Tax Underreporting Detection Using an Unsupervised Learning Approach

Herrera-Semenets, Vitali; Bustio-Martínez, Lázaro; González-Ordiano, Jorge Ángel; van den Berg, Jan

DOI
10.1007/978-3-031-75543-9_2

Publication date
2024

Document Version
Final published version

Published in
Advances in Soft Computing - 23rd Mexican International Conference on Artificial Intelligence, MICAI 2024, Proceedings

Citation (APA)
Herrera-Semenets, V., Bustio-Martínez, L., González-Ordiano, J. Á., & van den Berg, J. (2024). Tax Underreporting Detection Using an Unsupervised Learning Approach. In L. Martínez-Villaseñor, & G. Ochoa-Ruiz (Eds.), *Advances in Soft Computing - 23rd Mexican International Conference on Artificial Intelligence, MICAI 2024, Proceedings: 23rd Mexican International Conference on Artificial Intelligence, MICAI 2024, Tonantzintla, Mexico, October 21–25, 2024, Proceedings, Part II* (pp. 16-28). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 15247 LNAI). Springer. https://doi.org/10.1007/978-3-031-75543-9_2

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Tax Underreporting Detection Using an Unsupervised Learning Approach

Vitali Herrera-Semenets[1(✉)], Lázaro Bustio-Martínez[2], Jorge Ángel González-Ordiano[3], and Jan van den Berg[4]

[1] Advanced Technologies Application Center (CENATAV), La Habana, Cuba
vherrera@cenatav.co.cu
[2] Departamento de Estudios en Ingeniería para la Innovación, Universidad Iberoamericana Ciudad de México, Mexico City, Mexico
lazaro.bustio@ibero.mx
[3] Instituto de Investigación Aplicada y Tecnología, Universidad Iberoamericana Ciudad de México, Mexico City, Mexico
jorge.gonzalez@ibero.mx
[4] Intelligent Systems Department, Delft University of Technology, Delft , Netherlands
j.vandenberg@tudelft.nl

Abstract. Governmental adminstrative domains can potentially benefit from a wide variety of currently available big data analysis methods. The tax administration is such an area that requires massive data processing to identify hidden patterns and trends of possible tax evasion. The use of supervised methods can be effective in these cases, but the lack of available labeled data limits their practical application in real-world scenarios. An alternative is the use of unsupervised methods, which have potential benefits in certain cases. In this sense, unsupervised methods are considered to be feasible as a decision support tool in tax evasion risk management systems. This paper proposes an unsupervised approach to identify signs of tax evasion by detecting, possible, tax underreporting. The proposed strategy is evaluated on a data set associated with individual income tax statistics of the United States. The results achieved are considered to be useful in decision-making and preventive actions on cases reported as suspicious.

Keywords: Tax Underreporting · Tax Evasion · Unsupervised Classification · Clustering

1 Introduction

Tax evasion is a figure consisting of the non-payment of taxes established by law. It is an illegal activity in most legislation [15]. When tax is evaded by partial or total omission, voluntarily or involuntarily, there is an illegitimate decrease in tax revenues and damage to the Tax Administration. Taxpayers can reach tax evasion by altering, with false data, the personal income affidavit. This may

L. Martínez-Villaseñor and G. Ochoa-Ruiz (Eds.): MICAI 2024, LNAI 15247, pp. 16–28, 2025.
https://doi.org/10.1007/978-3-031-75543-9_2

have great impact, especially if there is no efficient control and inspection of the payment activities by taxpayers.

Deliberately underreporting income represents a specific manifestation of tax evasion. The use of supervised Machine Learning techniques has been popular in this scenario [17]. Tax authorities can use supervised classification models to select which taxpayers to audit. This allows them to detect possible tax evasion based on prior knowledge. One problem with supervised models is that they may fail to classify a sample associated with a class (tax evasion case) that has not been seen during training. Additionally, these techniques require labeled training datasets, i.e. the datasets should contain instances of verified *tax evasion*, as well as instances of *normal tax* behavior. A problem with supervised learning approaches is that labeled real-world datasets are not publicly available [13]. Thus unsupervised Machine Learning techniques might be an useful approach to identify tax evasion.

Unsupervised learning allow auditors to identify possible cases of tax evasion without prior knowledge of the domain. This means that new cases of tax evasion that have not been seen before can be identified. One of the difficulties of using unsupervised techniques is the validation of the results. In addition, several proposals do not consider the information that categorical features can provide. Such features are usually discarded, because most popular unsupervised algorithms cannot be directly applied for cases in which domain values are discrete and have no ordering defined [2]. Furthermore, the visualization of the results can also represent a problem, as it can be difficult to visually show the similarities and differences of taxpayers using feature vectors.

The main contribution of this paper is the proposal of an unsupervised learning method that demonstrates high effectiveness in detecting suspicious taxpayers within datasets containing both categorical and numerical features. The method excels in accuracy, providing auditors with a reliable tool for identifying potentially fraudulent activity. Additionally, the paper introduces a novel feature extraction strategy aimed at representing higher-dimensional feature vectors in a two-dimensional space. This strategy facilitates visual comparisons, enhancing decision-making for auditors by providing intuitive insights into taxpayer behavior patterns.

The remainder of this paper is structured as follows. Related work is described in Sect. 2. The proposed strategy is introduced in Sect. 3. In Sect. 4, the experimental results are discussed. Finally, the conclusions are outlined in Sect. 5.

2 Related Work

Detection of tax evasion remains a highly active research area. An unsupervised method utilizing spectral clustering to detect tax underreporting is proposed by De Roux et al. [4]. The authors employ spectral clustering to group tax declarations into clusters, followed by estimating the probability distribution of declared earnings within each cluster. Subsequently, suspicious tax declarations within a cluster are identified by applying a quantile threshold to the corresponding probability distribution.

Mehta et al. [10] proposed a method to detect taxpayers who evade indirect taxes by evading their tax returns. To model the problem, they used several attributes of tax returns filed by taxpayers and another attribute based on business interactions between taxpayers, which is calculated with the TrustRank algorithm and a graph-based representation. Then, a spectral grouping is carried out on the taxpayers, and those located at the limit of each group, by using the kernel density estimation, were marked as tax evaders.

Deep Learning, specifically autoencoder networks, has been used to detect suspicious journal entities in financial statement audits [14]. These algorithms use the reconstruction error of the trained network obtainable for a journal entry to detect anomalies that suggest the occurrence of financial fraud.

The strategy propposed by Vanhoeyveld et al. [16] use an anomaly detection approach to identify value-added tax fraud. Heuristics based on nearest neighbor and clustering are presented. For this, the authors use the euclidean distance, which may turn out to be a limitation in certain cases where categorical features are available.

There is more recent work combining clustering and representational Deep Learning to detect internally validated anomalies [13]. In this proposal, both anomaly detection algorithms are enhanced with knowledge of the relevant domain. A K-Means modification incorporating relevant domain knowledge, in terms of feature weights, is used to detect anomalies. The subset of instances reported by K-Means with non-suspicious tax-related behavior constitutes the training data set for an autoencoder. The set of anomalies identified by the autoencoder is also checked against the anomalies detected by K-Means. Then, the input dataset is extended with a binary feature indicating whether the corresponding instance is labeled as an anomaly by the autoencoder. Finally, a Decision Tree is trained on this extended (labeled) dataset to obtain an explainable surrogate model for anomaly detection.

2.1 Discussion

As observed in this section, techniques based on graphs, neural networks, and clustering are commonly employed to address the problem of tax fraud detection using an unsupervised approach. However, certain aspects limit the performance of these techniques. Graph-based techniques can only be applied if data describing relations or transactions between entities is available. In terms of understanding the classification results, Neural Networks often function as black boxes, so understanding why a model suggests the occurrence of possible fraud is unclear for an auditor [11]. Regarding clustering-based techniques, there are quite a lot of proposals using similarity measures that only support numerical features. This fact prevents these algorithms from taking advantage of categorical information.

Information that a categorical feature can offer (such as the locality or activity carried out by an entity) is very useful during a clustering or nearest neighbor approach. Finally, the works reviewed do not visualize how the taxpayers are grouped. This representation of the results would allow the auditor to carry out more detailed analyzes.

Considering such limitations, the main contribution of this work consists of a heuristic that enables the establishment of similarities between taxpayers using mixed data (categorical and numerical features) to identify suspicious taxpayers, possibly, linked to tax underreporting.

3 The New Unsupervised Learning Approach

The strategy proposed in this work consists of four stages. As shown in Fig. 1, each of these stages is performed sequentially.

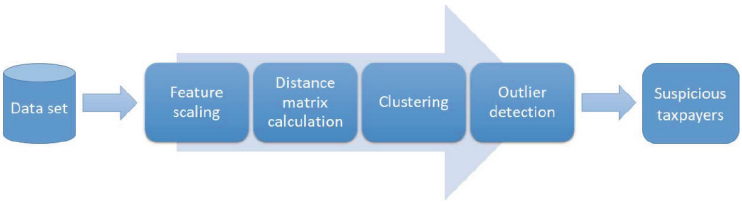


Fig. 1. Unsupervised strategy for suspicious taxpayers detection.

Tax-related data sets contain features (*i.e.* revenues, taxes, returns, etc.) that vary widely in magnitudes, units, and range. This fact can raise a problem if it is intended to apply algorithms that require calculating distances in a data set. If these features are used, the relevant learning algorithms may take into account the (numerical) magnitude of the features and neglect the units: as a consequence of this choice, features with high magnitudes will weigh much more for the distance calculation than features with low magnitudes. Feature scaling can solve this problem, since it allows to bring all the features to a standard level of magnitudes. The first stage of the proposed learning approach is precisely based on numerical features scaling using the standardization function Z-Score normalization. The effect of this type of rescalling is that the mean and standard deviation are 0 and 1, respectively. As can be seen in line 1 of Algorithm 1, the result of this stage is a standardized data set (D_s).

The second stage consists of computing the distances between the instances of D_s (see line 2 of Algorithm 1). It is important to consider that tax datasets are usually made up of mixed features (categorical and numeric). Both types of features are important to establish more precise similarities between the instances of the dataset. In this sense, to calculate the distance matrix, the Gower similarity coefficient was used [7], which allows working with mixed data. For two instances $x_i = (x_{i1}, ..., x_{ip})$ and $x_j = (x_{j1}, \ldots, x_{jp})$, the Gower distance d_{ij} can be computed according to Eq. 1. For each feature $f = (1, \ldots, p)$, a score $s_{ijf} \in [0, 1]$ is defined. If x_i and x_j are close to each other along feature f, then the score s_{ijf} is close to 1. Otherwise, if they are far apart along feature f, the score s_{ijf} is close to 0. The variable δ_{ijf} takes a value equal to 1, if x_i and x_j can be compared

Algorithm 1: Suspicious tax underrepoting detection

Input: D: *Data set*, L: *List of numerical features*
Output: S: *List of suspicious taxpayers*

1 D_s = Standardization(D,L)
2 M = Gower_Similarity(D_s)
3 C = OPTICS(M, $min\_samples = 2 \cdot n$)
4 **foreach** t ***in*** C **do**
5 | $t.tax\_rate = \frac{\theta_t}{\iota_t} \cdot 100$
6 **end**
7 **foreach** c ***in*** C **do**
8 | $\tau_c = \bar{\rho}_c - \sigma_c$
9 | **if** $t_c.tax\_rate < \tau_c$ **then**
10 | | S.add(t_c)
11 | **end**
12 **end**
13 **Return** S

along the feature f. If for some reason, such as missing values, instances x_i and x_j cannot be compared along the feature f, δ_{ijf} is set to 0.

$$d_{ij} = 1 - S_{ij} = \frac{\sum_{f=1}^{p} s_{ijf}\delta_{ijf}}{\sum_{f=1}^{p} \delta_{ijf}}. \tag{1}$$

The value of s_{ijf} is computed according to the feature type. If the feature type is numeric, s_{ijf} is computed using Eq. 2, where $R_f = max(x_{lf}) - min(x_l f)$ is the range of the feature f. For categorical feature type, s_{ijf} is computed using Eq. 3.

$$s_{ijf} = 1 - \frac{|x_{if} - x_{jf}|}{R_f}. \tag{2}$$

$$s_{ijf} = \begin{cases} 1 \; if \; x_{if} = x_{jf} \\ 0 \; if \; x_{if} \neq x_{jf} \end{cases} \tag{3}$$

Once the distance matrix M has been computed (see line 2 in Algorithm 1), the clustering process is performed in the third stage. At this stage, various criteria were considered to select a clustering algorithm that would make the work of auditors more enjoyable and better fit this context. The algorithm must comply with the following requirements: (i) it is not necessary to define a predetermined number of clusters; (ii) the algorithm requires defining a minimum number of parameters; (iii) clusters of various shapes can be discovered, even non-spherical ones; and (iv) it should be taken into account that the Gower distance was used, so it is not recommended to use clustering algorithms that work with the Euclidean distance (such as K-Means). These criteria led to selecting a density-based clustering algorithm, due to the advantage of offering the possibility to identify clusters of different shapes. DBSCAN is a density-based

algorithm very popular in the literature [6]. However, the group of researchers that developed the original algorithm later proposed another algorithm, called OPTICS [1], to address one of the DBSCAN's major weaknesses: the problem of non-detecting meaningful clusters in data of varying density. This relates to the following problem: the constant distance parameter *eps* in DBSCAN only regards points within *eps* from each other as neighbors. If the defined value *eps* is too small, a large part of the data will not be clustered, because it would be very difficult to obtain a dense region that represents a cluster. The non-clustered points are then labeled as outliers by the algorithm. On the other hand, if the defined value *eps* is very high, clusters will merge and the majority of objects will be in the same cluster. Therefore, it is very hard to predefine a perfect *eps* in DBSCAN. OPTICS does not require *eps* parameter to be defined or, what is the same, OPTICS does not require the density to be consistent across the dataset. This algorithm allows the use of a precomputed distance matrix, which is useful for the proposed strategy since a distance matrix is available at this stage. In addition, OPTICS only requires defining the minimum number of samples ($min\_samples$) necessary to form a dense region or cluster. Generally, $min\_samples$ should be greater than or equal to the dimensionality of the data set. The work presented by Sander et al. [12] recommends that if the data set has more than n-dimensions, where $n > 2$, should be chosen $min\_samples = 2 \times n$ (see line 3 of Algorithm 1). Considering the above, the OPTICS algorithm was selected to carry out the clustering process. In the last stage, outliers detection is performed. Most of the clustering-based proposals, select as outliers those taxpayers whose taxes reports are the lowest within the cluster to which they belong. However, this leaves out an important detail: the amount of taxes paid must be in correspondence with the amount of income reported by the taxpayer. This means that the person who pays the least tax in a cluster should not necessarily be an outlier, namely, if their income is also the lowest in the cluster. Based on this observation, it was proposed to compute for each taxpayer $t \in T$, where T is the set of all taxpayers with records in D, the tax rate $t.tax\_rate = \frac{\theta_t}{\iota_t} \times 100$, where θ_t represents the total taxes paid from the declared income ι_t (see lines 4–6 of Algorithm 1). Then, the mean $\bar{\rho}_c$ and the standard deviation σ_c of the tax rates in a cluster $c \in C$, where C is the set of clusters created by OPTICS, are used to compute the threshold $\tau_c = \bar{\rho}_c - \sigma_c$. It is important to highlight in this step that, in order to avoid biases in the calculation of the mean and the standard deviation, the maximum and minimum tax rates are removed. Finally, if a taxpayer t_c, belonging to cluster c, meets the condition $t_c.tax\_rate < \tau_c$, then t_c is considered suspicious of underreporting taxes and added to the list of suspects S (see lines 7–12 of Algorithm 1).

4 Experiments

This section outlines the experimental setup and presents and discusses the obtained results.

4.1 Experimental Setup

The experiments were conducted on a PC equipped with a 2.5 GHz Intel Quad-Core processor, 8 GB of RAM memory running Ubuntu 22.04 OS.

The Individual Income Tax Statistics (IITS) data set was used for evaluating the proposed strategy. The IITS data set is based on individual income tax returns filed with the Internal Revenue Service (IRS) [8], which is the federal agency of the United States Government responsible for tax collection and enforcement of tax laws. This unsupervised data set is made up of 152 features and 166 159 instances. Based on an auditor's judgment, seven features were selected to represent the data set (see Table 1).

Table 1. Selected features.

Feature	Description	Type
AGI_STUB	1 = $1 under $25,000 2 = $25,000 under $50,000 3 = $50,000 under $75,000 4 = $75,000 under $100,000 5 = $100,000 under $200,000 6 = $200,000 or more	Categorical
N1	Number of returns (approximates households)	Numerical
N2	Number of personal exemptions (approximates population)	Numerical
A00100	Adjusted gross income	Numerical
A00200	Salaries and wages amount	Numerical
A10600	Total tax payments amount	Numerical
A02650	Total income amount	Numerical

It is further noted that the data set does not represent individual taxpayers but rather contains statistics generated in various zip codes of different states. In this context, the experiment aims to identify those zip codes where cases of underreporting could occur, as the taxes received correspond to a very low proportion of the reported income, compared to other zip codes with similar feature values. This assertion is supported by the premise that similar zip codes should have a close value of ρ_t.

Given that each state applies different tax rates, the study was conducted locally, focusing on one specific state rather than the entire dataset. Alaska was chosen for this evaluation. In 2020, Alaska ranked among the six states with the highest corporate tax rates, exceeding 9% [3]. High tax rates are known to be a contributing factor to taxpayers evading tax payments [9].

Regarding the proposed strategy, the OPTICS algorithm, utilized for clustering, requires defining the *min_samples* parameter. Based on the strategy

described in the previous section and considering that the dataset comprises 7 features, the parameter $min\_samples = 14$ was defined.

4.2 Experimental Results

After processing the information from Alaska, the data was segmented into 7 clusters. It is important to highlight that the OPTICS algorithm generates a cluster (c_{-1}) for those instances that could not be assigned to any cluster. Therefore, these cases automatically become outliers that auditors can inspect in more detail. Bearing this in mind, the outlier detection stage is not performed on cluster c_{-1}. For the remaining 6 clusters, the strategy is applied in its entirety.

Cluster c_0 contains 32 instances of which 5 were labeled as outliers or suspicious, with AGI_STUB value equal to 1 (see Table 1 for reference to the range it represents). In Fig. 2 it can be seen how the tax percentage of zip codes with similar characteristics behave regarding to the threshold τ_{c_0} (represented with the red line in the Fig. 2) established for cluster c_0, highlighting the 5 suspicious zip codes in orange.

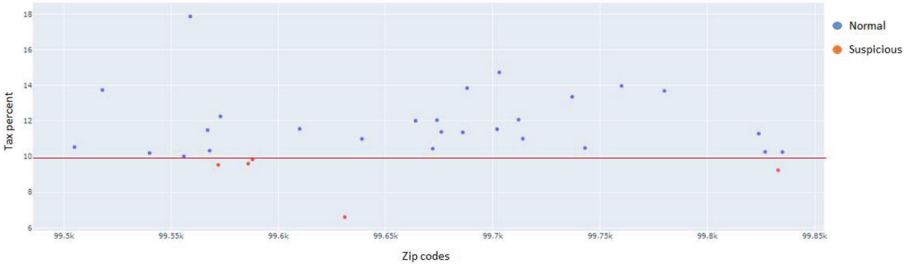


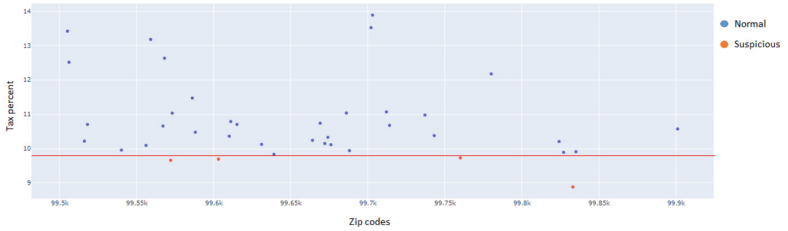
Fig. 2. Tax percentage behavior of zip codes in cluster c_0 (Color figure online)

A total of 39 instances make up cluster c_1, of which 4 were labeled as suspicious, with AGI_STUB value equal to 2. The tax percentage of each zip code in cluster c_1 can be seen in Fig. 3a.

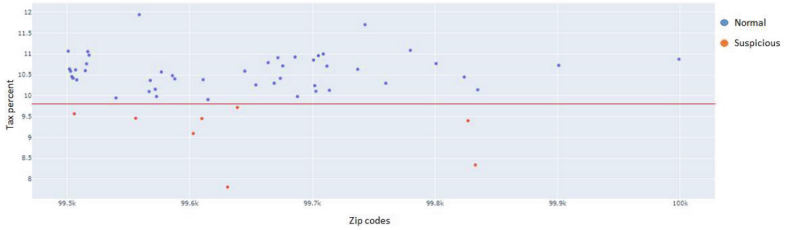
Each of the remaining 4 clusters (c_2, c_3, c_4 and c_5) contain 55 instances. As can be seen in Fig. 3b, 8 suspicious zip codes were identified in cluster c_2, with AGI_STUB value equal to 3.

A similar analysis can be established for clusters c_3 (see Fig. 3c), c_4 (see Fig. 3d) and c_5 (see Fig. 3e), with AGI_STUB value equal to 4, 6 and 5, respectively. In cluster c_3, 10 suspicious zip codes are reported. In the case of cluster c_4, 7 suspicious zip codes were detected. Finally, 8 suspicious zip codes were detected in cluster c_5.

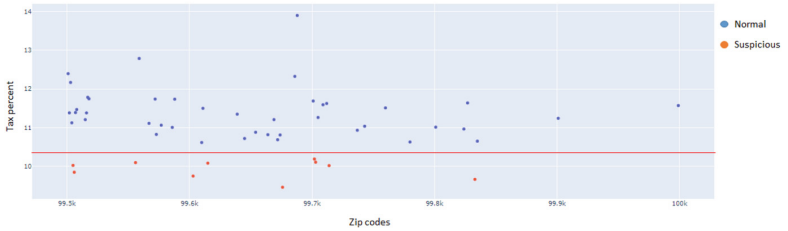
Despite the progressive tax system in the United States, the proposed strategy identifies zip codes with similar characteristics where tax underreporting may occur. Additionally, it is worth noting the significant influence of the categorical feature AGI_STUB in segmenting the data during the clustering process.



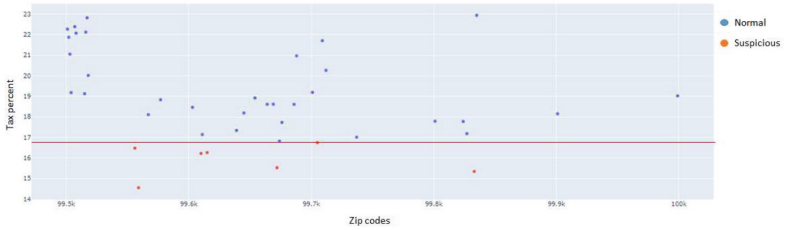
(a) Tax percentage behavior of zip codes in cluster c_1.



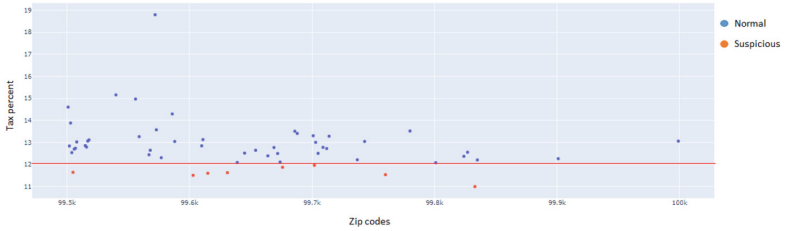
(b) Tax percentage behavior of zip codes in cluster c_2.



(c) Tax percentage behavior of zip codes in cluster c_3.



(d) Tax percentage behavior of zip codes in cluster c_4.



(e) Tax percentage behavior of zip codes in cluster c_5.

Fig. 3. Tax percentage behavior of zip codes in different clusters.

Each cluster shares a common AGI_STUB value, enabling the classification of groups based on adjusted gross income size.

Table 2 shows, for each cluster, the zip codes identified as suspicious. Among them, there are some repeat offenders in various clusters. For example, zip code 99603 is present in four clusters as a suspect, while 99833 appears as a suspect in all clusters. Considering that each cluster is related to the size of the adjusted gross income, it is of interest for an auditor to analyze what happens in these zip codes that appear to be associated with underreporting for different sizes of adjusted gross income.

Table 2. Suspicious zip codes.

Cluster	Zip codes
c_0	99572, 99586, 99588, 99631 and **99833**
c_1	99572, **99603**, 99760 and **99833**
c_2	99506, 99556, **99603**, 99610, 99631, 99639, 99827 and **99833**
c_3	99505, 99506, 99556, **99603**, 99615, 99676, 99702, 99703, 99714 and **99833**
c_4	99556, 99559, 99610, 99615, 99672, 99705 and **99833**
c_5	99505, **99603**, 99615, 99631, 99676, 99702, 99760 and **99833**

Although there is no evidence or feedback to know if these zip codes are actually linked to tax underreporting, it is appropriate to highlight that the proposed strategy follows a deterministic model, which evaluates the hypothesis that in a cluster of similar taxpayers, those who declare low income (below the computed threshold) are suspected of underreporting taxes. Therefore, the model detects suspicious zip codes, which does not imply that they are linked to tax underreporting. The specific analysis of each suspicious zip code to determine whether it is a case of tax underreporting, or not, must be carried out by an auditor. In fact, from a practical point of view, the idea pursued in this work is to obtain a strategy that complements the auditors' decision-making process about who they should audit.

For a better appreciation of the results, it is convenient to make a (two-dimensional) visual representation of the created clusters. The challenge is that the data vectors in these scenarios are usually n-dimensional, where $n > 2$, and in these cases it is necessary to convert them to a two-dimensional space.

The use of a feature extraction algorithm can be useful to reduce the dimensionality of the data. Principal Component Analysis (PCA) is one of the most popular methods of dimensionality reduction. However, the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm has begun to gain popularity for its use in dimensionality reduction for data visualization [5]. One of the most important differences between PCA and t-SNE is that it preserves only local similarities, while PCA preserves a large pairwise distance that maximizes

variance. Another characteristic of t-SNE is that it fits to the underlying data by performing different transformations in different regions. This characteristic also allows finding structures where other dimensionality reduction algorithms cannot.

Based on the previous considerations, the t-SNE algorithm was used to reduce the evaluated data set a two-dimensional space. The structure formed by the clusters is shown in Fig. 4. In this representation it can be seen how those instances (identified with c_{-1}) that could not be assigned to any of the clusters, are closer to clusters c_0 and c_1. This is consistent with the fact that the majority of instances in c_{-1}, specifically 91 %, share the same AGI_STUB value as c_0 (53 %) and c_1 (48 %). This gives us an idea of the information that categorical features can provide for data segmentation and how important they can be for cluster conformation. In addition, this visualization provides more information to an auditor to establish comparisons between the zip codes considered outliers and the clusters close to them.

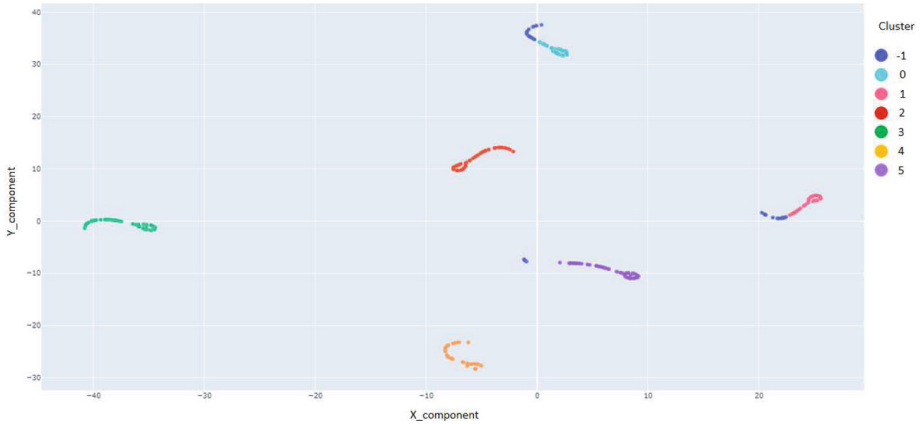


Fig. 4. Visual representation of each cluster instances in a two-dimensional space.

5 Conclusions

The results achieved show that the proposed strategy manages to adequately cluster similar zip codes using mixed data (categorical and numerical features). In addition, it is possible to detect suspicious zip codes whose tax behavior is (very) low compared to other zip codes with similar characteristics. This suggests that underreporting could be occurring. If the information of each taxpayer associated with a zip code were available, a more detailed study at the local level can be performed: applying the proposed approach to the taxpayers associated with suspicious zip code can reveal which ones influence this behavior. The above can support the process of selecting taxpayers to be audited.

Furthermore, visualizing the data in a two-dimensional space allows representing the structures of the clusters created. This makes it possible to associate the zip codes considered as outliers (c_{-1}) to other nearby clusters, which provides valuable information to auditors to perform specific analysis on each case.

Future works should extend the unsupervised strategy to a distributed environment that allows the massive and efficient processing of large volumes of data. Additionally, a study that considers certain features of a state such as demographic, economic, political, religious and social, as variables to predict how prone a state could be regarding to the occurrence of tax evasion is also intended. Such features should take into account privacy concerns and legislation.

References

1. Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J.: Optics: ordering points to identify the clustering structure. ACM SIGMOD Rec. **28**(2), 49–60 (1999)
2. Bai, L., Liang, J.: A categorical data clustering framework on graph representation. Pattern Recogn. **128**, 108694 (2022)
3. Center, T.P.: The state of state (and local) tax policy (2023). https://www.taxpolicycenter.org/briefing-book/how-do-state-and-local-corporate-income-taxes-work. Accessed 3 Mar 2023
4. De Roux, D., Perez, B., Moreno, A., Villamil, M.D.P., Figueroa, C.: Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 215–222 (2018)
5. Devassy, B.M., George, S.: Dimensionality reduction and visualisation of hyperspectral ink data using t-SNE. Forensic Sci. Int. **311**, 110194 (2020)
6. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, vol. 96, pp. 226–231 (1996)
7. Gower, J.C.: A general coefficient of similarity and some of its properties. Biometrics, 857–871 (1971)
8. IRS: Individual income tax statistics data set (2023). https://www.irs.gov/pub/irs-soi/19zpallnoagi.csv. Accessed 1 Mar 2023
9. Kassa, E.T.: Factors influencing taxpayers to engage in tax evasion: evidence from Woldia City administration micro, small, and large enterprise taxpayers. J. Innov. Entrepreneurship **10**(1), 1–16 (2021)
10. Mehta, P., Mathews, J., Bisht, D., Suryamukhi, K., Kumar, S., Babu, C.S.: Detecting tax evaders using TrustRank and spectral clustering. In: Abramowicz, W., Klein, G. (eds.) BIS 2020. LNBIP, vol. 389, pp. 169–183. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-53337-3_13
11. Pang, G., Shen, C., Cao, L., Hengel, A.V.D.: Deep learning for anomaly detection: a review. ACM Comput. Surv. (CSUR) **54**(2), 1–38 (2021)
12. Sander, J., Ester, M., Kriegel, H.P., Xu, X.: Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. Data Min. Knowl. Disc. **2**, 169–194 (1998)
13. Savić, M., Atanasijević, J., Jakovetić, D., Krejić, N.: Tax evasion risk management using a hybrid unsupervised outlier detection method. Expert Syst. Appl. **193**, 116409 (2022)

14. Schultz, M., Tropmann-Frick, M.: Autoencoder neural networks versus external auditors: detecting unusual journal entries in financial statement audits. In: Hawaii International Conference on System Sciences (2020)
15. Vâlsan, C., Druică, E., Ianole-Călin, R.: State capacity and tolerance towards tax evasion: first evidence from Romania. Adm. Sci. **10**(2), 33 (2020)
16. Vanhoeyveld, J., Martens, D., Peeters, B.: Value-added tax fraud detection with scalable anomaly detection techniques. Appl. Soft Comput. **86**, 105895 (2020). https://doi.org/10.1016/j.asoc.2019.105895, https://www.sciencedirect.com/science/article/pii/S1568494619306763
17. Wang, G., Ma, J., Chen, G.: Attentive statement fraud detection: distinguishing multimodal financial data with fine-grained attention. Decis. Support Syst., 113913 (2022)