Spending Segmentation and Outlier Detection in Brazilian Elections

L. G. C. Simões, F. A. N. Verri, T. Yoneyama

Instituto Tecnológico da Aeronáutica, São José dos Campos, SP, Brazil
leandro.simoes@ga.ita.br, verri@ita.br, takashi@ita.br

Abstract. The political campaigns in Brazilian elections are mostly financed by public money. Every candidate has to provide detailed accountability reports to the legal authorities, which must be analyzed in a short time frame in search of eventual fraud or suspicious transactions. In this work we have compiled a real data set from 2016 Brazilian elections for all city councils in the São Paulo state and used it to propose a framework of data segmentation analysis and validation. An exploratory data analysis is performed to determine the features distribution and to define the required feature pre-processing tasks. A clustering analysis using DBSCAN method is applied to a subset of the original data, focused on segmenting the spending data regarding contracts with car fuel providers and detecting potential outliers. Three clusters were identified and a ridge regression model was used to evaluate the most important features on cluster definition. One cluster was related to candidates that received zero votes and the remaining two discriminated suppliers if they had or not contracts almost exclusively related to candidate spending on car fuel. The hyperparameters from the clustering analysis were validated using a bootstrap method and a null hypothesis of data set structure randomness was rejected using a Monte Carlo approach.

CCS Concepts: • **Computing methodologies** → **Cluster analysis**.

Keywords: Clustering algorithms, fraud detection, machine learning, outlier detection

1. INTRODUCTION

The public campaigns for executive and legislative seats in the Brazilian elections are majorly financed by government budget. In a single election year, the global public spending can reach US$ 1 billion in a time frame shorter than six months. After the elections, all candidates must submit their accountability reports, which should be analyzed and approved by the Brazilian Superior Electoral Court in less than two months, prior to the elected candidates may take office.

Not only because of the high amount of public money involved, efficient methods for fraud detection during the campaign are mandatory to enable a democratic election result. Given the short time available and the responsibility while approving each report, such methods are fundamental to help the Court to effectively focus the scarce manpower available for this task.

1.1 Works on Fraud Detection

The use of automated methods to detect fraud is widespread in the financial system [West and Bhattacharya 2016; Sharma and Kumar Panigrahi 2012], with banks and credit card companies performing online checks if transactions are suspect or legitimate. Although a complex task, the long and constantly increasing history of transactions provide the major financial companies with a resourceful database of labeled entries, which enables supervised learning tasks.

Other areas that have seen an increased usage of fraud detection methods are reimbursement claims in healthcare [Hillerman et al. 2017; Ekin et al. 2018; Carvalho et al. 2017] and collusion detection in public procurements [Van Erven et al. 2017; Baldomir et al. 2018], however few examples labeled as fraud are available both due to the difficulty in experts manually identifying suspect cases as well as possibly long trials until the final accusatory sentence in prosecutions. This motivates the application of unsupervised methods in problems not related to the financial sector.

One of the methods that have been used for unsupervised fraud detection are Self-Organizing Maps (SOMs) [Kohonen 1990]. They are a type of artificial neural network that, motivated by some sensory processing done by the human brain, is able to map its input into a low-dimensional representation (usually two-dimensional). By grouping similar input samples close together, further methods can be used to identify dissimilar samples as possible outliers. This can be done by using clustering methods [López-Iturriaga and Sanz 2018], or by directly identifying [Olszewski et al. 2013] map regions with high values in the U-matrix, a matrix that represents the dissimilarity between each neuron and its neighbors.

A similar strategy can be used with a more complex unsupervised dimension reduction technique such as autoencoders, a deep neural network architecture that is trained to output the same data as its input. By symmetrically arranging the number of neurons in each layer, the central (latent) layer is designed with very few neurons so they can represent the input data into this low-dimensional space. The difference between the input and the output is defined as the autoencoder reconstruction error, which tends to be higher for uncommon or outlier training samples. By running a k-means clustering method on a subset containing all samples above a reconstruction error threshold, [Zamini and Montazer 2018] was able to better identify outliers in a credit card fraud dataset. Another work [Amarbayasgalan et al. 2018] has used the density-based clustering algorithm DBSCAN [Ester et al. 1996] to find clusters in the complete dataset on the low-dimensional space, defining clusters as outliers in case their members exceeded a certain reconstruction error threshold. Although powerful to capture non-linear relationships in this dimension reduction mapping, its use may reduce the model results explainability. This limitation was tackled by using hierarchical results of a clustering method applied to the autoencoder output to better identify and explain different types of fraud [Kim et al. 2019].

Unsupervised anomaly detection methods are another option to detect fraud on datasets without true labels. They can be roughly classified into nearest-neighbor based, clustering based or statistical methods, varying the metrics and methodologies used to define the dissimilarity between samples. By setting a threshold, each method can identify which samples are likely to be outliers. A review work [Goldstein and Uchida 2016] has compared 19 anomaly detection models applied to 10 datasets. Another method variant [Zhang and He 2017] was able to identify multiple types of fraud in a medicare dataset using a modified Local Outlier Factor [Breunig et al. 2000] method that penalizes samples pertaining to small clusters identified by DBSCAN.

Due to the intrinsic relational nature between different entities in some problems (i.e. companies and public entities in procurements), such databases can be modeled as graph networks. In order to make use of relations between entities in the learning process, some methods consider the graph structure in order to improve the model accuracy or computational cost [Van Erven et al. 2017].

Effective results were obtained [Yan et al. 2019] when applying clustering methods tied to an exploratory data analysis. When critically inspecting the processed data with support from expert personnel, such models were able to confirm already known suspects and also identify novel suspicious cases [Carvalho et al. 2017].

1.2 Our Approach

This study proposes methods to ease and automate the investigation of suspect spendings of candidates to the Brazilian city councils in the 2016 elections, considering the public database of spending

reports provided by the Brazilian Superior Electoral Court, which includes information related to the candidates and their service providers for electoral campaign.

The unsupervised clustering algorithm DBSCAN [Ester et al. 1996] was chosen for this study, presenting a compromise between affordable computational costs and proper identification of clusters with different density and non-spherical shapes, as found in the current data set. It is also capable of identifying outlier points, which may point as suspect transactions.

The model effectiveness is assessed by using clustering metrics, intra- and inter-cluster variables similarity and visual inspection after applying the non-linear dimension reduction algorithm t-distributed Stochastic Neighbor Embedding (t-SNE) [Maaten and Hinton 2008]. This method allows a multi-dimensional data set to be projected onto a 2-D space while trying to preserve non-linear relationships among their original dimensions. A validation of clustering hyperparameters is performed, along with a hypothesis test on randomness of data structure regarding cluster assignment.

2. DATA SET DESCRIPTION

The data set used in this work is based on the public repository[1] that lists all services contracted and people hired by candidates to the city councils of São Paulo state during the 2016 Brazilian elections. Each entry represents the unique aggregation of contracts between a given candidate and a given company or person, resulting in 359,453 entries between 79,363 candidates and 158,361 service providers, covering all 645 cities in São Paulo state. The 42 features under analysis can be classified into candidate, service provider and contract features. They include information related to candidate spending and received votes, supplier revenue from candidates and contract information.

3. EXPLORATORY DATA ANALYSIS AND PRE-PROCESSING

This section performs an Exploratory Data Analysis (EDA) on the data set presented in the previous section, defining additional transformations applied to the data set. Because each entry comprises of an aggregation of contracts and includes features related to the candidate and service provider, the EDA of these features must consider only unique information. This is important to remove bias due to the number of entries a candidate or provider has, otherwise such an entity present in a large number of contract entries would be counted multiple times.

The original 65 spending categories were aggregated into 10 features due to their similarity and are segregated into three perspectives (candidate, supplier and contract), each using a different reference value for normalization. They are defined as ratios in the interval $[0, 1]$, where all category ratios from a given entry and perspective sum to unity. Such pre-processing does not make a distinction if a contract was a donation to the candidate or a spending. Fig. 1 presents an overview of two typical category spending ratio histograms, for categories `Advertising` and `Car Rental`. For contract and supplier categories, most values are close to zero, seconded by values close to one (which means that contracts and suppliers are specialized in mostly one spending category), while values between them may have a uniform or modal distribution. Candidate spending categories behave differently, usually presenting a right-skewed distribution (meaning candidates spend in multiple categories, as expected), with higher values close to zero. The only unique distribution of a candidate category is `Advertising`, which presents a much more uniform histogram and also higher counts in a spending ratio of one (meaning that a large part of candidates spend only on advertising).

Many machine learning methods depend on data distribution assumptions, usually requiring or, at least, favoring normally-distributed data. A detailed analysis of all numerical features was performed

[1]Electoral data repository from Brazilian Superior Electoral Court,
http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1/repositorio-de-dados-eleitorais

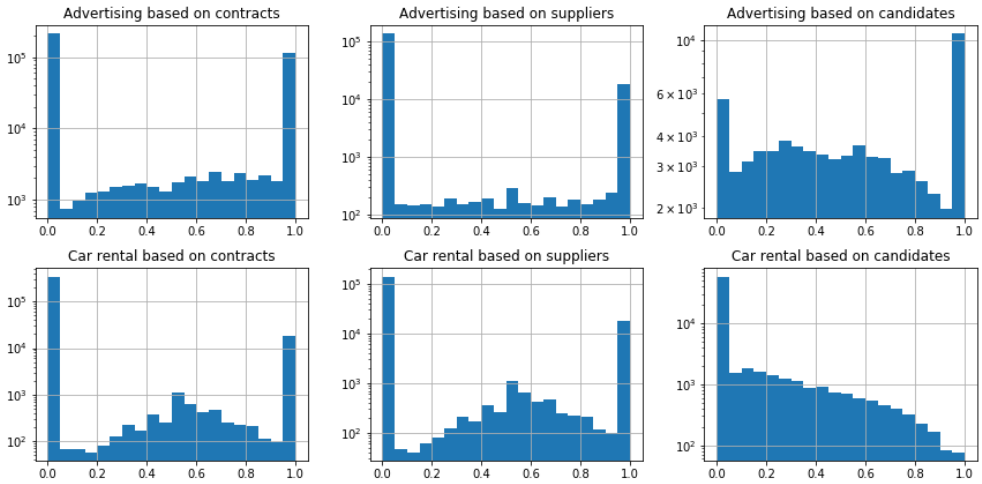

Fig. 1. Histogram of category ratios for Advertising and Car Rental categories, on all three subsets.

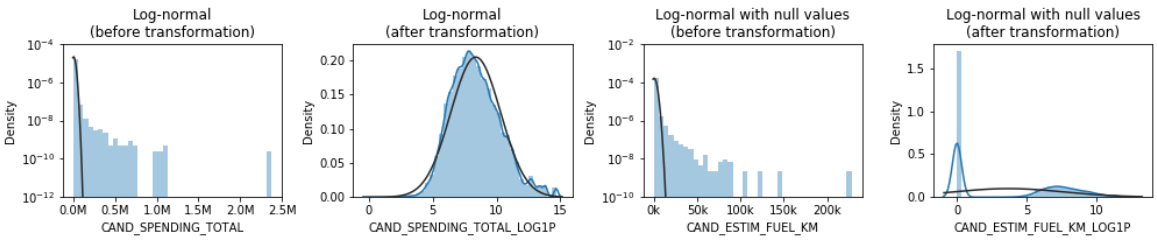

Fig. 2. Numerical features with log-normal distributions before and after log-transformation, exemplified for features with different presence of null values in the histogram. Blue bars represent the histogram frequency count, black line represents the fitted Gaussian distribution, and blue line represents the KDE.

and those identified as log-normal were transformed to a normal distribution by applying the log-transformation

$$T(p) = \ln(1 + p), \tag{1}$$

where p is the value to be transformed. Because $\ln(0)$ is undefined and $\ln(p)$ for $p \to 0$ may lead to numerical precision issues, the value to be log-transformed is added by one to avoid such problems. Fig. 2 exemplifies the Kernel Density Estimation (KDE) before and after log-transformation. For pure log-normal distributions such as `CAND_SPENDING_TOTAL`, a Gaussian distribution fits the data very well after transformation. Transformation of log-normal variables with a large quantity of null values (such as `CAND_ESTIM_FUEL_KM`) allows a reasonably good normal fit of the non-null values, however including the null values may lead to a wrong fit (as shown by the black line in Fig. 2).

Finally, all features with numerical values have their means centered at zero and are scaled in order to have a unity standard deviation. This allows a more balanced distance calculation among features with different scales, resulting in a better clustering.

4. MODEL DESCRIPTION AND CLUSTERING RESULTS

Let $\mathcal{X}_{train} = \{\vec{x}_1, \ldots, \vec{x}_M\}$ be the data set to be used as input in clustering model training, consisting of M feature vectors $\vec{x}_i \in \mathbb{R}^N$. Our goal is to assign data points from $\mathcal{X}_{train}$ to clusters that maximize inter-cluster and minimize intra-cluster variance.

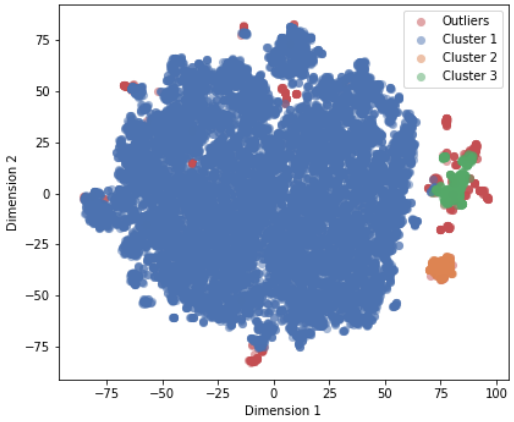

Fig. 3. Two-dimensional representation of $\mathcal{X}_{train}$ using t-SNE transformation, colored by cluster assignment C using DBSCAN algorithm.

From the pre-processed data set defined in section 3, the clustering analysis is performed by filtering out rows related to contracts or candidates with no spending in car fuel category. This results in a data set $\mathcal{X}_{train}$ with $M = 16541$ contract entries to be analyzed, or 4.6% of the complete data set. All $N = 42$ numerical features are considered in this analysis.

In order to run the DBSCAN clustering algorithm [Ester et al. 1996], two hyperparameters shall be tuned: the minimum number of points m_p in the neighborhood of a point so it can be considered a core point, and the maximum distance ϵ between two points so they can be considered in the same neighborhood. There are many heuristics suggested for defining m_p, such as $m_p = \ln(M)$ or $m_p \geq N + 1$. Using a small m_p may increase the identification of smalls clusters so, given the data set characteristics, the more conservative second approach was used to define $m_p = 45$.

After setting m_p, the parameter ϵ shall be carefully fine tuned: if set too small, most of the data will be considered as outliers and will not be assigned to any cluster; if set too large, multiple clusters of interest can merge into a single one. Because it is very dependent on the data set and distance metric used for clustering, the distance from each sample point to its k-th neighbor is sorted and plotted, considering $k = m_p - 1$. The ideal ϵ will be at the location where its value begins to increase rapidly, suggesting $\epsilon = 5$ as a good parameter.

Considering the hyperparameters $m_p = 45$ and $\epsilon = 5$, the DBSCAN algorithm has found a clustering partition C with three clusters using the Euclidean distance metric. All sample points were transformed into a two-dimensional space using the t-SNE method and colored by their cluster assignment, as shown in Fig. 3. It shows that the majority of the sample points are assigned to cluster 1, while two smaller clusters are clearly defined. The t-SNE transformation shows an apparent cluster in the top region, however DBSCAN was not able to separate it into a new cluster. Note that outliers are also identified, however they are mostly limited to the cluster surroundings, particularly to cluster 3.

In order to assess which features in $\mathcal{X}_{train}$ were most determinant to distinguish clusters in partitioning C, a set of Ridge regression models was trained to predict each cluster assignment based on $\mathcal{X}_{train}$, excluding the outlier group. For each Ridge model prediction, its coefficients were normalized by their sum and averaged over multiple runs by varying the Ridge model regularization parameter on a log scale between 10^{-15} and 10^{-6}. Finally, the range of each coefficient is calculated over the averaged coefficients for all cluster assignments. Fig. 4 compare boxplots for all clusters and outliers over four features with the largest coefficient ranges from the Ridge regression analysis. Features related to received and sent donations appear among the most important ones, however they have been

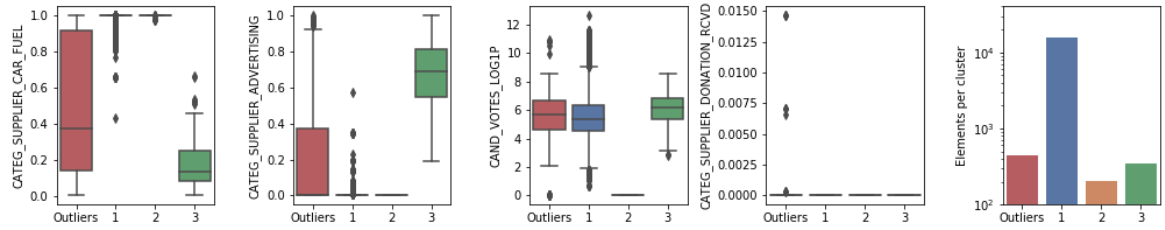

Fig. 4. Boxplots for the most important features to describe each cluster, as well as the element count for each partition (outliers and clusters 1, 2 and 3).

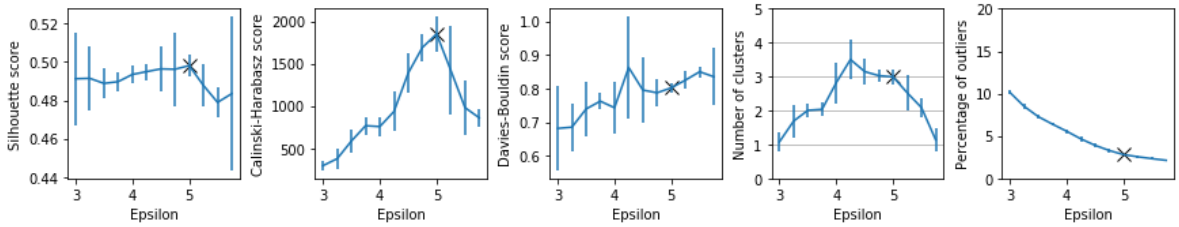

Fig. 5. Bootstrap method results for clustering validation of hyperparameter ϵ for DBSCAN algorithm with $m_p = 45$ and Euclidean distance metric. Mean values for each statistic, with error bars of size $\pm s$, where s is the sample standard deviation. The X marker indicates the mean value of each statistic for $\epsilon = 5$, as chosen in section 4.

favored with large coefficients in the Ridge models to compensate the fact that all inliers have these features set as zero. We can infer from a semantic cluster analysis that cluster 2 represents contracts signed with candidates that have received zero votes during election, while cluster 3 isolates contracts signed with suppliers where only 20% of their average revenue comes from car fuel services and 70% from advertising (in other words, they are not exclusive providers of car fuel services). Finally, cluster 1 represents contracts signed majorly with exclusive suppliers of car fuel services.

5. CLUSTERING ROBUSTNESS AND VALIDATION

An important step in the clustering process is to guarantee that the chosen hyperparameters have produced robust partitions and also to validate them against a null hypothesis of randomness on the input data [Halkidi et al. 2001; Sarle 1990].

In order to validate the clustering quality when choosing the ϵ parameter, a bootstrap method was employed by drawing N vector samples with replacement from $\mathcal{X}_{train}$. This process was repeated 200 times for each ϵ, fixing $m_p = 45$ and using an Euclidean distance. The scores and metrics used as statistics are Silhouette coefficient [Rousseeuw 1987], Calinski-Harabasz index [Caliński and Harabasz 1974], Davies-Bouldin index [Davies and Bouldin 1979], the number of clusters found and the percentage of outliers found. Since the outliers identified by DBSCAN are not a cohesive group, all metrics are calculated over only the inlier samples.

The results when varying $3 \leq \epsilon \leq 5.75$ in steps of 0.25 can be seen in Fig. 5, presenting the mean and sample standard deviation for each statistic. Both Silhouette and Calinski-Harabasz scores present high mean values and low dispersion at $\epsilon = 5.0$, suggesting a robust separation between the three identified clusters. On the other hand, the optimum DBSCAN maximum distance according to Davies-Bouldin score is $3 \leq \epsilon \leq 4$, which would result in only one or two clusters being found, with a significant increase in the percentage of outliers. Summarizing, the majority of the statistics under analysis indicate that $\epsilon = 5.0$ chosen in section 4 is a good compromise between number of clusters and their dispersion and separation.

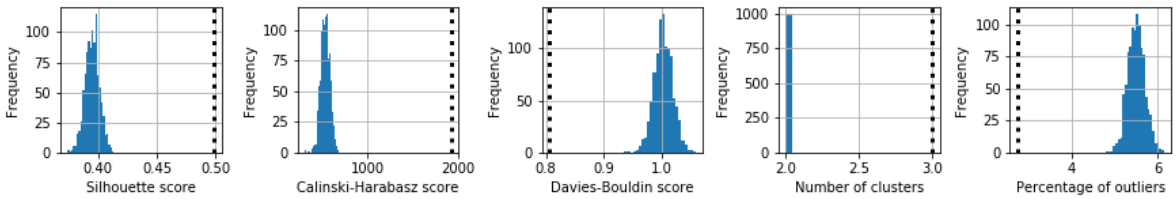

Fig. 6. Histogram of every statistic analyzed using Monte Carlo to test the null hypothesis H_0. The vertical dotted line represents the statistics calculated on the clustering C.

After validating the clustering robustness, it is important to test if the clustering algorithm has not simply found random structures in $\mathcal{X}_{train}$. This analysis requires the use of a quantitative statistical test that will accept or reject a null hypothesis H_0 stating that the data set has a random structure in the feature space. In other words, we must check if the clustering statistics calculated on $\mathcal{X}_{train}$ indicate a statistically significant better clustering than the probability density function (PDF) of the same statistics applied to a random data set. A Monte Carlo method is used to estimate the PDFs of all clustering metrics.

Since the population that generated $\mathcal{X}_{train}$ is unknown, the generation of a random data set for Monte Carlo analysis is not trivial. In this work, each feature of each random vector is independently sampled with replacement from the corresponding feature vector from $\mathcal{X}_{train}$. Because some features contain very few unique entries and some of them with different frequency counts, particularly the ones related to spending categories, this strategy allows the random data set to be drawn from a feature space closer to $\mathcal{X}_{train}$.

The application of the Monte Carlo method to verify the null hypothesis H_0 is shown in Fig. 6, estimating the PDFs of the same clustering indices used in the bootstrap analysis, but now calculated on 1000 random data sets using the same DBSCAN hyperparameters from section 4 . The vertical dotted line represents the same indices calculated on the partition C, as shown in section 4. Scores calculated on C present higher silhouette and Calinski-Harabasz scores and a lower Davies-Bouldin score than all scores from random data sets. This allows the rejection of the null hypothesis H_0 at level $\alpha < 0.05$, indicating that the conclusions obtained from partition C in section 4 are indeed significant, not caused by chance.

6. CONCLUSIONS AND FUTURE WORKS

This work defined a new data set of public spending that was used to identify patterns in contracts between candidates and service providers during their campaigns. A detailed exploratory data analysis was performed, assessing the distribution of every numerical feature. The required transformations were applied during pre-processing to ensure the data set was adequate for model fitting.

The unsupervised density-based clustering method DBSCAN was applied to a subset with contracts related to car fuel spending. A Ridge regression model was trained as a classifier to identify the most important features to define each cluster. The cluster partitions were validated by employing a bootstrap method to confirm the best DBSCAN hyperparameter ϵ and by defining and rejecting a null hypothesis of data set random structure using a Monte Carlo analysis.

The clustering results on the selected subset have shown two clusters of suppliers with different revenue share on car fuel category and a third one related to candidates with zero votes. Further exploratory clustering analysis could provide additional insight into the reason why the revenue of some car fuel suppliers is mainly from advertising.

Future works could test other distance metrics other than Euclidean, including the possibility to use precomputed mixed metrics that could deal with numerical and boolean variables at the same

time. Given that DBSCAN identifies cluster outliers, future works could investigate how they differ from their closest clusters. Finally, more recent clustering algorithms could also be used, such as HDBSCAN* and RNG-HDBSCAN* (both considered an evolution of DBSCAN).

REFERENCES

Amarbayasgalan, T., Jargalsaikhan, B., and Ryu, K. H. Unsupervised novelty detection using deep autoencoders with density based clustering. *Applied Sciences* 8 (9): 1468, 2018.

Baldomir, R. A., Van Erven, G. C., and Ralha, C. G. Brazilian government procurements: an approach to find fraud traces in companies relationships. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*. SBC, pp. 752–762, 2018.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. pp. 93–104, 2000.

Caliński, T. and Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3 (1): 1–27, 1974.

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

Davies, D. L. and Bouldin, D. W. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence* (2): 224–227, 1979.

Ekin, T., Ieva, F., Ruggeri, F., and Soyer, R. Statistical medical fraud assessment: exposition to an emerging field. *International Statistical Review* 86 (3): 379–402, 2018.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*. Vol. 96. pp. 226–231, 1996.

Goldstein, M. and Uchida, S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one* 11 (4): e0152173, 2016.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. On clustering validation techniques. *Journal of intelligent information systems* 17 (2-3): 107–145, 2001.

Hillerman, T., Souza, J. C. F., Reis, A. C. B., and Carvalho, R. N. Applying clustering and AHP methods for evaluating suspect healthcare claims. *Journal of Computational Science* vol. 19, pp. 97–111, 2017.

Kim, J., Kim, H.-J., and Kim, H. Fraud detection for job placement using hierarchical clusters-based deep neural networks. *Applied Intelligence* 49 (8): 2842–2861, 2019.

Kohonen, T. The self-organizing map. *Proceedings of the IEEE* 78 (9): 1464–1480, 1990.

López-Iturriaga, F. J. and Sanz, I. P. Predicting public corruption with neural networks: An analysis of spanish provinces. *Social Indicators Research* 140 (3): 975–998, 2018.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* 9 (Nov): 2579–2605, 2008.

Olszewski, D., Kacprzyk, J., and Zadrożny, S. Employing self-organizing map for fraud detection. In *International Conference on Artificial Intelligence and Soft Computing*. Springer, pp. 150–161, 2013.

Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* vol. 20, pp. 53 – 65, 1987.

Sarle, W. S. *Algorithms for clustering data*. Taylor & Francis Group, 1990.

Sharma, A. and Kumar Panigrahi, P. A review of financial accounting fraud detection based on data mining techniques. In *IJCA*. Vol. 39. pp. 37–47, 2012.

Van Erven, G. C., Carvalho, R. N., de Holanda, M. T., and Ralha, C. Graph database: A case study for detecting fraud in acquisition of brazilian government. In *2017 12th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, pp. 1–6, 2017.

West, J. and Bhattacharya, M. Intelligent financial fraud detection: a comprehensive review. *Computers & security* vol. 57, pp. 47–66, 2016.

Yan, J., Linn, K. A., Powers, B. W., Zhu, J., Jain, S. H., Kowalski, J. L., and Navathe, A. S. Applying machine learning algorithms to segment high-cost patient populations. *Journal of general internal medicine* 34 (2): 211–217, 2019.

Zamini, M. and Montazer, G. Credit card fraud detection using autoencoder based clustering. In *2018 9th International Symposium on Telecommunications (IST)*. IEEE, pp. 486–491, 2018.

Zhang, W. and He, X. An anomaly detection method for medicare fraud detection. In *2017 IEEE International Conference on Big Knowledge (ICBK)*. IEEE, pp. 309–314, 2017.