The International Journal of Digital Accounting Research
Vol. 24, 2024, pp. 57-91
ISSN: 2340-5058

Anomaly detection with the density based spatial clustering of applications with noise (DBSCAN) to detect potentially fraudulent wire transfers

Yongbum Kim. Ramapo College of New Jersey, USA, ykim3@ramapo.edu

Miklos Vasarhelyi. Rutgers University, USA, miklosv@business.rutgers.edu

Abstract. Most anomaly detection models are developed by using expert system methods that mimic human experts. The process to capture the expertise honed by fraud examiners is complicated and practically challenging, often resulting in suboptimal models. This study proposes a clustering-based model that captures hidden characteristics of potentially fraudulent wire transfers with less human intervention and expertise. Clustering methods classify and group observations with similar characteristics, excluding anomalies from major clusters. The choice of a clustering method and its parameters is often subjective and significantly affects a set of resulting clusters. In order to reduce the subjectivity of a clustering method while retaining its strength, this study proposes a clustering model with Density Based Spatial Clustering of Applications with Noise (DBSCAN) to detect potentially fraudulent wire transfers of an insurance company. The results show that the DBSCAN models identifies hidden relationships between the variables not only included but also excluded for the modeling with noise wire transfers while less human intervention is needed for clustering parameter selections.

Keywords: Anomaly detection, fraud detection, clustering, DBSCN, density based clustering, spatial clustering.

1. INTRODUCTION

According to the Association of Certified Fraud Examiners (ACFE 2022), organizations lost approximately five percent of their revenues due to fraud. Nearly half of those cases occurred due to lack of internal controls (29%) or override of

DOI: 10.4192/1577-8517-v24_3

Submitted June 2023
Accepted December 2023

existing controls (20%). This indicates that appropriate internal controls and monitoring are crucial to deterring potential fraud. The ACFE also reported that a significant amount of fraud cases was detected by internal and external audits (16% and 4%, respectively) and automated transaction/data monitoring (4%), underlining their importance to fraud prevention.

Internal controls, monitoring, and audits are not only essential to curbing fraud. When a fraud detection model flags a transaction being suspicious for further examination, it may be fraudulent, erroneous, or legitimate. In this study, the term "anomaly" will be used to refer to any flagged transaction that is not legitimate. The purpose of monitoring and detection models in this study is to correctly identify both erroneous and fraudulent instances out of a population of transactions.

Developing an anomaly monitoring and detection model has many obstacles to overcome. Compared with the whole population of a given dataset, the number of anomalous (either fraudulent or erroneous) transactions is often extremely small. This imbalance makes anomaly monitoring and detection a challenging task – often compared to "finding a needle in a haystack." Another major obstacle to developing an anomaly monitoring and detection model is the false positive problem; anomaly detection models often flag too many transactions as being suspicious. For example, if one percent of a firm's daily transactions are flagged as suspicious, it would be too many for an internal audit team to practically examine due to limited human resources. In order to better distinguish anomalous transactions, both supervised and unsupervised methods have been used in anomaly detection literature (Bolton & Hand, 2001, 2002).

Supervised methods require a labeled dataset, where each transaction is classified as either fraudulent or legitimate. This enables researchers to obtain prior knowledge about fraudulent transactions. However, this may not be a desirable situation in practice because the existence of sufficient labeled data implies that the company suffered from too many anomalies, suggesting that it might be too late to develop and apply an anomaly monitoring and detection model. In other words, for most companies, development of an anomaly monitoring and detection model is likely to start when they notice a few fraud cases that may not be sufficient to apply a supervised method (Major & Riedinger, 2002). Practically speaking, an anomaly detection model has to be developed without prior knowledge of frauds due to lack of a labeled dataset.

Unsupervised methods are practically applicable methods for developing an anomaly monitoring and detection model because they do not require a labeled dataset (Chandola et al., 2009; Kim & Vasarhelyi, 2012). They receive less attention in fraud detection literature because the result of an unsupervised anomaly detection model is not a substitute for direct evidence of fraud, and their effectiveness and efficiency are often difficult to measure and verify. Some examples of unsupervised methods are rule-based models with a suspicion scoring system and clustering techniques (Kim & Vasarhelyi, 2012; Kim & Kogan, 2014; Freiman et al., 2022).

Clustering is an unsupervised data mining technique used in fraud detection. Clustering methods divide given observations into a preset number of groups, called clusters. Observations within a cluster are similar to each other and significantly dissimilar from observations in clusters (Thiprungsri & Vasarhelyi, 2011; Sabau, 2012). Similarity of observations are represented by the distance between them. The shorter the distance between two observations is, the more similar their characteristics. Distance-based clustering methods, such as K-means, are efficient when observations form circular or elliptical distributions, but they are inefficient for nonconvex clusters. For example, as shown in Figure 1, K-means clustering fails to capture true shapes of existing clusters – inner and outer donuts. These arbitrarily shaped clusters, however, can be captured by a density-based clustering method, Density-Based Spatial Clustering of Applications with Noise (DBSCAN). In auditing, it is highly unlikely that auditors have knowledge about the shapes of legitimate and anomalous transactions, so it is reasonable to assume that legitimate transactions have either convex or nonconvex shapes. Hence, DBSCAN is a more suitable method for auditing in that they can identify legitimate observations groups with arbitrary shapes.

DBSCAN is different from other clustering methods. If an observation does not meet the preset criteria for clustering, it does not form a cluster and is labeled "noise" (Ester et al., 1996; Khan et al., 2014; Tatusch et al., 2020). These noisy observations are a collection of outliers that are dissimilar from the observations included in the clusters, which is one of the reasons that DBSCAN is suitable for identifying anomalies (Sheridan et al., 2020).

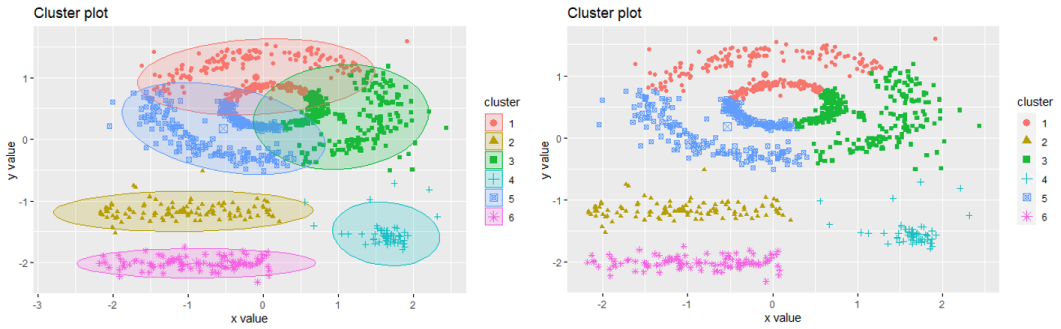

Figure 1. Clustering by K-means with and without shape-lines

In other clustering methods, each observation instead belongs to a cluster and the observations included in clusters below a given size threshold are considered anomalous. It often requires a subjective judgment to determine an appropriate threshold to define which clusters should be considered anomalous.

DBSCAN does not require prior knowledge about the structure of observations to determine the number of clusters to be formed. DBSCAN requires only two parameters: eps and minPts, where eps is a radius of a region and minPts is the number of observations within a region required to form a cluster. A cluster by DBSCAN has an arbitrary shape because it continues to grow as long as the minimum number of observations within a region is equal to or greater than minPts. Due to this nature, DBSCAN can capture nonconvex clusters (Ester et al., 1996; Khan et al., 2014; Tatusch et al., 2020).

In this study, the DBSCAN method was utilized to develop an anomaly monitoring and detection model that effectively discriminates between legitimate and suspicious transactions while minimizing false positive flags. The model also minimizes human intervention because it does not require prior knowledge about fraudulent transactions to form the necessary number of clusters.

The objective of this study is two-fold. Firstly, this research aims to illustrate the effectiveness of DBSCAN in detecting anomalous transactions. Secondly, this study contributes by demonstrating how anomaly detection activity using DBSCN can unveil hidden associations. This study yields valuable insights into identifying days and months with higher vulnerability, key initiators and approvers, specific lines of business susceptible, and distinct payees to fraud.

The rest of the paper is presented as follows. Prior literature about anomaly detection and clustering methods is summarized in Section II. Section III illustrates data description, model development, and analyses of the results. Section IV concludes with a discussion of future research.

2. LITERATURE REVIEW

2.1. Obstacles to Anomaly Detection

Auditing is a systematic process of objectively obtaining and evaluating evidence regarding assertions about economic actions and events. This concept was broadened by continuous auditing and monitoring, which emphasized timelier assurance and focused on transactional data (Vasarhelyi & Halper, 1991). Despite extensive academic work on continuous auditing and monitoring, the majority of research has focused on technical and theoretical proposals (Vasarhelyi & Halper, 1991; Kogan et al., 1999; Woodroof & Searcy, 2001; Rezaee et al., 2002; Murthy, 2004; Murthy & Groomer, 2004; Kim & Vasarhelyi, 2012). Relatively few papers conduct empirical studies on continuous auditing and monitoring due to a lack of applicable data (Bolton & Hand 2002; Phua et al., 2005; Kim & Vasarhelyi, 2012). Different from traditional auditing studies, empirical research on continuous auditing and monitoring requires transactional data, often considered a company's private asset and securely guarded to maintain a competitive position in the market. Consequently, companies are unwilling to offer their transactional data for continuous auditing and monitoring research (Kim & Vasarhelyi, 2012).

Other factors that discourage companies from offering their transactional data to researchers or the public are potential reputational damage and fraud that may arise from misuse of the disclosed information. Disclosure of fraud cases may have an adverse impact on a company's reputation in the market by giving a bad impression of having poor internal control systems, which may lead to a decrease in future revenues. Also, disclosure of fraud events may give detailed information about a company's operating and internal control systems that can be misused by potential fraudsters to penetrate that company's financial systems. Considering that most fraudsters were caught by simple mistakes, disclosure of such fraud cases may not be a good strategy for companies to prevent similar fraud in the future (Kim & Vasarhelyi, 2012; Kim & Alexander, 2014).

The nature of fraud detection is similar to that of continuous auditing in that the aim is to detect and correct anomalies in a timely manner. Since an anomaly includes both errors and frauds, anomaly detection is conceptually broader than fraud detection. Undetected fraud can cause a corporation to lose millions in revenue. As previously mentioned, the ACFE estimated that loss due to fraud was roughly five percent of an organization's revenue, though the true value might be higher, given the potential for undetected fraud. Although anomaly detection is often compared to looking for a needle in a haystack, its effectiveness was evidenced by the ACFE report.

2.2. Supervised and unsupervised methods in anomaly detection

Supervised anomaly detection methods are the most widely used methods in the anomaly detection literature. Under a supervised method, a fraud detection model is constructed with fraudulent and legitimate observations. Once a model is developed with a training set that is a portion of the classified data, it is tested with a test set using the remaining portion of the classified data. The test result is analyzed using various measures to gauge its prediction power (Bolton & Hand, 2002; Kim & Kogan, 2014; Kim & Vasarhelyi, 2012; Kou et al., 2004; Phua et al., 2005). Crucially, they utilize classified data, assuming that fraud patterns identified in a training dataset can determine fraudulency of each observation in a test dataset.

Despite the popularity of supervised methods in anomaly detection literature, they also have evident limitations. Prior knowledge about fraudulent and legitimate transactions is often unavailable in practice, and, when available, may include false positives and false negatives, which lead to suboptimal or inaccurate models. Even if a company examines all of its individual transactions, resulting models may not be broadly applicable (Bolton & Hand, 2002). Although an anomaly detection model is verified with a test dataset, the most accurate model may over-fit the data, decreasing external validity. An overfitted detection model may inefficiently or ineffectively detect fraud in future transactions or previously unknown types of fraud.

Unsupervised methods have received far less attention in the literature than supervised methods. An anomaly detection model using unsupervised methods purports to find transactions that do not exhibit expected behaviors. Although its major strength is that they do not require classified data, lack of classified data also poses a major weakness in that the model verification process is often challenging

due to lack of testable data. To overcome this weakness, various indirect verification methods, such as peer group analysis and break point analysis, were suggested in prior literature (Chandola et al., 2009; Bolton & Hand, 2002). In this study, profiling normal and anomalous transactions and comparing their characteristics were used as an indirect verification.

In practice, it is extremely rare to have sufficient labeled transactions when a company attempts to implement an anomaly monitoring and detection model. The labeled data is a collection of historical and documented transactions that shows that a transaction is either legitimate or fraudulent. Lack of labeled data serves as another challenge for the most companies that do not have preexisting anomaly detection systems. Consequently, it is imperative that companies are able to develop an anomaly detection model without labeled data.

Unsupervised methods show whether or not flagged transactions are potentially anomalous, after which internal auditors investigate for confirmation. In other words, they do not provide direct confirmatory evidence of anomalies. Despite the limitations of unsupervised methods, they may play a critical role at the initial implementation stage of anomaly detection, where the available data lacks labeled classification information. In addition, the results of unsupervised methods are broader than those of the supervised methods, which can help identify general patterns of anomalous transactions (Kim & Vasarhelyi, 2012; Liu & Vasarhelyi, 2013).

2.3. Clustering methods

Clustering techniques are the most widely applied unsupervised methods in research. Cluster analysis identifies a series of groups that share similar characteristics and identifies unusual observations that possess dissimilar characteristics. Anomalous observations can be defined in several ways: Anomalies are assumed to be observations that do not belong to any cluster, anomalies are observations whose distances are far from the nearest cluster centroid, or anomalies are defined as observations that belong to small or sparse clusters (Chandola et al., 2009; Thiprungsri & Vasarhelyi, 2011). This study adopted the first approach because it does not require further procedure to determine thresholds to define anomalies, which can facilitate application by practitioners.

The observations that do not belong to any clusters in this study are considered anomalies because they have distinctively different nature from those included in clusters. Anomalous transactions identified by a clustering analysis were labeled *potentially fraudulent*, which would demand further investigation to confirm their legitimacy Considering that companies usually process similar transactions in the course of their day-to-day operations, it will be reasonable to assume that they share similar characteristics. Accordingly, those dissimilar from others are outliers that are more prone to errors or fraud unless the company is newly opened and more likely to have new types of transactions.

Finding fraud is, however, often compared to "finding a needle in a haystack" because fraud often looks legitimate in appearance. This might be a reason why only four percent of occupational fraud cases were detected by automated transaction/data monitoring and most fraud cases were detected by nonstatistical methods, such as tips (42%) and management review (12%) (ACFE, 2022). Since clustering methods purport to separate observations by similarity, it will be challenging to detect fraud that is similar in appearance, especially when they belong to major groups. Hence, clustering methods are more suitable for flagging fraudulent transactions with characteristics that are different from legitimate observations. Observations that are flagged by the detection model in this study are labeled "potentially anomalous," which implies that their legitimacy can be identified only by further investigation. These flagged transactions are unusual in nature because they have dissimilar characteristics from those of major groups. This dissimilarity might result from their being erroneous or fraudulent. In this study, the DBSCAN results were analyzed by various variables to demonstrate the major factors that contributed to these dissimilarities.

A clustering analysis can be carried out with various clustering methods that have different logics for grouping observations with similar characteristics. Each clustering method generates different sets of clusters so that it is important to choose the best-fitting for the given dataset, which poses a major concern in selecting a clustering method. Once a specific clustering method is chosen, the next step is to determine parameters, such as the number of clusters. This task is often challenging and subjective unless prior knowledge about the given data is readily available. The parameter selection plays a critical role in determining the resulting clusters and is often time-consuming and laborious.

K-means clustering, one of the most well-known clustering techniques, is less efficient when observations consist of nonconvex shapes because it produces convex (circular or elliptical) clusters. Another disadvantage of K-means clustering is that the number of clusters must be predetermined. Considering that characteristics of anomalies are often unknown at the initial stage of developing fraud detection model, it may not be the best candidate for developing an anomaly monitoring and detection model. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) can tackle these problems in that it forms arbitrary shaped clusters and provides simpler ways to determine necessary parameters.

2.4. DBSCAN

DBSCAN is a density-based clustering method, which means it forms clusters based on how close observations are to each other. It extends a cluster as long as the minimum number of observations (minPts) exists within a given radius (eps). This algorithm captures arbitrarily shaped clusters. Unlike other clustering methods, observations do not belong to any clusters if they fail to meet the criteria to form a cluster (called *noise*) and they show distinctively different behavioral patterns from observations included in clusters (Ester et al., 1996; Khan et al., 2014; Tatusch et al., 2020). This conceptual framework is conceptually aligned with the definition of an anomaly, which is assumed to behave differently from majority groups.

In applying DBSCAN to anomaly detection, the focus lies on the noise rather than the clusters. If an observation belongs to a cluster, it shares similar characteristics with at least the minPts number of observations, suggesting that it is unlikely to be anomalous. However, those that do not belong to any clusters have fewer than the minPts number of neighboring observations and are considered potentially anomalous. Failing to form a cluster indicates that their characteristics are dissimilar from those included in clusters. This implies that the selection of two parameters affects the number of resulting noises that will increase for smaller eps and/or larger minPts.

In this study, one of the simplest, but most widely used, parameter selection methods was applied to minimize human intervention. This reduces subjectivity in the parameter selection, making DBSCAN more useful to practitioners who may not be familiar with using clustering techniques.

In this study, insurance payments to clients were explored with DBSCAN analysis to develop an anomaly detection model for identifying potentially fraudulent transactions. The characteristics of noises and major groups were compared to demonstrate that DBSCAN effectively discriminated between potential anomalous transactions.

3. METHODOLOGY

3.1. Data description

This study used records of wire payments made by a large U.S. insurance company. The dataset provided by the company consisted of 20,000 transactions whose effective dates ranged from 1/2/2008 to 12/31/2010 with 27 quantitative and categorical variables. Out of the 27 variables, 15 variables were excluded for the DBSCAN analyses because they had missing values whose causes were undeterminable. For example, a transaction might have no routing number for a variety of reasons: because it was omissible when it was sent to a well-known payee, because it was an optional field, because it was a left off in error, or because it was fraudulent. There were many ways to handle missing values for clustering, such as imputation and forming clustering only with observations without missing values, but it was unclear how the choice of missing value treatment would affect the result and which treatment method would best fit for this study. To simplify the DBSCAN process, this study excluded variables with missing values, relegating further investigation of missing value treatment to a future study. The 12 variables chosen for the clustering analyses in this study were Wire ID, Amount, Initiation date, Effective date, Account number of a line of business, Payee ID, month and day of initiation, month and day of the effective date, day of initiation date (Monday, Tuesday, Wednesday, Thursday, Friday, and non-working days such as Saturday, Sunday, and holidays), and day of the effective date.

A wire transfer was first initiated by an employee of a line of business (“initiator”) and approved by one or two employees (“approvers”), depending on its amount. Each initiator and approver had their own authorization limits, and a wire transfer beyond their authorization limits was forwarded to a senior employee with a higher authorization limit, such as a manager. Suspicious wire transfers also required a second approver to confirm their legitimacy. Employees only addressed transfers for the bank account groups to which they were assigned.

3.2. Model development

In order to verify the effectiveness of the DBSCAN model, the dataset was divided into two groups and their results were compared. One group was called the Training set which consisted of the first 10,000 wire transfers and the remaining was allocated to the other group, the Test set. This division was designed to mimic how fraud detection activity worked in real life, where we assumed that fraud detection activity would be initiated every 10,000 transactions.

3.2.1. Training set

Clustering requires extraordinary computational power which often leads to extreme heat that causes a system failure. In order to mitigate this problem and improve performance, unnecessary variables are removed. The 12 numeric variables, "dimensions", of each group were reduced to nine principal components (PCs) by a principal component analysis (PCA) often used for dimensionality reduction. The optimal number of dimensions was selected by analysis of the variances of the PCs (Table 1) and a scree plot (Figure 2). A common strategy when using a scree plot is selecting the number of PCs at the line's "elbow" where the percentage of explained variances shows a dramatic drop.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Standard Deviation	1.8711	1.3591	1.2761	1.1467	1.0611	0.9637	0.9071	0.7615	0.4482	0.2150	0.0605	0.0011
Proportion of Variance	0.2918	0.1539	0.1357	0.1096	0.0938	0.0774	0.0686	0.0483	0.0167	0.0039	0.0003	0.0000
Cumulative Proportion	0.2918	0.4457	0.5814	0.6910	0.7848	0.8622	0.9308	0.9791	0.9958	0.9997	1.0000	1.0000

Table 1. Analysis of the Variance of the PCs

The scree plot in Figure 2 may have two elbows at the dimensions of two and nine. Considering their cumulative variance (0.4457 and 0.99584, respectively), the significantly more explanatory elbow at nine PCs was chosen.

A DBSCAN model does not require that the number of clusters be determined in advance with ex ante expectations about observations' behavior patterns. Instead, it requires two parameters, eps and minPts, where the eps is a distance from a point to k nearest neighbor and the minPts is the number of k nearest neighbors within a specific distance (eps).

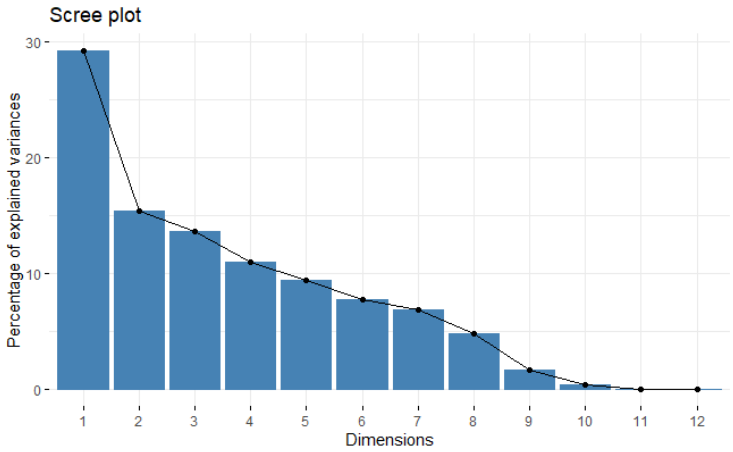

Figure 2. Accumulated variances of principal components

In a DBSCAN analysis, a dense region is formed if the number of points within an eps distance is the value of minPts or more. This process stops when no more dense formation is possible. This results in clusters with arbitrary shapes. The selection process of these two parameters may cause a subjectivity issue because there are many options for their determination. In order to minimize human interventions, two methods were utilized in this study: a kNNdistplot in R software and a natural log function.

The kNNdistplot is a commonly used function in determining an appropriate eps value. This function plots kNN distances that are distances from a point to its k nearest neighbor. When plotted, kNN distances are depicted against the points that are sorted by distance. The kNN distance at the "knee" in the plot is considered suitable for a DBSCAN model. As shown in Figure 3, the 12-NN distance at the knee was around three, so the eps for the DBSCAN model was set to three.

The other parameter, minPts, is often determined by using a natural log function. The minPts is a natural log value of the number of observations (Equation 1).

In this study, each dataset consisted of 10,000 observations, so minPts was set to 9, as it was the nearest integer value for the natural log value of 10,000 (9.21).

Equation 1. Computation of minPts

$$\text{minPts} = \ln(\text{the number of observations})$$

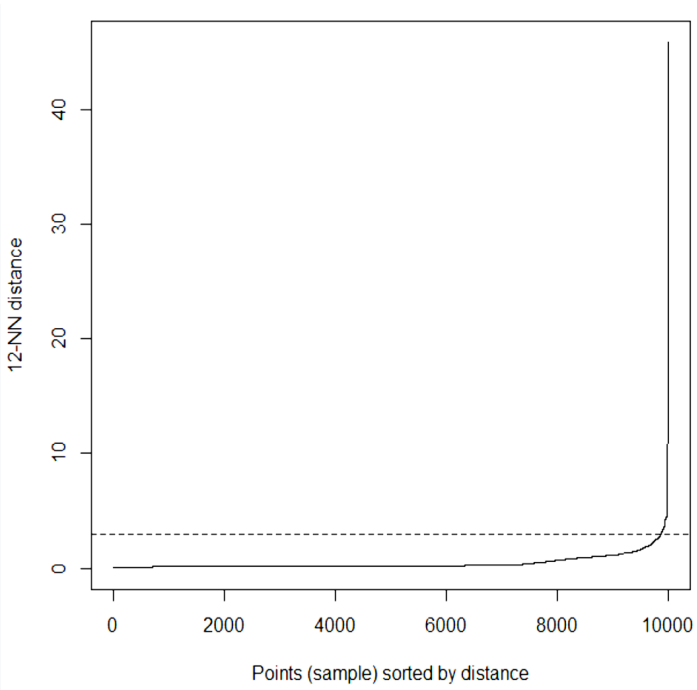

Figure 3. Training set – kNNdistplot

The DBSCAN on the Training set with an eps of three and a minPts of nine resulted in eight clusters with 46 points of noise (Table 2). Cluster 0 contains all points throughout the distribution that failed to form a cluster. As depicted in Figure 4, each cluster had arbitrary shapes, which differentiates DBSCAN from other clustering methods that form elliptical or circular clusters. The black dots are outliers that represent noisy wire transfers with behavioral patterns that are dissimilar from those in the eight clusters.

Cluster	0	1	2	3	4	5	6	7	8
Number of observations	46	9,751	34	64	9	27	9	9	51

Table 2. DBSCAN for the training set

In order to confirm that the characteristics of noise were distinctively dissimilar from those of clustered observations, descriptive statistics of the transfer payments were compared for variables used and excluded from the clustering.

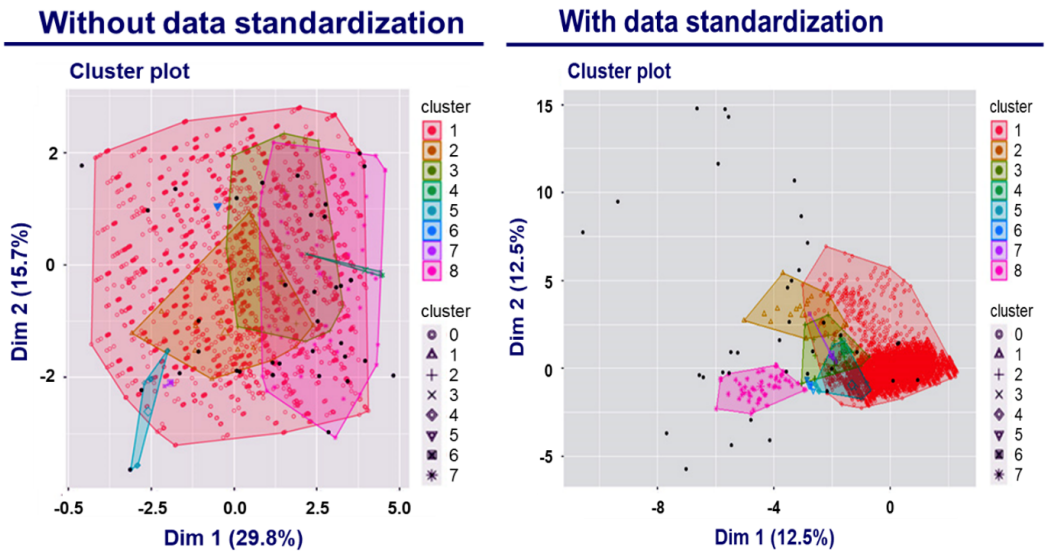

Figure 4. Visualization of the clusters

The variables used in the clustering were compared to show that the noise transfer payments were distinct from normal observations. Comparisons of the variables not included in the clustering were performed to examine whether the DBSCAN model could capture unknown characteristics of other variables.

		Included in the model	Excluded from the model
Descriptive statistics		• Amount	• Initiator - Authorization limit • Approver – Authorization limit
Frequency tests	Significant	• Day of initiation date • Day of effective date • Month of initiation date • Month of effective date • Payee ID	• Initiator ID • Initiator LOB • Approver ID • Approver LOB
	Insignificant	• Day of month of initiation date • Date of month of effective date	

Table 3. Variables in comparison

Variables included in the model

a. Descriptive statistics

A fraudulent wire transfer of a larger amount will be more damaging to an organization than a smaller transfer. As such, it was expected that noise transactions

would have higher amounts than the major transfers (i.e., non-noise wire transfers), and they would call for more attention. As expected, descriptive statistics of the 46 noise transfers showed higher average and median than those of normal transfers (Table 4). However, caution must be taken to interpret this result. Alternatively, fraudulent transactions may have similar amounts of their legitimate peers if fraudsters tried to mimic a pattern of legitimate transfers to avoid being caught. Considering that materiality is of major concern to auditors, this study focused on the fraudulent transfers with larger amounts (Kim & Kogan, 2014).

Type	_FREQ_	N	Nmiss	Min	Max	Range	Mean	Stdev	P25	P50	P75	LCLM	UCLM	Skew	Kurt
Major	9,954	9,954	-	0.0	10,000,000.0	10,000,000.0	69,630.4	233,585.2	14,135.7	31,817.2	61,643.6	65,041.1	74,219.7	21.0	631.1
Noise	46	46	-	0.0	150,000,000.0	150,000,000.0	19,845,778.9	41,181,495.1	4,574.7	67,139.0	14,913,862.1	7,616,384.5	32,075,173.4	2.0	2.7

Table 4. Training set – descriptive statistics: amount

b. Frequency test

Five out of the remaining 11 variables in the model displayed clear behavior differences between the major cluster transfers and the noise transfers: Day of initiation date, Day of effective date, Month of initiation date, Month of effective date, and Payee ID. These differences might not have been captured without DBSCAN's uniquely flexible clustering potential.

b.1. Variables with significant differences

The comparison of Day of initiation date showed that noise transfers were initiated during non-working days, such as Saturdays, Sundays, and holidays (Column 9 in Table 5). In the major clusters, wire transfers during non-working days accounted for only 0.36% of the total transfers, while the percentage of noise transfers that took place during the same periods was 4.35%. Considering that fraudsters needed opportunities to commit fraud, non-working days when no other employees were present could create better opportunities than working days. Another notable day that drew attention was Fridays (Table 5, column 5). While rare for clustered transfer payments (4.37%), 39.13% of noise transactions were initiated on Fridays. This might happen because wire transfers initiated on Fridays can take longer to verify than those initiated on other working days. As shown in Table 6, the percentage of noise wire transfers that took place on working days started increasing from Wednesdays, with the largest share initiated on Fridays.

	1	2	3	4	5	9	Total
Major	6,765	1,518	588	612	435	36	9,954
	67.96%	15.25%	5.91%	6.15%	4.37%	0.36%	100.00%
Noise	6	4	7	9	18	2	46
	13.04%	8.70%	15.22%	19.57%	39.13%	4.35%	100.00%
Total	6,771	1,522	595	621	453	38	10,000
>10X						> 10X	

Table 5. Training set – frequency test: day of initiation date

The second variable in comparison was Day of effective date that wire transfers were approved and ready for payments. Similar to day of initiation date, Table 6 shows that a greater percentage of noise transfers went into effect during non-working days (4.35% vs. 0.18% under Column 9 in Table 6). One explanation is that outside fraudsters might ask for an expedited process to avoid regular verification procedures during working days. Also, this might happen when insider fraudsters tried to minimize a chance of being monitored by another employee during working days. Another notable day worthy of attention was Mondays (Column 1 in Table 6). Compared to the major transfer payments (4.91%), 32.61% of noise transfers went into effect on Mondays. This might happen because wire transfers initiated on Fridays were verified and ready for payments on Mondays. As observed in Figure 4, there was one day of latency between initiation date and effective date, which implies that the insurance company tries to process initiated wires by the very next business day. Fraudsters might misuse this business practice by initiating fraudulent wire transfers on Fridays that had to be processed on Mondays, which were one of the busiest days for the company. It is highly likely that the company's employees had to spare less time on each wire transfer to meet the company's service policy.

	1	2	3	4	5	9	Total
Major	489	6,770	1,366	675	636	18	9,954
	4.91%	68.01%	13.72%	6.78%	6.39%	0.18%	100.00%
Noise	15	7	4	7	11	2	46
	32.61%	15.22%	8.70%	15.22%	23.91%	4.35%	100.00%
Total	504	6,777	1,370	682	647	20	10,000
>10X						> 10X	

Table 6. Training set – frequency test: day of effective date

The third and fourth variables in comparison were initiation month and effective month. This was to examine if the noise transactions took place more frequently in particular months of the year. The rationale behind this investigation was the motivation component of the fraud triangle (Motivation, Opportunity, and Rationalization). Motivation relates to a fraudster's incentive to commit fraud, such as financial hardship, or simple greed. Considering that people spend more money during holiday seasons, suspicious transfer payments were expected to occur more frequently in the months close to year-end. As shown in Table 7 and Table 8, the wire transfers in the major clusters were initiated and made effective evenly throughout the year, while about 50% of the noise wire transfers were processed in November and December. This finding indicates that potentially suspicious wire transfers were likely to take place in these two year-end months.

	1	2	3	4	5	6	7	8	9	10	11	12	Total
Major	680	803	1,022	770	747	895	815	928	824	748	826	896	9,954
	6.83%	8.07%	10.27%	7.74%	7.50%	8.99%	8.19%	9.32%	8.28%	7.51%	8.30%	9.00%	100.00%
Noise	3	4	0	4	1	3	0	3	4	1	14	9	46
	6.52%	8.70%	0.00%	8.70%	2.17%	6.52%	0.00%	6.52%	8.70%	2.17%	30.43%	19.57%	100.00%
Total	683	807	1,022	774	748	898	815	931	828	749	840	905	10,000
>10X													
>2X											> 2X	> 2X	

Table 7. Training set – frequency test: month of initiation date

	1	2	3	4	5	6	7	8	9	10	11	12	Total
Major	771	811	976	820	746	849	860	878	864	757	779	843	9,954
	7.75%	8.15%	9.81%	8.24%	7.49%	8.53%	8.64%	8.82%	8.68%	7.60%	7.83%	8.47%	100.00%
Noise	3	4	0	4	1	3	0	2	5	1	14	9	46
	6.52%	8.70%	0.00%	8.70%	2.17%	6.52%	0.00%	4.35%	10.87%	2.17%	30.43%	19.57%	100.00%
Total	774	815	976	824	747	852	860	880	869	758	793	852	10,000
>10X													
>2X											> 2X	> 2X	

Table 8. Training set – frequency test: month of effective date

The fifth variable, Payee ID, was used to examine if the noisy wire transfers were related to particular payees. As shown in Table 9, most of the noise wire transfers were related to new payees that did not have any wire transfers in the major clusters. Conversely, payees that received wire transfers more often were less likely to have noisy transfers. For example, the payee with the ID of 30432 had 144 wire transfers,

all of them were in the major clusters. Similarly, the payee with the ID of 30422 had 693 wire transfers and only two of them were noise.

	281	1367	1368	3821	3824	7167	10336	11072	15887	15893	16988
Major	0	0	0	0	0	0	0	0	0	0	2
	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%
Noise	1	1	1	1	1	1	1	1	1	1	0
	2.17%	2.17%	2.17%	2.17%	2.17%	2.17%	2.17%	2.17%	2.17%	2.17%	0.00%
Total	1	1	1	1	1	1	1	1	1	1	2
>10X	> 10X	> 10X	> 10X	> 10X	> 10X	> 10X	> 10X	> 10X	> 10X	> 10X	
New	New	New	New	New	New	New	New	New	New	New	

	17996	18737	19011	19874	19932	…	30422	30429	30432	30433	Total
Major	2	1	1	0	2	…	691	8	144	46	9,954
	0.02%	0.01%	0.01%	0.00%	0.02%	…	6.94%	0.08%	1.45%	0.46%	100.00%
Noise	0	1	0	2	0	…	2	0	0	1	46
	0.00%	2.17%	0.00%	4.35%	0.00%	…	4.35%	0.00%	0.00%	2.17%	100.00%
Total	2	2	1	2	2	…	693	8	144	47	10,000
>10X		> 10X		> 10X		…					
New				New							

Table 9. Training set – frequency test: Payee ID

b.2. Variables with significant differences

Not all variables in the DBSCAN analysis showed meaningful differences between the major and the noise groups. Day of month of initiation date (Table 10) and Date of month of effective date (Table 11) did not show clear behavioral differences between the major and noise wire transfers. Although these two variables did not produce significant differences, there was one observation that might require further investigation. In the result of Day of the month of initiation date, 21 noise wire transfers were initiated over the 10th, 11th, and 12th, and they all seemed to be effective on the 15th. There were no clear expectations or explanations about this phenomenon. Except for this case, the occurrences of the noise wire transfers were well distributed over days within months. The remaining four variables, Wire ID, Initiation date, Effective date, and Account number of a line of business did not show any significant results.

	1	2	…	10	11	12	13	14	15	16	…	28	29	30	31	Total
Major	393	529	…	280	296	354	452	485	343	282	…	310	338	232	106	9,954
	3.95%	5.31%	…	2.81%	2.97%	3.56%	4.54%	4.87%	3.45%	2.83%	…	3.11%	3.40%	2.33%	1.06%	100.00%
Noise	1	2	…	4	6	11	0	1	0	0	…	2	2	1	1	46
	2.17%	4.35%	…	8.70%	13.04%	23.91%	0.00%	2.17%	0.00%	0.00%	…	4.35%	4.35%	2.17%	2.17%	100.00%
Total	394	531	…	284	302	365	452	486	343	282	…	312	340	233	107	10,000
>10X			…								…					

Table 10. Training set – frequency test: day of month of initiation date

	1	2	3	4	…	14	15	16	17	18	…	28	29	30	31	Total
Major	226	373	429	427	…	453	521	353	299	300	…	281	339	322	112	9,954
	2.27%	3.75%	0.043	4.29%	…	4.55%	5.23%	3.55%	3.00%	3.01%	…	2.82%	3.41%	3.23%	1.13%	100.00%
Noise	1	1	2	1	…	1	12	0	0	1	…	5	3	2	0	46
	2.17%	2.17%	0.044	2.17%	…	2.17%	26.09%	0.00%	0.00%	2.17%	…	10.87%	6.52%	4.35%	0.00%	100.00%
Total	227	374	431	428	…	454	533	353	299	301	…	286	342	324	112	10,000
>10X					…						…					

Table 11. Training set – frequency test: day of month of effective date

Variables excluded from the model

Clustering methods are used to capture hidden characteristics of observations in a model without prior knowledge about them. The primary goal of this study was to capture the hidden relationships of the variables included in the model, but analyses were also conducted to determine whether DBSCAN can identify patterns in variables that were not included in the model. This section details the relationships found using DBSCAN on variables not included in the model.

a. Descriptive Statistics

Descriptive statistics of Initiator's Authorization Limit and Approver's Authorization Limit were compared to find any significant differences. As shown in Table 12, noisy wire transfers were not significantly different from the clustered wire transfers for Initiator - Authorization limit. The only significant difference was that the initiators of the noisy wire transfers had higher authorization limits for Max and P75. However, considering their magnitudes (1,071,352,074.6 vs. 1,718,387,206.7 for Max; 1,000,000,000.0 vs. 1,051,599,979.4 for P75), all of which were greater than one billion, their comparison did not seem meaningful. Furthermore, the initiator's authorization limit reached one billion at P25 which implied that the employees were allowed to initiate all wire transfers as a matter of

practical convenience. Only if the amount was absurdly high would it require approval.

Type	_FREQ_	N	Nmiss	Min	Max	Range	Mean	Stdev	P25	P50	P75	LCLM	UCLM	Skew	Kurt
Major	9,954	9,867	87	671,050,846	1,071,352,075	400,301,228	970,843,317	96,083,557	1,000,000,000	1,000,000,000	1,000,000,000	968,947,232	972,739,403	-2.8	5.8
Noise	46	45	1	671,050,846	1,718,387,207	1,047,336,360	1,019,104,188	131,107,824	1,000,000,000	1,000,000,000	1,051,599,979	979,715,038	1,058,493,339	2.7	20.5

Table 12. Training set – descriptive statistics: Initiator's authorization limit

In Table 13, differences between the clusters and the noise were found for Approver - Authorization limit. The approvers in the noisy wire transfers had significantly higher values for Min, Max, Mean, P25, P50, and P75. These findings were related to amounts of the noisy wire transfers. As discussed in the descriptive statistics of Amount, noisy wire transfers had higher amounts than those of the major wire transfers. Obviously, the approvers needed higher authorization limits for approval.

Type	_FREQ_	N	Nmiss	Min	Max	Range	Mean	Stdev	P25	P50	P75	LCLM	UCLM	Skew	Kurt
Major	9,954	9,923	31	373,933	189,310,454	188,936,521	7,497,905	21,715,395	999,999	999,999	999,999	7,070,591	7,925,219	4.7	26.7
Noise	46	45	1	999,999	851,261,995	850,261,996	98,748,027	183,071,506	5,000,000	5,000,000	170,260,804	43,747,257	153,748,796	3.3	11.7

Table 13. Training set – descriptive statistics: Approver's authorization limit

b. Frequency tests

This section compares four variables that were excluded from the model, Initiator ID, Initiator LOB (Line of Business), Approver ID, and Approver LOB, to test whether noisiness was related to specific initiators, approvers, and/or their LOBs. Comparisons of Initiator ID (Table 14) showed that specific initiators had significantly more noisy wire transfers and some of them were new initiators. For example, initiator 554809 was associated with 30.43% of the total noise wire transfers and the initiator 1006563 was new to wire transfer initiations. The related variable, Initiator LOB, in Table 15 also showed that specific Initiator LOBs were related to the noise wire transfers. For example, LOB 10025098 was a new LOB that had only noisy wire transfers. Another notable finding was that the number of noisy wire transfers was not proportional to the number of all wire transfers initiated by either the initiators or the LOBs. For example, initiator 1018694 had 2 noisy wire transfers that were 4.35% of the total number of the wire transfers initiated, while 19.57% of the wire transfers initiated by the initiator 1012738 were noise. Similar results were found in Initiator LOB. These findings may help improve the company's internal control systems by allocating more resources to specific initiators and their LOBs or by identifying problematic employees for discipline.

	375727	377989	489669	531936	554809	…	996669	1006563	1012675	1012738	1018694	1019485	1020227	…	1044825	1045537	1047624	1050431	1053527	Total
Major	10	76	1	7	72	…	16	0	0	18	684	50	81	…	5	24	37	5	14	9,954
	0.10%	0.76%	0.01%	0.07%	0.72%	…	0.16%	0.00%	0.00%	0.18%	6.87%	0.50%	0.81%	…	0.05%	0.24%	0.37%	0.05%	0.14%	100.00%
Noise	1	0	0	0	14	…	0	4	1	9	2	0	6	…	0	0	1	0	0	46
	2.17%	0.00%	0.00%	0.00%	30.43%	…	0.00%	8.70%	2.17%	19.57%	4.35%	0.00%	13.04%	…	0.00%	0.00%	2.17%	0.00%	0.00%	100.00%
Total	11	76	1	7	86	…	16	4	1	27	686	50	87	…	5	24	38	5	14	10,000
>10X	> 10X				> 10X	…		> 10X	> 10X	> 10X			> 10X	…						
New								New	New											

Table 14. Training set – frequency test: Initiator ID

	10023755	10023756	10023761	10023762	10024552	10025098	10025286	10025815	10027036	10032440	10034880	10042280	10049070	10068049	5025814	5025815	85025814	85025815	Total
Major	7	1	454	526	11	0	79	61	3926	45	13	200	19	4487	72	15	14	24	9954
	0.07%	0.01%	4.56%	5.28%	0.11%	0	0.79%	0.61%	39.44%	0.45%	0.13%	2.01%	0.19%	0.4508	0.72%	0.15%	0.14%	0.24%	100.00%
Noise	0	0	19	1	0	4	14	0	2	0	0	1	1	3	1	0	0	0	46
	0.00%	0.00%	41.30%	2.17%	0.00%	8.70%	30.43%	0.00%	4.35%	0.00%	0.00%	2.17%	2.17%	6.52%	2.17%	0.00%	0.00%	0.00%	100.00%
Total	7	1	473	527	11	4	93	61	3,928	45	13	201	20	4,490	73	15	14	24	10,000
>10X						> 10X	> 10X						> 10X						
New						New													

Table 15. Training set – frequency test: Initiator LOB

Comparisons for Approver ID (Table 16) and Approver LOB (Table 17) showed that noisy wire transfers were related to specific approvers and approver LOBs. As shown in Table 16, the approver 551036 approved 28.26% of the total noise, and many approvers, such as approver 952394, that approved noise were new to the approval process. The related variable, Approver LOB, also showed that specific Approver LOBs were common among noisy wire transfers. For example, LOB 10025098 was a new LOB that only approved noise. Similar to Initiator ID and Approval LOB, the number of noisy wire transfers was not in proportion to the number of all wire transfers approved by either the approvers or the LOBs. For example, approver 344248 had no noisy wire transfers despite approving 161 wire transfers, while 28.26% of the wire transfers approved by approver 551036 were noise. Similar results were found in Approver LOB. These findings may be useful to improve the company's internal control over their approval system.

	343375	344248	370462	489738	551036	554820	645107	…	943419	952394	953256	966385	968226	970535	990091	990637	992955
Major	31	161	11	8	79	0	57	…	37	0	0	47	96	74	685	0	0
	0.31%	1.62%	0.11%	0.08%	0.79%	0.00%	0.57%	…	0.37%	0.00%	0.00%	0.47%	0.96%	0.74%	6.88%	0.00%	0.00%
Noise	1	0	0	0	13	1	4	…	0	3	1	6	3	0	3	2	1
	2.17%	0.00%	0.00%	0.00%	28.26%	2.17%	8.70%	…	0.00%	6.52%	2.17%	13.04%	6.52%	0.00%	6.52%	4.35%	2.17%
Total	32	161	11	8	92	1	61	…	37	3	1	53	99	74	688	2	1
>10X					> 10X	> 10X	> 10X	…		> 10X	> 10X	> 10X				> 10X	> 10X
New						New				New	New					New	New

	…	1003195	1009382	1018530	1025214	1026338	1036956	Total
Major	…	10	10	39	4	8	92	9,954
	…	0.10%	0.10%	0.39%	0.04%	0.08%	0.92%	100.00%
Noise	…	1	0	0	2	0	0	46
	…	2.17%	0.00%	0.00%	4.35%	0.00%	0.00%	100.00%
Total	…	11	10	39	6	8	92	10,000
>10X	…	> 10X			> 10X			
New								

Table 16. Training set – frequency test: Approver ID

	10023709	10023755	10023756	10023759	10023761	10023762	10025098	10025286	10027036	10032440	10034880	10042280	10049070	10068041	10068049	10068051	Total
Major	11	100	8	6	363	697	0	79	3,926	45	11	202	19	108	4,161	218	9,954
	0.11%	1.00%	0.08%	0.06%	3.65%	7.00%	0.00%	0.79%	39.44%	0.45%	0.11%	2.03%	0.19%	1.08%	41.80%	2.19%	100.00%
Noise	0	0	0	0	15	4	4	14	2	0	0	1	3	0	3	0	46
	0.00%	0.00%	0.00%	0.00%	32.61%	8.70%	8.70%	30.43%	4.35%	0.00%	0.00%	2.17%	6.52%	0.00%	6.52%	0.00%	100.00%
Total	11	100	8	6	378	701	4	93	3,928	45	11	203	22	108	4,164	218	10,000
>10X							> 10X	> 10X					> 10X				
New							New										

Table 17. Training set – frequency test: Approver LOB

DBSCAN captured hidden characteristics of both the variables used in the model and those excluded from the model. Two of major obstacles to analyzing a large dataset in practice are excessive consumption of the computational resources for analysis and extensive processing time. In order to remedy this issue, this study utilized the PCA that was one of the widely used methods to reduce the dimensionality of datasets while retaining the maximal variability intrinsic to the original data (Hasan and Abdulazeez, 2021; Jolliffe and Jorge, 2016). When a company has new records, they have two options when using DBSCAN. They may run the clustering (1) with the whole dataset, including both old and new data, or (2) using only new data. The first option will produce more accurate clusters because noise this month may become clustered if similar transactions are executed next month. Unfortunately, this approach will face a computational resource issue sooner or later. Ideally, DBSCAN would be applied to the new data only but its results would still be similar to what would be found when running the whole dataset. In order to test this possibility, three DBSCAN analyses were conducted in the next section.

3.2.2. Test set

The analyses of the Test set were conducted with the following procedures. First, DBSCAN was applied to the Test set that consisted of the 10,001st to 20,000th wire transfers. Second, the Whole set (1st to 20,000th) was clustered with DBSCAN and the wire transfers that belonged to the Test set portion were extracted. Third, the Whole set was DBSCAN-clustered with the Training set parameters and then the Test set portion was extracted. The last method was to reduce the time to determine the DBSCAN parameters.

Table 18 shows the parameters and the numbers of noise transfers for each clustering. The same procedures in the Training set clustering were applied for consistent comparisons. The parameters selected for the Test set were the same as those of the Training set (eps=3 & minPts=9), while those for Whole set resulted in

the smaller eps and the larger minPts (eps=2 & minPts=10). As shown in Figure 5, the knee of the graph was nearest to 2, and ln (20,000) was 9.903 which was closer to 10.

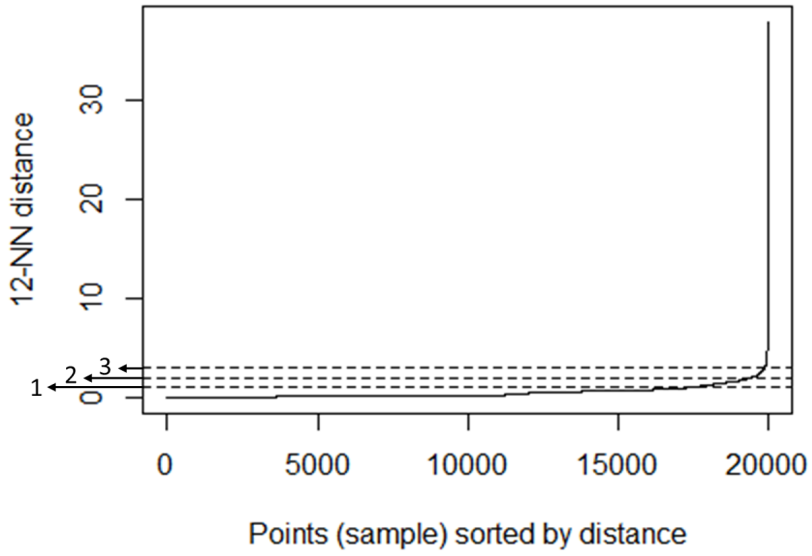

Figure 5. Whole set - kNNdistplot

With these parameters, each set was DBSCAN-clustered. The same procedures were applied to the Training set and the results were summarized in Table 18. The Standalone Test set had 34 noise wire transfers, while the number of noise wire transfers of the Whole set with eps of two and minPts of 10 was 277, of which 209 belonged to the Test set portion. With the parameters used for the Training set clustering, the Whole set resulted in 57 noise points and 47 of them were members of the Test set. The relationship between the three results was visualized with Venn diagram in Figure 6. Of the transfers that were flagged as noise, 31 were selected by all of the models, and 159 were flagged only by the Whole set with its own parameters. This result was not a surprise because the use of the stricter rule (i.e., more observations within a smaller area to form a cluster) caused more observations not to be clustered. The characteristics captured by the Standalone Test set and the Whole set with the Training set parameters were expected to be similar because their union accounted for 91.18% (= 31/34) and 65.96% (= 31/47), respectively. However, the Whole set with its own parameters might have different results because the union was only 14.83% of the total flagged noises (= 31/209).

	Parameters	Whole set (20,000 Obs)	Test set portion
Test set – Standalone	eps=3 & minPts=9	N/A	34
Whole set – Test set portion	eps=2 & minPts=10	277	209
Whole set with the Training set parameters – Test set portion	eps=3 & minPts=9	57	47

Table 18. Parameters and number of noise transfers

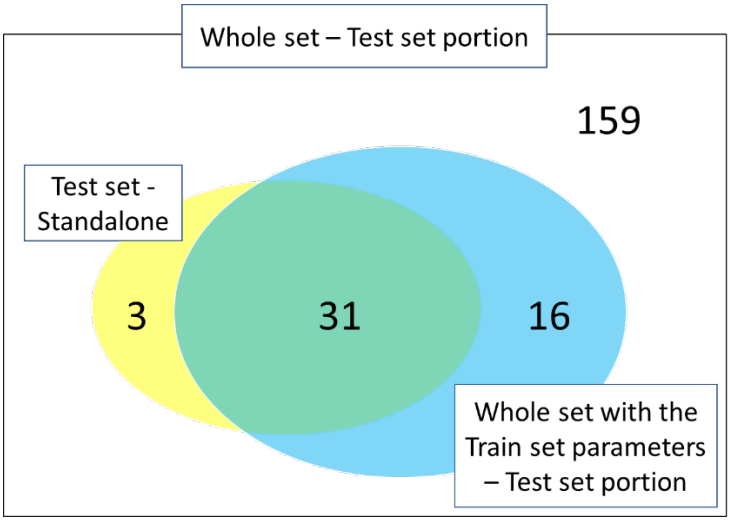

Figure 6. Numbers of noise wire transfers

The results of these three sets were compared in terms of the same set of the variables that were used for the Training set analyses for consistent comparisons.

Variables included in the model

a. Descriptive statistics

In all of the three clustering models, the major wire transfers were different from the noises in terms of all measures. As shown on Table 19, however, the values of the Whole set with its own parameters differed from the other two models. The averages and medians of noise transfers of each model showed were higher than those of clustered transfers.

	Type	_FREQ_	N	Nmiss	Min	Max	Range	Mean	Stdev	P25	P50	P75	LCLM	UCLM	Skew	Kurt
Test - Standalone	Major	9,966	9,966	-	2.8	4,093,893,198.0	4,093,893,195.0	48,393,341.5	231,833,327.0	11,525.6	51,037.1	300,138.8	43,841,194.6	52,945,488.4	7.5	73.5
	Noise	34	34	-	1,498.0	12,466,500,000.0	12,466,498,502.0	2,448,894,505.0	3,317,147,378.0	1,036,789.0	70,743,000.0	4,199,375,608.0	1,291,487,069.0	3,606,301,942.0	1.4	1.3
Whole set	Major	9,791	9,791	-	2.8	1,905,155,814.0	1,905,155,811.0	34,811,999.5	155,756,158.0	11,290.6	49,588.5	278,354.5	31,726,443.1	37,897,555.9	5.6	35.2
	Noise	209	209	-	131.6	12,466,500,000.0	12,466,499,868.0	1,075,149,129.0	1,747,219,291.0	76,569.0	17,068,807.0	1,713,905,970.0	836,886,099.0	1,313,412,159.0	2.9	12.1
Whole set - Train parameters	Major	9,953	9,953	-	2.8	2,923,437,690.0	2,923,437,687.0	45,022,901.2	207,900,535.0	11,500.8	50,906.3	300,138.8	40,938,019.2	49,107,783.2	6.5	50.6
	Noise	47	47	-	1,498.0	12,466,500,000.0	12,466,498,502.0	2,498,670,623.0	2,882,389,988.0	4,196,585.5	1,930,243,724.0	4,000,000,000.0	1,652,369,218.0	3,344,972,028.0	1.4	2.2

Table 19. Test set – descriptive statistics: Amount

b. Frequency tests

Similar to analyses using the Training set, the five variables included in the model were analyzed to capture the distinctive differences between the major and noise wire transfers. They were Day of initiation date, Day of effective date, Month of initiation date, Month of effective date, and Payee ID.

b.1. Variables with significant differences

First, all three models captured that Day of initiation date had more noise transfers initiated during non-working days (Table 20). Although the Whole set model with its own parameters flagged more noise cases, the relative ratio within the noise group was lowest among the models. This may indicate a false positive problem, which would mean that the other two models may be better choices if a company has limited resources to verify wire transfers. Another distinctive finding was that the Friday phenomenon found in the Training set did not exist in the Test set. This might imply that the behavior of the claimants changed, or the company had to limit the number of Friday wire transfers due to a shortage of employees.

		1	2	3	4	5	9	Total
Test - Standalone	Major	1,982	1,864	1,841	1,967	2,249	63	9,966
		19.89%	18.70%	18.47%	19.74%	22.57%	0.63%	100.00%
	Noise	4	6	5	3	9	7	34
		11.76%	17.65%	14.71%	8.82%	26.47%	20.59%	100.00%
	Total	1,986	1,870	1,846	1,970	2,258	70	10,000
	>10X						**> 10X**	
Whole set	Major	1,966	1,830	1,819	1,929	2,208	39	9,791
		20.08%	18.69%	18.58%	19.70%	22.55%	0.40%	100.00%
	Noise	20	40	27	41	50	31	209
		9.57%	19.14%	12.92%	19.62%	23.92%	14.83%	100.00%
	Total	1,986	1,870	1,846	1,970	2,258	70	10,000
	>10X						**> 10X**	
Whole set - Train parameters	Major	1,981	1,861	1,839	1,963	2,246	63	9,953
		19.90%	18.70%	18.48%	19.72%	22.57%	0.63%	100.00%
	Noise	5	9	7	7	12	7	47
		10.64%	19.15%	14.89%	14.89%	25.53%	14.89%	100.00%
	Total	1,986	1,870	1,846	1,970	2,258	70	10,000
	>10X						**> 10X**	

Table 20. Test set – frequency test: day of initiation date

Second, comparisons by Day of effective date also showed that non-working days had significantly higher noise ratios than those of working days (1.49% vs. 26.47% in the Test-Standalone; 1.12% vs. 22.49% in the Whole set with its own parameters; and 1.42% vs. 34.04% in the Whole set with the Training set parameters) as shown

on Table 21. However, the Monday phenomenon in the Training set was not observed in the Test set. This might be the result of a behavior change of the claimants, or possibly from shortage of initiators.

		1	2	3	4	5	9	Total
Test - Standalone	Major	2,297	2,392	1,670	1,689	1,770	148	9,966
		23.05%	24.00%	16.76%	16.95%	17.76%	1.49%	100.00%
	Noise	6	6	5	3	5	9	34
		17.65%	17.65%	14.71%	8.82%	14.71%	26.47%	100.00%
	Total	2,303	2,398	1,675	1,692	1,775	157	10,000
	>10X						**> 10X**	
Whole set	Major	2,254	2,368	1,655	1,655	1,749	110	9,791
		23.02%	24.19%	16.90%	16.90%	17.86%	1.12%	100.00%
	Noise	49	30	20	37	26	47	209
		23.44%	14.35%	9.57%	17.70%	12.44%	22.49%	100.00%
	Total	2,303	2,398	1,675	1,692	1,775	157	10,000
	>10X						**> 10X**	
Whole set - Train parameters	Major	2,297	2,392	1,668	1,686	1,769	141	9,953
		23.08%	24.03%	16.76%	16.94%	17.77%	1.42%	100.00%
	Noise	6	6	7	6	6	16	47
		12.77%	12.77%	14.89%	12.77%	12.77%	34.04%	100.00%
	Total	2,303	2,398	1,675	1,692	1,775	157	10,000
	>10X						**> 10X**	

Table 21. Test set – frequency test: day of effective date

Next, Table 22 and 22 show the comparisons of the third and fourth variables: Month of initiation date and Month of effective date. Similar to the Training set clustering result, all three models captured that the last two months at year-end had significantly higher noise ratios for both variables than other months. However, this distinction was less clear in the Whole set with its own parameters. The noise ratio in December was not significantly different from that of the major ratio (11.96% vs. 9.04%). The results for Month of effective date showed a similar result.

		1	2	3	4	5	6	7	8	9	10	11	12	Total
Test - Standalone	Major	886	626	956	926	720	911	994	667	822	882	674	902	9,966
		8.89%	6.28%	9.59%	9.29%	7.22%	9.14%	9.97%	6.69%	8.25%	8.85%	6.76%	9.05%	100.00%
	Noise	1	-	2	3	2	3	2	-	2	2	9	8	34
		2.94%	0.00%	5.88%	8.82%	5.88%	8.82%	5.88%	0.00%	5.88%	5.88%	26.47%	23.53%	100.00%
	Total	887	626	958	929	722	914	996	667	824	884	683	910	10,000
	>10X													
	>2X											**> 2X**	**> 2X**	
Whole set	Major	872	608	939	920	708	906	984	660	801	871	637	885	9,791
		8.91%	6.21%	9.59%	9.40%	7.23%	9.25%	10.05%	6.74%	8.18%	8.90%	6.51%	9.04%	100.00%
	Noise	15	18	19	9	14	8	12	7	23	13	46	25	209
		7.18%	8.61%	9.09%	4.31%	6.70%	3.83%	5.74%	3.35%	11.00%	6.22%	22.01%	11.96%	100.00%
	Total	887	626	958	929	722	914	996	667	824	884	683	910	10,000
	>10X													
	>2X											**> 2X**		

Whole set - Train parameters	Major	885	623	953	927	718	911	994	667	819	883	672	901	9,953
		8.89%	6.26%	9.58%	9.31%	7.21%	9.15%	9.99%	6.70%	8.23%	8.87%	6.75%	9.05%	100.00%
	Noise	2	3	5	2	4	3	2	-	5	1	11	9	47
		4.26%	6.38%	10.64%	4.26%	8.51%	6.38%	4.26%	0.00%	10.64%	2.13%	23.40%	19.15%	100.00%
	Total	887	626	958	929	722	914	996	667	824	884	683	910	10,000
	>10X													
	>2X											**> 2X**	**> 2X**	

Table 22. Test set – frequency test: month of initiation date

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**
Test - Standalone	Major	909	631	894	951	716	959	939	682	848	896	705	836	9,966
		9.12%	6.33%	8.97%	9.54%	7.18%	9.62%	9.42%	6.84%	8.51%	8.99%	7.07%	8.39%	100.00%
	Noise	1	-	2	3	2	3	2	-	2	1	10	8	34
		2.94%	0.00%	5.88%	8.82%	5.88%	8.82%	5.88%	0.00%	5.88%	2.94%	29.41%	23.53%	100.00%
	Total	910	631	896	954	718	962	941	682	850	897	715	844	10,000
	>10X													
	>2X											**> 2X**	**> 2X**	
Whole set	Major	895	613	876	946	704	953	930	674	827	885	673	815	9,791
		9.14%	6.26%	8.95%	9.66%	7.19%	9.73%	9.50%	6.88%	8.45%	9.04%	6.87%	8.32%	100.00%
	Noise	15	18	20	8	14	9	11	8	23	12	42	29	209
		7.18%	8.61%	9.57%	3.83%	6.70%	4.31%	5.26%	3.83%	11.00%	5.74%	20.10%	13.88%	100.00%
	Total	910	631	896	954	718	962	941	682	850	897	715	844	10,000
	>10X													
	>2X											**> 2X**		
Whole set - Train parameters	Major	908	628	891	952	714	959	939	682	845	897	703	835	9,953
		9.12%	6.31%	8.95%	9.56%	7.17%	9.64%	9.43%	6.85%	8.49%	9.01%	7.06%	8.39%	100.00%
	Noise	2	3	5	2	4	3	2	-	5	-	12	9	47
		4.26%	6.38%	10.64%	4.26%	8.51%	6.38%	4.26%	0.00%	10.64%	0.00%	25.53%	19.15%	100.00%
	Total	910	631	896	954	718	962	941	682	850	897	715	844	10,000
	>10X													
	>2X											**> 2X**	**> 2X**	

Table 23. Test set – frequency test: month of effective date

The result of the fifth variable, Payee ID, reconfirmed that the DBSCAN could pinpoint particular payees that were associated with noise wire transfers. This information could facilitate monitoring or detecting potential fraud or errors. As shown in Table 24, most of the noise wire transfers were related specific payees, many of whom were new payees. Caution must be taken when interpreting these results. The payee 20357 had no noise wire transfers in the models of Test-Standalone and Whole set with the Training set parameters. However, all of the 15 wire transfers sent to the payee were classified as noises by the model with the whole set.

		1715	**1924**	**2425**	**…**	**20357**	**20718**	**20981**	**…**	**30429**	**30430**	**30434**	**30435**	**Total**
Test - Standalone	Major	0	0	0	…	15	4	3	…	19	171	10	21	9,966
		0.00%	0.00%	0.00%	…	0.15%	0.04%	0.03%	…	0.19%	1.72%	0.10%	0.21%	100.00%
	Noise	1	1	1	…	0	0	0	…	0	0	0	1	34
		2.94%	2.94%	2.94%	…	0.00%	0.00%	0.00%	…	0.00%	0.00%	0.00%	2.94%	100.00%
	Total	1	1	1	…	15	4	3	…	19	171	10	22	10,000
	>10X	**> 10X**	**> 10X**	**> 10X**	…				…				**> 10X**	
	New	**New**	**New**	**New**	…				…					
Whole set	Major	0	0	0	…	0	3	2	…	19	170	10	21	9,791
		0.00%	0.00%	0.00%	…	0.00%	0.03%	0.02%	…	0.19%	1.74%	0.10%	0.21%	100.00%
	Noise	1	1	1	…	15	1	1	…	0	1	0	1	209
		0.48%	0.48%	0.48%	…	7.18%	0.48%	0.48%	…	0.00%	0.48%	0.00%	0.48%	100.00%
	Total	1	1	1	…	15	4	3	…	19	171	10	22	10,000
	>10X	**> 10X**	**> 10X**	**> 10X**	…	**> 10X**	**> 10X**	**> 10X**	…					
	New	**New**	**New**	**New**	…	**New**			…					

Whole set - Train parameters	Major	0	0	0	...	15	4	3	...	19	171	10	21	9,953
		0.00%	0.00%	0.00%	...	0.15%	0.04%	0.03%	...	0.19%	1.72%	0.10%	0.21%	100.00%
	Noise	1	1	1	...	0	0	0	...	0	0	0	1	47
		2.13%	2.13%	2.13%	...	0.00%	0.00%	0.00%	...	0.00%	0.00%	0.00%	2.13%	100.00%
	Total	1	1	1	...	15	4	3	...	19	171	10	22	10,000
	>10X	**> 10X**	**> 10X**	**> 10X**	...				...				**> 10X**	
	New	**New**	**New**	**New**	...				...					

Table 24. Test set – frequency test: Payee ID

b.2. Variables with insignificant differences

As mentioned, some variables in the DBSCAN model did not present meaningful differences between the major and the noise groups. As with the Training set, the clustered and noise wire transfers had no clear behavioral differences in Day of month of initiation date and Date of month of effective date.

Variables excluded from the model

As performed in the Training set model, this section illustrates that all three models captured differences in characteristics of variables not included in the models. Analyses of four variables not in the models showed meaningful implications that could be used for anomaly monitoring and detection.

a. Descriptive Statistics

Consistent with the Training set model, differences in Initiator's Authorization limit and Approver's Authorization limit were not significant due to extremely high, consistent authorization limits. The high authorization limits might have been set for a practical reason; it would be inconvenient to ask a senior whenever there was a wire transfer that exceeded their authorization limits. Instead, they might initiate the wire transfer because the approvers would further review it.

b. Frequency Tests

Analysis of the remaining four variables that were not used for the clustering models showed that there were strong associations between noise/cluster status and each variable. Comparisons of Initiator ID in Table 25 showed presence of specific initiators that tended to post noise wire transfers. As expected, the Test-Standalone and the Whole set with the Train parameters flagged a similar set of initiators while the Whole set showed different results. For example, two wire transfers from initiator 645860 were flagged by all three models, but one noisy transfer from initiator 645640 was detected only by the Whole set with its own parameters (Table 25).

Initiator LOB analyses uncovered that specific Initiator LOBs were related to noise

wire transfers. Although this finding showed a significant association between the initiator LOB and the noise wire transfers, its usefulness was highly limited because the majority of the wire transfers of the Test set were initiated by a single LOB, 10023761. As shown in Table 26, it initiated 91.46% of the total wire transfers in the Test set and the majority of noise wire transfers were also associated with it. This observation, however, does not mean that the relationship between the LOB and the noise wire transfers is not meaningful. Some LOBs had greater propensity for noise than others, which could be used at the initial stage of anomaly detection.

		374590	375727	531541	645640	645860	…	714368	718031	934834	…	5074149	JM	Total
Test - Standalone	Major	2	1	5	2	2	…	2	6	13	…	31	1	9,966
		0.02%	0.01%	0.05%	0.02%	0.02%	…	0.02%	0.06%	0.13%	…	0.31%	0.01%	100.00%
	Noise	0	0	0	0	2	…	0	0	3	…	0	0	34
		0.00%	0.00%	0.00%	0.00%	5.88%	…	0.00%	0.00%	8.82%	…	0.00%	0.00%	100.00%
	Total	2	1	5	2	4	…	2	6	16	…	31	1	10,000
	>10X					**> 10X**	…			**> 10X**	…			
	New						…				…			
Whole set	Major	2	1	5	1	2	…	0	6	13	…	31	1	9791
		0.02%	0.01%	0.05%	0.01%	0.02%	…	0.00%	0.06%	0.13%	…	0.32%	0.01%	100.00%
	Noise	0	0	0	1	2	…	2	0	3	…	0	0	209
		0.00%	0.00%	0.00%	0.48%	0.96%	…	0.96%	0.00%	1.44%	…	0.00%	0.00%	100.00%
	Total	2	1	5	2	4	…	2	6	16	…	31	1	10,000
	>10X				**> 10X**	**> 10X**	…	**> 10X**		**> 10X**	…			
	New						…	**New**			…			
Whole set - Train parameters	Major	2	1	5	2	2	…	2	6	13	…	31	1	9953
		0.02%	0.01%	0.05%	0.02%	0.02%	…	0.02%	0.06%	0.13%	…	0.31%	0.01%	100.00%
	Noise	0	0	0	0	2	…	0	0	3	…	0	0	47
		0.00%	0.00%	0.00%	0.00%	4.26%	…	0.00%	0.00%	6.38%	…	0.00%	0.00%	100.00%
	Total	2	1	5	2	4	…	2	6	16	…	31	1	10,000
	>10X					**> 10X**	…			**> 10X**	…			
	New						…				…			

Table 25. Test set – frequency test: Initiator ID

		10021241	10023527	10023755	10023756	10023761	10023762	10025652	10025815	10031040	10049070	10084210	5025814	5025815	85025814	85025815	Total
Test - Standalone	Major	155	57	3	3	9,127	384	41	22	5	9	13	111	11	24	1	9,966
		1.56%	0.57%	0.03%	0.03%	91.58%	3.85%	0.41%	0.22%	0.05%	0.09%	0.13%	1.11%	0.11%	0.24%	0.01%	100.00%
	Noise	3	0	0	0	19	3	1	0	0	2	0	4	0	2	0	34
		8.82%	0.00%	0.00%	0.00%	55.88%	8.82%	2.94%	0.00%	0.00%	5.88%	0.00%	11.76%	0.00%	5.88%	0.00%	100.00%
	Total	158	57	3	3	9,146	387	42	22	5	11	13	115	11	26	1	10,000
	>10X										> 10X		> 10X		> 10X		
	New																
Whole set	Major	155	46	3	2	8989	382	30	20	5	6	13	106	11	22	1	9791
		1.58%	0.47%	0.03%	0.02%	91.81%	3.90%	0.31%	0.20%	0.05%	0.06%	0.13%	1.08%	0.11%	0.22%	0.01%	100.00%
	Noise	3	11	0	1	157	5	12	2	0	5	0	9	0	4	0	209
		1.44%	5.26%	0.00%	0.48%	75.12%	2.39%	5.74%	0.96%	0.00%	2.39%	0.00%	4.31%	0.00%	1.91%	0.00%	100.00%
	Total	158	57	3	3	9,146	387	42	22	5	11	13	115	11	26	1	10,000
	>10X		> 10X		> 10X			> 10X			> 10X						
	New																
Whole set - Train parameters	Major	155	57	3	3	9113	386	41	22	5	9	13	111	11	23	1	9953
		1.56%	0.57%	0.03%	0.03%	91.56%	3.88%	0.41%	0.22%	0.05%	0.09%	0.13%	1.12%	0.11%	0.23%	0.01%	100.00%
	Noise	3	0	0	0	33	1	1	0	0	2	0	4	0	3	0	47
		6.38%	0.00%	0.00%	0.00%	70.21%	2.13%	2.13%	0.00%	0.00%	4.26%	0.00%	8.51%	0.00%	6.38%	0.00%	100.00%
	Total	158	57	3	3	9,146	387	42	22	5	11	13	115	11	26	1	10,000
	>10X										> 10X				> 10X		
	New																

Table 26. Test Set – Frequency Test: Initiator LOB

The last two variables in comparison were Approver ID (Table 27) and Approver LOB (Table 28). As observed in the Training set model, the three models showed

that specific approvers and approver LOBs were more prone to making noise wire transfers than others. The three models also identified the association between specific Approver LOBs and noise wire transfers. Similar to Initiator LOB, however, noise was concentrated on specific LOBs. For example, LOBs 10023759 and 10023761 accounted for 89.26% of the total wire transfers and 67.65% of the noise wire transfers so that caution must be taken in using the association between the noise wire transfers and Approver LOBs for monitoring and detecting anomalies.

		375727	382749	...	714368	815326	822867	936205	940579	963247	...	1025214	1025306	1026338	1036956	Total
Test - Standalone	Major	7	119	...	2	43	15	8	4	89	...	7	41	2	28	9,966
		0.07%	1.19%	...	0.02%	0.43%	0.15%	0.08%	0.04%	0.89%	...	0.07%	0.41%	0.02%	0.28%	100.00%
	Noise	1	0	...	0	3	0	0	0	6	...	0	1	0	0	34
		2.94%	0.00%	...	0.00%	8.82%	0.00%	0.00%	0.00%	17.65%	...	0.00%	2.94%	0.00%	0.00%	100.00%
	Total	8	119	...	2	46	15	8	4	95	...	7	42	2	28	10,000
	>10X	> 10X		...		> 10X				> 10X	...					
	New			...							...					
Whole set	Major	5	119	...	1	43	14	2	4	74	...	6	30	2	28	9,791
		0.05%	1.22%	...	0.01%	0.44%	0.14%	0.02%	0.04%	0.76%	...	0.06%	0.31%	0.02%	0.29%	100.00%
	Noise	3	0	...	1	3	1	6	0	21	...	1	12	0	0	209
		1.44%	0.00%	...	0.48%	1.44%	0.48%	2.87%	0.00%	10.05%	...	0.48%	5.74%	0.00%	0.00%	100.00%
	Total	8	119	...	2	46	15	8	4	95	...	7	42	2	28	10,000
	>10X	> 10X		...	> 10X			> 10X		> 10X	...		> 10X			
	New			...							...					
Whole set - Train parameters	Major	7	119	...	2	43	15	8	4	87	...	7	41	2	28	9,953
		0.07%	1.20%	...	0.02%	0.43%	0.15%	0.08%	0.04%	0.87%	...	0.07%	0.41%	0.02%	0.28%	100.00%
	Noise	1	0	...	0	3	0	0	0	8	...	0	1	0	0	47
		2.13%	0.00%	...	0.00%	6.38%	0.00%	0.00%	0.00%	17.02%	...	0.00%	2.13%	0.00%	0.00%	100.00%
	Total	8	119	...	2	46	15	8	4	95	...	7	42	2	28	10,000
	>10X	> 10X		...		> 10X				> 10X	...					
	New			...							...					

Table 27. Test set – frequency test: Approver ID

		10021241	10021854	10023437	10023527	10023755	10023756	10023759	10023761	10023762	10024524	10025652	10031040	10049070	10084110	10084170	10086250	Total
Test - Standalone	Major	1	43	1	56	30	11	1,035	7,868	732	4	41	5	15	119	4	1	9,966
		0.01%	0.43%	0.01%	0.56%	0.30%	0.11%	10.39%	78.95%	7.34%	0.04%	0.41%	0.05%	0.15%	1.19%	0.04%	0.01%	100.00%
	Noise	0	3	0	0	0	0	12	11	4	0	1	0	3	0	0	0	34
		0.00%	8.82%	0.00%	0.00%	0.00%	0.00%	35.29%	32.35%	11.76%	0.00%	2.94%	0.00%	8.82%	0.00%	0.00%	0.00%	100.00%
	Total	1	46	1	56	30	11	1,047	7,879	736	4	42	5	18	119	4	1	10,000
	>10X		> 10X											> 10X				
	New																	
Whole set	Major	1	43	1	45	28	10	952	7808	728	4	30	5	12	119	4	1	9,791
		0.01%	0.44%	0.01%	0.46%	0.29%	0.10%	9.72%	79.75%	7.44%	0.04%	0.31%	0.05%	0.12%	1.22%	0.04%	0.01%	100.00%
	Noise	0	3	0	11	2	1	95	71	8	0	12	0	6	0	0	0	209
		0.00%	1.44%	0.00%	5.26%	0.96%	0.48%	45.45%	33.97%	3.83%	0.00%	5.74%	0.00%	2.87%	0.00%	0.00%	0.00%	100.00%
	Total	1	46	1	56	30	11	1,047	7,879	736	4	42	5	18	119	4	1	10,000
	>10X				> 10X							> 10X		> 10X				
	New																	
Whole set - Train parameters	Major	1	43	1	56	30	11	1022	7866	734	4	41	5	15	119	4	1	9,953
		0.01%	0.43%	0.01%	0.56%	0.30%	0.11%	10.27%	79.03%	7.37%	0.04%	0.41%	0.05%	0.15%	1.20%	0.04%	0.01%	100.00%
	Noise	0	3	0	0	0	0	25	13	2	0	1	0	3	0	0	0	47
		0.00%	6.38%	0.00%	0.00%	0.00%	0.00%	53.19%	27.66%	4.26%	0.00%	2.13%	0.00%	6.38%	0.00%	0.00%	0.00%	100.00%
	Total	1	46	1	56	30	11	1,047	7,879	736	4	42	5	18	119	4	1	10,000
	>10X		> 10X											> 10X				
	New																	

Table 28. Test set – frequency test: Approver LOB

To summarize, all three models identified the significant characteristics of each variable both included and excluded with different degrees of efficiency. The model with the Whole set with its own parameters flagged significantly more noise wire transfers. However, this might imply that the model generated more false positives.

Considering the limited resources that a company can allocate to monitoring and detecting anomalies, this model may not be the best choice for practitioners. However, it may be most suitable if more thorough investigation is needed for fraud detection. Since the other two models uncover similar patterns of the variables in this study, applying the DBSCAN model directly to the Test set may be the better option.

4. CONCLUSION AND LIMITATIONS

This study introduced practitioner-friendly clustering DBSCAN analyses that monitor and detect anomalies in their transactional data. The proposed anomaly detection models demonstrated their capability to uncover hidden anomalous relationships in the transactions without prior knowledge about the data being used. DBSCAN analyses that had minimal requirements of domain knowledge to determine the input parameters could be advantageous for practitioners who often lack in-depth knowledge about their company's fraudulent transactions and their behavioral patterns (Ester et al., 1996). The subsequent analyses of the Train and Test sets also affirmed the reliability and consistency of DBSCAN as an anomaly detection method.

This study, however, has several limitations. First, the DBSCAN parameters used for the models might not be optimal. Determination of clustering parameters in applying other clustering methods requires prior knowledge about the anomalies in the dataset clustering method algorithm. However, this often serves as a barrier for practitioners attempting to develop fraud and error monitoring methods. The DBSCAN parameter selection process requires minimal human intervention, making it more approachable for practitioners in anomaly detection and monitoring. However, there could be a tradeoff between easy application and accuracy. The parameter selection methods used in future studies may utilize more sophisticated criteria. For example, instead of integer values for the eps, other rational numbers, such as 2.5, may be used to select the knee in a kNNDist plot.

Another limitation of this study was caused by the limited computational power of the computer used to execute the DBSCAN code. Due to the limited computational power, this study used only 20,000 wire transfers with 12 starting variables (cut to 9 principal components via PCA). Although the resulting principal components accounted for the majority of the variance (e.g., 97.91% in the Training set model), it was obvious that the dimensionality reduction led to a loss of variance of the

variables in the models. The DBSCAN models would have produced better results if raw data without conversion had been used. A future study could make use of a computer equipped with the enhanced computational capabilities for DBSCAN modeling to overcome this limitation.

Lastly, the modeling in this study assumed that fraud detection activities took place at regular intervals of a fixed number of transactions. In practice, however, a fraud detection model might be activated at preset time periods like monthly or quarterly intervals. This alteration could yield different outcomes, revealing additional relationships that remained undiscovered in this study. A future study could also consider potential seasonality that might influence the transaction volumes.

6. REFERENCES

ACFE (Association of Certified Fraud Examiners). (2022). Occupational Fraud 2022: A Report to the Nations. *ACFE*. https://legacy.acfe.com/report-to-the-nations/2022/ Accessed 21 April 2024.

Bolton, R. J., & Hand, D.J. (2001). Unsupervised profiling methods for fraud detection. *Credit Scoring and Credit Control*. VII, 235-255. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5b640c367ae9cc4bd072006b05a3ed7c2d5f496d. Accessed 21 April 2024.

Bolton, R. J., & Hand, D.J. (2002). Statistical Fraud Detection: A Review. *Statistical Science*. 17(3), 235-249. https://doi.org/10.1214/ss/1042727940.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Computing Surveys*, 41, 1-58. https://doi.org/10.1145/1541880.1541882.

Ester, M., Kriegel, H., Sander, J., & Xu, X. A. (1996). Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD*, 96, 226-231. https://dl.acm.org/doi/10.5555/3001460.3001507.

Freiman, J. W., Kim, Y., & Vasarhelyi, M.A. (2022). Full population testing: Applying multidimensional audit data sampling (MADS) to general ledger data auditing. *International Journal of Accounting Information Systems*, 46. https://doi.org/10.1016/j.accinf.2022.100573

Hasan, B.M.S, & Abdulazeez A.M. (2021). A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. *Journal of Soft Computing and Data Mining*, 2, 20-30. https://doi.org/10.30880/jscdm.2021.02.01.003

Jolliffe, I.T., & Jorge, C. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202 https://doi.org/10.1098/rsta.2015.0202.

Khan, K., Rehman, S.U., Aziz, K., Fong, & Sarasvady, S. (2014). DBSCAN: Past, Present, and Future. *The Fifth International Conference on the Application of Digital Information and Web Technologies*, 232-238 https://doi.org/10.1109/icadiwt.2014.6814687.

Kim, Y, & Kogan, A. (2014). Development of an Anomaly Detection Model for a Bank's Transitory Account System. *Journal of Information Systems*, 28, 145-165. https://doi.org/10.2308/isys-50699.

Kim, Y., & Vasarhelyi, M.A. (2012). A Model to Detect Potentially Fraudulent/Abnormal Wires of an Insurance Company: An Unsupervised Rule-Based Approach. *Journal of Emerging Technologies in Accounting*, 9, 95-110. https://doi.org/10.2308/jeta-50411.

Kogan, A., Sudit, E.F., & Vasarhelyi, M.A. (1999). Continuous online auditing: A program of research. *Journal of Information Systems*, 13, 87–103. https://doi.org/10.2308/jis.1999.13.2.87.

Kou, Y., Lu., C., Sirwongwattana, S., & Huang, Y. (2004). Survey of Fraud Detection Techniques. *IEEE International Conference on Networking, Sensing and Control*, 2, 749-754. https://doi.org/10.1109/ICNSC.2004.1297040.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A Comprehensive Survey of Data Mining Based Fraud Detection Research. *arXiv preprint arXiv:1009.6119.* https://doi.org/10.48550/arXiv.1009.6119.

Liu, Q., & Vasarhelyi, M.A. (2013). Healthcare Fraud Detection: A Survey and a Clustering Model Incorporating Geo-Location Information. *29th World Continuous Auditing and Reporting Symposium. Brisbane, Australia*. http://raw.rutgers.edu/docs/wcars/29wcars/Health%20care%20fraud%20detection%20A%20survey%20and%20a%20clustering%20model%20incorporating%20Geo-location%20information.pdf Accessed 21 April 2024.

Major, J. A., & Riedinger, D.R. (2002). EFD: A hybrid knowledge/statistical-based system for the detection of fraud. *The Journal of Risk and Insurance*, 69, 309–324. https://doi.org/10.1111/1539-6975.00025.

Murthy, U.S. (2004). An analysis of the effects of continuous monitoring controls on e-commerce system performance. *Journal of Information Systems*, 18, 29–47. https://doi.org/10.2308/jis.2004.18.2.29.

Murthy, U.S., & Groomer, M.S. (2004). A continuous auditing web services model for XML-based accounting systems. *International Journal of Accounting Information Systems*, 5, 139–163. https://doi.org/10.1016/j.accinf.2004.01.007.

Rezaee, Z., Sharbatoghlie, A., Elam, R., & McMickle, P.L. (2002). Continuous auditing: Building automated auditing capability. Au*diting: A Journal of Practice & Theory*, 21, 147–163. https://doi.org/10.2308/aud.2002.21.1.147.

Sabau, A.S. (2012). Survey of Clustering Based Financial Fraud Detection Research. Informatica Economica. http://revistaie.ase.ro/content/61/10%20-%20sabau.pdf Accessed 21 April 2024.

Sheridan, K., Puranik, T.G., Mangortey, E., Pinon-Fischer, O.J., Kirby, M., & Mavris, D.N. (2020). An application of DBSCAN clustering for flight anomaly detection during the approach phase. *AIAA Scitech 2020 Forum*, 1851. https://doi.org/10.2514/6.2020-1851.

Tatusch, M., Klassen, G., Bravidor, M., & Conrad, S. (2020). Predicting Erroneous Financial Statements Using a Density-Based Clustering Approach. *The 4th International Conference on Business and Information Management*, 89-94 https://doi.org/10.1145/3418653.3418673.

Thiprungsri, S., & Vasarhelyi, M.A. (2011). Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach. *The International Journal of Digital Accounting Research*, 11, 69 – 84. https://doi.org/10.4192/1577-8517-v11_4.

Vasarhelyi, M.A., & Halper, F.B. (1991). The continuous audit of online systems. *Auditing: A Journal of Practice & Theory*, 19, 110–125. https://www.researchgate.net/publication/255667612_The_Continuous_Audit_of_Online_Systems Accessed 21 April 21, 2024.

Woodroof, J., & Searcy, D. (2001). Continuous audit implications of Internet technology: Triggering agents over the web in the domain of debt covenant compliance. *The 34th Hawaii International Conference on System Sciences*, 8-pp. https://doi.org/10.1109/hicss.2001.927080.