Deep Learning-based Synthetic Data for Money Laundering Control Simulations

Edwin González[a], Olmer García-Bedoya[a] and Oscar M. Granados[b,*]

[a]*Department of Industries and Digital Technologies, Universidad Jorge Tadeo Lozano, Bogotá, Colombia*
[b]*Department of Economics and International Trade, Universidad Jorge Tadeo Lozano, Bogotá, Colombia*

ARTICLE INFO

Keywords:
ABM
GAN Networks
Synthetic Data
Money Laundering Control

ABSTRACT

The frequency of money laundering cases detected in different jurisdictions has generated a wide range of policies with limited results. We propose a deep learning strategy to create synthetic data that facilitates the simulation of money laundering schemes using an agent-based model to address this issue. We present a GAN networks methodology that makes synthetic data statistically significant to simulate money laundering situations in a network of non-banking correspondents. A set of procedures is built based on special rules to develop the simulation-based in traditional behaviors of suspicious agents. Those procedures identify features that support the method's robustness to consolidate some recommendations against money laundering because the data is the big problem to fight against this phenomenon.

1. Introduction

Money laundering is transforming money from illegal activities into legitimate money. This mechanism is possible because some agents lend their names to develop different asset investments, open financial institution accounts, or create shell companies. However, financial operations continue to stand out as one of the principal mechanisms. This process developed in various stages in which the first corresponds to the entry of money into the financial system through deposits or transfers made by many indeterminate agents in one or different places. Subsequently, after the money is deposited in several accounts, it is consolidated into an account. Those operations can be carried out through local and international financial institutions in different jurisdictions. In local operations, they can be operated in financial institutions, non-bank correspondents, or money transmitters, which have gained participation in several emerging markets such as Bangladesh, Colombia, Mexico, to list a few.

Regulations against money laundering play a fundamental role in the evolution of financial systems and the different service alternatives because of emerging technologies that have been implemented to give financial institutions greater strength and effectiveness. However, access to data to develop solutions against money laundering has become one of the reasons why options that arise from machine learning and deep learning cannot be easily deployed. It is for this reason that synthetic data is crucial. The literature has not jointly analyzed the use of synthetic data and money laundering. However, there have been several approaches to the two themes independently.

Against money laundering, economic methodologies with econometric, microeconomic, and game theory models have been used (Walker, 1999; Walker and Unger, 2009; Schneider and Windischbauer, 2008; Schneider, 2010; Tan et al., 2019; Imanpour et al., 2019; Ardizzi et al., 2018; Loayza et al., 2019; McCarthy et al., 2015). Corruption has also been modulated from complex systems, data science, and network science as a source crime of money laundering (Ribeiro et al., 2018; Luna-Pla and Nicolás-Carlock, 2020; Granados and Vargas, 2021), the use of social networks and profiling in the behavior of money laundering (Dreżewski et al., 2015; Demetis, 2018), as well as the combination of methodologies such as synthetic intelligence and network science to identify money laundering activities from different source crimes (Garcia-Bedoya et al., 2020; Guevara et al., 2020). In the case of synthetic data, some recent work has focused on using some deep learning methodologies to create synthetic data, especially for medical data, a sector that has advanced considerably in the use of LSTM and GAN networks (Yang et al., 2019; Beaulieu-Jones et al., 2019; Sandfort et al., 2019; Yoon et al., 2019) to list a few. However, those methodologies can be used in different topics such as facial recognition, image construction, and prediction.

[*] Corresponding author
ORCID(s): 0000-0002-4992-8972 (O.M. Granados)

The adversary generative adversary neural networks (GAN) become an essential tool to optimize a learning and development task for money laundering control. A supervised learning high-level algorithm GAN for the generation of synthetic data and integrated with multivariate statistical methods allow hypothetically contrasting the equality of averages and homogeneity of variances between real and synthetic data (Goodfellow et al., 2014). These multivariate statistical methods allow the new data generated by the GAN network to be of high quality. Additionally, they enable defining data to implement large-scale analyzes in which the objective is to control money laundering (González-Martínez, 2021). In addition, synthetic data begin from the access to historical data of reliable sources, where it is possible to train a GAN Network to learn from the data. As well as mapping unusual patterns and characteristics to extract their metrics, analyze them, and generate new information to develop new scientific proposals and methodologies.

In this way, the purpose of this paper is to generate statistically significant and consistent data, where the probabilistic benefits and the parameters for the generation of new information that facilitate the simulation of the behaviors of the agents involved in money laundering are determined. Simulations help build public and private policy schemes that could control a phenomenon that continues to grow worldwide. Society needs to implement new methodologies in the face of the increasing sophistication of criminal groups.

The remainder of this paper is organized as follows: In Section 2, we explain the construction of the synthetic data, where first, we described the original data structure, second, we present some aspects about synthetic data, and third, we integrated the deep learning methodology. In Section 3, we incorporated a basic agent-based model to simulate the agent transactions and identify patterns of suspicious activities in the context of money laundering control. Section 4 presents the results of our simulations using different model rules. Section 5 contains the conclusions of our work.

2. Synthetic Data Construction

2.1. Original Dataset

The dataset refers to the transactions carried out by a non-bank correspondent in a city in Colombia during 2019. We anonymized the data for confidentiality. There are 67,244 records with seven variables: agent identification, transaction value, transaction type, transaction status, date, data-phone type, and approval or rejection transaction code. For data preprocessing, we had several criteria into account about the quality of the records. The data eliminated was records with transactions value with zero, and for the type of transaction, we used only withdrawal, deposit, and transfer. In addition, the month, day, and time data was extracted from the date field, thus creating three additional variables to become factors in the analysis of the transactions. For the final dataset, we have 55,612 records with six variables: agent identification, transaction value, transaction type (deposit, withdrawal, or transfer), hour, day, and month. These variables experience a normalization transformation taking into account each variable's min and max value. The mean and variance parameters are saved for each variable, and the transactional type variable (categorical type) is transformed with one hot encoder where a binary variable represents each deposit, withdrawal, and transfer field.

2.2. Synthetic Data

The information access for the analysis of financial crimes is difficult because transactional and financial information can contain sensitive data (González Martínez, 2020). To analyze the growing number of financial crimes requiring alternatives and synthetic information becomes an option. In turn, the generation of synthetic data is a methodological alternative to label and protect sensitive third-party including financial institutions (Surendra and Mohan, 2017). Therefore, synthetic data allows generating new information and, in turn, identifying relationships and attributes. Synthetic data is classified into three categories. First, completely synthetic data. In this case, a generator constructed data for each field or variable with probabilistic statistical parameters. Second, partially synthetic data. For this process, statistical imputation, scale transformation, and dimension reduction transform the information. Third, hybrid synthetic data contains real and synthetic information from both data sources.

Early research developed a synthetic information generator for public and private sectors research. The proposal consisted of multiple imputations, i.e., conjugating and simulating with probabilistic sampling techniques called Bootstrap and carrying out simulations to sets of samples with or without replacement to estimate the value(Raghunathan et al., 2003). The estimated value was listed as the synthetic value with no record. Furthermore, another proposal was to find highly correlated auxiliary variables to replace the target variable. Subsequently, other techniques were developed to create synthetic values through multiple imputations using point estimates, supervised and unsupervised

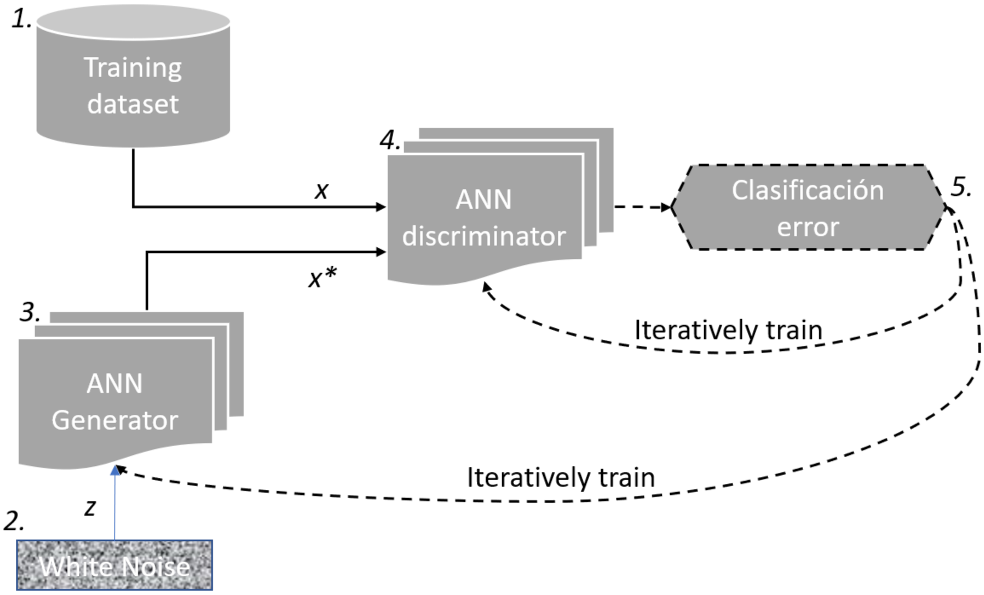

Figure 1: *GAN Architecture.*

learning models, and deep learning (Raghunathan et al., 2003). The following section explores a deep learning model to create synthetic data.

2.3. Generative Adversarial Network (GAN)

The generative adversarial network GAN consists of creating two synthetic neural networks(ANN). The first network (generator) is designed with a number of hidden layers of the form $(32, 64, 128, 256)$ with activation function *relu*. The second network (discriminator) is designed with a number of hidden layers of the form $(128, 64, 32, 1)$ with activation function *relu*. Also, learning the (GAN) takes a lot size of 500 with 100 epochs. This type of architecture allows the generative neural network to be more competitive and robust. However, the generative network can learn from the data and validate its performance to generate good quality synthetic data.

GAN networks are deep learning networks that aim that the two models of ANN compete with each other. The first model (generator), (g), is trained to generate false data, and the second model (discriminator), (d), is trained to distinguish the false data from the real ones. In GAN, there is a type of zero-sum cooperative game in which two or more players try to optimize their best play to beat their adversarial. This context is focused on the two neural networks so that they have a highly efficient level of learning given the convergence of the cost loss function between them. The generator learns through feedback received from the classifications of the discriminator (Langr and Bok, 2019). The discriminator aims to determine if the data is real (from the training data set) or false(created by the white noise generator). Figure 1 presents the learning process (Langr and Bok, 2019).

In this way, we can represent the scenario where the game competition function $(\min; \max)$ is expressed between the two networks with their goodness parameters of each one of them $v\left(\theta^{(g)}; \theta^{(d)}\right)$ (Equation 1). That is, during learning each player tries to maximize his own reward, so that the convergence is given by:

$$g^* = arg \min_g \max_d v\left(g, d\right) \tag{1}$$

Then, the function is defined by $f\left(x, v\left(g, d\right)\right)$:

$$v\left(\theta^{(g)}, \theta^{(d)}\right) = \mathbb{E}_{x\sim pdata}(x)\left[\log d(x)\right] + \mathbb{E}_{z\sim pz}(z)\left[\log\left(1 - d\left(g\left(z\right)\right)\right)\right] \tag{2}$$

Where: (x) represents the real training data, $p_{data}(x)$ of their distribution, (z) is noise with which the generator is fed to synthesize the data, and its distribution $p_z(z)$ which is determined as the distribution of the generated data $p_g(z)$. The first neural network called the *Generator* (g), consists of generating samples given at $x = g\left(z;\theta^{(g)}\right)$. Its adversary, the second neural network named *Discriminator* (d), tries to distinguish between samples drawn from the training data and samples drawn from the generator. The discriminator outputs a probability value given by $d\left(x;\theta^{(d)}\right)$, in it calculates or estimates a probability of x that a sample comes from the training data in instead of a fake sample taken from the (g) (Goodfellow et al., 2016) model.

Then, learning occurs through a zero-sum game, in which a function given by $v(\theta^{(g)},\theta^{(d)})$ determines the reward of the discriminator (Goodfellow et al., 2016). Also, the generator receives $-v(\theta^{(g)},\theta^{(d)})$ as its own reward, so the discriminator learn to classify the samples as real or false correctly. Simultaneously, the generator tries to trick the discriminator into believing that its samples are real. In addition, the convergence of the generator samples is indistinguishable from the real data, which means that the discriminator must have a high level of learning with a minimum error rate and maximize the probabilities for the discriminator to make estimates of high performance given at learning task as represented by equation (2). Additionally, the generator (g) implicitly defines a probability distribution p_g as the distribution of the samples $g(z)$ obtained when $z \sim p_z$ (Goodfellow et al., 2014). Therefore, it is sought that Algorithm 1 converges in a good unbiased estimator for p_{data}. This algorithm is based on the work of (Goodfellow et al., 2014)

Algorithm 1 The stochastic gradient descent method for the GAN_s, given a *m-th* sample size for a number of steps for the k discriminator. Where k is a hyperparameter.

```
For number of training iterations do
  for $k$ steps do
    • Sample minibatch of $m$ samples $\{z^{(1)},\ldots,z^{(m)}\}$ from noise prior $p_g(z)$
    • Sample minibatch of $m$ examples $\{x^{(1)},\ldots,x^{(m)}\}$ from data generating distribution $p_{data}(x)$
    • Update the discriminator by ascending its stochastic gradient:
      $\nabla_{\theta d}\frac{1}{m}\sum_{i=1}^{m}\left[logd\left(x^{(i)}\right)+log\left(1-d\left(g\left(z^{(i)}\right)\right)\right)\right]$
end for
    • Sample minibatch of $m$ noise samples $\{z^{(1)},\ldots,z^{(m)}\}$ from noise prior $p_g(z)$
    • Update the generator by descending its stochastic gradient:
      $\nabla_{\theta g}\frac{1}{m}\sum_{i=1}^{m}log\left(1-d\left(g\left(z^{(i)}\right)\right)\right)$
end for
```

In the above algorithm simulation, the best-performing mathematical formulation of the GAN game is a different formulation that is neither zero-sum nor the equivalent of maximum probability, i.e., it is a heuristic motivation. Learning rate improves the performance of (g) and (d). So the generator aims to increase the log probability that the discriminator will make an error, rather than decreasing the log probability that the discriminator will make a meaningful prediction. The simulation makes the derivative of the generator cost function concerning the $logits$ of the discriminator remain large even in the situation where the discriminator confidently rejects all samples from the generator (Goodfellow et al., 2016).

Additionally, the global optimization is given by $p_g = p_{data}$, where it is considered that the optimal discriminator (d) for any generator (g) (Goodfellow et al., 2014). Where (g) is fixed and the optimal discriminator (d) is given by:

$$d_g^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \tag{3}$$

The training criterion for the discriminator (d), given any generator (g), is to maximize the quantity $v\left(\theta^{(g)}, \theta^{(d(x))}\right)$.

$$\begin{aligned} v\left(\theta^{(g)}, \theta^{(d(x))}\right) &= \int_x p_{data}(x) \log(d(x))\,\mathrm{d}x + \int_z p_z(z) \log(1 - d(g(z)))\,\mathrm{d}z \\ &= \int_x p_{data}(x) \log(d(x)) + p_g(x) \log(1 - d(x))\,\mathrm{d}x \end{aligned} \tag{4}$$

For any $(a, b) \in \mathbb{R}^2$ given that $\{0, 0\}$, the function $y \rightarrow a\log(y) + b\log(1 - y)$ reaches max in $[0, 1]$ for $\frac{a}{a+b}$, where the discriminator need not be defined outside of $Supp\left(p_{data}\right) \cup Supp\left(p_g\right)$.

The training objective for (d), can be interpreted as the maximization of the logarithmic probability for estimating the conditional probability given by $p(Y = y \mid x)$, where Y indicates if (x) comes from the distribution p_{data} when $p(y = 1)$, otherwise, p_g is given when $p(y = 0)$. Where the adversary game (min; max) the equation is reformulated (2):

If $C(g) = \max_{(d)} v(g, d)$ then,

$$\begin{aligned} &= \mathbb{E}_{x\sim pdata}\left[\log d_g^*(x)\right] + \mathbb{E}_{z\sim pz}\left[\log\left(1 - d_g^*(g(z))\right)\right] \\ &= \mathbb{E}_{x\sim pdata}\left[\log d_g^*(x)\right] + \mathbb{E}_{x\sim pg}\left[\log\left(1 - d_g^*(x)\right)\right] \\ &= \mathbb{E}_{x\sim pdata}\left[\log \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}\right] + \mathbb{E}_{x\sim pg}\left[\log \frac{p_g(x)}{p_{data}(x) + p_g(x)}\right] \end{aligned} \tag{5}$$

The convergence of algorithm is given if (g) and (d) have enough capacity in each step of algorithm, then, the (g) allows reaching its optimum given (g), and p_g is updated to improve the criterion

$$\mathbb{E}_{x\sim pdata}\left[\log d_g^*(x)\right] + \mathbb{E}_{x\sim pg}\left[\log\left(1 - d_g^*(x)\right)\right] \tag{6}$$

then p_g converge to p_{data}

The GAN represent a limited family of p_g distributions through the function $v\left(z, \theta_g\right)$, where optimizes θ_g instead of the probabilities ofp_g, given that the network (d) has the ability to recognize false data from the generator and accept the actual data from the data set.

3. Basic Agent-based Model and Money Laundering Control

Simulation models that describe agent behavior have become a widely used tool in diverse socioeconomic issues, as well as in general social science research (Epstein and Axtell, 1996; Gilbert and Troitzsch, 2005) and specific issues of political science (Axelrod, 1997; Huckfeldt et al., 2004) or economics (Tesfatsion, 2002), to list a few. As well as in the ecological models (Grimm et al., 2006), the agent dynamics in socioeconomic contexts are driven by conditions that change over space and time. Namely, the agent behavior adapts to the changing context.

However, ABM methods have several approaches to consolidate the simulation. From expert systems to multiagent systems could help build a simulation framework for money laundering control. Next, we identify some elements of those methods to make a methodological proposal that we develop in this paper. First, we start from an intelligent system, specifically an expert system. This system is a kind of knowledge-based system designed to embody expertise in a particular specialized domain (Hopgood, 2001). We included a rule-based system, another knowledge-based system, where the base is represented in the form of a set of rules.

In this case, we used a kind of normative multiagent system since this allows a social order to be established as a pattern of interactions among agents, allowing the satisfaction of the interests of an agent (Castelfranchi, 2000). These interests can be a delegated goal, i.e., a good value for all agents or most of the members. In this case, the interest is based on the fact that the rules allow compliance with the parameters that non-bank correspondents have to make their activities. The agents attribute the goal to the normative system because some agents have socially delegated goals to the normative system. These objectives are the content of the obligations that regulate it (Boella and van der Torre, 2006).

To develop simulations in an agent-based model for a non-bank correspondent need the environment's definition, business rules, and the identification of suspicious transactions. Those components are defined by agents interacting with non-banking correspondents used through different services. Second, properties are factors that describe the non-banking correspondents such as deposits, withdrawals, transfers, time, day, and month. Third, the synthetic data is given for twenty branches environment. Fourth, the rules are defined to identify risk factors and patterns.

However, one of the main rules by the financial institution is that every transaction carried out is subject to a schedule from 7 am to 7 pm. The deposits range is between one to three million Colombian pesos, and withdrawals and transfers are between one to ten million Colombian pesos. For this amount, the agent has to fill a tax form. Additionally, we defined six rules for the agents' identification and suspicious transactions.

- Rule 1. Transaction Frequency by agent carried out during the year 2019.
- Rule 2. Transaction Frequency by agent and month.
- Rule 3. Transaction Frequency by agent, branch, and month.
- Rule 4. Transaction Frequency by agent, branch, month, day, and transaction type.
- Rule 5. Transaction Frequency by agent, branch, month, time, and transaction type.

Each rule has a greater degree of complexity that allows strengthening the identification of suspicious money laundering transactions.

4. Results

Based on the trained and optimized GAN model, the transformed data samples were taken from the non-bank correspondent where the predictions definition was by value, day, hour, and month. Furthermore, the predicted data is transformed to the original scale by taking the initial parameters of mean (μ) and variance (σ^2). The twenty synthetic datasets generated correspond to each non-banking correspondent. Also, all consolidated data are 958,225 records with seven variables, where each variable describes the identification number of the agent, branch, transaction value, month, day, time, and type of transaction. The synthetic information generated preserves the statistical benefits of the non-bank correspondent and his rules.

Additionally, the agent-based model allows identifying factors given to interactions between agents and their environment. First, the heat maps present behaviors, associations, correlations, and unusual patterns through statistical visual analysis. Second, the agent's behavior analysis shows the interaction processes with non-bank correspondents through a network visualization.

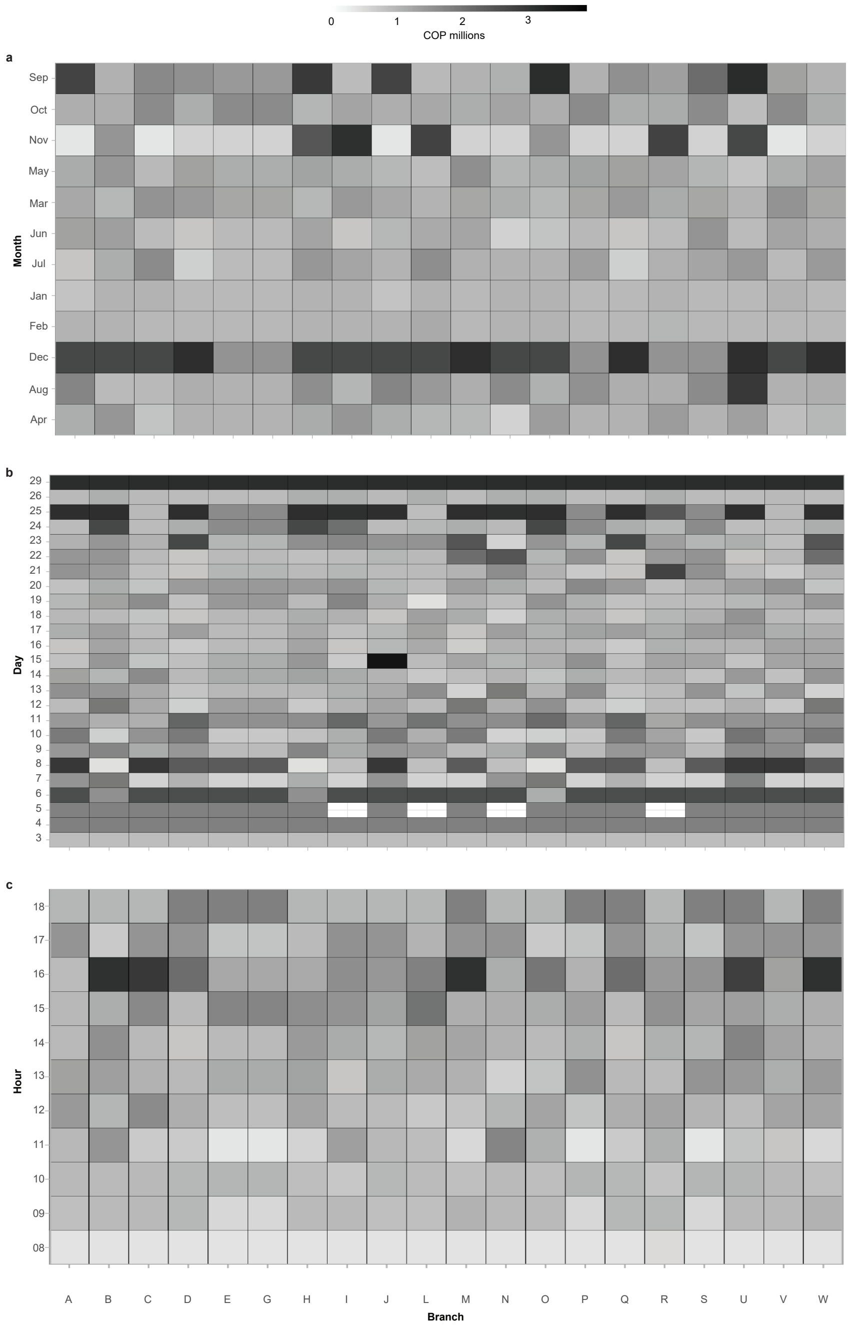

Figure 2: Heatmap Transactions. a. By Month. b. By Day. c. By Hour.

In December, it was identified that agents made operations of more than two million Colombian pesos (COP). In addition, branches A, H, J, O, and U during the month of September had several transactions greater than two million pesos and also in November for branches H, I, L, R, and U. The validation of these results allowed to identify some unusual patterns for branches H, U, and A, especially in the transaction value that becomes an alert process that can be classified as anomalies or suspicious transactions. On the other hand, transactions like deposits and withdrawals did not generate any alert given the scale of less than one million Colombian pesos (COP) (Figure 2a). However, using different branches by the same agent with a specific frequency could be classified as suspicious activity (Figure 3).

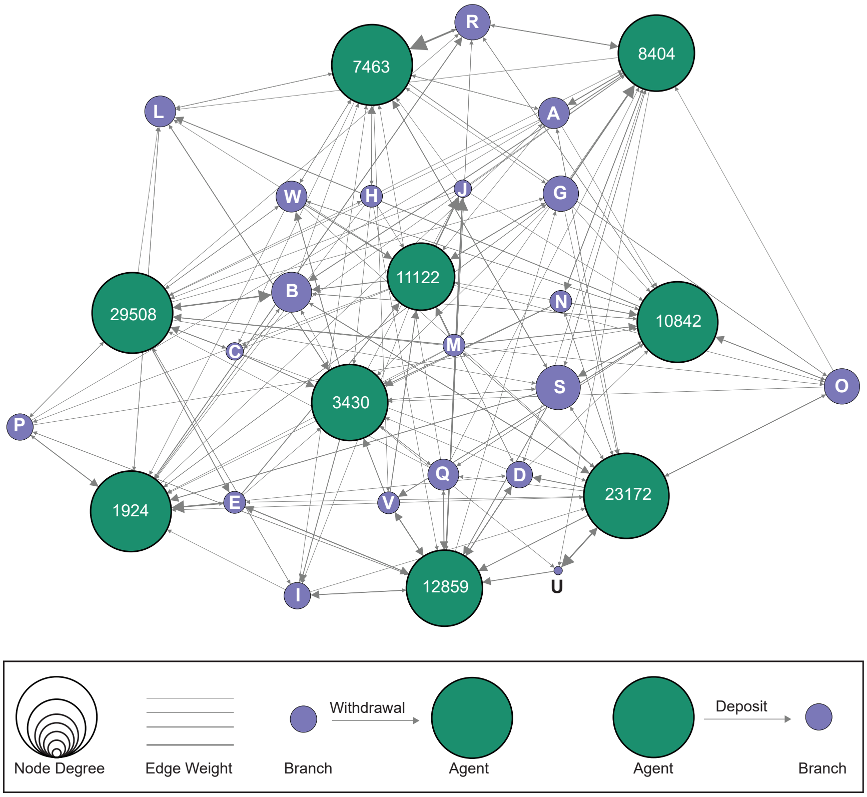

Figure 3: Simulation Networks. Transactions during June by branches and principal agents.

Here, some network science metrics allow identifying several of these elements, specifically, the average degree and the Weighted Average degree, as well as the network topology. However, several of these elements are beyond the scope of this research.

Furthermore, transactions are highly related to the days of the month and the simulated branch network (Figure 2b). Some days present an unusual behavior to the average of the days of the month, especially on the 15th, 25th, and 29th, with several transactions exceeding 2 million Colombian pesos, especially branch J. Similarly, for the 25th, some branches present transactions with values considerably higher than 2 million Colombian pesos. For the end of the month, the different types of transactions give high values, which means that the data show a seasonality typical of the paydays that occur in the Colombian economy. In addition, branches I, L, N, and R are null because they do not present records with significant values, i.e., given the scale of transactions, the values are low and do not show any alert. During the hours of the day, several patterns are relevant to mention. First, transactions are possible to carry out at any time (Figure 2c) transactions per hour describe the behavior of agents in carrying out their transactions of any type, but in the heat map it explains given the scale that the branches B, C, M, U, and W are transactions of the 16 hours where it presents an unusual behavior for the other hours of attention and branch.

To list a few results of Rule one, defined as the frequency of transactions per agent carried out during the year. Agent 3591 made 40 transactions during the year worth 47.5 million Colombian pesos, and agent 22206 made 37 transactions worth 41.2 million Colombian pesos (Table 1). But, what confirms this series of transactions? Although the simulation shows several characteristics of the agents, it is necessary to integrate it with some behavior patterns of the users of financial institutions, since the results of some agents do not allow us to identify if that money comes from an illegal activity but does generate suspicion that it uses so often the network of non-bank correspondents. Additionally, it could be crossed with tax information that confirms the type of economic activity that the agent has.

Each rule is more robust than the previous one, allowing finding new patterns. Rule two is the frequency of transactions per agent per month. For this case, we identified the first ten agents with the highest monthly transaction frequency (Table 2). To list a few, Agent 12859 made 33 transactions worth 35.6 million Colombian pesos during

Reg.	Agent	Frequency	Transaction Value (COP millions)
1	3591	40	47,5
2	3428	40	44,1
3	17184	40	43,5
4	7660	39	42,1
5	10436	38	42,2
6	29046	38	42,1
7	12859	38	41,4
8	19565	37	45,0
9	20579	37	44,2
10	22206	37	41,2

Table 1
Transaction frequency per principal agent

June, and Agent 3428 carried out 27 transactions in June for a value of 31.4 million Colombian pesos. This agent is in rule one made an undetermined number of transactions in the non-bank correspondent. The most frequent cases are in June with values above the average, where the hypothesis as to why the number of transactions occurs more frequently in this month and with high values could be semi-annual bonus payments to employees and the summer-break season. Rule three is defined as the frequency of transactions by agent, branch, and month (Table 3). The results of this rule take two agents (1000 and 1001) of the first twenty, where each one carried out a series of transactions between May and July for different values in Colombian pesos. It is a series of cases that are difficult to quantify but that allow the identification of the patterns that resemble suspicious activities and in which circumstances that drive of using non-bank correspondents for criminal activities such as money laundering could be traced because these agents exceed the average number of transactions and begin to use different branches.

Rule four is defined as the frequency of transactions by agents, branch, month, day, and type of transaction. It is possible to extract the information of the agents that carry out transactions during the same day (Table 4). Each agent represented in the table is the filter with the highest frequency of use and transactional services with amounts greater than 4 million Colombian pesos. For example, on June 15th, agent 24116 made four withdrawals for more than five million Colombian pesos at branch E. This example is an anomaly because the banking branch's users do not use the same branch at different hours with those values. Simulations could identify the money laundering methods in non-bank correspondents and any suspicious transactions.

Finally, Rule five is defined as the frequency of transactions by agent, branch, month, time, and type of transaction. This rule includes the time of the transaction where the pattern of the time that agents choose to carry out their operations is defined. In several of the results, 1 pm was the most frequent hour (Table 5). The principal reason for this pattern is that it is a part of the day when users usually use banking branches. In one of the cases, agent 11,125 carried out operations for 7.1 million Colombian pesos at branch G with various transactions and amounts allowed by the financial entity. However, if able to identify transactions by agent at this level of detail, it is possible to establish the frequency per branch and how often another branch or hours are used to carry out transactions. These are processes that facilitate the verification of operations that could be evaluated and consolidate the control of suspicious activities.

5. Conclusion

The main contribution of this work is the characterization of suspicious money laundering activities using a simulation model that evolves from synthetic data based on deep learning methodologies. We developed the agent interaction simulation, identifying risk factors for a non-bank correspondent. However, identifying these risk factors is complex because money laundering in non-bank correspondents can bear several similarities with financial institutions, especially in deposit and withdrawal transactions.

Agents make transactions in the non-bank correspondents without generating any alert because they are well documented on the rules established by the financial system. However, agent-based models recreate the transactional behavior given by the synthetic data created by GAN networks, which identify patterns of transactional behavior to generate large-scale information based on the original data. We could identify and quantify anomalies or suspicious activities based on the synthetic data analysis. Therefore, the knowledge extracted from this model allows us to identify

Reg.	Agent	Month	Frequency	Transaction Value (COP millions)
1	12859	June	33	35,6
2	3428	July	27	31,4
3	10842	June	27	29,3
4	23172	June	27	29,4
5	29508	June	27	28,0
6	1924	June	26	28,6
7	3430	June	26	28,6
8	7463	June	26	29,8
9	8404	June	26	28,9
10	11122	June	26	28,5

Table 2
Transaction frequency by agent and month

Reg.	Agent	Branch	Month	Frequency	Transaction Value (COP millions)
1	1000	W	May	1	1,3
2	1000	P	May	1	0,8
3	1000	M	June	2	2,7
4	1000	J	June	1	1,1
5	1000	B	June	1	0,9
6	1000	N	June	1	0,9
7	1000	E	June	1	0,8
8	1000	N	July	1	0,7
9	1000	H	June	1	0,7
10	1001	A	May	1	1,8
11	1001	B	May	1	1,3
12	1001	P	June	4	4,0
13	1001	R	June	2	2,4
14	1001	I	June	2	2,0
15	1001	G	June	1	1,5
16	1001	C	June	1	1,5
17	1001	O	June	1	1,3
18	1001	V	June	1	1,1
19	1001	E	June	1	1,1
20	1001	D	July	1	1,1

Table 3
Transaction frequency by agent, branch, and month

several methods and techniques that the money launderer uses in non-bank correspondents. Those simulation models could help authorities trace the origin to the destination of illegal money.

Additionally, the model identified several patterns from each rule because each rule is defined by the previous one, consolidating the model's robustness. When the model identifies a suspected money laundering agent, it can track operations by type of transaction, day, or hour. This feature could consolidate the mechanisms to money laundering control of financial institutions and non-bank correspondents that have grown considerably in different emerging markets.

The methodology can be deployed in the financial system because the business and risk rules defined in this document can be validated by the proposals in the control mechanisms of financial institutions. However, this modeling requires collaboration and information exchange between the financial sector and governments to counteract the flow of illegal money in the economies. The disadvantage that can arise is the complexity of identifying the final agents who receive the illegal money. Likewise, to specify new operating mechanisms for criminal organizations and corruption schemes that grow in different productive activities and continue to use the financial system to launder money. Furthermore, synthetic data does not allow statistical inference about the universe.

We conclude that the GAN networks generate good quality synthetic data that are statistically consistent. Using

Reg.	Agent	Branch	Month	Day	Type	Frequency	Transaction Value (COP millions)
1	24116	E	June	15	Withdrawal	4	5,1
2	17486	A	June	16	Withdrawal	4	4,7
3	29395	H	June	18	Withdrawal	4	4,2
4	30645	G	June	16	Deposit	4	3,5
5	5376	J	June	15	Withdrawal	3	4,7
6	8635	V	June	17	Withdrawal	3	4,6
7	19815	E	June	17	Withdrawal	3	4,5
8	45674	D	May	14	Withdrawal	3	4,5
9	7642	M	June	17	Withdrawal	3	4,4
10	31530	M	June	14	Withdrawal	3	4,4

Table 4
Transaction Frequency by Agent, Branch, Month, Day and Type

Reg	Agent	Branch	Month	Hour	Type	Frequency	Transaction Value (COP millions)
1	11125	G	June	13	Withdrawal	5	7,1
2	28430	L	June	13	Deposit	5	4,6
3	53211	W	June	13	Deposit	5	4,4
4	3292	M	June	14	Withdrawal	4	6,0
5	32861	N	June	13	Withdrawal	4	5,7
6	15173	I	June	13	Withdrawal	4	5,7
7	22215	G	June	13	Withdrawal	4	5,7
8	41555	O	June	13	Withdrawal	4	5,5
9	29720	G	June	13	Withdrawal	4	5,5
10	40028	M	June	13	Withdrawal	4	5,4

Table 5
Transaction Frequency by Agent, Branch, Month, Hour and Type

this synthetic data with an agent-based model facilitates the creation of robust methodologies for money laundering control on non-bank correspondents. In addition, agent-based models allow a technical explanation of how they carry out money laundering among non-bank correspondents without generating any alert or suspicious activities. However, more work is required to formally this claim, especially from two perspectives. First, expand the use of more sophisticated agent models such as BDI or multi-agent architectures that strengthen the simulation. Second, use data from financial institutions to simulate a wide range of suspicious activities between banks and non-bank correspondents.

References

Ardizzi, G., Franceschis, P.D., Giammatteo, M., 2018. Cash payment anomalies and money laundering: An econometric analysis of italian municipalities. International Review of Law and Economics 56, 105–121. doi:10.1016/j.irle.2018.08.001.

Axelrod, R., 1997. The complexity of cooperation. Princeton University Press.

Beaulieu-Jones, B.K., Wu, Z.S., Williams, C., Lee, R., Bhavnani, S.P., Byrd, J.B., Greene, C.S., 2019. Privacy-preserving generative deep neural networks support clinical data sharing. Circulation: Cardiovascular Quality and Outcomes 12. doi:10.1161/circoutcomes.118.005122.

Boella, G., van der Torre, L., 2006. Constitutive norms in the design of normative multiagent systems, in: Toni, F., Torroni, P. (Eds.), Computational Logic in Multi-Agent Systems, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 303–319.

Castelfranchi, C., 2000. Engineering social order, in: Omicini, A., Tolksdorf, R., Zambonelli, F. (Eds.), Engineering Societies in the Agents World, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 1–18.

Demetis, D.S., 2018. Fighting money laundering with technology: A case study of bank x in the UK. Decision Support Systems 105, 96–107. doi:10.1016/j.dss.2017.11.005.

Dreżewski, R., Sepielak, J., Filipkowski, W., 2015. The application of social network analysis algorithms in a system supporting money laundering detection. Information Sciences 295, 18–32. doi:10.1016/j.ins.2014.10.015.

Epstein, J.M., Axtell, R., 1996. Growing artificial societies: social science from the bottom up. Brookings Institution Press / The MIT Press.

Garcia-Bedoya, O., Granados, O., Burgos, J.C., 2020. AI against money laundering networks: the colombian case. Journal of Money Laundering Control 24, 49–62. doi:10.1108/jmlc-04-2020-0033.

Gilbert, N., Troitzsch, K., 2005. Simulation for the social scientist. McGraw-Hill Education.

González Martínez, E.F., 2020. Detección de fraude en tarjetas de crédito mediante técnicas de minería de datos. Editorial Académica Española.

González-Martínez, E.F., 2021. Generador de datos sintéticos para el monitoreo de transacciones con factores de riesgo de lavado de activos. Master's thesis.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, pp. 2672–2680.

Granados, O.M., Vargas, A., 2021. Financial Networks and Structure of Global Financial Crime. Springer International Publishing, Cham. pp. 131–152. URL: https://doi.org/10.1007/978-3-030-81484-7_8.

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S.K., Huse, G., Huth, A., Jepsen, J.U., Jørgensen, C., Mooij, W.M., Müller, B., Pe'er, G., Piou, C., Railsback, S.F., Robbins, A.M., Robbins, M.M., Rossmanith, E., Rüger, N., Strand, E., Souissi, S., Stillman, R.A., Vabø, R., Visser, U., DeAngelis, D.L., 2006. A standard protocol for describing individual-based and agent-based models. Ecological Modelling 198, 115–126. doi:https://doi.org/10.1016/j.ecolmodel.2006.04.023.

Guevara, J., Garcia-Bedoya, O., Granados, O., 2020. Machine learning methodologies against money laundering in non-banking correspondents, in: Communications in Computer and Information Science. Springer International Publishing, pp. 72–88. doi:10.1007/978-3-030-61702-8_6.

Hopgood, A.A., 2001. Intelligent systems for engineers and scientists. CRC Press.

Huckfeldt, R., Johnson, P.E., Johnson, P.E., Sprague, J., 2004. Political disagreement: The survival of diverse opinions within communication networks. Cambridge University Press.

Imanpour, M., Rosenkranz, S., Westbrock, B., Unger, B., Ferwerda, J., 2019. A microeconomic foundation for optimal money laundering policies. International Review of Law and Economics 60, 105856. doi:10.1016/j.irle.2019.105856.

Langr, J., Bok, V., 2019. GANs in Action: Deep Learning with Generative Adversarial Networks. Manning Publications.

Loayza, N., Villa, E., Misas, M., 2019. Illicit activity and money laundering from an economic growth perspective: A model and an application to colombia. Journal of Economic Behavior & Organization 159, 442–487. doi:10.1016/j.jebo.2017.10.002.

Luna-Pla, I., Nicolás-Carlock, J.R., 2020. Corruption and complexity: a scientific framework for the analysis of corruption networks. Applied Network Science 5. doi:10.1007/s41109-020-00258-2.

McCarthy, K.J., van Santen, P., Fiedler, I., 2015. Modeling the money launderer: Microtheoretical arguments on anti-money laundering policy. International Review of Law and Economics 43, 148–155. doi:10.1016/j.irle.2014.04.006.

Raghunathan, T.E., Reiter, J.P., Rubin, D.B., 2003. Multiple imputation for statistical disclosure limitation. Journal of official statistics 19, 1.

Ribeiro, H.V., Alves, L.G.A., Martins, A.F., Lenzi, E.K., Perc, M., 2018. The dynamical structure of political corruption networks. Journal of Complex Networks 6, 989–1003. doi:10.1093/comnet/cny002.

Sandfort, V., Yan, K., Pickhardt, P.J., Summers, R.M., 2019. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. Scientific Reports 9. doi:10.1038/s41598-019-52737-x.

Schneider, F., 2010. Turnover of organized crime and money laundering: some preliminary empirical findings. Public choice 144, 473–486.

Schneider, F., Windischbauer, U., 2008. Money laundering: some facts. European Journal of Law and Economics 26, 387–404. doi:10.1007/s10657-008-9070-x.

Surendra, H., Mohan, H., 2017. A review of synthetic data generation methods for privacy preserving data publishing. International Journal of Scientific & Technology Research 6, 95–101.

Tan, C., Behjati, R., Arisholm, E., 2019. A model-based approach to generate dynamic synthetic test data: A conceptual model, in: 2019 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), IEEE. doi:10.1109/icstw.2019.00026.

Tesfatsion, L., 2002. Agent-based computational economics: Growing economies from the bottom up. Artificial life 8, 55–82.

Walker, J., 1999. How big is global money laundering? Journal of Money Laundering Control 3, 25–37. doi:10.1108/eb027208.

Walker, J., Unger, B., 2009. Measuring global money laundering: "the walker gravity model". Review of Law & Economics 5. doi:10.2202/1555-5879.1418.

Yang, Y., Nan, F., Yang, P., Meng, Q., Xie, Y., Zhang, D., Muhammad, K., 2019. GAN-based semi-supervised learning approach for clinical decision support in health-IoT platform. IEEE Access 7, 8048–8057. doi:10.1109/access.2018.2888816.

Yoon, J., Jordon, J., van der Schaar, M., 2019. PATE-GAN: Generating synthetic data with differential privacy guarantees, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=S1zk9iRqF7.