Hacienda Pública Española/Review of Public Economics 239-(4/2021): 127-157

https://dx.doi.org/10.7866/HPE-RPE.21.4.4

Segmentation of Potential Fraud Taxpayers and Characterization in Personal Income Tax Using Data Mining Techniques

MARÍA DEL CAMINO GONZÁLEZ VASCO*
Instituto de Estudios Fiscales
MARÍA JESÚS DELGADO RODRÍGUEZ**
Universidad Rey Juan Carlos
SONIA DE LUCAS SANTOS***
Universidad Autónoma de Madrid

Received: March, 2019
Accepted: July, 2020

Abstract

This paper proposes an analytical framework that combines dimension reduction and data mining techniques to obtain a sample segmentation according to potential fraud probability. In this regard, the purpose of this study is twofold. Firstly, it attempts to determine tax benefits that are more likely to be used by potential fraud taxpayers by means of investigating the Personal Income Tax structure. Secondly, it aims at characterizing through socioeconomic variables the segment profiles of potential fraud taxpayer to offer an audit selection strategy for improving tax compliance and improve tax design. An application to the annual Spanish Personal Income Tax sample designed by the Institute for Fiscal Studies is provided. Results obtained confirm that the combination of data mining techniques proposed offers valuable information to contribute to the study of tax fraud.

Keywords: Personal income tax, Tax compliance, Data mining techniques, Multilayer perceptron, Decision trees, Fiscal fraud detection, Tax evaluation.

JEL Classification: H24, C55, C38.

1. Introduction

Fiscal fraud is an important issue that incurs expenses in terms of the loss of government revenues, which leads to less efficient tax programs and the inequity between evaders and

* ORCID ID: 0000-0002-5621-1638.
** ORCID ID: 0000-0003-3830-2701.
*** ORCID ID: 0000-0002-1490-3820.

honest filers (Alm, 2011, Slemrod, 2019). Tax Administration are under increasing pressure, since the financial crisis of 2008 and the large deficits that followed, to collect additional tax revenues and reduce fiscal fraud. Effective control of tax fraud requires addressing a fundamental statistical problem of non-detection, which can bias estimates of the overall amount of fraud and the relative fraud propensities of different socioeconomic groups. Tax fraud detection involves processing a large amount of information in search of fraudulent behavior that requires fast and efficient algorithms, among which data mining provides relevant techniques that can help tax administration to take preventive measures and improve tax design (Liao *et al*., 2012, Micci-Barreca and Ramachandran, 2004).

One of the taxes in which there is greater interest in controlling fraud is in the Personal Income Tax (PIT) since represents the second largest source of tax revenue after the social security contributions (EU, 2019). Papers that estimate Spanish Personal Income tax evasion show the necessity to increase the degree of compliance and that the fight against tax evasion must be a priority objective of Spanish Tax Office (Domínguez *et al*., 2014, 2015, 2017, Torregrosa, 2015). The structure and functioning of the PIT is complex, which makes difficult to control all the information that it is processed to take in consideration the personal and family situation of the taxpayer. In Spain, Personal Income Tax taxes mainly labor, self-employment income and savings income.[1] Tax liabilities corresponding to these categories are summed (intermediate tax liabities), and next an extended set of non-refundable tax credits are applied to figure out what is the amount that the taxpayers have to pay (tax liabiliies). Among other possibilities, tax fraud can be accomplished in the Personal Income Tax by over-claiming the amount of tax credits that can be declared by family and self-employment.[2] Although the amount of fraud per taxpayer in this concepts may not be high, the number of fraudulent taxpayers can be important. Auditing tax declarations is a slow and costly process, so that, tax authorities required to develop cost-efficient strategies to tackle this problem and improve tax design. This issue motivates our proposal. In our analysis we explore the applicability of the data mining techniques in developing a segmentation model that can contribute to tax design evaluation and the characterization of the segments of potential fraud taxpayers in the Personal Income Tax. Despite the increase in the use of these screening and classification models for detecting fraud patterns oriented at audit planning, there are no studies that focus on the identification of tax benefits in the income tax structure that are more likely to be used by potential fraud taxpayers. Additionally, this proposal to segment and characterize potential fraudulent taxpayers can also be applied to different types of taxes.

The proposal presented in this paper is an analytical framework that combines different types of machine learning techniques. As a prior step, we apply a principal component analysis (PCA) as a dimension reduction technique to reduce or eliminate statistical redundancy between the input variables without significant loss of information. Secondly, we test a Multilayer Perceptron (MLP) to assign scores /probabilities that result in fiscal fraud probabilities. Next we use decision trees (CHAID) to segment the sample records according to the probabilities assigned by the MLP. Finally, we characterize the resulting groups or segments according to sociodemographic and economic variables. The main insight, although quite evident, allow us to build a classification rule: those taxpayers who have tax benefits above 95% of their income group (in every of the 12 income stratum in the sample) should be labe-

led for further studies. We validate it using the annual Spanish PIT sample designed by the Institute for Fiscal Studies as study case to determine, first, the tax credits that are more likely to be used by potential fraud taxpayer. Secondly, the proposal makes possible to characterize through socioeconomic variables the profile of the segments of potential fraud taxpayer.

The rest of the paper is organized as follows. Section 2 provides a brief review of literature. Section 3 describes the Spanish PIT microdata included in the analysis. Section 4 illustrates the strategy and methodology of the study. Section 5 outlines the empirical results obtained. Section 6 concludes.

2. Literature review

In last decades, the techniques of data mining and artificial intelligence have been incorporated into the audit planning activities to explore and analyze large quantities of data in order to discover meaningful patterns and rules, oriented to classification and prediction. Both supervised learning methods (where a dependent variable is available for training the model) and unsupervised learning methods (where no prior information of dependent variable is available for use) can be potentially employed to solve this problem. The main data mining techniques used for tax fraud detection are logistic models, artificial neural networks, the Bayesian network, and decision trees, all of which provide primary solutions to the problems inherent in the detection and classification of fraudulent data.

Related to supervised learning methods, we find works such as Wu *et al*. (2012) that employ associations rules to enhance the performance for VAT evasion detection in Taiwan. Matos *et al*. (2014) that identify fraud patterns using association rules and two dimension-reduction methods to create a fraud scale to rank taxpayers using indicators obtained from several fiscal applications in Brazil. Other papers examine how commodity flows respond to destination sales taxes, allowing for tax evasion as a function of distance between trade partners (Fox *et al*., 2014), and identify clusters or groups of taxpayers who have similar behavior using a clustering algorithm named the SOM method (González and Velásquez, 2013). Among the papers that focus on income tax fraud detection we find Perez *et al*. (2019), that contributes by using neural network models with an application to the Spanish PIT. In the case of corporate taxation, papers like Ravisankar *et al*. (2011), Weatherford (2002) apply data mining techniques such as Probabilistic Neural Networks (PNN), and Logistic Regression (LR) to identify companies that resort financial statement fraud.

Other works tackle the problem of tax evasion and risk scoring using unsupervised learning techniques. Dias *et al*. (2016) classify taxpayers based on their risk of tax evasion. They propose a Cluster Analysis methodology to organize observations into homogeneous groups that allowed them to identify companies at risk in a more effective way. De Roux *et al*. (2018) makes a proposal to detect fraud for under-reporting urban delineation tax declarations by clustering and estimating the distributions of declarations in Colombia. Other papers also take advantage of clustering and classification techniques for constructing profiles of fraudulent behaviour, aimed at supporting audit planning (Bonchi *et al*., 1999), among others.

In this revision we have checked different approaches proposed for tax fraud detection. However, most of them are mainly supervised learning, or rely on the past behavior of taxpayers. In this cases, they use marked data indicating a fraud and use this information to create both predictions and classifications models. Results of these techniques are satisfactory, however these techniques can't be generally applied across tax fraud since data is not easily available. This study aims to fill this gap by developing a general strategy based on a combination of Principal Components Analysis, Neural Networks and Decision Trees that allows for the detection of over-claiming tax benefits and which can be applied to different types of taxes without the need to have access to tax fraud labeled historical data. Its aim is to support tax audit experts in defining critical elements to take into account when performing detection of fraudulent taxpayers.

3. The data set

3.1. The annual PIT sample

The application of the methodological proposal is based on the microdata contained in the annual Spanish PIT sample. In this particular paper we use the sample for the year 2013, which includes 2,161,647 records extracted from a population of 19,203,031 registers providing personal income tax returns (Picos, 2014). This database has been developed by the Spanish Institute of Fiscal Studies (Instituto de Estudios Fiscales, IEF), in collaboration with the Spanish National Tax Administration (Agencia Estatal de Administración Tributaria, AEAT), the entity in charge of extracting annual samples from its administrative registers of Spanish personal income tax.[3]

For the construction of this annual sample the minimum variance stratification under Neyman's allocation method has been used. Three stratification variables have been used in the sampling process:

a) Territorial stratum: the province. 46 provinces with common fiscal regime plus the Autonomous Cities of Ceuta and Melilla in addition to non-resident taxpayers who are taxed under Article 10 of Law 35/2006.

b) Income stratum: income level of the tax filers. The sample income was calculated as the sum of net incomes, imputed income and capital gains and losses. The sample income was divided into 12 groups for stratification:

   – *Income group 1:* negative up to 0 euros.

   – *Income group 2:* up to 6,000 euros.

   – *Income group 3:* from 6,000.01 to 12,000 euros.

   – *Income group 4:* from 12,000.01 to 18,000 euros.

   – *Income group 5:* from 18,000.01 to 24,000 euros.

– *Income group 6:* from 24,000.01 to 30,000 euros.

– *Income group 7:* from 30,000.01 to 36,000 euros.

– *Income group 8:* from 36,000.01 to 42,000 euros.

– *Income group 9:* from 42,000.01 to 48,000 euros.

– *Income group 10:* from 48,000.01 to 54,000 euros.

– *Income group 11:* from 54,000.01 to 60,000 euros.

– *Income group 12:* more than 60,000 euros.

c) The type of tax return stratum: separate or joint.

Hence, the "original weight" is calculated for each observation as the ratio between the size of the population of its belonging stratum h and its corresponding sample size $W_h = \frac{\mathrm{N_h}}{n_h}$. To select the sample, the tax returns were classified in each of the 49 x 12 x 2 = 1,176 strata. The sample size n was calculated for a sampling error (in the average of the income variable) less than 1.1% with a confidence level of 3 per 1,000. A restriction of statistical confidentiality has been imposed on design sizes. Therefore, the population for each stratum (N_h) was determined using the population quasi-variance of the sample income S_h^2. Finally, using the values N_h and S_h^2, the sample size of each stratum n_h was determined so that $\sum_h n_h = n$.

3.2. Scope and limitations of the study due to the data set

The main aim of this research is to explore the applicability of data mining techniques for fraud prediction and characterization. In this scope, using historical patterns to identify suspicious behavior similar to known fraud patterns would be ideal, but the lack of information about real fraud records makes this task unapproachable. This leads us to the first limitation of the study, which consists in the identification of fraud taxpayers and the generation of the target variable.

The second limitation has to do with the low frequency of the category of fraud taxpayers with respect to the category of non-fraud taxpayers in our database. The classification techniques used in this study are aimed at minimizing the number of individuals that are poorly classified (confusion matrix). This criterion leads the classification algorithms to assign all records to the category of non-fraud taxpayers, to minimize the number of misclassified records.

In order to identify potential fraud taxpayers we have applied a classification rule that searches data for anomalies that could indicate fraud or error. In data mining techniques, problems like fraud detection are usually framed as classification problems, predicting a discrete class label output given a data observation. In order to generate our dependent variable, the key data for the construction of this rule are the "tax credits", that is, the expenses

declared by taxpayers that decrease their tax liability. We calculate those tax credits as the difference between Intermediate Tax Liabilities and Tax Liabilities amounts in PIT.

The classification rule segments the sample population according to the income group criteria (1-12, Table 1) and labels as "potential fraud" those taxpayers who declare tax credits that are above 95 percentile of their income segment. From a statistical point of view, an observation is considered atypical if it is above the third quartile of the distribution (75th percentile) plus 1.5 times the interquartile range (that is, 75th percentile-25th percentile). We have calculated this threshold in tax credits for every income segment to conclude that the 95th percentile of tax credits is above this definition of outliers.

The results of this criterion are described in the Table 2, which show that the percentage of potential fraud taxpayers according to this classification rule is 1.6% (34,586 individuals). By construction, the number should be around 5%, but we only take into consideration those taxpayers with nonzero tax credits.

It is important to notice that this classification rule, by construction, is labeling taxpayers from every income strata, not just the ones in the upper strata. The reasoning behind this rule is that if a taxpayer in a certain income range is reporting tax credits above 95% of those tax filers in the same income stratum, the tax authority should review their tax return. The validity of this rule, of course, is highly conditioned on the dispersion of the variable "tax credits" within each income stratum. The greater the dispersion (variance) of the variable within the stratum, the higher percentiles of the distribution will be far from the rest.

There is a significant imbalance in the proportion of '1' values (potential fraud) in the target variable. Data mining classification algorithms, like the ones we use in this study, tend to tremble when faced with imbalanced classification data sets. With imbalanced data sets, an algorithm does not get the necessary information about the minority class to make an accurate prediction. This imbalance can lead the model to assign a false '0' (no potential fraud) forecast to all the taxpayers in the sample and, consequently, obtain a 98.40% of well classified registers (percentage near to 100% in terms of accuracy).

Table 1
POTENTIAL FRAUD AND NO-POTENTIAL FRAUD SAMPLE SIZES IN SPANISH PIT SAMPLE

Potential Fraud (PF)	Frequency	Cumulative Frequency	Cumulative Percent
0	98.40	2,127,061	98.40
1	1.60	2,161,647	100.00

Source: Own production using data drawn from the Spanish Personal Income Tax 2013 annual sample.

There are several approaches to handle class imbalance: the conventional solution is to resample the data so that the proportions of 1s and 0s are modified. To achieve this goal, we under-sample the subset of records assigned to the '0' class (non-potential fraud) using the income segment as stratification variable and selecting a random sample of the 20% for each seg-

ment. After merging this sample with the subset of records assigned to class '1', the resulting set contains 460,000 records with a potential fraud class '1' equal to 7.51% of the total group.

Table 2
POTENTIAL FRAUD AND NO-POTENTIAL FRAUD SAMPLE SIZES AFTER THE UNDER-SAMPLING PROCESS

Potential Fraud (PF)	Frequency	Cumulative Frequency	Cumulative Percent
0	92.49	425,454	92.49
1	7.51	34,546	100.00

Source: Own production using data drawn from the Spanish Personal Income Tax 2013 annual sample.

4. Analytical Framework

Below we detail the stages of the analytical framework proposed in this paper. The estimation strategy we will apply to the PIT sample is outlined into four stages:

— *Stage 1: Dimension reduction* to summarize the information provides by 469 variables of the sample using principal components analysis.

— *Stage 2: Scoring taxpayers* to assign evasion probabilities to each taxpayer in the sample. We use the factors obtained in Stage 1 as input variables in a Neuronal Network (Multilayer Perceptron).

— *Stage 3: Segmentation of the taxpayer population* according to the probabilities obtained by the Neuronal Network into different risk segments. We employ a decision tree algorithm (CHAID) to segment the taxpayer population.

— *Stage 4: Characterization of potential fraud taxpayers* through probability distributions of socioeconomic variables. In this way, we characterize the profile of the potential fraud taxpayer.

The four steps are carried out with the IBM SPSS Modeler software and SAS software. Each of these stages is examined in further detail below.

a. *Stage 1: Dimension reduction*

The objective of this stage is to find the smallest subset of dimensions that results in an accurate model, that is, the parsimonious solution. For our data set, which contains 469 input variables it is a necessary task to prevent the model to be over-trained or simply fail to be built, both alternatives possible with large data sets. Despite the fact that we are aware of the great amount of algorithms available for quantifying variable importance, most of these selection processes identify the significance of each variable individually, and skips the opportunity to incorporate the interaction between variables. In most cases, the interaction of two statistically insignificant variables may have a significant effect on the target variable.

The technique that best summarizes the information of all the input variables into orthogonal factors is Principal Components Analysis. The goal in principal components analysis is to find the minimum number of dimensions that are able to explain the largest variance contained in the initial set of indicators. We intend to simplify the information which gives us the correlation matrix to make it easier to interpret.

The principal components model proposed by Pearson (1901) and later on by Hotelling (1933) can be written as:

$$Y = XB + E \tag{1}$$

Where:

Y is an nxp matrix of the centered observed variables;

X is the nxj matrix of scores on the first j principal components;

B is the jxp matrix of eigenvectors;

E is an nxp matrix of residuals.

The goal of model is to minimize the trace of $E'E$. That means that the first j principal components are the best linear predictors of the original variables among all possible sets of j variables, although any nonsingular linear transformation of the first j principal components would provide an equally good prediction. We have computed the principal components from the correlation matrix instead of the covariance matrix for two reasons. First, because of the dependency of the covariance matrix on the units of the input variables. Second, the variance differences of the input variables.

The resulting principal components (that summarize the information of the initial set of 469 variables) will be the input data for the neural network.

b. ***Stage 2: Scoring taxpayers***

At this stage, each value of the principal component variables is fed into one of the neurons in the input layer of a Multilayer Perceptron. The multilayer perceptron (MLP) is a feed-forward, supervised learning network with up to two hidden layers. It is a function of one or more predictors that minimizes the prediction error of one or more targets. This model uses predictors and targets of both types, continuous and categorical. References to fundamentals of this type of artificial neural network can be found, among others, in Parlos *et al.* (1994), Bishop (1995), Ripley (1996), Haykin (1998) and Fine (1999). In this particular case, we use IBM SPSS Modeler software implementation of MLP[4].

The MLP consists of a system of simple interconnected neurons (or nodes). The nodes are connected by weights and output signals, which are a function of the sum of the inputs to the node modified by a simple nonlinear transfer function called "activation function". It is

the overlapping of many simple nolinear transfer functions that enables the MLP to approximate extremely non-linear functions.

Haykin (1998) stated that mathematically, we can describe a neuron k by the following equations:

$$y_k = F\left(\sum_{j=1}^{m} w_{kj}x_j + b_k\right) \tag{2}$$

Where:

— $x = (1, x_1, ..., x_m)$ are the network inputs (independent variables), where 1 corresponds to the bias of a traditional model.

— $w_k = (w_{k1}, ..., w_{km})$ are the weights of the inputs layer neurons to those of the intermediate or hidden layer.

— y_k is the output signal of the neuron (in our case, it refers to fraud probability).

— F is the unit activation function. In this work MLP uses the Hyperbolic Tangent function as activation function for the hidden layers $F(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ and the function $F(x) = \frac{e^x}{1 + e}$ for the output layer. Since we only have the binary variable "potential fraud" that only accounts for 0 or 1 values, the activation function can be written as: $F(1) = \frac{e}{1 + e}$ or $F(0) = \frac{1}{1 + e}$.

— b_k denote the bias has the effect of increasing or lowering the net input of the activation function.

The multilayer perceptron stems from back-propagation error learning. It is the most frequently utilised algorithm and besides it mostly makes use of the backpropagation algorithm, the conjugate gradient descent or the Levenberg-Marquardt algorithm. The advantages of the multilayer perceptron over other procedures can be attributed to the fact that all layers have the same linear structure, thereby rendering it more efficient.

In addition, unlike some other statistical techniques, it does not need to make prior assumptions concerning the data distribution. Another advantage to take into account is that it can model highly non-linear functions and can be trained to accurately generalize when presented with new, unseen data.

c. ***Stage 3: Segmentation of the taxpayer population***

At this stage we will follow McCormick *et al*. (2013) using a decision tree algorithm to obtain a population segmentation based on the probability of potential fraud obtained by the

MLP. Although MLP are strong performers this kind of neural networks do not present an easily-understandable model.

There are various implementations of decision trees we could use to segment the population in terms of the scores produced by MLP. These algorithms mainly differ in the splitting mechanism, that is, the method of finding the optimal partition and the number of new nodes that can be grown from a single node. We use a CHAID algorithm (Chi-square automated interaction detection; (Kass, 1980) and (Biggs *et al*., 1991) for this task. The CHAID uses the Chi-square independence test to decide on the splitting rule for each node and allows splitting into more than two subgroups. The Chi square test is only applicable to categorical data and therefore requires the discretization of all numerical input variables. For each input variable, the classes are merged into a "upper-class", based on their statistical similarity, and maintained if they are statistically dissimilar. These "upper class" variables are then compared to the potential fraud variable for dependency using the Chi-square independence test. The one with the highest significance is then selected as the splitting criteria for the node. The criteria in that case is the p-value of the *F* statistic for the difference in mean values between the g nodes generated by the split:

$$F = \frac{BSS/(g-1)}{WSS/(n-g)} \sim F_{(g-1),(n-g)} \tag{3}$$

Where:

— $WSS = \sum_{j=1}^{g}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$ is known as the residual or within sum square in analysis of variance, which is independent of the split variable. $\bar{y}_j$ is the mean value of the y_{ij} variables in node j and $g = 2$ groups that would be produced by the split.

— $BSS = TSS - WSS$, being $TSS = \sum_{j=1}^{g}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2$ the total sum of squares (before the split) and BBS the resulting between sum of squares.

For more details on how CHAID works see Ritschard (2013).

d. ***Stage 4: Characterization of potential fraud taxpayers***

The objective of this stage is to characterize in terms of socioeconomic variables the taxpayers in those segments with high probability of potential fraud. To achieve it, we use descriptive statistics methods such as comparative histograms or non-parametric contrasts to test the hypothesis of equality of the distributions between segments.

The Mann-Whitney U test (also called Wilcoxon rank-sum test, or Wilcoxon-Mann-Whitney test) is a nonparametric test of the null hypothesis that it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample. As a non-parametric test, it does not require the assumption of Normal

distribution and it is nearly as efficient as the T-test on Normal distributions. It was initially proposed by Wilcoxon (1945) for samples of equal sizes and extended to samples of arbitrary size as in other ways by Mann and Whitney (1947). Such as:

$$U = min(U_1, U_2) \tag{4}$$

where $U_1 = n_1 \cdot n_2 + \frac{n_1(n_1+1)}{2} - R_1$ and $U_2 = n_1 \cdot n_2 + \frac{n_2(n_2+1)}{2} - R_2$, being n_1 and n_1 the sample sizes of each of the two samples and R_1 and R_2 the sum of the ranges of each of the two samples. If $n_1 > 10$, $n_2 > 10$ the test can be approximated as:

$$Z = \frac{U - (n_1 \cdot n_2/2)}{\sqrt{\frac{n_1 \cdot n_2(n_1+n_2+1)}{12}}} \sim N(0,1) \tag{5}$$

The application of this test to the financial variables of our study (which do not follow a Normal distribution) yields results that allow us characterizing those segments of the population with the highest probability of fraud in terms of financial variables.

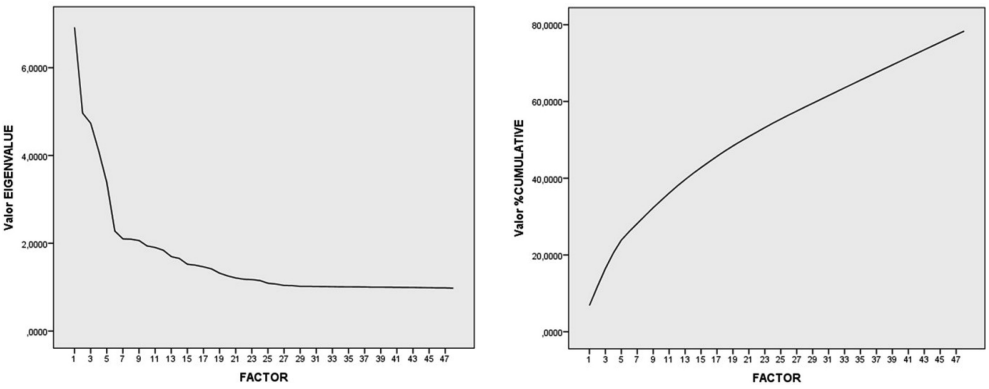
Below are the results obtained with the application of these four stages to the PIT sample.

5. Empirical application

a. ***Stage 1: Dimension reduction.***

Applying PCA to a set of more 469 variables related to the sample, we select the first 48 principal components which are able to explain just under 80% of the total variance of the initial group.

Graph 1
EIGENVALUES TABLE AND SCREE PLOT TABLE FROM PCA



Source: SPSS modeler outputs.

The initial communalities, i.e. the amount of common variance of the variables explained is 1. That is because PCA is based on the assumption that the whole variance can be explained by the factors. In Graph 1 we can find the eigenvalue corresponding to each factor and the % cumulative variance explained (scree plot).

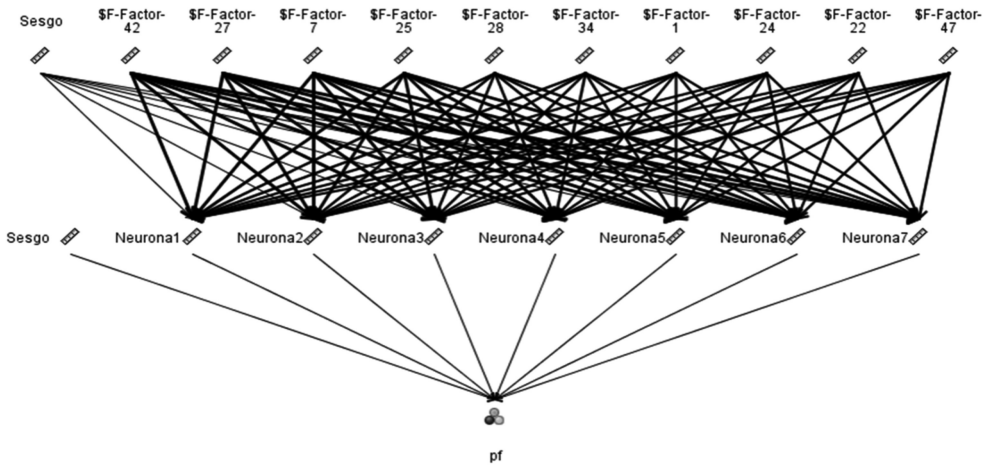
Eigenvectors are determined to transpose the given data. Every eigenvector has an eigenvalue that measures the amount of variance which is in the data in the direction of the eigenvector. The Scree Plot visualizes the number of components versus the cumulated percentage of variance explained. The results are based on the un-rotated components. As shown in the Graph 1, with 48 factors we are able to explain just under 80% of the total variance of the group.

In next section we introduce the 48 principal components in the input layer of the neural network.

b. ***Stage 2: Scoring taxpayers***

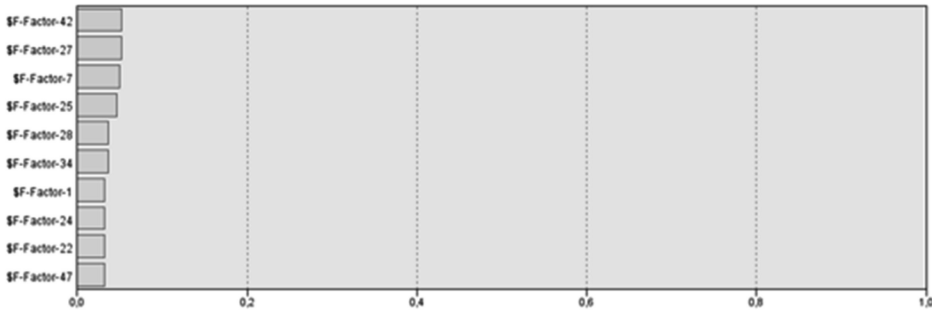
One of the main objectives of the paper named in the introduction is to identify those PIT items in the tax model structure (D-100) most likely to be used by potential fraud taxpayers. For this, we introduce the 48 principal components in the input layer of the neural network and as result we see in Graph 2 shows the architecture of the neural network that we will use to assign a probability (score) to each taxpayer based on its similarity with the group labeled as potential fraud taxpayers. Furthemore, the Graph 3 shows the predictor importance in the MLP, where that most important input variables are factors 42, 27, 7, 25, 28, 34, 1, 24, 22 y 47.

Graph 2
MLP STRUCTURE GRAPH WITH A HIDDEN LAYER USING THE FIRST 48-TH PRINCIPAL COMPONENTS AS INPUT VECTOR AND THE POTENTIAL FRAUD (pf) AS THE OUTPUT VECTOR. THERE ARE 7 NEURONS IN THE HIDDEN LAYER



Source: SPSS modeler outputs.

Graph 3
PREDICTORS IMPORTANCE IN THE MLP

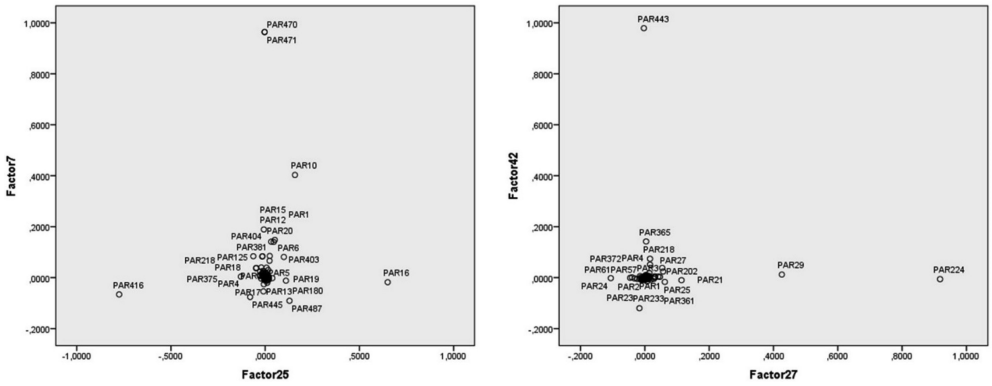


Source: SPSS modeler outputs.

Additionaly, we show the Varimax rotated factor loadings, in Graph 4, where most important predictors are those input variables that score significantly higher than the others on the axis associated to the factor. Graph 4 shows that factors 25, 7, 27 and 42 are strongly influenced by the input variables classified in Table 3, as deduced from its score on the axes. The information in the Table 3 is obtained by measuring the distance to the origin of coordinates of the corresponding PAR variables on the axis that represents the factor in graph4. The initial PAR variables that have a higher score on the axis of the factor are the most influential variables in the factor. Definitely, the result collected in Table 3, offer the list of initial variables that best discriminate those to identified as potential evaders.

Regarding factor 42 we found that the exempted income corresponding to the taxable base of savings (but taken into account for the calculation of the tax rate) scores very high on the axis corresponding to the factor. It is usually income obtained abroad and which by virtue of Agreements to avoid double taxation is declared exempt in the Spanish territory.

Graph 4
ROTATED FACTOR PATTERN FOR THE MOST IMPORTANT PREDICTORS IN THE MLP



Source: SPSS modeler outputs.

Table 3
MOST IMPORTANT FACTORS USED BY THE MULTILAYER PERCEPTRON

Factor 42	PAR443: Exempted income corresponding to the taxable base on savings.
	PAR 365: Negative net balance resulting from all capital gains and losses corresponding to 2013 treated as taxable income, with the limit of 10% of the net balance of the yields to be included in the general tax base plus income allocations.
Factor 27	PAR224: Income allocation from entities under the international fiscal transparency system (art.91 of the Spanish PIT Law).
	PAR29: movable capital yields to be integrated into the taxable base of savings. Tax-deductible expenses: deposit and administration of negotiable securities.
Factor 7	PAR 470: Benefits for investments in primary residence, State part.
	PAR 471: Benefits for investments in primary residence, Regional part.
Factor 25	PAR 16: Benefits for labor income (regulated in art. 20 of the Spanish PIT Law) whose annual net labor income is lower than 13.260 euros.
	PAR 487: Tax credit for the rent of the taxpayer's usual residence.
	Par 180: Net reduced return on economic activities: agricultural, livestock or forest in objective estimation.
	Par19: Additional benefits applied to income from work for active workers who are disabled.

Source: Own production.

As far as factor 27 is referred, the allocation of income in the international tax transparency regime is regulated as a special regime for the PIT Spanish regulation. This regime has its origins in the operations of fiscal engineering derived from the constitution of companies abroad (normally in what is known as tax havens or territories of low or no taxation) with the sole purpose of a transfer of income towards these companies by their partners, residents in Spanish territory, who, in this way, avoid the payment of the Spanish tax, so that the tax treatment is much more beneficial.

In factor 7, there are two variables strongly associated to the factor relative to tax credits for primary residence (state part and regional part). This tax advantage, which could no longer be applied for acquisitions made after 2013, is the one with the highest number of beneficiaries in the personal income tax, and has been one of the most questioned regarding tax fraud. Tax credits for primary residence fall into the category of tax fraud as long as they do not correspond exactly with the definition of primary residence. The habitual frauds in this benefit are generally related to a second residence that the taxpayer lists as primary residence or to the benefits for mortgages made on habitual dwellings to finance other expenses, such as reforms or the purchase of a car.

Concerning factor 25, the application of the benefits for labor income (regulated in art. 20 of the Spanish PIT Law) for those taxpayers whose annual net labor income is lower than 13,260 euros scores very high on the axis corresponding to the factor. In order to be entitled to this benefit, taxpayers may not have income excluding exempt income, other than that from work. This is a source for tax evasion as incomes are not accurately reported. For these purposes, the tax authorities conclude that the arithmetic sum of the different sources of income must be understood as such: income, capital gains and losses and income imputations, positive and negative for the year, referring to net income, i.e., prior to the application of any type of benefits established in the tax regulations, and without the legally established compensation limits being applicable to form the tax bases.

It is important to highlight that some items mentioned in the previous paragraphs, although significant in terms of neural network scores, partly because they reduce the amount to pay by the taxpayer, do not provide a real opportunity to commit fraud. This is because the control and verification of these items by the tax authority is quite simple and immediate, as it has information provided by third parties. Examples of these tax items are par19, par29 and par365. This is undoubtedly one of the limitations imposed by using the classification rule instead of working with the real sample of fraudsters, as would be desirable in this analytical framework.

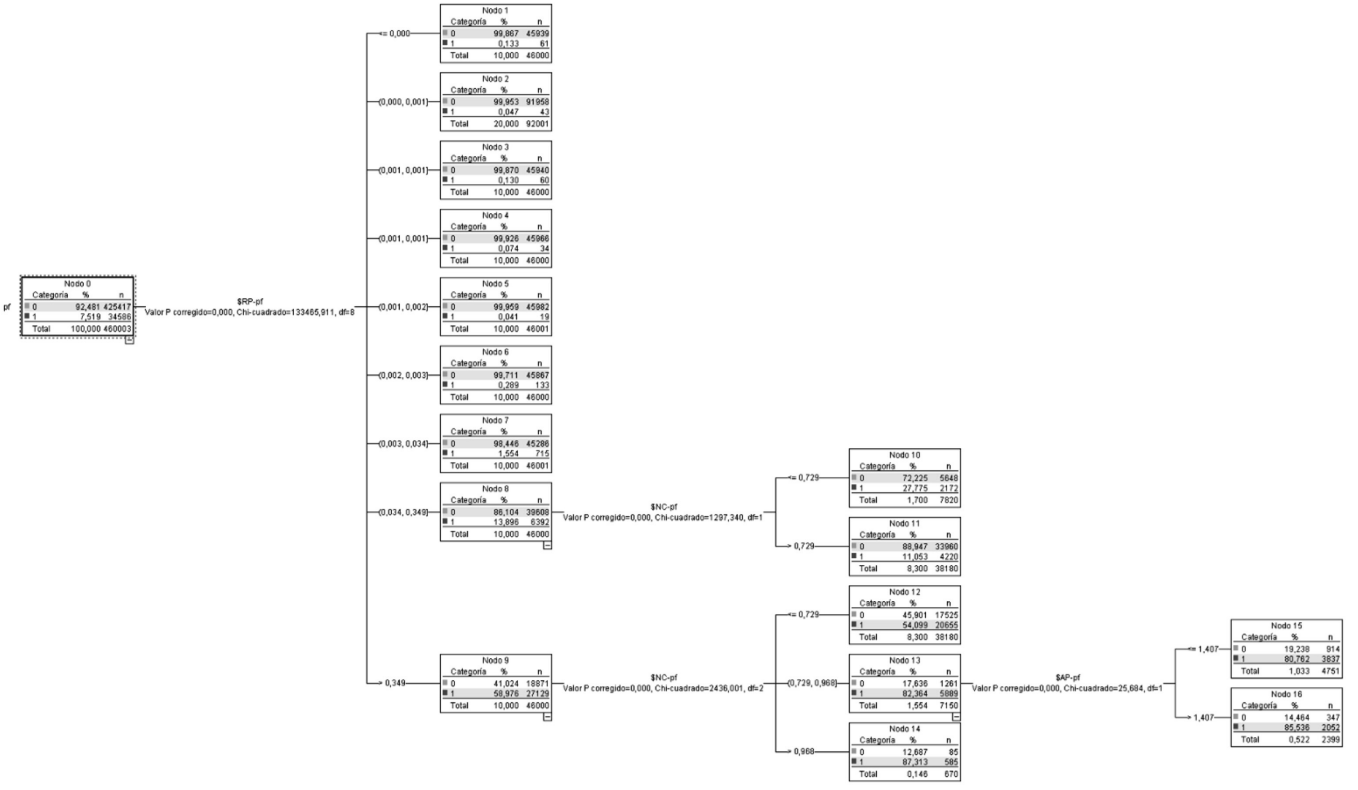
Once we have identified those tax items that are most important to assign a fraud probability to each taxpayer (based on our training sample), the next stage is to segment the population of taxpayers based on that probability of fraud. The objective is to be able to characterize those segments of the population with the highest probability of fraud. This characterization might be useful to focus audit and inspection tasks.

c. ***Stage 3: Segmentation of the taxpayer population***

In this section the objective is to assign each tax filler to a category or segment of the population grounded on their probability of fraud calculated by the neural network based on the information contained in the D-100 form. In this way, as soon as a declarant fills in the boxes of the D-100 form, it can be assigned to a segment or category, which differs from the rest by its probability of fraud. This probability is calculated following the information patterns of a training group, ideally real fraudsters, but in our case and in the absence of that reliable information we have generated "artificially" through a classification rule.

The result of CHAID structure for our analysis is presented in Graph 5. We are especially interested in nodes 14, 15 and 16 whose potential fraud percentages are between 80 and 87%. From now on, the name of the population segments will be that of the nodes of the CHAID tree that generated them. Next subsection explores the characterization of these population segments with high percentage potential fraud records, nodes 14, 15 and 16 (potential fraud percentage between 80 and 87%) with respect to those which, on the contrary, have a low potential fraud percentage (less than 0.2%: segments 5 and 6).

Graph 5
SEGMENTATION OF THE 2013 SPANISH PIT SAMPLE ACCORDING TO POTENTIAL FRAUD - PROBABILITIES ASSIGNED BY MLP(*)



(*) NC_PF is the probability of '0' value in potential fraud assigned by MLP.
RP_PF is the probability of '1' value in potential fraud assigned by MLP.

Source: SPSS modeler outputs.

d. ***Stage 4: Characterization of potential fraud taxpayers***

Once the segmentation of taxpayers has been developed, we are able to characterize potential evaders according to those variables whose distribution varies significantly between those segments which have a high percentage of potential fraud taxpayers and the rest of the sample. We show the differences in income, province, age, marital status and benefits for investments in main residence, which are the most influential variables in our model. We will focus on low percentage potential fraud segments (5 and 6 in graph 5) compared to high percentage potential fraud segments (14, 15 and 16 in Graph 5).

We use two ways to characterize the differences in the distribution of the variables: on the one hand we use a graphical approach through comparative histograms between segments; on the other hand, we use a non-parametric approach to test the hypothesis of equality of the distributions of both populations. The analyzed graphs are grouped in the Annex (A.1-A.10). This characterization is the second objective of this study, since establishing a profile of a potential fraud taxpayer is a useful instrument for selecting tax returns for audit tasks.

d.1. *Looking for differences between segments: a graphical approach*

One of most discriminant variables is the *income group*. Segments 5 and 6 (Graph 5) with low percentage records of potential fraud (less than 0.2%) show more taxpayers in low income groups. The descriptive analysis in Graph A.1 argues that the statistical distribution of the variable income on the population segments is different in those taxpayers labeled as potential fraud taxpayers. We see that those taxpayers placed in the potential fraud group are mainly concentrated in segments of income 11 (income between 54,000 and 60,000 euros per year) and 12 (more than 60,000 euros).

Regarding the province of *fiscal address*, the taxpayers in Madrid (28) and Barcelona (8) stand out, whose percentage is higher in the two groups, but which are much more prominent in the potential fraud class. Graphs A.2-A.4 in the Annex show the histograms of the province variable in classes potential *fraud = 0* versus potential *fraud = 1* for the whole PIT sample. The bins in the histogram corresponding to Madrid (28) and Barcelona (08) accumulate many more frequencies of observations high percentage potential fraud segments. The frequency distribution of province variable is more biased about Madrid and Barcelona provinces in this segment (80% of potential fraud).

Graphs A.5-A.6 show the histograms of the *marital status* variable in class *potential fraud = 0* versus *potential fraud = 1* for the whole PIT sample. The single category (Marital Status '1') is much less frequent in the high percentage potential fraud segment compared to segments 5 and 6. This distribution is similar to those segments with high percentage potential fraud population. As we see between the two classes (potential fraud taxpayers versus non potential fraud taxpayers), the percentage of married people for the population labeled as potential fraud rises with respect to non-potential fraud to the detriment of single people.

With regard to *age variable*, we observe that although the variable follows a Normal distribution in both groups (potential fraud versus non potential fraud), the values of kurtosis and skewness are very different between them (Graph A.7). The Kurtosis of the distribution in segment 15 leads to focus the attention on population with ages between 40 and 55 years old in 2013 (Graph A.8).

The *sex variable* presents a markedly different distribution in the group of taxpayers labeled as "potential fraud". In this group there is a greater imbalance between men and women, the percentage of men being much higher than in the non-potential fraud group (Graph A.9). If we have a look at the tax fillers segments, the distribution of frequencies between the two categories of sex variable is more balanced in segments with low percentage of potential fraud (5 y 6, first row in Graph A.10). On the contrary, in segments with high percentage of potential fraud (11, and 12, second row) the imbalance between men and women is more evident and biased toward men.

Equally important are the differences related to *benefis for investment in primary residence*. According to Graph A.13 both segments with low percentage of potential fraud (5 and 6, first row) accumulate all the records in the '0' bin. On the contrary, for the high potential fraud segments, more than 70% of the taxpayers accounts for deductions of investment in main residence between 600 and 800 euros in 2013, and approximately 15% accounts for benefits between 800 and 1,000 euros. Figures for the total sample are represented in Graphs A.11 and A.12 with the same conclusions.

Graphs A.1-A.13 reveal that high percentage of potential fraud segments are significantly different and accumulates more records in the categories of male, ages between 40-55, married, the province of residence Madrid and Barcelona, income greater than 54,000 euros (sum of net incomes, imputed income and capital gains and losses greater than 54,000 euros, ie, income groups 11 and 12) and benefits on investment in main residence between 600 and 1,000 euros.

Next subsection formalizes differences between segments using non-parametric test to contrasts population homogeneity in financial variables.

d.2. *Testing for differences between segments in financial variables: a non-parametric approach*

In this section we formalize that there are significant differences in the distribution of certain input financial variables depending on the segment of taxpayers. This is equivalent to testing whether two samples (segments of taxpayers) are pulled from the same population. The first point to keep in mind is that most of these financial variables contained in the Spanish PIT sample (and analyzed in previous sections) do not follow a Normal distribution, and therefore, the application of any parametric approach (the T-Student approach for example, where the distribution functions were assumed to be known Normal) for this purpose is open to very serious objections. This is the reason why we are going to use a non-parametric approach.

The following Table 4 contains those financial variables that showed significant differences between segments.

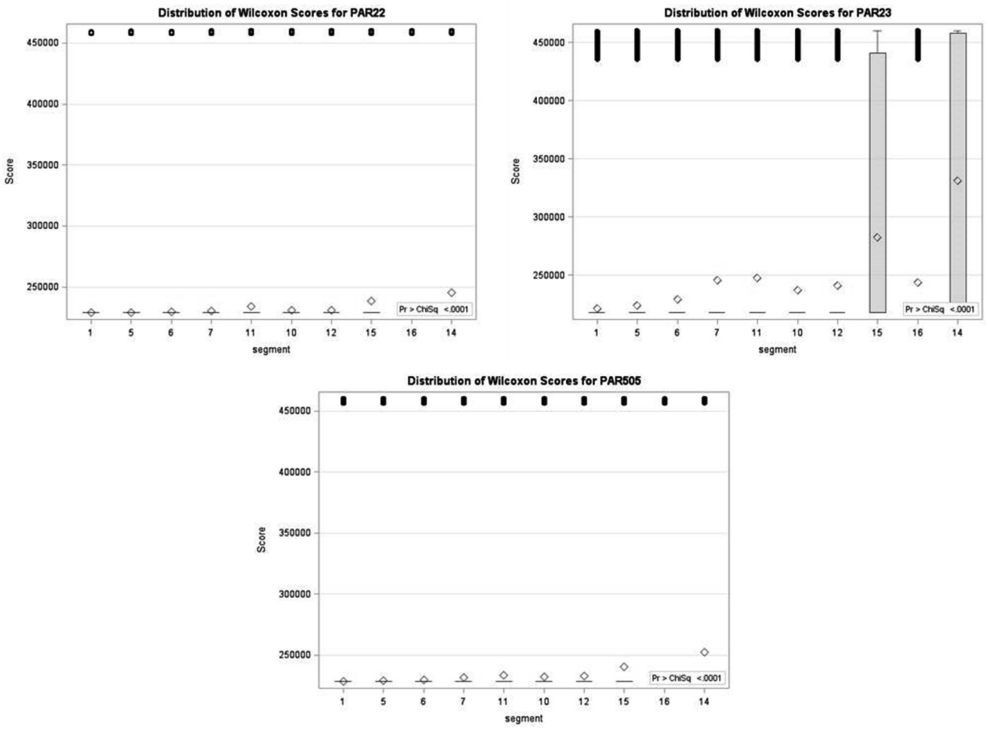
Table4
FINANCIAL VARIABLES THAT HOWED SIGNIFICANT DIFFERENCES BETWEEN SEGMENTS

Variable	Label
Par22	Interest on financial assets entitled to the bonus provided for the 11th Provisional Regulation of the Corporate Income Tax Law.
Par23	Dividends and other income from participation in equity of entities.
Par505	Benefits for double international taxation, due to the income obtained and taxed abroad.

Source: Own production.

As seen in the distribution of Wilcoxon scores for variables for Table 4, in Graph 6, segments with high percentage of potential fraud have higher means in these three variables. This result makes us focus attention on those taxpayers with investments abroad who have these investments as sources of income and take advantage of benefits for double international taxation, due to the income obtained and taxed abroad.

Graph 6
DISTRIBUTION OF WILCOXON SCORES FOR VARIABLES: PAR22, PAR23 AND PAR505



Source: SPSS modeler outputs.

6. Conclusions

Researching on tax fraud is a challenging and daunting task, due to the complexity of the patterns involved and the size of the data sets. In this study, a combination of data mining techniques is proposed to explore the probability of "potential fraud", a binary target variable built from the analysis of Tax Liabilities and Tax Credits for every income segment in the Spanish Personal Income Tax.

The study involves a methodology aimed at, from the information obtained from the boxes of the D-100 PIT form, assigning a segment or category of tax filers according to the fraud probability. The entire methodology is built to be "trained" from a training group of "real fraud taxpayers". As in this specific case we do not have the real training group, we have applied the assignment rule and we have obtained as a result the group we have identified as "potential tax evaders". If in future there is reliable information about a group of tax fillers having fraudulent information declared on D-100 form, that group would be taken as the real training group for the model and the Potential Fraud variable would be taken as observed.

One of the problems of this approach is the large number of predictor variables which are multicolinear in nature. In our particular case, we have used 469 variables related to the D-100 form related to the Spanish Personal Income Tax in 2013. Applying Principal Components Analysis as a dimension reduction technique we have been able to obtain 48 orthogonal regressors that best summarize 80% of the total variance of the group. These principal components are the independent variables that are taken into account for predicting the probability of potential fraud using the Multilayer Perceptron (MLP) model. With the challenge of providing more information related to the initial input variables and their relationship with the MLP scores, we have used the probabilities obtained by the MLP as input variables for a CHAID algorithm, to produce a taxpayers segmentation in terms of potential fraud percentage.

The application of the method to the 2013 sample showed that most interesting input variables used by the Multilayer Perceptron to assign potential fraud scores are Tax Credits expenses: deposit and administration of negotiable securities, Tax Credits for investments in main residence and Rental deduction: benefits for the rent of the taxpayer's usual residence. These tax credits are under revision what confirms the interest of the proposed method. Additionally, the Tax Agency intends to continue giving impetus to various projects that increase the availability of international tax information to reduce fraud, which are among others:

— The automatic exchange of financial account information abroad with ownership of residents in Spain has been generalized to the extent that in 2018 it has meant that a very significant number of jurisdictions have been incorporated into the Common Reporting Standard project developed by the OECD and promoted by the Global Transparency Forum on Information Exchange.

— The receipt of information regarding the so-called "Country-by-Country Report" as part of the OECD/G20 Project on the erosion of the tax base and the transfer of benefits (BEPS "Base Erosion and Profit Shifting).

These two measures are compatible with the results of the non-parametric analysis of the financial variables in this study in which is recommended to focus attention on taxpayers with investments abroad who have these investments as sources of income and take advantage of deductions for double international taxation, due to the income obtained and taxed abroad.

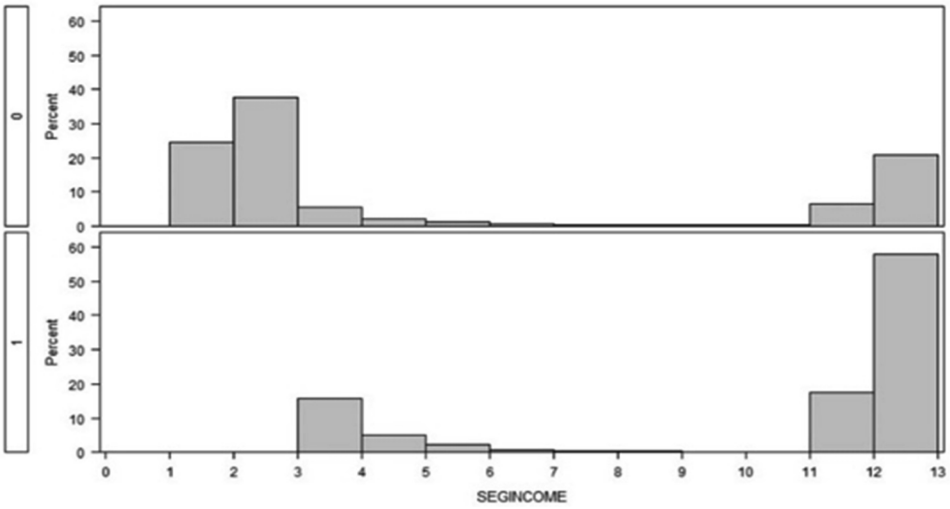
Our proposal also allows potential fraud taxpayers to be characterized in terms of socio-economic variables, results are in line with work like the presented by Domínguez *et al*. (2017), we find that the lowest degree of compliance is found in Barcelona and Madrid and people in the top part of the earnings distribution are found much less compliance.

In this paper, we have presented a proposal that allows Tax Administration to prioritize 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 with tax benefits like the Spanish Corporate Tax. Future work includes the evaluation of this proposal also in the panel of Personal Income Taxpayers and in more recent Spanish PIT samples to compare results.

Annex

Graph A.1
DISTRIBUTION OF POTENCIAL FRAUD BY INCOME

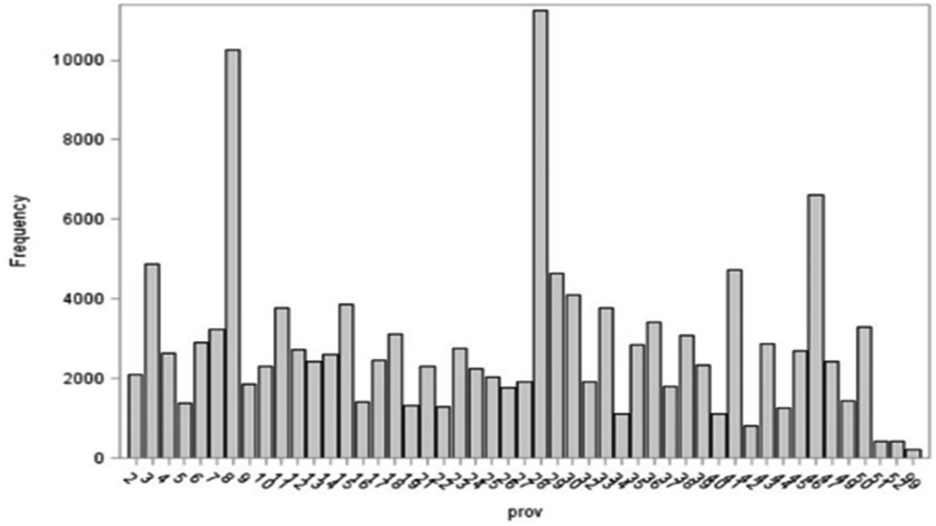
Comparative Analysis of income segment in potential fraud (1) versus not potential fraud (0) classes



Source: SPSS modeler outputs.

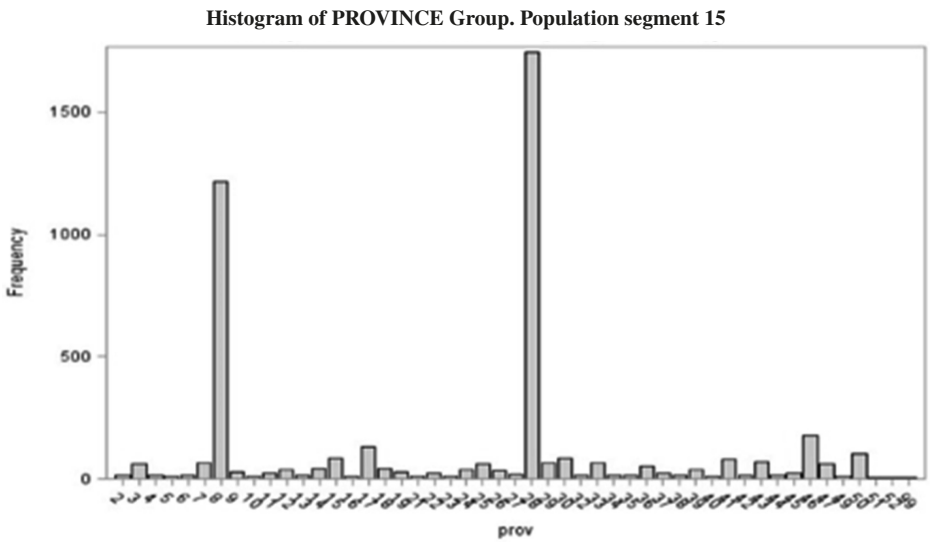
Graph A.2
DISTRIBUTION OF POTENCIAL FRAUD BY PROVICE IN SEGMENT 5
(99.9% of non potencial fraue)

Histogram of PROVINCE Group. Population segment 5



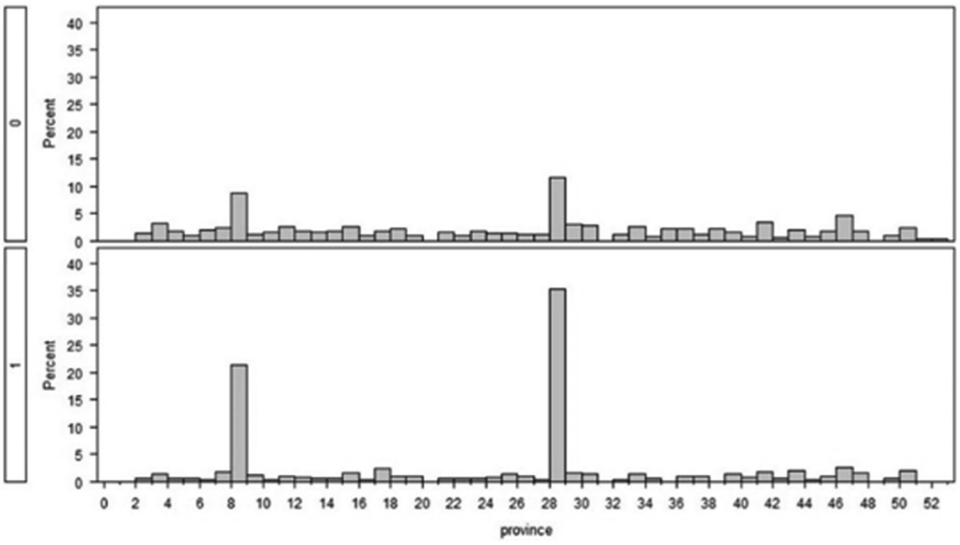
Source: SPSS modeler outputs.

Graph A.3
DISTRIBUTION OF POTENCIAL FRAUD BY PROVINCE
(in segment 15-80% of non potencial fraud)



Source: SPSS modeler outputs.

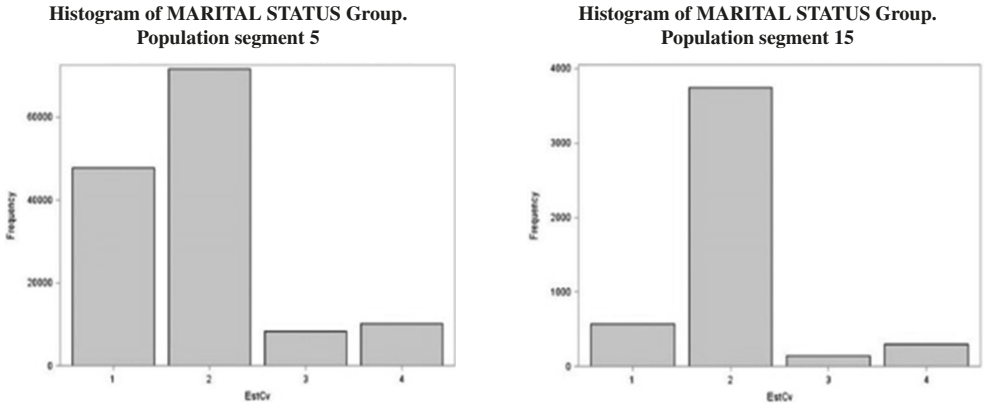
Graph A.4
DISTRIBUTION OF POTENCIAL FRAUD BY PROVINCE



Source: SPSS modeler outputs.

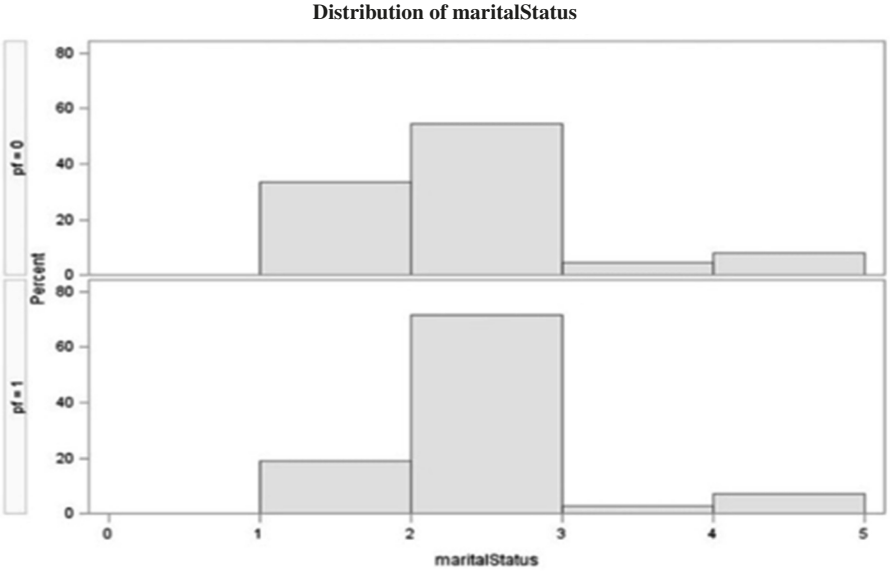
Graph A.5
DISTRIBUTION OF POTENCIAL FRAUD BY MARITAL ESTATUS IN SEGMENT 5 (left) AND SEGMENT 15 (right). THE CATEROGIES ARE: 1. SINGLE, 2. MARRIED, 3. WIDOWED AND 4. DIVORCED OR LLEGALLY SEPARATED

(99% of non potencial fraud) *(80% of potencial fraud)*



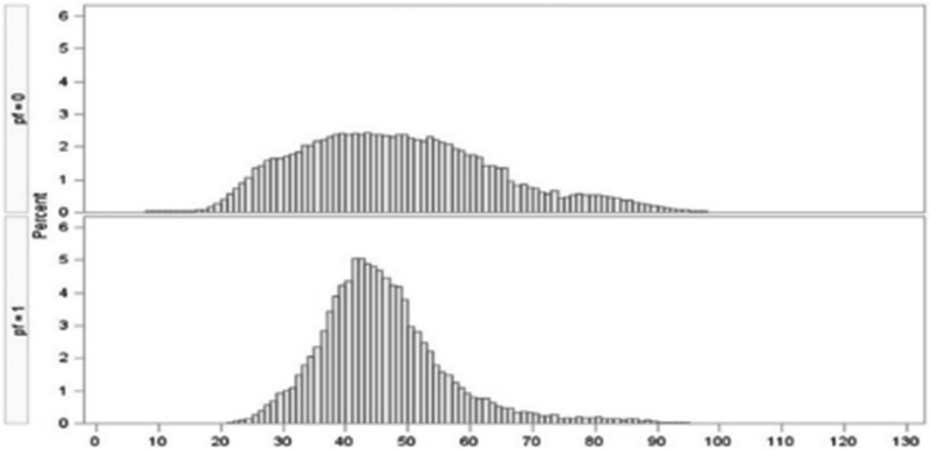
Source: SPSS modeler outputs.

Graph A.6
DISTRIBUTION OF POTENCIAL FRAUD BY MARITAL STATUS



Source: SPSS modeler outputs.

Graph A.7
DISTRIBUTION OF POTENCIAL FRAUD BY AGE
Comparative Analysis of age in potential fraud (1) versus not potential fraud (0) classes
The UNIVARIATE Procedure
Distribution of age

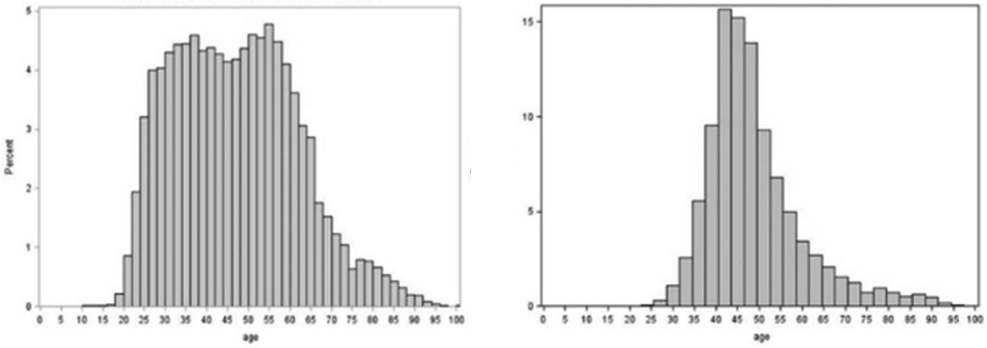


Source: SPSS modeler outputs.

Graph A.8
DISTRIBUTION OF AGE IN SEGMENT 6 (left) AND SEGMENT 15 (right)

(99% of non potencial fraud) *(80% of potencial fraud)*

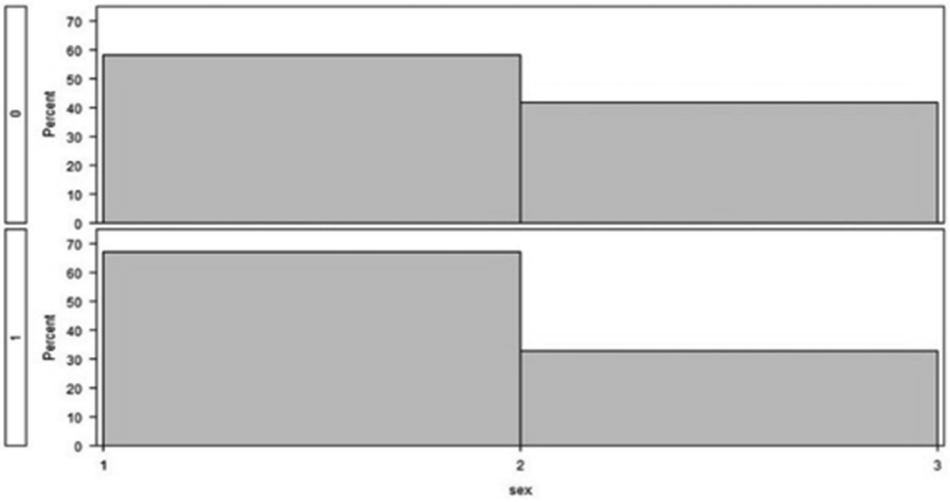
Histogram of AGE. Population segment 6 **Histogram of AGE. Population segment 15**



Source: SPSS modeler outputs.

Graph A.9
DISTRIBUTION OF POTENCIAL FRAUD BY SEX

Comparative Analysis of SEX in potential fraud (1) versus not potential fraud (0) classes

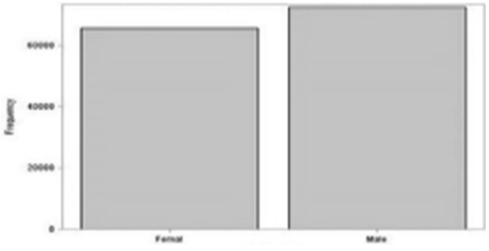


Source: SPSS modeler outputs.

Graph A.10
DISTRIBUTION OF SEX IN SEGMENT 6 (left) AND SEGMENT 15 (right)

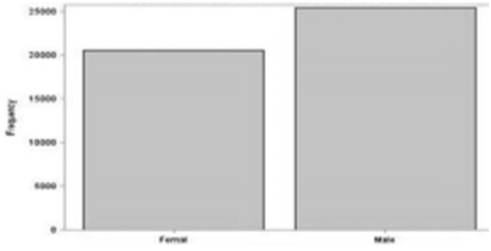
(99% of non potencial fraud)

Bar chart variable SEX. Population segment 5



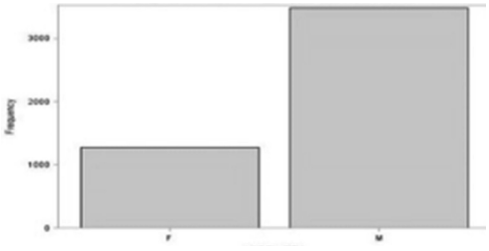
(80% of potencial fraud)

Bar chart variable SEX. Population segment 6

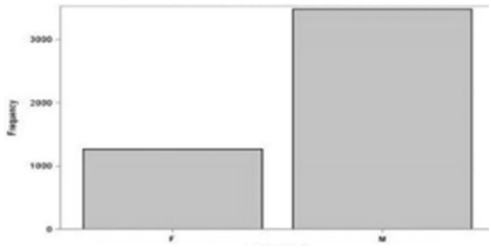


(85% of potencial fraud)

Bar chart variable SEX. Population segment 15



Bar chart variable SEX. Population segment 16

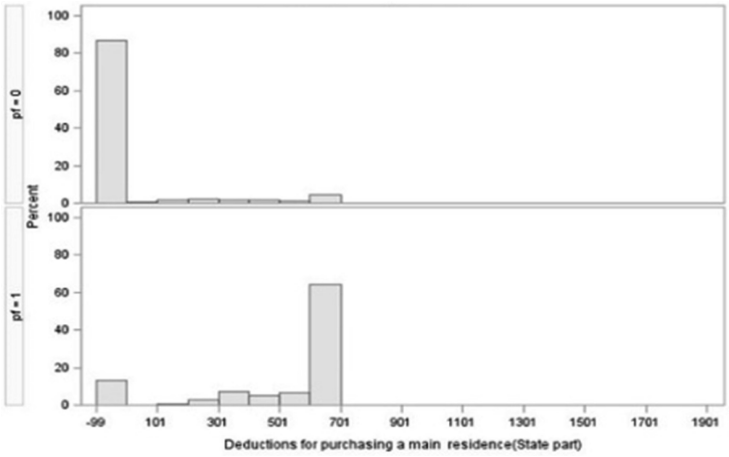


Source: SPSS modeler outputs.

Graph A.11
DISTRIBUTION OF POTENCIAL FRAUD BY DEDUCTION FOR INVESTMENT IN MAIN RESIDENT (State part)
Comparative Analysis of deductions for investments in primary residence state part in potential fraud (1) versus not potential fraud (0) classes

The UNIVARIATE Procedure

Distribution of par470m

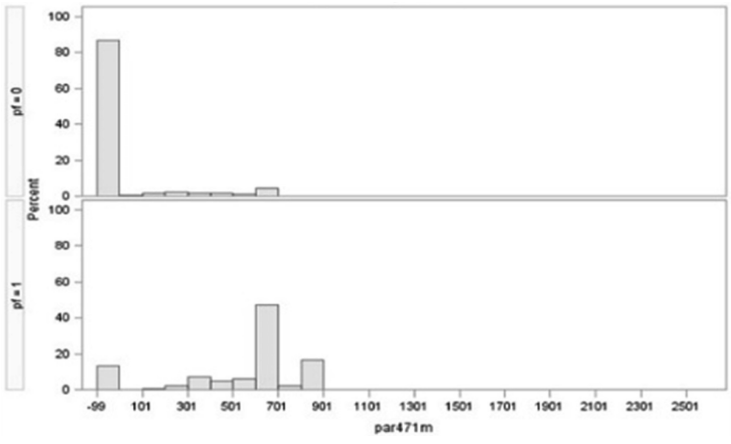


Source: SPSS modeler outputs.

Graph A.12
DISTRIBUTION OF POTENCIAL FRAUD BY DEDUCTION FOR INVESTMENT IN MAIN RESIDENT (Regional part)
Comparative Analysis of deductions for investments in primary residence regional part in potential fraud (1) versus not potential fraud (0) classes

The UNIVARIATE Procedure

Distribution of par471m



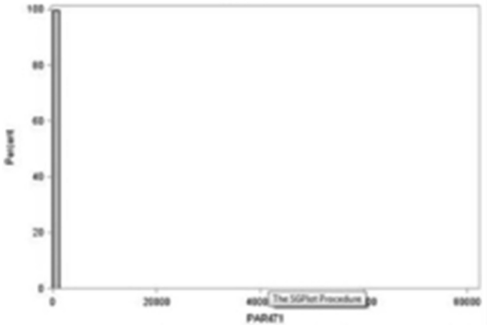
Source: SPSS modeler outputs.

Graph A.13
DISTRIBUTION OF POTENCIAL FRAUD BY DEDUCTION FOR INVESTMENT IN MAIN RESIDENT

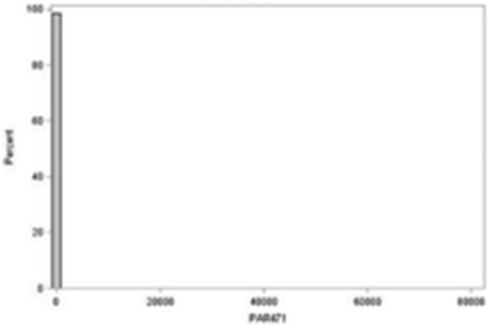
Segment 5 ***Segment 6***

(99,9% of non potencial fraud)

Histogram of DEDUCTION OF INVESTMENT IN MAIN RESIDENCE. Population segment 5



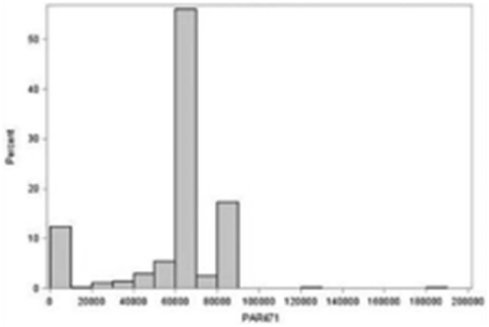
Histogram of DEDUCTION OF INVESTMENT IN MAIN RESIDENCE. Population segment 6



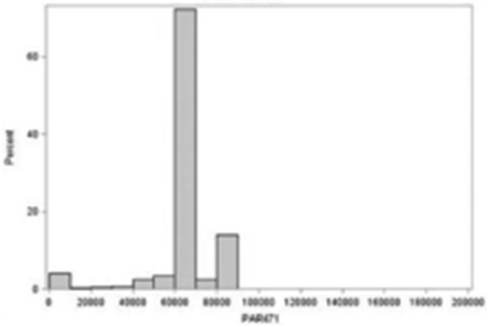
Segment 15 ***Segment 16***

(80% of potencial fraud) *(85% of potencial fraud)*

Histogram of DEDUCTION OF INVESTMENT IN MAIN RESIDENCE. Population segment 15



Histogram of DEDUCTION OF INVESTMENT IN MAIN RESIDENCE. Population segment 16



Source: SPSS modeler outputs.

Notes

1. See Hernández de Cos and López-Rodríguez (2014), López-Rodríguez and García-Ciria (2018) and García-Miralles *et al*. (2019) for descriptions of Spanish Personal Income tax in the context of the European Union and the OECD.

2. Other tax fraud possibilities are individuals who do not file tax returns and those who under-report income or overclaim deductions.

3. The design of the register of the microdata PIT sample file as well as the meaning of each of the variables named in this study are publicly available. The Personal Income Tax Sample is based on tax returns from the corresponding tax model (D-100) filled yearly by taxpayers in the Common Fiscal Territory of Spain. All variables in the sample follow the tax model structure (D-100) and its name consists of the prefix PAR followed by sequential number according to the tax item of the form. The registry design of the sample is easily available on the annex I of Pérez *et al*. (2016).

4. A description is available at the document IBM (2016) can be found on this link: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerApplications.pdf.

References

Alm, J. (2011), "Measuring, explaining and controlling tax evasion: lessons from theory, experiments, and field studies", *International Tax and Public Finance*, 19(1): 54-77.

Biggs, D., De Ville, B. and Suen, E. (1991), "A method of choosing multiway partitions for classification and decision trees", *Journal of Applied Statistics*, 18(1): 49-62.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Microsoft Research, Cambridge, U. K.

Bonchi, F., Giannotti, F., Mainetto, G. and Pedreschi, D. (1999), "Using Data Mining Techniques in Fiscal Fraud Detection", in: Mohania M. and Tjoa A. M. (eds.), *DataWarehousing and Knowledge Discovery. Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg.

De Roux, D., Pérez, B., Moreno, A., Villamil, M. P. and Figueroa, C. (2018), "Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach", *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Dias, A., Pinto, C., Batista, J. and Neves, M. E, (2016), "Signaling Tax Evasion, Financial Ratios and Cluster Analysis", *Working Paper* 51, OBEGEF, Observatorio de Economia y Gestáo de Fraud, Coimbra.

Domínguez, F., López, J. and Rodrigo F. (2014), "El hueco que deja el diablo: Una estimación del fraude en el IRPF con Microdatos Tributarios", EEE2014-01, Madrid.

Domínguez, F., López, J. and Rodrigo F. (2015), "Fraude en el IRPF por fuentes de renta, 2005-2008: del impuesto sintético al impuesto dual", EEE2015-14, Madrid.

Domínguez F., López J. and Rodrigo, F. (2017), "Tax evasion in Spanish Personal Income Tax by income sources, 2005-2008", *European Journal of Law and Economics*, 44: 47-65.

EU (2019), "Taxation trends in the European Union", *DG Taxation and Customs Union*, Publication Office of the European Union, Luxembourg.

Fine, T. L. (1999), "Feedforward Neural Network Methodology", 3rd ed. New York: Springer - Verlag.

Fox, W. F., Luna, L. and Schaur, G. (2014), "Destination taxation and evasion: Evidence from US inter-state commodity flows", *Journal of Accounting and Economics*, 57(1): 43-57.

García-Miralles, E., Guner, N. and Ramos, R. (2019), "The Spanish Personal Income Tax: facts and parametric estimates", *Working Paper* 1904, CEMFI, Madrid.

González, P. C. and Velásquez, J. D. (2013), "Characterization and detection of taxpayers with false invoices using data mining techniques", *Expert Systems with Applications*, 40(5): 1427-1436.

Haykin, S. (1998), *Neural Networks: A Comprehensive Foundation*, 2nd ed., New York: MacmillanCollege Publishing.

Hernández de Cos, P. and López-Rodríguez, D. (2014), "Estructura impositiva y capacidad recaudatoria: un análisis comparativo con la UE", *Documentos ocasionales del Banco de España* 1406, Banco de España.

Hotelling, H. (1933), "Analysis of a complex of statistical variables into principal components", *Journal of Educational Psychology*, 24(6): 417.

IBM (2016), *IBM SPSS Modeler 18.0 Algorithms Guide*, IBM Corporation, USA.

Kass, G. V. (1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data", *Applied Statistics*, 29(2): 119-127.

Liao, S. H., Chu, P-H. and Hsiao, P-Y. (2012), "Data mining techniques and applications - A decade review from 2000 to 2011", *Expert Systems with Applications*, 39(12): 11303-11311.

López-Rodríguez, D. and García-Ciria, C. (2018), "Estructura impositiva de España en el Contexto de la Unión Europea", *Documentos Ocasionales* 1810, Banco de España, Madrid.

Mann, H. B. and Whitney, D. R. (1947), "On a test of whether one of two random variables is stochastically largers than the other", *Annals of Mathematical Statistics*, 22: 125-128.

Matos, T., de Macebo, J. A. and Monteiro, J. M. (2014), "An empirical method for discovering tax fraudsters: a real case study of Brazilian fiscal evasion", *Proceedings of the 19th International Database Engineering & Applications Symposium*, New York.

McCormick, K., Abbott, D., Brown, M. S., Khabaza, T. and Mutchler, S. R. (2013), *IBM SPSS modelercookbook*, Packt Publishing.

Micci-Barreca, D. and Ramachandran, S. (2004), "Improving tax administration with data mining" *White paper*, Elite Analytics LLC.

Parlos, A. G., Chong, K. T. and Atiya, A. F. (1994), "Application of the recurrent multilayer perceptron in modeling complex process dynamics", *IEEE Transactions on Neural Networks*, 5(2): 255-266.

Pearson, K. (1901), "LIII. On lines and planes of closest fit to systems of points in space", *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559-572.

Pérez, C., Villanueva, J., Burgos, M. J., Bermejo, E. and Chairi, L. K. (2016), "La muestra del IRPF de 2013: descripción general y principales magnitudes", Instituto de Estudios Fiscales, *Working Paper* 9/2016.

Pérez, C., Delgado, M. J. and de Lucas, S. (2019), "Tax fraud detection through neural networks: an application using a sample of Personal Income Taxpayers", *Future Internet*, 11(4): 86-95.

Picos Sánchez, F. (2014), "Microdatos fiscales en España: caracterización y aplicaciones prácticas", Riuma, Repositorio Institucional de la Universidad de Málaga.

Ravisankar, P., Ravi, V., Raghava Rao, G. and Bose, I. (2011), "Detection of financial statement fraud and feature sele,ction using data mining techniques", *Decision Support Systems*, 50(2): 491-500.

Ritschard, G. (2013), "CHAID and Earlier Supervised Tree Methods", in: J.J. McArdle and G. Ritschard (eds.), *Contemporary Issues in Exploratory DataMining in Behavioral Sciences*, Routeledge, New York, 48-74.

Ripley, B. D. (1996), "Pattern Recognition and Neural Networks", Cambridge Univ. Press, Cambridge, G. B.

Slemrod, J. (2019), "Tax Compliance and Enforcement", *Journal of Economic Literature*, 57(4): 904-54.

Torregrosa, S. (2015), "Bypassing progressive taxation Fraud and base evasion in the Spanish Income Tax (1970-2001)", *Documents de Treball de lÍEB* 2015/31, Barcelona. Institut d'Economía de Barcelona.

Weatherford, M. (2002), "Mining for fraud", *Intelligent Systems IEEE*, 17(4): 4-6.

Wilcoxon, F. (1945), "Individual Comparisons by Ranking Methods", *Biometrics Bulletin*, 1: 80-83.

Wu, R. S., Ou, C. S., Lin, H. Y., Chang, S. I. and Yen, D. C. (2012), "Using data mining technique to enhance tax evasion detection performance", *Expert Systems with Applications*, 39(10): 8769-8777.

Resumen

En este trabajo se propone un marco analítico que combina técnicas de minería de datos para obtener una segmentación de la muestra en función de la probabilidad de fraude potencial. En este sentido, el objetivo de este estudio es doble. En primer lugar, trata de determinar los beneficios fiscales con mayor probabilidad de ser utilizados por los potenciales contribuyentes defraudadores mediante la investigación de la estructura del IRPF. En segundo lugar, pretende caracterizar a través de variables socioeconómicas los perfiles de los segmentos de potenciales contribuyentes defraudadores para ofrecer una estrategia de selección de auditorías para mejorar el cumplimiento fiscal y mejorar el diseño de los impuestos. Se realiza una aplicación a la muestra anual del IRPF español diseñada por el Instituto de Estudios Fiscales. Los resultados obtenidos confirman que la combinación de técnicas de minería de datos propuesta ofrece información valiosa para contribuir al estudio del fraude fiscal.

Palabras clave: Impuesto sobre la Renta, cumplimiento fiscal, técnicas de minería de datos, perceptor multicapas, árboles de decisión, detección del fraude fiscal, evaluación fiscal.

Clasificación JEL: H24, C55, C38.