# Machine Learning Methodologies Against Money Laundering in Non-Banking Correspondents

Jorge Guevara(✉), Olmer Garcia-Bedoya, and Oscar Granados

Universidad de Bogota Jorge Tadeo Lozano Carrera, 4 22-61 Bogota, Colombia
{jorgei.guevarap,olmer.garciab,oscarm.granadose}@utadeo.edu.co

**Abstract.** The activities of money laundering are a result of corruption, illegal activities, and organized crime that affect social dynamics and involved, directly and indirectly, several communities through different mechanisms to launder illegal money. In this article, we propose a machine learning approach to the analysis of suspicious activities in non-banking correspondents, a type of financial agent that develops some financial transactions for specific banking customers. This article uses several algorithms to identify anomalies in a transaction set of a non-banking correspondent during 2019 for an intermediary city in Colombia. Our results show that some methodologies are more appropriate than others for this case and facilitate to identify the anomalies and suspicious transactions in this kind of financial intermediary.

**Keywords:** Money laundering · Financial services · Machine learning · Non-banking correspondents

## 1 Introduction

Money laundering activities have used different mechanisms to launder their illegal money. This money tries to access the legal economy in almost all countries through goods trading, commodities exploitation, real estate transactions, financial transactions, or illegal activities as smuggling. In Colombia, as a global level, the exact amount of money laundering is impossible to estimate because the economic activities are susceptible to money launderers, that who use a dynamical structure. The literature has investigated this old problem with different methodologies. For Colombian case as a criminal and socio-economic approach [26], as a legal approach [27,28], as a political approach, as an economic approach [16], or as a result of drug production [3] to list a few. Other scholars proposed a multidisciplinary approach between artificial intelligence and network science [10] because money laundering is a result of corruption, financial crime, or drug trafficking, and other activities that needs a group of methodologies to fight a growing problem.

© Springer Nature Switzerland AG 2020
H. Florez and S. Misra (Eds.): ICAI 2020, CCIS 1277, pp. 72–88, 2020.
https://doi.org/10.1007/978-3-030-61702-8_6

For other countries, several scholars have implemented diverse machine learning methodologies. First, with a basic statistical model to identify the correlation between risk assessment and suspicious transactions [6]. Second, with the analysis of user's transactions to characterize them based on the behavior of all transactions in a specific dataset [1] or some patterns in suspicious financial transactions or money flow with transaction mining algorithms and frequent pattern mining algorithms [8,9]. Third, with the analysis of bank statements to find patterns that could resemble techniques used to the money laundering process [25]. Fourth, the classification of machine learning algorithms from anti-money laundering typologies, link analysis, behavioral modeling, risk scoring, anomaly detection, and geographic capability to identify the different features and mechanisms of money laundering [5]. Fifth, a combination of structural coupling theory with some data methodologies and algorithms to identify the information redundancy that could affect the money laundering control process [7]. Sixth, the Support Vector Machine (SVM) a machine learning method that trains a data set to identify the outliers [21].

On the other hand, 1.7 billion adults worldwide do not have a basic transaction account according to the World Bank's Universal Financial Access Initiative, the non-bank correspondents (NBC) are the first link for financial inclusion and to reduce poverty. Habitually, they could more close than the traditional financial service providers and could improve the trust in financial service providers trough everyday activities of the adult population. Thus, the non-bank correspondents as a strategy to create a new option for financial inclusion and to integrate customers to the formal financial system are non-financial establishments belong to everyday sectors whose principal activity involves cash. This mechanism facilitates the financial operations of customers in different developing countries, but the non-bank correspondents cannot identify suspicious transactions or implement an anti-money laundering scheme. Additionally, the non-banking correspondents have increased their operations in different regions like China, India, Bangladesh, Pakistan, Indonesia, Turkey, Congo, South Africa, Brazil, Colombia, Ecuador, Mexico, among others with some particular conditions as economic informality that facilitate to criminal organizations the money laundering through traditional or new schemes that they use in financial institutions. In consequence, this kind of correspondent is vulnerable to criminal organizations and illegal activities.

This situation has attracted the attention of scholars from a wide range of disciplines. However, the non-banking correspondent is a new phenomenon to study in the anti-money laundering techniques, and data science methodologies motivate us to create an effective approach. We propose an implementation of algorithmic analysis and visualization that facilitates identifying suspicious activities in non-banking correspondents. Our contribution is detecting unusual transactions in a non-banking correspondent using data analysis and machine learning techniques in a real dataset as well as, empirical rules and others have already known about money laundering. In this article, we apply some tools from machine learning and visualization techniques to identify suspicious transactions

in non-bank correspondents, a growing service in Colombia and other developing countries. We used unsupervised machine learning algorithms because test data does not have exit labels.

This article is divided as follows. In Sect. 2, we present some basic terminology employed in the text. Section 3 describes several features of the dataset structure and analysis. The following Sect. 4 presented the results of implemented methodologies and we close the paper with Sects. 5 and 6 that consist of concluding comments and provide directions for future work.

## 2 Preliminaries

For the reader convenience, in this section, we include some basic notions employed in this work and related to anti-money laundering and several characteristics of non-banking correspondents. First, we explain the SARLAFT normative and the irregular transactions that the financial sector must report to UIAF (Financial Information and Analysis Unit). Secondly, we present how the non-banking correspondents work.

### 2.1 Risk Management System for Money Laundering and Terrorism Financing (SARLAFT)

Two stages describe the risk management system for money laundering and terrorism financing:

- Risk Prevention. The objective is to prevent the inclusion of the resources from activities related to money laundering or financing of terrorism into the financial system(ML/FT). Thus, all clients should give some information to financial institutions with this purpose.
- Risk Control. Institutions like banks should report irregular transaction related to ML/FT, for that there must have a system that monitors and reduce the risk over these activities.

According to the Colombian Criminal Code of the 64 related crimes with money laundering, the financial transfer is one of the schemes in which financial institutions become vulnerable since they are the ideal agent for this purpose compromising their assets and their reputation.

### 2.2 Detection and Prevention Methods to Money Laundering and Terrorist Financing

Empirically, unusual transactions have been detected that could be related to ML/FT, although some are current transactions or transactions without due process. In this case, the unusual or suspicious transactions contemplated in the Colombian Criminal Code that could be made at NBC are:

- Withdrawal close to the limits defined in Resolution 285 of 2007 of the Financial Information and Analysis Unit - UIAF. Those amounts are not reported but should be analyzed in some way, for example, one or two transactions close to the limits.
- Withdrawals with the same debit card or with several debit cards at different moments on the same day.
- Reiterative transactions (ex: withdrawals, transfers, deposits) over financial products during a specific period.
- Withdrawals with different cards at the same time (users have a limit of transactions).
- Several debit transactions to the same account during a specific period.
- Balance inquiries repeatedly.

### 2.3 Non-Banking Correspondents

The non-bank correspondents (NBC) as a strategy to promote the financial inclusion and to integrate customers to the formal financial system are non-financial establishments belong to everyday sectors whose principal activity involves cash as gas stations, drugstores, retail stores, among others. This mechanism facilitates the financial operations of customers, develops new customer interactions in services and products, and creates new benefits through business commissions to the owners of those non-financial establishments.

### 2.4 Operations in a Non-Banking Correspondent

In the standard scheme, non-banking correspondents carry out several operations like cash in, cash out, and bill payments. In this case, the operations are:

◇ *Saving Deposit.* Deposit to a savings account.
◇ *Current Deposit.* Deposit to a current account.
◇ *Balance Inquiry.* The balance inquiry process associated with current accounts, saving accounts or financial products as credit cards.
◇ *Collection.* The collection process involves pursuing payments of debts or services that have been owed by individuals or companies. This process can be with automatic validation, i.e., the process uses a bar-code or QR code to validate the operation. In the other case, the process is semi-automatic because it needs a specific number as a payment reference that the cashier capture manually. Finally, the collection process involves payments of financial services as consumer credit, mortgages, and credit cards.
◇ *Withdrawal.* A retreat of money that has been deposited in an account previously. This account could be current, savings or special account.
◇ *Transfer.* A transfer is the movement of money from one account to another or others.

## 3 Materials

Here we explain some relevant information about the datasets, describing the raw data and some data mining processes required to obtain our samples.

### 3.1 Data Collection

A non-bank correspondent active since 2015 provided the data set used in this article. During 2019, at the end of every day, the non-bank correspondent collected the transaction receipts by date. Additionally, the digital data was downloaded monthly and stored by date. The owner of the non-bank correspondent provided all this evidence to integrate the datasets and initiate the analysis.

The dataset for this research was collected from a non-banking correspondent in Colombia during 2019. From the POS machine (provided by the financial institution) connected to the information system of the financial institution, the non-banking correspondent receives and sends information for each transaction. Each transaction issues a receipt, which is delivered to the transaction holder and the second one as back up for the transaction information that managed by the non-bank correspondent. The data stored in the receipt such as date, time, transaction type, and transaction number is information about each transaction, and this data linked to any of the financial institution's products like debit cards, credit cards, personal accounts, business accounts, or debt numbers.

To analyze and visualize the data and subsequently apply analytical and machine learning techniques, it was important to have the receipts data in a dataset. The financial institution has an information system so that each non-bank correspondent can obtain the operations information except the product number that affects the transaction, i.e. if the transaction is a deposit to a current or savings account, a deposit to a business account or other financial instrument does not record the product number. On the other hand, in the debit card withdrawals, the product number was omitted. However, those data found on the receipt.

To have the complete data in digital and to use adequately the database, it was necessary to integrate some data of receipts to the digital file. For this, we made several tests with the Google Vision API [11] through the OCR service (Optical Character Recognition) and programming code to capture the missing data employing a receipt image. Although the tool could detect the text, it did not have the accuracy to take the complete data due to some physical features of receipts as the thermally printed process. With those findings, we resorted to transcribing the receipt data into the database manually. To confirm the veracity of the process, we compared 100 random samples between transactions in the database and receipts with a result of 100% coincidence.

### 3.2 Data Analysis

The database has numeric and categorical columns, as well as a column with the transaction date and time. With this structure, the idea is to find unusual transactions that may be related to money laundering in any of the 64 crimes established in the Colombian criminal code.

Although the transaction amounts in a non-bank correspondent have limits and are low compared to transactions at banking institution branches, we

observed some particular patterns such as similar transactions, transactions carried out on specific days and times, or with a certain periodicity. The 80% of transactions are withdrawals and savings account deposits (Table 1), additionally, 72% of transactions are closed values, considering that those values usually do not belong to debt or service payments. We try to approach these transactions with behaviors that could become unusual.

**Table 1.** Relative frequency by transaction type

| Type | Code | Frequency |
|---|---|---|
| Withdrawal | WITHDRAWAL | 0.4286 |
| Savings account deposit | DEPOSIT_SAVINGS | 0.3734 |
| Collection | COLLECTION | 0.1111 |
| Current account deposit | CURRENT_DEPOSIT | 0.0643 |
| Credit card collection | CREDIT_CARD_PASS | 0.0096 |
| Transfer | TRANSFER | 0.0066 |
| Other collections | PURSE | 0.0064 |

Transactions have value limits, which change depending on the transaction type. Thus, the maximum value for withdrawal is ten Colombian million peso, for transfer is six Colombian million peso, and for other transactions is three Colombian million peso. In consequence, we observed that not consider this difference generates variables' disproportion problem. To resolve this, we divided the raw database considering the three transaction limits. The first dataset for withdrawals (WITHDRAWAL), second for transfers (TRANSFER), and third for other transactions (VARIOUS). From the value, we identified the asymmetric behavior of distributions for the withdrawal database (Fig. 1a) and various database (Fig. 1b), and also how some closed values widely exceed the mean.

Concerning categorical variables, we identify the characteristics count for each variable and especially as savings account product as the most relevant (Fig. 2a), as well as the deposit in this product is the transaction type most relevant (Fig. 2b) for the Various dataset. Meanwhile, the density distribution and various values outside the average ranges are characteristic of the categorical variable of the withdrawal dataset. Namely, the box-plot helped us to indicate whether a distribution skewed and whether there are potential unusual observations (outliers) in the withdrawal dataset (Fig. 2c). Additionally, we eliminated some variables because they did not provide relevant information, and we added other variables extracted from the relevant variables that could strengthen the database. For the withdrawal case, we used the time, and we divided into other variables such as a month, day of the month, day of the week, hour, and time bands, all to obtain other information from the dataset as the frequency by day to list a few (Fig. 2c). Summarize, the dataset consists of 64,000 records,

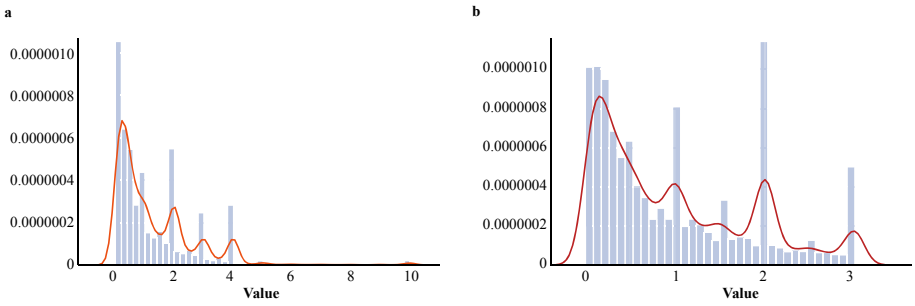


**Fig. 1.** Value histogram. a. WITHDRAWAL dataset. b. VARIOUS dataset.

and after preprocessing and cleaning, the dataset has 52,512 records and nine variables. We divided this raw dataset into three datasets as we explained above.

## 4 Unusual Transactions Detection with Machine Learning

This section presents visualization techniques, machine learning algorithms, and the modeling and testing layer of the CRISP-DM [31] methodology. We used unsupervised machine learning algorithms to detect unusual transactions taking account that the data does not have target labels.

Suspicious or unusual movements found in the data analysis are statistical. However, to improve the objective, we used machine learning algorithms to carry out the tests and take advantage of the computational resources, efficiency, and self-learning of those algorithms to detect, visualize, and predict these transactions. The algorithms used for testing are Isolation Forest based on decision trees [15] and One-Class SVM based on support vector machines (SVM) [17]. Of each one, we present a methodological structure, the importance and function of different parameters, and the way to validate them. Although K-Prototype based on clustering is not an algorithm to detect anomalies [13], but helps us to identify some behaviors through segmentation.

According to dataset characteristics, we used the grid search method through the GridSearchCV library of Sklearn in Phyton [19] to determine the optimal values of the hyper-parameters and validate algorithms in the Isolate Forest and One-Class SVM models. This method provides a tool to generate its scoring objects for each algorithm as a result of a score from the analyzed data, which classifies the unusual values as -1 and the normal ones as 1. Besides, with the estimator's function (*score_samples*), we can access the score used to compare the results of the two anomaly detection algorithms. Finally, as a complementary analysis with K-Prototype, we used the elbow method to determine the number of clusters in each dataset [4].

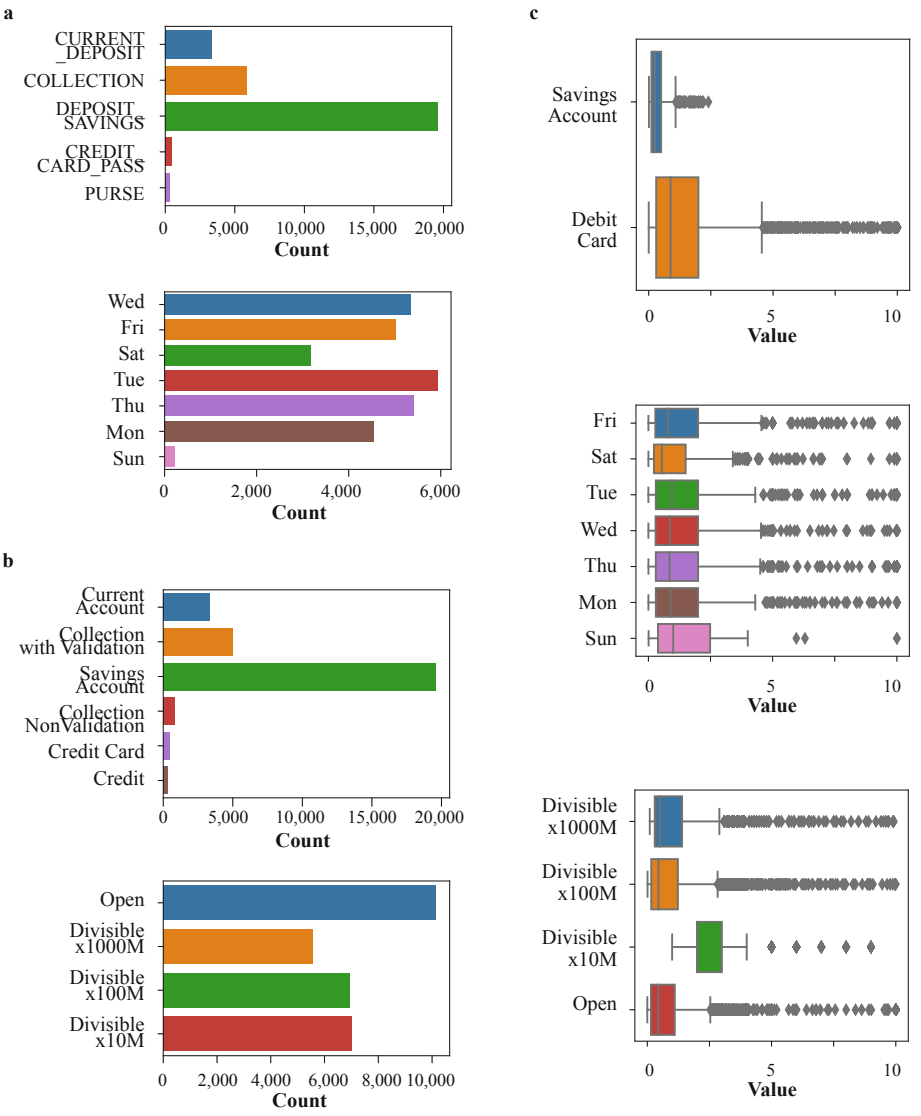


**Fig. 2.** Categorical variables. a. Transaction type in VARIOUS dataset. b. Product type in VARIOUS dataset. c. Boxplot (Box and whisker plot) in WITHDRAWAL dataset.

### 4.1 Isolation Forest

It is an unsupervised learning algorithm based on decision trees to detect anomaly in a dataset. Statistically, an anomaly is an observation or event that deviates significantly from other events to raise suspicions that a different mean generated it [15]. Habitually, algorithms define the profile of that is average data

and then isolate that is not, but Isolation Forest uses the opposite approach to detect anomalies, i.e., it isolates unusual data immediately. In consequence, anomaly detection is a process of two-stage. The first one builds isolation trees using sub-samples of the training set, i.e., trees are defined by recursive partitioning the training set until instances are isolated or trees reached a limit height $l$, around to the average tree height defined by the sub-sampling size [15]. Thus, the algorithm structure is:

**Input:** $X$ - input data, $T$ - number of trees, $\psi$- subsampling size
**Output:** a set of $T$ iTrees
**Initialize** `Forest`**:**
  set height limit $l$ = ceiling(log2 $\psi$);
  **for** *i = 1 to t* **do**
    $X' \leftarrow$ sample($X$, $\psi$)
    Forest $\leftarrow$ Forest $\cup$ iTree($X'$, 0, $l$)
  **end**
**return Forest**

**Algorithm 1.** iForest ($X$, $T$, $\psi$)

The second stage passes the test instances through isolation trees to obtain an anomaly score for each instance [15]. This score is a result of the expected path length for each test instance. Thus, this algorithm is more specific and requires less computational resources.

The evaluation of this algorithm for the WITHDRAWAL dataset (Fig. 3) presents the dispersion diagram, where indicates the anomalous transactions. We find that the transaction's value is not the only factor defining data features since there are anomalous and normal points throughout the value range.

### 4.2 One-Class SVM

The support vector machine (SVM) is a linear classifier based on the margin maximization principle and define the methodology of this algorithm [2,30]. The SVM accomplishes the classification task by constructing the hyperplane that optimally separates the data into two categories, which can be used for both classification and regression tasks and should receive a labeled training data set, defined as follows: $x_1, ...x_n \in \mathcal{X}$, where $n \in \mathbb{N}$ is the number of observations and $\mathcal{X}$ is some set [23]. In this case, the model trained with data of only one class, i.e., the positive information [18]. In other words, this algorithm tries to classify one class of objects and distinguish it from the other possible objects. However, it has to train to reject this object and to define it as an outlier [24]. Also, it infers the standard class properties and from these properties predicts which data is different [17].

To test the detection effect of the classifier based on the One-Class SVM algorithm, we use this methodology with the training and testing data for the WITHDRAWAL dataset respectively, and the detection result shows in Fig. 4. Newly,

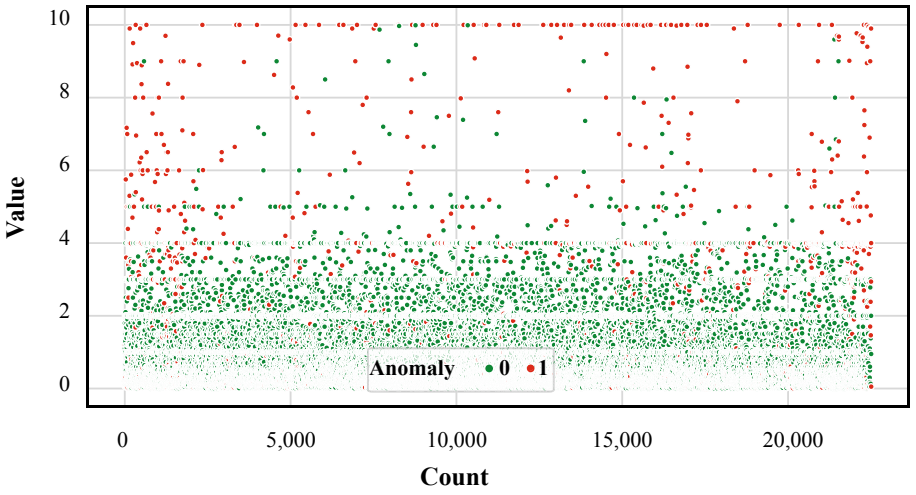


**Fig. 3.** iForest dispersion diagram for WITHDRAWAL dataset.

the transaction's value is not the only factor defining data features because there are anomalous and normal points throughout the range as well as in the previous algorithm.

### 4.3 Complementary Analysis

We compared the top ten transactions marked as anomalous by the two algorithms mentioned above (Table 2), and we find how values that we call closed have the highest frequencies in both cases. Also, we observed that values do not represent only the maximum or minimum values in each dataset because of algorithms included other additional variables to the transaction value. As a result, the One-class SVM algorithm shows a homogeneous classification of anomalous and normal values (Fig. 4) than the Isolation Forest algorithm (Fig. 3).

Additionally, as a technique for statistical analysis, we used K-prototype algorithm. This algorithm cluster objects with mixed, numeric, and categorical attributes in a way similar to k-means. Clustering consists of a set of objects that it with the greatest number of similar characteristics grouped, but in this case, objects clustered against k-prototypes instead of k-means, i.e., k-means, based on Euclidean distance, used for the numerical data and k-modes used for categorical variables [12]. However, when applied to numeric data the k-prototypes algorithm is similar to k-means (Table 3).

The purpose of this process was to demonstrate how numeric and categorical attributes interact with each other in the process of clustering and obtain other issues that complement the anomalies analysis. Figure 5 shows clustering in seven groups of the WITHDRAWAL dataset, defining by the elbow technique [4]. These seven clusters show some aspects that could help us to complement the

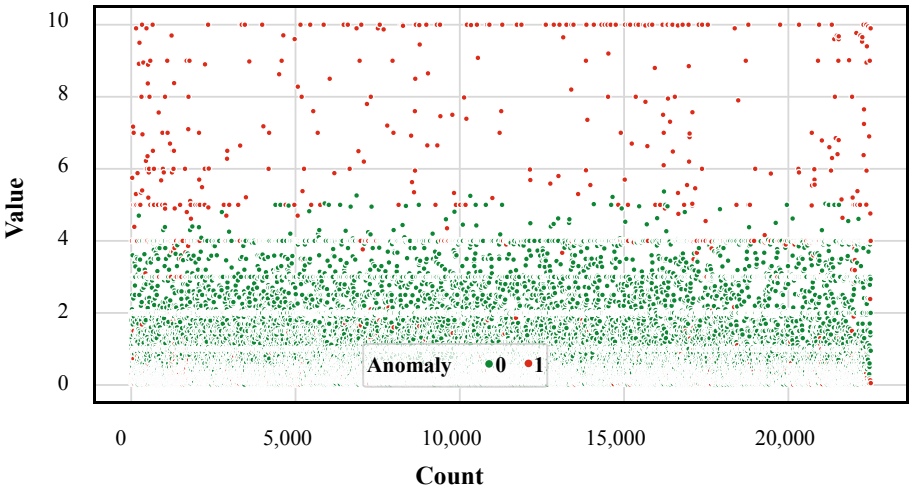


**Fig. 4.** One-Class SVM dispersion diagram for WITHDRAWAL dataset.

**Table 2.** Relative frequency of closed values for WITHDRAWAL dataset

| iForest | | One-Classs SVM | |
|---|---|---|---|
| Value | Frequency | Value | Frequency |
| 4,000,000 | 0.103234 | 4,000,000 | 0.091852 |
| 100,000 | 0.048737 | 9,999,999 | 0.036543 |
| 200,000 | 0.042534 | 1,000,000 | 0.028148 |
| 9,999,999 | 0.032787 | 3,000,000 | 0.021235 |
| 300,000 | 0.025698 | 100,000 | 0.020247 |
| 1,000,000 | 0.024369 | 200,000 | 0.019753 |
| 50,000 | 0.023039 | 2,000,000 | 0.015802 |
| 5,000,000 | 0.021267 | 50,000 | 0.014815 |
| 400,000 | 0.019938 | 5,000,000 | 0.013827 |
| 2,000,000 | 0.019495 | 150,000 | 0.013827 |

above results. We analyze some elements in cluster 0 to list a few. Several clients withdraw the maximum value of (9,999,999) as a strategy to evade the tax law control of the Financial Information and Analysis Unit (UIAF), which is from 10 million. This situation presented every month, weekday, and day that the non-banking correspondent operates and is a particular feature that needs monitoring. Additionally, withdrawals with specific values or closed values (75% values in cluster 0 are equal to four Colombian million peso) could be a signal of suspicious operations if they integrated with their frequency and temporality, like

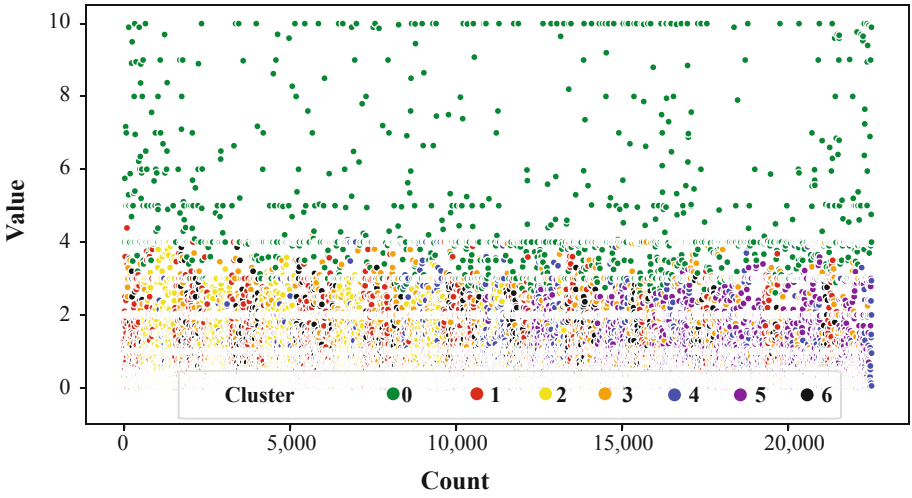


**Fig. 5.** Clustering for WITHDRAWAL database.

the case of the maximum values in other clusters. Namely, the cluster analysis is a powerful tool to complement the anomalies' analysis as long as characteristics from clusters can be identified and integrated because of each cluster covers a homogeneous space in the entire sample space.

Finally, we confirm a baseline approach with the statistical analysis. In Fig. 1a, we identified a long tail from five Colombian million pesos to ten million values on the x-axis, which indicates that those values skewed towards 10 million. Also, the first graph in Fig. 2c confirms that withdrawals with debit cards have unusual values starting at 5 Colombian million pesos. Therefore, if we compare the results of One-Class SVM in Fig. 4, we observed that the values on the y-axis after 5 million are unusual, which is confirming that seen above. Another example (1a) shows how some closed values like 1, 2, 3, and 4 million deviate from normal values. We could compare those values with Table two where the most common unusual values are 1, 2, 3, and 4 million.

## 5 Discussion

Anomaly detection algorithms can detect anomalous transactions directly and consistently with results of previous exploratory analyzes, while the clustering algorithm identified different behaviors in datasets. Although the detection of anomalies in bank transactions is not something new, if it is limited in the case of non-bank correspondents since beyond being a financial service in non-financial establishments, they have the particularity of developing in socio-economic, cultural, and particular geographical environments. These environments require the definition of policies and the design of technological tools that reduced the growing problem of money laundering. Likewise, algorithms allowed identifying

**Table 3.** Cluster statistical description for withdrawal database

| Measure | Value | Month | Day_month | Weekday | Time |
|---|---|---|---|---|---|
| Cluster 0. Count 2,226 | | | | | |
| min | 2,500,000 | 1 | 1 | 0 | 8 |
| 25% | 3,500,000 | 5 | 10 | 1 | 14 |
| 50% | 4,000,000 | 7 | 15 | 2 | 15 |
| 75% | 4,000,000 | 9 | 21 | 3 | 16 |
| max | 9,999,999 | 12 | 31 | 6 | 19 |
| **Cluster 1.** Count 3,768 | | | | | |
| min | 1,600,000 | 1 | 1 | 2 | 7 |
| 25% | 250,000 | 3 | 4 | 4 | 10 |
| 50% | 600.000 | 5 | 8 | 4 | 11 |
| 75% | 1,500,000 | 6 | 12 | 5 | 14 |
| max | 4,390,000 | 10 | 19 | 6 | 18 |
| **Cluster 2.** Count 3,203 | | | | | |
| min | 3,000 | 1 | 11 | 0 | 8 |
| 25% | 250,000 | 2 | 20 | 1 | 14 |
| 50% | 630,000 | 3 | 25 | 2 | 15 |
| 75% | 1,500,000 | 5 | 28 | 3 | 16 |
| max | 4,000,000 | 7 | 31 | 5 | 18 |
| **Cluster 3.** Count 4,016 | | | | | |
| min | 3,500 | 1 | 1 | 0 | 8 |
| 25% | 260,000 | 4 | 8 | 0 | 9 |
| 50% | 734,000 | 7 | 13 | 1 | 10 |
| 75% | 1,900,000 | 9 | 19 | 2 | 10 |
| max | 4,190,000 | 12 | 30 | 3 | 12 |
| **Cluster 4.** Count 2,634 | | | | | |
| min | 2,000 | 1 | 12 | 2 | 8 |
| 25% | 230,000 | 5 | 21 | 3 | 9 |
| 50% | 600,000 | 7 | 25 | 4 | 10 |
| 75% | 1,500,000 | 10 | 28 | 5 | 11 |
| max | 4,190,000 | 12 | 31 | 5 | 16 |
| **Cluster 5.** Count 3,496 | | | | | |
| min | 1,300 | 6 | 4 | 0 | 11 |
| 25% | 210,000 | 8 | 16 | 1 | 14 |
| 50% | 598,000 | 10 | 20 | 2 | 15 |
| 75% | 1,527,500 | 11 | 25 | 4 | 16 |
| max | 3,590,000 | 12 | 31 | 5 | 18 |
| **Cluster 6.** Count 3,166 | | | | | |
| min | 3,000 | 1 | 1 | 0 | 11 |
| 25% | 240,000 | 3 | 4 | 1 | 15 |
| 50% | 650,000 | 5 | 8 | 1 | 16 |
| 75% | 1,535,000 | 7 | 12 | 2 | 16 |
| max | 3,860,000 | 10 | 18 | 4 | 18 |

in greater detail the operations of several users, especially in the use of some financial products. In this case, we identified some user operations exceeded the maximum amounts defined by the Colombian Tax Code, as well as an unusual situation: the use of an account every 35 h during the period, something that is not usual in personal accounts (Fig. 6). The anomaly detection algorithms allowed us to improve results by adjusting hyper-parameters and available validations. However, for anomaly detection, we need to combine algorithm processes that robustness the tool and obtain detecting results more specific and for each cluster.

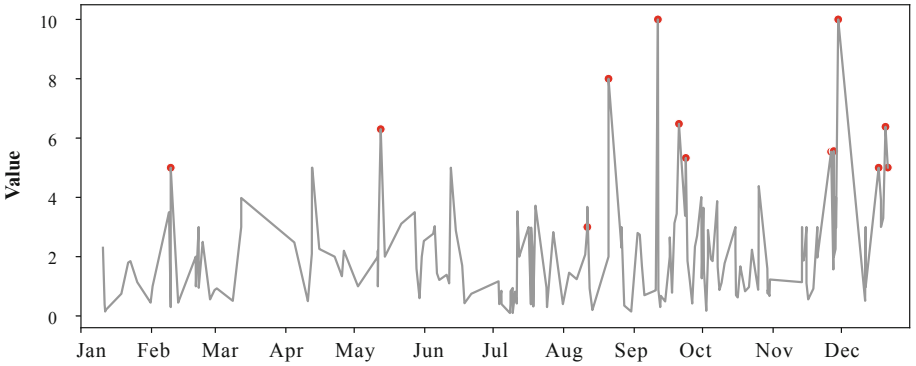


**Fig. 6.** Anomalies example of a specific user.

On the other hand, when identifying movements of a non-bank correspondent, a new problem arises that adds to money laundering and that is the robbery vulnerability [14,22]. They have this risk because of an increase in operations levels that makes them targets of common and organized crime. Some statistical metrics defined that non-banking correspondents could have operations on average between five to ten times greater than a store, a gas station, or a drugstore in a traditional operation day. Also, non-banking correspondents could have operations as a banking branch. In consequence, non-banking correspondents must have professional advice from financial institutions in cash management, money laundering, and other risks because of the criminal organizations that dedicate to money laundering and other illegal activities evolve permanently. Namely, those situations could affect the benefits and reduce the number of non-banking correspondents.

## 6 Conclusion

The main focus of this article is how machine learning and visualization algorithms serve to identify anomalies in a set of transactions of a non-banking

correspondent. This work sheds light on this problem by introducing some algorithms that can create a new perspective about financial transactions in a kind of financial agent that grows in different countries, especially, in developing countries as Bangladesh, Brazil, China, Colombia, Ecuador, India, Mexico, Pakistan, among others, as well as how is the money laundering activities in specific socioeconomic environments.

We identified some research topics to explore. First, the analysis of agents in big cities or some hot-spots in countries with a special growth of these kinds of financial intermediaries. Another approach for future research would be algorithms to identify anomalies using supervised and unsupervised machine learning techniques to confront the results, for example, Naive Bayes and Adaboosting algorithms to list a few. Third, according to the dataset type, we propose an implementation of a probabilistic graphical modeling technique (PGM) as a Bayesian network for unsupervised data that facilitate to model uncertainties by using Directed Acyclic Graphics [20,29]. Fourth, we could deepen to identify anomalies in each cluster using the result of segmentation.

In summary, our findings can help to define monitoring policies to develop preventive actions and reduce money laundering in non-bank correspondents, but it is important the constant support of financial institutions and expert personnel to counter a growing situation.

**Acknowledgements.** We have the support of administrative personnel and cashiers of some non-banking correspondents who contributed empirically to understand the business particularities.

## References

1. Rao, A.A., Kanchana, V.: Dynamic approach for detection of suspicious transactions in money laundering. Int. J. Eng. Technol. **7**(3.10), 10–13 (2018). https://doi.org/10.14419/ijet.v7i3.10.15619
2. Adankon, M.M., Cheriet, M.: Support vector machine. In: Li, S.Z., Jain, A. (eds.) Encyclopedia of Biometrics, pp. 1303–1308. Springer, Boston (2009). https://doi.org/10.1007/978-0-387-73003-5_299
3. Bayona-Rodríguez, H.: Money laundering in rural areas with illicit crops: empirical evidence for Colombia. Crime Law Soc. Change **72**(4), 387–417 (2019). https://doi.org/10.1007/s10611-019-09822-z
4. Bholowalia, P., Kumar, A.: EBK-means: a clustering technique based on elbow method and k-means in WSN. Int. J. Comput. Appl. 105(9) (2014). https://doi.org/10.5120/18405-9674
5. Chen, Z., Van Khoa, L.D., Teoh, E.N., Nazir, A., Karuppiah, E.K., Lam, K.S.: Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review. Knowl. Inf. Syst. **57**(2), 245–285 (2018). https://doi.org/10.1007/s10115-017-1144-z
6. Cindori, S., et al.: Money laundering: correlation between risk assessment and suspicious transactions. Financ. Theor. Pract. **37**(2), 181–206 (2013). https://doi.org/10.3326/fintp.37.2.3

7. Demetis, D.S.: Fighting money laundering with technology: a case study of bank x in the UK. Decis. Support Syst. **105**, 96–107 (2018). https://doi.org/10.1016/j.dss.2017.11.005
8. Drezewski, R., Dziuban, G., Hernik, L., Paczek, M.: Comparison of data mining techniques for money laundering detection system. In: Proceedings - 2015 International Conference on Science in Information Technology: Big Data Spectrum for Future Information Economy, ICSITech 2015, pp. 5–10 (2016). https://doi.org/10.1109/ICSITech.2015.7407767
9. Drężewski, R., Sepielak, J., Filipkowski, W.: System supporting money laundering detection. Digital Invest. **9**(1), 8–21 (2012). https://doi.org/10.1016/j.diin.2012.04.003
10. García-Bedoya, O., Granados, O., Cardozo, J.: Ai against money laundering networks: the Colombian case. J. Money Laundering Control (2020). https://doi.org/10.1108/JMLC-04-2020-0033
11. GoogleCloud: Vision AI. https://cloud.google.com/vision
12. Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Discovery **2**(3), 283–304 (1998). https://doi.org/10.1023/A:1009769707641
13. Ji, J., Bai, T., Zhou, C., Ma, C., Wang, Z.: An improved k-prototypes clustering algorithm for mixed numeric and categorical data. Neurocomputing **120**, 590–596 (2013). https://doi.org/10.1016/j.neucom.2013.04.011
14. Kumar, A., Parsons, A., Urdapilleta, E., Nair, A.: Expanding Bank Outreach through Retail Partnerships: Correspondent Banking in Brazil. The World Bank, Washington (2006)
15. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422. IEEE (2008). https://doi.org/10.1109/ICDM.2008.17
16. Loayza, N., Villa, E., Misas, M.: Illicit activity and money laundering from an economic growth perspective: A model and an application to Colombia. J. Econ. Behav. Organ. **159**, 442–487 (2019). https://doi.org/10.1016/j.jebo.2017.10.002
17. Manevitz, L.M., Yousef, M.: One-class SVMs for document classification. J. Mach. Learn. Res. **2**, 139–154 (2001)
18. Muller, K., Mika, S., Ratsch, G., Tsuda, K., Scholkopf, B.: An introduction to kernel-based learning algorithms. IEEE Trans. Neural Networks **12**(2), 181–201 (2001). https://doi.org/10.1109/72.914517
19. Pedregosa, F., et al.: Scikit-learn: Machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
20. Pham, D.T., Ruz, G.A.: Unsupervised training of Bayesian networks for data clustering. Proc. Roy. Soc. Math. Phys. Eng. Sci. **465**(2109), 2927–2948 (2009). https://doi.org/10.1098/rspa.2009.0065
21. Prakash, A., Apoorva, S., Amulya, K.H., Kavya, T.P., Prashanth Kumar, K.N.: Proposal of expert system to predict financial frauds using data mining. In: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1080–1083, March 2019. https://doi.org/10.1109/ICCMC.2019.8819709
22. Sánchez-González, C., Prada-Araque, D., Erazo-Inca, F.: El aporte de los corresponsales no bancarios a la inclusión financiera. Desarrollo Gerencial **12**(1), 1–23 (2020). 10.17081/dege.12.1.3599
23. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. **13**, 1443–1471 (2001). https://doi.org/10.1162/089976601750264965

24. Shin, H.J., Eom, D.H., Kim, S.S.: One-class support vector machines–an application in machine fault detection and classification. Comput. Ind. Eng. **48**(2), 395–408 (2005). https://doi.org/10.1016/j.cie.2005.01.009
25. Singh, K., Best, P.: Anti-money laundering: using data visualization to identify suspicious activity. Int. J. Acc. Inf. Syst. **34**, 100418 (2019). https://doi.org/10.1016/j.accinf.2019.06.001
26. Thoumi, F.E.: Political Economy and Illegal Drugs in Colombia. Lynne Rienner Publishers, Boulder (1995)
27. Thoumi, F.E., Anzola, M.: Asset and money laundering in Bolivia, Colombia and Peru: a legal transplant in vulnerable environments? Crime Law Soc. Change **53**(5), 437–455 (2010). https://doi.org/10.1007/s10611-010-9235-8
28. Thoumi, F.E., Anzola, M.: Can AML policies succeed in Colombia? Crime, Law and Soc. Change **57**(1), 1–14 (2012). https://doi.org/10.1007/s10611-011-9331-4
29. Thulasiraman, K., Swamy, M.: 5.7 acyclic directed graphs. In: Graphs: Theory and Algorithms, p. 118. Wiley, New York (1992)
30. Vishwanathan, S., Murty, M.N.: SSVM: a simple SVM algorithm. In: Proceedings of the 2002 International Joint Conference on Neural Networks, IJCNN 2002 (Cat. No. 02CH37290), vol. 3, pp. 2393–2398. IEEE (2002)
31. Wirth, R., Hipp, J.: CRISP-DM: towards a standard process model for data mining. In: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, pp. 29–39. Springer, London (2000)