Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach

Daniel de Roux
Universidad de los Andes
Bogotá, Colombia
d.de1033@uniandes.edu.co

Boris Pérez
Universidad de los Andes
Bogotá, Colombia
br.perez41@uniandes.edu.co
Univ. Francisco de Paula Santander
Cúcuta, Colombia
borisperezg@ufps.edu.co

Andrés Moreno
Universidad de los Andes
Bogotá, Colombia
dar-more@uniandes.edu.co

Maria del Pilar Villamil
Universidad de los Andes
Bogotá, Colombia
mavillam@uniandes.edu.co

César Figueroa
Secretaría de Hacienda Distrital
Bogotá, Colombia
cfigueroa@shd.gov.co

ABSTRACT

Tax fraud is the intentional act of lying on a tax return form with intent to lower one's tax liability. Under-reporting is one of the most common types of tax fraud, it consists in filling a tax return form with a lesser tax base. As a result of this act, fiscal revenues are reduced, undermining public investment.

Detecting tax fraud is one of the main priorities of local tax authorities which are required to develop cost-efficient strategies to tackle this problem. Most of the recent works in tax fraud detection are based on supervised machine learning techniques that make use of labeled or audit-assisted data. Regrettably, auditing tax declarations is a slow and costly process, therefore access to labeled historical information is extremely limited. For this reason, the applicability of supervised machine learning techniques for tax fraud detection is severely hindered.

Such limitations motivate the contribution of this work. We present a novel approach for the detection of potential fraudulent tax payers using only unsupervised learning techniques and allowing the future use of supervised learning techniques. We demonstrate the ability of our model to identify under-reporting taxpayers on real tax payment declarations, reducing the number of potential fraudulent tax payers to audit. The obtained results demonstrate that our model doesn't miss on marking declarations as suspicious and labels previously undetected tax declarations as suspicious, increasing the operational efficiency in the tax supervision process without needing historic labeled data.

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219878

CCS CONCEPTS

• **Mathematics of computing** → **Distribution functions**; • **Information systems** → **Data analytics**; **Clustering**; • **Applied computing** → *Economics*;

KEYWORDS

Unsupervised machine learning, Anomaly detection, Spectral clustering, Kernel density estimation, Tax fraud detection

ACM Reference Format:
Daniel de Roux, Boris Pérez, Andrés Moreno, Maria del Pilar Villamil, and César Figueroa. 2018. Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3219819.3219878

1 INTRODUCTION

Tax fraud is a global phenomenon, affecting society as a whole. This phenomenon can be described as an intentional act of lying on a tax return with intent to obtain illegal financial benefit and lowering one's tax liability [2, 26]. Recent studies [6, 8] have estimated that governments around the world lose approximately US$500 billion annually. Fiscal revenue losses are particularly greater in low to mid income countries in the sub-Saharan Africa, Latin America and the Caribbean and South Asia regions, which also are the countries that have a greater reliance on tax revenues for their fiscal planning. As a result, these countries are the most affected by budget shortages, limiting the reach of their public investment. It is crucial for governments to adopt cost effective tax fraud detection strategies to distinguish between fraudulent and non-fraudulent activities, thereby classifying fraudulent taxpayers or activities and enabling tax authorities to take actions to decrease the impact of fraud [3, 5, 16, 28].

Local tax authorities are responsible for designing policies that are directed to ensure a horizon of sustainable finances. One key aspect to attain this goal is to enable the adequate and efficient verification of the tax payers' compliance. Moreover, they are responsible for finding effective and efficient methods and models to

select taxpayers for tax audits. These audits are often developed to determine the consistency between the amount payed by a taxpayer and the real value to pay.

According to Castellón in [5], tax authorities have traditionally tackled tax fraud through two approaches: auditors experience and rule-based systems. The former approach consists of randomly selecting tax declarations and auditing them based on the experience, intuition, and domain knowledge of the tax auditors. The latter uses methodologies based on rule-based systems. A rule-based system, as described by Baesens et al. in [4], is often implemented in the form of a set of *if-then* rules that detect fraud cases. These rules are developed by a burdensome process, where auditors, after a through review, detect a tax fraud case, generalize its characteristics and include a rule into the tax fraud knowledge base. Once applied to incoming cases it triggers an alert or signal when fraud is detected. However, these traditional techniques have two important disadvantages: (1) They rely mainly on past experiences, therefore are unable to discover new fraud mechanisms by themselves. (2) The subjective judgment of experts makes the knowledge bases for rule-based systems expensive to build, maintain and update [4].

A more recent approach to detect tax fraud is through the use of data mining techniques, which provide mechanisms to extract and generate knowledge from substantial volumes of data to support the detection of fraudulent behavior, improving the use of resources [3, 5, 9, 11, 28]. Also, this is one of the best established applications of data mining in both government and industryTax authorities in most countries have frequently adopted data mining techniques to assist them in identifying taxpayers who evade obligations [28].

There are works, in the literature, using supervised learning techniques because of the use of labeled data or audit-assisted data. Also there have been substantial efforts in the development of models for fraud detection and risk scoring using data mining techniques. However, auditing tax declarations is a slow and costly process, therefore consolidating a historic labeled data set takes valuable time and resources, as a result, access to labeled information is extremely limited. This restricts the applicability of this kind of models for tax fraud detection.

These issues motive the objective of this paper: a screening technique for the detection of under-reporting tax declarations without having historic labeled data. Under-reporting is a type of tax fraud characterized by filling a tax return with a lesser tax base; or by claiming deductions and exemptions that are not applicable [19, 27].

The main insight, although quite evident, allows us to build a flexible yet effective model: similar tax declarations should pay a similar tax base, therefore, this model is based on clusters of declarations. Once tax declaration are grouped, an estimation of how anomalous the reported tax base is made taking into account the declaration's group. We validated it using the Urban Delineation tax as study case in collaboration with the District Treasury Secretary ("Secretaría de Hacienda Distrital").

The rest of the paper is structured as follows. Section 2 presents the related work in the context of fraud, credit risk and data mining. Section 3 presents a description of the proposed technique used in our approach to detect fraudulent tax payments without the need of marked data. Section 4 shows these techniques applied in a specific case, and the most important findings. Finally, section 5 presents the conclusions of this work.

2 RELATED WORK ON FRAUD DETECTION USING DATA MINING

To date, the most common fraud detection tools used by tax authorities are rule-based systems [15]. Tax fraud experts use their knowledge and recorded historical fraud cases to define a set of rules to apply. When a new tax payment matches one of the defined rules, the system shows an alarm indicating a potentially fraudulent taxpayer.

According to Kültür in [15], this approach is successful for identifying fraudulent transactions but it lacks agility. This is also supported by Krivko in [14], who states that a long delay is required before a rule can be added, and during this time, fraud strategies or even some government laws may change, making the rule obsolete.

For the discovery of under-reporting fraudsters, many approaches have previously been proposed in both academia and industry. These approaches are focused on fraud detection and risk scoring. They use data mining techniques which can be classified as supervised learning techniques, unsupervised learning techniques and semi-supervised learning techniques, as show in Table 1. All these works use mainly supervised and semi-supervised techniques. None of these works are purely unsupervised.

Related to supervised learning, we identified mainly two types of works. The first type consists in a unique algorithm to detect fraudsters or to score the risk of default [1, 23, 28], and the second type, uses several algorithms and ensemble algorithms to produce better results [17, 25, 29].

In the first group, works such as Wu et al. in [28] employ associations rules to enhance the performance for VAT evasion detection in Taiwan. They search for patterns or rules of relationships between the associated attributes of VAT evasion. Sanchez et al. in [23] extract a set of fuzzy association rules from a dataset containing both fraudulent and genuine credit card transactions. Another work such as Aleskerov et al. in [1] proposes a system based on a neural network learning module for credit card fraud detection based on the analysis of previous spending data. They used marked data to train the model and to do experimentation in order to determine the accuracy of their model.

In the second group, proposals such as Twala in [25] explore the predictive behaviour of an ensemble of five classifiers for different types of noise in terms of credit risk prediction accuracy. The data used was artificially corrupted, and no real data was used. Matos et al. in [17] identify fraud patterns using association rules and two dimension-reduction methods (i.e. PCA and SVD), then they created a fraud scale to rank taxpayers according to their potential fraud behavior using real taxpayer data. Results were validated by tax auditors specialized in fraud detection. Yeh and Lien in [29] review six data mining techniques (discriminant analysis, logistic regression, Bayes classifier, nearest neighbors, artificial neural networks, and classification trees) and their applications on credit scoring. In terms of classification accuracy, artificial neural networks obtain the best results compared to the other five methods. .

Other works tackle the problem of tax evasion and risk scoring using semi-supervised learning techniques. These proposals can

Table 1: Types of models for fraud detection and risk scoring

Learning technique	Category	References
Supervised learning	Unique algorithm	[21], [23], [1]
	Ensemble algorithm	[17], [25], [29], [12]
Unsupervised learning	Fraudulent behavior	[31], [5], [20]
	Non-typical activities	[30], [3]

be divided into two categories: the first one related to identification of taxpayers fraudulent behaviour, and the second one, the identification of non-typical activities.

In the first category, works such as [5, 31] used supervised learning algorithms to classify fraudster operations combined with a clustering methodology to obtain smaller classification errors. The data was marked by tax audit experts. More concretely, Zhang in [31] presents an approach to detect whether the tax declared by an enterprise is legitimate or not. They used ensemble ISGNN (Iteration learning Self-Generating Neural Network), which is a type of Self-Organizing Neural Network, to solve the problem of fraud detection in tax declarations. This technique builds the model in an unsupervised fashion and then uses marked data samples for adjusting the weights of the model and validating the results.

Castellón and Velásquez in [5] uses self organizing maps (SOM), neural gas (NG) and decision trees, for characterization and identification of behavioral patterns in micro, small, medium and large companies. Similar taxpayers are grouped by the SOM and NG algorithms, then decision trees are used in each group to classify the company in a supervised fashion.

The second category deals with the identification of non-typical activities through the use of unsupervised learning algorithms. Some of them, [3, 30] build groups based on previous activities and then try to establish if a new activity is typical or not.

Dias et al. in [3] classify tax payers based on their risk of tax evasion. They propose a Cluster Analysis methodology to organize observations into homogeneous groups. They use KMeans with three clusters or homogeneous groups of observations. This number of clusters allowed them to identify companies at risk in a more effective way. Zaslavsky and Strizhak in [30] use all transactions in a payment system and divide it into two disjoint subsets: legal transactions and fraudulent ones. The main idea is to create a cardholder's profile using patterns of legal cardholder transactions and patterns of fraudster transactions. In this work, the identification of a typical cardholder behaviour requires a set of some of his previous transactions.

According to the academic literature review on data mining techniques in financial fraud detection presented by Ngai et al. in [19], the use of unsupervised learning approaches for outlier detection is rather low, as most works deal with marked and historic data sets.

To conclude, we mention that many approaches have been proposed for tax fraud detection and risk scoring. However, these approaches are mainly of supervised learning, or rely on the past behaviour of tax payers or credit users. In other words, they use marked data indicating a fraud, and use this information to create both prediction and classification models. Results of these techniques are satisfactory, however, marked data is very hard to obtain since it requires prior auditing, and the results hardly generalize across different types of taxes. It is a well know fact that auditing is a long and costly process, and thus data sets for different types of taxes are usually small and inference is limited.

These techniques can't be generally be applied, since access to tax fraud labeled historical data is unavailable to most local tax authorities which are only starting to consolidate marked data sets. In addition, there is a lack of clarity in the literature about how data should be processed in this context. This makes it hard to replicate and evaluate these proposals.

The above-mentioned issues motivate this proposal, which consists of a general strategy based solely on unsupervised learning which allows for the detection of under-reporting tax payments and which can be applied to different types of taxes. Its aim is to support tax audit experts in defining critical elements to take into account when performing detection of fraudulent taxpayers.

3 DETECTING UNDER-REPORTING TAX DECLARATIONS

In many scenarios, taxpayers must declare the amount according to the tax base in some process and pay a percentage of that amount. This implies that many tax payers under-report earnings to reduce their taxes, as they have no incentive to report the actual amount. On the other hand, tax authorities have many responsibilities, including detection of fraudulent behaviour. This task is heavily time and resource consuming. Moreover, it is difficult to have labeled tax payers as fraudulent and non fraudulent since auditing is a long and costly process. This greatly reduces the use of traditional supervised analysis, which usually requires vast amounts of data to produce desirable results.

These issues motivate our proposal, which provides a strategy to detect and score taxpayers under-reporting their tax base in order to pay less taxes than they should. The proposal follows a general *unsupervised* methodology that doesn't rely on apriori auditing. It is intended to allow screening of suspicious tax declarations. The main assumption of this work is that similar tax declarations according to their features should pay more or less the same amount of money.

Suppose a tax declaration $X \in \mathbb{R}^{p+1}$ is characterized by p variables or features. Some features may be continuous and others categorical. Now, the $p+1$ entry of X is the tax base that the taxpayer declares. His tax fare is a function of this value, including deductions and exceptions. To detect under-reporting tax declarations, our model seeks to find outliers in the distribution of declared

tax bases. We propose to do this in a three-fold process: First, a clustering phase is made, grouping similar tax declarations according to the values of their features. Second, we adjust a probability distribution to the tax bases reported in each cluster. Finally, we detect suspicious declarations using a quantile of the adjusted distribution. We will present now each of the phases of this process.

3.1 Clustering of the tax declarations

It is very important to only compare tax declarations that are similar to each other. Therefore, we first construct clusters of similar tax declarations. In the literature, many algorithms, such as $K-means$, are used for clustering [13]. However, the former doesn't work well in the presence of both continuous and categorical variables [10]. Some other algorithms for clustering, such as gaussian mixtures, rely on the knowledge of the distribution of underlying variables [10], thus discouraging the use of these methods as the distribution of the variables is almost always unknown. Therefore, we propose a non-parametric method that allows us to include both continuous and categorical data in the clustering process, without assuming any apriori distribution of the data variables. For the construction of the clusters, we propose to use *spectral clustering* to obtain the desired clusters. Spectral clustering is a widely used technique for a finding maximal sub-graphs or cliques in weighted graphs. For an extensive review of this technique, see [18]. To use this algorithm, we construct a strongly connected graph G whose vertices are tax declarations and their edge's weights are the distances between each tax declaration.

The graph is constructed as follows: Let $X_1, ..., X_i, ..., X_n \in \mathbb{R}^{p+1}$ be tax declarations. Let $j \in \{1, .., p\}$ and X_i^j be the $j-th$ feature of X_i , $1 \leq i \leq n$. Let r_j be the range of $j-th$ coordinate along the n observations. In other words, $r_j = max\,(X_1^j, ..., X_n^j) - min\,(X_1^j, ..., X_n^j)$. Now, given two observations X_l and X_k, if the $j-th$ feature is a categorical one, define the $j-th$ coordinate distance as in Equation 1:

$$d_j(X_k, X_l) = \begin{cases} 0 & \text{if } X_k^j, X_l^j \text{belong to the same class.} \\ 1 & \text{else} \end{cases} \tag{1}$$

If the $j-th$ feature is continuous define the the $j-th$ coordinate distance as in Equation 2:

$$d_j(X_k, X_l) = \frac{|X_k^j - X_l^j|^{w_j}}{r_j^{w_j}}. \tag{2}$$

The coefficients r_j assure that the coordinate distance is bounded by 1 for each j. The exponents $w_j \in (0, 1)$ in the continuous case are chosen to take into account the scale and variance of the X^j, and assure that the coordinate distance varies properly from 0 to 1. Notice that d_j is a metric for each j. Define the distance of X_k, X_l as:

$$d(X_k, X_l) = \sum_j \alpha_j d_j(X_k, X_l) \tag{3}$$

As d is a finite sum of metric functions it is also a metric. All of the distances d_j are bounded by 1 so that the different scales of the features don't affect the overall distance. The coefficients α_j are used to give more or less weight to the features.

Using this metric, we construct the adjacency matrix A of the graph G whose vertices are the observations X_i and where the weight of the vertex between X_k and X_l is $d(X_k, X_l)$.

$$A_{kl} = d(X_k, X_l) \tag{4}$$

Spectral clustering uses the eigenvectors of the *normalized symetric laplacian* to find the clusters. For more details, we refer the reader to [18]. The amount of clusters m is a parameter of the algorithm and is usually found by the elbow rule or using a cluster quality measure, such as the cluster silhouette.

Notice that our method didn't assume that any of the features are distributed in a given way, and also allows us to use both categorical and continuous features.

3.2 Estimation of the distribution of declared earnings

In order to find quantiles that discriminate tax under-reporting, we first need to adjust a probability density distribution to the declared tax base of the tax declarations. We use the method of kernel density estimation. Kernel density estimation is a widely used technique in non-parametric statistics for approximation of unknown probability distributions. Let $(x_1, .., x_n)$ be an independent and identically distributed sample drawn from an unknown distribution f. The kernel density estimator $\hat{f}_n^h$ of f is defined as

$$\hat{f}_n^h(x) := \frac{1}{n}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big) \tag{5}$$

K is known as the kernel function. In our implementation, the gaussian distribution was used as the kernel as it is the most widely used in the literature. h is called the bandwidth and is a measure of the weight that each point x_i has on the overall distribution. The main result of kernel density estimation [24] is that under mild assumptions of f and K, the estimated density $\hat{f}_n^h$ converges to f in the sense that

$$\int |(\hat{f}_n^h(x) - f(x))| \to 0 \text{ as } n \to \infty \tag{6}$$

For a cluster $q \in \{1, .., m\}$ let $X_{1,q}, X_{2,q} ..., X_{q_l,q}$ be the tax declarations that belong to this cluster. We estimate the density of the declared earnings $\hat{f}_q$ by using kernel density estimation on the values $X_{i,q}^{p+1}$, $i \in \{1, 2, .., q_l\}$:

$$\hat{f}_q(x) := \frac{1}{q_l}\sum_{i=1}^{q_l} K\Big(\frac{x - X_{i,q}^{p+1}}{h}\Big) \tag{7}$$

Drawing independently from this distribution, we estimate the $5-quantile$ ζ_5^q of $\hat{f}_q$. Finally, given a tax declaration X and its cluster q, we can classify it as anomalous if

$$X^{p+1} < \zeta_5^q \tag{8}$$

In other words, mark the tax declaration X as suspicious of under-reporting if the tax base is too small compared to the tax base of other declarations in the same cluster of X.

4 CASE OF STUDY: URBAN DELINEATION TAX DECLARATIONS

The strategy described in Section 3 was used for detecting under-reporting tax payers for the Urban Delineation tax in Bogotá, Colombia. The Urban Delineation tax is applied to the execution of licensed constructions, or the recognition of previously undeclared constructions. Builders must declare the construction budget and pay a tax consisting of 2.3% of the budget. Since the tax base is self reported by the taxpayer, there is a high risk that builders under-report their budget in order to avoid paying a higher tax base. Thus, the validation of this work will be focused on automatically help auditors to detect suspicious tax declarations that under-reported their construction budgets on tax declarations.

4.1 Dataset

Our data set consists of 1,367 tax declarations of building projects in the city of Bogotá, Colombia. Each invoice has 3 continuous variables: area of the allotment, total area built and project cost, and 3 categorical variables: city zone, number of habitable floors and social stratum. The description of project cost variable is not presented for confidentiality concerns. A brief description of the variables can be seen in Table 2 and Table 3.

Table 2: Description of the 3 categorical variables used

Variable	Unique values	Mode
city zone	105	27
social stratum	7	3
num.of habitable floors	24	3

Table 3: Description of the continuous variables used

Variable	Median	Unique values
area of the allotment	169 m^2	979
total area built	278 m^2	1,265

Data preparation is as follows: declarations with empty features are removed from the data set. All the project costs are brought to present value using the construction price index by year. In this way, all costs are comparable.

4.2 Cluster detection

Using the previously described variables, we construct the distance matrix of the observations following steps described in Section 3. The w_j are set to $\frac{1}{2}$ for the continuous variables, except for the area of allotment and the total area built which are set to $\frac{1}{8}$ and $\frac{1}{5}$ respectively. The area of allotment and the total area built have high variance, so lower values of w_j are chosen to assure that the feature distance properly varies from 0 to 1. All the α_j are set to 1, except for the already mentioned variables for which the α_j were set to 2 by suggestion of tax auditors. This is done to capture the fact that the area of allotment and total area built are more significant than the other features when comparing buildings.

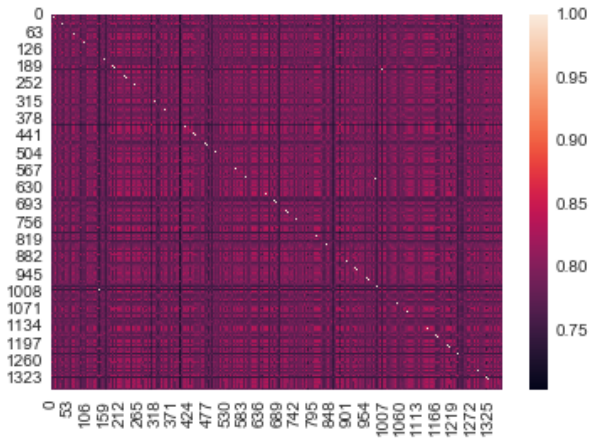


Figure 1: Heatmap of the obtained similarity matrix

Table 4: Number of buildings in each cluster and the 5-quantile used to discriminate between fraudulent and non fraudulent invoices in each cluster.

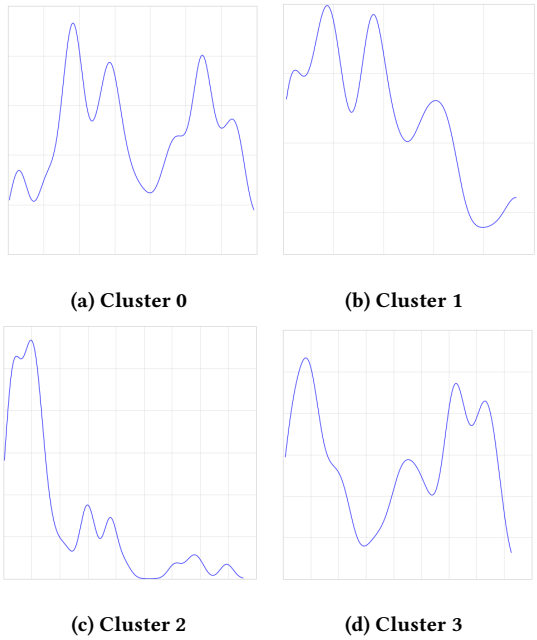
Cluster	Size	5-quantile of declared tax base
0	737	USD $ 1,783.80
1	196	USD $ 9,538.77
2	193	USD $ 6,533.30
3	241	USD $ 11.470,98

Figure 1 shows the heatmap of the obtained adjacency matrix. As the scale shows, the buildings are fairly away from each other. Moreover, we can observe the presence of various black lines that account for the existence of buildings that are away from every other building in the data set. This suggests that certain type of buildings are under represented in our data set and thus various types of building clusters are not accounted for in our approach. This problem could be addressed by increasing the size of the data set.

We build 4 clusters following the strategy described in subsection 3.1. The number of clusters was constructed in a greedy hierarchical way. For each cluster, if the size of the cluster exceeds 20% of the total amount of declarations, we run the spectral clustering algorithm varying the amount of clusters to generate from 0 to 7. We then select the partition with the lowest quality measure, and we repeat the procedure until clusters can't be further improved. We assume that we begin with only 1 cluster (all the data belongs to the same cluster). The number of declarations in each cluster is shown in Table 4.

For each cluster, we estimate the density distribution of the project cost following subsection 3.2. Figure 2 shows the distribution of the declared building cost along the four clusters.

Using the estimated density, we compute the 5-percentiles of the distribution for each cluster, and classify tax declarations as fraudulent following the rule presented on subsection 3.2. The 5-percent threshold is used as it is common in the statistics literature



(a) Cluster 0 **(b) Cluster 1**

(c) Cluster 2 **(d) Cluster 3**

Figure 2: Estimated distribution of building costs along 4 clusters. Scales are omitted for data privacy.

Table 5: Pairwise D-statistic for the two-sample Kolmogorv-Smirnov test of the area of allotment between clusters. All the p-values are lower that $2e^{-16}$ for the null hypothesis "the two samples are drawn from the same distribution".

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Cluster 0	0	0.976	0.948	0.761
Cluster 1	0.976	0	0.712	0.944
Cluster 2	0.948	0.712	0	0.685
Cluster 3	0.761	0.944	0.685	0

to label outliers as the data which is below the 5-percentile of a given distribution. The percentiles are presented on Table 4 for each cluster.

4.3 Validation

The validation of our results was done in two steps.

4.3.1 First step: quality review. The quality of the clusters created was examined. Clusters should exhibit statistical difference in the features of the declarations that belong to them. Figure 3 shows the box plots of different features along clusters, and suggest that the clusters effectively have statistical different features. This result is confirmed using a two-way Kolmogorov-Smirnov test, which is a standard non-parametric technique to test for equality of continuous, one dimensional probability distributions [7]. Table 5 shows the D-statistic for the area of allotment and Table 6 shows the D-statistic for the total area built. A value higher than 0.1 strongly suggest statistical difference of the two samples.

Table 6: Pairwise D-statistic for the two-sample Kolmogorv-Smirnov test of the total area built between clusters. All the p-values are lower that $2e^{-16}$ for the null hypothesis "the two samples are drawn from the same distribution".

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Cluster 0	0	0.683	0.781	0.797
Cluster 1	0.683	0	0.296	0.30
Cluster 2	0.781	0.296	0	0.559
Cluster 3	0.797	0.30	0.559	0

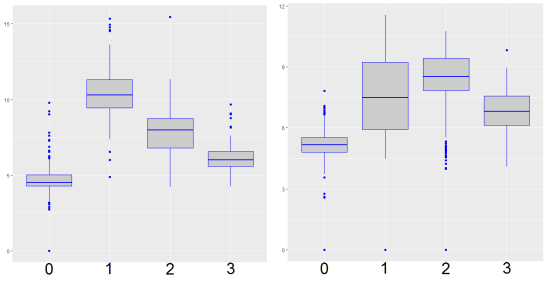
In the process of constructing the clusters, the silhouette was used as cluster quality measure. Cluster silhouette is a method of validation of clustered data. Suppose the observations are clustered into m classes. For each observation i, let $a(i)$ be the average distance of i with all other data in the same cluster. Let $b(i)$ be the lowest average distance of i to all points in any other cluster of which i is not a member. Equation 9 defines the silhouette of i.

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{9}$$

Notice that $-1 \leq s(i) \leq 1 \forall\ i$. A value of $s(i)$ close to -1 indicates that i would be better assigned in a different cluster than it is. A value close to 1 indicates that i is appropriately clustered. The silhouette of the clusters is the average of $s(i)$ over all the observations. The Silhouette measure obtained at the end of the process was 0.08. Recall that the silhouette varies between -1 and 1, and a higher value indicates a better quality of clusters. This is fairly satisfactory due to the small amount of features available that describe the buildings. However, it is far from perfect value. Nonetheless, it is positive, suggesting that the declarations are not wrongly assigned to the clusters, and are on average in the border of natural clusters [22].

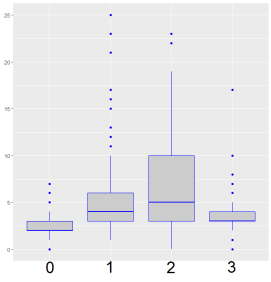
4.3.2 Second step: experts review. Our discrimination of under-reporting declarations was evaluated with the help of an auditor of the Department of Finance of Bogotá.

We took 10 building construction declarations and presented them to the auditor for evaluation. Five of them were marked by our model as under-reporting. The auditor's strategy to determine if the declarations were under-reporting was to verify if the variables of total area built and social stratum were compatible with the taxable base declared. Of the non under-reporting declarations, none were marked by the auditor as suspicious. This suggests that our model doesn't miss on marking declarations as suspicious. Of the under-reporting declarations marked by our model, one was marked as suspicious by the auditor. Since confirming whether or not these invoice are indeed under-reporting requires an auditing, which usually takes 4 to 6 months, it is unclear if our model exhibits a low type 1 error. However, this result was expected since the proposed model is using more information about the constructions than the auditor's own intuition. In any case, our model, at the very least, proposes a priority for tax auditing.

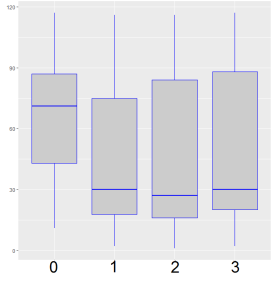


(a) Box plot of the logarithm of the area of allotment along clusters.

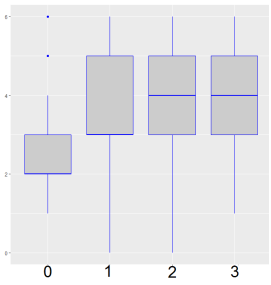
(b) Box plot of the logarithm of the total area built along clusters.



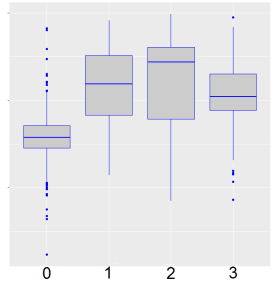
(c) Box plot of the number of habitable floors.



(d) Box plot of the city zone codes.



(e) Box plot of the social stratum codes.



(f) Box plot of the logarithm of the declared tax base. Scale is omitted for data privacy.

Figure 3: Box plots of the different features of the declarations in the different clusters.

5 CONCLUSION

Supervised machine learning techniques tend to fail in the context of tax fraud detection since tax authorities, at least in the Colombian case, have extremely low amounts of historic labeled data due to the high cost in time and resources of auditing. This greatly hinders the generalization power of supervised algorithms and thus their usefulness.

In this paper, we presented a technique that allows tax authorities to prioritize in a data-driven way their audits without requiring historic labeled data. It should be noted that this strategy can be used for tax fraud screening in other type of taxes as long as the fundamental premise assumed for the Urban Delineation tax holds: similar tax invoices should pay a similar amount of taxes.

Although our model shows promising results, many limitations exist. The model assumes that observations of buildings in the same cluster are similar. Moreover, the feature set isn't comprehensive: the available variables are far from completely describing the buildings. This makes some buildings look similar even though in reality they are not, and comparing their costs is unfair. Another limitation was the low amount of data. Buildings come in many different types, and thus many building clusters are expected. Certain types of building are sub-represented in our data set and hence fall in a cluster of very different buildings while they should be on a cluster of their own type. Also, results heavily depend on our fundamental premise. If the premise doesn't hold there is no guarantee that our approach will yield reasonable results. This limits the scenarios or type of tax where this methodology can be applied.

The accuracy of our model for screening of suspicious tax declarations cannot be evaluated directly since our data is not marked with "suspicious", "not suspicious" labels. A partial evaluation was done with the help of a tax auditor. This evaluation is partial since auditing us required to completely determine if a tax declaration is actually under-reporting or not. Such a process usually takes 4 to 6 months.

In addition, the 5-quantile used to discriminate between fraudulent and non fraudulent invoices in each cluster can be improved through two strategies: by integrating into the choice the intuition of the tax auditors after multiple reviews of cases effectively verified ,or by marking as suspicious the left-most mode of the distribution. This latter second strategy is viable in our case since the obtained densities (Figure 2) are multimodal. Intuitively, many project managers could under report the cost of their building by more or less the same margin, thus difficulting their detection. However, this accumulation will appear as a "bump" in the density of declared costs. Following this intuition we could instead mark as suspicious the buildings that appear on the left-most mode of the distribution. In general, when heavy tails are present in the distributions, the outlier selection will be too sensitive to the selected threshold and marking the whole tail as suspicious could help make the approach more robust.

Future work includes the evaluation of our model in a scenario where tax declarations are better characterized by their features and where more data is available. VAT declarations are a good candidate for this work. Finally, we plan to evaluate our approach by comparing it with other approaches based on marked data to measure the error rate of our model.

ACKNOWLEDGEMENTS

This research was carried out by the Center of Excellence and Appropriation in Big Data and Data Analytics (CAOBA), supported by the Ministry of Information Technologies and Telecommunications of the Republic of Colombia (MinTIC) through the Colombian Administrative Department of Science, Technology and Innovation (COLCIENCIAS) within contract No. FP44842-anex46-2015. We would like to make an special acknowledgement to the Taxation Authority of Bogota at the Bogota's Department of Finance (SHD), to the office of telecommunications of the National Planning Department (DNP), and to the CAOBA-Uniandes team, both faculty and students.

REFERENCES

[1] E. Aleskerov, B. Freisleben, and B. Rao. 1997. CARDWATCH: a neural network based database mining system for credit card fraud detection. In *Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFEr)*. IEEE, New York City, NY, USA, 220–226. https://doi.org/10.1109/CIFER.1997.618940

[2] Investing Answers. 2017. Tax Fraud. http://www.investinganswers.com/financial-dictionary/laws-regulations/tax-fraud-4116. (2017).

[3] Maria Elisabete Neves António Dias, Carlos Pinto, João Batista. 2016. Signaling Tax Evasion, Financial Rations and Cluster Analysis. (2016). http://www.gestaodefraude.eu/wordpress/wp-content/uploads/2016/01/wp051.pdf

[4] Bart Baesens, Veronique Van Vlasselaer, and Wouter Verbeke. 2015. *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection*. John Wiley & Sons, 111 River Street, Hoboken, NJ 07030-5774.

[5] Pamela Castellón González and Juan D. Velásquez. 2013. Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications* 40, 5 (apr 2013), 1427–1436. https://doi.org/10.1016/j.eswa.2012.08.051

[6] Alex Cobham and Petr Janskỳ. 2017. *Global distribution of revenue loss from tax avoidance: Re-estimation and country results*. Technical Report. WIDER Working Paper.

[7] W.J. Conover. 1971. *Practical Nonparametric Statistics*. John Wiley & Sons, 111 River Street, Hoboken, NJ 07030-5774. https://books.google.com.co/books?id=NV4YAAAAIAAJ

[8] Ernesto Crivelli, Ruud De Mooij, and Michael Keen. 2016. Base erosion, profit shifting and developing countries. *FinanzArchiv: Public Finance Analysis* 72, 3 (2016), 268–301.

[9] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. Advances in Knowledge Discovery and Data Mining. In *Advances in Knowledge Discovery and Data Mining*, Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy (Eds.). American Association for Artificial Intelligence, Menlo Park, CA, USA, Chapter From Data Mining to Knowledge Discovery: An Overview, 1–34. http://dl.acm.org/citation.cfm?id=257938.257942

[10] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics, New York.

[11] David J. Hand. 1998. Data Mining: Statistics and More? *The American Statistician* 52, 2 (1998), 112–118. https://doi.org/10.1080/00031305.1998.10480549 arXiv:http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.1998.10480549

[12] Sung Ho Ha and Ramayya Krishnan. 2012. Predicting repayment of the credit card debt. *Computers & Operations Research* 39, 4 (apr 2012), 765–773. https://doi.org/10.1016/j.cor.2010.10.032

[13] Anil K. Jain. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31, 8 (2010), 651 – 666. https://doi.org/10.1016/j.patrec.2009.09.011 Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

[14] M. Krivko. 2010. A hybrid model for plastic card fraud detection systems. *Expert Systems with Applications* 37, 8 (2010), 6070 – 6076. https://doi.org/10.1016/j.eswa.2010.02.119

[15] Yiğit Kültür and Mehmet Ufuk Çağlayan. 2017. Hybrid approaches for detecting credit card fraud. *Expert Systems* 34, 2 (4 2017), e12191. https://doi.org/10.1111/exsy.12191

[16] Bin Liu, Guang Xu, Qian Xu, and Nan Zhang. 2012. Outlier Detection Data Mining of Tax Based on Cluster 1. *Physics Procedia International Conference on Medical Physics and Biomedical Engineering* 33 (2012), 1689–1694. https://doi.org/10.1016/j.phpro.2012.05.272

[17] Tales Matos, José Antonio F. de Macedo, and José Maria Monteiro. 2014. An Empirical Method for Discovering Tax Fraudsters. In *Proceedings of the 19th International Database Engineering & Applications Symposium on - IDEAS '15*. ACM Press, New York, New York, USA, 41–48. https://doi.org/10.1145/2790755.2790759

[18] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2001. On Spectral Clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*. MIT Press, One Rogers Street, Cambridge, MA 02142-1209, 849–856.

[19] E.W.T. Ngai, Yong Hu, Y.H. Wong, Yijun Chen, and Xin Sun. 2011. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems* 50, 3 (feb 2011), 559–569. https://doi.org/10.1016/j.dss.2010.08.006

[20] Jon T.S. Quah and M. Sriganesh. 2008. Real-time credit card fraud detection using computational intelligence. *Expert Systems with Applications* 35, 4 (nov 2008), 1721–1732. https://doi.org/10.1016/j.eswa.2007.08.093

[21] Mehdi Samee Rad and Asadollah Shahbahrami. 2016. Detecting high risk taxpayers using data mining techniques. In *Signal Processing and Intelligent Systems (ICSPIS), International Conference of*. IEEE, IEEE, New York City, NY, USA, 1–5.

[22] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[23] D. Sánchez, M.A. Vila, L. Cerda, and J.M. Serrano. 2009. Association rules applied to credit card fraud detection. *Expert Systems with Applications* 36, 2 (2009), 3630 – 3640. https://doi.org/10.1016/j.eswa.2008.02.001

[24] Lanh Tat Tran. 1989. The L₁ Convergence of Kernel Density Estimates under Dependence. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 17, 2 (1989), 197–208. http://www.jstor.org/stable/3314848

[25] Bhekisipho Twala. 2010. Multiple classifier application to credit risk assessment. *Expert Systems with Applications* 37, 4 (2010), 3326 – 3336. https://doi.org/10.1016/j.eswa.2009.10.018

[26] Jau-Hwang Wang, You-Lu Liao, Tyan-muh Tsai, and Garfield Hung. 2006. Technology-based financial frauds in Taiwan: issues and approaches. In *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, Vol. 2. IEEE, IEEE, New York City, NY, USA, 1120–1124.

[27] How Stuff Works. 2017. How Tax Evasion Works. https://money.howstuffworks.com/personal-finance/personal-income-taxes/tax-evasion1.htm. (2017).

[28] Roung-Shiunn Wu, C.S. Ou, Hui ying Lin, She-I Chang, and David C. Yen. 2012. Using data mining technique to enhance tax evasion detection performance. *Expert Systems with Applications* 39, 10 (2012), 8769 – 8777. https://doi.org/10.1016/j.eswa.2012.01.204

[29] I-Cheng Yeh and Che hui Lien. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications* 36, 2, Part 1 (2009), 2473 – 2480. https://doi.org/10.1016/j.eswa.2007.12.020

[30] Vladimir Zaslavsky and Anna Strizhak. 2006. CREDIT CARD FRAUD DETECTION USING SELF- ORGANIZING MAPS. *An International Journal* 18 (2006), 48–63. http://connections-qj.org/system/files/18.03{_}Zaslavsky{_}Strizhak.pdf

[31] K. Zhang, A. Li, and B. Song. 2009. Fraud Detection in Tax Declaration Using Ensemble ISGNN. In *2009 WRI World Congress on Computer Science and Information Engineering*, Vol. 4. IEEE, New York City, NY, USA, 237–240. https://doi.org/10.1109/CSIE.2009.73