

UNIVERSIDADE DO VALE DO RIO DOS SINOS - UNISINOS
UNIDADE ACADÊMICA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO
EM COMPUTAÇÃO APLICADA
NÍVEL MESTRADO

WILLIAM FERREIRA MORENO OLIVERIO

**A HYBRID MODEL FOR FRAUD DETECTION ON PURCHASE ORDERS BASED
ON UNSUPERVISED LEARNING**

São Leopoldo
2018

William Ferreira Moreno Oliverio

**A HYBRID MODEL FOR FRAUD DETECTION ON PURCHASE ORDERS BASED
ON UNSUPERVISED LEARNING**

Proposta de dissertação apresentada como
requisito parcial para a obtenção do título de
Mestre, pelo Programa de Pós-Graduação em
Computação Aplicada da Universidade do Vale
do Rio dos Sinos – UNISINOS

Orientador: Prof. Dr. Sandro José Rigo

São Leopoldo

2019

A FICHA CATALOGRÁFICA DEVE SER ELABORADA POR BIBLIOTECÁRIO COM
REGISTRO NO CRB.

Catologação na publicação: Bibliotecária responsável – CRB XX/XXXX

(Esta folha serve somente para guardar o lugar da verdadeira folha de aprovação, que é obtida após a defesa do trabalho. Este item é obrigatório, exceto no caso de TCCs.)

To my wife Debora, who provided me with unconditional love and support.
To my loved and caring daughter, Beatriz.
I love you both to the moon and back.

ACKNOWLEDGEMENTS

First, I would like to thank God who has given me enlightenment and the opportunity to meet so many special people who helped me achieve so many things in my life.

I would like to thanks for the unconditional support from my wife Debora Oliverio, who always stayed by my side and believed in me, even in the moments that even I was not able to. I love you so much!

Thanks to my daughter Beatriz, who game me energy and serenity on the most difficult moments, by just holding my hand and giving a kiss before going to bed, she was able to give me all the energy I needed during all the times I was studying at night.

Thanks to my parents, who taught me the values I carry until today.

I would like to especially thank my advisor, Sandro Rigo, which always supported me, aided in the moments when the path was not clear and gave me tranquillity and confidence that any work can be done with a proper plan in place.

I would like to thanks my friends from FJ Informatica, Silvio, Amauri, Francisco and Pericles. You changed my life, from opening the door of your own house to teaching me things that I didn't learn from my own family. Without you, I would not become the professional I am and would never be writing this dissertation. I will always have you in my heart.

I also would like to mention my colleagues Ismael, Juares, Lucas and so many others. Your company was really enjoying, and made all the time we worked together really enjoying.

I would like to thanks UNISINOS for the quality of teaching which was far above my expectations.

Finally, I would like to thanks my manager Rodolpho and Paul, who gave me several insights for topics related to this study as well as supporting me to get the time-off required to work on this study.

ABSTRACT

Fraud on the purchasing area is an issue which impacts companies all around the globe. This issue is treated with audits. However, due to the massive volume of the data available, it is impossible to verify all the transactions of a company. Therefore only a small sample of the data is verified. Due to the small number of frauds compared to the standard transactions, frequently, these fraudulent transactions are not included in the sample and hence are not verified during the audit. This work presents a new approach using the techniques of signature detection associated with clustering for an increased probability of inclusion of fraud-related documents in the sample. Due to the non-existence of a public database for fraud detection related to the purchase area of companies, this work uses real procurement data to compare the probability of selecting a fraudulent document into a data sample. Our work compares random sampling versus the sampling obtained from the proposed model. We also explore what would be the best clustering algorithm for this specific problem. The proposed model improves the current state-of-the-art since it does not require pre-classified datasets to work, is capable of operating with a very high number of data records, and does not need manual intervention.

Keywords: Fraud detection, procurement, non-supervised machine learning, clustering, signature detection.

RESUMO

A fraude na área de compras é uma questão que afeta empresas de todo o mundo. Esse problema é tratado com auditorias. No entanto, devido ao grande volume de dados disponíveis, é impossível verificar todas as transações de uma empresa. Portanto, apenas uma pequena amostra dos dados é verificada. Devido ao pequeno número de fraudes em comparação com as transações padrão, frequentemente essas transações fraudulentas não são incluídas na amostra e, portanto, não são verificadas durante a auditoria. Este trabalho apresenta uma nova abordagem utilizando as técnicas de detecção de assinatura associadas ao clustering para aumentar a probabilidade de inclusão de documentos relacionados à fraude na amostra. Devido à inexistência de um banco de dados público para detecção de fraudes relacionadas à área de compras das empresas, este trabalho utiliza dados de aquisições reais para comparar a probabilidade de selecionar um documento fraudulento em uma amostra de dados. Nosso trabalho compara amostragem aleatória versus a amostragem obtida a partir do modelo proposto. Também exploramos qual seria o melhor algoritmo de clustering para esse problema específico. A abordagem proposta melhora o atual estado da arte, uma vez que não requer conjuntos de dados pré-classificados, é capaz de operar com um número muito elevado de registros de dados e não precisa de intervenção manual.

Palavras-chave: Detecção de fraudes, agrupamento, detecção de assinaturas.

LIST OF FIGURES

Figure 1: K-Means algorithm logic	32
Figure 2: Proposed approach overview.....	41
Figure 3: Database relationship of the extracted tables from ERP	44
Figure 4: Historical AVG and STD calculated during signature matching	45
Figure 5: Historical AVG and STD calculated before signature matching	46
Figure 6: Example of a signature identification through SQL	48
Figure 7: Steps performed on clustering generation module	49
Figure 8: PCA result, variance per component	54
Figure 9: Variation of internal indexes based on the clustering parameter.....	57
Figure 10: Process flow of a purchase order	74
Figure 11: Examples of scenario matching	75
Figure 12: Proposal of fraud detection based on Fuzzy C-means	77

LIST OF TABLES

Table 1: Summary of clustering algorithms and related scores for the requirements of the approach	30
Table 2: Selected algorithms for implementation.....	31
Table 3: Summarized score for the selected papers.....	39
Table 4: Countries selected for the study	42
Table 5: List of tables used to retrieve the information from the ERP system	43
Table 6: List of metrics calculated on data preparation module	46
Table 7: List of signatures to be included in the model.....	47
Table 8: Parameters used by each clustering algorithm.....	50
Table 9: Example of the output of the generated signatures	54
Table 10: Average run times of clustering algorithms.....	56
Table 11: Best internal index value achieved per clustering algorithm.....	56
Table 12: Average values for signatures among the clusters with the most significant values highlighted.....	60
Table 13: Clusters and number of signatures with top scores.....	61
Table 14: Distribution of digits according to Benford's Law	72
Table 15: Cluster partition overlapping	77
Table 16: Comparison of different algorithms on purchase prediction	81
Table 17: Summary of the clusters generated	83
Table 18: Comparison of related works.....	83

LISTA OF ACRONYMS

ABNT	Associação Brasileira de Normas Técnicas
AVG	Average
BRA	Business Risk Audit
BW	Business Warehouse
CLARA	Clustering Large Applications
CURE	Clustering Using REpresentatives
ECC	ERP Central Component
ERP	Enterprise Resource Planning
GPU	Graphical Processing Unit
NGO	Non-Governmental Organization
OECD	Organization for Economic Co-operation and Development
PAM	Partitioning Around Medoids
PCA	Principal Component Analysis
PFC	Potential Fraud Candidate
PO	Purchase Order
PPGCA	Programa de Pós-Graduação em Computação Aplicada
PSO	Particle Swarm Optimization
SOM	Self-Organizing Map
SQL	Structured Query Language
STD	Standard Deviation
UNISINOS	Universidade do Vale do Rio dos Sinos

CONTENTS

1 INTRODUCTION	21
1.1 Motivation	21
1.2 Research question	22
1.3 Methodology	22
1.4 Structure	23
2 BACKGROUND	24
2.1 Fraud in the procurement area	24
2.1.1 Bid Rigging	24
2.1.2 Double Payment	25
2.1.3 Kickback fraud	25
2.1.4 Non-accomplice vendor	26
2.1.5 Personal Purchases	26
2.1.6 Redirect Payment fraud	26
2.1.7 Shell Company	26
2.1.8 Pass through	27
2.2 Signature matching	27
2.3 Clustering	28
2.3.1 Cluster validation	31
2.3.2 K-MEANS	32
2.3.3 CURE	33
2.3.4 DBSCAN	34
2.3.5 CLICK	34
2.3.6 CLIQUE	35
2.3.7 HDBSCAN	35
2.3.8 BIRCH	36
3 RELATED WORKS	38
4 PROPOSED APPROACH	40
4.1 Approach overview	40
4.2 Data extraction and cleansing	41
4.3 Data preparation and feature generation	45
4.3.1 Historical AVG and STD calculated during signature matching	45
4.3.2 Historical AVG and STD calculated before signature matching	46
4.4 Signature generation	46
4.5 Cluster generation and validation	49
4.6 Final selection	51
5 EXPERIMENTAL RESULTS	52
5.1 Technical aspects	52
5.2 Dataset	52
5.3 Metrics	53
5.4 Baseline	53
5.5 Results	53
5.5.1 Manual classification	58
5.5.2 Automated classification	60
5.5.3 Manual verification of POs	61
5.6 Limitations	64
6 CONCLUSION	66
6.1 Contributions	66
6.2 Future work	67
REFERENCES	68
APPENDIX A – RELATED WORKS	72
A business process mining application for internal transaction fraud mitigation	73

Fraud detection in ERP systems using Scenario matching	74
Reducing false positives in fraud detection: Combining the red flag approach with process mining	75
Fuzzy C-Means for Fraud Detection in Large transaction data sets	76
Screening for bid-rigging – does it work?	77
Prediction of Enterprise Purchases using Markov models in Procurement Analytics Applications	79
K-Means Algorithm: Fraud Detection Based on Signalling Data	81
Comparison of the related works	83
APPENDIX B – SIGNATURE DETAILS.....	89

1 INTRODUCTION

Enterprise Resource Planning (ERP) are systems that provide complete automation for most business processes. While the automation increases the efficiency of the company, it opens possibilities for internal fraud if the controls available on the system are not robust enough to prevent them.

One of the common ways to identify frauds performed in an organization is through an auditing process. Companies listed on stock markets have a requirement to be audited both internally and externally (LONDON STOCK EXCHANGE, 2012; NEW YORK STOCK EXCHANGE, 2014). During an audit, some of the main activities performed are collect data samples and implement an analysis of the data. The audit is usually done with the ERP and other interconnected systems used in a business process which could be impacted by fraud.

One of the business processes in which fraud is observed is the purchase of goods and services from other companies. These operations may include goods and services which ranges from the most expensive machinery to office supplies, which creates a much diversified and one high number of purchase orders. For some companies, this number can be in the range of several thousand a day.

Deal with the high volume of transactions and the diversity of its nature represents a massive challenge from an audit perspective since it is not feasible to go manually through all these purchases documents to identify fraudulent cases. In parallel, a random selection of documents to be used during the audit have a small chance of identifying the frauds since, by definition, fraudulent transactions should be an exception, with a small number of fraudulent documents.

To address this issue, different approaches were implemented, including process mining (BAADER, G.; KRCMAR, H., 2018), scenario matching (ISLAM et al., 2015) and artificial neural networks associated with logistic regression (LEE, Y.; HSIAO, Y.; PENG, C., 2015) to cite a few.

1.1 Motivation

The primary motivation for this work is the fact that fraud is already impacting, on average, 5% of company revenue, with total reported fraud going from 6.3 Billion USD in 2016 to 7 Billion US\$ on 2018. This number only covers cases of formally reported frauds, with the real number expected to be much higher. Besides, the Association of Certified Fraud Examiners, (ACFE, 2016; ACFE, 2018) a worldwide American organization that studies internal fraud cases indicates that frauds are being performed for 16 months in average before being detected.

A report by Ardent Partners highlights that the average procurement department of an organization manages 60.6% of total enterprise expenditure (WESTERSKI, A. et al., 2015). Another report from Price Waterhouse Coopers, a global audit company, states that procurement is a 28% increase on the number of UK companies experiencing procurement fraud as well as having only 1% of frauds being detected by data analytics (PWC, 2018).

Specifically, about the purchasing area, 77% of the frauds related to purchasing area are related to corruption of suppliers or third parties (ACFE, 2018), so it is an area in which is very difficult to identify the frauds due to the involvement of the external parties on the fraudulent schemes.

These reports indicate fraud is a problem increasing over the years, and in some countries, more specifically in the United Kingdom, it is growing specifically on procurement area (ACFE, 2018).

With the high impact of fraud as mentioned above, the massive amount of information generated in enterprise ERPs and the inability of auditors to verify on a manual approach all the information available, an automatic solution needs to be implemented to help the identification of the frauds in a faster way.

Because in purchasing area suppliers are often involved in the process, it is more difficult to identify a fraud only with the data available on the ERP system. Hence instead of trying to identify the frauds in a categorical way, the objective would be to identify a better data sample in order to be used by the professional auditors in the field.

Most of the research on the fraud detection area involves some classification approach, with no studies performed on clustering or time-series approaches (YUE et al., 2007; CARLSSON C., HEIKKILA M., WANG X., 2018; BAADER, G.; KRCMAR, H, 2018). By the best of our knowledge, no similar studies were performed using clustering associated with signature matching on the fraud detection scope for frauds on ERP systems. Also, among the related work papers studied, there is not a single solution which can be used to identify possible fraudulent documents without manual intervention on datasets with a very high number of records, besides having access to a previously classified dataset.

1.2 Research question

As briefly introduced in the previous section, there is a significant amount of work developed in the field of fraud detection on ERP systems, but for the best of the author knowledge, there are no experiments focusing on sampling selection for fraud detection.

This research aims to address the following research questions related to the field of data sampling for fraud detection in the purchasing area.

RQ1: Can the proposed approach increase the probability of identifying a possibly fraudulent document, when compared with the heuristic currently used in the audit area?

RQ2: Which clustering algorithm provides a better result on the domain of fraud detection in procurement?

1.3 Methodology

This work started with the requirement of having a solution to select which documents would be included in a sample in order to be used on an audit process since random sampling do not provide satisfactory results. Based on the requirement, an initial study was performed to understand the most common types of fraud in the procurement area. This study is presented in chapter 2.

The next step was to review related works on fraud detection specifically on the procurement area. Next, the selected papers were compared and the gaps in the state of the art were identified. This is presented in chapter 3.

To answer RQ1 and RQ2 we proposed an approach which combines clustering and signature matching, presented in chapter 4. Experiments were conducted and are evaluated in chapter 5.

1.4 Structure

This text is structured as follows. Chapter 2 provides the background for the work, covering the types of fraud in the procurement area and providing a review of the chosen clustering algorithms for this study. Chapter 3 provides a review of the related works on the fraud detection area, comparing the approaches and identifying the gap in state-of-the-art solutions. Chapter 4 presents a detailed description of the proposed approach. In chapter 5, the experiment results are presented. Chapter 6 presents a summary of the findings as well as the directions for future research on this field.

2 BACKGROUND

This chapter reviews the basics of the fraud types in the procurement area, non-supervised learning and clustering algorithms. It supports the work by providing more detailed information about how the frauds are performed on the procurement department as well as how clustering could be used in order to identify these transactions among the regular ones.

This chapter is organized into three sections. Primarily, section 2.1 introduces aspects related to the fraud in the procurement area, the most common groups of frauds and some symptoms associated with the frauds. Section 2.2 presents some important concepts about signature matching, the original definition and how it can be applied to the fraud detection field. Section 2.3 provides more detailed information on the clustering algorithms used in this study as well as how to compare the performance of clustering algorithms for a given dataset.

2.1 Fraud in the procurement area

To discuss fraud identification in procurement, it is imperative to have a clear definition of which activities can be considered fraud in the procurement area of a company. Besides, it is important to identify for each of these frauds the symptoms which can be used to classify a transaction as fraudulent or provide a strong indication of fraudulent activities involving a vendor or employee.

According to (ACFE, 2016), the definition of occupational fraud is: “the use of one's occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization's resources or assets”. During our study, the methodology used was to consolidate the information related to fraud detection in procurement and consolidate the information into a list containing the following aspects:

- Fraud Scenario;
- Symptoms to identify the fraud;
- Possibility of detecting the symptoms using the data from an ERP;
- Technical difficulty to identify this symptom on the ERP.

In (BAADER, G.; KRCMAR, H, 2018) the types of fraud related to procurement were described, as well as their related symptoms. However, the paper provides no indication on which of these symptoms can be identified purely using the data from an ERP environment, which is the main source of data for fraud detection on this work.

On the next topics, the details for all the documented types of frauds involving procurement will be detailed as well as how they can be identified.

2.1.1 Bid Rigging

The Organization for Economic Co-operation and Development (OECD) is an international organization with 36 countries as active members, 2500 employees, and which the objective is to perform analysis and provide recommendations and guidelines for economic and social improvements. The OECD published the guideline for fighting bid-rigging in public procurement (OECD, 2009). According to this guideline and also according to (IMHOF, D.; KARAGÖK, Y.; RUTZ, S., 2018), the definition of bid-rigging as “Bid rigging (or collusive tendering) occurs when businesses, that would otherwise be expected to compete, secretly

conspire to raise prices or lower the quality of goods or services for purchasers who wish to acquire products or services through a bidding process”.

Although the definition of bid-rigging is clear, there are several different ways the companies can operate to reach the same result, including the following:

- **Cover bidding:** Most frequent bid rigging is implemented in public procurement, in which companies submit a bid higher than the value of the designated winner or with unacceptable terms.
- **Bid suppression:** In which companies agree to simply do not participate in the bidding process to ensure that a designated winner wins the bid.
- **Bid rotation:** Companies continue to win with a designated winner for each of the bids thus looking like a real competition.
- **Market allocation:** Companies agree to not compete for certain specific customers or geographic areas, so each company does not interfere with the market of other companies.

The main indications of bid-rigging, according to (IMHOF, D.; KARAGÖK, Y.; RUTZ, S., 2018) and (HUBER, M.; IMHOF, D., 2018), are:

- The high variation between the price of the winning company and the remaining bids who lost the contract.
- High iteration between companies participating in a bid-rigging scheme.
- Low variation of prices related to costs (prices are more rigid).
- Low coefficient variation for the bids for a specific contract.
- A similar number of winning bids among the overall number of contracts for a specific customer or geographical area.

2.1.2 Double Payment

As detailed in (BAADER, G.; KRCMAR, H, 2018), the double payment fraud is an attempt to pay out an invoice several times. The payment is often made twice to an accomplice of the fraudster. Another possibility is that an invoice that was already paid in the past can be paid again.

After the double payment is performed, the fraudster contacts the supplier who received the double payment and asks for the amount to be returned, which is done via cheque and can be changed for money or to provide a credit for the company.

Since the company has a credit with that supplier, the fraudster can use the credit to clear any future invoice and make the payment of this invoice to himself using a shell company or a redirect payment type of fraud.

2.1.3 Kickback fraud

This type of fraud needs to have a cooperation of an internal perpetrator to perform the approval of the invoices. On a typical situation, a vendor submits a fictitious invoice or a real invoice with inflated prices and an internal employee, which participates in the fraud, approves the

invoice. In the end, the vendor shares part of the surplus with the internal employee (BAADER, G.; KRCMAR, H, 2018).

2.1.4 Non-accomplice vendor

As in (BAADER, G.; KRCMAR, H, 2018) the non-accomplice vendor is “a legitimate supplier not involved in the fraud case, used to defraud the company”.

This type of fraud requires that the perpetrator pays an invoice from a legitimate supplier with a value above the one stated on the invoice. The perpetrator then contacts the supplier and asks for a refund and intercept the money. Another variant of this fraud type is the acquisition of goods which should not be used by anyone in the company, the payment of the purchase and the subsequent return of the goods.

2.1.5 Personal Purchases

As in (BAADER, G.; KRCMAR, H, 2018), the personal purchase can be defined as the act of “making private purchases at the expense of the company”.

In summary, the perpetrators request the procurements of goods or services for their personal use but including those are company liabilities. It is common to have these requests concealed as business requirements or with invoices with fake descriptions to conceal the nature of the services and goods procured.

Usually, the perpetrators are the ones authorizing the purchases or use the company-owned credit card.

2.1.6 Redirect Payment fraud

Legitimate purchases by the company are changed in order that the payment is performed to the fraudster bank account (ISLAM, A. et al., 2010). The perpetrator needs to have access to change the master data (e.g. bank account) to make this fraud take place and usually cover the track by changing the bank account to the original information.

This type of fraud can be linked to the double payment, where one payment is performed to the fraudulent bank account and another one to the supplier to avoid suspects since the correct supplier will not complain about a payment not made.

2.1.7 Shell Company

As per detailed on (BAADER, G.; KRCMAR, H, 2018), a shell company is a “fictive entity without active business activities or significant assets”.

In order to receive the payments, usually, bank accounts are created on behalf of the shell company.

Shell companies can be used in legal activities, for example, holdings for groups owning several different companies but can be used for tax evasion activities and money laundering as well. Due to their frequent illegal usage, several countries are already taking measures on regulating this type of company, including United States (FINANCIAL CRIMES ENFORCEMENT NETWORK, 2016) which needs to know the real identity of a shell company owner before opening a bank account and United Kingdom which made mandatory that

overseas territories and crown dependencies to tell the true name of Shell Companies, however by 2020, all this information should be made available through a public register to avoid anonymous use of a shell company.

2.1.8 Pass through

Like a shell company, but in this case, an internal employee is responsible for setting up the shell supplier company which only receives the invoices and payments, but the goods and sent by a third-party company on behalf of the fake supplier. (BAADER, G.; KRCMAR, H, 2018).

2.2 Signature matching

Several papers were analysed in order to identify a single definition of signature matching.

In the paper of (ZAREMSKI, A., WING, J., 1993) the signature matching technique is used to identify similar programs codes based on specific requirements and it is described as: *“Signature matching is the process of determining when a library component ‘matches’ a query”*.

The definition of the concept is expressed in formulation 2.1 and is considered as the implication (Query Signature, Match Predicate, Component Library) \rightarrow Set of Components

$$\text{Signature Match}(q, M, C) = \{c \in C : M(c, q)\} \quad (2.1)$$

In other words, given a query “q”, a match predicate “M”, and a library of components “C”, signature matching returns a set of components which satisfies the matching predicate (ZAREMSKI, A., WING, J., 1993).

In the paper (JONASSON, J., OLOIFSSON, M., MONSTEIN, H. J., 2007) the signature matching is used for classifying bacteria into known pathogenic species or non-pathogenic species which can be found everywhere in nature. The signature matching concept is based on collecting the sequencing of the RNA of the analysed bacteria and compare against a database of known sequences of RNA, so the signature matching is defined as when bacteria RNA matches the pattern of previously identified and classified bacteria.

(SMITH, R. et al, 2009) applied signature matching on the identification of network packages based on Graphical Processing Unit (GPU) in order to increase the speed of the signature matching, with the main benefits of increasing the performance of intrusion detection systems, traffic shaping and quality of service. The signature matching on this paper works by identifying a series of patterns on the network packages being analysed against a database of known patterns.

Across the papers reviewed there is not a generic definition of signature. The papers provide a definition focused on the problem being analysed by the paper. However, in all the papers there is a three-step process involved, which includes:

- **Feature Extraction:** Analysis of the input data and generation of a series of features which can be used to classify the data currently being processed into a unique category.
- **Comparison against known data:** Compare the features generated in step 1 against a list of known features and related classes in order to check if there is a

match between the feature extracted in the previous step and the list of known features.

- **Classification:** In case of a match, the data currently being processed is classified as being relevant or not to the signature currently being processed.

2.3 Clustering

According to (XU, R., WUNSCH II, D., 2005) classification algorithms can be either supervised or unsupervised.

Supervised classification assigns new inputs to a finite number of discrete supervised classes using a mathematical function $y = y(x, w)$ on which x is the input data and w is a vector of adjustable parameters which are adjusted by a learning algorithm, also called inducer, which aims to reduce the loss. Once the learning algorithm reaches convergence or terminates, the classifier is created (XU, R., WUNSCH II, D., 2005).

Unsupervised learning, also called clustering or exploratory data analysis, can be described as a type of classification where no labelled data is available. “The goal of clustering is to separate a finite unlabelled data set into a finite and discrete set of natural, hidden data structures, rather than provide an accurate characterization of unobserved samples generated from the same probability distribution” (XU, R., WUNSCH II, D., 2005).

According to (POPAT, S.; EMMANUEL, M., 2014) “Clustering is an automatic learning technique which aims at grouping a set of objects into clusters so that objects in the same clusters should be as similar as possible, whereas objects in one cluster should be as dissimilar as possible from objects in other clusters.”.

Clustering algorithms can be classified into several categories. In the paper (POPAT, S.; EMMANUEL, M., 2014), clustering algorithms are classified into three main groups:

- Partitional clustering;
- Density-based clustering;
- Hierarchical clustering.

The paper (ELAVARASI, S., AKILANDEAWARI, J., SATHIYABHAMA, B., 2011) goes beyond and includes two more categories:

- Spectral clustering;
- Grid-based clustering.

Finally, (XU, D., TIAN, Y., 2015) extends the number of categories by adding further four more categories, reaching a total of nine categories:

- Fuzzy theory clustering;
- Distribution based clustering;
- Graph-based clustering;
- Model-based clustering.

For the specified subject of fraud detection, a survey performed by (SABAU, A. S., 2012) shows that clustering algorithms used in works aiming to detect fraud can be classified into 3 groups:

Partitional Clustering: present in 72% of the papers included in the survey and mainly using K-Means and DBSCAN.

Hierarchical Clustering: observed in 24% of the papers included in the survey, with different agglomerative and divisive clustering as the main algorithms being used.

Visualization Techniques: only present in 4% of the papers included in the survey, focused on displaying the data through different visuals.

In the paper (XU, D., TIAN, Y., 2015) 45 clustering algorithms were compared regarding their performance using seven metrics briefly described below:

Complexity(time): How the performance of the algorithm varies according to the complexity of the data set.

Scalability: Capacity of the algorithm to achieve acceptable results as the size of the data set increases.

Large-scale data: Capacity to handle data large in volume, rich in variety, high in velocity and doubt in veracity, also known as big data or non-structured data.

For high dimensional data: Capacity of the algorithm to process data with a large number of dimensions.

The shape of the data set: Capacity of the algorithm to handle non-convex data.

Sensitivity to the sequence of inputting data: Capacity of the clustering algorithm to take the sequence of the data records into account during the clustering process.

Sensitivity to noise/outliers: How the final classification result can be impacted by the presence of noise/outliers on the data set.

These metrics were compared to the problem of fraud detection on the procurement area and 3 metrics were considered as not relevant to our specific problem:

Large-scale data: Due to the fact that datasets are all structured and there are no big-data elements as part of the approach.

The shape of the data set: Since the dataset is structured as normalized and convex, it is applicable to all the algorithms on the study.

Sensitivity to the sequence of inputting data: Although the sequencing of some activities is a very important part during the fraud detection, on our approach, this is being handled on the module related to signature identification. Hence the clustering algorithm does not need to handle the sequencing of data.

Based on the remaining 4 metrics, clustering algorithms will be chosen based on the following minimal requirement:

Complexity(time): Due to the high data volume involved in this case study, the performance should be feasible to be executed within acceptable time limits. Hence algorithms with overall poor performance will be discarded.

Scalability: Only algorithms with high capacity will be chosen due to the very high variety of data and possible combinations of different signatures identified in the previous module.

For high dimensional data: Only algorithms capable of dealing with multiple dimensions will be selected since this is the shape of the data sets being used in the experiments.

Sensitivity to noise/outliers: At least a medium score on this metric will be required due to the presence of outliers/noise on the source data.

Based on the four metrics above, each algorithm received a score which varies between 0 and 2, for 0 being related to poor performance on a specific metric and 2 for a good performance on the metric. Algorithms with intermediary results received a score of 1. Table 1 shows the summarized version of the algorithms surveyed on the paper of (XU, D., TIAN, Y., 2015) and the results according to the requirements of our approach.

Table 1: Summary of clustering algorithms and related scores for the requirements of the approach

Category	Algorithm	Complexity (time)	Scalability	For high dimensional data	Sensitivity to noise/outliers	Score	Metrics with acceptable performance
Based on Partition	K-Means	Low	Middle	No	Highly	5	3
Based on Partition	K-Medoids	High	Low	No	Little	0	0
Based on Partition	PAM	High	Low	No	Little	0	0
Based on Partition	CLARA	Middle	High	No	Little	3	2
Based on Partition	CLARANS	High	Middle	No	Little	1	1
Based on Hierarchy	BIRCH	Low	High	No	Little	4	2
Based on Hierarchy	CURE	Low	High	Yes	Little	6	3
Based on Hierarchy	ROCK	High	Middle	Yes	Little	3	2
Based on Hierarchy	Chameleon	High	High	No	Little	2	1
Based on Fuzzy Theory	FCM	Low	Middle	No	Highly	5	3
Based on Fuzzy Theory	FCS	High	Low	No	Highly	2	1
Based on Fuzzy Theory	MM	Middle	Low	No	Little	1	1
Based on distribution	DBCLASD	Middle	Middle	Yes	Little	4	3
Based on distribution	GMM	High	High	No	Little	2	1
Based on density	DBSCAN	Middle	Middle	No	Little	2	2
Based on density	OPTICS	Middle	Middle	No	Little	2	2
Based on density	Mean-shift	High	Low	No	Little	0	0
Based on graph theory	CLICK	Low	High	No	Highly	6	3
Based on graph theory	MST	Middle	High	No	Highly	5	3
Based on Grid	STING	Low	High	Yes	Little	6	3
Based on Grid	CLIQUE	Low	High	Yes	Moderately	7	4
Based on Fractal Theory	FC	Low	High	Yes	Little	6	3
Based on Model	COBWEB	Low	Middle	No	Moderately	4	3
Based on Model	SOM	High	Low	Yes	Little	2	1
Based on Model	ART	Middle	High	No	Highly	5	3
Based on Kernel	kernel K-means	High	Middle	No	Little	1	1
Based on Kernel	kernel SOM	High	High	No	Little	2	1
Based on Kernel	kernel FCM	High	Middle	No	Little	1	1
Based on Kernel	SVC	High	Low	No	Little	0	0
Based on Kernel	MMC	High	Low	No	Little	0	0
Based on Kernel	MKC	High	Low	No	Little	0	0
Based on Swarm Intelligence	ACO based (LF)	High	Low	No	Highly	2	1
Based on Swarm Intelligence	PSO based	High	Low	No	Moderately	1	1
Based on Swarm Intelligence	SFLA based	High	Low	No	Moderately	1	1
Based on Swarm Intelligence	ABC based	High	Low	No	Moderately	1	1
Based on Quantum Theory	QC	High	Middle	No	Little	1	1
Based on Quantum Theory	DQC	Middle	Middle	No	Little	2	2
Spectral Clustering	SM	High	Middle	Yes	Little	3	2
Spectral Clustering	NJW	High	Middle	Yes	Little	3	2
Based on affinity propagation	AP	High	Low	No	Little	0	0
Based on density and distance	DD	High	Low	No	Little	0	0

Source: Created by the author

According to table 1, there is not a single algorithm which has a high score on all the four metrics considered for the comparison, with the only algorithm which scores high results on 3 of the 4 metrics being the grid-based classifier CLIQUE with a total score of 7.

For this study, all the algorithms which received a score of 6 or more will be included with the addition of K-Means and DBSCAN as a baseline since according to (SABAU, A. S., 2012) both were used on a significant number of papers related to fraud detection.

Finally, there were two grid-based algorithms with a score of at least 6, hence the STING algorithm will be removed from the list of algorithms tested in favour of CLIQUE due to high overall score which CLIQUE achieved.

In addition to these algorithms, HDBSCAN was included on the list of selected clustering algorithms since it is a novelty in the area and the experiments presented in (MCINNES, L., HEALY, J., ASTELS, S., 2017) shows promising results when compared with other clustering algorithms.

Table 2 contains the list of the selected algorithms which will be used during the implementation of the proposed approach, covering six different algorithm categories.

Table 2: Selected algorithms for implementation

Category	Algorithm	Complexity (time)	Scalability	For high dimensional data	Sensitivity to noise/outliers	Score	Metrics with acceptable performance
Based on Partition	K-Means	Low	Middle	No	Highly	5	3
Based on Hierarchy	BIRCH	Low	High	No	Little	4	2
Based on Hierarchy	CURE	Low	High	Yes	Little	6	3
Based on Hierarchy	ROCK	High	Middle	Yes	Little	3	2
Based on Hierarchy	HDBSCAN	N/A	N/A	N/A	N/A	N/A	N/A
Based on density	DBSCAN	Middle	Middle	No	Little	2	2
Based on density	OPTICS	Middle	Middle	No	Little	2	2
Based on Grid	CLIQUE	Low	High	Yes	Moderately	7	4

Source: Created by the author

Details of the selected clustering algorithms can be found on the next sections of this document as well as the details of how the clusters will be evaluated.

2.3.1 Cluster validation

According to (RENDON, E. et al, 2011) there are two approaches to calculate the cluster validity:

External criteria: Where the cluster results are validated against a pre-specified structure, i.e. external information that is not included in the data set.

Internal criteria: The cluster results are compared without relying on any information from outside the data set (prior knowledge).

Since the object of this study does not have any information to validate the index apart from the information already available on the dataset, internal criteria should be used to identify which clustering algorithm and which parameters for each clustering algorithm generate the best clusters.

Papers reviewed use different indexes to validate the quality of generated clusters and today there is no state-of-the-art index which can be used. The findings supported by (XU, R., WUNSCH II, D., 2005) state that: *“Validation criteria provide some insights on the quality of clustering solutions. But even how to choose the appropriate criterion is still a problem requiring more efforts.”*

In the paper (MAULIK, U., BANDYOPADHYAY, S., 2002) the performance of 4 indexes were compared using 5 different datasets. It was presented that \mathcal{I} index had significantly better results by identifying the correct number of clusters on all datasets, while Calinski Harabasz (CH) index identified the correct number of clusters on 3 datasets and Davies-Bouldin (DB) index which correctly identified the number of clusters on 2 datasets.

On the paper (RENDON, E. et al, 2011) the performance of 6 different indexes was compared using 13 datasets. The outcome is that index NIVA had a slightly better performance by identifying the correct number of clusters on 12 datasets versus 11 datasets identified against the Davies-Bouldin (DB) index and Silhouette index. CH index identified the correct number of clusters on 10 datasets was included in the study as one of the baselines. Finally, on (MILLIGAN, G. W., COOPER, M. C., 1985) 30 indices were compared across and ranked across 108 datasets, with CH index achieving the best performance overall.

Based on these three papers it can be concluded that depending on the datasets used, the index which provides the best results may change, however, among the papers reviewed, CH index, Silhouette index and DB index were among the best scores among several others, hence these three indexes will be used in the implementation of the model to validate the quality of the generated clusters.

2.3.2 K-MEANS

First mentioned in (MCQUEEN, J., 1967), this algorithm is described as “a process for partitioning an N-dimensional population into k sets based on a sample. The process, which is called 'k-means', appears to give partitions which are reasonably efficient in the sense of within-class variance.”

It is an algorithm calculated with Euclidean distances, which takes a K number of clusters received as input parameter and partition a dataset which contains N objects into K clusters (CHAUHAN, P., SHUKLA, M., 2014).

The steps for the K-Means algorithm are described in figure 1.

Figure 1: K-Means algorithm logic

```

Step 1: Randomly select object  $k$  from dataset  $D$  as initial centre of cluster;
Step 2: for each data point in dataset do:
    Step 3: Calculate distance of data point from each cluster centre
    Step 4: Based on that distance find closest cluster and put that data point in that cluster
Step 5: end for
Step 6: After assigning the objects re-calculate the mean of each clusters and update its value
Step 7: Go to step 2 until stopping criterion is match

```

Source: CHAUHAN, P., SHUKLA, M. (2014)

This algorithm is computationally inexpensive (MCQUEEN, J., 1967) hence can be used on big datasets as well as being easy to implement, however it does have disadvantages as well, mainly the need of specifying the number of clusters before running the algorithm and the inability of handling outliers (POPAT, S.; EMMANUEL, M., 2014).

In the paper (MIN, X., LIN, R., 2018) the K-Means is used to categorize the data from phone calls to identify fraudulent phones. A differential of this paper was the usage of Principal Component Analysis (PCA) to reduce the number of entries to be used as the input for the K-Means algorithm and the statistical analysis of the data with Hopkins statistics as described in (BANERJEE, A., DAVE, R., 2004).

In the paper (CHAUHAN, P., SHUKLA, M., 2014), K-Means algorithm is used as an outlier detection tool for data streams, rather than a clustering tool which was the original purpose of the algorithm. On the paper, the authors provide a total of 7 different approaches on how to identify outliers based on hybrid models using K-means coupled with SVMs, DBSCAN as well as using modified versions of the K-Means.

2.3.3 CURE

According to (GUHA, S.; RASTOGI, R.; SHIM, K., 2001), the CURE algorithm stands for Clustering Using REpresentatives. It is a hierarchy-based algorithm which was designed to overcome the limitations of clustering algorithms when handling datasets with clusters of significantly different sizes and non-spherical shapes.

Partitional algorithms usually work to partition the data in a way to optimize the criterion function, with the square-error criterion being the most common at the time of the paper creation. The square-error works in a way that clusters as compact as possible, however, in datasets with large clusters, the square-error criterion could split a large cluster in order to minimize the error value. (GUHA, S.; RASTOGI, R.; SHIM, K., 2001) defines partitional clustering algorithms as “*a sequence of partitions in which each partition is nested into the next partition in the sequence*”. On agglomerative clustering each data point is considered as an independent cluster, the clusters which are closest to each other are merged until the desired number of K clusters is reached.

The measures used to calculate the distance between these clusters to select which ones will be merged could be either the mean distance, on which the centre of the cluster is used to calculate which clusters are the closest ones, or the minimum distance which consider the clusters which has any of its members closest to a member of another cluster.

Both methods work well when clusters are compact and separated but their results can vary significantly for clusters positioned very close to each other or with non-uniform forms.

CURE uses a novel algorithm which is based on the following phases:

- 1 - Initially the algorithm scans a minimum sample size of the dataset and then the partitional cluster is used.

- 2 – Data for each partitioned is clustered and outliers eliminated

- 3 – After the outlier elimination, the data is cluster together to generate the final clusters.

Finally, CURE has a worst-case time complexity of $O(n^2 + n m \log n)$ with n the number of input data points and m being the number of clusters on the input data.

2.3.4 DBSCAN

DBSCAN was originally described in the paper (ESTER, M., et al., 2010) as a density-based clustering algorithm designed to identify clusters of arbitrary shape.

Density-based clustering is based on the principle that the density of points in an area which delimitates a cluster is significantly higher than the area in which there is no cluster available. Hence, based on the density, a data point can be classified as a member of a cluster or as noise.

Before starting to review the logic of the algorithm, it is important to clarify two important aspects. The parameters required for the algorithm and the concept of each data point can be classified. The algorithm needs to receive two input parameters:

Eps: Maximum distance so the points can be considered as a neighbour in order to calculate the density.

MinPts: Minimum number of points within Eps distance so the point can be considered a core point.

The main concept for DBSCAN is how the points of a given dataset are classified, which each point can be classified as either:

Core Point: Any point which has more than the number of points established on *MinPts* within distance *Eps*. These points are inside the cluster.

Border Point: Any point which does not have the minimum number of points on *MinPts*, but is in the neighbourhood of a core point.

Noise Point: Any point which is not a core point or a border point.

In the case of a point being classified as a core point, the algorithm will look for any other core point within distance *Eps*. In case such point exists, the point currently being processed would be assigned to the same cluster of the closest core point, otherwise a new cluster would be generated.

In the case of a point being classified as border point, it will be classified as a member of the cluster of the core point within range *Eps*. Finally, points classified as noise will not be a member of any cluster and will be categorized as noise/outliers.

2.3.5 CLICK

CLICK stands for CLuster Identification via Connectivity Kernels. It was first introduced in the paper of (SHARAN, R., SHAMIR, R., 2000). It is a clustering algorithm based on the graph theory which was initially introduced aiming to solve problems related to cluster genes into clusters in order to deduct the gene functionalities.

It is based on the minimum cut weight to form clusters. The graph is weighted, and the edge weights are assigned a new interpretation, by combining probability and graph theory (XU, R., WUNSCH II, D., 2005). The weighted edge is between the nodes i and j are calculated as shown below where S_{ij} is the similarities between the two nodes.

$$e_{ij} = \log \frac{\text{Pr ob}(i, j \text{ belong to the same cluster} | S_{ij})}{\text{Pr ob}(i, j \text{ does not belong to the same cluster} | S_{ij})} \quad (2.2)$$

CLICK then checks for similarities within the cluster following Gaussian distribution with different means and then checks similarities between clusters again with Gaussian distribution but with different variances.

This result of a new equation based on the previous one but rewritten using Bayes' theorem where p_0 is the prior probability of two objects to belong to the same cluster and $\mu_B, \sigma_B^2, \mu_W, \sigma_W^2$ are the means and variances for the similarities between clusters and intra-cluster.

$$e_{ij} = \log \frac{p_0 \sigma_B}{(1 - p_0) \sigma_W} + \frac{(S_{ij} - \mu_B)^2}{2\sigma_B^2} - \frac{(S_{ij} - \mu_W)^2}{2\sigma_W^2} \quad (2.3)$$

CLICK then works on a recursive way on the current subgraph and creates a kernel list, containing the components which satisfy a criterion function.

If a subgraph contains a single node, they are classified as singletons and are further processed by carrying out a series of singleton adoptions and cluster merges in order to generate the resulting clusters (XU, R., WUNSCH II, D., 2005).

2.3.6 CLIQUE

CLIQUE algorithm was initially presented on the paper (AGRAWAL, R., et al., 2005) and stands for CLustering In QUest, a data mining research project at IBM Almaden.

It is a grid-based algorithm, which was created with the main objective of clustering datasets with a high number of dimensions, the capability to identify clusters which do not need to be on circular or in a spherical space and to have a scalable performance as the number of points and dimensions increases.

CLIQUE works on a three-step process. First, it uses a bottom-up approach to identify rectangular cells across all the dimensions based on how dense the data is. This process could be very processing-intensive, and its performance is increased by the application of the Minimal Description Length (MDL) principle. This technique groups together the dense units which are lying on the same subspace and then checks for each of the remaining sub-spaces, what is the fraction of the database which is covered by the group being processed. The subspaces with the highest coverage are selected and the remaining ones are pruned. On the next step, CLIQUE works to identify the clusters by working on the connected components in a graph, on which the vertices stand for dense units. Finally, CLIQUE generates the minimum descriptions for the generated clusters by merging the rectangles previously identified (XU, R., WUNSCH II, D., 2005).

2.3.7 HDBSCAN

HDBSCAN algorithm which first introduced in the paper of (CAMPELLO, R., MOULAVI, D., SANDER, J., 2013). It extends the DBSCAN algorithm by converting the original algorithm into a hierarchical clustering one. From the generated hierarchy, simplified clusters are generated, which outperforms the state of the art of density-based cluster algorithms in several datasets.

According to (MCINNES, L., HEALY, J., ASTELS, S., 2017) the main steps performed by HDBSCAN are:

1. Transform the space according to the density and sparsity of the data.

On this step, the data is transformed into a graph using mutual reachability distance in order to make the data set more robust to outliers and noise in the data.

2. Build the minimum spanning tree.

The aim of this step is to find areas of dense data within the dataset being processed. This is performed by Prim's minimum spanning tree algorithm, which eliminates the edges of the graph created in the previous step until a threshold provided to the algorithm is achieved.

3. Construct a cluster hierarchy.

On this step, the minimum spanning tree created on the previous step is converted into a hierarchy of connected components by sorting the edges of the spanning tree and iterating through the tree, creating a new merged cluster for each stage.

4. Condense the cluster tree

On this step, the previous hierarchy is simplified by generating clusters which have a minimum number of members which is provided as a parameter to the HDBSCAN algorithm.

Clusters which does not have the minimum number of elements are classified as points falling out of a cluster.

5. Extract the stable clusters

On the last step, the final clusters are selected based on the assumption that when selecting a cluster, the algorithm can not select any other cluster which is descendant of it.

On top of this assumption, the clusters are selected based on persistence which is calculated based on a lambda value, calculated as $1/\text{distance}$ so the clusters selected are the ones that have a higher persistence.

2.3.8 BIRCH

BIRCH stands for Balanced Iterative Reducing and Clustering using Hierarchies. It was first introduced in the paper of (ZHANG, T., RAMAKRISHNAN, R., LIVNY, M., 1996) and was implemented based on two motivations: being able to deal with large data sets and to be robust enough to outliers.

To achieve these goals, a new data structure, clustering feature tree, is used. This structure is used to store the summaries of the original data and contains the number of data objects in the cluster, the linear sum of the objects and the squared sum of the objects.

This algorithm eliminates outliers by identifying elements sparsely distributed in the feature space. Finally, after the CF tree is created an agglomerative hierarchical cluster is used to perform the global clustering (XU, R.; WUNSCH II, D., 2005).

3 RELATED WORKS

Based on the literature review performed on background topics, we performed a survey for fraud detection approaches with a focus on the procurement area. In order to do this survey, we searched for articles using the search pattern “Fraud Detection” and “Procurement”. In addition, articles were searched using the terms “Unsupervised learning” and “Clustering”.

We included in the survey the papers proposing or implementing models related to fraud detection on procurement area without making the usage of labelled data or to be models related to another subject, but which could be modified in order to fit for the purpose of fraud detection in the procurement area.

Each paper was classified based on six different topics:

Scalable: How the solution is capable of handling large data sets (above 1 million data records).

Real-time capable: Capability of the solution to process the information just created and reach a decision in a reduced amount of time.

Domain knowledge requirement: Level of domain knowledge of the user operating the model in order to understand the information provided and make a decision.

Adaptable: Capacity of the solution to handle data for which it was not previously prepared or configured, for example, on identifying new types of fraud.

Automated: Ability of the model to reach the final decision without human interaction.

Metrics available: Which metrics were used in order to reach a result comparison.

In addition, the gaps identified for each work were documented. The details of the survey are available on appendix A. Table 3 shows the results of the survey for the selected papers. Scores were classified in colours for easier comprehension, on which a green score indicates the paper achieves the requirement of the metric, the scores in yellow partially achieve the objective and in red does not achieve the minimum requirement.

Table 3: Summarized score for the selected papers

Paper	Scalable	Real time capable	Domain knowledge required	Adaptable	Automated	Metrics available
The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data	Yes	Partial	Yes	Yes	No	No
A business process mining application for internal transaction fraud mitigation	No	No	Yes	Yes	No	Yes
Fraud detection in ERP systems using Scenario matching	Partial	Yes	No	No	Yes	Yes
Reducing false positives in fraud detection: Combining the red flag approach with process mining	Partial	Partial	Partial	No	No	Yes
Fuzzy C-Means for Fraud Detection in Large transaction data sets	Partial	Partial	Partial	Yes	Partial	No
Screening for bid rigging – does it work?	No	No	Yes	Yes	No	Yes
Prediction of Enterprise Purchases using Markov models in Procurement Analytics Applications	Yes	Partial	No	Yes	Yes	Yes
K-Means: Fraud Detection Based on Signaling Data	Yes	No	Yes	Yes	No	Yes

Source: Created by the author

The main gaps identified considered are the following. First, the requirement of having manual intervention during the analysis of the data, drastically limiting the scalability of the solution for high data volumes and real-time analysis. Second, the requirement of domain knowledge to understand the results of the model in order to identify fraudulent documents. Finally, the lack of a standardized data set in order to compare the performance of the selected models.

The only model which matches all the evaluation items is the paper of (WESTERSKI, A. et al., 2015), which focus on predicting the next purchase of a specific user instead of identifying the document as fraudulent or not. Nevertheless, the approach can be modified in order to be used on the fraud detection domain.

4 PROPOSED APPROACH

In this chapter, we describe the approach proposed for this work. After an overview with the main aspects of the proposal, the components are detailed and commented.

4.1 Approach overview

Based on the study performed on the related works, we propose a hybrid approach for fraud detection on purchase orders inspired on some key components observed. The approach incorporates the concepts observed in studied approaches, including scenario matching, Bedford analysis, process mining to be used when the infrastructure and applications on the company are operational, and non-supervised learning.

As described in (JANS et al., 2011) there is currently a lack of datasets related to internal transactional fraud. Therefore this work will be focused on the non-supervised learning algorithms to generate a list of purchase orders with the highest chances of fraud. The scenario matching part of this approach was inspired on the work of (ISLAM et al., 2010) with additional scenarios of known fraud types added according to the work of (JANS et al., 2011).

The concept of using non-supervised learning was based on several articles described in (SABAU, A., 2012), showing that several authors are investigating this approach of fraud detection. The differential of the approach proposed is that instead of trying to identify each purchase document as fraudulent or normal, we aim to provide a sample of the purchase orders with the following two conditions:

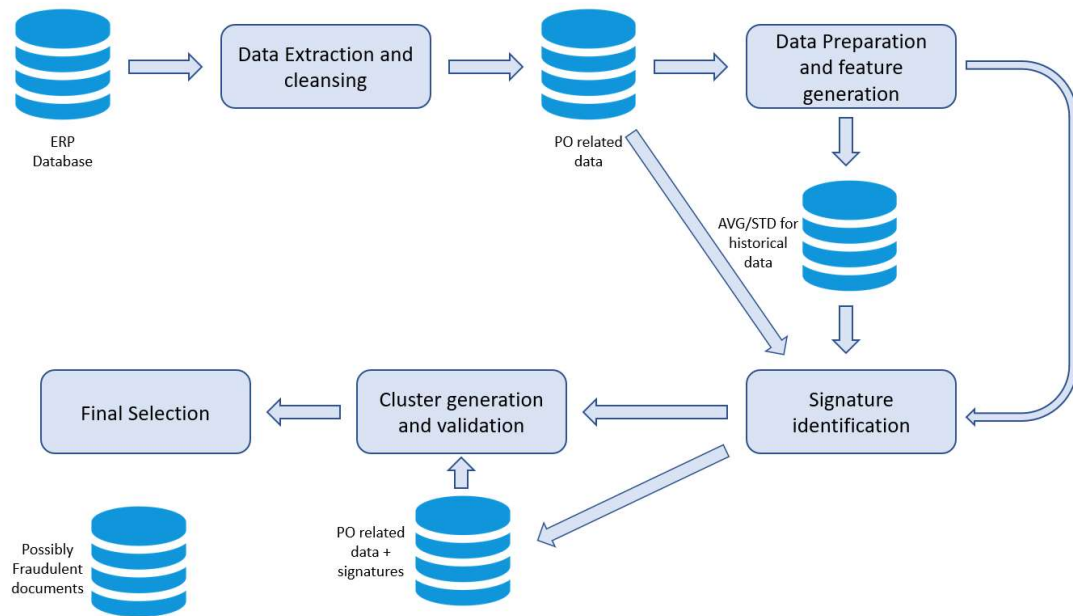
- The highest likelihood of fraud based on the signature detection techniques;
- The highest possible financial impact on the corporation.

Finally, due to the lack of standardized data sets for fraud detection, this work will create a methodology of how to validate which documents can be considered fraudulent or with an indication of being fraudulent, so the solution can be validated and used in future studies.

Figure 2 illustrates the main steps defined in the proposed approach. Also are described the main components and therefore, their interaction. The approach starts extracting the data from the ERP environment to another database used specifically for the model. After that are generated the AVG and STD for the POs which will be stored on an additional table. Next, the signature identification module will generate the signatures based on both the original PO related data as well as the AVG and STD values. Finally, the clusters will be generated, and the clusters with the highest overall score will be selected.

The main advantage of the proposed approach is the combination of the high accuracy of the signature detection for known types of frauds with the possibility of the identifying new types of frauds based on the combination of the symptoms which alone does not indicate fraud, but when combined could provide useful insights on identifying fraudulent documents.

Figure 2: Proposed approach overview



Source: Created by the author

The main components illustrated in figure 2 are commented below:

- **Data Extraction and cleansing:** Module responsible for extracting the data from several tables from ECC and BW solution. This will include master data related to vendors, transactional data related to purchase orders and system logs.
- **Data preparation and feature generation:** Prepare the data to be processed by changing values to the same currency for countries with multiple currencies, date formatting change as well as summarized information for PO requestor, approvers and suppliers created, and applying Benford analysis to the summarized data.
- **Signature identification:** Identify specific signatures on each document, example for retroactive PO generation, sudden spikes on purchases for a specific supplier, etc.
- **Cluster generation:** Cluster the data generated on previous steps and identify outliers on the generated data set.
- **Final selection:** Create the final sample of PO documents to be included in the scope, as well as the list of suspect suppliers, PO creators and PO approvers.

In the next items, each component is described in detail.

4.2 Data extraction and cleansing

The first module will be responsible for extracting the data from the ERP system used by the company who is the object of this case study.

This phase has two main challenges: data volume and data availability. The data volume for a big multinational company can be massive. On the company used as an example of this implementation, the tables which hold the changes performed on the most important documents

currently have more than 3 billion records. This makes the extraction of the whole data unfeasible due to time and machine capacity constraints.

In order to solve this problem two main activities were performed: a) Filter the logs only for the changes related to purchase requests, purchase orders, goods receipt, invoice receipt, payments, material movements and their related master data. b) Reduce the number of countries in scope for this experiment.

In our approach, we decided to select the countries based on the corruption perception index 2017. This index is a survey held by the Transparency International Organization, a non-governmental organization founded in 1993, currently present in more than 100 countries and which main objective is to fight corruption. The corruption perception index is a survey of all the countries and their related corruption fight score which is calculated using several metrics. Based on this score, a global corruption rank is created. By the time of the writing on this document, the most updated version of this study is the global corruption perception index 2017 (TRANSPARENCY INTERNATIONAL, 2017). Based on the ranking the approach decided was to choose a total of 9 countries which can be found in table 4. The countries were chosen using the methodology below:

- 3 among the less corrupted countries;
- 3 with average scores;
- 3 among the most corrupted countries.

By selecting countries with a low, medium and high level of corruption, the data selected should have a similar likelihood of fraud from the whole population. The countries selected for the study can be found in table 4.

Table 4: Countries selected for the study

Position	Country	Corruption Prevention Index	Classification
3	Switzerland	85	Very Clean
11	Germany	81	Very Clean
13	Australia	77	Very Clean
60	Croatia	49	Average Results
61	Romania	48	Average Results
61	Malaysia	47	Average Results
149	Bangladesh	28	Highly Corrupt
149	Kenya	28	Highly Corrupt
168	Venezuela	18	Highly Corrupt

Source: Created by the author

Brazil was not included in this study since it has a score of 37, which is slightly below the score of an average country and the fact that the vast majority of the purchase orders in Brazil are related to the procurement of raw materials to be used in other factories across the globe. This lead to a unique procurement behaviour which does not exist in any other country which the company operates, hence a decision was taken to select countries which have similar ways of working.

Regarding data availability, we have some important issues. Based on recommendations from the ERP manufacturer, data which is considered old needs to be archived, which means that the data is deleted from the ERP system and stored on a secondary database system in case it needs to be accessed for any reasons during the retention period, which varies according to each

country legislation. This makes the process more complex since the performance of the system which holds the archived data is not as good as the productive ERP, leading to several errors during massive data extraction.

In order to avoid these issues, the approach chosen was to limit the historical period of data being analysed to 2 years since after this period the data is archived and deleted from the database. To extract the data from the ERP database, a data extraction tool was used. This tool provided read-only access to the ERP database.

A total of 25 tables were read from the ERP database and copied into another database used specifically for this experiment. These tables were chosen since they hold the information related to purchase orders which are usually verified manually during an audit.

The list of these tables can be found in table 5 but in summary, it will include data related to:

- Customers, vendors, materials, plants master data;
- Purchase requisitions and purchase orders;
- Invoices;
- Material movements;
- Payments;
- Accounting documents;
- Logs of database changes.

Table 5: List of tables used to retrieve the information from the ERP system

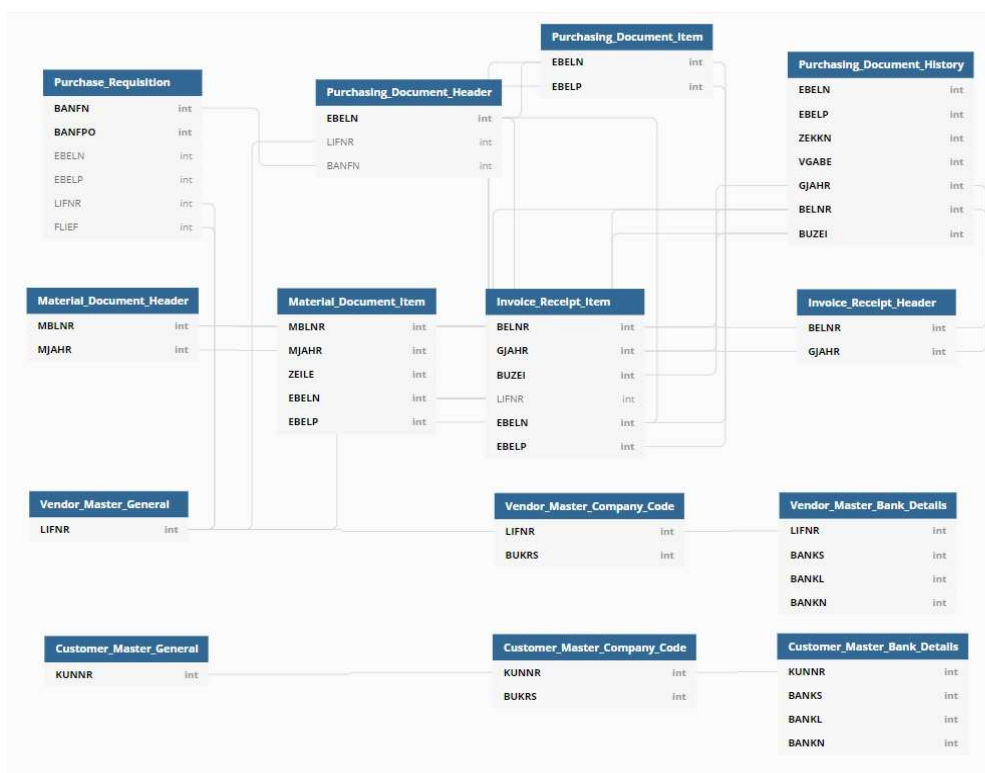
Table	Table Description
BSAK	Accounting: Secondary Index for Vendors (Cleared Items)
BSIK	Accounting: Secondary Index for Vendors
CDHDR	Change document header
CDPOS	Change document items
EBAN	Purchase Requisition
EKBE	History per Purchasing Document
EKKO	Purchasing Document Header
EKPO	Purchasing Document Item
KNA1	General Data in Customer Master
KNB1	Customer Master (Company Code)
KNBK	Customer Master (Company Code)
LFA1	Vendor Master (General Section)
LFB1	Vendor Master (Company Code)
LFBK	Vendor Master (Bank Details)
MARA	General Material Data
MARC	Plant Data for Material
MARD	Storage Location Data for Material
MKPF	Header: Material Document
MSEG	Document Segment: Material
PAYR	Payment Medium File
RBKP	Document Header: Invoice Receipt

RSEG	Document Item: Incoming Invoice
T001	Company Codes
T001W	Plants/Branches
TVKO	Organizational Unit: Sales Organizations

Source: Created by the author

In addition to the tables listed above, the data connections can be found in figure 3. Only the primary keys and foreign keys were included in the data model in order to improve the visualization.

Figure 3: Database relationship of the extracted tables from ERP



Source: Created by the author

The last step on this module is to perform the data cleaning. This step performs two main activities:

Remove unused columns: this is required since the tools used to extract the data from the ECC database copy all the columns of the tables. By removing columns which are not necessary for the model, the database size and processing time can be reduced.

Reduce the size of changelogs: this activity aims to reduce the number of records for tables CDHDR and CDPOS which holds the logs for data change on SAP ECC. The design of the SAP system creates an entry on these tables in case of any changes on the most important information on the system, for example, when a bank account is changed for a supplier, a sales order is deleted or the product price is updated, an entry is created on these tables.

The module will hence keep only the logs for the transactions and tables which are relevant to this study, greatly reducing the size of the data being analysed to less than 10% of the data available on the ERP system.

4.3 Data preparation and feature generation

During the signature identification module, several signatures will be processed by comparing the value of a specific purchase order with the data of previous purchase orders related to the same supplier, material, requestor and approver involved to identify significant differences between the data of the PO currently being processed versus the historical data of the same supplier, material, requestor and approver.

This information is required to classify if the signatures identified on the next module are related to the whole universe of data or to a specific case, for example, in case of a sudden increase in the cost of a specific material, you could have two different situations:

Market price change: On this example, there is no fraud since the price of the good purchased should, on average, be higher than previous purchases, no matter which supplier, requestor or approver were selected.

Fraud between approver and supplier: On this case, the price should only be increased when the combination of the fraudulent supplier and approver were selected, with the remaining suppliers/approvers having the cost significantly smaller.

The calculation of the AVG and STD for the historical PO data can be either performed before the signature identification module (as proposed on this model) or during the calculation of the signature. Below will be described the differences between each approach based purely on the performance of each approach since the result of the calculations will be for both approaches. Considering N as the number of POs included on the original data set, we have the pseudo-code used to calculate the data and the Big O notation for the related performance.

4.3.1 Historical AVG and STD calculated during signature matching

Since all N POs will be processed during the signature matching, and for each PO, all the historical data will have to be read up to the current PO. The number of documents read can be between 1 (for the first PO processed) and N (for the last PO processed), hence it will be described as $N/2$, with a total Big O notation of $O(N * (N/2))$.

The algorithm for this case can be found in figure 4.

Figure 4: Historical AVG and STD calculated during signature matching

```

For N
  Read PO data from database up to current PO
    Calculate AVG and STD
  Write to database

```

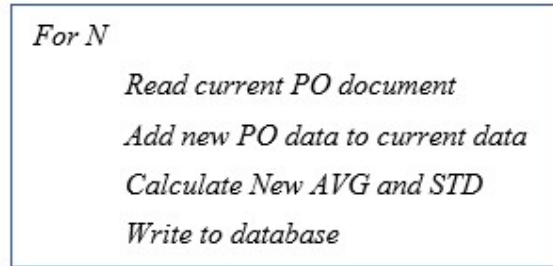
Source: Created by the author

4.3.2 Historical AVG and STD calculated before signature matching

Since the list of the PO documents will need to be read-only because the historical AVG and STD are calculated before the signature matching phase, this approach has a Big O performance of $O(N)$ which is significantly faster than $O(N*(N/2))$ achieved when performing this calculation in parallel with the signature matching.

The algorithm for this case can be found in figure 5.


Figure 5: Historical AVG and STD calculated before signature matching



Source: Created by the author

Due to the data volume used on this solution being significantly large, it was chosen to select the option which has the best performance for the implementation in the model. The details of the metrics calculated are included in table 6.

Table 6: List of metrics calculated on data preparation module



Historical data	Supplier	Material
Average unit price	X	
Average unit price		X
Average unit price	X	X
Average quantity purchased	X	
Average quantity purchased		X
Average quantity purchased	X	X

Source: Created by the author

4.4 Signature generation

In the paper (PORTNOY, L., 2000) it is presented the concept that transactions which have the same type of fraud will be close together under some metric. These metrics will be calculated on this module to provide input for the cluster generation module. During this phase, the signatures will be identified based on the known types of frauds or suspicious behaviour related to purchasing phase.

The signatures will be split into two different groups: time-dependent and time-independent. Time-independent signatures do not consider the time of the events as an input in order to generate a score. An example could be a PO which was not approved through the standard approval workflow. For this case, the time when the PO was approved or the actions which took place before or after the approval does not impact the score of the signature. Time-dependent signatures have the opposite behaviour. On this group, it can be included a signature

called retrospective PO, on which the date on the invoice received is earlier than the PO creation date, so time plays a decisive role in this signature.

Finally, the identification of a document on a signature does not necessarily mean that it is a fraudulent document. This is happening because each signature is one of the possible symptoms which can identify a fraudulent transaction, however, there may be valid business reasons for this. An example of this case could be a purchase of an expensive spare part of a production machine which failed during the weekend. During this scenario, the responsible manager which should approve the purchases through an automated workflow was not available on the company and, due to the urgency, the factory manager (which has the same hierarchy level and approval powers) could authorize the purchase manually in a process parallel to the approval workflow, so even if this purchase document have signatures raised on this step, it was a completely normal process.

On table 7 are included the list of signatures that will be included in the model. This list was based on the papers of (ISLAM et al., 2010) and (BAADER, G.; KRCMAR, H, 2018) but was enhanced with the knowledge of the auditors of the company which is subject of this study.

Table 7: List of signatures to be included in the model



Area	Description
Goods receipt	Invoice for undelivered goods/services
Invoices and payment	Sequential invoice numbers
Invoices and payment	Changes performed on payment terms before a purchase
Invoices and payment	The invoice amount is higher than the order
Purchase order	Price above average for supplier and material
Purchase order	Quantity above average for supplier and material
Purchase order	Price above average for supplier
Purchase order	Quantity above average for supplier
Purchase order	Price above average for the material
Purchase order	Quantity above average for the material
Purchase order	Purchase Order is approved outside the standard workflow
Purchase order	Purchase order blocked
Purchase order	Retrospective PO (goods receipt)
Purchase order	Retrospective PO (invoice receipt)
Purchase order	Purchases are divided into several partial purchases in order to bypass the approval process
Purchase order	Price of the good increased after PO creation
Supplier	Supplier blocked before a purchase
Supplier	Vendors without phone number
Supplier	Vendors without address
Supplier	A vendor with the same bank account as another vendor
Supplier	Difference between supplier creation and first sale
Supplier	Bank data changed for supplier
Supplier	Sudden business activity with old "sleeping" supplier (sudden activity in non-active accounts)
Supplier	Difference between sales dates

Source: Created by the author

The logic details for each of the signature is included in appendix B.

Each of the signatures mentioned in table 7 will be identified through Structured Query Language (SQL) and should identify which purchase order is affected by each signature on the list.

This requires that each signature have a SQL code written as shown in figure 6.

Figure 6: Example of a signature identification through SQL

```
-- CHECK 099 -- INVOICE FOR UNDELIVERED GOODS/SERVICES

IF OBJECT_ID('Output_DB.dbo.CHECK099_INVOICE_UNDELIVERED_GOOD','U') IS NOT NULL
DROP TABLE Output_DB.dbo.CHECK099_INVOICE_UNDELIVERED_GOOD

SELECT INVO_RCPT.EBELN AS INVO_RCPT_EBELN,
       INVO_RCPT.EBELP AS INVO_RCPT_EBELP,
       INVO_RCPT.WERKS AS INVO_RCPT_WERKS,
       PURC_HEAD.BUKRS AS PURC_ORDE_BUKRS,
       PURC_HEAD.LIFNR AS PURC_ORDE_LIFNR,
       PURC_ITEM.MATNR AS PURC_ORDE_MATNR,
       GOOD_RCPT.VGABE AS GOOD_RCPT_VGABE
INTO Output_DB.DBO.CHECK099_INVOICE_UNDELIVERED_GOOD
FROM Main_DB.dbo.PARAMETERS
INNER JOIN MAIN_DB.DBO.EKKO AS PURC_HEAD ON PURC_HEAD.BUKRS = PARAMETERS.VALUE
                                     AND PURC_HEAD.LIFNR NOT LIKE 'PV%' --exclude plant vendor
                                     AND PURC_HEAD.LIFNR NOT LIKE 'PC%' --exclude product centre
                                     AND PURC_HEAD.LIFNR NOT LIKE 'IC%' --exclude intra-company
                                     AND PURC_HEAD.BSTYP = 'F'

INNER JOIN MAIN_DB.DBO.EKPO AS PURC_ITEM ON PURC_ITEM.EBELN = PURC_HEAD.EBELN

INNER JOIN MAIN_DB.DBO.EKBE AS INVO_RCPT ON INVO_RCPT.EBELN = PURC_HEAD.EBELN
                                     AND INVO_RCPT.VGABE = '2'

LEFT JOIN MAIN_DB.DBO.EKBE AS GOOD_RCPT ON GOOD_RCPT.EBELN = INVO_RCPT.EBELN
                                     AND GOOD_RCPT.EBELP = INVO_RCPT.EBELP
                                     AND GOOD_RCPT.VGABE = '1'

WHERE PARAMETERS.FIELD = 'BUKRS'

-- DELETE THE NULL VALUES TO KEEP ONLY THE GOOD RECEIPTS WITHOUT INVOICE RECEIPTS
DELETE FROM Output_DB.dbo.CHECK099_INVOICE_UNDELIVERED_GOOD WHERE GOOD_RCPT_VGABE IS NOT NULL
```

Source: Created by the author

The last step of the module is the scaling of the data since according to (MOHAMAD, I. B., USMAN, D. 2013) the results of K-Means algorithm increased significantly when data is standardized versus non-standardized data. The process chosen to standardize the data is Min-Max scaling, which works by converting the lowest value for each figure to zero and the highest value to 1. The normalized value can be represented as:

$$MM(X_{ij}) = \frac{X_{ij} - X_{\min}}{X_{\max} - X_{\min}} \quad (4.1)$$

The final product of the module will be a single table containing one PO per line and one column per signature. This output table will be used as input on the cluster generation module.



4.5 Cluster generation and validation

In the cluster generation module, the scores from the signatures previously identified will be submitted clustered using nine cluster algorithms, several different parameters for each algorithm and validated against three internal indexes in order to choose the which algorithm and which parameter provided the best internal index scores.

In order to reduce the time to implement the clustering algorithms, make the results reproducible and to reduce the chances of coding errors, the algorithms were imported from open source libraries as shown below:

Sci-kit learn library (PEDREGOSA, F. et al., 2011):

- K-Means
- DBSCAN
- BIRCH
- OPTICS
- Spectral Clustering

Py Clustering library (NOVIKOV, A., 2019):

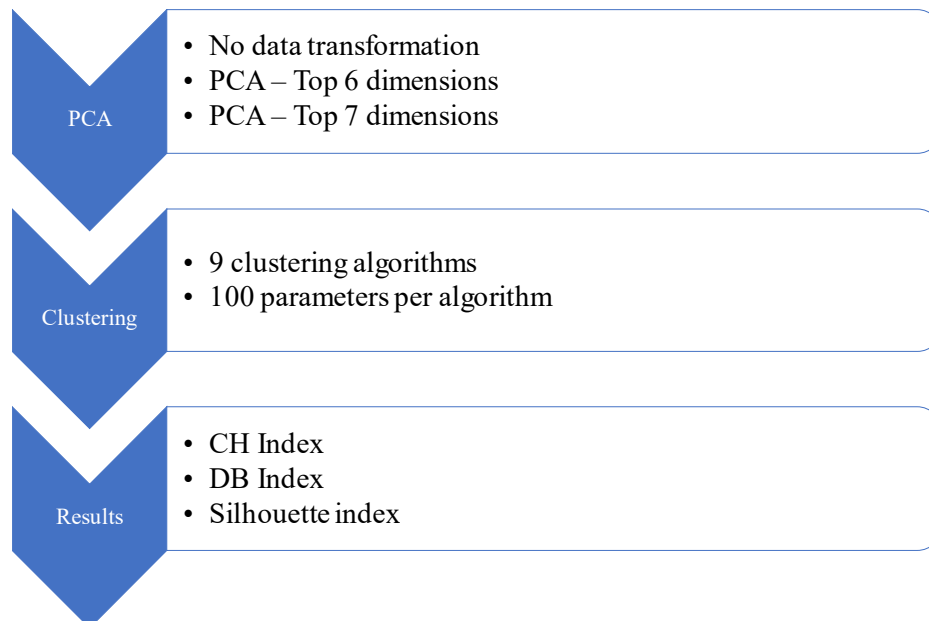
- CURE
- CLIQUE
- ROCK

HDBSCAN library (MCINNES, L., HEALY, J., ASTELS, S., 2017)

- HDBSCAN

This module will have a three-step approach which can be found in figure 7.

Figure 7: Steps performed on clustering generation module



The first step of the cluster generation module is to reduce the dimensionality of the dataset. According to the papers of (DING, C., XIAOFENG, 2004) and (MIN, X., LIN, R., 2018), this approach proved to achieve better results when compared to executing the clustering algorithms without the prior execution of PCA.

Based on the analysis of the data performed, it was decided to execute the clustering algorithms with three variations of the dataset:

- Original data set, without PCA;
- PCA algorithm executed and top 6 dimensions were selected;
- PCA algorithm executed and top 7 dimensions were selected.

The PCA algorithm used in this model was imported from Sci-kit learn (PEDREGOSA, F. et al., 2011) library.

Step two of the clustering module is to run all the 9 clustering algorithms for each of the three datasets previously created using parameter values for each clustering algorithm.

The parameter for each clustering algorithm is essential to achieve the best possible clustering results, for example, on K-Means algorithm the parameter K needs to be adjusted in order to identify what is the best number of clusters. Table 8 shows the parameters chosen for each clustering algorithm.

Table 8: Parameters used by each clustering algorithm

Algorithm	Parameter	Meaning
K-Means	K	Number of clusters to be created by the clustering algorithm
DBSCAN	eps	Minimum distance for a member to become part of a cluster
BIRCH	Threshold	The maximum distance between an element and existing subclusters
HDBSCAN	min_cluster_size	The minimum number of elements which can generate a cluster
CURE	number_clusters	Number of clusters to be created by the clustering algorithm
OPTICS	max_eps	The maximum distance of which a member can be part of a cluster
Spectral Clustering	n_clusters	Number of clusters to be created by the clustering algorithm
CLIQUE	tau	Density threshold
ROCK	cluster_numbers	Number of clusters to be created by the clustering algorithm

Source: Created by the author

After the execution of the clustering algorithm is performed, the internal indexes CH Index, DB Index and Silhouette index will be calculated and saved so the results could be compared and the best clustering algorithm and best parameter selected.

The best execution of among all will be the one where the best values for at least two of the three internal indexes points to the same combination of PCA, clustering algorithm and parameter.

In case the combination of the best values of the internal indexes are related to different combinations of PCA, clustering algorithm and parameter, all of the three executions will be selected for further evaluation.

4.6 Final selection

The final selection will decide which are the clusters with the highest probability of fraud by checking the average values of the signatures for the signatures of all the members which belongs to each cluster.

Clusters with low average values would be considered as low risk and clusters with the highest average values would be considered as higher risk.

5 EXPERIMENTAL RESULTS

Experiments were performed on a real-world data set to investigate the behaviour of the proposed approach using procurement data created over a period of 1 year by 9 different countries.

Initially the data set used for this experiment will be presented, followed by the baseline used for the evaluation and finally, the results achieved.

5.1 Technical aspects

The proposed model was implemented using Python and SQL languages. The following software was used:

SQL Server 2017: relational database management system developed by Microsoft with support to SQL language (<https://www.microsoft.com/en-us/sql-server/sql-server-2017>).

Anaconda: Open source Python distribution for scientific computing (<https://www.anaconda.com/distribution/>).

Live Compare: Tool to extract data from the ERP environment into the database used by this experiment (<https://www.intellicorp.com/livecompare>).

The following libraries and modules were used for Python language:

Pandas: This module provides high-performance, easy-to-use data structures and data analysis tools (<https://pandas.pydata.org/>).

Numpy: Is a package for scientific computing in Python (<https://numpy.org/>).

Sci-kit Learn: Scikit-learn is a free software machine learning library (<https://scikit-learn.org/stable/>).

Py-Clustering: Python and C++ data mining library with a focus on clustering and neural networks (<https://pypi.org/project/pyclustering/>).

HDBSCAN: Library that implements the HDBSCAN clustering algorithm (<https://hdbscan.readthedocs.io/en/latest/index.html>).

The implementation of the model was performed on a server configured with a CPU Xeon E7-8837 (4 cores) and 32 Gb of memory.

5.2 Dataset

The data set used for this experiment consists of 147,898 purchase orders created by 9 countries during the year of 2018.

The following types of procurement orders were not included in the dataset:

Intra-Company purchase orders – When one company of the group purchase goods or services from another company of the same group.

Plant Vendors – Similar to intra-company but more related to goods purchased from other factories in the group.

Repetitive direct procurement – Most important raw materials used in the factory, on which the procurement details are covered under contracts and the orders are created automatically to keep the minimum stock levels of raw materials.

5.3 Metrics

To answer RQ1 we will consider the number of PO documents which can be verified for fraud comparing random sampling versus the proposed model.

RQ2 will be calculated using internal criteria through Calinski-Harabaz, Silhouette and Davies-Bouldin indexes comparing all the clustering algorithms included in the study and the data sets involved.

5.4 Baseline

The baseline for RQ1 will be the time required from a professional auditor to search for frauds on a specific PO with the same scope as the proposed model. Since this information is yet to be confirmed, we will consider that each PO document will require 30 minutes for fraud verification. This number will be used on a linear equation to show the progression of how many POs can be verified over time using random sampling.

The baseline for RQ2 will be the usage of the K-Means algorithm due to it being the most used clustering algorithm on the fraud detection area.

5.5 Results

Data were extracted from the ERP environment using the LiveCompare tool and copied into a SQL Server 2017 database which will be used to perform this experiment.

Since this experiment was performed with data from different countries, as a consequence, the amounts involved were related to different currencies as well.

To solve this issue, the currencies were converted to a single currency using the average exchange rate for the period being processed.

Next step performed was the generation of the signatures shown in table 10 using a SQL script running on a SQL Server 2017 database.

Signatures were calculated at a PO level, which means that each PO will have scores for each of the signatures calculated individually.

Table 9 shows a sample of the scores achieved during the signature generation phase.

It is important to indicate that the supplier codes and document numbers were changed to sequential numbers/characters to ensure anonymity.

Table 9: Example of the output of the generated signatures

Company	PO Number	Supplier	PO Date	Amount	Signature 1	Signature 1	Signature 2	Signature 3	Signature 4	Signature 5	Signature 6	Signature 7
AU11	1	A	01/05/2018	710.72	0	0	0	0	0	1344	1	2
AU11	2	B	01/05/2018	506	0	0	0	0	0	310	1	14
AU11	3	C	01/05/2018	97	0	0	0	0	0	1664	4	15
AU11	4	D	01/05/2018	352.27	0	0	0	0	0	450	1	15
DE10	5	E	01/05/2018	45.45	0	0	0	0	1	1666	1	25
DE10	6	F	01/05/2018	431.82	0	0	0	0	0	469	1	13
DE10	7	G	01/05/2018	272.73	0	0	0	0	0	1691	1	0
AU11	8	H	05/06/2018	560	0	0	0	0	0	1706	2	12
AU11	9	I	05/06/2018	50	0	0	0	0	1	1	1	8
AU11	10	J	05/06/2018	240	0	0	0	0	0	1706	2	12
AU11	11	K	05/06/2018	208.64	0	0	0	0	0	1664	2	21
AU11	12	L	05/06/2018	49.09	0	0	0	0	0	1664	2	21
AU11	13	M	05/06/2018	45.45	0	0	0	0	0	1664	2	21
AU11	14	N	05/06/2018	190.55	0	0	0	0	0	390	1	21
BD10	15	O	05/06/2018	45	0	0	0	0	0	1506	1	12
BD10	16	P	05/06/2018	574	0	0	0	0	1	783	1	20
AU11	17	Q	05/06/2018	17.53	0	0	0	0	0	1726	1	0
AU11	18	R	05/06/2018	45.45	0	0	0	0	0	855	3	19
KE12	19	S	05/06/2018	45	0	0	0	0	0	1666	1	14
KE12	20	T	05/06/2018	200	0	0	0	0	0	1655	1	29

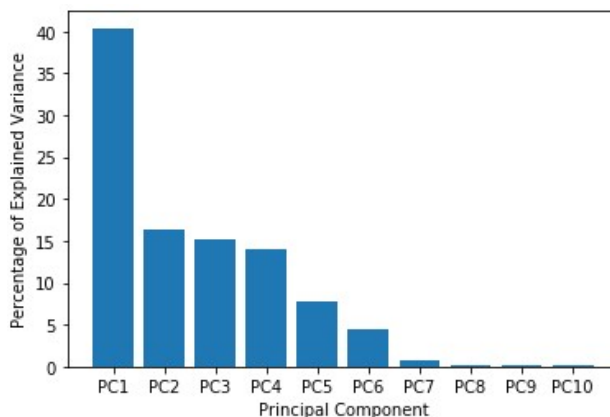
Source: Created by the author

After the signatures were calculated, the results of the signatures were scaled using the MinMaxScaling method from Scikit learn library.

The next step performed was the execution of the PCA to check if it could be used on the data to reduce the dimensionality of the dataset but keeping most of the variance.

Results are available in figure 8 which shows that keeping 6 dimensions would maintain 97.9% of original data variance and 7 dimensions would maintain 98.6%.

Based on the results from PCA, the data was executed on the clustering algorithms in three different formats: a) No PCA executed; b) PCA executed and keeping the top 6 principal components; c) PCA executed and keeping the top 7 principal components.

Figure 8: PCA result, variance per component

Source: Created by the author

In order to ensure a possible implementation of this methodology in a productive system, a run-time of 12 hours was selected as being the maximum time a clustering algorithm could run. In case a clustering algorithm takes more than 12 hours, it will be considered unfit for purpose.

The following clustering algorithms were executed:

- K-Means
- DBSCAN
- BIRCH
- HDBSCAN
- CURE
- OPTICS
- Spectral Clustering
- CLIQUE
- ROCK

Spectral Clustering, CLIQUE and ROCK failed due to lack of enough memory. Virtual memory of the server running this algorithm was increased to 64 Gb but the issue remained the same. Due to this issue, these clusters algorithms were classified as not fit-for-purpose for the dataset being processed and the server capacity available.

Algorithm CURE and OPTICS failed with lack of memory error only when the processing the data on its original format (with 25 columns), however when selecting the top dimensions from PCA the algorithm started the execution without this error however both algorithms failed to finish within a 12 hours run-time limit, hence both were classified as not fit-for-purpose.

BIRCH algorithm was able to finish within the 12 hours of run-time, however when the value of the parameter threshold was to be above 0.11, the execution didn't finish within the 12 hours limit, hence only 6 executions were possible with this algorithm.

K-Means and DBSCAN were executed both with 100 executions without any problems to report, with the average run-times of 245 and 725 seconds respectively when the entire dataset was processed, with the run-time significantly reduced when PCA was applied.

HDBSCAN was executed initially 100 times however without any issues, however the values of the associated signatures were still improving, which suggested the optimal value of the parameter minimum cluster size had not been achieved yet, hence the algorithm was executed an additional 126 times until the value of the results of the internal validation indexes started to reduce.

Table 10 shows the average execution times per clustering algorithm. Algorithms considered as fit-for-purpose for the dataset currently being processed are highlighted in green, while the algorithms considered as being not fit-for-purpose are highlighted in amber.

Table 10: Average run times of clustering algorithms

Algorithm	Average run-time in seconds					
	Source data		PCA - 7 Dimensions		PCA - 6 Dimensions	
	Number of executions	Average run time	Number of executions	Average run time	Number of executions	Average run time
K-Means	100	245.70	100	173.41	100	236.58
DBSCAN	100	725.14	100	446.92	100	526.66
BIRCH	6	10.83	6	6.56	6	5.89
HDBSCAN	241	210.18	241	13.29	241	7.56
CURE	-	Lack of memory	-	Timeout - 12 hours	-	Timeout - 12 hours
OPTICS	-	Lack of memory	-	Timeout - 12 hours	-	Timeout - 12 hours
Spectral Clustering	-	Lack of memory	-	Lack of memory	-	Lack of memory
CLIQUE	-	Lack of memory	-	Lack of memory	-	Lack of memory
ROCK	-	Lack of memory	-	Lack of memory	-	Lack of memory

Source: Created by the author

Next step performed was to compare which clustering algorithm generated the best results for the internal indexes as a measure of the quality of the clusters generated. Only the clustering algorithms K-Means, DBSCAN, BIRCH and HDBSCAN will be considered from this point onwards.

Calinski-Harabasz scores the best result for the K-Means algorithm, irrespectively of the data used as an input being the original data set or the two datasets created using the top components via PCA.

Davies-Bouldin and Silhouette indexes scored the best values for HDBSCAN algorithm for the original dataset. The same algorithm applied to dataset generated via PCA had significantly worst results.

The clustering algorithm chosen as the one with the best result achieved was HDBSCAN since it had the best scores for two of the three internal validation indexes, hence RQ2 can be answered with HDBSCAN being the best clustering algorithm to be used with the dataset considering the dataset used due to three points:

- 1 – Capacity to handle a large dataset with multiple dimensions
- 2 – Having the best overall performance
- 3 – Having the best results for the DB index and Silhouette with slightly inferior results on CH index when compared to the baseline K-Means clustering algorithm.

Table 11 shows the best results achieved by each clustering algorithm.

Table 11: Best internal index value achieved per clustering algorithm

PCA Performed / Dimensions	Algorithm	Max of CH Index	Min of DB Index	Max of Silhouette Index
YES-6	BIRCH	7,184.1	1.18336	0.31691
YES-6	DBSCAN	31,300.0	0.78809	0.49017
YES-6	HDBSCAN	12,179.7	0.75381	0.49686
YES-6	K-Means	99,112.2	0.84192	0.48378
YES-7	BIRCH	9,036.4	1.09877	0.32323
YES-7	DBSCAN	31,300.0	0.79820	0.49207
YES-7	HDBSCAN	56,817.8	0.63224	0.53223
YES-7	K-Means	99,112.2	0.84155	0.48378
No PCA applied	BIRCH	15,582.1	0.68663	0.34748
No PCA applied	DBSCAN	15,621.0	0.68818	0.48885
No PCA applied	HDBSCAN	87,980.2	0.53937	0.55652
No PCA applied	K-Means	99,112.2	0.84163	0.48378

Source: Created by the author

Considering that HDBSCAN was selected as the clustering algorithm which will be used for clustering the data for this specific dataset, the next step is to select the parameter which provides the best values for the internal indexes.

HDBSCAN was executed with parameter **min_cluster_size** initially starting with the value of 10 elements per cluster. This value was increased until the internal validation indexes stop to improve.

In order to have a better visualization of the behaviour of the indexes, CH was scaled to fit the same variation of the other two indexes (between 0 and 1). The scaled used was:

Scaled CH Index 0 = Real CH index value of 0

Scaled CH Index 1 = Maximum CH achieved for the data processed (87975.91)

Figure 9: Variation of internal indexes based on the clustering parameter



Source: Created by the author

According to figure 9, there are two possible candidates as the best parameters to be used. With min_cluster_size parameter with a value of 400 elements per cluster, the following values are achieved among the three indexes.

DB index = 0.54

Silhouette index = 0.56

CH index = 18128 (scaled value of 0.21)

The second option is with `min_cluster_size` parameter with a value of 3900 elements per cluster, which generates the following indexes:

DB index = 0.66

Silhouette index = 0.54

CH index = 79698 (scaled value of 0.91)

Comparing the values for both executions, we have the following conclusion:

Parameter 400 has a better score on two of the three indexes when compared to parameter 3900, with DB index having a score 22.2% better and Silhouette index 3.57% better, however, the CH index having a score 333% better for the parameter value of 3900.

In addition, parameter 400 generated a total of 33 clusters (plus outliers) while the parameter 3900 generated 5 clusters (plus outliers), making the classification easier to comprehend.

Based on the fact of the significant increase of the CH index and the reduced number of clusters, parameter value of 3900 was chosen as the one that produces the best clusters for the HDBSCAN for the dataset used.

The next step on the methodology would be to select which are the clusters which should be considered as high risk or low risk. This classification can be achieved by either manual analysis or in an automated way without human intervention.

5.5.1 Manual classification

The scores for the signatures were manually analysed and the clusters were classified based on the knowledge of the user performing the analysis. The result of this analysis is shown below:

Cluster 0: This cluster has the highest average values for retrospective PO for goods receipts and the second high for invoice receipts as well as all the vendors without phone numbers in the master data.

The retroactive POs, even not following the default policy of the company, can not be an indication of fraud by themselves even when combined with the fact of missing.

Based on this, we can not see any pieces of evidence of possible fraud on this cluster and hence classify it as low risk. This cluster contains 8836 POs.

Cluster 1: The only signature with high values on this cluster is the number of suppliers sharing a bank account. Even not being a common item, there are times that multiple vendors can share the same bank account, for example for payment of taxes, when a vendor is created for each tax but all of them should be paid to the government on the same account.

Since there are no further signatures associated, we can classify this cluster as low risk. This cluster contains 74257 POs.

Cluster 2: There are no indications in any signatures on this cluster, hence it is classified as low risk. It contains 4333 POs.

Cluster 3: This cluster contains the highest signatures for the sequential invoices, which means that the suppliers involved send most of their invoices to the company object of this study coupled with the fact that all the POs were approved outside the default approval workflow. The only case foreseen for this are the invoices submitted by contractors directly hired by the company, hence most of their invoices (if not all of them) are sent to the same company. This behaviour by itself is not worrying and the cluster would be considered as low risk, however as a control to ensure this is really the case the suppliers which are part of this group could be verified to be really contractors in order to completely eliminate this risk.

In summary, this cluster is classified as low-risk and contains 3952 POs.

Cluster 4: Cluster 4 is very similar to cluster 3, with the only signature active is that all the POs were approved outside the default approval workflow. It contains 35068 documents and it is classified as low risk.

Outliers: The outlier is the cluster which should be included as the scope for an audit in this area since it has the highest score for 16 out of the 22 signatures available and hence there is a risk for several types of frauds. It contains 21452 POs and is classified as the only high-risk cluster.

After a review of the results, it was identified that the company is currently performing a project to change the approval workflow for procurement orders, which is being implemented on a market by market basis. This could be the main reason for the high number of clusters showing POs not being approved through the standard workflow and hence be a false positive.

The values for the signatures Invoice for undelivered goods/services didn't identify a single case across all the POs which were part of this experiment.

The values for signatures difference between supplier creation and first sale and quantity above average for supplier and material were considered normal across all clusters.

The average results for the signatures used to perform the manual classification can be found in table 12 with significant values highlighted in green and signatures not considered in amber.

Table 12: Average values for signatures among the clusters with the most significant values highlighted

Signatures	Clusters generated					
	Outliers	0	1	2	3	4
Invoice for undelivered goods/services	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Invoice amount is higher than order	12.97786	9.86555	9.03913	3.11562	3.98305	7.66559
Sequential invoice numbers	4.86654	0.51494	0.03721	8.84537	9.08629	0.06773
Sleeping vendor	0.02694	0.00000	0.00000	0.00000	0.00000	0.00000
Po performed outside standard workflow	0.57230	0.00000	0.00000	0.00000	1.00000	1.00000
Retrospective po (goods receipt)	0.25396	1.75928	0.18738	0.04154	0.34514	0.02310
Retrospective po (invoice receipt)	5.27009	5.17587	0.79992	1.21348	0.24848	1.05569
Vendors without phone number	0.17775	1.00000	0.00000	0.00000	0.00000	0.00000
Vendors without address	0.00685	0.00000	0.00000	0.00000	0.00000	0.00000
Difference between supplier creation and first sale	977.30146	673.27230	1134.05867	965.35680	1339.25607	1363.75918
Vendor with the same bank account as another vendor	1.05836	1.20032	1.23555	0.99792	0.98634	1.12513
Average difference between sales dates	42.45432	37.17021	28.67281	14.82022	10.71027	25.74612
Price increased after po creation	4.18561	0.71885	0.85943	0.31109	0.03264	0.05322
Price above average for supplier and material	1.12240	0.01829	0.03571	0.00000	0.00000	0.01178
Quantity above average for supplier and material	0.06075	0.02842	0.03704	0.00449	0.00026	0.00467
Price above average for supplier	10.87754	0.70644	1.21253	4.77470	0.80708	0.68366
Quantity above average for supplier	5.82948	0.33985	0.39617	0.22728	0.10648	0.27615
Price above average for material	12.99958	0.01829	0.03579	0.00237	1.42592	0.19439
Quantity above average for material	0.10398	0.06614	0.03742	0.00176	0.04113	0.01031
Multiple po for same creator, supplier, amount and	0.18050	0.00000	0.00000	0.00000	0.00000	0.00000
Bank data changed for supplier	0.00009	0.00000	0.00000	0.00000	0.00000	0.00000
Supplier blocked before a purchase	0.10656	0.00283	0.01212	0.00092	0.00177	0.03214
Supplier marked for deletion after a PO creation	0.10544	0.00283	0.01236	0.00092	0.00177	0.03291
Changes performed on payment terms before a purchase	0.58559	0.13886	0.11474	0.07339	0.01645	0.12065
Purchase order blocked	0.01380	0.00136	0.00059	0.00000	0.00000	0.00137
Number of PO per cluster	21452	8836	74257	4333	3952	35068

Source: Created by the author

5.5.2 Automated classification

Using the automated classification, the model should select the clusters with the best scores on each signature.

In order to make the methodology adaptable to any datasets, the approach decided was to choose 20% of the clusters (rounded to the nearest number) as the sample data to be audited.

Since the dataset used for this exercise, when clustered with HDBSCAN resulted in 6 clusters (considering the outliers as an additional cluster), the automated selection would select 1 cluster as to become the sample to be audited.

Table 13 shows the clusters with the highest scores on each signature. On this case cluster -1 (outliers) would be selected as the scope of the audit.

On this case, both the manual analysis and automated analysis reached the same conclusion.

Table 13: Clusters and number of signatures with top scores

Cluster	Signatures with highest score
Outliers	16
0	3
1	1
2	0
3	2
4	1

Source: Created by the author

5.5.3 Manual verification of POs

Taking the outliers as the POs with the highest likelihood of fraud, a total of 3 samples including 5 POs each were reviewed directly on the ERP environment. The size of the sample should be reduced since the activity of checking the details of the POs on the ERP environment is very time-consuming. Details of the sample are below:

Sample 1: Documents which had multiple POs created on the same day, to the same vendor, by the same person and with the same amount, the bank account used by these vendors were the same for other vendors, these POs were approved outside of the standard approval workflow and the supplier didn't have any contact phone.

Sample 2: Supplier created and a PO was performed only 2 days after the supplier creation, the bank account of the vendor is shared with other 25 vendors.

Sample 3: Supplier sending sequential invoices, with POs approved outside the standard approval workflow.

The manual analysis of the documents returned the following conclusions:

Sample 1: Multiple POs were created to the supplier since each PO was to deliver goods at different addresses of the company since it was the same good, the POs had the same amount as well. The bank accounts were shared with other different companies, however, all of them were part of the same group. This was created so that goods could be purchased from any of the locations the company operates but paid to the headquarters offices.

The addresses of the companies were checked using google maps and all cases matched.

Also, these invoices were related to a country where you can check the business details on a government website by providing the address of the company. All details matched perfectly. No pieces of evidence of fraud have been found and hence is considered a false-positive.

Sample 2: The POs were created to a vendor which was just included in the system 2 days ago. When checking the bank accounts details, the same was shared with several other companies which appear to be of the same group.

When checking the web site for the supplier, the same is not working anymore, so it is already a point to be taken into account.

The address provided for these suppliers are based in shopping centres, so the web site of these shopping centres was accessed and the name of the supplier was searched on the list of available stores/companies available for trading for each shopping centre. The result is that the box which should be occupied by the supplier is listed as empty and available for rent.

In addition, no description of the goods purchased is available on the PO. There are enough pieces of evidence that would justify a deeper investigation of this case.

Sample 3: The multiple invoices were related to a company related to electrical maintenance and invoices were created in a sequential way. The e-mail from the supplier was a yahoo email account which appears to be the name of a person and does not have any relation with the supplier.

Looking up the company name on google didn't provide any results and when checking the address via google maps, there is a physiotherapy business operating in the place of what should be the electrical maintenance company.

There are enough pieces of evidence that would justify a deeper investigation of this case.

In order to answer the RQ1, the heuristic currently being used in the audit area to select a data sample should be used and compared with the results achieved with the proposed methodology.

The audit methodology currently in use in the company is a risk-based approach, which means that not all the business processes are audited, only those who could have substantial associated risk, otherwise, the size and duration of the audits would become unfeasible.

This methodology has 3 steps:

Risks: The first step on this methodology is to identify which are the risks (from a business point of view) which should be covered under the scope of the audit.

Controls: Based on the concept that every business risk should be mitigated and/or controlled, a list of expected controls are created based on the risk list generated in the previous step.

This list of controls contains all the controls which should be in place (from the audit team point of view) so the risk could be controlled or mitigated in an appropriate way.

Test of controls: The last step is to create the test of controls. The test of controls formalizes how the controls previously defined should be tested to ensure it is working on the way it should.

Below is one example of the three steps:

Risk: Inadequate control over procurement of goods and services

Control: Adequate approvals are provided for each purchase, based on the company delegation of authority (the financial limit each person could approve a transaction within the company).

Test of control: Ensure that all POs are approved according to the release strategy configured in the system.

For each test of control, the methodology establishes the size of the data sample which should be used to analyse the effectiveness of the test of control. Daily business process should have a sample size of 30 documents, the weekly process should have a sample size of 10 documents and monthly processes should have a sample size of 3 documents.

This initial sample will look for any exceptions for the test of control, which in the example above, should be a PO which was not approved within the release strategy (default approval workflow of the system).

In case of any exception or any evidence identify which could suggest an exception, the sample size should be doubled.

If at any time, two or more exceptions are identified, the control is declared as ineffective in the audit and actions should be taken.

The audit methodology states that sample data selection could be either random or based on a professional judgement, which will vary according to the knowledge of the auditor.

Below are examples of three different approaches for data sampling which are the most common among the auditors.

Approach 1: Random sampling, selecting POs created up to 3 months before the audit period. Selecting the data created only during the last month before the audit reduces the initial scope to 38222 POs.

Approach 2: Perform filters on the amount of the PO and select for sample random documents above a value which is considered significant (POs above 10 thousand dollars). Selecting only the POs which are above 10 thousand US dollars, reduces the initial scope to 50269 POs.

Approach 3: Random sampling selecting only the documents which were approved outside the formal approval workflow. By selecting only the POs approved outside the formal approval workflow, reduces the initial scope to 51297 POs.

The heuristic chosen by the approach 1 does not increase the likelihood of identifying fraud among the universe since restricting the sample to only 3 months does not reduce or eliminate the risk of a fraud being performed outside of the selected period.

The approach 2, selected a heuristic which does not increase the chances of identifying frauds among the universe of POs, however, this heuristic restrict the initial scope only to POs which could have a higher financial impact of fraud. This means that even in the case a fraud

is happening on the POs with an amount lower than 10 thousand dollars, it will not be a fraud so significant which could impact the operation of the company as a whole.

The approach 3, should, in theory, increase the chances of identifying a fraudulent transaction using the logic that if a purchase was not approved through the correct levels and procedures, that should have an increased likelihood of fraud.

This fact, however, is challenged by the consideration that more than 48 thousand documents were approved outside the formal approval workflow.

The cluster -1 (outliers) used as the sample provided by this methodology identified 21452 documents as being candidates for frauds.

The final answer to RQ1 will be based on the three items below:

- The heuristics currently being used by the auditors provide a much higher number of POs as the universe where the sample should be selected when compared with the clusters classified as high-risk from the proposed solution.
- The current approach in use does not have any verification of a given PO has any indication of being a low or high risk when compared to the remaining population, with the exception of the approach 3 which only verifies the approval of the PO can be classified as incomplete since in case of the approver of the PO is involved in any fraudulent scheme and is approving the POs, they would not be audited.
- The analysis of 3 sets of POs from the high-risk cluster identified some evidence which would justify further investigation. This fact corroborates to the fact of the POs included in this cluster are indeed high-risk.

Considering that the number of POs included in the high-risk cluster of this solution is the lowest one among the remaining heuristics, the fact that this cluster was generated based on symptoms which are related to fraud in the literature and the actual verification of the POs classified as high-risk identified evidence which would require a further investigation, we can answer RQ1 that the proposed model has a higher chance of identifying a fraudulent PO when compared to the heuristics currently in use.

5.6 Limitations

During the implementation of the proposed methodology, some limitations were identified, which will be clarified on this section.

Dependency on signatures: Although the signatures were implemented to be as generic as possible and cover as many possible fraud types, however, if a type of fraud does not impact the score of a significant number of signatures, it will not be identified by this methodology.

Automated classification of clusters: the final selection of which clusters should be used as a sample for the auditors is based on selecting the clusters with the top scores for each signature. The clusters with the highest number of top scores are selected to become the sample, however, this approach has two issues:

Some signatures may have a top score, but are so small which should not be considered as a real symptom. For example, the signature quantity above average for supplier and material

has the highest score for the cluster -1 (outliers), however, the value of 6% of quantities is considered a normal figure.

The second issue is related to signatures with similar values, this happened for the signature retrospective PO (invoice receipt), when cluster -1 (outliers) was selected as having the top score, however, the score of cluster 0 was only slightly lower than the value for cluster -1 and was not included in the automatic selection.

6 CONCLUSION

This work started with a review of the state-of-the-art associated with fraud detection on the procurement area, compared the proposed solutions and identified the gaps and challenges. Based on the study performed, we identified several approaches for fraud detection, including clustering, process mining, descriptive statistics and rule-based approaches. However, when including the requirement of having an automated approach which does not require manual analysis of the data, two approaches were used: rule-based model and clustering. Since to the best of our knowledge, there was no work being performed consolidating both of these techniques for fraud detection on procurement, we adopted this approach.

We used the reliability of rule-based models for identifying known types of frauds with the flexibility of clustering which can identify associations between key signatures on data to identify even unknown types of frauds. The model does not require labelled data to work and can be completely automated.

Experiments were performed using real procurement data to compare the clustering algorithms which provides the best results and which metrics could be used to choose the best clustering algorithms.

Manual analysis was performed using 3 samples of POs which scored the highest signature scores from within the cluster selected as high-risk and two of the PO samples has evidence which can indicate that further investigation is required.

Finally, the signatures which were previously used to generate the clusters and automatically identify the groups of POs among clusters in order to identify documents with the highest likelihood of fraud can be used as additional information by the auditors during the audit itself so documents with different types of associated signatures could be selected in the sample.

6.1 Contributions

The first contribution of this work is the submission of the article “A HYBRID APPROACH FOR FRAUD DETECTION ON PURCHASE ORDERS” to the 20th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL) hold by the University of Manchester, UK.

In addition, this work provided a new approach for the selection of data to be used in audit in the procurement area in a methodology which can either be completely automated or be improved with human analysis in the final phase with better results.

This approach compared several clustering algorithms available and identified the most suitable ones for the problem of fraud detection in clustering, including the methodology of evaluation of which algorithm could provide the best results so the same could be used on future work.

In addition, this work included an implementation detection approach of the symptoms related to fraud which are available in chapter 2, providing the list of all the tables where this information could be found on an SAP ECC based ERP system, so the same could be replicated if desired.

Finally, the implementation of the approach can be either executed in a completely automated way or with manual analysis of the generated clusters by an auditor to select which

clusters could be considered of high-risk. For the experiment performed, the result of both analyses was exactly the same, however, this approach gives the flexibility to either run a model in an automated way to get results in almost realtime or performing a manual analysis of the data in order to select the most appropriate clusters.

6.2 Future work

A future work related to this topic would be the creation of a dataset for the procurement area. A possibility which could be used is the list of tables of an ERP environment available on table 5.

The list of tables includes most of the data related to purchases (based on the ERP system called SAP ECC). Since the main issue is related to the availability of public data related to the system, future authors could be generating fraudulent data on a training environment and making a labelled dataset available.

Due to the issue of not having public procurement dataset available, at this moment there are no papers comparing the different approaches currently available for fraud detection. Such a study could make a proper comparison of the different approaches currently available.

Another point identified as future work is that even reducing the universe of POs to less than 15% of the original sample size, the number of POs still classified as high-risk is very large, with more than 20 thousand documents. Subsequent work could focus on different approaches of how to narrow the selection further down using different approaches.

REFERENCES

- ASSOCIATION OF CERTIFIED FRAUD EXAMINERS (ACFE). **Report to the nations.** Austin, USA, 2016. Available on: <https://www.acfe.com/rtn2016/docs/2016-report-to-the-nations.pdf> Access on: March 3rd 2019
- ASSOCIATION OF CERTIFIED FRAUD EXAMINERS (ACFE). **Report to the nations.** Austin, USA, 2018. Available on: <https://s3-us-west-2.amazonaws.com/acfe-public/2018-report-to-the-nations.pdf> Access on: March 3rd 2019
- AGRAWAL, R. et al. Automatic Subspace Clustering of High Dimensional Data. **Data Mining and Knowledge Discovery**. [S.1], V. 11, n. 1, p. 5-33, 2005
- BAADER, G.; KRCMAR, H. Reducing false positives in fraud detection: Combining the red flag approach with process mining. **Int. J. Account. Inf. Syst.**, [S.1], V. 31, n., p. 1-16, 2018
- BANERJEE, A.; DAVE, R. Validating clusters using the Hopkins statistic. In: 2004 IEEE International Conference on Fuzzy Systems, 07., 2004, Budapest, Hungary, Hungary. **Proceedings...** IEEE, 2004, p. 149–153
- CAMPELLO, R., MOULAVI, D., SANDER, J. Density-Based Clustering Based on Hierarchical Density Estimates. In: Advances in Knowledge Discovery and Data Mining, Berlin, Heidelberg, Germany. **Proceedings...** PAKDD, 2013, vol. 7819
- CARLSSON C.; HEIKKILA M.; WANG X. Fuzzy C-Means for Fraud Detection in Large Transaction Data Sets. In: 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). 10., 2018, Rio de Janeiro, RJ, Brazil. **Proceedings. . .** IEEE, 2018, p. 1–6, 2018, IEEE
- CHAUHAN, P.; SHUKLA, M., A review on outlier detection techniques on data stream by using different approaches of K-Means algorithm. In: 2015 International Conference on Advances in Computer Engineering and Applications, 03., 2015, Ghaziabad, India. **Proceedings...** IEEE, 2015, p. 580-585
- DING, C., XIAOFENG, H. K-means Clustering via Principal Component Analysis. **Computational Research Division, Lawrence Berkeley National Laboratory.** p.29, 2004
- DURTSCHI, C.; HILLISON, W.; PACINI, C. The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data. **Journal of forensic accounting**, [S.1], V. 5, n. 1, p.17-34, 2004
- ELAVARASI, S.; AKILANDEAWARI, J.; SATHIYABHAMA, B. A Survey on Partitional Clustering Algorithm. **International Journal of Enterprise Computing and Business Systems**. [S.1], V. 1, n. 1, p. 1-13, 2011
- ESTER, M., et all. Density-Based Clustering Algorithms for Discovering Clusters in Large Spatial Databases with noise. **Kdd**. [S.1], V. 96, n. 34, p. 226-231, 1996
- FINANCIAL CRIMES ENFORCEMENT NETWORK. Customer Due Diligence Requirements for Financial Institutions. **Federal Register**. p. 29397- 29458, 2016

- GUHA, S.; RASTOGI, R.; SHIM, K. Cure: an efficient clustering algorithm for large databases. **Information systems**. [S.1], V. 26, n. 1, p. 35-58, 2001
- HUBER, M.; IMHOF, D. Machine Learning with screens for detecting Bid-Rigging Cartels. **Working Papers SES**. [s. l.], n. 494, p. 1-28, 2018
- IMHOF, D.; KARAGÖK, Y.; RUTZ, S. Screening for bid rigging-does it work? **Journal of Competition Law and Economics**. [s. l.], V. 14, n. 2, p. 235–261, 2018.
- ISLAM, A. et al. Fraud detection in ERP systems using Scenario matching. In: IFIP International Information Security Conference, 09., 2010, Brisbane, QLD, Australia. **Proceedings**. . . Springer, 2010, p.112-123
- JANS, M. et al. A business process mining application for internal transaction fraud mitigation. **Expert Systems with Applications**. [S.1], V. 38, n. 10, p. 13351-13359, 2011
- JONASSON, J., OLOIFSSON, M., MONSTEIN, H. J. Classification, identification and subtyping of bacteria based on pyrosequencing and signature matching of 16S rDNA fragments. **Apmis**. [S.1], V. 110, n. 3, p. 263-272, 2002
- LEE, Y. et al. Using Mahalanobis–Taguchi system, logistic regression, and neural network method to evaluate purchasing audit quality. **Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture**. [S.1], V. 229, n. 1\suppl, p. 3-112, 2012
- LONDON STOCK EXCHANGE (LSE) **Corporate Governance for Main Market and AIM Companies**. London, UK, 2012 Available on: <https://www.londonstockexchange.com/companies-and-advisors/aim/publications/documents/corpgov.pdf> Access on: March 3rd 2019
- MACQUEEN, J. Some Methods for classification and Analysis of Multivariate Observations. In: 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, Oakland, CA, USA. **Proceedings...** p. 281–297, 1967
- MAULIK, U., BANDYOPADHYAY, S. Performance evaluation of some clustering algorithms and validity indices. **IEEE Transactions on Pattern Analysis and Machine Intelligence**. [S.1], V. 24, n. 12, p. 1650-1654, 2002
- MCINNES, L., HEALY, J., ASTELS, S. HDBSCAN: Hierarchical density based clustering. **The Journal of Open Source Software**. [S.1], V. 2, n. 11, 2017. <https://doi.org/10.21105/joss.00205>
- MILLIGAN, G.; COOPER, M. An examination of procedures for determining the number of clusters in a data set. **Psychometrika**. [S.1], V. 50, n. 2, p. 159-179, 1985
- MIN, X., LIN, R. K-Means Algorithm: Fraud Detection Based on Signalling Data. In: 2018 IEEE World Congress on Services (SERVICES), 07., 2018, San Francisco, CA, USA. **Proceedings**. . . IEEE, 2018, p. 21-22

MOHAMAD, I.; USMAN, D. Standardization and its effects on K-means clustering algorithm. **Research Journal of Applied Sciences, Engineering and Technology**. [S.1], V. 6, n. 17, p. 3299-3303, 2013

NEW YORK STOCK EXCHANGE (NYSE). Corporate Governance Listing Standards. New York, USA, 2014. Available on:
https://www.nyse.com/publicdocs/nyse/listing/NYSE_Corporate_Governance_Guide.pdf
 Access on: March 3rd 2019

NOVIKOV, A. PyClustering: Data Mining Library. **Journal of Open Source Software**. [S.1], V. 4, n. 36, p. 1230, 2019

ORGANISATION FOR ECONOMIC CO-OPERATIONS AND DEVELOPMENT (OECD). **Guidelines for fighting bid rigging in public procurement: Helping governments to obtain best value for money**. Paris, France, 2009 Available on:
<http://www.oecd.org/daf/competition/cartels/42851044.pdf> Access on: March 3rd 2019

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**. [S.1], V. 12, p. 2825-2830, 2011

POPAT, S.; EMMANUEL, M. Review and Comparative Study of Clustering Techniques. **International Journal of Computer Science and Information Technologies**. [S.1], V. 5, n. 1, p. 805-812, 2014

PORTNOY, L. Intrusion detection with unlabelled data using clustering. Undergraduate thesis (Bachelors). Columbia University. 2000.

PRICEWATERHOUSECOOPERS (PWC). **PwC's Global Economic Crime Survey 2018: UK findings Pulling fraud out of the shadows**. London, UK, 2018. Available on:
<https://www.pwc.co.uk/forensic-services/assets/gecs/global-economic-crime-survey-2018-uk-findings.pdf> Access on: March 3rd 2019

RENDON, E. et al. Internal Versus External Cluster Validation Indexes. **International Journal of computers and communications**. [S.1], V. 5, n. 1, p. 27-34, 2011

RENDON, G.; RENDON, J. Auditability in public procurement: an analysis of internal controls and fraud vulnerability. **International Journal of Procurement Management**. [S.1], V. 8, n. 6, p. 710-730, 2015

SABAU, A. Survey of Clustering based Financial Fraud Detection Research. **Informatica Economica**. [S.1], V. 16, n. 1, p. 110-122, 2012

SHARAN, R., SHAMIR, R. CLICK: a clustering algorithm with applications to gene expression analysis. **Proceedings of International Conference on Intelligent Systems for Molecular Biology**. [S.1], p. 307-316, 2000

SMITH, R. et al. Evaluating GPUs for network packet signature matching. In: 2009 IEEE International Symposium on Performance Analysis of Systems and Software, Boston, MA, USA. **Proceedings...** IEEE, 2009, p. 175-184

TRANSPARENCY INTERNATIONAL. **CORRUPTION PERCEPTION INDEX**. 2017, Available on:

https://www.transparency.org/news/feature/corruption_perceptions_index_2017#table

Access on: March 3rd 2019

WALLACE, W. Assessing the quality of data used for benchmarking and decision-making. **The Journal of Government Financial Management**. [S.1], V. 51, n. 3, p. 16, 2002

WESTERSKI, A. et al. Prediction of enterprise purchases using Markov models in procurement analytics applications. **Procedia Computer Science**. [S.1], V. 60, p. 1357-1366, 2015

XU, D. and TIAN, Y. A Comprehensive Survey of Clustering Algorithms. **Annals of Data Science**. [S.1], V. 2, n. 2, p. 165-193, 2015

XU, R.; WUNSCH II, D. Survey of Clustering Algorithms **IEEE Transactions on Neural Networks**. [S.1], V. 16, n. 3, p. 645-678, 2005

YUE, D. et al. A review of data mining-based financial fraud detection research. In: 2007 International Conference on Wireless Communications, Networking and Mobile Computing, Shanghai, China. **Proceedings...** IEEE, 2007, p. 5519-5522

ZAREMSKI, A., WING, J. Signature Matching: A key to reuse. In: 1st ACM SIGSOFT symposium on Foundations of software engineering, Los Angeles, California, USA. **Proceedings...** ACM, 1993, p. 182-190

ZHANG, T., RAMAKRISHNAN, R., LIVNY, M. BIRCH: an efficient data clustering method for very large databases. **ACM Sigmod Record Proceedings...** [S.1], V. 25, n. 2, p. 103-114, 1996

APPENDIX A – RELATED WORKS

In this appendix, we are going to present a detailed review of the papers selected during our research phase.

The effective use of Benford's Law to assist in detecting fraud in accounting data

In the paper of (DURTSCHI, C.; HILLISON, W.; PACINI, C., 2004) the usage of the Benford's law is analysed under the optics of application on the audit field, to which situations it can be useful and which ones it should not be used.

The Benford law is an empirically observed phenomena identified first Simon Newcomb in 1881 and later on by (BENFORD, X., 1938), on which there are more numbers starting with lower digits rather than higher ones and the mathematical way of representing it is described below, in equation 3.1, for the first number of a value which is different than zero

$$P(d) = \text{Log}_{10}(1 + 1/d) \quad (3.1)$$

In equation 3.1, item d is a number between 1 and 9, and the item P is the observed probability. In addition, table 14 shows the expected values for the Benford distribution for the four first positions of a given number, which in short means that a probability of the first digit of a given number being a 1 is roughly 30% while the probability of the first number being a 9 is only 4.5%

Table 14: Distribution of digits according to Benford's Law

Digit	1st place	2nd place	3rd place	4th place
0		.11968	.10178	.10018
1	.30103	.11389	.10138	.10014
2	.17609	.19882	.10097	.10010
3	.12494	.10433	.10057	.10006
4	.09691	.10031	.10018	.10002
5	.07918	.09668	.09979	.09998
6	.06695	.09337	.09940	.09994
7	.05799	.09035	.09902	.09990
8	.05115	.08757	.09864	.09986
9	.04576	.08500	.09827	.09982

Source: DURTSCHI, C.; HILLISON, W.; PACINI, C. (2004)

(WALLACE, W., 2002) suggests that if the mean of a set of numbers is larger than the median and the skewness value is positive, the data set likely follows a Benford distribution.

In addition, the study provides an analysis on which sets of data the Benford distribution can be applied or not. The same is summarized below, in table 15.

Table 15: Application of Benford's Law to different data sets

When Benford Analysis Is Likely Useful	Examples
Sets of numbers that result from mathematical combination of numbers - Result comes from two distributions	Accounts receivable (number sold * price), Accounts payable (number bought * price)
Transaction-level data - No need to sample	Disbursements, sales, expenses
On large data sets - The more observations, the better	Full year's transactions
Accounts that appear to conform - When the mean of a set of numbers is greater than the median and the skewness is positive	Most sets of accounting numbers
When Benford Analysis Is Not Likely Useful	Examples
Data set is comprised of assigned numbers	Check numbers, invoice numbers, zip codes
Numbers that are influenced by human thought	Prices set at psychological thresholds (\$1.99), ATM withdrawals
Accounts with a large number of firm-specific numbers	An account specifically set up to record \$100 refunds
Accounts with a built in minimum or maximum	Set of assets that must meet a threshold to be recorded
Where no transaction is recorded	Thefts, kickbacks, contract rigging

Source: DURTSCHI, C.; HILLISON, W.; PACINI, C. (2004)

The study explains the methods for identifying the distribution of any data set can be considered anomalous. For this activity, the article proposes the method of calculating the standard deviation and after that applying a Z statistic test.

The negative aspects of the Benford law are that it requires a substantial number of deviant elements in a dataset so the same can be considered anomalous. This can be considered a problem since a small number of fraudulent transactions on a given data set can be classified as normal, even with a substantial amount on each independent fraudulent transaction.

Another limitation of the Benford analysis is that since it is based on how the numbers are distributed among a specific value or dataset, it cannot identify missing records which are absent.

A business process mining application for internal transaction fraud mitigation

In the paper of (JANS et al, 2011) the different types of fraud are split between internal and external as well as between transactional fraud and financial statement fraud, with the definitions defined below:

- Internal fraud: when the fraud is committed by a company employee
- External fraud: when the fraud is performed by someone outside the organization
- Financial statement fraud: 'the intentional misstatement of certain financial values to enhance the appearance of profitability and deceive shareholders or creditors'
- Transactional fraud: 'The intention with transaction fraud is to steal or embezzle organizational assets'

Due to frauds related to financial statements needs to fill with the government and in case of fraud, they need to be well analysed and documented to go to prosecution, there is a significant amount of data sets which can classify the items of a financial statement as regular or fraudulent.

The same applies to external frauds, since usually an external person/company needs to be charged with the fraud.

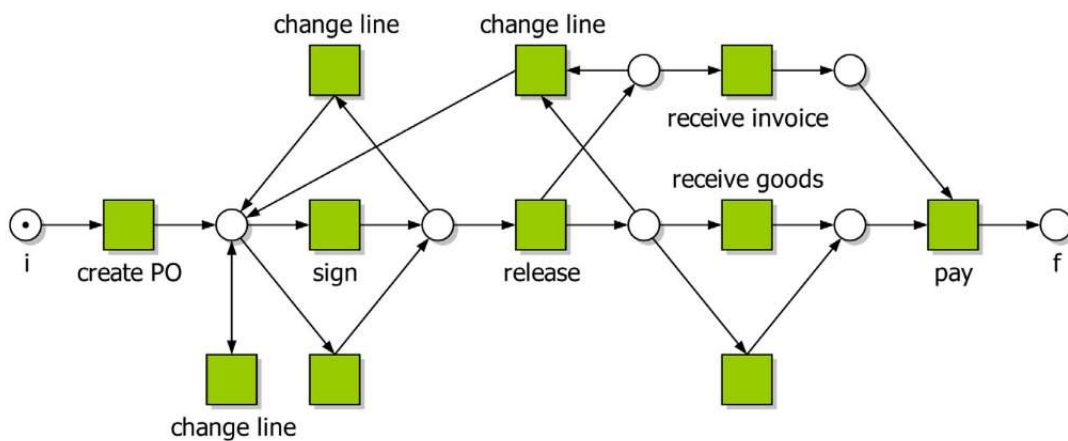
This behaviour though does not happen with transactional frauds executed internally in a company since according to (JANS et al., 2011) the stakeholders of a company usually lose faith in a company where there are stories related to internal frauds.

Due to the items above, the paper makes it clear about this being the reason for the lack of supervised data sets for internal transactional fraud.

The study performed a data-mining exercise on a sample of purchase orders documents from a European financial institution. This exercise selected randomly 10.000 records from an ERP system.

The focus of the paper was to show that through process mining, you can identify several behaviours which can be suspect of fraud, mainly regarding the sequencing of activities as the example shown on the Petri net below, in figure 10.

Figure 10: Process flow of a purchase order



Source: JANS et al. (2011)

What was noticed though, is that the rules and verifications that should be performed to differentiate a regular transaction from a transaction that can be categorized as suspect needs to be provided by a domain expert, making it clear that this approach requires a close iteration between the person developing/operating the application and the domain experts.

Fraud detection in ERP systems using Scenario matching

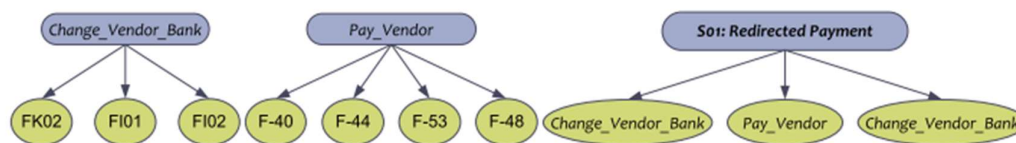
In the paper of (ISLAM et al., 2010) is presented a model based on patterns of computer activity, described as “signatures”.

The signatures are the definition of the pattern of a known type of fraud, which contrasts with statistical techniques which aim to identify irregular or statistically inconsistent data.

The paper also reviews all the signature specification languages available, their pros and cons. Next, the concept of signature is described as the union of activities and sequences, which means that a signature is more than just a specific activity in a given point in time, but a sequence of activities performed with have a specific meaning.

Figure 11 shows the signature of a redirected payment fraud, in which the banking details of the supplier is changed to a fraudulent account, the payment is performed but the funds are transferred to the fraudulent account in the system and finally, the account configuration is reverted to the supplier correct account.

Figure 11: Examples of scenario matching



Source: ISLAM et al. (2010)

This approach, although not flexible, can identify a specific fraud with a very level of confidence.

The approach used to implement this solution was to extract the data from the ERP system database tables which contains the information related to the signatures being processed, but in summary is comprised of three groups of tables: a) **Master data tables** – Holds the master data related to a customer/vendor; b) **Transactional data tables** – Holds all the transactional information related to a specific business process; c) **Log tables** – Keep records of all changes performed on critical tables and fields in the system.

To perform the identification of the signatures, the authors developed a system in which the signatures were written and based on these signatures, generated SQL code to read the data from the ERP tables and identify the number of occurrences of each scenario.

It was demonstrated that utilizing the signature-based approach, known types of frauds could be detected on ERP systems.

Reducing false positives in fraud detection: Combining the red flag approach with process mining

In the paper of (BAADER, G.; KRCMAR, H, 2018) the approach of combining the traditional red flag approach, which is recommended by most fraud auditing standards, with process mining.

The paper provides a detailed description of several types of fraud based on an extensive literature review in order to distinguish a fraudulent transaction versus a normal one.

One the first phase of the model, a SQL script was created in order to go through the data available and identify all the records which match the possible types of fraud.

On the next phase, the data was loaded into a process mining tool which uses fuzzy mining to display the discovered model in a graphical way in order to enable the users of the solution to make decisions if the documents previously raised on the previous phase are fraudulent or just normal transactions.

The main idea of the model is that using the process mining-related information in a visual way, the documents that were previously raised as a red flag could have the process flow of its execution analysed, and based on it, confirmed as a fraudulent document or discarded as a false positive raised by the red flag methodology.

To generate the test data, two different groups of students received a task to generate different frauds on the ERP system in a way that the other group was not able to detect the frauds.

After the period of generating the frauds was finished, each student group used the prototype solution to analyse the data created by the other group in order to detect the frauds.

The result was a detection rate of 48.38% across the groups and a false negative rate of 0.37% which was the lowest among the related work.

Fuzzy C-Means for Fraud Detection in Large transaction data sets

In the paper of (CARLSSON C., HEIKKILA M., WANG X., 2018) an analysis of the most common frauds in the procurement area as well as a new approach of using fuzzy logic to improve the quality of the clustering techniques is presented.

The paper was a case study on a multinational company with 75000 employees and more than 16 Billion USD in annual sales and was performed on internal master data from employees and vendors, transactional data related to purchases as well as data retrieved from the suppliers of this company.

One of the key aspects is the limitation of the classical clustering algorithms which limits each data record to belong to only one cluster versus fuzzy clustering which allows the same data record to belong to multiple clusters with a degree of similarity, where this degree is determined in $[0, 1]$ and the sum of all the degrees of similarity is equal to 1.

This difference can be crucial for data records which are in the limit between clusters, for example between a cluster of documents without indication of fraud and a cluster of documents with the indication of fraud. This document could be classified as non-fraudulent even with significant pieces of evidence which put the document close to a fraudulent group.

Tests were performed with a dataset of 32,313 transactions and 61 features which were clustered using Fuzzy C-means.

On the first part, the C-means algorithm was executed for 2 and then for 5 clusters with different values for the imprecision parameter, showing that the number of members increases as the value of the parameter is increased from 1.1 to 2.1.

The behaviour between two specific clusters increased significantly as the imprecision parameter changed from 1.1 to 1.9 as can be seen in table 15. This rapid change in behaviour of the clustering algorithm shows that choosing the correct imprecision parameter value is crucial for achieving better results.

Table 15: Cluster partition overlapping

Imprecision parameter	Number of overlapping	% of overlapping
m = 1.1	299	5.65
m = 1.5	3676	69.42
m = 1.9	4525	85.46

Imprecision parameter	Number of overlapping	% of overlapping
m = 1.1	518	9.78
m = 1.5	3164	59.75
m = 1.9	5187	97.96

Source: ISLAM et al. (2010)

Finally, the paper proposes an approach for handling fraud detection through the usage of Fuzzy C-means algorithm according to figure 12.

Figure 12: Proposal of fraud detection based on Fuzzy C-means

- Step 1:** *Classifies the data which fraud is very unlikely to occur (long contracts, regular vendors, recently audited areas)*
- Step 2:** *Collect the remaining data and label them as anomalies, considering that among these records exists a subset of records which can be categorized as possible frauds*
- Step 3:** *Collect profiles of fraudulent transactions, identify features for these fraud methods*
- Step 4:** *Use fuzzy C-means to collect transactions which are like the fraudulent transactions and store as possible fraudulent transactions*
- Step 5:** *Work through the transactions identified on step 4 with methods to identify fraud, for example Bayesian belief networks or naïve Bayesian to certify non-fraud transactions and report the remaining documents still classified as possible fraud for further actions.*

Source: ISLAM et al. (2010)

There was no implementation of a prototype for this approach and no performance metrics were provided in the paper.

Screening for bid-rigging – does it work?

In the paper of (IMHOF, D.; KARAGÖK, Y.; RUTZ, S., 2018) the screening method for detecting bid-rigging fraud in public procurement in Swiss is presented.

The study used the approach of using the test statistic marked so-called variance screen, which in short studies the variance of the final price versus the costs of the goods or services, which in a collusion, tends to very less responsive than in a competitive environment, or in another word, there is a certain degree of price rigidity.

The variance screen method is comprised of two main indicators: a) Quantity related markers: Studies the market share between companies involved in the contracts to identify patterns not related to competitive markets; b) Price related markers: Studies the prices and their variation over time and companies to identify symptoms of collusion.

This approach was proved to be able to detect cartels in the fuel market and baby products in Italy as well as construction and fish cartels in the United States among other successful cases.

The study received the list of all the contracts approved by one region of Swiss and only data related to road construction was selected for the case study.

The initial step was to clear the data from contracts which were not related to road construction (for example safety equipment and road signals) to ensure that only road construction contracts remained.

For this study quantity related markers were not used since the companies involved could be working in other areas than road construction as well as on other regions of Swiss, hence there is no reliable data for company market share, hence only price-related markers were used.

The first one was the coefficient of variation (CV_j), which is defined as standard deviation (σ_j) divided by the arithmetic mean (μ_j) of all bids for submitted for contract j :

$$CV_j = \frac{\sigma_j}{\mu_j} \quad (3.3)$$

As per the literature provided on the paper, low values of the coefficient of variation indicates price rigidity, which in turn relates to a suspect bidding behaviour. More specifically the variation of the coefficient over time can indicate periods of collusion, however for this specific study there is no variation over time and values were found to be non-conclusive, with only weak evidence that bids for invitation are more likely to have collusion than public bids.

The next test performed was the cover bidding screen, which checks the differences between the prices of the winning company versus the prices of the companies who lost the contract using the ratio between the difference in the two lowest bids ($\Delta_{j,l}$), and the standard deviation of the losing bids ($\sigma_{j,lb}$). This yields the following formula for the measure of relative distance (RD_j) as per below:

$$RD_j = \frac{\Delta_{j,l}}{\sigma_{j,lb}} \quad (3.4)$$

With an RD close to 1 there is no indication of behaviour between the winning bid and the remaining bidders, however, values much larger than 1 indicates that cover bidding might have occurred.

Again, there is no strong evidence for clear collusion, but the RD of public bids as 1.2 and the RD of invitation-only reaching 1.92, there is another evidence that the likelihood of collusion to happen on invitation only collusion.

Since the analysis of the population did not return any specific findings, the paper started a new screening for partial collusion, which is checking the interaction between companies

involved in the bids. This kind of analysis, however, does require to have substantial interactions between groups or subgroups of companies.

The approach had four steps:

Step 1: Isolate companies with suspicious behaviour by using the CV and RD values.

All contracts with $RD > 1$ and $CV < 0.06$ are considered suspect contracts. These figures were empirical verifications based on the analysis of contracts of companies identified as bid-rigging cartels in Swiss. This resulted in a total of 80 conspicuous contracts of which 80% were related to invitation-only bids. The same activity was performed with significantly more conservative parameters $RD > 1.3$ and $CV < 0.03$ which still resulted in many contracts. A third execution with values $RD > 1.15$ and $CV < 0.05$ was executed as well. The values were selected since they were the average between the proposed model and the most conservative one.

Step 2: Identify companies participating together with these suspicious companies on bids. To perform the activities, it was identified all companies who participated in at least 10% of these contracts to remove companies who bid in a sporadic way and are unlikely to be part of a cartel. In all the three scenarios there was the exact same 17 companies involved in the contracts selected from a total of 138 companies who participated in these contracts. In parallel, the companies who were bidding for the contracts were classified in a matrix to identify how many times they participated together in the same bid and the result is that 6 companies were selected as participating frequently on the same bids.

Step 3: This step checked the geographical distribution of the contracts, and when compared the contracts where these companies were operating. From a total of 8 different regions in Swiss, 2 regions were particularly suspicious since the companies were all bidding to the projects and the number of contracts each company won on each region was roughly the same.

Step 4: Create a graphical method to visualize the information of previous steps.

The conclusion according to the paper is that: “each of the suspect firms has, on an average and simultaneously with another suspect firm, submitted bids for roughly ten conspicuous contracts. Additional analysis shows that suspect firms exclusively submitted bids for fourteen contracts and that 91 per cent of all submitted bids came from the suspect group of firms. These results and figures point in the direction of a high degree of entanglement between the suspect firms.”

Prediction of Enterprise Purchases using Markov models in Procurement Analytics Applications

In the paper of (WESTERSKI, A. *et al.*, 2015) a model used to predict the purchase requests using Markov chains to identify future purchases and improve the efficiency of the procurement department.

The model receives the data related to the purchases of Singapore government from 2010 to 2013 and consists of a total of 141,286 purchase orders.

An initial analysis was performed on the POs to identify the requestor pattern for the information. It was identified that 59% of the purchase orders had only one item, meaning that that one PO was created to solve one specific problem at a time.

In addition, was identified that the descriptions of the purchase order for the same good had minor differences since the goods required had to be input manually instead of a standardized description for each item.

In order to solve this problem, the data was pre-processed using hierarchical clustering and calculated the similarity between the descriptions using the q-gram distance so similar descriptions could be grouped together.

After the pre-processing, the data was submitted to the algorithms below in order to predict the next future purchase and the prediction of multiple purchases.

Random Sampling: Used as a baseline for predicting the next purchase of a requestor. The simplest method, choosing the predicted values among the list of purchases from a user in the past in an aleatory way.

Probability Distribution: On this algorithm, first the Probability Density Function (PDF) is calculated for the orders of each requestor, then based on the PDF, the Cumulative Distribution Function (CDF) is calculated which serve as the basis for prediction of future orders.

Simple Sequential Sampling: Used as a baseline for predicting multiple purchases of a requestor.

Like random sampling, but for multiple values. On this algorithm, the data is organized as a sequence of values for each individual requestor. Next, an item of the sequence is randomly chosen, this item and the following ones are selected as the sequence of purchases for a requestor.

Markov Chain: On this algorithm, the prediction is based on a single previous state recorded by a requester. During the training phase of the algorithm, a matrix is built with the probability of moving between states, which are the unique purchase descriptions from the requestor history of the Markov chain.

Before the execution of the model two optimizations were done on the data:

Creation date optimization, where multiple purchases performed by the same requestor on the same day are treated as a single purchase with multiple items, this is performed to ensure the interval between orders is on the same degree of magnitude.

Frequency count optimization, which focuses on reducing the size of a chain for a specific requestor by eliminating all the transitions with a frequency below a specific threshold.

If all the transitions for the requestor are eliminated, the user is removed from the model since there is not enough data in order to perform a prediction

In order to construct the matrix, the data is analysed one requestor at a time, from the oldest requests based on the creation date to the newest one.

If one purchase description is followed by another purchase of the same description, the state does not change and hence no changes are made to the matrix.

The probability for a transition from a given state to another is calculated based on the number of times a given transition was observed in training data in relation to the total amount of time any transition from the original state has occurred.

To evaluate the results of the different algorithms and parameters, the results were based on the two metrics below:

$$precision = \frac{\text{count of correctly predicted feature values}}{\text{count of all unique predicted feature values}} \quad recall = \frac{\text{count of correctly predicted feature values}}{\text{count of all unique feature values in test subset}}$$

Based on the sequence of purchase orders to be predicted as 20 orders, the results can be verified in figure 16.

Table 16: Comparison of different algorithms on purchase prediction

Setup	Ignored Requesters (% of all requesters)	AVG Precision/Recall for Requester (Requester Count / % of Dataset orders / precision / recall)		
		Precision ≥ 0	Precision > 0	Precision > 0.5
0.5 train+ Markov+ 20 order set	98.16%	212 / 15.04% / 0.34 / 0.09	96 / 5.70% / 0.74 / 0.19	67 / 1.61% / 0.97 / 0.26
0.5 train+ CDF+ 20 order set	35.58%	6438 / 97.50% / 0.04 / 0.03	1319 / 49.41% / 0.20 / 0.17	120 / 0.88% / 0.95 / 0.80
0.5 train+ Random Sampling+ 20 order set	35.58%	6438 / 97.50% / 0.04 / 0.03	1258 / 46.28% / 0.19 / 0.17	110 / 0.58% / 0.94 / 0.84
0.5 train+ Sequence Prediction+ 20 order set	35.58%	6438 / 97.50% / 0.04 / 0.04	1367 / 40.10% / 0.17 / 0.19	98 / 0.54% / 0.93 / 0.91
0.5 train+ Markov+ 20 order set+ clustering	72.06%	2356 / 78.16% / 0.32 / 0.08	1235 / 52.47% / 0.61 / 0.15	600 / 12.88% / 0.96 / 0.24
0.5 train+ CDF+ 20 order set + clustering	37.39%	6214 / 97.05% / 0.15 / 0.13	3134 / 83.76% / 0.29 / 0.25	382 / 2.59% / 0.91 / 0.75
0.5 train+ Random Sampling+ 20 order set + clustering	37.39%	6214 / 97.05% / 0.00 / 0.00	14 / 1.05% / 0.01 / 0.02	0 / 0.00% / 0.00 / 0.00
0.5 train+ Sequence Prediction+ 20 order set + clustering	37.39%	6214 / 97.05% / 0.00 / 0.00	6 / 0.31% / 0.02 / 0.03	0 / 0.00% / 0.00 / 0.00

Source: WESTERSKI, A. et al. (2015)

The experimental results show that the association of Markov chains with clustering greatly increase the coverage of the population at the cost of a slight decrease in accuracy.

K-Means Algorithm: Fraud Detection Based on Signalling Data

In the paper of (MIN, X., LIN, R., 2018) a model to identify frauds on the telecommunication area is presented.

The model received the logs of 130,000 phone calls which were reduced to 26,670 after the consolidation of the phone numbers. There were 97 features associated with the data which was reduced to 67 after deleting the blank and duplicated features.

Due to the high number of features, Principal Component Analysis (PCA) was used to reduce the number of features from 67 to 6 features, which contained 92% of the original information. The number of features selected was based on the ratio of features/loss of information.

The data set was then submitted to Hopkins statistic test and returned a value of 0.998, suggesting a very high tendency of clustering.

The data were then clustered using the K-Means algorithm, with parameter K (number of clusters) between = 1 and 10. Then, the Sum of Squared Errors (SSE) was calculated for each value of K and using the elbow method, the value of K = 7 was chosen as the best value.

The clusters generated with K-Means algorithm resulted in one very large cluster which contained 75% of the data, 5 small clusters and one extremely small cluster. Details for the clusters are presented below and the overall distribution of records among the clusters can be found in table 4.

The features of the biggest cluster (cluster 0) were analysed, and it was found that the average call per number is 1.09 calls a day, with an average call time of 100 seconds, a very small variance between the call times, with high call success rate and talk time responsible to 73% of the total time call, all being characteristics of normal phone calls.

Cluster one and two had 4317 elements altogether with a very high number of calls, low talk time, low call success rate, low average duration with very high variation, suggesting a significant number of calls not completed and with a significantly higher number of target numbers, both clusters were classified as fraudulent phone calls.

Cluster three had only 48 elements with an average number of calls of 678.15 calls per day, with only 37% of the calls completed and a talk time ratio of only 0.2. This suggested that this cluster was related to an automatic calling machine.

Cluster four was like cluster zero, with slightly increase call times, number of calls and moderate call success ratio. This cluster was classified as inconclusive.

Cluster five was similar to cluster zero as well but with significantly longer call duration and talk times. It was classified as just normal long calls.

Cluster six was again similar to cluster zero, but included on fixed phone numbers, with remaining features very similar to cluster zero hence was classified again as normal calls.

The outcome of the model is that 3 clusters were categorized as normal calls, containing 82.8% of the calls, 1 cluster categorized as inconclusive with 0.84% of the calls and 4 clusters categorized as fraudulent with 16.37% of the calls, which is available on table 17.

Table 17: Summary of the clusters generated

Cluster	Classification	Records	Percentage
0	Normal	20126	75.46%
1	Fraudulent	449	1.68%
2	Fraudulent	3868	14.50%
3	Fraudulent	48	0.18%
4	Normal	223	0.84%
5	Unconclusive	661	2.48%
6	Normal	1295	4.86%

Source: MIN, X., LIN, R., THE, A. (2018)

Comparison of the related works

In this section, we compare the related works presented in the current chapter.

Each paper will be analysed based on 8 different topics:

Scalable: How the solution is capable of handling large data sets (above 1 million data records).

Real-time capable: Capability of the solution to process the information just created and reach a decision in a reduced amount of time.

Domain knowledge requirement: Level of domain knowledge of the user operating the model in order to understand the information provided and make a decision.

Adaptable: Capacity of the solution to handle data for which it was not previously prepared or configured, for example, on identifying new types of fraud.

Automated: Ability of the model to reach the final decision without human interaction.

Metrics available: Which metrics were used in order to reach a result comparison.

Gaps: The gaps identified in the model proposed.

Table 18 shows a comparison of the papers previously described in the related work section.

Table 18: Comparison of related works

The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data (DURTSCHI, C.; HILLISON, W.; PACINI, C., 2004)		
Item	Score	Details
Scalable	Yes	Very good scalability due to the simplicity of the algorithm
Real-time capable	Partial	Capable of identifying anomalies on data acquired in real-time, however, the data needs to be grouped until a minimal population is reached to perform the analysis which could lead to delays

Domain knowledge requirement	Yes	Benford analysis will only provide an indication of which digits are not following the expected distribution, hence manual analysis is required to identify a possible fraud
Adaptable	Yes	Since Bedford Law is purely based on statistical distribution, it can detect anomalies irrespectively if the type of fraud performed is unknown
Automated	No	The solution can be automated but in order to reach any meaningful decision, human interaction is required
Metrics available	No	No metrics available to the fraud detection were found
Core concept		Benford distribution algorithm
Gaps		Not a complete model for fraud detection on its own, but a tool that can be used as part of fraud detection solutions

A business process mining application for internal transaction fraud mitigation (JANS et al., 2011)		
Item	Score	Details
Scalable	No	Due to the requirement of a manual analysis of the documents process flow in order to reach a result, cannot be applied for a large volume of data
Real-time capable	No	Not suitable due to the requirement of manual process flow analysis
Domain knowledge requirement	Yes	Domain knowledge required to perform the analysis of the process flow and reach a decision
Adaptable	Yes	Since the model does not have any rule specified, it can identify unknown types of frauds
Automated	No	While part of the model can be executed without human interaction (log extraction and processing), manual analysis of the data is key in order to identify fraudulent documents
Metrics available	Yes	From the documents used on the data sample, it was found that - 0.77% breaking on price change rules - 2.5% not following the correct approval process - 0.21% not following the segregation of duty
Core concept		Log extraction and visual analysis

Gaps		<p>Rely on the sequence of activities to identify frauds, not considering any other data.</p> <p>The dependency of manual analysis and the accuracy of the results is directly related to the knowledge of the user operating the solution</p>
------	--	--

Fraud detection in ERP systems using Scenario matching (ISLAM et al., 2010)		
Item	Score	Details
Scalable	Partial	Although not having any manual intervention to reach the final results, the performance will decrease based on the number and complexity of the rules implemented
Real-time capable	Yes	Capable of processing data in real-time since every single document can be processed individually
Domain knowledge required	No	All the logic is implemented on the model, no domain knowledge is required to analyse the results
Adaptable	No	No adaptability due to rule-based model
Automated	Yes	Model is executed without any manual intervention
Metrics available	Yes	Random data was generated, and records were identified as per below: Change Vendor Bank: 2.9% PO Approval: 6.6% Pay Vendor: 12.3% Good Receipt: 19.8% Create Invoice: 6.7% Create Vendor: 20.6% Approve Invoice: 3.2 Create Customer: 6.6% Create PO: 15.5% Credit to Customer: 5.9%
Core concept		Rule-based model
Gaps		Detection rate highly dependent on fixed rules, making the model unable to identify frauds previously identified and configured

Reducing false positives in fraud detection: Combining the red flag approach with process mining (BAADER, G.; KRCMAR, H, 2018)		
Item	Score	Details
Scalable	Partial	<p>The performance will decrease based on the number and complexity of rules.</p> <p>The visual analysis included as part of the process will increase the execution time based on the number of documents flagged by the rules.</p>
Real-time capable	Partial	Capable of processing data in real-time only if a user is available to verify the data raised by the rules
Domain knowledge required	Partial	Part of the knowledge required to perform the analysis is included in the rules, but domain knowledge is still required to make a final decision during visual analysis

Adaptable	No	No adaptability since the first phase of the model is rule-based
Automated	No	User needs to check the data raised by the rules though visual analysis
Metrics available	Yes	True positives: 48.38% False positives: 0.37% True negatives: 51.61% False negatives: 99.99%
Core concept		Rule-based model associated with visual analysis
Gaps		Detection rate highly dependent on fixed rules, making the model unable to identify frauds previously identified and configured

Fuzzy C-Means for Fraud Detection in Large transaction data sets (CARLSSON C., HEIKKILA M., WANG X., 2018)		
Item	Score	Details
Scalable	Partial	There is no manual intervention to process the data, but the proposed model does not define the steps after the clustering to reach the final decision
Real-time capable	Partial	The clustering part of the algorithm can be executed in real-time, but since is not a complete model, may not be able to reach a final decision in real-time
Domain knowledge required	Partial	Not clear since the model is not complete, but since up to the clustering execution it does not require manual intervention, the result will be partial
Adaptable	Yes	The model is adaptable since there are no fixed rules
Automated	Partial	The model can be automated up to the clustering part
Metrics available	No	No metrics provided
Core concept		Fuzzy clustering
Gaps		An incomplete model with a prototype implementation or any metrics. It lacks the final part of the model to make the final decision if a transaction is fraudulent or not

Screening for bid-rigging – does it work? (IMHOF, D.; KARAGÖK, Y.; RUTZ, S., 2018)		
Item	Score	Details
Scalable	No	Model is based on statistical analysis on a phased approach. Due to the amount of manual analysis involved can be challenging to implement on large data sets
Real time capable	No	Since most activities are manual

Domain knowledge required	Yes	Very deep knowledge is required both on the domain of the data as well as on the statistical field
Adaptable	Yes	Since there are no rules and analysis are mostly manual
Automated	No	All the analysis and decisions are manual activities
Metrics available	Yes	From a dataset of 1491 bids, 282 tenders and 138 firms, six firms were identified with a high degree of entanglement, indicating the presence of a cartel
Core concept		Descriptive statistical analysis and correlation
Gaps		Strictly manual model based on individual specific statistical and domain knowledge

Prediction of Enterprise Purchases using Markov models in Procurement Analytics Applications (WESTERSKI, A. et al., 2015)		
Item	Score	Details
Scalable	Yes	Due to no manual intervention or specific configuration required as data increases
Real-time capable	Partial	Although not requiring manual intervention, the solution requires the process of all the historical data, leading to possible performance issues
Domain knowledge required	Yes	Domain knowledge is required since there is no information on the model on how to select fraudulently
Adaptable	Yes	No fixed rules included in the model
Automated	Yes	Model is executed without any manual intervention
Metrics available	Yes	The model was able to detect the description of the next purchase of the requestor with an accuracy of 78.16%
Core concept		Clustering and Markov chains
Gaps		Model is only able to predict what would be the next purchase of a given requestor, without any link indication if the same is fraudulent or not

K-Means: Fraud Detection Based on Signalling Data (MIN, X., LIN, R., 2018)		
Item	Score	Details

Scalable	Yes	Due to no manual intervention or specific configuration required as data increases
Real-time capable	No	Although not requiring manual intervention, only the clustering of the data can be automated, leaving the classification of the clusters to the manual interpretation of the user of the system
Domain knowledge required	Yes	Domain knowledge is required to analyse the clusters created and decide which are the clusters related to fraudulent calls
Adaptable	Yes	No fixed rules included in the model
Automated	No	Manual intervention is required to perform the analysis on the clusters generated
Metrics available	Yes	The model was able to detect 4 clusters with 16.37% of the phone calls analysed
Core concept		K-Means algorithm, PCA and Elbow method for identifying the ideal number of clusters
Gaps		The model can perform the clustering of the information in a relatively autonomous way but require significant analysis to understand the differences between the clusters and to identify the clusters as normal or fraudulent calls.

Source: Created by the author

APPENDIX B – SIGNATURE DETAILS

In this appendix, we are going to present the logic used to generate the signatures used to identify fraud symptoms on this methodology.

Invoice for undelivered goods/services

As a company policy, an invoice can only be paid after the goods or services purchased were received.

This signature verifies if the goods or services were received by the company and recorded on the ERP system before the payment to the supplier is performed.

Sequential invoice numbers

Checks the invoice numbers received from a supplier over a period of one year and subtracts the lowest invoice number from the highest invoice number and divide the quotient by the number of invoices received.

Based on the feedback from the audit team, suppliers with an average gap of at least 10 invoices between the invoices submitted is considered normal.

In the case of a supplier providing sequential invoices, this could be an indication of a ghost supplier which only invoices one company.

Changes performed on payment terms before a purchase

Payment term is the definition of when a supplier would be paid after the goods/services and the invoice related to a PO is received by the company.

Changes in the default payment term for reducing the term for a supplier right before a PO is created could be an indication of favouring a supplier in order to receive some advantage in the process.

This signature will return a binary flag if the payment terms for the supplier was changed up to 3 days before a PO is created.

The invoice amount is higher than the purchase order

On this check, the quantity of goods received is compared against the purchase order.

This signature will return the difference in percentage amount between the quantity order and the quantity on the invoice only for the cases when the quantity in the invoice is higher than the quantity on the procurement order to avoid false positives when a supplier does not have in stock all the quantity requested but can make a partial delivery, eg: when a company orders 100 boxes of paper but the supplier only have 80 to deliver at the required date.

Price and quantity above average

According to the details provided in chapter 4, the signatures will be generated based on the PO created over a period of one year, however, this specific signature will be generated based on the PO created one year prior to the period used in the model.

The objective is to identify any specific materials or suppliers with drastic increases in price over time.

This signature is calculated based on the averages for the material being procured, supplier and the combination of supplier and material procured.

Purchase Order is approved outside the standard workflow

Any type of PO should be approved before the request is performed to a supplier. In order to approve a PO, several different paths could be taken, however, the safest one, from the point of audit, would be the approval through the standard approval workflow.

The output of this signature will be a binary output of whether or not the PO was approved through the standard approval workflow.

Purchase order blocked

POs can be blocked at any given time due to several reasons, for example, due to quality problems of the goods delivered, legal issues with contracts, etc.

This signature will return a binary value of whether or not the PO was blocked from its creation until its closure.

Retrospective PO

Every good or service should only be procured from a supplier after a PO is approved.

This signature will verify if the date the invoice was created (which is different from the date when the invoice is received by the company) by the supplier happened before the approval of the PO.

These cases could indicate a supplier is being favoured or at least an internal policy which is not being followed.

Price increased after PO creation

This signature will check if the unit price of the procured good/service increased after the PO creation. The output of the signature is the difference in percentage between the initial unit price of the PO and the highest unit price.

In the case of a price decrease, the value would be zero.

Supplier blocked before a purchase

Like a PO, suppliers can be blocked for several reasons, quality of materials, ability to deliver on time, contract issues, etc.

Cases when a supplier is blocked, and a PO is created right after that could be considered suspicious for a possible extortion of the supplier.

The signature will return a binary output of whether or not the supplier was blocked/unblocked up to 3 days before a PO is created.

Vendor without phone or address available

According to literature reviewed, suppliers involved in corruption cases had frequently missing either phone numbers or address, hence this signature will return a binary output if a supplier does not have the address or phone numbers properly maintained.

Vendor with the same bank account as another vendor/ employee

This signature will check if the bank details (country, bank, branch and account number) are the same as another vendor or employee.

Cases flagged as having a positive value on this signature could be considered suspicious, however, there could be legitime cases normal, for example, the payments for goods ordered from one supplier location could be paid to the headquarters banking account due to a contract agreement.

The signature will return the number of other vendors or employees which have the same banking account as the vendor supplying the goods for the invoice currently being processed.

Difference between supplier creation and first sale

This signature will check the difference between the creation of a new supplier and the first PO created against the supplier.

Usually, there is a significant gap between creating a supplier on the ERP master data and the creation of the first PO from that supplier. This time is spent on checking the required documentation from the supplier and getting the necessary approvals.

Events of suppliers being created and POs raised right after should be considered suspicious.

Bank data changed for supplier

A supplier can change the bank they use to operate and hence, changes on supplier bank data can be considered a normal activity.

What should be considered though is that even being a normal activity, the frequency a supplier changes their bank details is very low, hence changes on supplier bank data should be considered suspicious activity and subject to investigation.

This signature only provides the number of times a supplier had changed their banking details over the period of one year.

Sudden business activity with old "sleeping" supplier (sudden activity in non-active accounts)

This signature checks which are the vendors which were previously actively trading with the company which is subject of this study stopped trading for a long period (in this case it was chosen a period of 2 years) and started trading again.

It will have a binary output of whether or not the vendor associated with the PO being processed fits into this pattern.

Difference between sales dates

It contains the average number of days between the POs created by the vendor over a period of one year.

The signature will check how many days have passed between the PO currently being processed and the previous PO from the same supplier.

ANEXO A ARTIGOS PUBLICADOS