Anomaly Detection in Brazilian Federal Government Purchase Cards Through Unsupervised Learning Techniques *

Breno Nunes[1][0000−0002−5854−9288], Tiago Colliri[2][0000−0001−9233−4662], Marcelo Lauretto[3][0000−0001−5507−2368], Weiguang Liu[4], and Liang Zhao[5][0000−0002−1502−6604]

[1] Brazilian Office of the Comptroller General (CGU), Brasilia, Brazil
breno.nunes@cgu.gov.br

[2] Dept. of Computer Science, University of Fortaleza (UNIFOR), Fortaleza, Brazil
tcolliri@alumni.usp.br

[3] School of Arts, Science and Humanities, Univ. of Sao Paulo, Sao Paulo, Brazil
marcelolauretto@usp.br

[4] School of Computer Science, Zhongyuan Univ. of Technology, Zhengzhou, China
weiguang.liu@zut.edu.cn

[5] Faculty of Philosophy, Science and Letters, Univ. of Sao Paulo, Ribeirao Preto, Brazil
zhao@usp.br

Abstract. The Federal Government Purchase Card (CPGF) has been used in Brazil since 2005, allowing agencies and entities of the federal public administration to make purchases of material and provision of services through this method. Although this payment system offers several advances, in the technological and administrative aspect, it is also susceptible to possible cases of card misuse and, consequently, waste of public funds, in the form of purchases that do not comply with the terms of the current legislation. In this work, we approach this problem by testing and evaluating unsupervised learning techniques on detecting anomalies in CPGF historical data. Four different methods are considered for this task: K-means, agglomerative clustering, a network-based approach, which is also introduced in this study, and a hybrid model. The experimental results obtained indicate that unsupervised methods, in particular the network-based approach, can indeed help in the task of monitoring government purchase card expenses, by flagging suspect

* This research was funded by CEPID-CeMEAI – Center for Mathematical Sciences Applied to Industry (grant 2013/07375-0, Sao Paulo Research Foundation–FAPESP), and was carried out at the Center for Artificial Intelligence (C4AI-USP), with also support from the Sao Paulo Research Foundation (FAPESP) under grant number: 2019/07665-4 and by the IBM Corporation. This work is also supported in part by FAPESP under grant numbers 2015/50122-0, the Brazilian National Council for Scientific and Technological Development (CNPq) under grant number 303199/2019-9, and the Ministry of Science and Technology of China under grant number: G20200226015.

transactions for further investigation without requiring the presence of a specialist in this process.

Keywords: Anomaly detection · Government purchase cards · Unsupervised learning · Complex networks.

1 Introduction

In the last years, the credit card market has been rapidly increasing in Brazil and worldwide, trading a total amount of more than R$ 1.1 trillion in 2019, only in Brazil, which represents a growth of almost 20% over the previous year [1]. The availability of easy to use and practicality led government agencies from several countries to also adopt the use of this payment method. However, with the credit card market transacting increasingly larger volumes, the detection of fraud and anomalies has become one of the biggest challenges faced by companies in this branch [18]. It is estimated that, annually, fraud and anomalies in the use of credit cards generate losses of billions of dollars worldwide.

In Brazil, the Federal Government Purchase Card (CPGF, in Portuguese) was instituted in 2001 by means of a presidential decree, but its use only started in 2005 when it was regulated by another presidential decree [17]. The CPGF allows agencies and entities of the federal public administration that are part of the fiscal budget and of the social security to make purchases of material and provision of services, under the terms of the current legislation. The CPGF is issued in the name of the management unit of each agency, with the nominal identification of its bearer. Due to its specificity, the main concern regarding the CPGF is not related to fraud – in the sense of theft and/or cloning cases – but in the detection of possible anomalies that can help to identify expenses not covered by the current legislation, thus raising possible misuse of CPGF and, consequently, public funds. In this sense, in an attempt to bring transparency to CPGF expenses, the Brazilian Transparency Portal [19] presents a series of information regarding its transactions, such as: expenses by type of card, by cardholders and favored. Nevertheless, despite the efforts from the Brazilian Transparency Portal to transform the public spending technical data, especially expenses made on the CPGF, into information intelligible to the society in general, this promotion of social control does not fully achieve its objective, since this information, as it is currently presented, does not point out possible anomalies in the application of public resources. Therefore, studies detecting possible anomalies in these data can help to identify non-compliant transactions and potential misuse of the CPGF.

Machine learning techniques for detecting fraud/anomalies can be divided into two main approaches: *supervised* and *unsupervised* ones [18]. In the supervised approach, the dataset used to train the model needs to be labeled as legitimate or fraudulent transactions. In the case of credit cards, such labels often do not exist in the dataset, and the suggested approach, in this case, is the unsupervised one. Several studies have been developed over the years, in different areas, on the identification of anomalies using unsupervised machine learning models. One of these works [14], which can be applied in different areas such as financial transactions and industrial machines, focus on the detection of anomalies in time series, proposing a fast and efficient anomaly detection scheme focusing on fluctuation features, thus being capable of handling non-extreme fluctuation anomalies involved in periodic patterns. In the specific area of fraud detection in financial transactions, an unsupervised deep learning model was used to classify Brazilian exporters regarding the possibility of committing fraud in exports [16]. There is also a study [8] which combines unsupervised and supervised learning, in a complementary form: while supervised techniques learn from the fraudulent behaviors of the past, unsupervised techniques are aimed at detecting new types of fraud in the present. Another particularly related study [3] proposes a methodology for detecting anomalies in government purchases, in the form of public bids, through the use of frequent pattern mining, temporal correlation and joint multi-criteria analysis, obtaining results that are in line with those previously and independently gathered by specialists from the Office of the Comptroller General (CGU), through former audits and inspections.

In this work, we investigate anomalies detection in an unsupervised manner, by using real data from CPGF with the main objective of public transparency, i.e., in an attempt to identify purchases that do not comply with the current legislation for the set of goods and services that can be purchased through this payment method. To this end, four different machine learning techniques are tested: two clustering methods (K-means and agglomerative clustering), a network-based approach, which is also introduced in this work, and a hybrid model, based both on networks and the K-means method. We apply these techniques on historical CPGF expenses real data from Brazilian public agencies, occurred in 2018 and 2019. The experimental results obtained indicate that unsupervised methods have the potential to help in the task of monitoring government purchase card expenses, by flagging suspect transactions for further investigation without requiring the presence of a specialist in this process.

2 Materials and Methods

The dataset used in this study was built from original data obtained in the Brazilian Transparency Portal [19], comprising a total amount of R$ 55,048,553 in CPGF expenses, from 211,896 different transactions, performed by 175 public agencies, in 2018 and 2019. We focus our analyses on data from two public agencies which were among the top largest spenders during this period: Army Command and Ministry of Agriculture, Livestock and Food Supply (Table 1).

In Fig. 1, we show the frequency histogram of all values and also the ones from the 10 most frequent types of expense, from Ministry of Agriculture CPGF data. One can observe, from this figure, that the values are concentrated in smaller transactions. We also note that, overall, the values do not seem to follow a normal distribution, and may present different curve shapes, according to the expense type.

Table 1. Summary of the data used in the analysis

Public agency	Transactions	Total amount (R$)
Army Command	13,550	5,478,177
Ministry of Agriculture	9,450	2,930,626

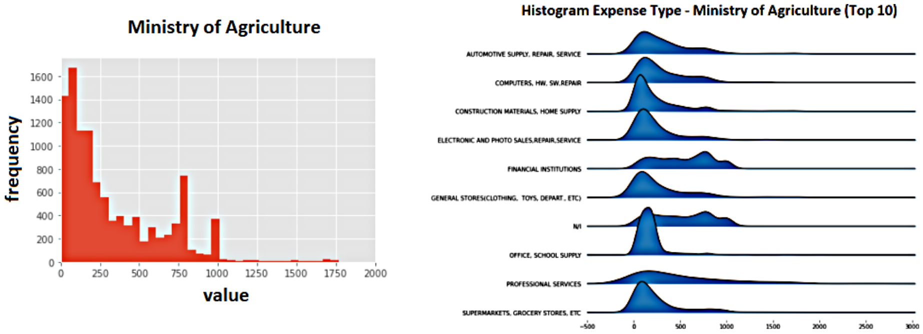


Fig. 1. Frequency histogram of all values and of the 10 most frequent types of expense, from Ministry of Agriculture CPGF data.

For the anomaly detection task, four different unsupervised techniques are tested: K-means [15], agglomerative clustering [12], a network-based approach, also introduced in this study, and a hybrid model, based on complex networks and the K-means method. Following, we describe each one of them in more details.

2.1 K-means method

The idea of the K-means algorithm is to group n objects of a sample into k clusters by minimizing the moment of inertia of each cluster, i.e., by making the data within each cluster as close as possible. K-means uses the squared Euclidean distance to calculate the inertia or, as it is also known, the Within-Cluster Sum of Square - WCSS. Therefore, the total inertia of a dataset can be denoted as follows:

$$\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad , \tag{1}$$

which means that for all points x that are in cluster C_i, the square of the distances from each point to the center of the cluster μ_i is summed up, and to find the total inertia, we just sum all the inertia values of each of the k clusters in the dataset.

One of the problems of the K-means algorithm is that it is extremely sensitive to the centroids starting conditions. This problem can be addressed through the initialization scheme called K-means $^{++}$ [6]. In this method, the first centroid is selected randomly, and the others are selected based on the distance to the first one. In this way, K-means $^{++}$ reduces the likelihood of bad starting conditions by selecting the initial centroids that are usually far from each other. The best number of clusters k in this technique is found using the elbow method [4], and the initialization parameter is adjusted by the K-means $^{++}$ technique.

2.2 Agglomerative clustering method

Agglomerative clustering is a bottom-up hierarchical clustering algorithm where the hierarchy of clusters is represented as a tree, called a dendrogram. The root of the dendrogram is the single cluster that contains all the samples, whereas the leaves are the clusters where each sample would be considered as a cluster. In this study, the dendrogram is created by using the Ward linkage method [20], that seeks to minimize the sum of the square of the differences within all clusters. This is an approach which minimizes the variance and, in this respect, is somewhat similar to the K-means function, but from a hierarchical perspective.

For determining the number of clusters to be considered for this technique, we set a threshold value and draw a horizontal line that cuts the tallest vertical line in the dendrogram, and take the number of vertical lines which are intersected by the line drawn using the threshold as the number of clusters to be used in the analysis. In this case, as one can observe from Fig. 2, the number of clusters has been defined as 3, both for Ministry of Agriculture and Army Command data.

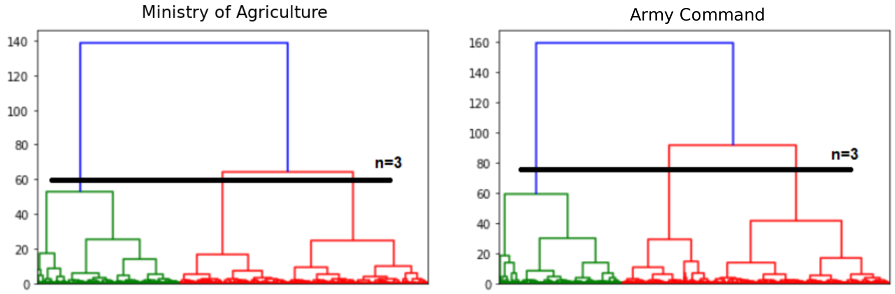


Fig. 2. Dendrograms generated from Ministry of Agriculture (left) and Army Command data (right), for the hierarchical agglomerative clustering analysis.

2.3 Network-based approach

In the last years, driven by technological advances and by the increase in the number of data available to be analyzed, the area of *complex networks* has emerged as a topic capable of unifying complex systems, and is currently present in several branches of science [5]. A network can be defined as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of nodes, $\mathcal{V} = \{v_1, v_2, ..., v_n\}$, and $\mathcal{E}$ is a set of tuples,

$\mathcal{E} = \{(i, j) : i, j \in \mathcal{V}\}$, representing the edges between each pair of nodes. One of the most evident characteristics in complex networks is the presence of communities. The notion of community is straightforward: each community is defined as a subgraph (or subnet) whose vertices are densely interconnected and, at the same time, these vertices have little connection with the rest of the network. A traditional way of detecting communities in a network is through the use of the *modularity* measure, which, broadly speaking, compares the number of connections between vertices which share a same characteristic with the expected number of connections when occurred randomly [7]. The *fast greedy* algorithm [9], for example, determines the optimal number of communities in the network by maximizing the modularity score of the graph.

Although there are currently already different graph-based methods for detecting anomalies [2], for this study, we have opted for developing and testing a novel one for the considered task. The network-based approach introduced in this work is inspired on a technique originally conceived for detecting periodicity in time series, such as in weather-related data [13]. The same technique had also already inspired the developing of a network-based model to identify up and down price trend patterns in the stock market, using real data [10, 11] and, in this study, we adapt it in order to detect anomalies in Federal Government Payment Cards (CPGF) data.

We start by grouping the daily CPGF expenses data, in the form of weekly averaged values, considering only the days where there were at least one card transaction. Afterwards, we calculate the variations in the weekly time series, using a sliding window, in the form of an array X. This array is then sorted in ascending order, by weekly variations, and split into n ranges with the same length l. In the original study [13], l is defined as $\sqrt{t}$, where t is the total length of X and, in this work, we opt to set $l = \sqrt[3]{t}$, thus increasing the number of ranges n in the series, to better adapt the technique for the outlier detection task. These ranges can be represented by a list $R = [r_0, r_1, r_2 ..., r_{n-1}]$ ordered from low to high ranges. More specifically, each range $r_q \in R$ is an interval defined as follows:

$$r_q = \begin{cases} [\min(X), X_{[1/n]}] & \text{if } q = 0; \\ ]X_{[(q-1)/n]}, X_{[q/n]}] & \text{if } q \in (1, \dots, n-2); \\ ]X_{[(q-1)/n]}, \max(X)] & \text{if } q = n-1, \end{cases} \tag{2}$$

where $X_{[qt]}$ denotes the empirical qt-quantile of X, $qt \in [0, 1]$.

Every range $r_q \in R$ will be mapped as a vertex $v_q \in \mathcal{V}$ in the network $\mathcal{G}(\mathcal{V}, \mathcal{E})$. For this end, we label every range of R as a sequence of integers $[0, 1, 2 ..., n-1]$ in the same order they appeared in R, that will be used to represent the indexes of an adjacency matrix A, and also to store the variation ranges in R as node attributes in the set of nodes $\mathcal{V}$, from network $\mathcal{G}$. Initially, A is considered a null matrix, of size $n \times n$. Two vertices v_q and v_w will become connected if there are two consecutive observations x_i and x_{i+1} in the series located in the different ranges r_q and r_w, i.e., if they ever appeared consecutively in the series, chronologically. In this case, we have that $A_{q,w} = A_{w,q} = 1$. The graph resulting from repeating this process to all pairs of consecutive observations in X is an undirected one. The overall procedure for generating the network is described in Algorithm 1.

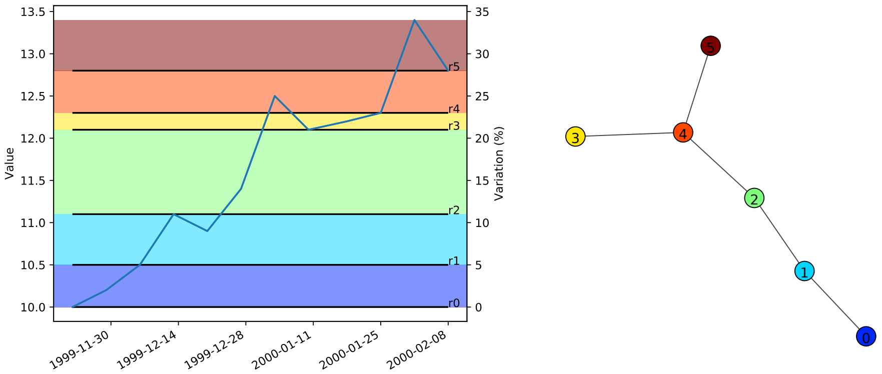
Algorithm 1 Adjacency matrix A generation.

```
Input: CPGF spending data $X$, with weekly variations
Output: adjacency matrix $A$
procedure GET ADJACENCY MATRIX A($X$)
    $X \leftarrow$ sort by weekly variation, in ascending order
    $R \leftarrow$ ranges($X, n$), with same length $l$ and values rounded down
    $X \leftarrow$ sort again by chronological order
    $A \leftarrow$ zero matrix ($n \times n$)
    $t \leftarrow$ length of $X$
    for $i = 0$ to $t - 1$ do
        $r1 \leftarrow$ range($R, x_i$)
        $r2 \leftarrow$ range($R, x_{i+1}$)
        if $r1 \neq r2$ then
            $A_{r_1,r_2} = 1$
            $A_{r_2,r_1} = 1$
    return $A$
```

In Fig. 3, we provide an illustration showing the application of the network-based technique on a simple time series, comprising only 12 observed weekly values. In this case, the array is divided into 6 equal parts with two observations in each of them, i.e., $\sqrt[3]{12}$, creating variation ranges $R = [r_0, r_1, r_2, r_3, r_4, r_5]$. These values are then mapped as nodes in the network $\mathcal{G}$, with the links between them being generated according to whether they ever appeared consecutively in the original time series, pairwise. Please note that, according to this rule, the node 4 is the one with highest degree in the network.

The resulting network $\mathcal{G}$ can be used to identify anomalies and spending oscillation patterns in the CPGF time series, based on the topological structure of the data. Given that two vertices in the network $\mathcal{G}$ will be connected by a link only if they are immediately next to each other, chronologically, at any point in the weekly time series, then the variation ranges in $\mathcal{G}$ will indicate how distant a certain weekly spending value is from the weekly spending value from a certain number of weeks ago, used as sliding window, and the network topological structure will reflect the overall spending oscillation pattern for the CPGF data, with subsequent and similar variation ranges tending to be connected and closer to each other. In this way, most weekly variation values in X will be less or equal the highest range returned by the nodes in $\mathcal{G}$. When there is an exception to this rule, i.e., when a spending data instance in X presents a weekly variation bigger than the variation ranges represented in $\mathcal{V}$, then it will be labeled as an outlier in the results output. This procedure is described in Algorithm 2.

It is also worth mentioning that, since the variation ranges in R are generated based on a sliding window, there is the possibility of regulating the model's sensitivity when detecting sudden or progressive spending increases, with smaller



(a) Time series

(b) Resulting network

Fig. 3. Example demonstrating how the variation ranges in the original dataset X are defined and then mapped as nodes in the network $\mathcal{G}$, in the network-based approach. For the sake of simplification, there are only 12 weekly observations in this time series, and the variations are calculated based on the initial value, instead of on a sliding window. (a) After sorting the series by variation values, in ascending order, the array is split into 6 equal parts, hence delimiting the ranges r_0 to r_5. (b) Each range will become a node in the network $\mathcal{G}$, and two nodes will be connected only if they ever appeared consecutively in the original time series, chronologically.

Algorithm 2 Anomaly detection.

```
Input: CPGF weekly spending data $X$, variation ranges network $\mathcal{G}(\mathcal{V}, \mathcal{E})$
Output: Data instances' labels list
procedure GET LABELS($X$)
    $l \leftarrow [\ ]$
    $t \leftarrow$ length of $X$
    for $i = 0$ to $t - 1$ do
        if $x_i \leq$ max of variation ranges represented in $\mathcal{V}$ then
            $l_i \leftarrow$ corresponding variation range $r_v$ in $\mathcal{V}$
        else
            $l_i \leftarrow$ outlier label
    return $l$
```

window values resulting in a higher sensitivity to detect possible anomalies over time. Another useful feature of the proposed approach is in the fact that the communities detected in the resulting network $\mathcal{G}$ can be used as some sort of spending variation "thermometer", so to speak, such that by looking at which community a certain spending value belongs to, then one can use it as an indicator of how far its variation is from the spending average from a certain number of weeks ago, used as sliding window.

2.4 Hybrid approach

We also test a hybrid model in the anomaly detection task on CPGF data, consisting of taking the weekly averaged values and variation ranges generated by the network-based approach, pairwise, and grouping them using the K-means algorithm. For this end, as preprocessing, all variation values are converted to the logarithmic scale. In this approach, the best number of clusters k is achieved using the elbow method [4] as well, being defined as $k = 4$. To reduce the sensibility to the centroids starting condition, the K-means $^{++}$ method is also used to adjust the initialization parameter.

3 Experimental Results

In this section, we present the results obtained when applying the four techniques to detect anomalies in the CPGF data.

We start by showing the results obtained from the K-means and agglomerative clustering techniques, in Fig. 4. In both of these methods, the identification of points of attention as possible anomalies in the use of CPGF had been pointed out by specialists from the business area, who are familiar with the purchasing process in this modality, through the analysis of the generated plots. In this manner, the points identified in red are those that can be further investigated as possible cases of CPGF misuse. As one can observe, the clusters and anomalies returned by the two grouping approaches for the considered period are very alike, with the main difference between them being in the number of clusters generated.

From the results obtained by both K-means and agglomerative clustering techniques, the specialists from the business area were able to identify a total of 10 transactions as possible cases of CPGF misuse for the Army Command, with two of them amounting to around R$ 30,000 each. As for the results obtained for the Ministry of Agriculture, the specialists were able to identify 5 transactions as suspected, with values around R$ 2,000 each. It is also worth noting that the clusters generated by these techniques were more influenced by the transaction values than by the type of expenditure, as can be observed by their flat shape in Fig. 4.

With regard to the network-based approach, we present, in Fig. 5, the variations range networks resulting from the application of this method on the CPGF time series, using a 12-weeks sliding window, i.e., around 3 months of card transactions. Each node in these networks corresponds to one variation range in the time series. For detecting communities in the network, we use the fast greedy algorithm [9]. Both networks, in this case, present 3 communities, which can be used as some sort of "thermometer" to indicate how far a certain weekly value is from the weekly average from 12-weeks ago. In this way, for monitoring purposes, values from communities with higher variation ranges would trigger more attention than values from lower and in-between communities.

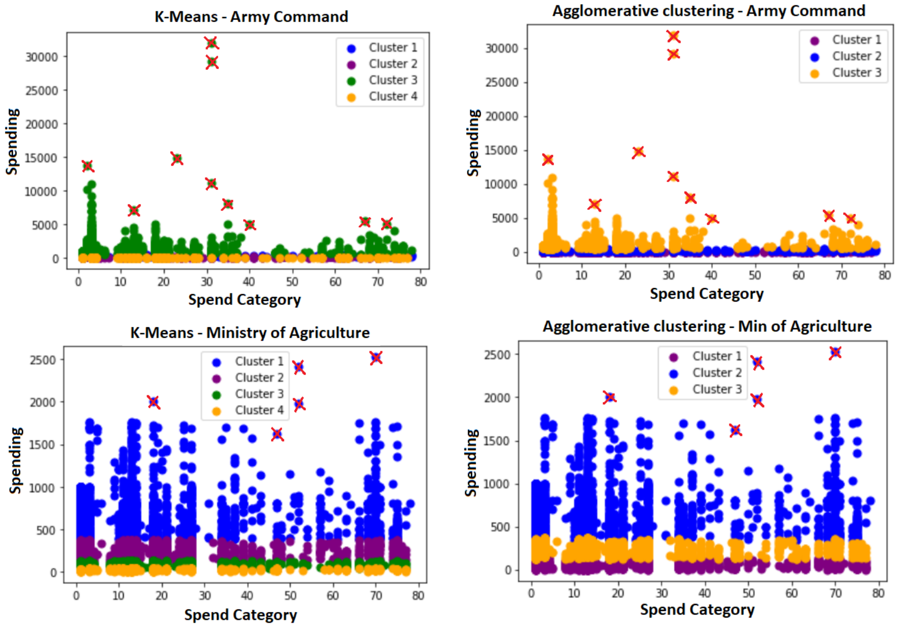


Fig. 4. Results obtained with the K-means and agglomerative clustering techniques. The highlighted values indicate possible anomalies in the use of CPGF, identified by specialists from the business area.

In Fig. 6, we present the results obtained with the network-based approach, showing the CPGF weekly time series, for each public agency, and respective outliers detected during the considered period. The main difference we note in these results, when compared to the ones provided by the K-means and agglomerative clustering methods, is that this method was able to point out, regardless of the identification from specialists in the business area, values considered outliers in the time series, based on the chosen 12-weeks sliding window variation. This feature is important as it can contribute to optimize the anomalies-identification process in this area, thus allowing the specialists team to have more focus on monitoring and investigation activities.

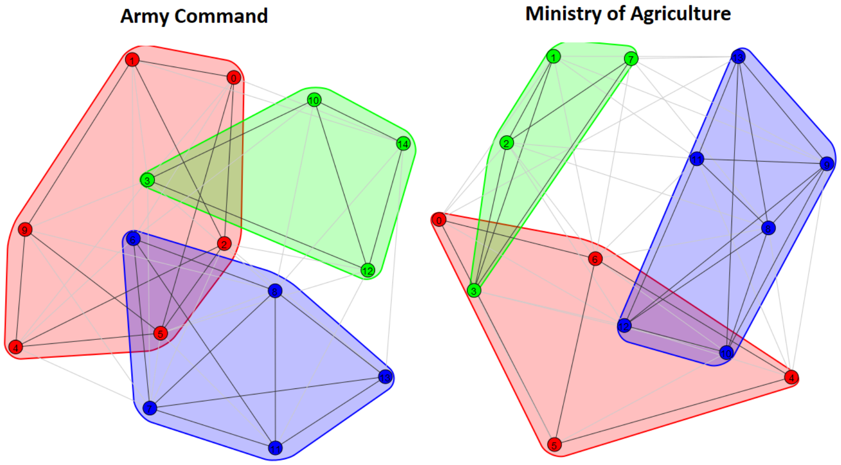


Fig. 5. Variation ranges networks resulting from CPGF data of Army Command and Ministry of Agriculture.

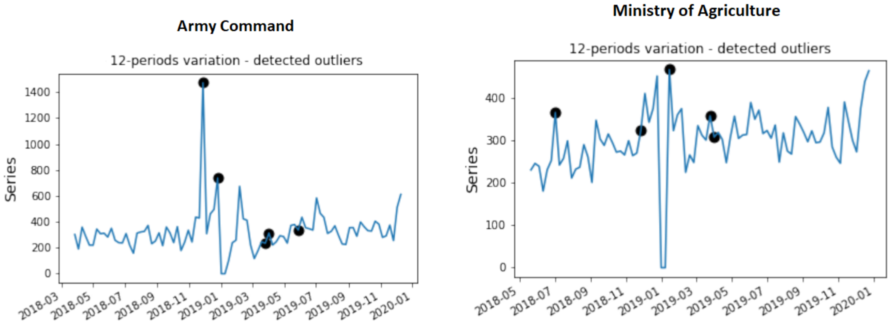


Fig. 6. Results obtained with the network-based approach, using a 12-weeks sliding window. The points in black indicate possible anomalies in the use of CPGF returned by the model for the considered period.

In Fig. 7, we show the weekly variations and the communities identified in the CPGF time series, still according to the chosen 12-weeks sliding window. As mentioned earlier, this plot, with the identification of communities, can be used as an alert indicator for certain values of the time series, where communities with higher variation ranges would raise the alert level for possible suspect transactions, with values increasingly becoming more distant from the average of 12 weeks ago. Hence, this type of representation, which is also provided in the results from the network-based approach, can be helpful, as an additional graphical resource, in the activity of monitoring the CPGF transactions.

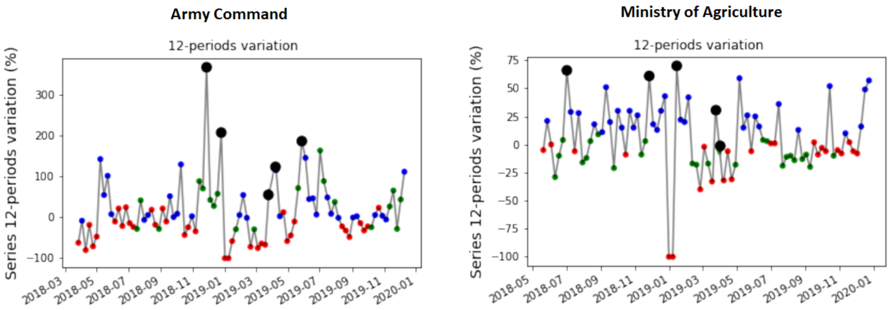


Fig. 7. Weekly variations in the network-based approach. The colors denote the communities, that can be used as an indicator of how much the transaction values deviate from those of a certain number of weeks ago (12 weeks, in this case, which is the chosen sliding window).

Finally, we show, in Fig. 8, the results obtained with the hybrid approach, which consists of taking the information generated by the network-based method, i.e., the weekly averaged values and variation ranges, pairwise, and grouping them using the K-means algorithm. In this technique, the values highlighted can be considered outliers. However, in the same manner of the grouping models, this identification had been carried out by specialists from the business area.

4 Final Remarks

In this work, four different unsupervised learning techniques were tested and evaluated on detecting anomalies in data from Brazilian Federal Government purchase cards (CPGF). The obtained experimental results indicate that the K-means and agglomerative clustering methods were not able to provide a clear output, in terms of information to help in the identification of values that can configure possible misuse of the card. On the other hand, the network-based approach, also introduced in this study, demonstrated to be more successful in this task, as it was the only technique able to automatically infer some points

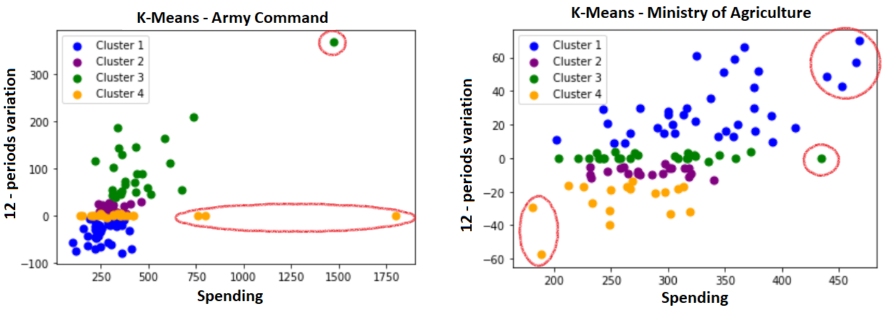


Fig. 8. Results obtained with the hybrid approach. The highlighted values indicate possible anomalies in the use of CPGF, identified by specialists from the business area.

of attention that can be used as a starting point for further investigations. The hybrid model obtained more clear results when compared to the first two clustering techniques, and has potential to be used as well in the initial process of identifying misuse of CPGF.

It is important to note that the results obtained in the anomaly detection task, in this case, depend heavily on the definition of some factors inherent to the knowledge of the business areas from the Controller General office, such as: what are the more suitable sliding window values to make comparisons, what are the more suitable grouping periods for transactions (weekly, biweekly etc.) and what would be the threshold, for each public agency, to define which payments worth a deeper investigation and which ones can be neglected. All of these parameters have proved to be relevant to the results obtained in this study.

We believe that one of the strengths of this research lies in its social contribution aspect, especially when we consider that currently only a few amount of people know how to properly use the Brazilian Transparency Portal and gather relevant information from this source. Therefore, we expect that our work can contribute to improve this scenario, by bringing more attention to the importance of this tool for the population in general, in order to increase the government accountability level in Brazil.

Finally, from the values identified as outliers in the models, especially those returned by the network-based and hybrid approaches, one can evolve to develop a more detailed measurement in order to ascertain whether in fact these identified transactions represent a misuse of the CPGF. In this way, the overall process could hence evolve to a semi-supervised learning approach and, later, to supervised models.

References

1. ABECS: Brazilian association of credit card and services companies. www.abecs.org.br. [accessed on July, 2, 2020] (2020)

2. Akoglu, L., Tong, H., Koutra, D.: Graph based anomaly detection and description: a survey. Data Mining and Knowledge Discovery **29**(3), 626–688 (2015)
3. de Andrade, P.H.M.A., Meira, W., Cerqueira, B., Cruz, G.: Auditing government purchases with a multicriteria anomaly detection strategy. Journal of Information and Data Management **11**(1) (2020)
4. Bholowalia, P., Kumar, A.: EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications **105**(9) (2014)
5. Bornholdt, S., Schuster, H.G.: Handbook of graphs and networks. From Genome to the Internet, Willey-VCH (2003 Weinheim) (2001)
6. Bradley, P.S., Fayyad, U.M.: Refining initial points for k-means clustering. In: ICML. vol. 98, pp. 91–99. Citeseer (1998)
7. Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. IEEE Transactions on Knowledge and Data Engineering **20**(2), 172–188 (2007)
8. Carcillo, F., Le Borgne, Y.A., Caelen, O., Kessaci, Y., Oblé, F., Bontempi, G.: Combining unsupervised and supervised learning in credit card fraud detection. Information Sciences (2019)
9. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. Physical Review E **70**(6), 066111 (2004)
10. Colliri, T., Zhao, L.: A network-based model for optimizing returns in the stock market. In: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS). pp. 645–650 (2019). https://doi.org/10.1109/BRACIS.2019.00118
11. Colliri, T., Zhao, L.: Stock market trend detection and automatic decision-making through a network-based classification model. Natural Computing pp. 1–14 (2021)
12. Day, W.H., Edelsbrunner, H.: Efficient algorithms for agglomerative hierarchical clustering methods. Journal of Classification **1**(1), 7–24 (1984)
13. Ferreira, L.N., Zhao, L.: Detecting time series periodicity using complex networks. In: 2014 Brazilian Conference on Intelligent Systems. pp. 402–407. IEEE (2014)
14. Li, J., Di, S., Shen, Y., Chen, L.: Fluxev: A fast and effective unsupervised framework for time-series anomaly detection. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. pp. 824–832 (2021)
15. Likas, A., Vlassis, N., Verbeek, J.J.: The global k-means clustering algorithm. Pattern recognition **36**(2), 451–461 (2003)
16. Paula, E.L., Ladeira, M., Carvalho, R.N., Marzagao, T.: Deep learning anomaly detection as support fraud investigation in Brazilian exports and anti-money laundering. In: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 954–960. IEEE (2016)
17. Republica, P.: Decreto 5.355 de 25 de janeiro de 2005. www.planalto.gov.br/ccivil03/ato2004-2006/2005/decreto/d5355.htm. [accessed on May, 7, 2021] (2021)
18. Rezapour, M.: Anomaly detection using unsupervised methods: credit card fraud case study. Int J Adv Comput Sci Appl **10**(11) (2019)
19. da Uniao, B.C.G.: Portal da transparencia. Gastos por cartoes de pagamento. www.portaltransparencia.gov.br/cartoes?ano=2019. [accessed on June, 27, 2020] (2020)
20. Ward Jr, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association **58**(301), 236–244 (1963)