and, where appropriate, leveraging partnerships that provide specialized expertise in distributed AI engineering and operations.

Fourth, any implementation should be accompanied by governance mechanisms that frame the use of AI. Developing validation protocols, documentation practices, and transparency mechanisms can help address regulatory compliance, ethical safeguards, and protection of sensitive data. Such measures may also help manage risks associated with automation and support stakeholder trust within and beyond the health network.

Finally, maintaining a close connection between research and practice may be beneficial. The thesis illustrates an iterative approach in which initial orientations can be adjusted as empirical constraints become clearer. In this spirit, collaborative structures bringing together researchers, practitioners, and decision-makers may support continuous adaptation of models and more sustainable integration into organizational processes.

In summary, these considerations aim to support a responsible, gradual, and context-aware exploration of AI-based anomaly detection technologies. The intended objective is to improve the resilience and efficiency of healthcare supply chains while maintaining transparent governance and alignment with the operational realities of Québec's health network.

5.3 FUTURE WORK

Although this thesis offers theoretical and practical contributions, several avenues remain open for future work. These directions aim to deepen the research and to address limitations observed when integrating AI into healthcare procurement management. A priority is to reduce the methodological constraint imposed by the semi-supervised evaluation setting, specifically the reliance on binary anomaly labels from a single Administrative Process Specialist, by implementing multi-expert validation, quantifying inter-rater variability, and/or incorporating

institutional reference signals where available (e.g., audit-related flags), without assuming that any single source provides definitive ground truth.

A first avenue is the development or adaptation of anomaly detection algorithms designed for distributed environments. The *Spark*-based experiments highlighted the difficulty of transplanting certain classical models (such as LOF or One-Class SVM) into massively parallel architectures. Future work could therefore focus on algorithms that are native to *Spark* (or similar platforms), with the aim of handling larger data volumes more efficiently while maintaining acceptable analytic fidelity.

A second perspective concerns integrating additional data sources. The present analysis relied primarily on SAP data; future work could explore whether enriching the pipeline with other sources: such as logistics timestamps, supplier attributes, contract metadata, or external market indicators, improves contextualization and reduces ambiguity in anomaly interpretation. Such integration would also introduce new governance and data quality challenges that would require explicit handling.

Third, it may be relevant to explore advanced deep learning and semi-supervised methods. While this thesis prioritized interpretable models suited to public sector contexts, specialized architectures (e.g., autoencoders or graph neural networks) might capture more complex anomaly patterns. Any such exploration would likely need accompanying explainability and validation strategies to support acceptability in sensitive environments such as healthcare.

Another important avenue is the design of interactive decision-support tools. Beyond model performance, real-world effectiveness depends on how alerts are presented, interpreted, and acted upon. Future work could therefore involve developing dashboards with clear visualizations, filtering capabilities, and feedback loops that allow users to contextualize and review anomalies, and that support continuous improvement of detection and post-processing rules.

Finally, continued attention to ethical, organizational, and regulatory considerations remains essential. Future research could explore governance protocols that address data protection, accountability, transparency of algorithmic decisions, and fairness considerations in public resource management, recognizing that these requirements may vary across institutions.

In sum, future research can build on this thesis by moving from experimental evidence toward more durable, transparent, and scalable integration of AI tools into health-system procurement functions. Achieving this would likely involve (i) advancing distributed-native anomaly detection methods with the aim of improved scalability, (ii) integrating diverse data sources to strengthen contextual interpretation, and (iii) establishing multi-expert validation protocols and/or institutional reference signals to reduce reliance on a single expert's judgments and to better characterize uncertainty and generalizability.

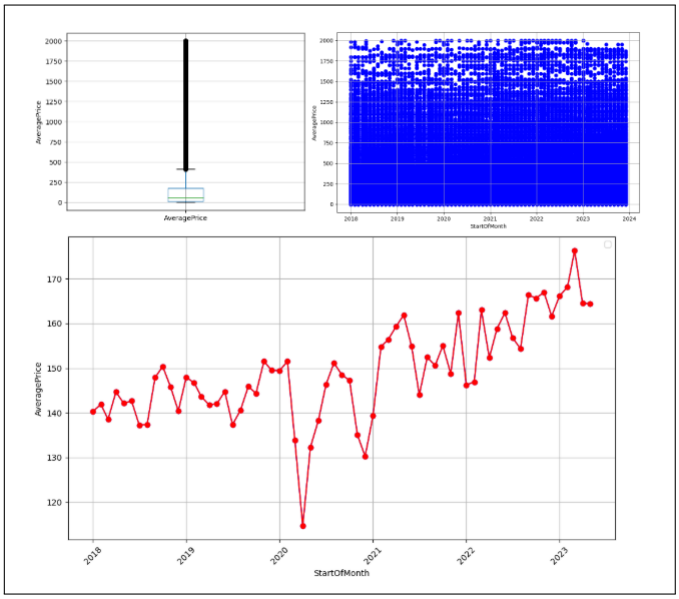
APPENDIX A

EXPLORATORY ANALYSIS: DESCRIPTION AND VISUALIZATION

A.1 ORDERS

Table A.1 : Descriptive analysis of orders data

	Date	Product price	Ordered value
Nombre	477734	477734	477734
Moyenne	2020-12-17	203.87	572.51
Minimum	2018-01-01	0	0
Q1	2019-06-01	16.48	47.97
Q2	2021-01-01	59.41	141.20
Q3	2022-07-01	187.46	384.43
Maximum	2023-12-01	2 051 885.19	5 795 862.07
Écart-type	NaN	3427.46	17 157.43



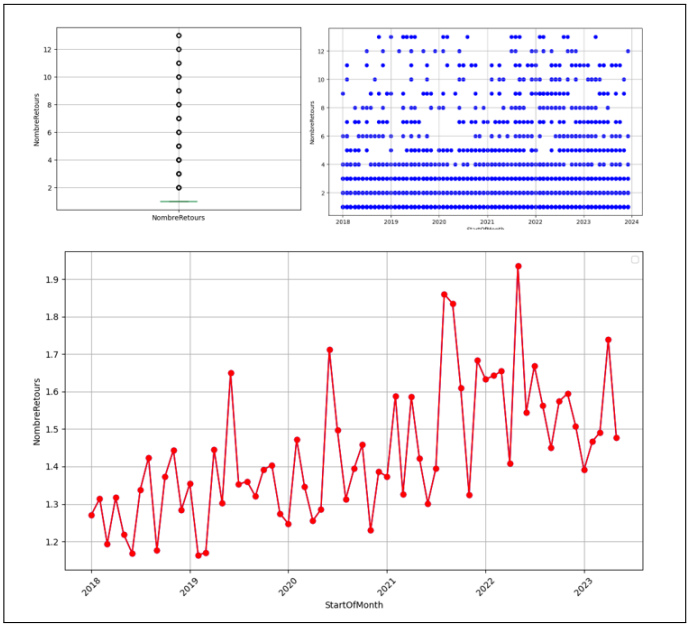
© Colin Bouchard

Figure A.1 : Orders data visualization

A.2 RETURNS

Table A.2 : Descriptive analysis of returns data

	Date	Number of returns
Number	13088	13088
Mean	2018-01-01	1
Minimum	2019-05-01	1
Q1	2020-11-01	1
Q2	2021-01-01	1
Q3	2022-04-01	1
Maximum	2023-12-01	59
Std. dev.	NaN	2.0657638186



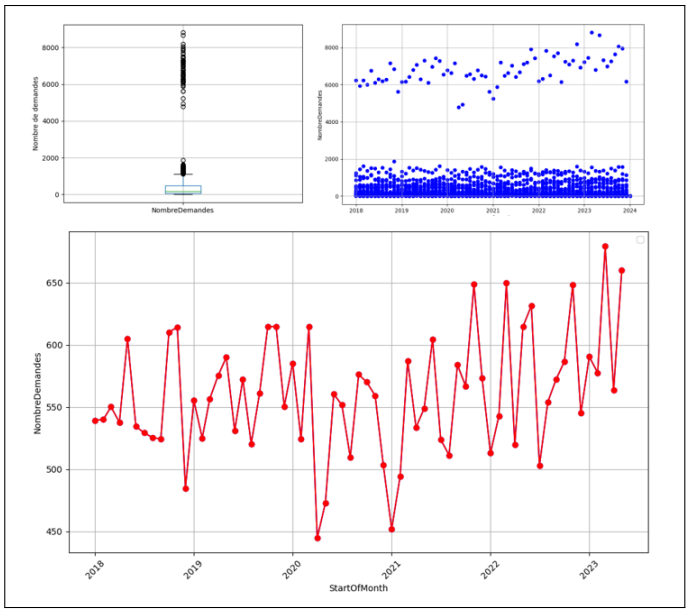
© Colin Bouchard

Figure A.2 : Returns visualization

A.3 PURCHASE REQUESTS

Table A.3 : Descriptive analysis of purchase requests data

	Date	Number of purchase requests
Number	1822	1822
Mean	2018-01-01	554.6319253977
Minimum	2019-05-01	1
Q1	2020-11-01	68
Q2	2021-01-01	160
Q3	2022-04-01	486
Maximum	2023-12-01	8815
Std. dev.	NaN	1316.9531541898



© Colin Bouchard

Figure A.3 : Purchase requests data visualization

A.4 ORDERED QUANTITIES

Table A.4 : Descriptive analysis of ordered quantities data

	Date	Quantities ordered
Number	431884	431884
Mean	2018-01-01	2.1694756706
Minimum	2019-05-01	0
Q1	2020-11-01	1
Q2	2021-01-01	1
Q3	2022-04-01	2
Maximum	2023-12-01	19
Std. dev.	NaN	2.2701715267

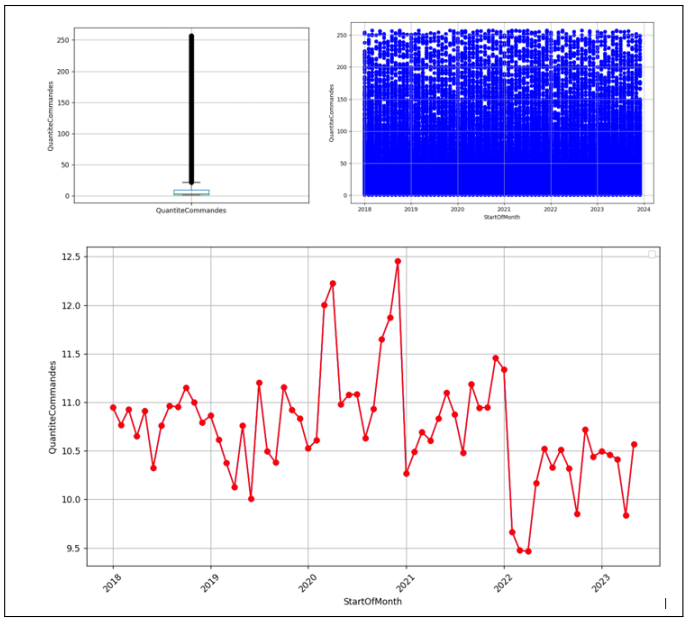


Figure A.4 : Ordered quantities data visualization

A.5 RECEIVED QUANTITIES

Table A.5 : Descriptive analysis of received quantities data

	Date	Monthly quantities received
Number	421193	421193
Mean	2018-01-01	61.69
Minimum	2019-05-01	0.02
Q1	2020-11-01	1
Q2	2021-01-01	3
Q3	2022-04-01	9
Maximum	2023-12-01	4 474 263.21
Std. dev.	NaN	7 314.55

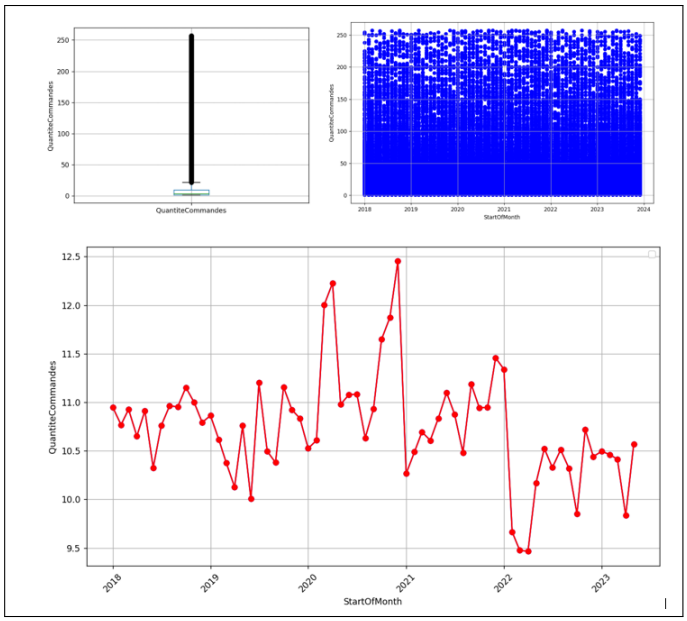
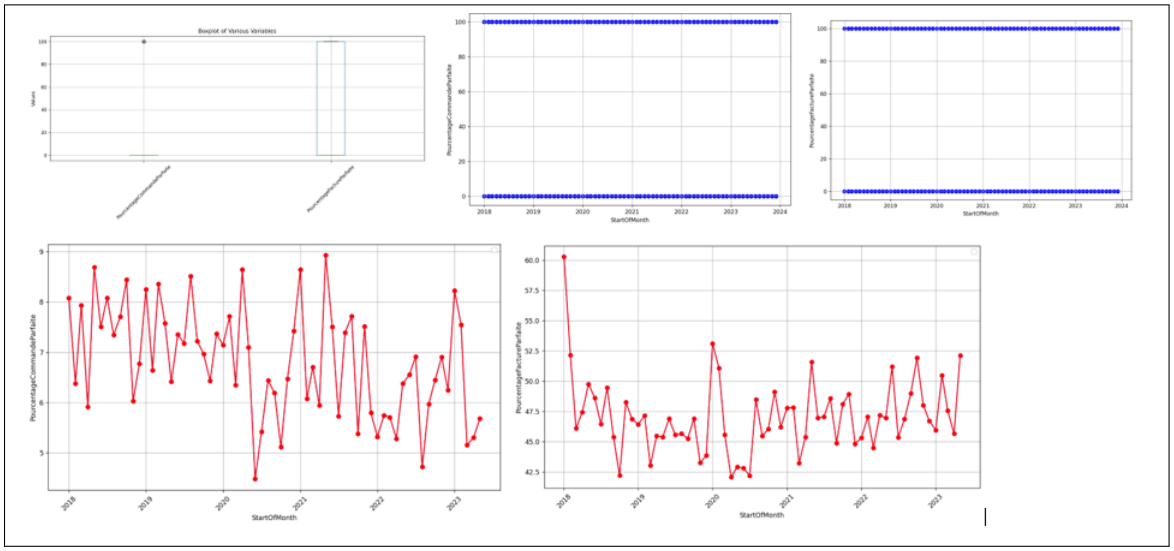


Figure A.5 : Received quantities data visualization

A.6 SUPPLIER PERFORMANCE MANAGEMENT KPIS

Table A.6 : Descriptive analysis of supplier performances KPIs data

	Date	Percentage of perfect orders	Percentage of perfect invoices
Number	421193	55008	54388
Mean	2018-01-01	26.22	25.65
Minimum	2019-05-01	0	0
Q1	2020-11-01	7	6
Q2	2021-01-01	14	13
Q3	2022-04-01	29	28
Maximum	2023-12-01	1161	1161
Std. dev.	NaN	44.58	44.89



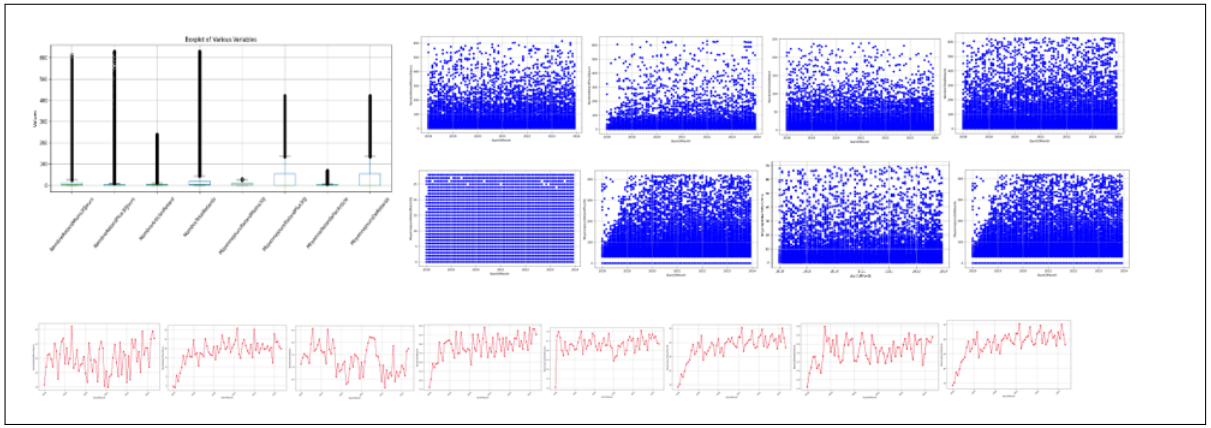
© Colin Bouchard

Figure A.6 : Supplier performances KPIs data visualization

A.7 BACK ORDERS PREVENTION KPIS

Table A.7 : Descriptive analysis of delay-related KPIs

	Date	Nb delays < 30 days	Nb delays > 30 days	Nb late products	Total delays	Avg days < 30 days	Avg days > 30 days	Avg delay per product	Avg delay (days)
Number	55506	55506	55506	55506	55506	38608	24557	31976	24557
Mean	2018-01-01	202.02	58.18	26.66	58.18	9.45	124.28	14.35	124.28
Minimum	2019-05-01	0	0	0	0	1	31	0	31
Q1	2020-11-01	0	0	0	1	4	45	0.66	45
Q2	2021-01-01	2	0	1	4	8	68	1.68	68
Q3	2022-04-01	13	4	5	25	13	123	4.92	123
Maximum	2023-12-01	128536	65991	11240	130278	30	7668	27817	7668
Std. dev.	NaN	2763.01	756.26	260.85	3013.87	6.63	181.39	194.08	181.38



© Colin Bouchard

Figure A.7 : Back orders prevention KPIs data visualization

For each dataset, the data show heterogeneity. For our anomaly detection objectives (RQ1–RQ2), this heterogeneity means that simple global thresholds on raw values would be unreliable: the same absolute value can be normal in one context and anomalous in another. This motivates the use of multivariate, context-aware detectors (RC, IF, LOF, OC-SVM) that learn typical behaviour jointly across variables instead of relying on univariate cut-offs.

APPENDIX B

EXPLORATORY ANALYSIS: DENSITY TEST

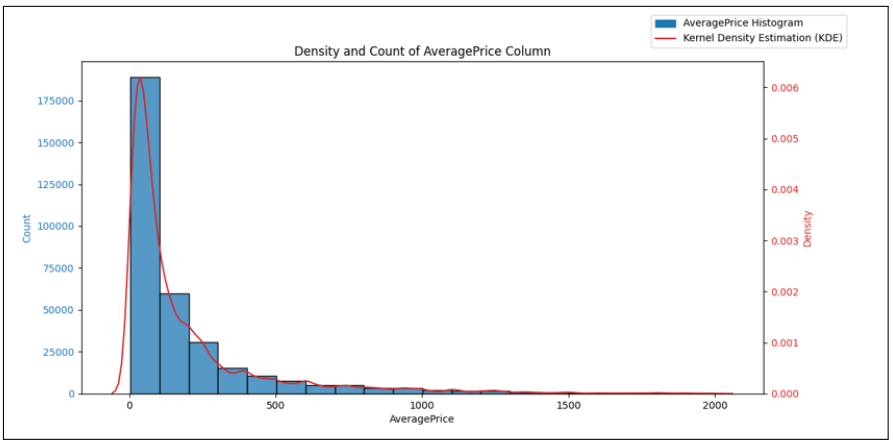


Figure B.1 : Product price KDE visualization

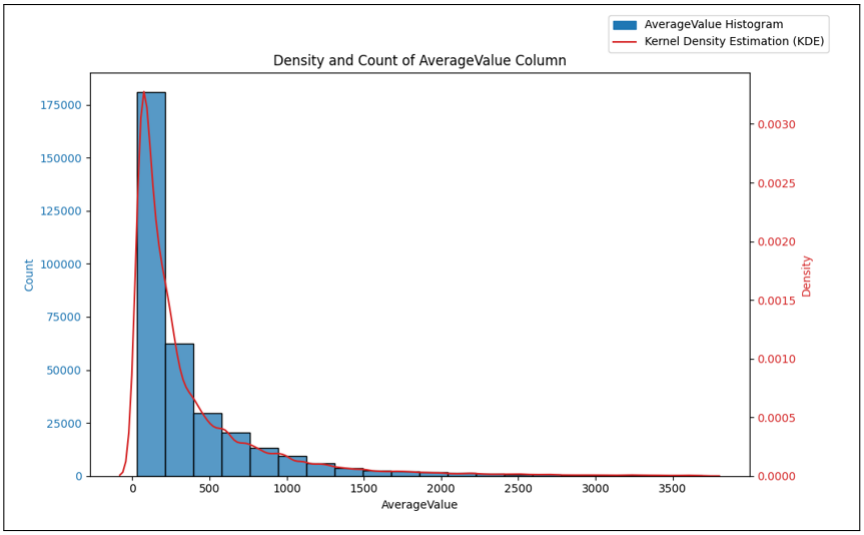
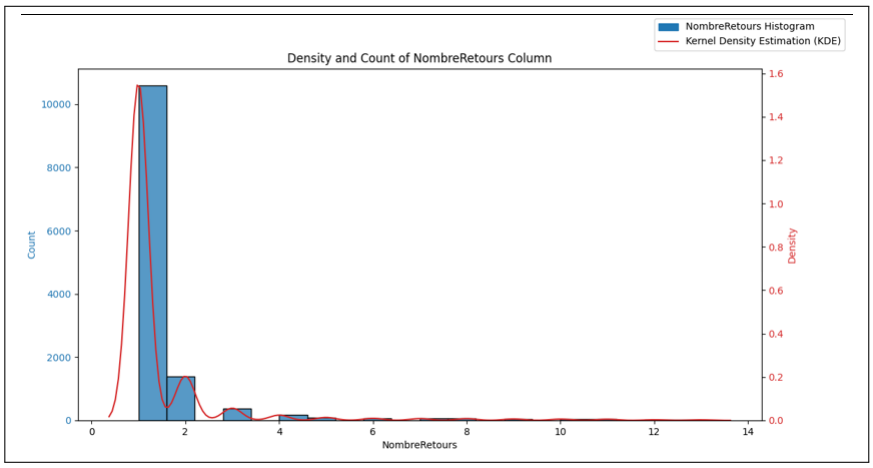
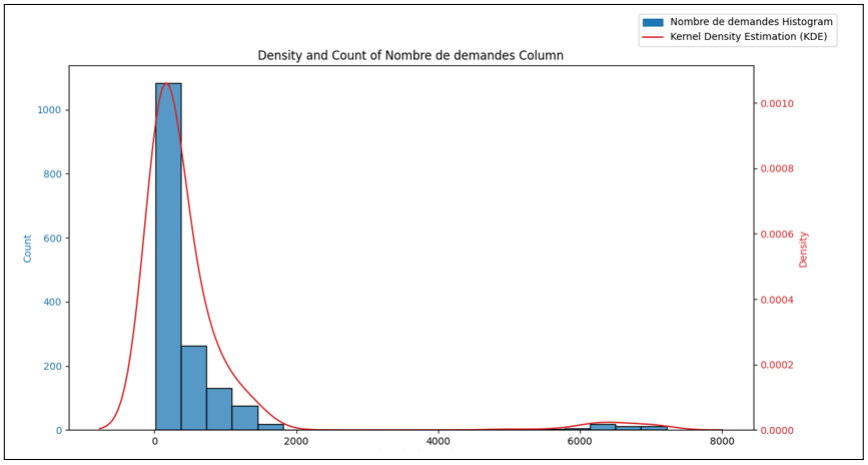


Figure B.2 : Ordered value KDE visualization



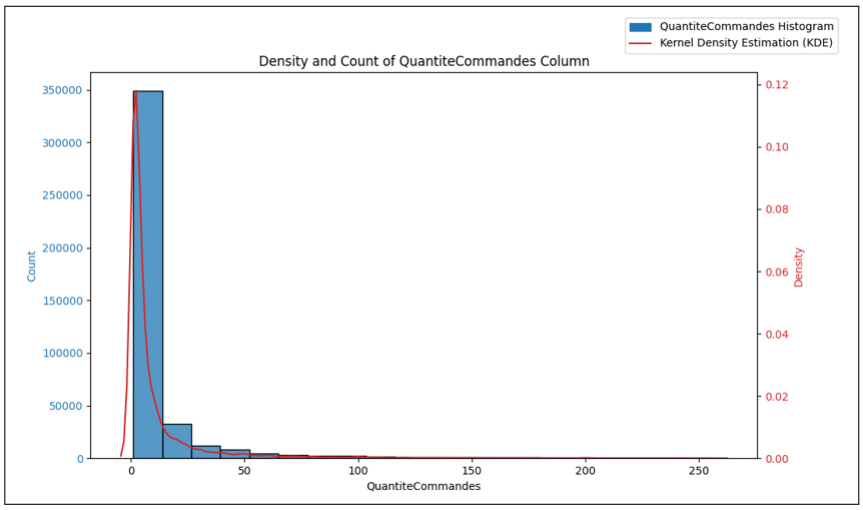
© Colin Bouchard

Figure B.3 : Returns Kernel KDE visualization



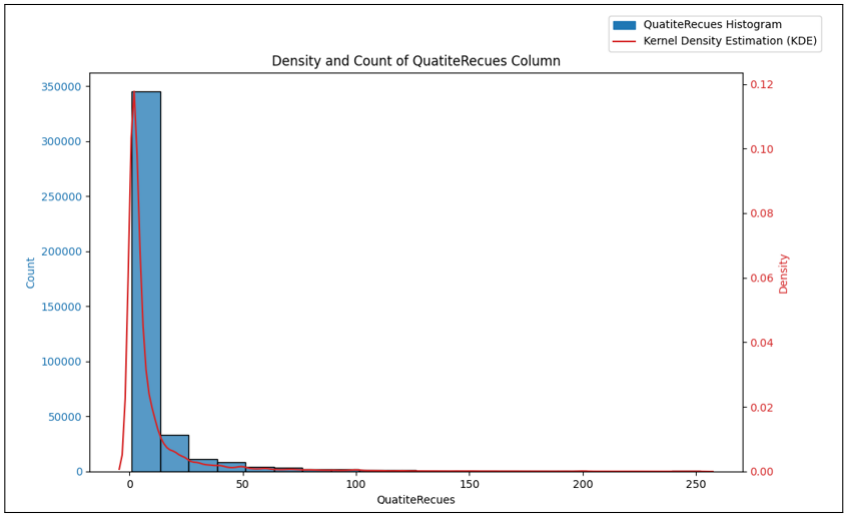
© Colin Bouchard

Figure B.4 : Purchase requests KDE visualization



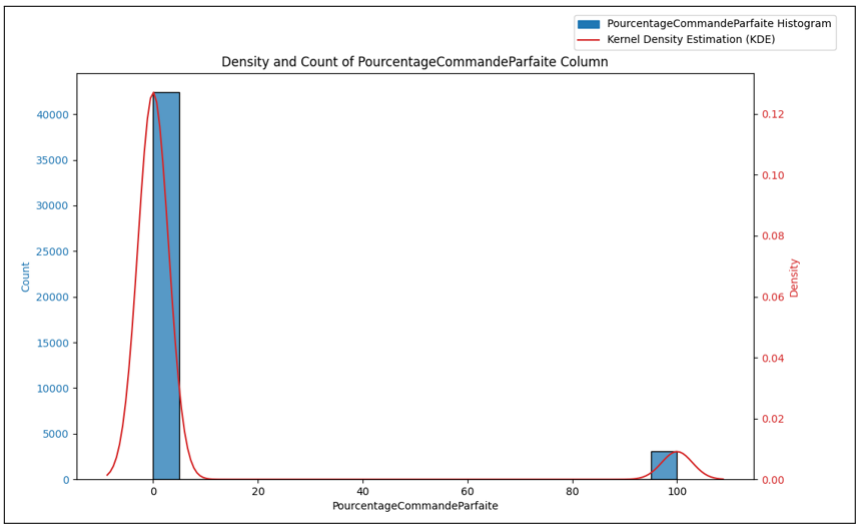
© Colin Bouchard

Figure B.5 : Ordered quantity KDE visualization



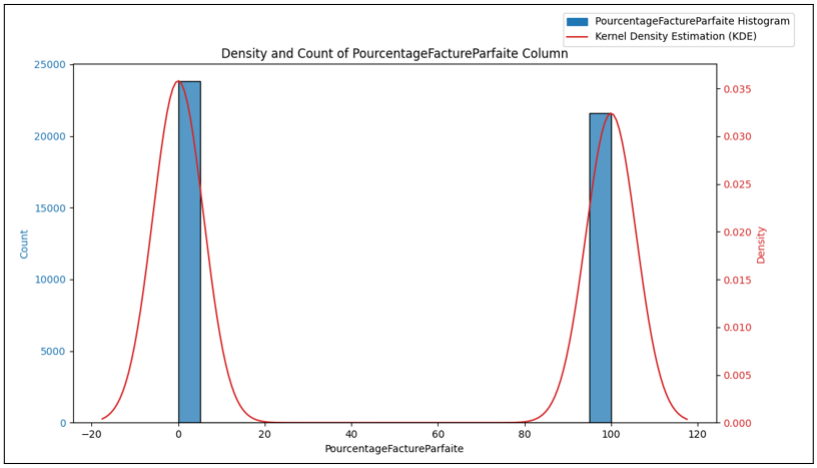
© Colin Bouchard

Figure B.6 : Received quantity KDE visualization



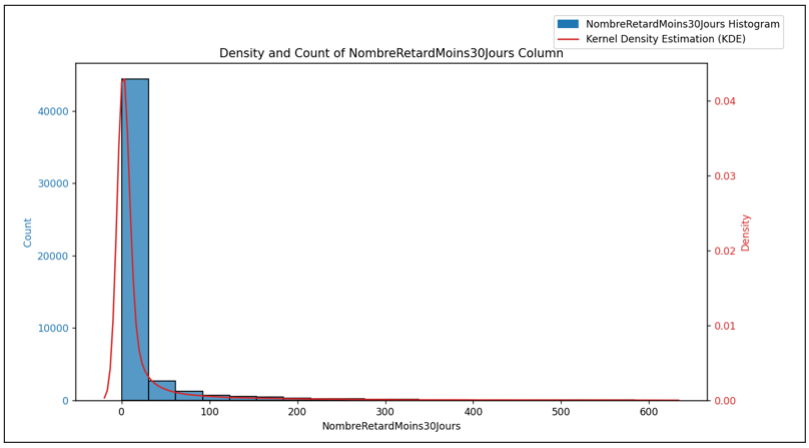
© Colin Bouchard

Figure B.7 : Perfect order KDE visualization



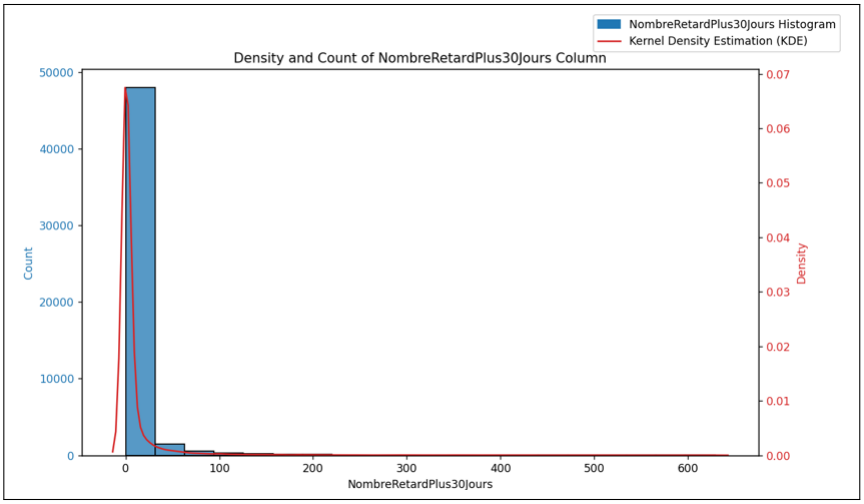
© Colin Bouchard

Figure B.8 : Perfect bill KDE visualization



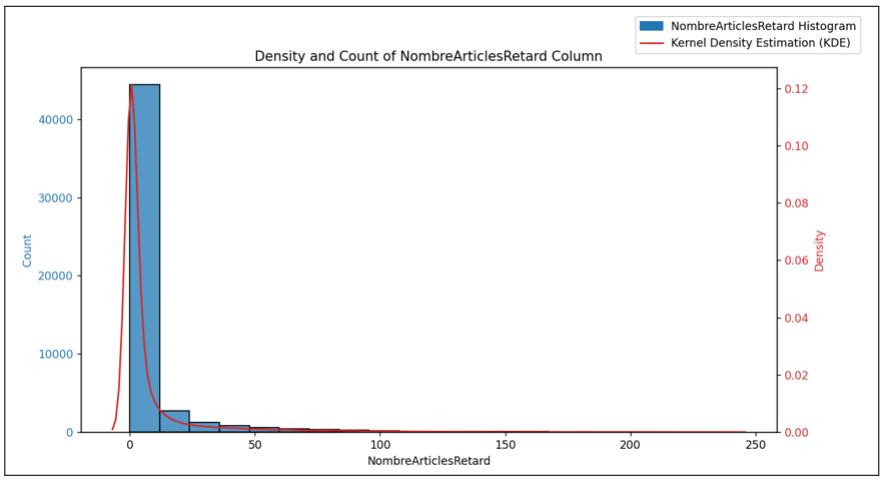
© Colin Bouchard

Figure B.9 : Number of delays (less than 30 days) KDE visualization



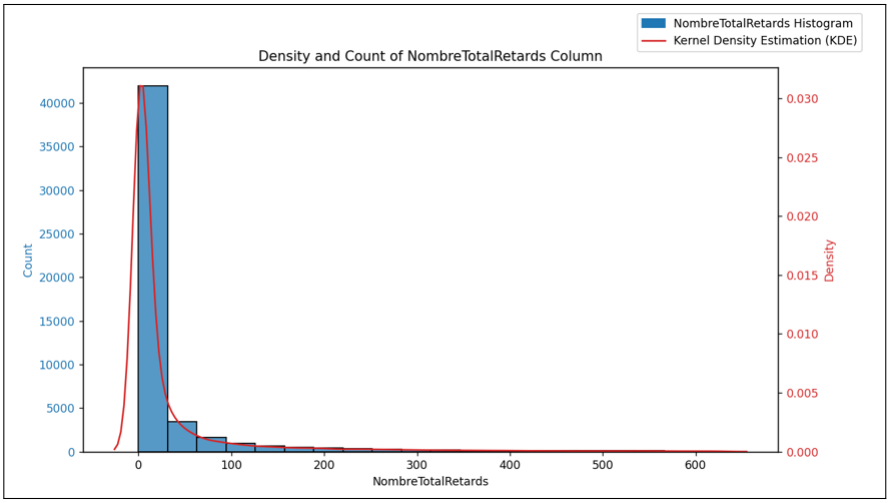
© Colin Bouchard

Figure B.10 : Number of delays (more than 30 days) KDE visualization



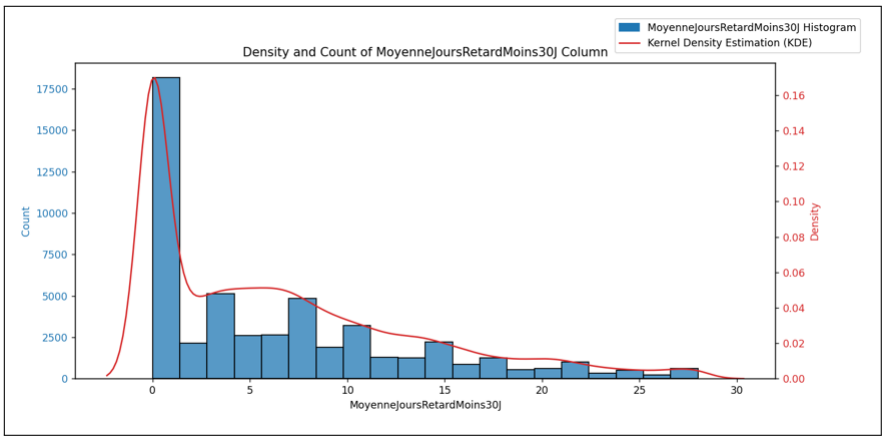
© Colin Bouchard

Figure B.11 : Late product KDE visualization



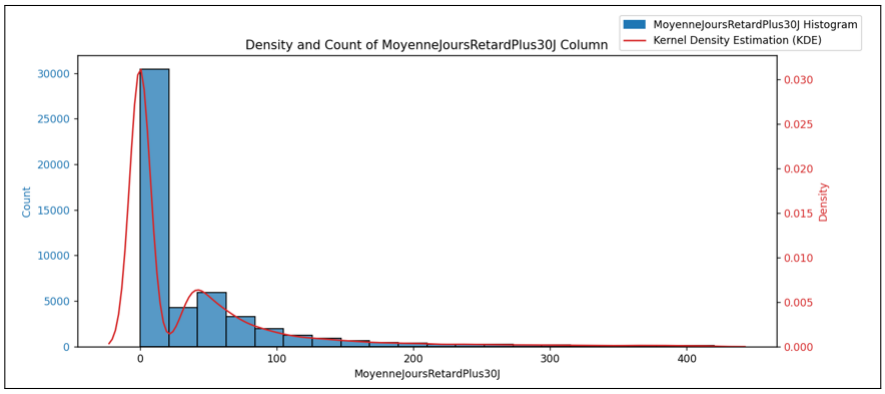
© Colin Bouchard

Figure B.12 : Total number of delays KDE visualization



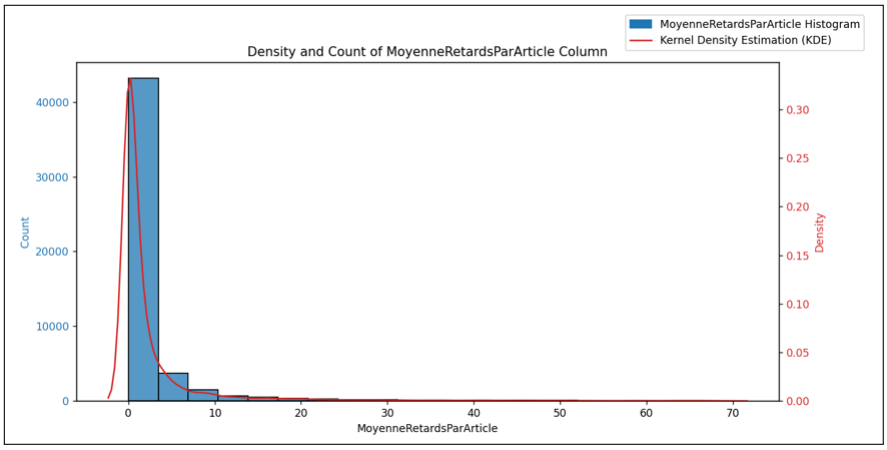
© Colin Bouchard

Figure B.13 : Average delay (less than 30 days) KDE visualization



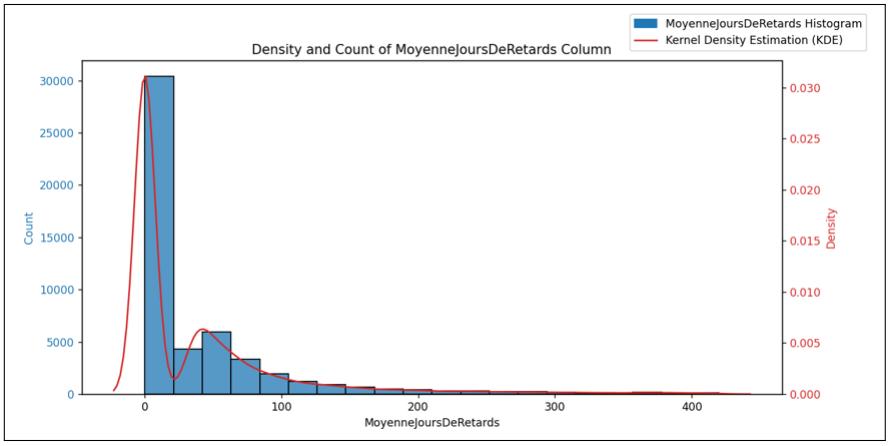
© Colin Bouchard

Figure B.14 : Average number of delay (more than 30 days) KDE visualization



© Colin Bouchard

Figure B.15 : Average delay per product KDE visualization



© Colin Bouchard

Figure B.16 : Average days of delay per product KDE visualization

Appendix B presents the results of the Kernel Density Estimation analysis, which provides a smooth, non-parametric estimate of the underlying probability density function for the main variables[223]. These KDE curves highlight the overall shape of the distributions, including asymmetry, heavy tails, and possible multimodality that are not visible with simple histograms or normality assumptions [223]. The observed density profiles indicate that several variables deviate from a standard Gaussian shape, which supports the use of unsupervised machine learning methods for anomaly detection, as these approaches do not rely on strict parametric assumptions and can better capture complex distributional structures in the data.

APPENDIX C

EXPLORATORY DATA ANALYSIS: TEMPORAL DEPENDENCE TEST (AUTOCORRELATION FUNCTION)

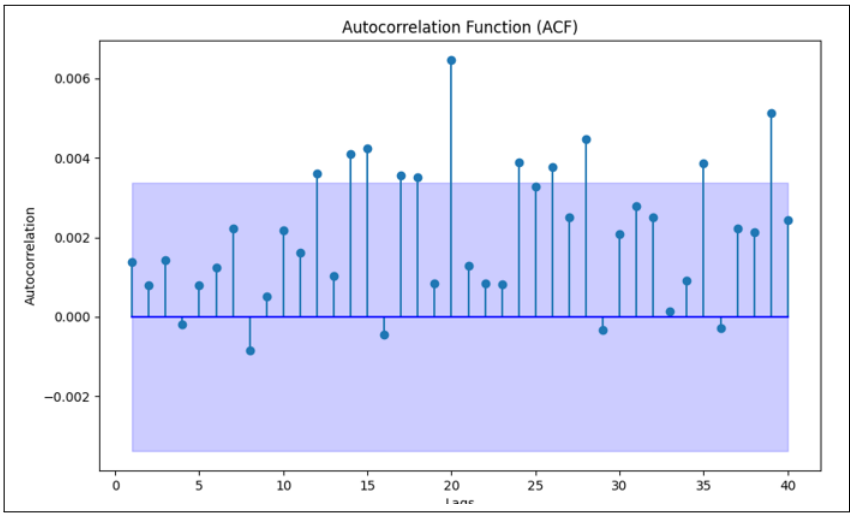
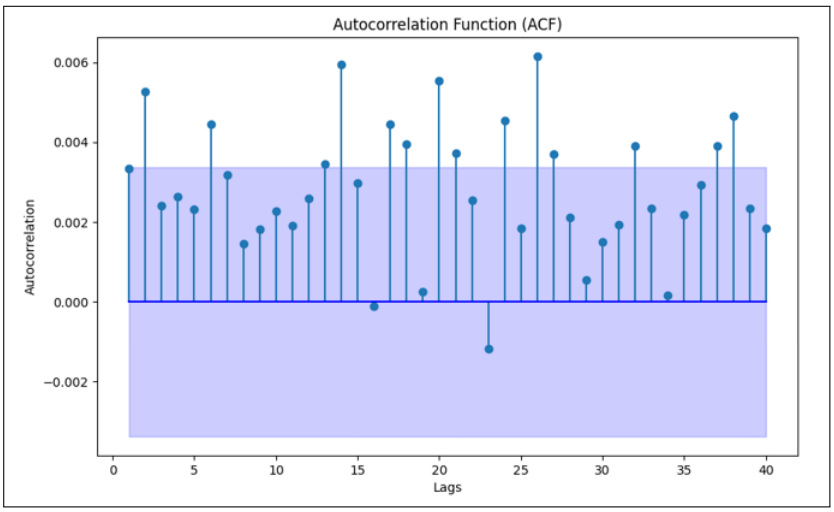
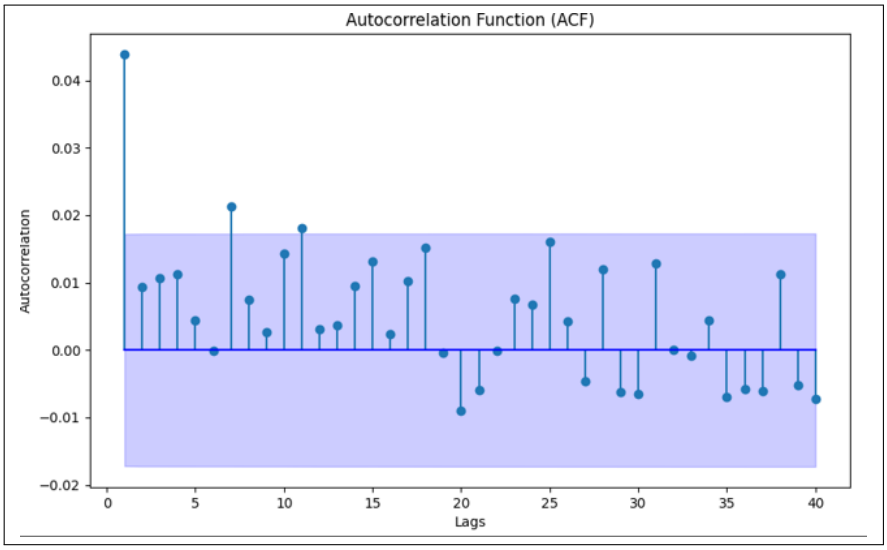


Figure C.1 : Auto-correlation function of products prices data



© Colin Bouchard

Figure C.2 : Auto-correlation function of orders values data



© Colin Bouchard

Figure C.3 : Auto-correlation function of returns data

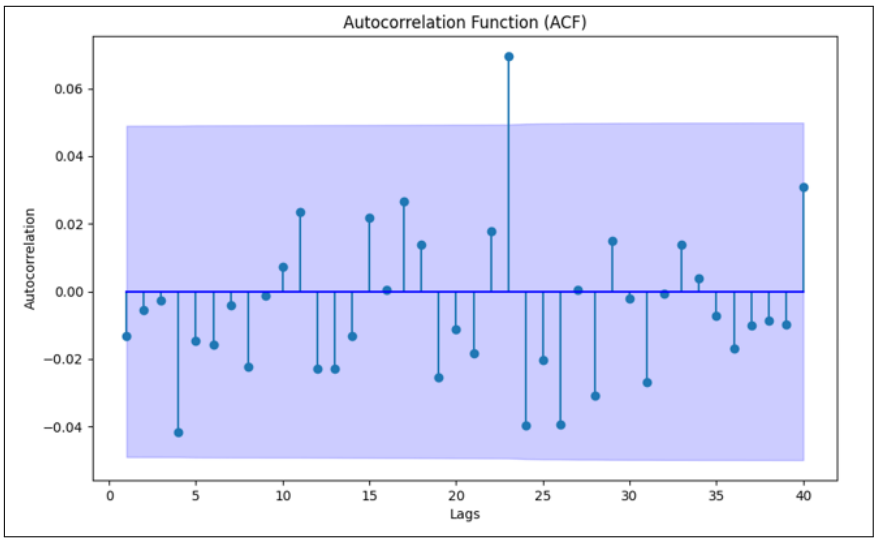


Figure C.4 : Auto-correlation function of purchase requests data

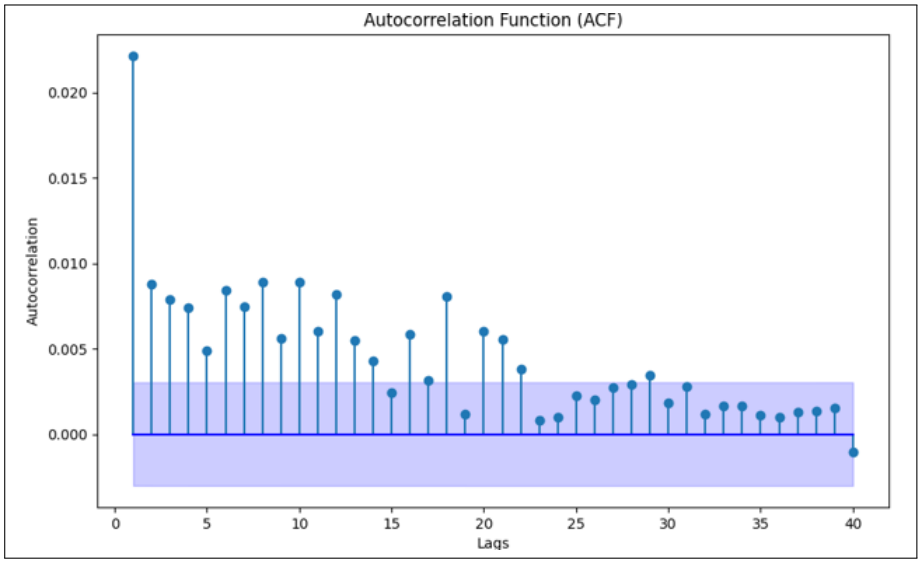


Figure C.5 : Auto-correlation function of ordered quantity data

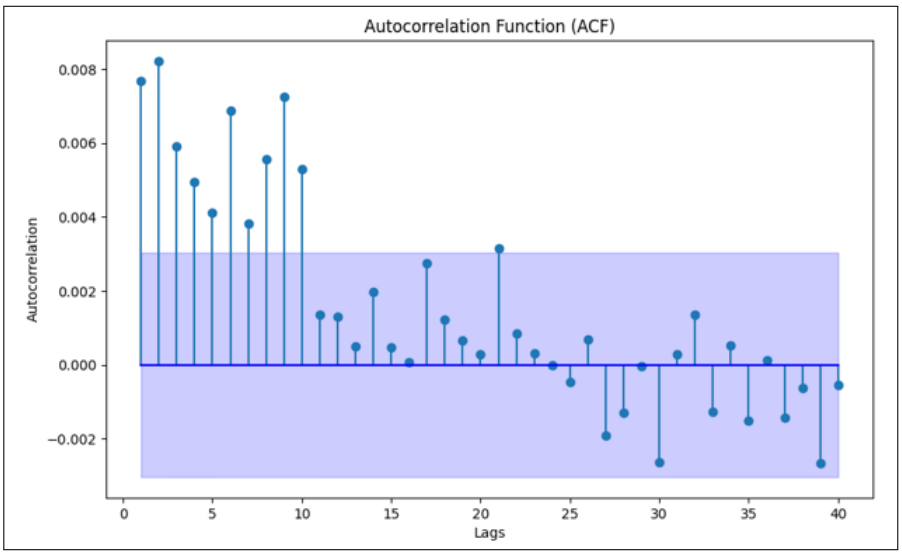


Figure C.6 : Auto-correlation function of received quantity data

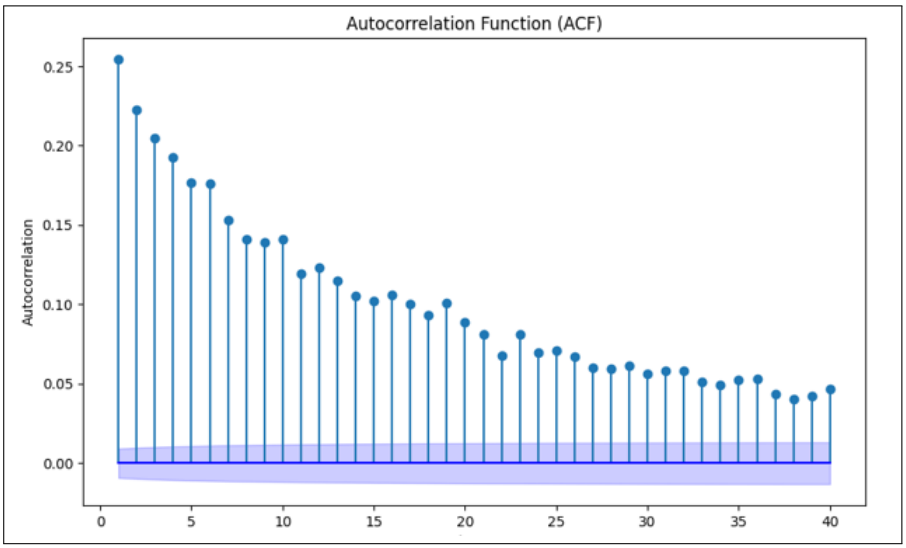
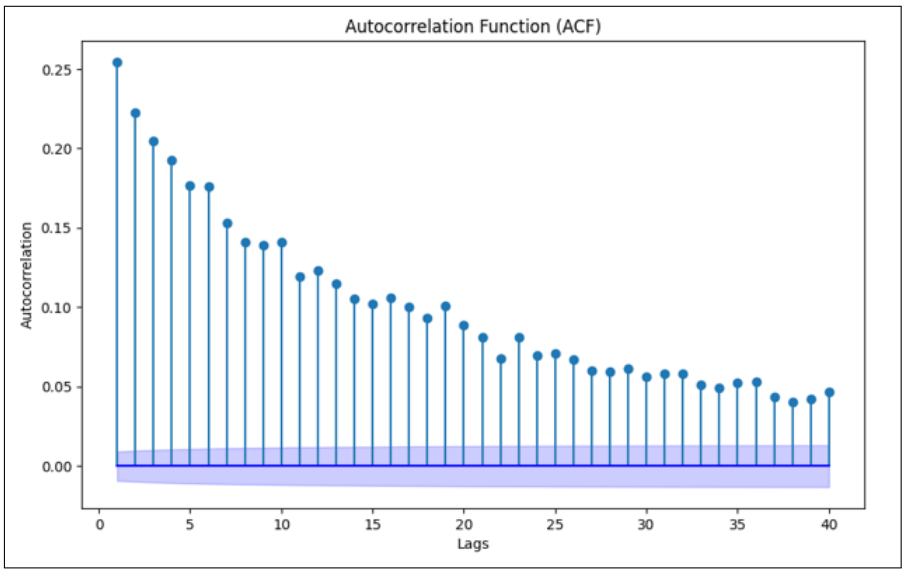
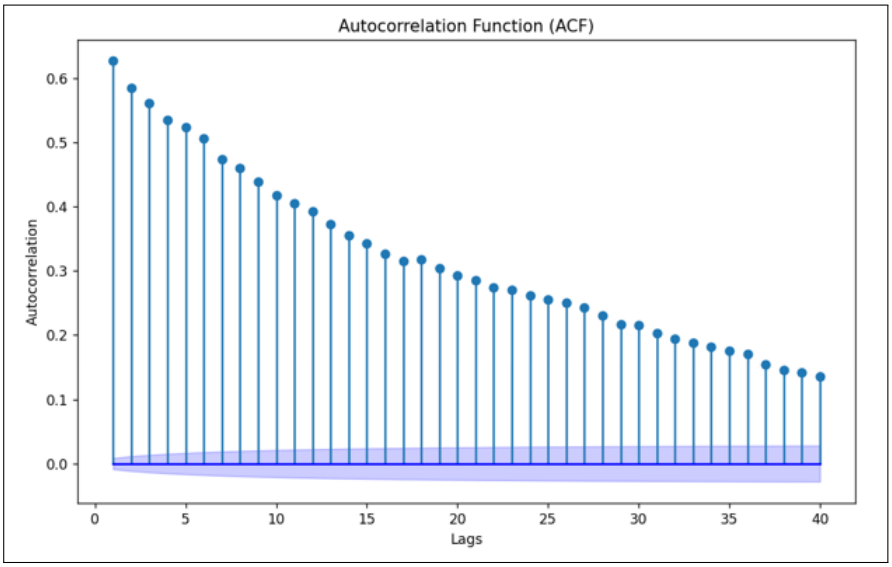


Figure C.7 : Auto-correlation function of perfect order data



© Colin Bouchard

Figure C.8 : Auto-correlation function of perfect bill data



© Colin Bouchard

Figure C.9 : Auto-correlation function of number of delays (less than 30 days) data

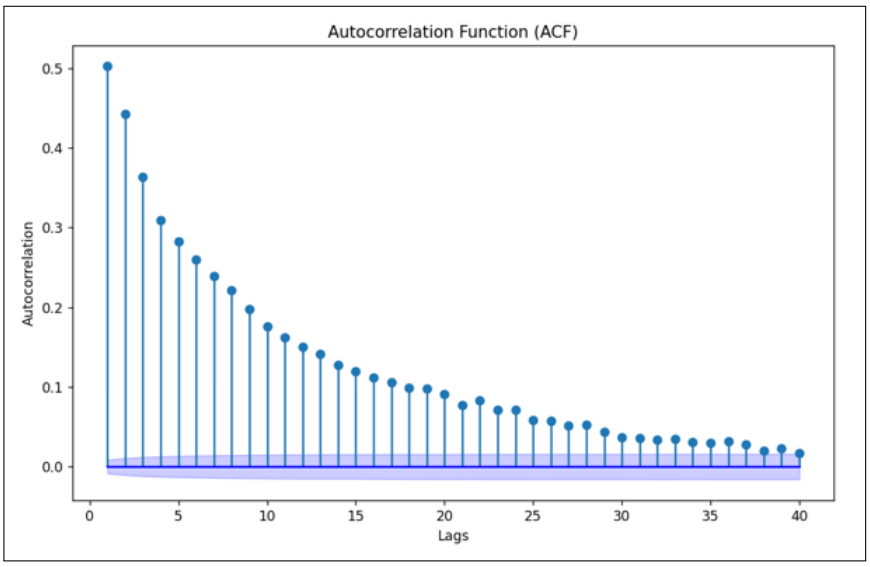


Figure C.10 : Auto-correlation function of number of delays (more than 30 days) data

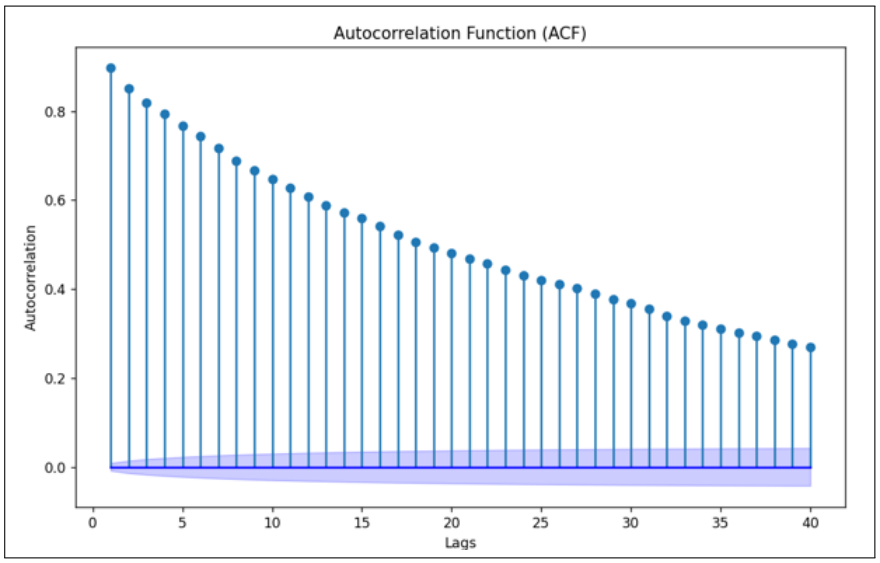


Figure C.11 : Auto-correlation function of late product data

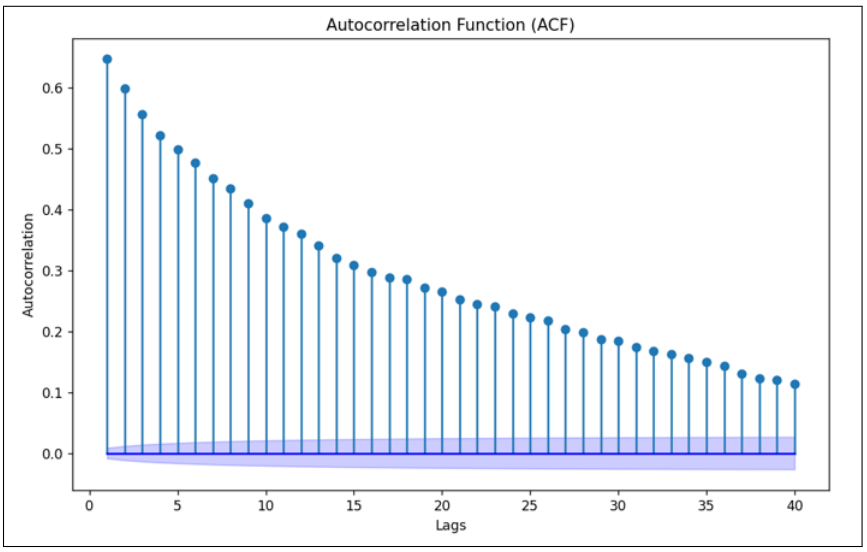


Figure C.12 : Auto-correlation function of total number of delays data

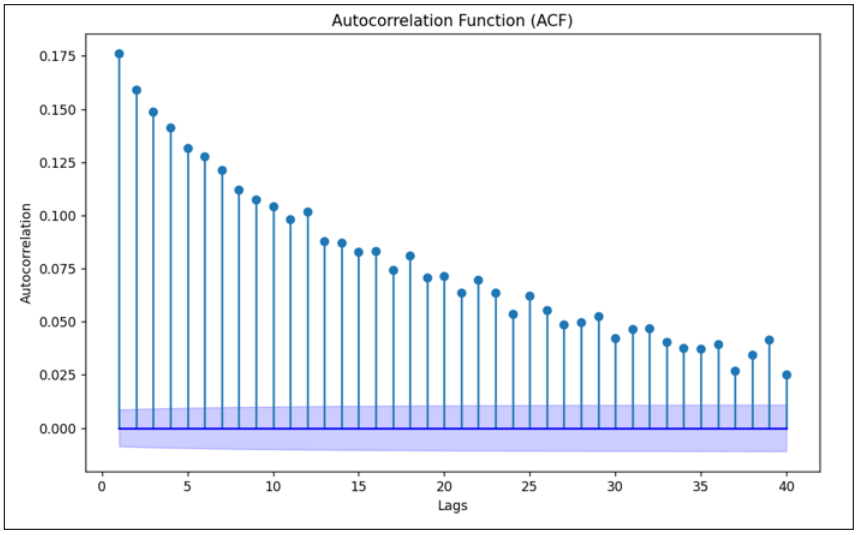
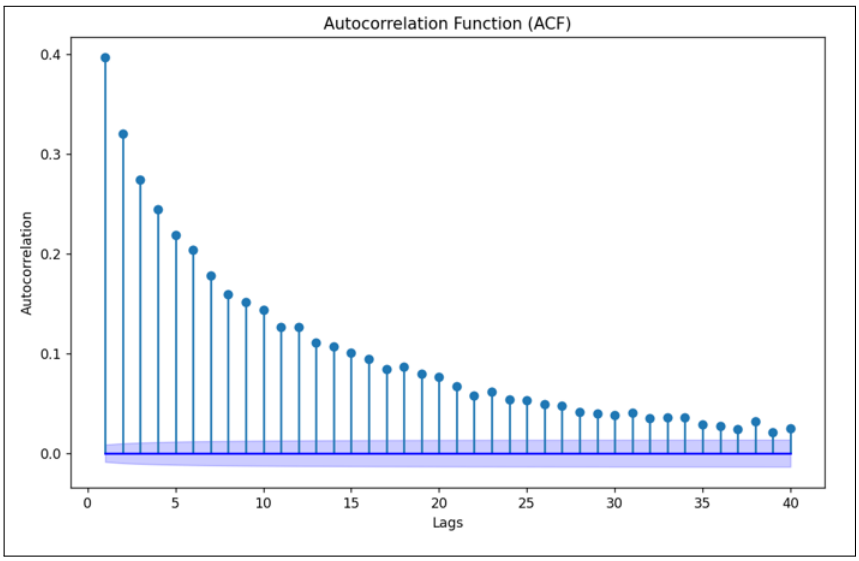
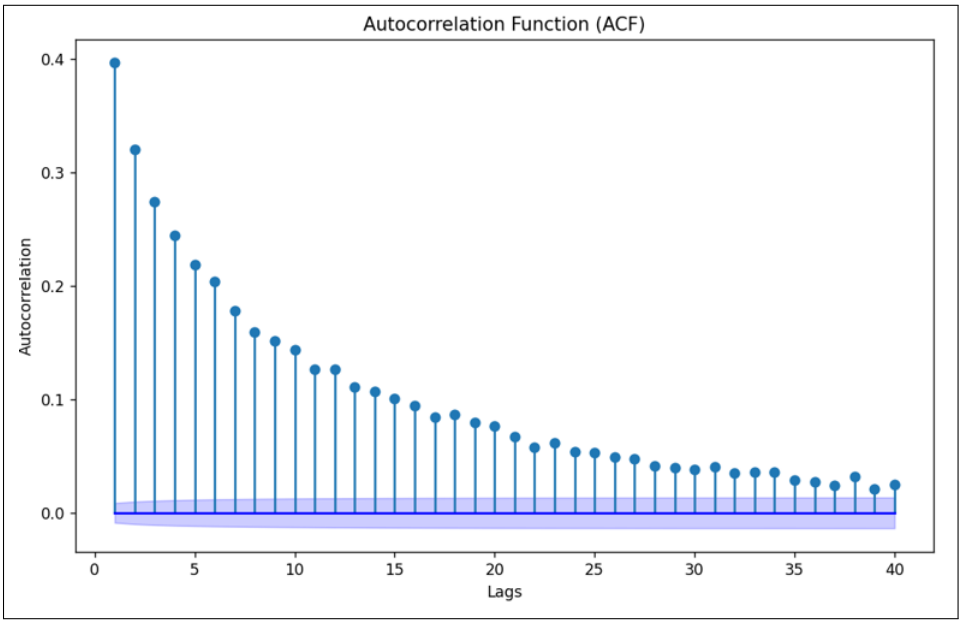


Figure C.13 : Auto-correlation function of average delay (less than 30 days) data



© Colin Bouchard

Figure C.14 : Auto-correlation function of average number of delay (more than 30 days) data



© Colin Bouchard

Figure C.15 : Auto-correlation function of average delay per product data

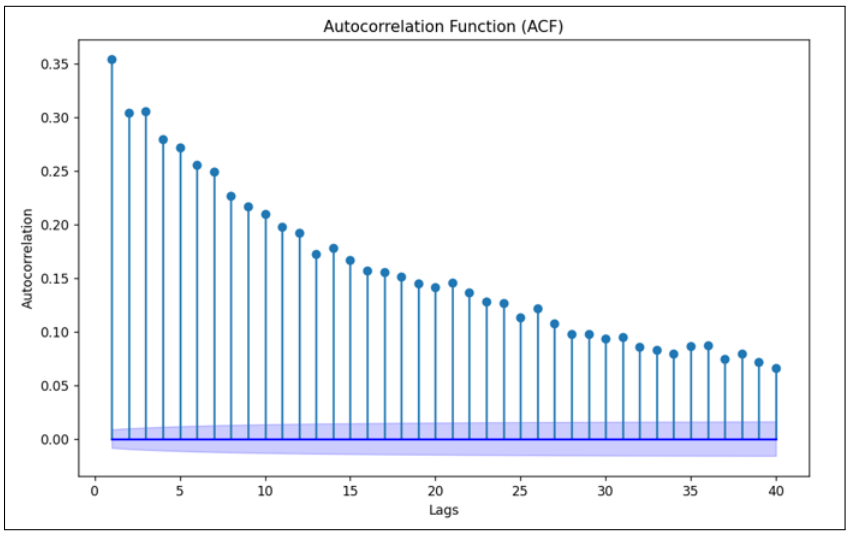


Figure C.16 : Auto-correlation function of average days of delay per product data

Appendix C applies the autocorrelation function to examine the correlation between observations in a series at different time lags. Strong autocorrelation over several lags may indicate the presence of regularities or repeating patterns in the data, which is important for modelling and forecasting[224, 225]. This appears to be the case for several of the variables. The presence of autocorrelation and (weak) stationarity over fiscal years supports our strategy of building time-aggregated, history-aware features for each supplier or product rather than treating observations as snapshots. For anomaly detection, this justifies formulating criteria such as “unusual annual average order quantity or price relative to historical behaviour” and evaluating models on temporally ordered splits, while leaving explicit time-series anomaly detectors as future work (RQ3).

APPENDIX D

EXPLORATORY ANALYSIS: NORMALITY TEST (SHAPIRO–WILK)

Table D.1 : Normality test (Shapiro–Wilk)

Variable	Statistic	p-value	Normal distribution
Average price	0.618783434232076	< 0.01	False
Average purchase value	0.6548340519269308	< 0.01	False
Number of returns	0.34822893142700195	< 0.01	False
Number of requests	0.3906937837600708	< 0.01	False
Ordered quantity	0.6602349678675333	< 0.01	False
Received quantity	0.40842755635579425	< 0.01	False
Percentage of perfect orders	0.2681267062822978	< 0.01	False
Percentage of perfect invoices	0.6350814501444498	< 0.01	False
Number of delays (< 30 days)	0.8854314088821411	< 0.01	False
Number of delays (> 30 days)	0.9054062962532043	< 0.01	False
Number of delays per product	0.8986338973045349	< 0.01	False
Total number of delays	0.7700250744819641	< 0.01	False
Average delay days (< 30 days)	0.8700205087661743	< 0.01	False
Average delay days (> 30 days)	0.9085018038749695	< 0.01	False
Average delay days per product	0.6246798038482666	< 0.01	False

The results of the Shapiro–Wilk test[215, 226], presented in Appendix D, reveal p-values lower than 0.01 for all variables, indicating that their distributions deviate significantly from normality. This finding is essential for guiding the choice of analytical methods toward those that do not assume a normal distribution of the data. Anomaly detection techniques are often dependent on specific assumptions about the data distribution. When the data do not follow a normal distribution, certain methods that rely on this assumption, such as the Z-score[2] and Grubbs' test[3], may not be appropriate. For this reason, it is recommended to turn to machine learning methods, which do not necessarily require the assumption of a normal data distribution. Because all variables deviate strongly from normality, parametric outlier tests such as Z-score and Grubbs, whose thresholds assume Gaussian tails, would yield biased anomaly scores if used as primary detectors. This justifies basing our core anomaly detection on methods that do not rely on normality assumptions, such as Robust Covariance (RC) in Elliptic Envelope (EE), Isolation Forest (IF), Local Outlier Factor (LOF) and One-Class SVM (OC-SVM), and relegating Z-score/Grubbs to a preliminary, sanity-check role only.

APPENDIX E

EXPLORATORY DATA ANALYSIS: SINGULARITY TEST (MATRIX RANK)

Table E.1 : Singularity test (matrix rank)

Variables	Matrix rank	Singular
Average price	5	*False*
Average purchase value	5	*False*
Number of returns	3	*False*
Number of requests	3	*False*
Ordered quantity	3	*False*
Received quantity	3	*False*
Percentage of perfect orders	4	*False*
Percentage of perfect invoices	4	*False*
Number of delays (< 30 days)	7	*True*
Number of delays (> 30 days)	7	*True*
Number of delays per product	7	*True*
Total number of delays	7	*True*
Average delay days (< 30 days)	7	*True*
Average delay days (> 30 days)	7	*True*
Average number of delay per product	7	*True*
Average delay days per product	7	*True*

Appendix E highlights the results of a singularity test carried out on various supply and logistics variables. This analysis reveals a clear distinction between operational variables, which are non-singular, and those related to delays, which are identified as singular. This observation indicates a strong correlation among the variables associated with delays, suggesting possible redundancy of information [227]. Such redundancy may hinder subsequent statistical analyses, notably because of the risk of multicollinearity, which affects the reliability of results obtained by linear regression and other similar analytical methods [219]. The near-singular structure and strong multicollinearity among delay-related variables indicate that regression-type approaches or distance measures in the original space would be unstable. In our pipeline, this motivates applying PCA and favouring anomaly detectors that can exploit correlated structure (RC, OC-SVM, IF, LOF) while avoiding interpretations that depend on independent contributions of each delay feature.

APPENDIX F

EXPLORATORY DATA ANALYSIS: MULTIVARIATE TEST

Table F.1 : Regression results (R^2, F statistic and p-value)

Variables	R^2	F statistic	p-value
Average price	0.007	1215	< 0.01
Average purchase value	0.008	1363	< 0.01
Number of returns	0.005	62.28	< 0.01
Number of requests	0.216	220.7	< 0.01
Ordered quantity	0.002	647.7	< 0.01
Received quantity	0.001	580.5	< 0.01
Percentage of perfect orders	1	2.562e+36	< 0.01
Percentage of perfect invoices	1	4.133e+32	< 0.01
Number of delays (< 30 days)	1	7.381e+31	< 0.01
Number of delays (> 30 days)	1	9.012e+29	< 0.01
Number of delays per product	1	2.377e+30	< 0.01
Total number of delays	1	5.940e+26	< 0.01
Average delay days (< 30 days)	1	5.133e+29	< 0.01
Average delay days (> 30 days)	1	1.161e+31	< 0.01
Average delay days per product	1	5.650e+28	< 0.01

Appendix F shows, through a multivariate analysis, that operational variables such as average price and number of requests have a moderate but statistically significant influence on the response variable, as indicated by low to moderate R^2 values and p-values < 0.01. In contrast, the variables related to delays display a perfect R^2 of 1 with extremely high F statistics, indicating an exact correlation with the response variable[228]. This contrast highlights potential overfitting or extreme specificity of the percentage variables (perfect orders/invoices) and of the delay (count/average) variables, requiring cautious interpretation to avoid misleading conclusions based on an apparent perfect correlation.

APPENDIX G

EXPLORATORY DATA ANALYSIS: ELLIPTICITY TEST

Table G.1 : Ellipticity analysis of variables

Variables	Eigen-ratio	Skewness	Kurtosis	Ellipticity risk
Average price	116140028384410	1.39377188	6.85441766	30%
Average purchase value	116140028384410	1.39377188	6.85441766	30%
Number of returns	1529799543275101	1.66051702	11.48425180	30%
Number of requests	54722059691867	1.50641833	9.17514456	30%
Ordered quantity	5221565608944	1.77016521	13.22273770	30%
Received quantity	5220329913954	1.74064055	12.90568180	30%
Percentage of perfect orders	5042814501475	0.99699189	4.44709033	30%
Percentage of perfect invoices	5042814501475	0.99699118	4.44709033	30%
Number of delays (< 30 days)	4066922344262	0.39536160	5.25854973	60%
Number of delays (> 30 days)	4066922344262	0.39536160	5.25854973	60%
Number of delays per product	4066922344262	0.39536160	5.25854973	60%
Total number of delays	4066922344262	0.39536160	5.25854973	60%
Average delay days (< 30 days)	4066922344262	0.39536160	5.25854973	60%
Average delay days (> 30 days)	4066922344262	0.39536160	5.25854973	60%
Average delay days per product	4066922344262	0.39536160	5.25854973	60%

Appendix G shows, through the ellipticity test, that variables such as average price and average purchase value exhibit a moderate ellipticity risk of 30%, with slightly asymmetric distributions and heavier-than-normal tails, suggesting some deviation, but not an extreme one, from a multivariate normal distribution. In contrast, the variables related to delays present a higher ellipticity risk of 60%, indicating distributions that are further from normality, with increased variability and more frequent extremes. The ellipticity risk was determined using the eigen-ratio as well as skewness and kurtosis[229, 230]. These distinctions highlight the importance of adapting analytical methods to account for these differences in data distribution, especially for the delay variables, which display a more complex structure. The ellipticity results show that delay variables in particular do not follow simple elliptical contours, which limits the reliability of purely Mahalanobis-based distance rules in the raw space. This strengthens the choice of combining Robust Covariance (RC) in Elliptic Envelope (EE) (for approximately elliptical regions) with density- and isolation-based methods such as Local Outlier Factor (LOF) and Isolation Forest (IF), which remain effective when clusters are skewed or heavy-tailed, directly serving RQ1 on robust multi-criteria anomaly detection.

APPENDIX H

EXPLORATORY DATA ANALYSIS: COMPLEXITY TEST

Table H.1 : Complexity test heatmap

Variables / Dataset	DataFrame size	Filtering (seconds)	Sorting (seconds)	Mean calculation (seconds)
Average price and value	100	0.000000000	0.000000000	0.000000000
Average price and value	1000	0.000000000	0.001001659	0.000000000
Average price and value	10000	0.000998259	0.001997948	0.000999689
Average price and value	100000	0.005015373	0.016986370	0.011100077
Number of returns	100	0.000997782	0.000000000	0.000000000
Number of returns	1000	0.001000643	0.000000000	0.000000000
Number of returns	10000	0.001000881	0.000997782	0.001002073
Number of requests	100	0.000000000	0.000998259	0.000000000
Number of requests	1000	0.000996828	0.007452249	0.000000000
Ordered quantity	100	0.000000000	0.000960112	0.000000000
Ordered quantity	1000	0.000000000	0.001004696	0.000000000
Ordered quantity	10000	0.001000881	0.001005411	0.000994682
Ordered quantity	100000	0.005177975	0.015009165	0.000995874
Received quantity	100	0.000999212	0.000000000	0.000000000
Received quantity	1000	0.000999212	0.000969887	0.000000000
Received quantity	10000	0.001033783	0.000999689	0.000000000
Received quantity	100000	0.005963087	0.014038086	0.000991821
Percentage of perfect orders	100	0.000000000	0.000000000	0.000000000
Percentage of perfect orders	1000	0.000000000	0.000999689	0.000000000
Percentage of perfect orders	10000	0.001000166	0.000999928	0.001003027
Percentage of perfect invoices	100	0.000000000	0.000000000	0.000000000
Percentage of perfect invoices	1000	0.000991821	0.001000404	0.000000000
Percentage of perfect invoices	10000	0.001000881	0.001003504	0.000995398
Number of delays (< 30 days)	72	0.000998735	0.001074314	0.000000000
Number of delays (> 30 days)	72	0.001002789	0.000000000	0.000000000
Late products	72	0.000000000	0.000998020	0.000000000
Average delay days (< 30 days)	72	0.000000000	0.000000000	0.000000000
Average delay days (> 30 days)	72	0.000000000	0.000000000	0.000000000
Total number of delays	72	0.000000000	0.000000000	0.000000000
Average delays per product	72	0.001019716	0.000000000	0.000000000

Appendix H examines the complexity of operations on various datasets, revealing that the time required for filtering, sorting, and mean calculation increases with dataset size[231]. Sorting stands out as the operation most affected by the growth in data volume, especially when moving from small to large datasets. This analysis highlights the importance of optimizing algorithms to efficiently handle large amounts of data, particularly in contexts where performance and response time are critical. It underscores the direct relationship between data complexity and processing time, a key factor in the design and improvement of data management systems.

APPENDIX I

EXPLORATORY DATA ANALYSIS: STATIONARITY TEST (AUGMENTED DICKEY–FULLER)

Table I.1 : Stationarity test (Augmented Dickey–Fuller)

Variables	Statistic	p-value	Stationary
Average price	0.6545046170552572	< 0.01	True
Average purchase value	-70.2133621188546	< 0.01	True
Number of returns	-42.69499431352292	< 0.01	True
Number of requests	-26.482032633790027	< 0.01	True
Ordered quantity	-51.1814485920867	< 0.01	True
Received quantity	-70.16313707259405	< 0.01	True
Percentage of perfect orders	-40.61579410574195	< 0.01	True
Percentage of perfect invoices	-51.9827512831878	< 0.01	True
Number of delays (< 30 days)	-7.866971827984114	< 0.01	True
Number of delays (> 30 days)	-8.918648864324053	< 0.01	True
Number of delays per item	-8.227089062485762	< 0.01	True
Total number of delays	-7.0578523521789505	< 0.01	True
Average delay days (< 30 days)	-6.794904370225435	< 0.01	True
Average delay days (> 30 days)	-8.356264728427387	< 0.01	True
Average delay days per item	-8.227089062485762	< 0.01	True

Appendix I shows, via the Augmented Dickey–Fuller test, that all of the analyzed variables are stationary, with p-values lower than 0.01. Results at monthly resolution suggest several series may be compatible with (weak) stationarity; this supports, but does not guarantee, stationarity assumptions for forecasting. This stationarity means that the statistical properties of these time series, such as their mean and variance, remain constant over time, which facilitates the use of temporal predictive models without requiring prior transformations to stabilize the series[221, 222]. This is a positive indication for future analyses, providing a solid basis for the application of advanced time-series analysis techniques.

APPENDIX J

ETHICAL CERTIFICATION

This thesis has received scientific and ethical certification from the Research Ethics Committee of the Centre intégré universitaire de santé et de services sociaux (CIUSSS) du Saguenay–Lac-Saint-Jean. The official information regarding this project is as follows:

This master thesis is based on the project entitled *"Amélioration de la prise de décision en approvisionnement par l'Intelligence Artificielle : Une approche basée sur le prétraitement et l'hyperparamétrisation pour la détection d'anomalie"*, conducted at the Centre intégré universitaire de santé et de services sociaux (CIUSSS) du Saguenay–Lac-Saint-Jean. The project was approved under project number **2024–001 (APPRO–AI)** and certificate number (Nagano) **2024–577**. Scientific and ethical approval was granted by the Research Ethics Committee (CER) on April 30, 2024, and remains valid until April 30, 2027, subject to renewal and continued compliance with applicable requirements.

Although this research does not directly involve human participants, it relies on the use and management of data requiring strict ethical oversight to ensure confidentiality, integrity, and responsible use. The ethical evaluation covered data collection, processing, analysis, and storage, confirming compliance with institutional and legal standards, particularly regarding information protection and risks related to secondary data use. The approval granted attests that the project meets the research ethics requirements of the CIUSSS and UQAC, and adheres to principles of scientific integrity and ethical conduct.

REFERENCES

[1] SciKit-Learn Developers. Comparing anomaly detection algorithms for outlier detection on toy datasets. Accessed on November 24th 2025. [Online]. Available: https://scikit-learn.org/stable/auto_examples/miscellaneous/plot_anomaly_comparison.html

[2] A. E. Curtis, T. A. Smith, B. A. Ziganshin, and J. A. Elefteriades, "The mystery of the z-score," *Aorta*, vol. 4, no. 04, pp. 124–130, 2016.

[3] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.

[4] Gouvernement du Québec. (2023) SSSS: Gouvernance et organisation des services. Accessed on November 24th 2025. [Online]. Available: https://www.msss.gouv.qc.ca/reseau/systeme-de-sante-et-de-services-sociaux-en-bref/gouvernance-et-organisation-des-services/

[5] ——. (1971) Loi sur la santé et les services sociaux. Accessed on November 24th 2025. [Online]. Available: https://www.legisquebec.gouv.qc.ca/fr/document/lc/s-4.2

[6] ——. (2023) SSSS: Contexte. Accessed on November 24th 2025. [Online]. Available: https://www.msss.gouv.qc.ca/reseau/systeme-de-sante-et-de-services-sociaux-en-bref/contexte/

[7] ——. (2015) Loi modifiant l'organisation et la gouvernance du réseau de la santé et des services sociaux notamment par l'abolition des agences régionales. Accessed on November 24th 2025. [Online]. Available: https://www.legisquebec.gouv.qc.ca/fr/document/lc/o-7.2

[8] ——. (2023) Loi sur la gouvernance du système de santé et de services sociaux (lgssss). Accessed on November 24th 2025. [Online]. Available: https://www.legisquebec.gouv.qc.ca/fr/document/lc/G-1.021

[9] ——. (2023) Projet de loi n° 15, loi visant à rendre le système de santé et de services sociaux plus efficace. Accessed on November 24th 2025. [Online]. Available: https://www.assnat.qc.ca/fr/travaux-parlementaires/projets-loi/projet-loi-15-43-1.html?appelant=MC

[10] ——. (2025) Transformation du système de santé et de services sociaux. Accessed on November 24th 2025. [Online]. Available: https://www.quebec.ca/sante/systeme-et-services-de-sante/organisation-des-services/systeme-quebecois-de-sante-et-de-services-sociaux/transformation-systeme-sante/reseau-sante-efficace-population

[11] ——. (2018) Centres intégrés de santé et de services sociaux (cisss) et centres intégrés universitaires de santé et de services sociaux (ciusss). Accessed on November 24th 2025. [Online]. Available: https://www.quebec.ca/sante/systeme-et-services-de-sante/organisation-des-services/cisss-et-ciusss

[12] Ministère de la Santé et des Services Sociaux. (2020) Cadre de référence pour l'approvisionnement responsable. Accessed on November 24th 2025. [Online]. Available: https://publications.msss.gouv.qc.ca/msss/fichiers/2020/20-733-01W.pdf

[13] Gouvernement du Québec. (2025) Plan santé. Accessed on November 24th 2025. [Online]. Available: https://cdn-contenu.quebec.ca/cdn-contenu/gouvernement/MCE/memoires/Plan_Sante.pdf

[14] Centre intégré universitaire de santé et services sociaux du Saguenay—Lac-Saint-Jean. (2018) Politique d'acquisition de biens et/ou de services et de gestion contractuelle. Accessed on November 24th 2025. [Online]. Available: https://santesaglac.gouv.qc.ca/medias/2022/09/po-rf.001_-_politique_dacquisition_de_biens_et-ou_de_services_et_de_gestion_contractuelle_2018-06-14.pdf

[15] Gouvernement du Québec. (2006) Loi sur les contrats des organismes publics. C-65.1. Accessed on November 24th 2025. [Online]. Available: https://www.legisquebec.gouv.qc.ca/fr/document/lc/c-65.1

[16] Statistique Canada. (2021) Répercussions de la covid-19 sur les travailleurs de la santé : prévention et contrôle des infections (rctspci). [Online]. Available: https://www23.statcan.gc.ca/imdb/p2SV_f.pl?Function=getSurvey&SDDS=5340

[17] Nations Unies. (2020) Covid-19 : des interruptions de services de santé essentiels signalés dans 90% des pays. Accessed on November 24th 2025. [Online]. Available: https://news.un.org/fr/story/2020/08/1076142

[18] CNESST. (2023) Covid-19 et autres infections respiratoires. [Online]. Available: https://www.cnesst.gouv.qc.ca/fr/prevention-securite/milieu-travail-sain/covid-19-autres-infections-respiratoires

[19] Statistique Canada. (2022) Enquête sur l'équipement de protection individuelle (eepi). [Online]. Available: https://www23.statcan.gc.ca/imdb/p2SV_f.pl?Function=getSurvey&SDDS=5332#a2

[20] ——. (2022) Demande et offre d'équipement de protection individuelle. [Online]. Available: https://www150.statcan.gc.ca/t1/tbl1/fr/tv.action?pid=1310078601

[21] ——. (2022) Accès des travailleurs de la santé à l'équipement de protection individuelle pendant la pandémie de covid-19. [Online]. Available: https://www.statcan.gc.ca/o1/fr/plus/2296-acces-des-travailleurs-de-la-sante-lequipement-de-protection-individuelle-pendant-la

[22] ——. (2023) Enquête canadienne sur l'incapacité, 2022 : Guide des concepts et méthodes. [Online]. Available: https://www150.statcan.gc.ca/n1/pub/89-654-x/89-654-x2023004-fra.htm

[23] ——. (2022) Tendances dans le secteur de la fabrication liées à la pandémie de covid-19 et aux perturbations de la chaîne d'approvisionnement. [Online]. Available: https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2022050-fra.htm

[24] Affaires Mondiales Canada. (2021) Vulnérabilité des industries canadiennes aux perturbations dans les chaînes d'approvisionnement mondiales. [Online]. Available: https://www.international.gc.ca/trade-commerce/economist-economiste/analysis-analyse/supply-chain-vulnerability.aspx?lang=fra

[25] Association pour la sante publique du Quebec. (2023) Une santé publique forte au coeur d'un système de santé efficace et efficient. [Online]. Available: https://www.assnat.qc.ca/Media/Process.aspx?MediaId=ANQ.Vigie.Bll.DocumentGenerique_190251&process=Default&token=ZyMoxNwUn8ikQ+TRKYwPCjWrKwg+vIv9rjij7p3xLGTZDmLVSmJLoqe/vG7/YWzz

[26] Agence de la santé publique du Canada. (2023) Une vision de la surveillance de la santé publique au canada d'ici 2030 : Guide de discussion technique. [Online]. Available: https://www.canada.ca/fr/sante-publique/programmes/

consultation-vision-evaluation-sante-publique-avenir-canada/document-technique.html

[27] S. Painuly, S. Sharma, and P. Matta, "Artificial Intelligence in e-Healthcare Supply Chain Management System: Challenges and Future Trends," in *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*. IEEE, 2023, pp. 569–574.

[28] A. Kumar, V. Mani, V. Jain, H. Gupta, and V. Venkatesh, "Managing healthcare supply chain through artificial intelligence (AI): A study of critical success factors," *Computers and Industrial Engineering*, vol. 175, p. 108815, 2023.

[29] P. Zhou and S. Abbaszadeh, "Towards Real-Time Machine Learning for Anomaly Detection," in *2020 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*. IEEE, 2020, pp. 1–3.

[30] A. Lavin and S. Ahmad, "Evaluating Real-Time Anomaly Detection Algorithms – The Numenta Anomaly Benchmark," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2015, pp. 38–44.

[31] S. Jagadeesan, D. K. Malik, S. Bharti, S. P. Singh, R. K. Ibrahim, and M. B. Alazzam, "Artificial Intelligence in Supply Chain Management in Industry 4.0," in *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. IEEE, 2023, pp. 2722–2726.

[32] A. E. Glaser, J. P. Harrison, and D. Josephs, "Anomaly detection methods to improve supply chain data quality and operations," *SMU Data Science Review*, vol. 6, no. 1, p. 3, 2022.

[33] Bureau du verificateur general du Canada. (2017) Gérer le risque de fraude. [Online]. Available: https://www.oag-bvg.gc.ca/internet/Francais/parl_oag_201705_01_f_42223.html

[34] Services publics et Approvisionnement Canada. (2021) Transparence, gestion des risques et intégrité : Comité permanent des comptes publics. [Online]. Available: https://www.tpsgc-pwgsc.gc.ca/trans/documentinfo-briefingmaterial/pacp/2021-05-27/p3-fra.html

[35] Institut du Québec. (2021) Le bond des postes vacants au

québec montre que la pénurie de main-d'œuvre s'accentue, particulièrement en santé. [Online]. Available: https://institutduquebec.ca/le-bond-des-postes-vacants-au-quebec-montre-que-la-penurie-de-main-doeuvre-saccentue-particuliereme

[36] Gouvernement du Canada. (2021) Audit of procurement. Veterans Affairs, Audit and evaluation division. Accessed on November 24th 2025. [Online]. Available: https://publications.gc.ca/collections/collection_2021/acc-vac/V32-434-2020-eng.pdf

[37] Cabinet Demers Beaulne. (2023) Fraude en approvisionnement. [Online]. Available: https://www.demersbeaulne.com/2023/04/03/fraude-en-approvisionnement/

[38] BDO Canada. (2019) Fraude au sein de la chaîne d'approvisionnement : conséquences pour les organisations et réduction de l'exposition. [Online]. Available: https://www.bdo.ca/fr-ca/insights/guide-leading-supply-chain-practices

[39] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[40] J. Zhan and X. Fang, "Anomaly detection in social-economic computing," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE, 2011, pp. 695–703.

[41] G. P. Cachon and M. Fisher, "Supply chain inventory management and the value of shared information," *Management science*, vol. 46, no. 8, pp. 1032–1048, 2000.

[42] G. Nenes, S. Panagiotidou, and G. Tagaras, "Inventory management of multiple items with irregular demand: A case study," *European Journal of Operational Research*, vol. 205, no. 2, pp. 313–324, 2010.

[43] F. A. Roesch and P. C. Van Deusen, "Anomaly detection for analysis of annual inventory data: A quality control approach," *Southern Journal of Applied Forestry*, vol. 34, no. 3, pp. 131–137, 2010.

[44] R.W. Tatko, C.R. White, G.M. Williams, R.K. Myers, "Backorder prediction," IBM Corporation U.S. Federal for Defense logistics agency, R&D, Tech. Rep., 2021.

[45] Gouvernement du Canada. (2022) Pénurie d'instruments médicaux. Accessed on November 24th 2025. [Online]. Available: https://www.canada.ca/fr/sante-canada/services/medicaments-produits-sante/instruments-medicaux/penuries.html

[46] K. S. Park, "Inventory model with partial backorders," *International Journal of Systems Science*, vol. 13, no. 12, pp. 1313–1317, 1982.

[47] V. M. Pillai and T. C. Pamulety, "Impact of backorder on supply chain performance–an experimental study," *IFAC Proceedings Volumes*, vol. 46, no. 9, pp. 1932–1937, 2013.

[48] National Academies of Sciences, Engineering, and Medicine., *Supply Chain Challenges and Solutions amid COVID-19*. Washington, DC: The National Academies Press, 2025, accessed on November 24th 2025. [Online]. Available: https://nap.nationalacademies.org/read/29153/chapter/1

[49] Berkeley University Glossary. (2023) Anomaly. [Online]. Available: https://undsci.berkeley.edu/glossary/anomaly/

[50] C. C. Aggarwal, *An introduction to outlier analysis*, 2nd ed. Cham: Springer International Publishing, 2016.

[51] S. Thudumu, P. Branch, J. Jin, and J. Singh, "A comprehensive survey of anomaly detection techniques for high dimensional big data," *Journal of Big Data*, vol. 7, no. 1, p. 42, 2020. [Online]. Available: https://doi.org/10.1186/s40537-020-00320-x

[52] M. E. K. Niessen, J. M. Paciello, and J. I. P. Fernandez, "Anomaly detection in public procurements using the open contracting data standard," in *2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG)*. IEEE, Apr. 2020, pp. 127–134.

[53] A. Ahmadi, M. S. Pishvaee, and S. A. Torabi, "Procurement management in healthcare systems," in *Operations research applications in health care management*. Cham: Springer International Publishing, 2017, pp. 569–598.

[54] L. F. Carvalho, C. H. Teixeira, W. Meira, M. Ester, O. Carvalho, and M. H. Brandao, "Provider-consumer anomaly detection for healthcare systems," in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, Aug. 2017, pp. 229–238.

[55] Paul, P. O., Ogugua, J. O., Eyo-Udo, N. L., "Procurement in healthcare: Ensuring efficiency and compliance in medical supplies and equipment management," *Journal of Healthcare Procurement and Supply Chain Management*, vol. 6, no. 1, pp. 19–34, 2024.

[56] N. Privett and D. Gonsalvez, "The top ten global health supply chain issues: Perspectives from the field," *Operations Research for Health Care*, vol. 3, no. 4, pp. 226–230, 2014.

[57] S. Cho and H. Zhao, "Healthcare supply chain," *Handbook of Healthcare Analytics: Theoretical Minimum for Conducting 21st Century Research on Healthcare Operations*, pp. 159–185, 2018.

[58] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: methods, systems and tools," *IEEE communications surveys & tutorials*, vol. 16, no. 1, pp. 303–336, 2013.

[59] W. Hilal, S. A. Gadsden, and J. Yawney, "Financial fraud: a review of anomaly detection techniques and recent advances," *Expert systems with applications*, vol. 193, p. 116429, 2022.

[60] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Deep learning for medical anomaly detection–a survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 7, pp. 1–37, 2021.

[61] P. Kamat and R. Sugandhi, "Anomaly detection for predictive maintenance in industry 4.0-A survey," in *E3S web of conferences*, vol. 170. EDP Sciences, 2020, p. 02007.

[62] C. Jacinto, D. López, I. Aguilera-Martos, D. García-Gil, I. Markova, M. Garcia-Barzana, M. Arias-Rodil, J. Luengo, and F. Herrera, "Anomaly detection in predictive maintenance: A new evaluation framework for temporal unsupervised anomaly detection algorithms," *Neurocomputing*, vol. 462, pp. 440–452, 2021.

[63] M. Krichen, "Détection des anomalies environnementales via les capteurs des smartphones," Doctoral dissertation, Sfax University, ReDCAD Laboratory, 2021, ph.D Thesis.

[64] J. Sipple and A. Youssef, "A general-purpose method for applying explainable ai for anomaly detection," in *International Symposium on Methodologies for Intelligent Systems*.

Springer, 2022, pp. 162–174.

[65] C.-W. Ten, J. Hong, and C.-C. Liu, "Anomaly detection for cybersecurity of the substations," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 865–873, 2011.

[66] B. Genge, P. Haller, and C. Enăchescu, "Anomaly detection in aging industrial internet of things," *IEEE Access*, vol. 7, pp. 74 217–74 230, 2019.

[67] J. Shi, G. He, and X. Liu, "Anomaly detection for key performance indicators through machine learning," in *2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, 2018, pp. 1–5.

[68] M. Malik, S. Abdallah, and M. Ala'raj, "Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review," *Annals of Operations Research*, vol. 270, pp. 287–312, 2018.

[69] L. Stojanovic, M. Dinic, N. Stojanovic, and A. Stojadinovic, "Big-data-driven anomaly detection in industry (4.0): An approach and a case study," in *2016 IEEE international conference on big data (big data)*. IEEE, 2016, pp. 1647–1652.

[70] R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, and M. Imran, "Real-time big data processing for anomaly detection: A survey," *International Journal of Information Management*, vol. 45, pp. 289–307, 2019.

[71] T. Ebel, E. Larsen, Ketan Shah. (2013) Strengthening health care's supply chain: A five-step plan. [Online]. Available: https://www.mckinsey.com/industries/healthcare/our-insights/strengthening-health-cares-supply-chain-a-five-step-plan

[72] A. W. Snowdon, M. Saunders, and A. Wright, "The Emerging Features of Healthcare Supply Chain Resilience: Learning from a Pandemic," *Healthcare Quarterly (Toronto, Ont.)*, vol. 25, no. 2, pp. 44–53, 2022.

[73] D. Friday, D. A. Savage, S. A. Melnyk, N. Harrison, S. Ryan, and H. Wechtler, "A collaborative approach to maintaining optimal inventory and mitigating stockout risks during a pandemic: capabilities for enabling health-care supply chain resilience," *Journal of Humanitarian Logistics and Supply Chain Management*, vol. 11, no. 2, pp. 248–271,

2021.

[74] H. D. Nguyen, K. P. Tran, S. Thomassey, and M. Hamad, "Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management," *International Journal of Information Management*, vol. 57, p. 102282, 2021.

[75] L. F. Carvalho, C. H. Teixeira, W. Meira, M. Ester, O. Carvalho, and M. H. Brandao, "Provider-consumer anomaly detection for healthcare systems," in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2017, pp. 229–238.

[76] V. Haldane, C. De Foo , S.M. Abdalla, A.S. Jung, M. Tan, S. Wu, A. Chua, M. Verma, P. Shrestha, S. Singh, T. Perez, "Health systems resilience in managing the COVID-19 pandemic: lessons from 28 countries," *Nature Medicine*, vol. 27, no. 6, pp. 964–980, 2021/06/01.

[77] D. Gohil and S. V. Thakker, "Blockchain-integrated technologies for solving supply chain challenges," *Modern Supply Chain Research and Applications*, vol. 3, no. 2, pp. 78–97, 2021.

[78] M. Ratia, J. Myllärniemi, and N. Helander, "Robotic process automation-creating value by digitalizing work in the private healthcare?" in *Proceedings of the 22nd International Academic Mindtrek Conference*, 2018, pp. 222–227.

[79] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422.

[80] N. A. Campbell, "Robust procedures in multivariate analysis I: Robust covariance estimation," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 29, no. 3, pp. 231–237, 1980.

[81] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Support vector method for novelty detection," in *Advances in Neural Information Processing Systems (NIPS 12)*, 2000, pp. 582–588.

[82] Y. Chen, X. S. Zhou, and T. S. Huang, "One-class SVM for learning in image retrieval," in

Proceedings 2001 international conference on image processing (Cat. No. 01CH37205), vol. 1. IEEE, 2001, pp. 34–37.

[83] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000, pp. 93–104.

[84] R. A. Fisher, *Statistical methods for research workers*. Springer New York, 1970.

[85] Gouvernement du Québec. (2020) Loi sur le centre d'acquisitions gouvernementales. RLRQ c C-7.01, Accessed on November 24th 2025. [Online]. Available: https://www.legisquebec.gouv.qc.ca/fr/document/lc/C-7.01

[86] M. O. Silva, L. G. L. Costa, L. D. Gomide, G. Bezerra, G. P. Oliveira, M. A. Brandão, A. Lacerda, and G. Pappa, "Overpricing analysis in brazilian public bidding items," *Journal on Interactive Systems*, vol. 15, no. 1, 2024. [Online]. Available: https://doi.org/10.5753/jis.2024.3831

[87] D. Bandaly, A. Satir, and L. Shanker, "Impact of lead time variability in supply chain risk management," *International Journal of Production Economics*, vol. 180, Jul. 2016. [Online]. Available: https://doi.org/10.1016/j.ijpe.2016.07.014

[88] V. Kumar, D. Ekwall, and L. Wang, "Supply chain strategies for quality inspection under a customer return policy: A game theoretical approach," *Entropy*, vol. 18, no. 12, p. 440, 2016. [Online]. Available: https://doi.org/10.3390/e18120440

[89] K. Hong, Y. Ren, F. Li, W. Mao, and Y. Liu, "Unsupervised anomaly detection of intermittent demand for spare parts based on dual-tailed probability," *Electronics*, vol. 13, no. 1, p. 195, 2024. [Online]. Available: https://doi.org/10.3390/electronics13010195

[90] A. Herreros-Martínez, R. Magdalena-Benedicto, J. Vila-Francés, A. J. Serrano-López, S. Pérez-Díaz, and J. J. Martínez-Herráiz, "Applied machine learning to anomaly detection in enterprise purchase processes: A hybrid approach using clustering and isolation forest," *Information*, vol. 16, no. 3, p. 177, 2025. [Online]. Available: https://doi.org/10.3390/info16030177

[91] J. Liu, W. Gu, Z. Chen, Y. Li, Y. Su, and M. R. Lyu, "Mtad: Tools and benchmarks for multivariate time series anomaly detection," 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2401.06175

[92] M. Gabellini, L. Civolani, F. Calabrese, and M. Bortolini, "A deep learning approach to predict supply chain delivery delay risk based on macroeconomic indicators: A case study in the automotive sector," *Applied Sciences*, vol. 14, no. 11, p. 4688, 2024. [Online]. Available: https://doi.org/10.3390/app14114688

[93] E. Guijarro, E. Babiloni, and M. Cardós, "On the estimation of the fill rate for the continuous (s, s) inventory system for the lost sales context," *PLOS ONE*, vol. 17, no. 2, p. e0263655, 2022. [Online]. Available: https://doi.org/10.1371/journal.pone.0263655

[94] N. Usman, E. Utami, and A. D. Hartanto, "Comparative analysis of elliptic envelope, isolation forest, one-class svm, and local outlier factor in detecting earthquakes with status anomaly using outlier," in *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*. IEEE, Feb. 2023, pp. 673–678.

[95] J. Lever, M. Krzywinski, and N. Altman, "Classification evaluation," *Nature Methods*, vol. 13, no. 8, pp. 603–604, 2016. [Online]. Available: https://doi.org/10.1038/nmeth.3945

[96] N. Elmrabit, F. Zhou, F. Li, and H. Zhou, "Evaluation of machine learning algorithms for anomaly detection," in *2020 international conference on cyber security and protection of digital services (cyber security)*. IEEE, 2020, pp. 1–8.

[97] A. Cools, M. A. Belarbi, and S. A. Mahmoudi, "Benchmarking of anomaly detection methods for industry 4.0: Evaluation, ranking, and practical recommendations," *Big Data and Cognitive Computing*, vol. 9, no. 5, p. 128, 2025.

[98] R. Xin, H. Liu, P. Chen, and Z. Zhao, "Robust and accurate performance anomaly detection and prediction for cloud applications: a novel ensemble learning-based framework," *Journal of Cloud Computing*, vol. 12, no. 1, p. 7, 2023. [Online]. Available: https://doi.org/10.1186/s13677-022-00383-6

[99] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[100] C. Marzban, "The ROC curve and the area under it as performance measures," *Weather and Forecasting*, vol. 19, no. 6, pp. 1106–1114, 2004.

[101] S. Yang, G.Berdin, "The receiver operating characteristic (roc) curve," *The Southwest Respiratory and Critical Care Chronicles*, vol. 5, no. 19, pp. 34–36, 2017.

[102] H.R. Sofaer, J.A. Hoeting, C.S. Jarnevich, "The area under the precision–recall curve as a performance metric for rare binary events," *Methods in Ecology and Evolution*, vol. 10, no. 4, pp. 565–577, 2019.

[103] J. Keilwagen, I. Grosse, and J. Grau, "Area under precision-recall curves for weighted and unweighted data," *PLOS ONE*, vol. 9, no. 3, p. e92209, 2014.

[104] P. Probst and A.-L. Boulesteix and B. Bischl, "Tunability: Importance of hyperparameters of machine learning algorithms," *The Journal of Machine Learning Research*, vol. 20, no. 53, pp. 1934–1965, 2019.

[105] H. Li, P. Chaudhari, H. Yang, M. Lam, A. Ravichandran, R. Bhotika, and S. Soatto, "Rethinking the hyperparameters for fine-tuning," *arXiv preprint arXiv:2002.11770*, 2020.

[106] A. M. Vincent and P. Jidesh, "An improved hyperparameter optimization framework for automl systems using evolutionary algorithms," *Scientific Reports*, vol. 13, no. 1, p. 4737, 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-32027-3

[107] F. Karl, T. Pielok, J. Moosbauer, F. Pfisterer, S. Coors, M. Binder, L. Schneider, J. Thomas, J. Richter, and M. Lang, "Multi-objective hyperparameter optimization–an overview," *arXiv preprint arXiv:2206.07438*, 2022.

[108] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, "Hyperparameter optimization for machine learning models based on bayesian optimization," *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, 2019.

[109] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[110] scikit-learn developers, "EllipticEnvelope — scikit-learn documentation," https://scikit-learn.org/stable/modules/generated/sklearn.covariance.EllipticEnvelope.html, accessed on January 8 2026.

[111] J. Yang, S. Rahardja, P. Fränti, "Outlier detection: How to threshold outlier scores?" in *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, Dec. 2019, pp. 1–6.

[112] G. P. Spathoulas and S. K. Katsikas, "Methods for post-processing of alerts in intrusion detection: A survey," *International Journal of Information Security Science*, vol. 2, no. 2, pp. 64–80, 2013.

[113] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang, "Post-processing temporal action detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 837–18 845.

[114] J.-C. Wu, C.-M. Jiang, and C.-L. Huang, "Post-processing for improving hyperspectral anomaly detection accuracy," in *Image and Signal Processing for Remote Sensing XXI (Vol. 9643)*. SPIE, 2015, pp. 227–232.

[115] K.U. Simmer and J. Bitzer and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, ser. Digital Signal Processing, M. Brandstein and D. Ward, Eds. Berlin, Heidelberg: Springer, 2001, pp. 39–60.

[116] F. Jiru, "Introduction to post-processing techniques," *European Journal of Radiology*, vol. 67, no. 2, pp. 202–217, 2008.

[117] M. S. Alam and M. N. Islam and A. Bal, and M.A. Karim, "Hyperspectral target detection using gaussian filter and post-processing," *Optics and Lasers in Engineering*, vol. 46, no. 11, pp. 817–822, 2008.

[118] S.H. Sun and A.Sankararaman and B. M. Narayanaswamy, "Online adaptive anomaly thresholding with confidence sequences," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 235, 2024, pp. 47 105–47 132. [Online]. Available: https://proceedings.mlr.press/v235/sun24h.html

[119] M. Chen and C. Gao and Z. Ren, "Robust covariance and scatter matrix estimation under huber's contamination model," *The Annals of Statistics*, vol. 46, no. 5, pp. 1932–1960, 2018.

[120] G. P. Spathoulas and S. K. Katsikas, "Reducing false positives in intrusion detection systems," *Computers & Security*, vol. 29, no. 1, pp. 35–44, 2010.

[121] R. A. Bauder, T. M. Khoshgoftaar, A. Richter, and M. Herland, "Predicting medical provider specialties to detect anomalous insurance claims," in *2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI)*. IEEE, 2016, pp. 784–790.

[122] A. Westerski, R. Kanagasabai, E. Shaham, A. Narayanan, and M. S. J. Wong, "Explainable anomaly detection for procurement fraud identification—lessons from practical deployments," *International Transactions in Operational Research*, vol. 28, no. 6, pp. 3276–3302, 2021.

[123] D. Patel, D. Phan, and M. Mueller, "Time Series Anomaly Detection Toolkit for Data Scientist," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, 2022, pp. 3202–3204.

[124] M. Feurer and F. Hutter, *Hyperparameter optimization*. Springer International Publishing, 2019.

[125] Y.-Y. Lau, M. A. Dulebenets, H.-T. Yip, and Y.-M. Tang, "Healthcare supply chain management under covid-19 settings: The existing practices in hong kong and the united states," in *Healthcare*. MDPI, 2022, p. 1549.

[126] S. Fantini, *Business Intelligence avec SQL Server 2019 - Maîtrisez les concepts et réalisez un système décisionnel*. Edition ENI, 2019.

[127] X. Yan, C. Zhang, and S. Zhang, "Toward databases mining: Pre-processing collected data," *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 545–561, 2003.

[128] Y. Kenji, T. Jun-Ichi, "A unifying framework for detecting outliers and change points from non-stationary time series data," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2002, pp. 676–681.

[129] C. Bouchard, J. Maitre, and B. Bouchard, "Anomaly detection using ai in healthcare procurement data across multiple criteria with tailored post-processing filters," 2025, preprint included in this master's thesis (Chapter 2); manuscript in preparation and intended for submission.

[130] ——, "Ai-based anomaly detection in healthcare procurement using apache spark: a post-processing filter approach," 2025, preprint included in this master's thesis (Chapter 3); manuscript in preparation and intended for submission.

[131] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications," *The VLDB Journal*, vol. 8, no. 3, pp. 237–253, 2000.

[132] K. Pearson, "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

[133] S. Adams and P. A. Beling, "A survey of feature selection methods for gaussian mixture models and hidden markov models," *Artificial Intelligence Review*, vol. 52, pp. 1739–1779, 2019.

[134] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes," *Inequalities*, vol. 3, no. 1, pp. 1–8, 1972.

[135] S. Lange and M. Riedmiller, "Deep auto-encoder neural networks in reinforcement learning," in *The 2010 international joint conference on neural networks (IJCNN)*. IEEE, 2010, pp. 1–8.

[136] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[137] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[138] L. K. Lok, V. A. Hameed, and M. E. Rana, "Hybrid machine learning approach for anomaly

detection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 27, no. 2, p. 1016, 2022.

[139] E. J. Gumbel, *Statistical theory of extreme values and some practical applications: a series of lectures*. US Government Printing Office, 1948, vol. 33.

[140] ——, *Statistics of extremes*. Columbia university press, 1958.

[141] D. Figueiredo, L. Silva, A. Santos, and C. Malaquias, "Living with outliers: How to detect extreme observations in data analysis," *Revista Brasileira de Informação Bibliográfica em Ciências Sociais*, vol. 99, p. 1, 2023.

[142] C. Liu, L. Pan, Z. Gu, J. Wang, Y. Ren, and Z. Wang, "Valid probabilistic anomaly detection models for system logs," *Wireless Communications and Mobile Computing*, vol. 2020, p. 8827185, 2020. [Online]. Available: https://doi.org/10.1155/2020/8827185

[143] S. Madhuri and M. U. Rani, "Anomaly detection techniques," *SSRN Electronic Journal*, 2018.

[144] G. Pang, C. Shen, L. Cao, and A. Hengel, "Deep Learning for Anomaly Detection: A Review," *ACM computing surveys*, vol. 54, no. 2, 2021.

[145] Z. Z. Darban, G. I. Webb, S. Pan, C. C. Aggarwal, and M. Salehi, "Deep learning for time series anomaly detection: A survey," *arXiv preprint arXiv:2211.05244*, 2022.

[146] K. Lagemann, C. Lagemann, B. Taschler, and S. Mukherjee, "Deep learning of causal structures in high dimensions under data limitations," *Nature Machine Intelligence*, vol. 5, no. 11, pp. 1306–1316, 2023. [Online]. Available: https://doi.org/10.1038/s42256-023-00744-z

[147] B. Lusch, J. N. Kutz, and S. L. Brunton, "Deep learning for universal linear embeddings of nonlinear dynamics," *Nature Communications*, vol. 9, no. 1, p. 4950, 2018. [Online]. Available: https://doi.org/10.1038/s41467-018-07210-0

[148] E. Duede, "Deep learning opacity in scientific discovery," *Philosophy of Science*, vol. 90,

no. 5, pp. 1089–1099, 2023. [Online]. Available: https://www.cambridge.org/core/product/C46306D902A7AC87FD192D996639784A

[149] L. E. Rangkiti, F. K. Lubis, I. Muda, "Supply chain management: A review of anomaly detection techniques and the benford's law," *Russian Law Journal*, vol. 11, no. 6, 2023.

[150] Z. Li, Y. Zhu, and M. Van Leeuwen, "A survey on explainable anomaly detection," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 1, pp. 1–54, 2023.

[151] C.-H. Chang, J. Yoon, S. O. Arik, M. Udell, and T. Pfister, "Data-efficient and interpretable tabular anomaly detection," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 190–201.

[152] S. Russell, S. Hauert, R. Altman, and M. Veloso, "Ethics of artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 415–416, 2015.

[153] N. Bostrom and E. Yudkowsky, *The ethics of artificial intelligence*. Chapman and Hall, 2018.

[154] L. Wang, "Heterogeneous data and big data analytics," *Automatic Control and Information Sciences*, vol. 3, no. 1, pp. 8–15, 2017.

[155] J. Fan, F. Han, and H. Liu, "Challenges of Big Data analysis," *National Science Review*, vol. 1, no. 2, pp. 293–314, 2014.

[156] N. Truong, K. Sun, S. Wang, F. Guitton, and Y. Guo, "Privacy preservation in federated learning: An insightful survey from the gdpr perspective," *Computers & Security*, vol. 110, p. 102402, 2021.

[157] M. E. K. Niessen, J. M. Paciello, and J. I. P. Fernandez, "Anomaly Detection in Public Procurements using the Open Contracting Data Standard," in *2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG)*. IEEE, 2020, pp. 127–134.

[158] N. Modrušan, K. Rabuzin, and L. Mršic, "Review of public procurement fraud detection techniques powered by emerging technologies," *International Journal of Advanced*

Computer Science and Applications, vol. 12, no. 2, 2021.

[159] M. Veale and I. Brass, "Administration by algorithm? Public management meets public sector machine learning," *Algorithmic Regulation, Oxford University Press*, 2019.

[160] A. Protogerou, S. Papadopoulos, A. Drosou, D. Tzovaras, and I. Refanidis, "A graph neural network method for distributed anomaly detection in IoT," *Evolving Systems : An Interdisciplinary Journal for Advanced Science and Technology*, vol. 12, no. 1, pp. 19–36, 2020.

[161] K. P. Tran, H. D. Nguyen, and S. Thomassey, "Anomaly detection using long short term memory networks and its applications in supply chain management," in *IFAC-PapersOnLine*, vol. 52, no. 13, 2019, pp. 2408–2412.

[162] P. P. Ippolito, "Hyperparameter tuning: the art of fine-tuning machine and deep learning models to improve metric results," in *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*. Springer International Publishing, 2022, pp. 231–251.

[163] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[164] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, p. 281–305, 2012.

[165] C. Chen, X. Chen, C. Ma, Z. Liu, and X. Liu, "Gradient-based bi-level optimization for deep learning: a survey," *arXiv preprint arXiv:2207.11719*, 2022.

[166] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: a review of bayesian optimization," *Proceedings of the IEEE*, vol. 104, no. 1, p. 148–175, 2016.

[167] T. H.W. Bäck, A. V. Kononova, B. V. Stein, H. Wang, K. A. Antonov, R. T. Kalkreuth, J. D. Nobel, D. Vermetten, R. D. Winter, F. Ye, "Evolutionary algorithms for parameter optimization—thirty years of continuous black-box optimization," *Evolutionary Computation*,

vol. 31, no. 2, p. 81–115, 2023.

[168] D. M. Belete and M. D. Huchaiah, "Grid search in hyperparameter optimization of machine learning models for prediction of hiv/aids test results," *International Journal of Computers and Applications*, vol. 44, no. 9, pp. 875–886, 2022.

[169] T. Agrawal, "Hyperparameter optimization using scikit-learn," in *Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient*. Berkeley, CA: Apress, 2020, pp. 31–51.

[170] O. Kramer, "Scikit-learn," in *Machine Learning for Evolution Strategies*, ser. Studies in Big Data. Cham: Springer International Publishing, 2016, vol. 20, pp. 45–53.

[171] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2012, pp. 2951–2959.

[172] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.

[173] N. W. M. Binois, "A survey on high-dimensional gaussian process modeling with application to bayesian optimization," *arXiv preprint arXiv:2111.05040*, 2021.

[174] S. R. Young, D. C. Rose, T. P. Karnowski, S.-H. Lim, and R. M. Patton, "Optimizing deep learning hyper-parameters through an evolutionary algorithm," in *Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments (MLHPC '15)*. ACM / IEEE (workshop), 2015, pp. 1–5.

[175] O. Schmidt, S. Safarani, J. Gastinger, T. Jacobs, S. Nicolas, and A. Schülke, "On the performance of differential evolution for hyperparameter tuning," *arXiv preprint arXiv:1904.06960*, 2019.

[176] Y. Huan, F. Wu, M. Basios, L. Kanthan, L. Li, and B. Xu, "Ieo: Intelligent evolutionary optimisation for hyperparameter tuning," in *Advances in Machine Learning – Applications, Tools and Techniques*. Cham, Switzerland: Springer International Publishing, 2020.

[177] B. Bischl, M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, J. Thomas, T. Ullmann, M. Becker, A.-L. Boulesteix, D. Deng, and M. Lindauer, "Hyperparameter optimization: Foundations, algorithms, best practices and open challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13, no. 3, p. e1484, 2023.

[178] J. Soenen and E. V. Wolputte and L. Perini and V. Vercruyssen and W. Meert and J. Davis, H. Blockeel, "The effect of hyperparameter tuning on the comparative evaluation of anomaly detection algorithms," in *Workshop on Learning with Imbalanced Domains (LID) / Odd Workshop*, 2021.

[179] M. Mohit, G. Dasarathy, and A. Spanias, "Bayesian optimization in high-dimensional spaces," *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, vol. 6, no. 4, pp. 1–25, 2021.

[180] A. Garg, W. Zhang, J. Samaran, S. Ramasamy, and C.-S. Foo, "An evaluation of anomaly detection and diagnosis in multivariate time series," *arXiv preprint arXiv:2109.11428*, 2021. [Online]. Available: https://arxiv.org/abs/2109.11428

[181] J. Qiu, H. Shi, Y. Hu, and Z. Yu, "Enhancing anomaly detection models for industrial applications through svm-based false positive classification," *Applied Sciences*, vol. 13, no. 23, p. 12655, 2023. [Online]. Available: https://doi.org/10.3390/app132312655

[182] K. J. Prabhod, "The role of artificial intelligence in reducing healthcare costs and improving operational efficiency," *Quarterly Journal of Emerging Technologies and Innovations*, vol. 9, no. 2, pp. 47–59, 2024.

[183] P.J.M. Ali, R.H. Faraj, "Data normalization and standardization: a technical report," Koya University Publishing, Tech. Rep. 1, 2014.

[184] L. V. D. Maaten, E. O. Postma, and H. J. V. D. Herik, "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research*, vol. 10, no. 66-71, p. 13, 2009.

[185] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[186] J. B. M. Fazekas, "Improving public procurement outcomes," *World Bank Group*, 2021. [Online]. Available: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/656521623167062285/improving-public-procurement-outcomes-review-of-tools-and-the-state-of-the-evidence-base

[187] L. Li, "The role of inventory in delivery-time competition," *Management science*, vol. 38, no. 2, pp. 182–197, 1992.

[188] G. Graman and D. Rogers, "Delivery delay variation in multi-echelon inventory problems," *Journal of the Operational Research Society*, vol. 48, no. 10, pp. 1029–1036, 1997.

[189] R. Lolla, M. Harper, J. Lunn, J. Mustafina, J. Assi, C.K. Loy, D. Al-Jumeily OBE, "Machine Learning Techniques for Predicting Risks of Late Delivery," in *The International Conference on Data Science and Emerging Technologies*. Springer Nature Singapore, 2022, pp. 343–356.

[190] A. Veneziano, "Non conformity of goods in international sales - a survey of current caselaw on CISG," *International Business Law Journal / RDAI*, p. 39–66, 1997.

[191] P. Walley, K. Silvester, and R. Steyn, "Managing variation in demand: lessons from the UK National Health Service," *Journal of healthcare management*, vol. 51, no. 5, pp. 309–320, 2006.

[192] G. G. F. Bergadano, *AI for Cybersecurity : robust models for authentication, threat and anomaly detection*. MDPI, 2023, vol. 16, no. 7, special issue reprint. [Online]. Available: https://directory.doabooks.org/handle/20.500.12854/112521

[193] J. Shreffler and M. R. Huecker, *Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios*. StatPearls Publishing, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK557491/

[194] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 1–11, 2015.

[195] K. Levy, K. E. Chasalow, and S. Riley, "Algorithms and Decision-Making in the Public

Sector," *Annual Review of Law and Social Science*, vol. 17, no. 1, pp. 309–334, 2021.

[196] M. Martínez-García and E. Lemus, "Health systems as complex systems," *American Journal of Operations Research*, vol. 3, no. 1A, pp. 113–126, 2013.

[197] National Academies of Sciences, Engineering, and Medicine, *Understanding medical product supply chains*. Washington, DC: The National Academies Press, 2022.

[198] F. G. Bruno and A. C. Santos, "A global scoping review on the patterns of medical fraud and abuse," *International Journal of Health Policy and Management*, vol. 12, no. 3, pp. 101–121, 2024.

[199] A. Fernández-Medina, D. Ramírez-García, and P. Sánchez, "Discovering anomalies in big data: A review focused on machine learning," *Frontiers in Big Data*, vol. 6, 2023.

[200] H. Huang, P. Wang, J. Pei, J. Wang, S. Alexanian, and D. Niyato, "Deep learning advancements in anomaly detection: A comprehensive survey," *arXiv preprint arXiv:2503.13195*, 2025.

[201] M. E. K. Niessen, J. M. Paciello, and J. I. Pane Fernández, "Anomaly detection in public procurements: Detecting fraud through procurement data mining using unsupervised models," in *Proceedings of the Open Contracting Data Standard Conference*, 2023.

[202] E. F. Agyemang, "Anomaly detection using unsupervised machine learning algorithms: A simulation study," *Scientific African*, vol. 26, p. e02386, 2024.

[203] M. Alvarez, J.-C. Verdier, D. K. Nkashama, M. Frappier, P.-M. Tardif, and F. Kabanza, "A revealing large-scale evaluation of unsupervised anomaly detection algorithms," *arXiv preprint arXiv:2204.09825*, 2022.

[204] V. I. Madai and D. C. Higgins, "Artificial intelligence in healthcare: Lost in translation?" *arXiv preprint arXiv:2107.13454*, 2021.

[205] M.J. Leming, E.E. Bron, R. Bruffaerts, Y. Ou, J.E. Iglesias, R.L. Gollub, H. Im, "Challenges of implementing computer-aided diagnostic models for neuroimages in a clinical setting,"

NPJ Digital Medicine, vol. 6, no. 1, 2023.

[206] M. T. R. Laskar, J. Huang, V. Smetana, C. Stewart, K. Pouw, A. An, S. Chan, and L. Liu, "Extending isolation forest for anomaly detection in big data via k-means," *arXiv preprint arXiv:2104.13190*, 2021.

[207] S. Zhang, V. Ursekar, and L. Akoglu, "Sparx: Distributed outlier detection at scale," *arXiv preprint arXiv:2206.01281*, 2022.

[208] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, "A review of local outlier factor algorithms for outlier detection in big data streams," *Big Data and Cognitive Computing*, vol. 5, no. 1, p. 1, 2021.

[209] Y. Yan, L. Zhang, L. Akoglu, X. Wang, and J. Huang, "Distributed local outlier detection in big data," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 1537–1546.

[210] S. S. Khan and M. G. Madden, "One-class classification: Taxonomy of study and review of techniques," *Knowledge Engineering Review*, vol. 29, no. 3, pp. 374–397, 2014.

[211] M. Nardi, L. Valerio, A. Passarella, "Centralised vs decentralised anomaly detection: When local and global patterns matter," in *Proceedings of Machine Learning Research, vol. 154*, 2021, pp. 1–12.

[212] T. L. Marzetta, G. H. Tucci, and S. H. Simon, "A random matrix–theoretic approach to handling singular covariance estimates," *arXiv preprint arXiv:1010.0601*, 2010.

[213] M. Khodak, R. Tu, T. Li, L. Li, M.-F. Balcan, V. Smith, and A. Talwalkar, "Federated hyperparameter tuning: Challenges, baselines, and connections to weight-sharing," *arXiv preprint arXiv:2106.04502*, 2021.

[214] Castillo, Enrique, Peteiro-Barral, Diego, Guijarro-Berdinás, Bertha, Fontenla-Romero, Óscar, "Distributed one-class support vector machine," *International Journal of Neural Systems*, vol. 23, no. 4, p. 1–24, 2013.

[215] P. Royston, "Approximating the shapiro-wilk w-test for non-normality," *Statistics and computing*, vol. 2, pp. 117–119, 1992.

[216] F. E. Grubbs, "Sample criteria for testing outlying observations," *The Annals of Mathematical Statistics*, vol. 21, no. 1, pp. 27–58, 1950.

[217] V. Barnett and T. Lewis, *Outliers in Statistical Data*, 3rd ed. John Wiley & Sons, 1994.

[218] B. Iglewicz and D. C. Hoaglin, *How to Detect and Handle Outliers*, ser. ASQC Basic References in Quality Control. ASQC Quality Press, 1993, vol. 16.

[219] D. A. Belsley, V. Klema. (1974) Detecting and assessing the problems caused by multicollinearity: a use of the singular-value decomposition. Accessed on November 24th 2025. [Online]. Available: https://www.nber.org/system/files/working_papers/w0066/w0066.pdf

[220] Y. Chen, "An analytical process of spatial autocorrelation functions based on moran's index," *PLoS One*, vol. 16, no. 4, p. e0249589, 2021.

[221] R. Mushtaq, "Augmented dickey fuller test," Ph.D. dissertation, Université Paris, 2011, ph.D Thesis.

[222] C. F. Baum, "sts15: Tests for stationarity of a time series," *Stata Technical Bulletin*, vol. 57, pp. 36–39, 2000.

[223] G.R. Terrell and D.W. Scott, "Variable kernel density estimation," *The Annals of Statistics*, vol. 20, no. 3, pp. 1236–1265, 1992.

[224] G.E.P. Box and G.M. Jenkins and G.C. Reinsel, *Time Series Analysis: Forecasting and Control*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall, 1994.

[225] R.J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.

[226] G. S. Mudholkar, D. K. Srivastava, and C. Thomas Lin, "Some p-variate adaptations of the shapiro-wilk test of normality," *Communications in Statistics-Theory and Methods*, vol. 24,

no. 4, pp. 953–985, 1995.

[227] R. Zaka, "Strongly consistent determination of the rank of matrix," *EERI Research Paper Series*, no. 4/2003, 2003.

[228] M. S. Lewis-Beck and A. Skalaban, "The r-squared: Some straight talk," *Political Analysis*, vol. 2, pp. 153–171, 1990.

[229] F. W. Huffer and C. Park, "A test for elliptical symmetry," *Journal of Multivariate Analysis*, vol. 98, no. 2, pp. 256–281, 2007.

[230] D. Xu, J. Cui, R. Bansal, X. Hao, J. Liu, W. Chen, and B. S. Peterson, "The ellipsoidal area ratio: an alternative anisotropy index for diffusion tensor imaging," *Magnetic resonance imaging*, vol. 27, no. 3, pp. 311–323, 2009.

[231] L. Tang, H. Lv, F. Yang, and L. Yu, "Complexity testing techniques for time series data: A comprehensive literature review," *Chaos, Solitons & Fractals*, vol. 81, pp. 117–135, 2015.