Expert Systems with Applications 40 (2013) 1427–1436

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Characterization and detection of taxpayers with false invoices using data mining techniques

Pamela Castellón González [a], Juan D. Velásquez [b,*]

[a] Servicio de Impuestos Internos, Government of Chile, Chile
[b] Department of Industrial Engineering, Universidad de Chile, Av. República 701, P.O. Box 8370439, Chile

ARTICLE INFO

Keywords:
False invoices
Fraud detection
Data mining
Clustering
Prediction

ABSTRACT

In this paper we give evidence that it is possible to characterize and detect those potential users of false invoices in a given year, depending on the information in their tax payment, their historical performance and characteristics, using different types of data mining techniques. First, clustering algorithms like SOM and neural gas are used to identify groups of similar behaviour in the universe of taxpayers. Then decision trees, neural networks and Bayesian networks are used to identify those variables that are related to conduct of fraud and/or no fraud, detect patterns of associated behaviour and establishing to what extent cases of fraud and/or no fraud can be detected with the available information. This will help identify patterns of fraud and generate knowledge that can be used in the audit work performed by the Tax Administration of Chile (in Spanish *Servicio de Impuestos Internos* (**SII**)) to detect this type of tax crime.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Tax evasion and tax fraud[1] have been a constant concern for tax administrations, especially when pertaining to developing countries (Davia, Coggins, Wideman, & Kastantin, 2000). While it is true that taxes are not the only source of government funding, the fact is that they send a very important signal about the commitment and effectiveness with which the State can carry out its functions and restrict access to other sources of income.

In particular, the value added tax (VAT), implemented in over 130 countries at different stages of economic development has become a key component of tax revenues, raising about 25% of the world's tax revenue (Harrison & Krelove, 2005). In the case of Chile, taxes provide about 75% of the resources from which the State each year pays its expenses and investments, collecting during 2011 a total of USD $41.6 billion dollars.[2] VAT represents 45% amounting to USD $18.7 billion dollars and generating over 400 million invoices a year, of which 56% is issued in paper format and 44% in electronic format (Bergman, 2010).

The phenomenon of false invoices in respect of VAT is explained by the mechanics of determining the tax payable. When a company receives a false invoice, it simulates a purchase that never existed, thus increasing its tax credit fraudulently and decreasing VAT payment. Also, there is a decrease of payment in the income tax due to increased costs and expenditures declared.

The falsity of the document may be *material* if the physical elements that make up the invoice have been adulterated, or *ideological* when the materiality of the document is not altered, but the operations recorded in it are adulterated or nonexistent. The latter is more complex and difficult to detect because it involves fictitious transactions in which an audit is required to examine the sales books and corrections, or cross referencing the information with suppliers. Moreover, these cases are more expensive for SII, as they require a greater amount of time dedicated to collecting and testing evidence, which is harder to find.

The best known cases of material falsification are the physical adulteration of the document, the use of *hanging* invoices in which an invoice is counterfeited to impersonate a taxpayer of good behavior, and the use of a double set of tax invoices, which has two same-numbered invoices, but one of which is fictional and for a higher amount. In ideological falsification, invoices are used to register a nonexistent operation or adulterate the contents of an existing operation.

According to a method used by the SII to estimate VAT evasion (Schneider & Enste, 2000) resulting from false invoices and other credit enlargements applied in the period 1996–2004, evasion by false invoices has historically represented between 15% and 25% of total VAT evasion, increasing significantly in years of economic

* Corresponding author. Tel.: +56 2978 4834; fax: +56 2689 7895.
E-mail address: jvelasqu@dii.uchile.cl (J.D. Velásquez).
URL: http://wi.dii.uchile.cl/ (J.D. Velásquez).

[1] Usually refers to *tax avoidance* when referring to behaviors that, within the law, prevent or reduce taxes, while *evasion or tax fraud* involves a violation of the law to obtain the same results.

[2] Considering only Central Government tax revenue (excluding CODELCO, the main government cooper miner company, municipalities and social security).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.08.051

crisis. This is why in the crisis of 1998–1999 the participation rate increased to 38%, reaching an amount close to USD $1 billion dollars. This becomes relevant since recently there was a global economic crisis that hit Chile in late 2008 and the middle of 2009, causing an increase in the rate of VAT evasion to 23%, in the amount of evasion of USD $4 billion dollars.

It also requires that resources be invested in well-focused monitoring, detecting those taxpayers who have greater compliance risk and not bother or waste time and resources on those who do comply (Slemrod & Yitzhaki, 2002). For this, data mining techniques offer great potential, because they allow the extraction and generation of knowledge from large volumes of data to detect and characterize fraudulent behavior and failure to pay tax, in the end improving the use of resources (Fayyad, Piatestky-Shapiro, & Smyth, 1996).

This paper is organized as follows. Section 2 describes how artificial intelligence techniques have facilitated the detection of tax evasion in tax administrations. Section 3 describes the data mining techniques applied. Section 4 describes the type of information used and the main results obtained in the characterization and detection of fraud in the issuance of invoices, and Section 5 presents the main conclusions and future lines of research.

2. Related work

Fraud in its various manifestations is a phenomenon that no modern society is free of. All governments, regardless of whether they are large or small, public or private, local or multinational, are affected by this reality, which seriously undermines the principles of solidarity and equality of citizens before the law and threatens business.

There are many fields and industries affected by this phenomenon. A study conducted by Chena, Huang, and Kuo (2009) in 2006, surveyed 150 medium and large Chilean companies to consult on this issue. The results show that 41% of them were victims of fraud in the past two years. This poses great challenges in prevention and opportunities for detection (Bonchi, Giannotti, Mainetto, & Pedreschi, 1999), given that fraud is usually higher than reported by companies, because somehow disturbs the image of the company towards customers and suppliers. In many cases there are even companies that are not known to have been victims of fraud.

Many fraud detection problems involve a large amount of information (Lundin, Kvarnstomr, & Jonsson, 2003). Processing these data in search of fraudulent transactions requires a statistical analysis which needs fast and efficient algorithms, among which data mining provides relevant techniques, facilitating data interpretation and helping to improve understanding of the processes behind the data (Myatt Glenn, 2007). These techniques have facilitated the detection of tax evasion and irregular behavior in other areas such as banking, telecommunications, insurance, IT, money laundering, and in the medical and scientific fields, among others (Cechhini, Aytug, Koehler, & Pathak, 2010).

To detect tax fraud, tax institutions began using random selection audits or focusing on those taxpayers who had no previous audits in recent periods and selecting cases according to the experience and knowledge of the auditors. Later methodologies were developed based on statistical analysis and construction of financial or tax ratios which evolved into the creation of rule-based systems and risk models (OECD, 1999). These transform tax information into indicators which permit ranking of taxpayers by compliance risk. In recent years, the techniques of data mining and artificial intelligence have been incorporated into the audit planning activities (US Government Accountability Office, 2004; OECD, 2004b), mainly to detect patterns of fraud or evasion, which are used by tax authorities for specific purposes.

The internal revenue service (**IRS**), the institution responsible for administering taxes in the United States, has used data mining techniques (US Government Accountability Office, 2004) for various purposes, among which are measuring the risk of taxpayer compliance, the detection of tax evasion and criminal financial activities (Dubin, 2007), electronic fraud detection, detection of housing tax abuse, detection of fraud by taxpayers who receive income from tax credits and money laundering (OECD, 2004a; OECD, 2004b; Watkinsa et al., 2003). The Fig. 1 shows part of the data mining techniques used by tax administrations as well as the logistic regression models, decision trees, neural networks, clustering algorithms and visualization techniques such as link analysis.

In the Australian Tax Office, the *Compliance Program* is based on a risk model which uses statistical techniques and data mining in order to make comparisons, to find associations and patterns by logistic regression, decision trees and SVM (US Government Accountability Office, 2004; US Government Accountability Office, 2008). A case of interest has been the approach used by Denny and Christen (2007), of discovering small clusters or unusual subpopulations, called *Hot Spots*, using techniques such as the self organizing map (SOM) to explore its features, clustering algorithms like k-means and visuals that are easy to understand for non-technical users.

In New Zealand, the existing model associates the degree of compliance with attention to auditing, which coincides with that used by the Australian counterpart (OECD, 2004a). The plan includes an analysis of the economic, international, population, ethnic diversity and family structure. For its part, Canada uses neural networks and decision trees to distinguish the characteristics of taxpayers who evade or commit fraud, based on the results of past audits, to detect patterns of noncompliance or evasion (OECD, 2004b).

In Latin America, Peru was one of the first to apply these techniques to detecting tax evasion (García & Valderrama, 2007; Torgler, 2005), adding to the selection system of the Maritime Customs of Callao an artificial intelligence tool based on neural networks. During 2004, this model was improved through the application of fuzzy rules and association for pre-processing variables and classification and regression trees (CART) to select the most relevant variables.

Brazil has developed project risk analysis and applied artificial intelligence (HARPIA) jointly with the Brazilian Federal Revenue and universities in the country (Digiampietri et al., 2008). This project consists of developing a detection system of atypical points to help the regulators to identify suspicious transactions based on a graphic display of information on historical imports and exports and a system of export product information based on Markov chains, to help importers in the registration and classification of their products, avoid duplication and to calculate the probability that a string is valid in a given domain.

In the case of Chile the first trial was developed in 2007 (Lückeheide, Velásquez, & Cerda, 2007), using the SOM and k-means to segment VAT taxpayers according to their F29 statements and characteristics. Later, in 2009, following the international trend, risk models were built of different stages of the life cycle of the taxpayer, in which neural networks, decision trees and logistic regression techniques are applied. The first trial was further developed to identify potential users of false invoices through artificial neural networks and decision trees, mainly using information from tax and income declarations in micro and small enterprises.

3. Data mining techniques applied

For purposes of characterization and identification of patterns three data mining techniques are applied: self organizing maps

Technique Applied	USA	Canada	Australia	UK	Bulgaria	Brazil	Peru	Chile
Neural Networks	✓	✓		✓	✓		✓	✓
Decision Tree	✓	✓	✓				✓	✓
Logistic Regression	✓		✓	✓	✓			
SOM			✓					✓
K-means			✓					✓
Support Vector Machines	✓		✓					
Visualization Techniques	✓					✓		
Bayesian Networks			✓					
K-Nearest Neighbour			✓					
Association Rules							✓	
Fuzzy Rules							✓	
Markov Chains						✓		
Time Series		✓						
Regression				✓				
Simulations	✓							

Fig. 1. Data mining techniques used by tax administrations to detect tax fraud.

(SOM), neural gas (NG) and decision trees. Backpropagation neural networks and Bayesian networks are subsequently used for detection, and are described below.

3.1. Self-organizing maps

The self-organizing map (SOM) (Vesanto, 2000) is one of the models most widely used in artificial neural networks for analysis and visualization of high dimensional data, based on unsupervised competitive learning. Specifically, the network consists of a set of neurons arranged in a grid dimension 'a', usually rectangular, cylindrical or toroidal, which generates an output space of dimension 'd', with 'a' less or equal than 'd', on which neighborhood relations are defined, and whose aim is to discover the underlying structure of the data entered into it. By construction all the same neurons receive input at any given time.

During training, neurons in the network generate some activity from the stimulation of the input data, allowing a more specific identification of which areas or which neurons have learned to represent certain input patterns. Activity patterns generated in the same area have similar characteristics and can be grouped into a single category or cluster, based on a distance measure, usually Euclidean. The winning neuron output layer or *best matching unit* (BMU) is one whose weight vector is most similar to input information.

This tool is usually applied to clustering and segmentation, creating groups of objects with behavior similar to each other but different from the objects of another group.

3.2. Neural gas

Neural gas (NG) (Filippone, Camastra, Masulli, & Rovetta, 2008) is a relatively new algorithm for unsupervised neural networks, focused on vector quantization of arbitrary structures. The major difference with the SOM is that this method does not define a grid that imposes topological relationships between units of the network and each neuron can move freely through the data space. This freedom allows the algorithm a better ability to approximate the distribution of data in the input space, as the neurons are not required to have to maintain specific neighborly relations. However, having some background on the number of groups is expected to be required.

During the network training, the neurons change their position and adapt themselves to the data cloud. In this algorithm, each input pattern generates an excitation in each unit of the network. In each iteration a random data vector is presented to all neurons. For each data vector the nearest neuron is found, according to the Euclidean distance. This neuron is called *winning*. In the next step the neighborhood (diameter) of the winning neuron is established, which decreases exponentially with the number of iterations.

3.3. Classification trees

Classification trees (Murthy, 1998) are one of the non-parametric supervised learning methods most commonly used, being notable for their simplicity and applicability to different areas and interests. In general, the tree construction algorithms differ in the strategies used to partition nodes and prune the tree. In our case, we use trees based on CHAID methodology, which generate a different number of branches from a node considering both continuous and categorical variables. Basically the algorithm consists in forming all possible pairs and combinations of categories, grouping the categories that behave homogeneously with respect to the response variable in a group and maintaining separate those categories that behave differently.

For each possible pair, we calculate the statistics for their cross with the dependent variable (chi-square statistic for categorical target fields or F statistic for continuous outputs). The pair with the lowest value of this indicator will be a new category of two merged values, provided it is not statistically significant. For merged categories further consolidation of the values of the predictor is done, but this time with one category less, the process ending when no more mergers can be effected because statistically significant results occur.

3.4. Multilayer perceptron neural network

The multilayer perceptron model (MLP) (Parlos, 1994) is an artificial neural network model of layers used for classification and grouping, based on human brain function through an interconnected set of vertices. The network must find the relationship between input attributes and the desired output for each case. This is done through a learning method called *backpropagation* which minimizes the prediction error by adjusting the weights of the network. This method has two stages. The first departures are calculated based on the inputs and the weights assigned to the initial network, for which the prediction error is calculated. In the second phase, the error is calculated backward through the network from the output units to the input units, getting an error in each unit. In this way the weights are updated through a gradient descent method. This process is iterative, so that after repeating the algorithm several times, the network will converge to a state that allows the classification of all training patterns, which minimizes the error.[3]

[3] Usually calculates the mean square error.

3.5. Bayesian networks

Bayesian networks (Friedman, Geiger, & Goldszmidt, 1997; Heckerman, Geiger, & Chickering, 1995) are directed acyclic graphs, used to predict the likelihood of different outcomes, based on a set of facts. The network consists of a set of nodes representing the variables of the problem and a set of directed arcs connecting the nodes and indicating a relationship of dependency between the attributes of the observed data. Bayesian networks describe the probability distribution that governs a set of variables, specifying assumptions of conditional independence with conditional probabilities. Typically, this problem is divided into two parts: structural learning, which is to obtain the network structure, and parametric learning, in which through known graph structure, we obtain the probabilities for each node. Their main advantage is that the probability of occurrence of a given event based on a set of actions can be obtained, giving a clear view of the relationship through a web graph.

4. Data, analysis and results

Unlike the previous study developed by SII related to this problem, this paper aims to complement the use of tax information with additional variables related to its historical performance and its performance in the year of analysis, and include aspects concerning direct associates, such as agents, partners and legal representatives and their characteristics, such as level of coverage, age, and whether electronic invoices or full accounting are used, among others. Moreover, a model for medium and large companies is developed, where there is less knowledge of how to operate regarding the use of false invoices, since they have more complex evasion procedures. This will build models differentiated by the size of the taxpayer, grouping on one hand the micro and small enterprises and on the other medium and large enterprises.

4.1. Data and attribute selection

The year 2006 is chosen for characterization and detection, because the audits are performed up to a period of three years previous, which makes it difficult to use the latest information, as in 2010 cases were still being generating that could have used false invoices from 2007 onwards. Thus, for the characterization of contributors the universe of taxpayers is considered to be all those taxpayers who had filed at least one VAT return between 2005 and 2007, which corresponds to 582,161 enterprises. In the case of detection, information is used from those audits where there is certainty that the invoices were checked in 2006, independent of when that was done, considering a total of 1692 companies. Table 1 shows a taxonomy with the taxpayers consider in our analysis.

One of the biggest drawbacks to defining cases with and without fraud occurring is related to the way in which information is recorded, for example the date of the start and completion of the audit, the revised tax periods and the result are known, but the information on the periods in which differences occur is not automated. Therefore, to see if the false invoice detected corresponded specifically to 2006, the notes and comments made by the auditor would have to be reviewed and corrections done in the codes related to the invoices of that year. Cases with and without fraud were categorized into three types: "0" indicates that the taxpayer was audited and no false invoices found in any of the periods reviewed, "1" indicating that the taxpayer did not use false invoices in the year of analysis but did in other periods reviewed (usually the previous year or the next) and "2" indicating that the taxpayer used false invoices in the year of study.

To construct the feature vector, 20 codes were selected from the Monthly VAT Tax Payment Form (F29) related to the operative payment of VAT, 31 codes from the Annual Income Tax Form (F22) associated with the generation of taxable income class and business financial data and 31 tax ratios relating the VAT and Income Tax information with profitability and the company's liquidity, among others. Regarding the behavior and features of the company, this generates 92 indicators that can signal good or bad behavior over time related to its historical performance, its particular characteristics and information generated at different stages of the life cycle as show in Table 2.

In the pre-processing of data, using a rule to carry out data cleansing, those cases that exceed the mean plus five times the standard deviation are considered as outliers, leaving only those cases with a positive value of each code. In most cases, the distribution of variables with which they worked was declining, where a large percentage of taxpayers pay low amounts of taxes, and only a small group pays high amounts. The same applies to the behavioral variables, because they constitute the misconduct of only a small group of taxpayers. Therefore, eliminating cases with higher values removes those taxpayers who generally have worse behavior, which are the focus group of the study.

Since the declaration of payment of VAT is done monthly and the declaration of income tax is done on an annual basis, the first transformation is to consider the annual total sum of the monthly amounts for each F29 code during the year to make them comparable with income tax information. Regarding the completeness of null data, the VAT information is more complete than the income tax information, because these codes should only be filed by taxpayers who are full accounting. Therefore, VAT debit and credit information is used to complete the revenue and cost data for the period, due to the direct relationship between them. For the rest of the income fields, the median is used for taxpayers in the same sales code section. Finally, due to the decreasing distribution of the tax variables, logarithmic transformation is applied to reduce the impact of extreme data. To avoid variables with a greater range of values downplaying others with a smaller range, it is necessary to normalize the variables in ways that are comparable with each other, using the min–max standard deviation in the range [0.1].

Additionally, prior to selecting the variables used in behavioral models it is necessary to reduce them through principal component analysis (PCA). As a Result 15 principal components in the micro and small enterprises are generated, explaining 61.3% of the variance in the data. Similarly, 16 principal components for medium and large businesses are generated that explain 59.9% of the variance in the data.

Since our interest was to generate behavioral variables related to the use and sale of false invoices and not other behaviors, those variables were selected that have a medium–high correlation with the variable use of false invoices. Those variables were discarded

Table 1
Number of taxpayers used in the analysis.

Taxpayers	Micro-small	Medium-large	Total enterprises
Enterprises active in period 2005–2007	558.319 (96%)	23.842 (4%)	582.161 (100%)
Companies audited by invoices in 2006 resulting in fraud or no fraud	1.280 (76%)	412 (24%)	1.692 (100%)

Table 2
Type of information used to construct the feature vector.

Concept	Type of information
Payment of taxes	VAT tax declarations (F29), declarations of income tax (F22), tax rates and income taxes
Personal characteristics	Age of taxpayer, age of company, level of coverage, electronic biller, computer accounting, economic activities, declares online, change of subject, whether domiciled and owns branches
Historical	Selective audits, previous offenses, address problems
Behavior and within year of study	Failures to attend, accusations and closures, losses of RUT, destruction of documents, debt regularization, loss of invoices, invoices investigated and/or closures, warnings
Life cycle	Start-ups, verification of activities, stamping of documents, changes information, expiration of prior suspension of activities
Relationships	Agents, legal representatives, partners, family, suppliers, accountants, associations and representations (assets, history of offenses, investigations, closures)

that have more than a 10% chance that the Pearson correlation coefficient is zero, except for some codes of interest such as the total debits, total tax credits and VAT payments.

Similarly, we discarded those variables that have a large percentage of null values. In this way 42 variables are selected in the micro and small segment and 36 variables in the medium and large segment for analysis. In the first group, 35% of the code variables correspond to the VAT, 35% of code variables are related to income tax and 30% to variables related to behavior. In the second group on the other hand these percentages vary by 31%, 38% and 31%, respectively, with a higher prevalence of variables related to income tax.

After removing the outliers and inconsistent cases, the final data set is composed of 532,755 taxpayers who are micro and small enterprises and 22,609 medium and large enterprises, eliminating 4.6% of the first group and 3.4% of the second.

4.2. Modeling

In order to effect the characterization and identification of patterns, in a first stage, data mining techniques are applied to the universe of companies, in order to identify relationships between their payment of taxes and behavioral variables associated with the use of false invoices. Then classification techniques are applied in those cases where the condition of fraud and no fraud is known, in order to identify specific patterns of this group of taxpayers. Finally, classification tools are applied to detect cases with and without fraud with the information generated.

4.2.1. Characterizing the universe of companies

Initially, the SOM method is applied to the universe of taxpayers, to identify clusters or groups of taxpayers who have similar behavior. The working hypothesis assumed that when considering only the behavioral variables related to the use of false invoices combined with tax variables, it was possible to detect groups of taxpayers who had good or bad fiscal behavior, and know how they made their tax payment.

For the generation of experiments the *R-SOM* package is used, based on a rectangular network topology, with three input neurons and 24×24 output neurons in micro and small enterprises, and 36×36 output neurons in medium and large enterprises, with a maximum of 100 iterations. In the first group a random sample of 100,000 businesses is considered due to computational constraints. As a Result 5 clusters are generated in the segment of micro and small enterprises and 6 clusters in the medium and large enterprises, as shown in Fig. 2.

The clusters obtained in the first group are mainly differentiated by the use of sales slips and/or invoices, the VAT payment level, the level of reported costs, the level of formality of the company, participation in other companies and some tracking issues. The medium and large group is differentiated by the use of sales slips and/ or invoices, level of use of tax credit balances, credit notes and invoices of fixed assets, liabilities and assets as well as the results of previous audits and the level of formality of the company.While some patterns of behavior were found, these were not specifically related to the use of false invoices. Moreover, behavioral variables associated with historical characteristics and irregularities do not vary much from one group to another.

Neural gas is then applied, considering the same number of clusters as the Kohonen map, using the '*R-cclust* package, which generates an array with the characteristics of the centroids of each variable and a vector classification, marking the group each taxpayer belongs to. In this case, the groups generated are also influenced by the payment of taxes, but with major differences in terms of behavior. This allows the differentiation of which groups have better and worse behavior, and relates it to their tax payment, although the cases of false invoices were not necessarily found in the same group.

While these techniques can characterize the universe of taxpayers and identify some distinguishing patterns, considering those variables most correlated with the use of false invoices tends to give more prominence to tax than behavior variables, creating groups that differ in the type of transaction (sales with invoices and/or sales slips), the level of activity (high-low level of sales, costs) and tax payment (high-low), due to greater variability in these variables compared to those of behavior. Accordingly, the following patterns were identified associated with bad and good performance, considering the common points obtained by both methods, as shown in Fig. 3.

4.2.2. Characterizing cases with fraud or without fraud

When analyzing the distribution of each variable, it is noted that fraud cases are usually found among the extreme cases of each. For this reason, it is determined to apply decision trees to all audit data with known results, since it permits the identification of the cutoff point of each variable against which there is a change of behavior, the consideration of extreme cases and the generation of rules that can be validated and implemented.

The type of tree used is the Chi-square automatic interaction detection (CHAID), which allows non-binary classifications and the generation of branches from a node considering both continuous and categorical variables. This method requires access to significant sample sizes, since when divided into multiple groups there is a risk of finding empty or unrepresentative groups if there are not sufficient cases in each combination of categories. In addition, the exhaustive method is evaluated, which seeks to address some weaknesses of the traditional CHAID.

As shown en Fig. 4, the factors that have the greatest impact in the micro and small enterprises were the result of previous audits and the percentage of purchases supported by invoices. This indicates that those who have been audited more times in the past and nothing was found, and whose purchases are not based primarily on invoices, are less likely to use false invoices than those whose purchases were mainly recorded by sales slips and had

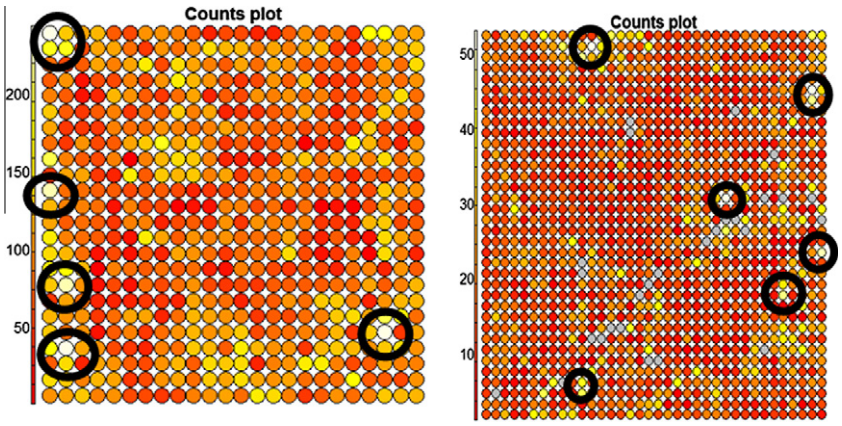

Fig. 2. Map resulting from SOM application in MI-SM (left) and ME-LA (rigth).

MICRO AND SMALL				
Variable	Period	Concept	Good	Bad
Sales Slip Debits	t	VAT	↑	
Payment of VAT			↑	↓
Credits			↓	↑
Tax credits balances			↓	↑
Ratio debts/credits	t	Ratio Income VAT		↓
Ratio income/assets				↓
Stamped invoices	t-2	Stamping	↑	↓
Stamping frecuency			↑	
Activity checks	< t	Lyfe cycle	↓	↑
Crimes and irregularities	< t	Historical Behavior	↓	
Crimes indirectly related			↓	
Positive previous audit				↑
MEDIUM AND LARGE				
Variable	Period	Concept	Good	Bad
Costs and expenditures	t	Income	↑	
Assets			↑	
Liabilities			↑	↓
Credits	t	VAT	↑	
Tax credit balances			↑	
Number of sales slips				↓
Ratio costs/assets	t	Ratio Income VAT		↑
Ratio earnings/assets				↑
Ratio Invoice debits/total debits				↑
Formalization of accounting	t	Characteristics	↑	↓
Coverage			↑	↓
Legal representatives			↑	
Previous closures	< t	Historical Behavior		↑
Stamping restricted				↑
Accusations and closures				↑
Failures to attend				↑
Audits			↑	↓

Fig. 3. Variables associated with good and bad behavior in the universe of taxpayers.

productive audits in the past. In fact, these two variables alone identify a number of end nodes with a preponderance of cases without fraud.

Additionally, the variable indicating a greater preponderance of crimes and irregularities associated with historical invoices combined with the frequency of stamping generates end nodes with

a preponderance of cases with false invoices. In particular node 12, which contains nearly half the cases (46%), is decomposed into several branches according to the value taken by the average credit by invoice issued (the higher this indicator, the greater potential there is to commit fraud). Similarly, the preponderance of cases of fraud in each branch depends on the number of invoices issued, VAT paid, the total debits per invoice/sales slips, the relationship between costs and assets and the level of participation in other companies.

This technique was effective in identifying patterns associated with fraud and without fraud, since the end nodes consisted mainly of cases of a single type, or were otherwise combined with cases with output value "1", which more closely approximate the behavior of fraud cases "2". Considering the patterns and rules that are repeated in each branch of the tree to differentiate between cases with fraud and without fraud, Fig. 5 shows the behaviors associated with each of them in each segment, which summarizes the main variables considered and the relationships that generate nodes with and without the use of false invoices.

The most important variables to distinguish cases of fraud in the micro and small enterprises were the result of previous audits, the total VAT determined, the percentage of credit supported by invoices, the relationship between tax credit balances and credits, total debits by invoice/sales slips and the relationship between stamped and issued invoices. The medium and large variables correspond to the total of tax credit balances, percentage of credit supported by invoices, the number of legal representatives, level of formalization of accounting and the relationship between costs and assets, among others.

4.3. Fraud detection

For detection, artificial neural networks, decision trees and Bayesian networks were applied. To avoid over-adjustment of the network, the data was divided into two sets, a training set and a testing set, using the 70/30 rule. Moreover, these three methods were implemented using the *SPSS Clementine* technological tool.

One of the complexities of neural networks is to determine the number of layers and hidden nodes and the number of epochs or iterations. To determine these parameters different numbers of cycles and nodes in the hidden layers were considered, in order to establish the appropriate values through trial and error. For the iterations the values used are 1000, 5000, 10,000 and 20,000. In the case of the nodes, using the number the software calculates by default based on the model and other data corresponding to half the number of input nodes, gives 3 and 20 nodes, respectively.

In the case of Bayesian networks two methods are evaluated for constructing the network, the TAN algorithm and the Markov Blanket estimation algorithm available in the SPSS Clementine software. Additionally, a previous pre-processing of the variables is used to identify which are the most relevant variables and improve the processing time and performance of the algorithm. Likewise, an independent test of maximum likelihood and a chi-square test for parametric learning are used.

The experimental results are presented in Table 3, which contains the following indicators obtained in group testing: (1) Sensitivity-indicates the proportion of cases with fraud classified correctly, (2) Specificity-indicates the proportion of cases without fraud where the classification was correct, (3) Consistency-indicates

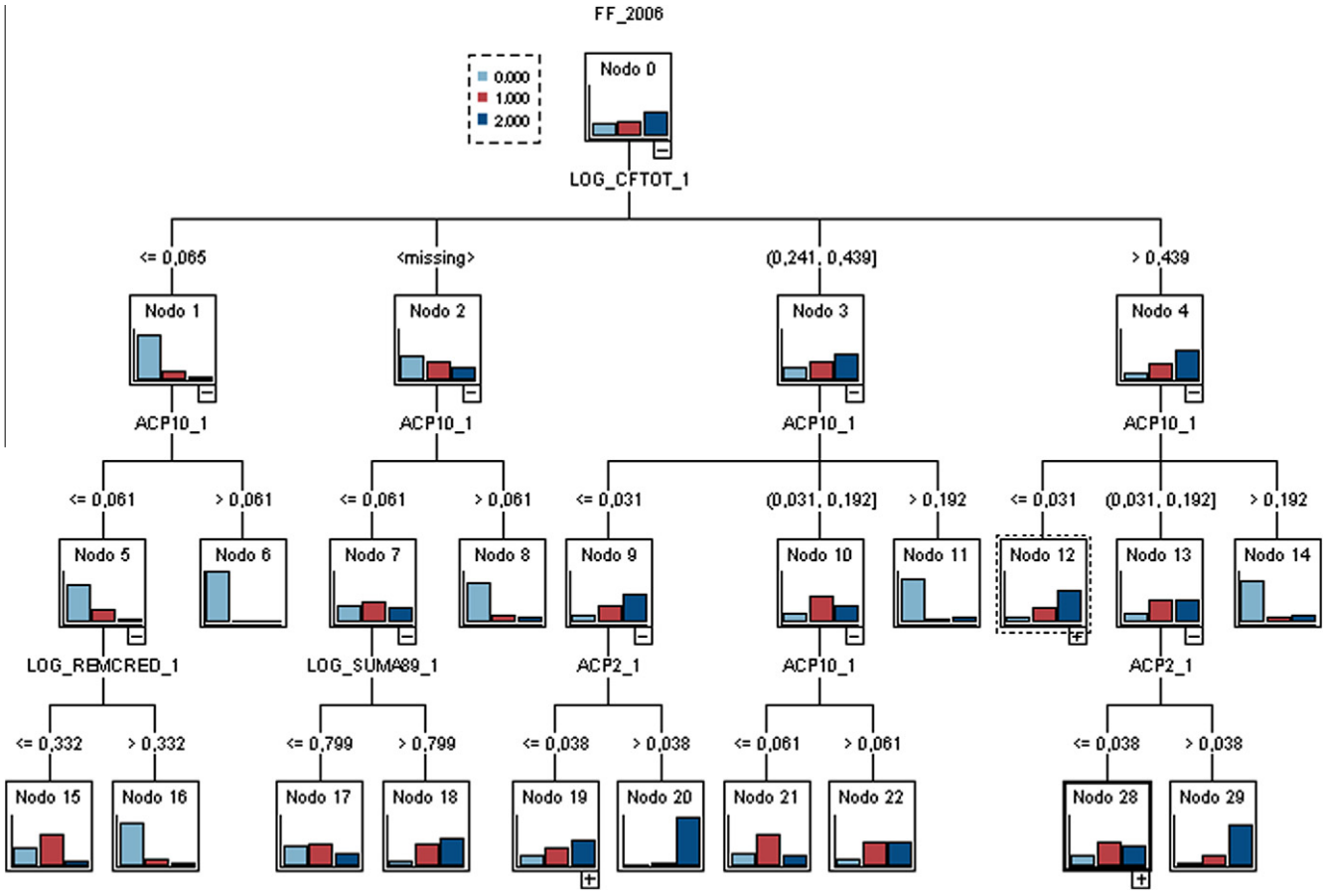

Fig. 4. Main branches of decision tree in micro and small companies.

MICRO AND SMALL				
Variable	Period	Concept	No Fraud	Fraud
Invoice Debits	t	VAT		↑
Issued invoices			↓	
VAT				↑
Ratio credits invoices/total credits	t	Ratio Income VAT	↓	↑
Ratio tax credit balances/ credit mean			↑	
Ratio costs/assets				↑
Stamping frecuency	t-2	Stamping		↑
Ratio issued invoices/stamping invoices			↓	↑
Crimes and irregularities	< t	Historical Behavior		↑
Negative previous audits			↑	↓
Positive previous audits				↑
MEDIUM AND LARGE				
Variable	Period	Concept	No Fraud	Fraud
Tax credit balances	t	VAT	↑	↓
Ratio credit invoices/total credits	t	Ratio Income VAT	↓	↑
Ratio costs/assets			↓	↑
Age of company	t	Characteristic		↓
Formalization of accounting				↓
Economic activities				↑
Amount of orders to pay	< t	Historical Behaviour		↑
Failures to answer notifications				↑
Irregularities with invoices			↓	↑

Fig. 5. Variables associated with fraudulent and non-fraudulent behavior by false invoices.

Table 3
Experiments in detection fraud by false invoices.

Exp.	Segmento	Method	Sensitivity (%)	Specifity (%)	Consistency (%)	Error rate (%)
1	MI-SM	NN	92.6	72.9	87.2	12.8
2	MI-SM	BN	82.3	64.1	77.9	22.1
3	MI-SM	DT	89.0	79.0	87.0	13.0
4	ME-LA	NN	88.8	59.1	72.5	27.5
5	ME-LA	BN	73.3	66.7	70.3	29.7
6	ME-LA	DT	79.0	85.0	82.0	18.0

the proportion of cases with and without fraud in which the classification was correct and (4) Error Rate-indicates the proportion of cases with and without fraud which were assigned to an incorrect class.

In both segments, the best detection results of cases with false invoices were obtained with the neural network method. In the group of micro and small enterprises, experiment 1 showed that 92.6% of fraud cases were assigned to the correct class, while in the group of medium and large enterprises the proportion of fraud cases correctly allocated was 88.8%. Moreover, the power of generalization of the model was quite good, as test results were similar to those obtained in the network training, where detection of cases without fraud was 93.7% and 87.4%, respectively.

The neural network generated for the micro and small enterprises indicates a preponderance of variables associated with the payment of VAT and behavior, and to a lesser extent to income-related variables. The most relevant correspond to background information obtained from the activity checks, the relationship between tax credit balances and average credits, the total debits by invoices issued, the relationship between money income and

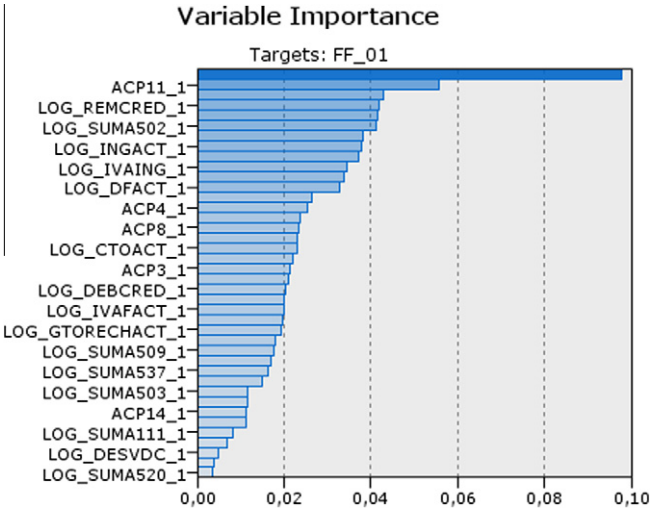

Fig. 6. Level of importance of the variables in micro and small companies according to the neural network.

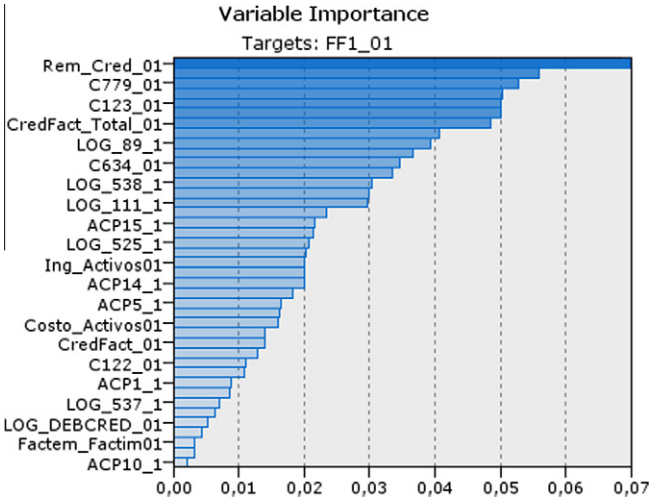

Fig. 7. Level of importance of the variables in medium and large companies according to the neural network.

assets and the relationship between VAT paid and the income declared as shown en Fig. 6. In the case of medium and large companies, the most important variables correspond to the relationship between tax credit balances and average credits, accounts payable to related companies, total liabilities, the proportion of tax credits associated with invoices and the VAT determined in the period as shown en Fig. 7.

5. Conclusions and future work

The clustering and classification methods used to characterize the taxpayers who have good or bad fiscal behavior associated with the use of false invoices show that it is possible to identify some distinguishing characteristics between one group and another, which accord with what happens in reality. Particularly the neural gas method found that it was possible to identify some relevant variables to differentiate between good or bad behavior, not necessarily associated with the use and sale of false invoices. Kohonen's method however, did not provide any behavioral patterns associated with the use of false invoices, but rather clusters were detected in relation to taxation, in which the variables with the largest number of zeros and variance proved to have more impact in shaping the groups.

The decision tree method applied to cases in which the result of fraud and no fraud was known was a good technique to detect variables that could distinguish between fraud and no fraud. This is because when analyzing the distribution of variables in each group, it is noted that fraud cases tend to take more extreme values of the variables, so it was possible to distinguish ranges in which there is a chance of having or not having fraud. On the other hand, the results were consistent with those observed in reality, according to the expert view.

Thus, in the case of micro and small enterprises the variables that allowed distinguishing between fraud and no fraud were mainly related to the percentage of tax credits generated by invoices with respect to total credit and previous audits with negative results. To the extent that the taxpayer was audited several times in the past and nothing was found, they are more likely to have no fraud in the future. On the other hand, where their credit is more associated with other items than invoices (fixed or other assets) they are less likely to use invoices to support their claims. Other important variables were the number of invoices issued during the year and its relation to the invoices stamped in the past two years, the total amount of VAT declared during the year, the ratio of average tax credit balances and positive prior audits and historical crimes and irregularities associated with invoices.

In the medium and large companies, the most important variables were the amount of surplus credit accumulated in prior periods, the percentage of credit associated with invoices, the relationship between costs and assets, the level of informality in their accounting and the age of the company, as well as the number of irregularities associated with previous invoices and the amount of orders to pay and historical failures to answer notifications.

In relation to the detection models, those which performed better were the multilayer perceptron neural network models, which for purposes of the study had an input layer containing the explanatory variables, an intermediate layer of processing and an output layer. In the case of micro and small businesses the percentage of correctly detected fraud cases was 92%, while in the case of medium and large enterprises, this percentage was 89%.

Given this result, and considering that in practice only a rather small group of companies in a year can be monitored, we recommend a combination of the results obtained with neural networks, decision tree and bayesian networks, in order to select for audit those that appear labeled as fraud in the neural network and have the highest odds of committing fraud under the Bayesian network and decision tree.

According to studies made by the SII, about 20% of taxpayers use false invoices to evade taxes. No information disaggregated by type of taxpayer exists but considering the percentage of classification of cases with and without fraud by neural network models, it is estimated that the universe of potential users of false invoices is 116,000 micro and small enterprises and 4768 medium and large enterprises, generating a potential collection of USD $210 million dollars.

Finally, to test the actual detection model developed, and being consistent with the previous point, its implementation in activities in the field is vital to determine the level of accuracy in the classification of taxpayers selected in the sample. The implementation of a pilot program that will target the two economic sectors studied is recommended, which shall be conclusive in terms of the real effectiveness of the model.

For future work, we recommend generating new historical behavioral variables related to specific audits and level of coverage of these, considering other methods for preprocessing and selection of variables as well as cross-validation techniques to explore and implement other data mining techniques to improve the detection of cases with and without fraud.

Acknowledgement

We are very grateful to the Chilean Millennium Institute of Complex Engineering Systems (ICM: P-05-004-F, CONICYT: FBO16), which partially supported this paper.

References

Bergman, M. (2010). *Tax evasion and the rule of law in Latin America: The political culture of cheating and compliance in Argentina and Chile*. Penn State University Press.

Bonchi, F., Giannotti, F., Mainetto, G., & Pedreschi, D. (1999). A classification-based methodology for planning audit strategies in fraud detection. In *Proceedings of the 5th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '99* (pp. 175–184). ACM.

Cechhini, M., Aytug, H., Koehler, G., & Pathak, P. (2010). Detecting management fraud in public companies. *Management Science, 56*, 1146–1160.

Chena, H., Huang, S., & Kuo, C. (2009). Using the artificial neural network to predict fraud litigation: Some empirical evidence from emerging markets. *Expert Systems with Applications, 36*, 1478–1484.

Davia, H. R., Coggins, P., Wideman, J., & Kastantin, J. (2000). *Accountant's guide to fraud detection and control* (2nd ed.). Wiley.

Denny, W., & Christen, P. (2007). Exploratory multilevel hot spot analysis: Australian taxation office case study. In *Conferences in research and practice in information technology* (Vol. 70, pp. 73–80). CRPIT Press.

Digiampietri, L. A., Roman, N. T., Meira, L. A. A., Filho, J. J., Ferreira, C. D., Kondo, A. A., et al. (2008). Uses of artificial intelligence in the Brazilian customs fraud detection system. In *Proceedings of the 2008 international conference on digital government research* (pp. 181–187). Digital Government Society of North America.

Dubin, J. (2007). Criminal investigation enforcement activities and taxpayer noncompliance. *Public Finance Review, 35*, 500–529.

Fayyad, U., Piatestky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery indatabases. *American Association for Artificial Intelligence*, 37–54.

Filippone, M., Camastra, F., Masulli, F., & Rovetta, S. (2008). A survey of kernel and spectral methods for clustering. *Pattern Recognition, 41*, 176–190.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29*, 131–163.

García, V., & Valderrama, J. (2007). Toward a more efficient tax policy. In M. Giugale, V. Fretes-Cibils, & N. J. L. (Eds.), *An opportunity for a different PERU prosperous, equitable, and governable* (pp. 103–134). Washington, DC, USA: The World Bank.

Harrison, G., & Krelove, R. (2005). *VAT refunds: A review of country experience*. International Monetary Fund (IMF).

Heckerman, D., Geiger, D., & Chickering, D. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning, 20*, 197–243.

Lückeheide, S., Velásquez, J. D., & Cerda, L. (2007). Segmentación de los contribuyentes que declaran iva aplicando herramientas de clustering. *Revista de Ingeniería de Sistemas, 21*, 87–110.

Lundin, E., Kvarnstomr, H., & Jonsson, E. (2003). Synthesizing test data for fraud detection systems. In *Proceedings of the 19th Annual Computer Security Applications Conference* (pp. 384–394). CSAC Press.

Murthy, S. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery, 2*, 345–389.

Myatt Glenn, J. (2007). *Making sense of data, a practical guide to exploratory data analysis and data mining*. Wiley Interscience.

OECD (1999). *Compliance measurement, practice note. Centre for Tax Policy and Administration, Tax Guidance Series. General Administrative Principles – GAP004 Compliance Measurement*. OECD Press.

OECD (2004a). *Compliance risk management, managing and improving tax compliance. forum on tax administration compliance subgroup. Centre for Tax Policy and Administration*. OECD Press.

OECD (2004b). *Compliance risk management, audit case selection systems. forum on tax administration compliance subgroup*. Centre for Tax Policy and Administration. OECD Press.

Parlos, A. (1994). Application of the recurrent multilayer perceptron in modeling complex process dynamics. *IEEE Transactions on Neural Networks, 5*, 255–266.

Schneider, F., & Enste, D. (2000). Shadow economies: Sixe, causes and consequences. *Journal of Economic Literature, XXXVIII*, 77–114.

Slemrod, J., & Yitzhaki, S. (2002). Tax avoidance, evasion, and administration. *Handbook of Public Economics, 3*, 1423–1470.

Torgler, B. (2005). Tax morale in Latin America. *Public Choice, 122*, 133–157.

US Government Accountability Office (2004). *Data mining: Agencies have taken key steps to protect privacy in selected efforts, but significant compliance issues remain*. GAO Press.

US Government Accountability Office (2008). *Lessons learned from other countries on compliance risks, administrative costs, compliance burden and transition*. Report to Congressional Requesters. GAO Press.

Vesanto, J. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks, 11*, 586–600.

Watkinsa, R. C., Reynoldsa, K. M., Demaraa, R., Georgiopoulosa, M., Gonzaleza, A., & Eaglina, R. (2003). Tracking dirty proceeds: Exploring data mining technologies as tools to investigate money laundering. *Police Practice and Research: An International Journal, 4*, 163–178.