TUCS

Dorina Marghescu | Minna Kallio | Barbro Back

Using financial ratios to select companies for tax auditing: An exploratory analysis

Turku Centre *for* Computer Science

TUCS Technical Report
No 996, December 2010

Using financial ratios to select companies for tax auditing: An exploratory analysis

Dorina Marghescu

Åbo Akademi University, Department of Information Technologies,
Turku Centre for Computer Science

Minna Kallio

Åbo Akademi University, Department of Information Technologies,
Turku Centre for Computer Science

Barbro Back

Åbo Akademi University, Department of Information Technologies,
Turku Centre for Computer Science

TUCS Technical Report
No 996, December 2010

Abstract

Tax auditing procedures include an investigation of the accounting records of a company and of other sources of information in order to assess whether the taxation has been based on correct and complete information. When there are found discrepancies between the accounting information and the real situation, the taxation should be corrected so that the eventual tax defaults are assessed and debited. The paper analyzes to what extent the financial performance of a company can be used as an indicator of tax defaults. We focus on one type of tax, namely employer's contribution or payroll tax, and four financial ratios. We evaluate different models built on a set of Finnish companies by using a binomial logistic regression analysis. The study is exploratory, meaning that it aims at understanding the characteristics of the companies with tax defaults, rather than confirming a hypothesis or predicting the likelihood that the company is in the high-risk group. In addition, the analysis is at a preliminary stage in the sense that it does not include all relevant characteristics of the companies, but only a limited amount, namely four ratios representing financial performance. However, the study is useful because it provides evidence that a certain group of companies that have employer's contribution defaults presents particular characteristics and that more variables are needed to fully capture the particularities of the companies of interest. Methodologically, the study shows that the logistic regression is useful for modeling the differences between the companies with tax defaults and the companies without tax defaults, and that the pre-processing of the data in terms of filtering out the companies into meaningful groups is as important in modeling as the selection of appropriate variables.

Keywords: tax auditing, financial performance, financial ratios, binomial logistic regression

TUCS Laboratory
Data Mining and Knowledge Management laboratory

1 Introduction

In Finland, tax audits are one form of tax control undertaken by the Tax Administration to ensure that the taxes are imposed in the correct amount and at the right time. Tax auditing procedures include an investigation of the accounting records of a company and of other sources of information in order to assess whether the taxation has been based on correct and complete information [1]. Taxation is based on the information provided by taxpayers and other sources; therefore, when, through audit, there are found discrepancies between the provided information and the real situation, the taxation should be corrected so that the eventual tax defaults are assessed and debited. The tax defaults here refer to those tax liabilities that are not timely paid due to misreporting; they are different from the tax corrections that are done routinely by the Tax Administration.

Tax audits are expensive and thus, tax authorities must select the taxpayers for auditing carefully. It is important that the tax audits target those companies that have indeed significant tax defaults. Hence, finding effective and efficient methods and models for selecting the companies for tax auditing is an important task, interesting for both public authorities and academia. The literature on this topic is not very generous. One reason is that the task is very difficult to be accomplished by a single method, but rather by a multitude of methods and during a highly interactive process involving both domain experts and database systems [2]. In addition, tax crimes are detected by using complex procedures that are usually conducted by cooperating authorities. These procedures usually employ all kinds of information, not only financial information or data that are stored in computers' databases. Another reason is the confidential nature of the subject and of the data under analysis [3,4].

Tax audits' scope includes all categories of taxes or only some of them. In this paper, we focus on the employer's contribution or payroll tax[1]. We investigate whether there are any relationships between the presence of employer's contribution defaults and the financial performance of companies. The study aims at determining the extent to which the financial performance of a company, measured by four particular ratios, signals the presence of employer's contribution defaults. The study is exploratory, meaning that it aims at understanding the characteristics of the companies with tax defaults, rather than confirming a hypothesis or predicting the likelihood that the company is in the high-risk group. In addition, the analysis is at a preliminary stage in the sense that it does not include all relevant characteristics of the companies, but only a limited amount, namely four ratios representing financial performance. However, the study is useful because it provides evidence that a certain group of companies that have employer's contribution defaults presents particular characteristics and that more variables are needed to fully capture the particularities of the companies of interest. Methodologically, the study shows that the logistic regression is useful for modeling the differences between the companies with tax defaults and the ones without tax defaults, and that the pre-processing of the data in terms of filtering out the companies into meaningful groups is as important in modeling as the selection of appropriate variables.

The rest of the paper is structured as follows. Section 2 presents a summary of approaches used in the selection of taxpayers for auditing. Section 3 discusses briefly the financial performance ratios for bankruptcy prediction. Section 4 describes an empirical study that uses logistic regression analysis to investigate the relationships between financial performance and employer's contribution. Section 5 describes ten models obtained by analyzing the data. Section 6 discusses the results and the limitations of the study. Section 7 concludes the paper.

2 Selection of Taxpayers for Auditing

The selection of tax payers for inspection regards the identification of profiles of companies that are likely to provide erroneous or fraudulent tax returns and the specification of models that estimate the

[1] In this paper, we use the term employers' contribution [21]. This refers to what is defined in other sources as the payroll tax.

probability that a company has a high-risk of being inspected [2,5]. One approach to select companies for tax auditing is to assign a risk score to each company. The score measures the likelihood of that company to have discrepancies between the data provided and the real situation [2,5,6].

Another researched approach is to test data against the Benford's law [7] in order to detect anomalies in lists of numbers representing the financial indicators provided by companies to the tax authority. Benford's law is the expected frequency distribution of the 0-9 digits in a given dataset [8]. This approach is used in [4,8,9] for detecting accounting manipulations.

A clustering approach for selecting companies for tax auditing using the Self-Organizing Map technique has been explored in [10]. The dataset is partitioned based on eight variables suggested by the Finnish Tax Authority. A model is successful if a high-risk cluster containing a high proportion of inspected companies with large tax defaults is identified. Moreover, the number of uninspected companies assigned to the high-risk cluster should be reasonable, because the model assumption is that the uninspected companies that belong to the high-risk cluster are similar to the ones inspected and, therefore, likely to generate similar tax corrections if audited.

3 Financial Performance

Financial performance indicators are usually defined as ratios in order to allow comparisons across companies and over time [11]. The ratios are extensively used in bankruptcy prediction models. The empirical research in the area of bankruptcy prediction started with [12-14]. Their aim was to discover from the data the characteristics of the companies likely to fail. The explanatory variables are usually financial ratios representing profitability, solvency and liquidity ratios [11]. Empirical studies using Finnish data are, for example, conducted to determine the explanatory power of ratios in bankruptcy prediction using different multivariate methods [15] and to benchmark companies using the Self-Organizing Map [16].

The bankruptcy prediction models are important, for example, in bank lending because banks need to predict the probability of default of a firm that solicits a credit [17]. Moreover, the financial ratios are recently explored for being used to detect false financial statements [18].

The topic of bankruptcy prediction is also relevant to that of tax auditing. Similarly with credit risk assessment, the financial ratios could be used to detect companies that present large tax defaults. The motivation of using financial ratios for identifying companies with tax defaults is suggested by studies such as [9] that provide evidence that in Finland some companies report lower income in order to avoid taxes. Reporting lower income for the purpose of tax evasion could be therefore tracked by analysing the financial performance of the companies. On the other hand, a Finnish Police report mentions that companies involved in economic crimes usually have a short life-cycle, presenting a tendency to go bankrupt shortly after they start up [19]. Thus, a low performance of companies and the likelihood of becoming bankrupt in the near future may signal possible tax evasion behaviour.

4 An Empirical Study

Inspired by the above studies and evidence, we evaluate the power of four ratios to indicate tax defaults. We focus on employer's contribution defaults (ECD). The ECD are a remarkable and surprisingly common area in grey economy. It has been estimated that in EU there are over 30 million employees getting "grey salaries" and the Finnish government has worked over fifteen years to solve this problem [20]. The employer's contributions are defined in [21] as being the amounts paid to the public administration in addition to, and because of, paying wages to an employee.

4.1 Data

The dataset consists of a sample of Finnish limited companies of a particular business line and industry in 2004. By a filtering procedure, we removed the companies for which no financial statements data were available to us. Here, we analyze only the inspected companies because they, unlike the uninspected ones, can be classified based on the presence of employer's contribution defaults.[2] The inspected companies consist of three classes: (1) *Clean*: companies that have been inspected and no defaults have been found; (2) *Employer's contribution defaults*: companies that have been found with employer's contribution defaults, and possibly with other tax defaults too; and (3) *Other tax defaults*: companies that have been found with other tax defaults than employer's contribution. Table 1 presents the companies in the dataset under analysis.

Table 1: Classification of the companies

Classification	Explanation	Count
Class 1	Clean	129
Class 2	Employer's contributions defaults	193
Class 3	Other tax defaults	152

4.2 Variables

The explanatory variables used in the models are two profitability ratios, one solvency ratio, and one liquidity ratio. The ratios[3] are selected based on the data availability and their relation with bankruptcy prediction models. The dependent variable is binary and indicates the presence of the employer's contribution defaults.

4.3 Method

For modeling the relationships between the financial performance and the ECD we use binomial logistic regression in SPSS Modeler [22]. Based on the classification of companies and different pre-processing operations, we built several binomial logistic regression models. The pre-processing steps are described in Section 4.3.1, and a brief review of the binomial logistic regression is presented in Section 4.3.2.

The models are evaluated in terms of (1) the explanatory power of the independent variables and (2) the classification performance. The importance of each independent variable in the model is measured by the Wald statistic. However, this test is sensitive to very large estimates, leading to type II errors, i.e., variables that have large coefficients estimates are regarded as not significant (when they are in fact significant). The model fit is assessed using three statistics: (1) *Hosmer and Lemeshow test*, (2) *Omnibus test*, and (3) *Nagelkerke R-squared*. Hosmer and Lemeshow test is used to assess the overall fit of the model. It calculates a chi-square statistic and if its value is not significant it indicates that the differences between the predicted and observed probabilities are not significant, therefore the model adequately fits the data. The Omnibus test is also based on the chi-square method. It tests whether the predictors influence significantly the dependent variable. If the test indicates significance, then at least one of the predictors contributes significantly to predicting the response. The Nagelkerke R-squared is a measure of explained variance in the dependent variable, similar to the R-squared in multiple regression analysis. Thus, it is an attempt to measure the strength of association between the independent and dependent variables. However, its value is sensitive to the frequency distribution of the dependent variable. Due to the fact that not all datasets are balanced, this measure is more difficult to interpret than the corresponding R-squared in multiple regression models. Nevertheless, a value close to 1 is desirable. If the value is close to zero, it means that the model is underspecified, and therefore, more variables should be added to the model.

[2] The problem setting is therefore different from that of bankruptcy prediction, where the number and identities of the companies that went bankrupt are known.

[3] The names and definitions of the independent variables are confidential.

The classification performance of the model is measured by the overall accuracy rate, the true positive rate and the false positive rate. The true positive rate measures the proportion of correctly classified companies as having employer's contribution defaults. The false positive rate calculates the proportion of companies with no employer's contribution defaults that are misclassified.

4.3.1 Pre-processing

First, we perform univariate and bivariate statistical analyses in order to test whether the ratios are normally distributed, present outliers, and are correlated. We also study the discriminatory power of the four ratios. The outliers are defined according to the definition used by the box-plot technique [23, p. 28]. The box plot uses the interquartile range (IR) as a measure of spread of the distribution. The interquartile range is the distance between the upper and lower quartiles (the 75th and the 25th percentiles respectively). The outliers are the values that are either larger than P75+1.5 x IR or smaller than P25-1.5 x IR, where P75 and P25 are the 75th and 25th percentiles, respectively [24, pp. 25-27]. To measure the correlations between variables we computed the Spearman's correlation coefficient. To test the differences between classes in terms of the financial ratios, we use the Median test, with the significance level 0.05 for the two-tailed region of rejection. The Median test is a non-parametric procedure used for testing whether two or more groups differ in central tendencies measured by the median in each group [25].

4.3.2 Binomial logistic regression analysis

Binomial logistic regression is employed to investigate the relationships between the *set* of independent variables and the dependent variable. This method is a type of regression that is used when the dependent variable is binary, that is, it has only two values, for example, positive or negative, 0 or 1, etc. In our problem setting, a company may or may not have employer's contributions defaults. The method is useful for (1) predicting the dependent variable based on a number of independent variables (predictors) of any type; (2) determining the percent of variance in the dependent variable explained by the predictors; (3) ranking the relative importance of the predictors; and (4) understanding the influence of the explanatory variables [26]. In this study, the focus is especially on the last three tasks, the first one, prediction, is only employed to evaluate the models (in fact, only classification as we do not test the models on new datasets).

The dependent variable is determined as the estimated probability that the event associated with the dependent variable occurs [22]. In our case, the event is the presence of employer's contributions defaults. Probability of this event is defined based on the logistic curve and the values of the independent variables as in the following equation.

$$P(event) = \frac{1}{1+e^{-z}},$$

where Z is the linear combination of the *n* independent variables included in the model, i.e., $Z = B_0 + B_1X_1 + B_2X_2 + ... + B_nX_n$. *Z* is also known as an unobservable factor that influences the event of interest. For our data, the model is defined by equation (1).

$$P(ECD=1) = \frac{1}{1+e^{-(B_0+B_1R_1+B_2R_2+B_3R_3+B_4R_4)}} \qquad (1),$$

where ECD=1 indicates the presence of employer's contributions defaults, and $R_i, i=1,...,4$ are the financial ratios. To estimate the coefficients of the logistic model, $B_i, i=0,...,4$, the maximum-likelihood method is used. This method is an iterative algorithm that selects the coefficients that make the observed results most likely. Once the coefficients are estimated, the predicted probabilities can be calculated by replacing the values in the equation (1).

Norušis [22, p. 322] warns about the differences in the probability estimates that may appear due to the fact that the model has been built based on a different mix of cases in the sample than in the population. The probability estimates are calculated based on the proportion of cases in the sample that experienced the event of interest. If this proportion does not reflect the one in the population, then the estimated probabilities are not correct for the population, because the estimate of the constant term in the model differs. However, the coefficients' estimates are correct and the warning concerns only the validity of the prediction or classification accuracy of the models. This means that, for example, if in the whole population only 3% of the instances present the event of interest, but the model is based on a sample in which 30% of the instances are in the category of interest, then the probability estimates, that is, the estimated value of the dependent variable, can exhibit differences from the real values. However, the model is reliable in terms of explanatory power of each independent variable. Therefore, in our models, the coefficients estimates are easy to interpret, however, the probability estimates have to be interpreted with caution, since the true proportion of the companies with ECD in the whole population is unknown.

5 Results

5.1 Initial data pre-processing

Based on the univariate data analyses, we found that no ratio is normally distributed. This result is in line with [27] which points out that rarely financial data are normally distributed. All four ratios have outliers, especially abnormally smaller values than typical values. Three of the ratios have also very large values (Ratios 2, 3 and 4). In this dataset, 95 companies have outliers in one or more of the ratios.

Because outliers may distort the analysis results, we analyze separately the dataset of 474 companies and the dataset of 379 companies that does not include the outliers. We denote these datasets by *'474 dataset'* and *'379 dataset'*, respectively. In the '379 dataset', there are 51 companies that have all four ratios equal to zero. They are also likely to distort the analysis results, because they come mainly from missing or incomplete data. Therefore, we analyze separately the dataset after these 51 companies are removed. We refer to this dataset as *'328 dataset'*. In addition, we consider a fourth subset obtained by removing the companies with all ratios zero from the '474 dataset'. This set is referred to as *'423 dataset'*.

The Spearman's correlation coefficients indicate that the four ratios are correlated among themselves, this fact posing the problem of multicolinearity in the logistic regression model. A solution to this problem can be to perform the principal components analysis to obtain new variables that are not correlated to each other, but preserving the variance in the data. This approach will be considered for future work. The correlations between the four ratios and the dependent variable (the presence of ECD) are presented in Table 2.

Table 2: Correlation coefficients' sign and significance in different datasets

	ECD in datasets		
	'474'	*'379'*	*'328'*
Ratio 1			+*
Ratio 2	-*	-**	
Ratio 3	-**	-**	
Ratio 4	-**	-**	
Legend:	+* positive relationship significant at 0.05		
	-** negative relationship significant at 0.01		
	-* negative relationship significant at 0.05		

Table 3 presents the results of the Median test for measuring the significance of the ratios in discriminating between the two classes of companies in the four datasets. With bold are marked the significant variables at level 0.05.

Table 3: Significant differences between groups in different datasets using Median test

Dataset	*Ratio 1*	*Ratio 2*	*Ratio 3*	*Ratio 4*
‘474’	0.08	**0.00**	**0.01**	**0.00**
‘423’	0.81	0.33	0.30	0.08
‘379’	0.38	**0.00**	**0.05**	**0.00**
‘328’	0.59	0.26	0.82	0.5

The change in significance of Ratios 2, 3 and 4 may indicate that the companies with zero-value ratios that belong to the ‘474’ and’379’ datasets can be the ones which make the difference between the two groups of companies investigated.

We performed the logistic regression modeling on different datasets to see what pre-processing approach could be more useful in the task of obtaining meaningful patterns. First, we analyze separately the ‘474’, 379’, 328’, and ‘423’ datasets. Second, in the ‘474 dataset’ we add four binary variables that signal the presence of ratios with zero values ($Z_i, i = 1,...,4$). One binary variable is added for each ratio. The binary variables Zi are entered in the model as categorical variables. Because there are only two possible values for each Zi, in this case the first value, i.e., 0 is represented as a dummy variable. The dummy variable is interpreted as follows; Zi=0 corresponds to Ri=0, and Zi=1 corresponds to Ri≠0, $\forall i = 1,...,4$.

A third approach is to remove from the ‘474 dataset’ companies with other tax defaults than employer’s contributions to observe whether the four ratios explain the difference between clean companies and the ones with ECD. Similarly, we remove from the ‘474 dataset’ the clean companies and perform logistic regression on the remaining dataset to observe whether the financial ratios and the four binary variables added explain the differences between companies with ECD and those with other tax defaults. Moreover, we examine the extent to which the financial ratios influence the probability that a company has other tax defaults than employer’s contributions, when the companies with ECD are removed from the dataset.

Fourth, because the ‘474 dataset’ is imbalanced, i.e., the number of companies with ECD is lower than the rest of the companies, we create a balanced dataset by randomly sampling a fixed number of companies from the other two classes, clean and other tax defaults. In addition, in a similar manner we also create a balanced dataset that includes only the clean companies and the companies with ECD. Table 4 presents the above datasets and a short description of them.

Table 4: Datasets used in binary logistic regression

Model id	**Dataset**	**Description**
1	‘474’	The original dataset.
2	‘423’	The original dataset without the companies that have all ratios equal to zero.
3	‘379’	The ‘474’ set without outliers.
4	‘328’	The ‘423’ set without outliers.
5	‘474 + 4B’	The original dataset to which four binary variables are added that signal the ratios that are equal to zero.
6	‘474+4B balance’	‘474+4B’ to which a fixed number of companies from class clean and with other tax defaults than ECD are deleted by randomly sampling from the mentioned groups, so that the dataset becomes balanced, i.e., the number of companies with ECD equals the number of the rest of the companies.
7	‘474+4B clean vs. EC’	‘474+4B’ from which the companies with other tax defaults than ECD are removed.
8	‘474+ 4B other vs. EC’	‘474+4B’ from which the clean companies are removed.
9	‘474+ 4B clean vs. other’	‘474+4B’ from which the companies with ECD are removed.
10	‘474+4B clean vs. EC balance’	‘474+4B clean vs. EC’ to which a fixed number of clean companies are duplicated by random sampling, so that the dataset becomes balanced.

5.2 Models

For each dataset in Table 4, we create a binomial logistic regression model. The evaluation of each model regards the model estimates and Wald significance level of each independent variable. Moreover, the goodness-of-fit and classification performance measures are presented in a separate table where the measures are denoted from 1 to 8 as follows.

1: Hosmer and Lemeshow test; a p-value higher than .05 shows a good model fit.
2: Omnibus test; a p-value smaller than .05 shows a good model fit.
3: Nagelkerke R-squared test; a value close to 1 is desirable showing good model specification.
4: Baseline accuracy % of a naive model.
5: Model accuracy % = the classification accuracy rate of the model.
6: True positive rate %.
7: False positive rate %.
8: Cutoff = the threshold value of the probability estimated by the model that a company is found with ECD.

Model 1

The first model is built for the entire dataset selected for analysis (Table 1). This consists of 474 companies, out of which 193 have ECD. The dataset is imbalanced, contains outliers and companies with all ratios equal to zero. The results show that none of the independent variables has a significant contribution for estimating the probability of ECD (Table 5). Moreover, none of the evaluation measures exhibits good model fit or classification accuracy (Table 6). The results mean that a model that is based on these four ratios and includes outliers and companies with all ratios equal to zero (due to misreporting or no activity) does not provide reliable estimates, and therefore cannot be used for explanation or prediction of present or future ECD, respectively.

Table 5: The estimates and their significance levels in Model 1

Variable	Estimate’s sign	Wald significance level
R1	-	0.444
R2	+	0.291
R3	-	0.326
R4	+	0.385
Constant	-	0.070

Table 6: Goodness-of-fit and classification accuracy in Model 1

	Goodness of fit			Classification accuracy				
Measures	1	2	3	4	5	6	7	8
Model 1	0.009	0.068	0.025	59.03	58.6	0.5	1.4	0.5

The poor fit of the model is illustrated in Figure 1 that displays the estimated probability of ECD for the companies with ECD (marked with 1) and without ECD (marked with 0). The plot shows that the model is not able to discriminate between these two types of companies, both of them being assigned low probabilities of having ECD.

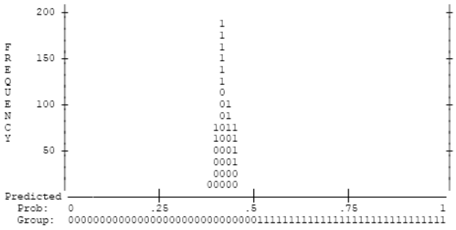


Figure 1. Classification performance in Model 1

Model 2

The second model is built for the dataset after removing the companies with all ratios equal to zero. Ratio R1 appears to be significant and it is positively related with the probability of ECD (Table 7). Moreover, two goodness-of-fit measures (HL and Omnibus tests) show good model fit, but the classification accuracy of the model is not satisfactory (Table 8). The HL test indicates that the difference between predicted and observed probabilities is not significant, thus the model fits the data. The Omnibus test shows that at least one variable in the model is able to contribute significantly to the dependent variable. However, the third measure shows that the model is underspecified, thus more variables are needed to fully capture the variance in the dependent variable.

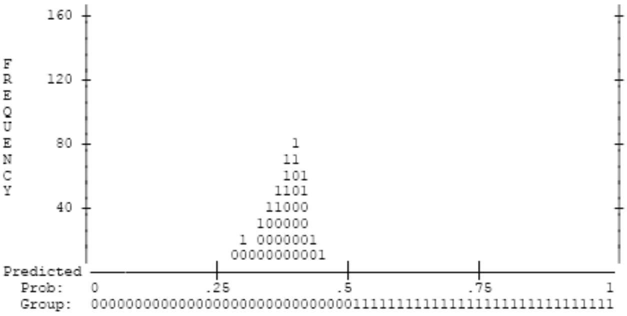
Table 7: The estimates and their significance levels in Model 2

Variable	Estimate's sign	Wald significance level
R1	+	**0.088**
R2	+	0.739
R3	-	0.330
R4	-	0.253
Constant	-	0.000

Table 8: Goodness-of-fit and classification accuracy in Model 2

	Goodness of fit			Classification accuracy				
Measures	1	2	3	4	5	6	7	8
Model 2	**0.761**	**0.046**	0.031	63.6	63.4	0.6	0.7	0.5

The classification performance of the model is illustrated in Figure 2. The graph shows that the two types of companies are not well separated based on this model.



```
       160 +                                                            +
           |                                                            |
 F         |                                                            |
 R     120 +                                                            +
 E         |                                                            |
 Q         |                                                            |
 U         |                                                            |
 E      80 +                       1                                    +
 N         |                      11                                    |
 C         |                      101                                   |
 Y         |                     1101                                   |
        40 +                    11000                                   +
           |                   100000                                   |
           |                 1 0000001                                  |
           |                0000000001                                  |
Predicted  ---------------+--------------+--------------+--------------
   Prob:   0             .25             .5            .75             1
   Group:  000000000000000000000000000000111111111111111111111111111111
```

Figure 2. Classification performance in Model 2

Model 3

The third model is built for the dataset after removing the outliers. The companies with all ratios zero are included in the model. Only Ratio R4 appears significant, and it is negatively related with the probability of ECD (Table 9). Only Omnibus test shows good model fit, in particular that at least one variable (i.e., R4) is significant. However, the model is underspecified and the differences between predicted and observed probabilities are significant. On the other hand, the prediction accuracy is higher than in the previous models, but the rate of false positives also increases in this model (Table 10).

Table 9: The estimates and their significance levels in Model 3

Variable	Estimate's sign	Wald significance level
R1	+	0.967
R2	-	0.794
R3	-	0.122
R4	-	**0.016**
Constant	-	0.240

Table 10: Goodness-of-fit and classification accuracy in Model 3

	Goodness of fit			Classification accuracy				
Measures	1	2	3	4	5	6	7	8
Model 3	0.043	**0.002**	0.059	55.9	**62.8**	**45.5**	**23.6**	0.5

Figure 3 shows that the model is able to separate some of the companies with the ECD from the rest. However, for the chosen cutoff value 0.5 of the estimated probability, the separation is not very good, resulting in a high false positive rate. At a higher cutoff value, the false positive rate will decrease, on the expense of a small true positive rate.

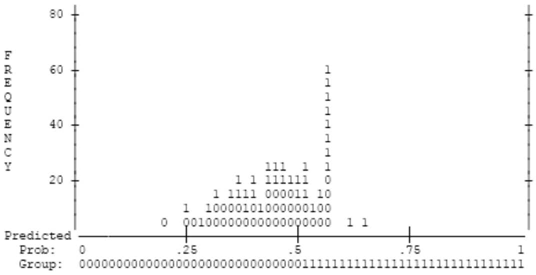


Figure 3. Classification performance in Model 3

Model 4

Model 4 is based on the dataset after removing both outliers and companies with all ratios equal to zero. Table 11 presents the signs of the estimates in the model and the Wald significance levels; significant associations are in bold. The ratio R1 is positively related with the presence of ECD; the higher R1, the higher the probability that the company has ECD. Ratio R4 is negatively associated with the presence of ECD; a lower R4 indicates a higher probability that the company has ECD.

Table 12 presents measures of model performance. The model fits the data (measures 1 and 2), but it is underspecified (measure 3). The classification accuracy (61.6%) measures the extent to which this model is able to distinguish between the two types of companies; it is very close to that of a naïve model that predicts the class with the highest frequency (i.e., companies without ECD). The rate of false positives is small (9.5%), while only 16.4% of the companies with ECD are correctly classified.

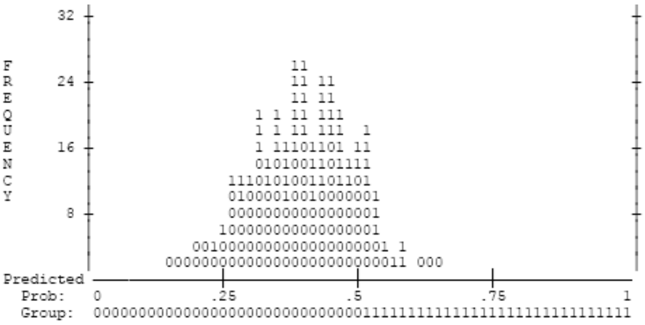
Table 11: The estimates and their significance levels in Model 4

Variable	Estimate's sign	Wald significance level
R1	+	**.009**
R2	-	.652
R3	-	.393
R4	-	**.082**
Constant	-	.010

Table 12: Goodness-of-fit and classification accuracy in Model 4

	Goodness of fit			Classification accuracy				
Measures	1	2	3	4	5	6	7	8
Model 4	**0.778**	**0.018**	0.049	61.0	**61.6**	**16.4**	**9.5**	0.5

Figure 4 shows graphically the classification performance. The separation between the two types of companies is not clear in this model, though the model is at some extent useful and significant.



```
          32 +                                                            +
             |                                                            |
F            |                                                            |
R         24 +                      11                                    +
E            |                      11 11                                 |
Q            |                      11 11                                 |
U            |                  1 1 11 111                                |
E         16 +                  1 1 11 111  1                             +
N            |                  1 11101101 11                             |
C            |                  0101001101111                             |
Y            |               1110101001101101                             |
           8 +               01000010010000001                            +
             |               00000000000000001                            |
             |              100000000000000001                            |
             |           0010000000000000000001 1                         |
             |        000000000000000000000000011 000                     |
Predicted ---------------+---------------+---------------+---------------
  Prob:   0             .25              .5             .75              1
  Group:  000000000000000000000000000000111111111111111111111111111111
```

Figure 4. Classification performance in Model 4

Model 5

Model 5 is based on the original dataset to which four binary variables are added. The four dummies, denoted by Z1, Z2, Z3 and Z4, are introduced to observe whether the zero values of the ratios can have an impact on the estimation of ECD presence. The results show that two of the dummies, corresponding to the ratios R3 and R4, are significantly negatively related to the estimated probability (Table 13). Moreover, all three goodness-of-fit measures indicate model fit (Table 14). However, the low value of Nagelkerke measure (3) shows that the model is underspecified, meaning that more variables are needed in the model.

The classification accuracy of this model is the best among all analyzed models. The rate of false positives is also one of the lowest when compared to the true positives rate. The graphic illustrating the separation of the companies based on this model is shown in Figure 5.

Table 13: The estimates and their significance levels in Model 5

Variable	Estimate's sign	Wald significance level
R1	+	.206
R2	+	.656
R3	-	.347
R4	-	.368
Z1	-	.875
Z2	+	.231
Z3	-	**.098**
Z4	-	**.008**
Constant	+	.001

Table 14: Goodness-of-fit and classification accuracy in Model 5

	Goodness of fit			Classification accuracy				
Measures	1	2	3	4	5	6	7	8
Model 5	**0.808**	**0**	**0.115**	59.3	**64.1**	**25.4**	**9.3**	0.5

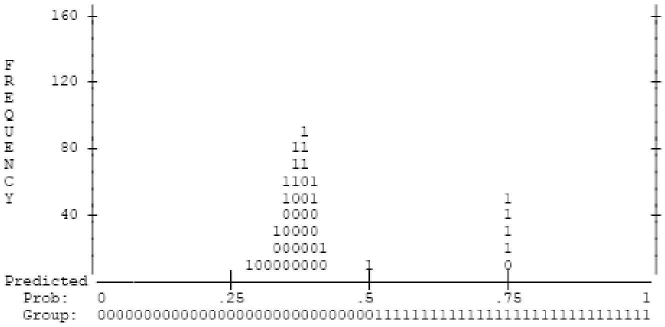


Figure 5. Classification performance in Model 5

Model 6

Model 6 is the balanced version of Model 5, so that the number of companies with ECD equals the number of companies without ECD. There are no important differences between Model 6 and Model 5, except that the overall accuracy is lower in Model 6, while the true positives and false positives rates are higher (Tables 15,16). In both models, there is a group of companies with ECD that clearly separates from the rest of companies (Figure 6). These account for 31.6% of the companies with ECD.

Table 15: The estimates and their significance levels in Model 6

Variable	Estimate's sign	Wald significance level
R1	+	.558
R2	+	.551
R3	-	.314
R4	-	.142
Z1	+	.595
Z2	+	.498
Z3	-	**.051**
Z4	-	**.065**
Constant	+	.000

Table 16: Goodness-of-fit and classification accuracy in Model 6

	Goodness of fit			Classification accuracy				
Measures	1	2	3	4	5	6	7	8
Model 6	**0.421**	**0**	**0.13**	50.1	**59.4**	**31.6**	**12.9**	0.5

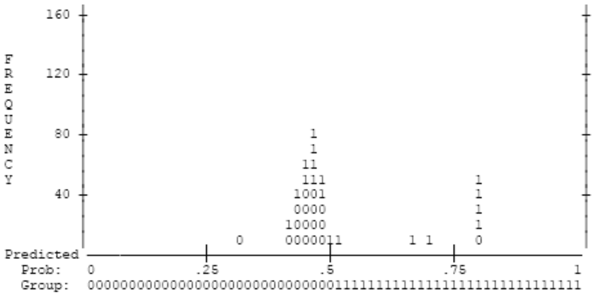


Figure 6. Classification performance in Model 6

Model 7

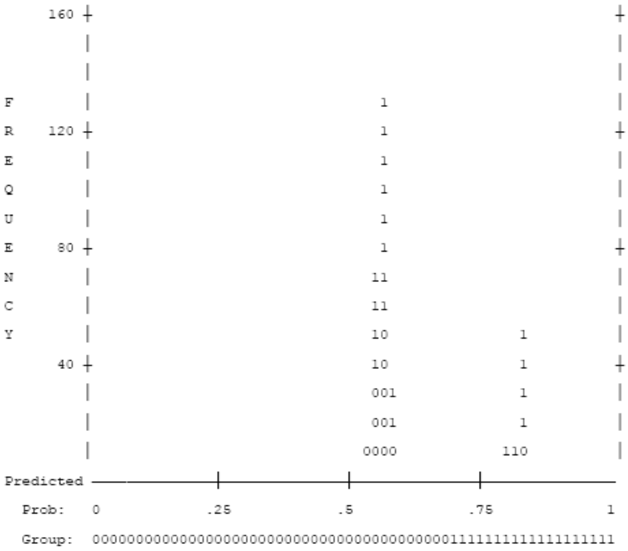
Model 7 uses only the companies with ECD vs. the clean companies. Only the two dummy variables corresponding to ratios R3 and R4 are significant. They are negatively associated with the dependent variable (Table 17). It means that the lower Z3 and Z4, the more likely is that the companies have ECD. In other words, the companies which have R3 and R4 equal to zero are likely to have ECD. However, the evaluation of the model shows that the model accuracy is lower than of a naïve model, where the most frequent class (companies with ECD) is assigned to all companies (Table 18). Therefore, the model can classify correctly based on this rule only 28.5% of the companies with ECD, while producing 9.3% of misclassifications of the clean companies. This fact is also illustrated in Figure 7, which shows that the correctly classified companies with ECD are clearly separated from the rest, while the model produces also a number of false alarms.

Table 17: The estimates and their significance levels in Model 7

Variable	Estimate's sign	Wald significance level
R1	+	.763
R2	-	.606
R3	-	.331
R4	+	.772
Z1	+	.788
Z2	+	.133
Z3	-	**.066**
Z4	-	**.032**
Constant	+	.000

Table 18: Goodness-of-fit and classification accuracy in Model 7

	Goodness of fit			Classification accuracy				
Measures	1	2	3	4	5	6	7	8
Model 7	**0.108**	**0.001**	**0.108**	59.9	53.4	28.5	9.3	0.7



```
       160 +                                                                  +
           |                                                                  |
           |                                                                  |
F          |                                     1                            |
R      120 +                                     1                            +
E          |                                     1                            |
Q          |                                     1                            |
U          |                                     1                            |
E       80 +                                     1                            +
N          |                                    11                            |
C          |                                    11                            |
Y          |                                    10                 1          |
        40 +                                    10                 1          +
           |                                    001                1          |
           |                                    001                1          |
           |                                   0000              110          |
Predicted  ---------------------+----------------+----------------+--------------
  Prob:    0                   .25               .5              .75            1
  Group:   000000000000000000000000000000000000000000000000011111111111111111111
```

Figure 7. Classification performance in Model 7

Model 8

Model 8 compares the companies with ECD with the ones having other type of tax defaults (e.g., income tax or VAT). The results show that these two categories of companies are different with respect to Ratio 1, Ratio 4 and the dummy corresponding to R4 (Table 19). The positive sign of the estimate for R1 indicates that the higher R1, the higher the probability that the company has ECD. In addition, the negative signs for R4 and Z4, indicate that the lower R4, or if it is equal to zero, then the probability of having ECD increases. The model fits the data well (Table 20), but it is underspecified (lower value of Nagelkerke measure). Though the overall accuracy is lower than of a naïve model, there are 23.5% of the companies with ECD that can be characterized by a high R1 and low or zero R4, when the threshold value is set to 0.8. In addition, the rate of false positives is very low. Figure 8 shows that the group of 23.5% companies with ECD that are correctly classified using this model is well separated from the rest of the companies.

Table 19: The estimates and their significance levels in Model 8

Variable	Estimate's sign	Wald significance level
R1	+	**.024**
R2	+	.518
R3	-	.586
R4	-	**.011**
Z1	-	.757
Z2	+	.554
Z3	-	.425
Z4	-	**.048**
Constant	+	.000

Table 20: Goodness-of-fit and classification accuracy in Model 8

	Goodness of fit			Classification accuracy				
Measures	1	2	3	4	5	6	7	8
Model 8	**0.584**	**0**	**0.165**	55.9	55.4	23.8	4.6	0.8

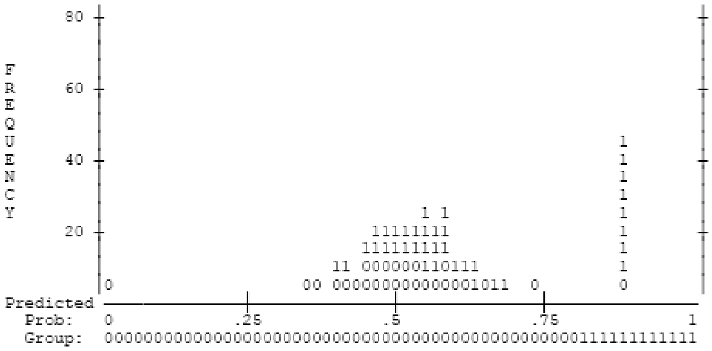


Figure 8. Classification performance in Model 8

Model 9

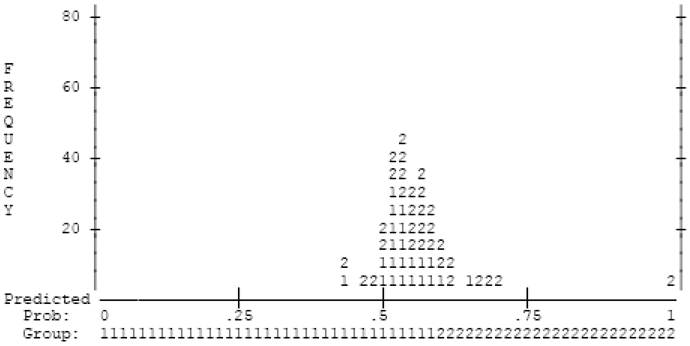
Model 9 aims at observing differences between the companies with other tax defaults than ECD and the clean companies. The results show that none of the four ratios and their corresponding dummies is able to discriminate between the two types of companies (Table 21). The model therefore does not fit the data (Table 22). Figure 9 illustrates that the clean companies are hardly separable from the companies with other type of tax defaults than ECD.

Table 21: The estimates and their significance levels in Model 9

Variable	Estimate's sign	Wald significance level
R1	-	.196
R2	-	.463
R3	-	.455
R4	+	.226
Z1	+	.819
Z2	+	.422
Z3	-	.237
Z4	+	.610
Constant	-	.561

Table 22: Goodness-of-fit and classification accuracy in Model 9

	Goodness of fit			Classification accuracy				
Measures	1	2	3	4	5	6	7	8
Model 9	0.047	0.085	0.064	54.1	51.2	17.8	9.3	0.6



```
           80 +                                                            +
              |                                                            |
              |                                                            |
F             |                                                            |
R          60 +                                                            +
E             |                                                            |
Q             |                                                            |
U             |                               2                            |
E          40 +                              22                            +
N             |                              22 2                          |
C             |                              1222                          |
Y             |                              11222                         |
           20 +                             211222                         +
              |                             2112222                        |
              |                         2   11111122                       |
              |                         1 2211111112 1222                 2|
Predicted      --------------+--------------+--------------+---------------
  Prob:        0            .25            .5             .75             1
  Group:       111111111111111111111111111111111112222222222222222222222222
```

Figure 9. Classification performance in Model 9

Model 10

Finally, Model 10 is the balanced version of Model 7. The results are similar with Model 7, but in addition to the dummies for R3 and R4, the dummy of Ratio 2 is also found significant (Table 23). The signs of the estimates indicate that, if R3 and R4 are equal to zero and R2 is high, then the company is likely to have ECD. The model fits the data and it has a relatively good accuracy performance when the cutoff is set to 0.7, compared with other models (Table 24). Figure 10 illustrates that there is a group of companies (28.5% of the companies with ECD), which follow this pattern.

Table 23: The estimates and their significance levels in Model 10

Variable	Estimate's sign	Wald significance level
R1	+	.669
R2	-	.284
R3	-	.276
R4	+	.690
Z1	-	.916
Z2	+	**.092**
Z3	-	**.017**
Z4	-	**.014**
Constant	+	.000

Table 24: Goodness-of-fit and classification accuracy in Model 10

	Goodness of fit			Classification accuracy				
Measures	1	2	3	4	5	6	7	8
Model 10	**0.435**	**0**	**0.134**	50	**60.1**	**28.5**	**8.3**	0.7

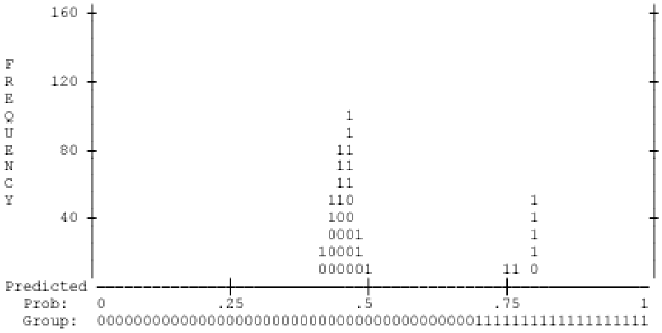


Figure 10. Classification performance in Model 10

6 Summary and discussion

Table 25 presents the significance levels of the Wald statistic in all ten models. Typically, a variable is considered significant if the value in the table is lower than 0.05. However, values below 0.1 can also be regarded as acceptable. The values in bold indicate the variables significant at the 0.1 level. Generally, the Ratios 2 and 3 are not significant; therefore they can be removed from a future analysis. However, the dummy corresponding to Ratio 3 appears in many models as significant; similarly dummy Z2 is significant in Model 10. Ratios 1, 4 and dummy Z4 are significant in some models.

Table 25: Wald significance levels for individual variables in the models corresponding to different datasets

	Models (Datasets)									
Variable	1	2	3	4	5	6	7	8	9	10
R1	.444	**.088**	.967	**.009**	.206	.558	.763	**.024**	.196	.669
R2	.291	.739	.794	.652	.656	.551	.606	.518	.463	.284
R3	.326	.330	.122	.393	.347	.314	.331	.586	.455	.276
R4	.385	.253	**.016**	**.082**	.368	.142	.772	**.011**	.226	.690
Z1					.875	.595	.788	.757	.819	.916
Z2					.231	.498	.133	.554	.422	**.092**
Z3					**.098**	**.051**	**.066**	.425	.237	**.017**
Z4					**.008**	**.065**	**.032**	**.048**	.610	**.014**
Constant	.070	.000	.240	.010	.001	.000	.000	.000	.561	.000

Table 26 presents the signs of the model estimates for each independent variable, including the dummies. A negative value in Table 26 indicates that there is an inverse relationship between the predictor and the dependent variable, e.g., the relationships between R4 and Z4 and the presence of ECD in models 3-8 and 10. In these models, a small value of R4 is associated with a high probability of having employer's contributions defaults. In addition, the fact that R4 is different than zero (Z4) decreases the probability that the company has ECD; in other words, if a company has R4 equal to zero, it has a higher probability that has ECD.

Table 26: The signs of the coefficients' estimates in different models

	Models (Datasets)									
Variable	1	2	3	4	5	6	7	8	9	10
R1	-	+	+	+	+	+	+	+	-	+
R2	+	+	-	-	+	+	-	+	-	-
R3	-	-	-	-	-	-	-	-	-	-
R4	+	-	-	-	-	-	+	-	+	+
Z1	n.a.	n.a.	n.a.	n.a.	-	+	+	-	+	-
Z2	n.a.	n.a.	n.a.	n.a.	+	+	+	+	+	+
Z3	n.a.	n.a.	n.a.	n.a.	-	-	-	-	-	-
Z4	n.a.	n.a.	n.a.	n.a.	-	-	-	-	+	-
Constant	-	-	+	-	+	+	+	+	-	+

A similar interpretation is given to the negative estimates for the Z3 variable. In all models where this variable is present, the estimate is negative, indicating that a value of R3 different than zero decreases the probability that the company has ECD. In other words, a company with R3 equal to zero has a higher probability that has ECD. In models 5-7 and 10 this relationships is found as significant at level 0.1.

On the other hand, a positive value in Table 26 indicates the presence of a direct relationship between the independent and dependent variables. For example, in models 2,4 and 8, the ratio R1 is found as significantly positively contributing to the probability that the event of interest occurs. In addition, the estimates for the Z2, indicating values of the ratio R2 different than zero, have relatively large values, even if they do not appear significant at the 0.1 level, except in model 10. Especially in model 10, the estimate indicates that values of R2 different than zero are associated with a higher probability of ECD.

From the point of view of tax auditors, the variables' significance and contribution in the models are very important in the task of selecting companies for inspection. Therefore, our analysis can provide useful insights of how important a variable is in the model and in what way each variable influences the probability that a company presents ECD. In addition, one can also investigate the odds ratios of each variable for obtaining a more precise measure of the effect size.

However, it is also important to evaluate the overall validity of the model, by calculating and examining the goodness-of-fit and classification accuracy measures (Table 27). In Table 27, there are reported the significance levels for the HL and Omnibus tests, the Nagelkerke R-squared value, and the classification accuracy measures. We marked with bold the models that are found with acceptable goodness-of-fit and classification accuracy. Higher values of the Hosmer and Lemeshow test indicate that the predicted probabilities are not significantly different than the observed probabilities. According to this test, models 2, 4-8 and 10 are matching the data well. The Omnibus test shows that all models are significant at 0.1 in that at least one variable in the model significantly contributes to the prediction of the dependent variable. The Nagelkerke R-squared is a value that shows the strength of association between the predictors and the dependent variable. Models 5-8 and 10 have relatively higher values compared with other models, but they are very low compared to 1, which indicates that all models are underspecified and more variables are needed to be included in the analysis.

Table 27: Goodness-of-fit and classification accuracy measures of the models

	Models (Datasets)									
Measure	1	2	3	4	5	6	7	8	9	10
Goodness-of-fit tests										
HL test, p-value	.009	**.761**	.043	**.778**	**.808**	**.421**	**.108**	**.584**	.047	**.435**
Omnibus test, p-value	.068	**.046**	**.002**	**.018**	**.000**	**.000**	**.001**	**.000**	.085	**.000**
Nagelkerke R square	.025	.031	.059	.049	**.115**	**.130**	**.108**	**.165**	.064	**.134**
Classification accuracy										
Baseline accuracy %	59.3	63.6	55.9	61.0	59.3	50.1	59.9	55.9	54.1	50
Model accuracy %	58.6	63.4	**62.8**	**61.6**	**64.1**	**59.4**	53.4	55.4	51.2	**60.1**
True positive rate %	.5	.6	**45.5**	**16.4**	**25.4**	**31.6**	28.5	23.8	17.8	**28.5**
False positive rate %	1.4	0.7	**23.6**	**9.5**	**9.3**	**12.9**	9.3	4.6	9.3	**8.3**
Cutoff	.5	.5	.5	.5	.5	.5	.7	.8	.6	.7

The classification accuracy measures show that the models 3-6 and 10 have a better overall accuracy than the baseline models. Model 3 has the highest true positive rate, but also the highest rate of false alarms. This model corresponds to the dataset in which outliers are removed, but the companies with all ratios zeros are present in the dataset. In Model 4, the dataset consists of companies that have at least one of the ratios different than zero and they are not among the outliers. This model has a good rate of false alarms, but the recall rate of companies with ECD is lower than in other models. The remaining three models, namely 5, 6 and 10 have similar classification performances.

In summary, our study shows that the four chosen financial ratios are useful to some extent for identifying companies with employer's contribution defaults. If we compare Models 5 and 9, we see that the two dummies corresponding to Ratios 3 and 4 are important for discriminating among a group of companies with ECD and the rest of companies. Moreover, Models 4 and 8 shows that when Ratio 1 is higher and Ratio 4 is zero (low value or missing data), there is a possible sign that the company has employer's contribution defaults. On the other hand, the selected independent variables are not able to distinguish between clean companies and companies with other tax defaults than ECD. In addition, Ratios 2 and 3 seem insignificant and therefore can be removed from analysis, but holding the dummies in the models.

The results however should be interpreted with caution, because there are a number of limitations that can affect them. One limitation is the correlation among variables. To avoid this problem, one can eliminate the variables that are highly correlated with other variables or transform the variables into new ones using principal components analysis. Another limitation is the fact that there are many other variables representing financial performance or variables regarding closely the domain of interest, namely employer contribution, or employment, that are not included in the model. For future, we intend to collect all necessary data for creating a fully-specified model. Regarding the method, one can also use multinomial regression analysis in order to observe in more detail the differences between all types of tax defaults based on a certain set of variables. In addition, in order to create valid models that could be used in prediction, a separation of the dataset into training, test and validation subsets is necessary so that the model performance is reliably assessed. However, our study was of exploratory nature, aiming at discovering whether the financial performance can be used as an indicator of tax defaults. The results also show that the pre-processing of the data plays a crucial role in modeling. Selecting the appropriate set of companies for analysis is therefore important and challenging for succeeding in identifying the companies with tax defaults.

7 Conclusions

In this paper we analyzed to what extent the financial performance of a company can indicate tax defaults. The aim was to study whether the financial performance of companies can be used by the tax authority as an indicator for selecting companies for inspection in order to detect tax fraud or tax inconsistencies. The financial health of the companies was measured using four ratios, two of them defining the profitability, one - liquidity and another - the solvency of a company. We focused on one type of tax, namely employer's contribution or payroll tax. The analysis was conducted on a real dataset consisting of Finnish limited companies observed in 2004. The data was obtained from the Finnish Tax Authority. The dataset has been further pre-processed, thus resulting ten subsamples of companies that constituted the basis of our models. We created and evaluated the models based on the binomial logistic regression. The evaluation concerned the goodness-of-fit, classification accuracy, and the importance of the four ratios in predicting the probability that a company has employer's contribution default.

By developing a series of models based on different datasets, we found that the models which include dummy variables signaling the values of zero in the financial ratios are more useful for identifying the companies with employer's contributions defaults. In particular, the developed models indicate that the companies that report two particular ratios as being zero or give incomplete data are more likely to be found with discrepancies in the employer's contributions. Moreover, the four ratios and the derived dummies are more useful for identifying the companies with employer's contributions defaults from the rest of the companies, than for identifying the companies with other types of tax inconsistencies.

However, for generating correct classifications for all of the companies with employer's contribution defaults, more variables are necessary to be added because the models are underspecified. Future work will focus on selecting relevant ratios that can improve the models. Despite the limitations of the models, the study shows that this approach, if supplemented with all relevant ratios, could be used to select the companies with tax defaults for auditing.

In conclusion, the logistic regression models show that the four financial performance ratios are able to capture the differences between the companies with employer's contributions defaults and the rest of the companies only to a limited extent, namely they indicate that when two particular ratios are equal to zero they may signals inconsistencies in the employer's contributions. Moreover, when one particular ratio has higher values and other ratio is zero, there is a possible sign that the company has employer's contribution defaults. Future work is intended to identify other relevant ratios that can improve the models. Moreover, because ratios are correlated among themselves, we intend to use first principal components analysis to obtain new independent variables and then apply logistic regression.

Acknowledgements. We gratefully acknowledge the financial support from Tekes project no. 33/31/08 and from the Academy of Finland project no. 127592.

References

1. Finnish Tax Administration: Good Tax Auditing Practice (2010), http://www.vero.fi/?article=4137
2. Bakin, S., Hegland, M., Wiliams, G.: Mining taxation data with parallel BMARS. Parallel Algorithms and Applications, 15(1-2), 37–55, (2000)
3. Bolton, R.J., Hand, D.J.: Statistical fraud detection: A review. Statistical Science, 17(3), 235–255, (2002)
4. Watrin, C., Struffert, R., Ullmann, R.: Benford's law: An instrument for selecting tax audit targets? Review of Managerial Science, 2, 219–237, (2008)
5. McCalden, J.D.: Patent WO 01/22316 A1: Method and apparatus for selecting taxpayer audits. Patent, World Intellectual Property Organization, International Bureau, (2001)
6. Gillen, M.A., Packer, S.M.: New IRS strategic initiative: Increased audit on its way? The Legal Intelligencer, September 2, (2009)
7. Benford, F.: The law of anomalous numbers. Proc. Am. Philos. Soc., 78(4), 551–572, (1938)
8. Nigrini, M.J., Mittermaier, L.J.: The use of Benford's law as an aid in analytical procedures. Auditing: A Journal of Practice & Theory, 16(2), 52–67, (1997)
9. Niskanen, J., Keloharju, M.: Earnings cosmetics in a tax-driven accounting environment: Evidence from Finnish public firms. The European Accounting Review, 9(3), 443–452, (2000)

10. Kallio, K., Back, B.: The self-organizing map in selecting companies for tax audit. In: Proc. of the 32nd Annual Congress of the European Accounting Association, Tampere, (2009)
11. Salmi, T., Martikainen, T.: A review of the theoretical and empirical basis of financial ratio analysis. The Finnish Journal of Business Economics, 4, 426–448, (1994)
12. Beaver, R.: Financial ratios as predictors of failure. J. Accounting Research, 4, 71-111, (1966)
13. Altman, E.: Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. Journal of Finance, 13, 589-609, (1968)
14. Ohlson, J.: Financial ratios and the probabilistic prediction of bankruptcy. J. Accounting Research, 18, 109-131, (1980)
15. Back, B., Laitinen, T., Sere, K., van Wezel, M.: Choosing bankruptcy predictors using discriminant analysis, logit analysis and genetic algorithms. TUCS Technical report 40, (1996)
16. Eklund, T.: The Self-Organizing Map in Financial Benchmarking. TUCS PhD Thesis, Åbo Akademi University, (2004)
17. Atiya, A.F.: Bankruptcy prediction for credit risk using neural networks: A survey and new results. IEEE Trans. On Neural Networks, 12 (4), 929-935, (2001)
18. Spathis, C., Doumpos, M., Zopounidis, C.: Detecting falsified financial statements: A comparative study using multicriteria analysis and multivariate statistical techniques. European Accounting Review, 11(3), 509–535, (2002)
19. Keskusrikospoliisi: Rakennusalan yrityksiin kohdistuvan ja niitä hyödyntävän rikollisuuden teematilannekuva, Report no. KRP/RTP 393/213/2010, (2010) http://www.intermin.fi/intermin/hankkeet/turva/home.nsf/
20. Kosonen, E. (2010): http://www.palkkatyolainen.fi/pt2004/pt0403/p040407-t2.html
21. Finnish Tax Administration: Tax Glossary (2010), http://www.vero.fi/?path=488,684&domain=VERO_ENGLISH
22. Norušis, M. J.: PASW Statistics 18.0 Statistical Procedures Companion. Prentice Hall, (2010)
23. Hartwig, F., Dearing, B.E.: Exploratory Data Analysis. Sage Publications Inc, (1982)
24. Cleveland, W.S.: Visualizing Data. AT&T Bell Laboratories, Murray Hill, New Jersey, (1993)
25. Siegel, S., Castellan, N.J.Jr.: Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill Book Company, (1988)
26. Garson, G. D.: Logistic regression, from Statnotes: Topics in Multivariate Analysis. Retrieved 26 April 2010 http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm , (2010)
27. Brooks, C.: Introductory Econometrics for Finance. 2nd ed., Cambridge Univ. Press, (2008)

Turku
Centre *for*
Computer
Science

Joukahaisenkatu 3-5 B, 20520 Turku, Finland | www.tucs.fi

University of Turku

- Department of Information Technology
- Department of Mathematics

Åbo Akademi University

- Department of Information Technologies

Turku School of Economics

- Institute of Information Systems Sciences

ISBN 978-952-12-2537-6
ISSN 1239-1891