OBEGEF – Observatório de Economia e Gestão de Fraude

WORKING PAPERS

#51 Signaling Tax Evasion, Financial Ratios and Cluster Analysis >>

António Dias, Carlos Pinto, João Batista,
Maria Elisabete Neves

OBEGEF
Observatório de Economia
e Gestão de Fraude

>> FICHA TÉCNICA

SIGNALING TAX EVASION, FINANCIAL RATIOS AND CLUSTER ANALYSIS

WORKING PAPERS Nº 51 / 2016
OBEGEF – Observatório de Economia e Gestão de Fraude

António Dias[1], Carlos Pinto[2], João Batista[2], Maria Elisabete Neves[3]
Editor: Edições Húmus
1ª Edição: Fevereiro de 2015
ISBN: 978-989-755-190-1

Localização web: http://www.gestaodefraude.eu
Preço: gratuito na edição electrónica, acesso por download.
Solicitação ao leitor: Transmita-nos a sua opinião sobre este trabalho.

Paper in I2FC:2014 - Multiple Perspectives of the Shadow Economy

[1] Tax and Customs Authority. acgdias@gmail.com
[2] Tax and Customs Authority
[3] ISCAC-Coimbra Business School

>> ÍNDICE

>> RESUMO

A presente investigação pretende desenvolver uma metodologia de análise de dados que através da utilização de indicadores financeiros permita identificar práticas de relato financeiro fraudulento e eventual evasão fiscal.
Trata-se de um estudo de carácter exploratório em que são utilizadas técnicas estatísticas apropriadas à extracção e tratamento de dados, designadamente a Análise Cluster, com o objectivo de agrupar as observações em grupos homogéneos em função de características que indiciem incumprimento fiscal.
Foram utilizados os dados declarados por empresas que nos exercícios de 2010 e 2011 operaram no sector dos Minerais de Construção - Extracção e Fabricação.
Analisamos nove indicadores financeiros que consideram características relacionadas com a rentabilidade, liquidez e actividade, capazes de sinalizar eventual manipulação dos elementos contabilísticos e consequente relato financeiro fraudulento.
Os resultados validam a metodologia utilizada e apontam para a existência de capacidade classificatória em quatro das variáveis estudadas o que permite a construção de grupos homogéneos que podem ser utilizados para identificar as empresas devem ser alvo de auditoria fiscal.

>> ABSTRACT

This research intends to develop a methodology for analyzing data through the use of financial indicators to classify taxpayers according the risk of possible tax evasion.
We have used the data reported by companies in the years 2010 and 2011 operated in a specific sector, which is Construction Minerals - Extraction and Fabrication.
This study by using the Cluster Analysis intends to classify the observations into homogeneous groups according to their characteristics which suggest tax evasion. We have considered, according to recent literature, characteristics related with profitability, liquidity and activity to signal a possible manipulation of accounting data and subsequent fraudulent financial reporting.
The results point out the existence of classificatory capacity in four variables. Specifically, our evidence proposes the construction of homogeneous groups which can then be used to identify companies to be tax auditing.

JEL Classification: H26, C38, C81
Keywords: Data Mining, Tax Evasion, Cluster analysis, Financial Ratios, Fraudulent Financial Reporting

>> INTRODUCTION

This paper is built upon the predictions of the tax evasion issues and aims to develop a methodology capable of classifying the taxpayers as a function of the risk of possible tax evasion. This methodology involves financial ratios classified into three classes, profitability, liquidity and activity.

In Portugal, in essence, the taxation is based on information provided by taxpayers, competing to the tax administration to ensure that taxes are paid by the value and at the right time.

In this sense, Fiscal audits are often developed in order to determine whether the amounts declared by companies and established in accounting, correspond to which in fact should be presented. However, this type of procedure is complex, time-consuming and expensive and it is not possible to use in a generally way.

Against a backdrop of scarce resources it is essential that tax audits covering just on companies that present concrete evidence of tax evasion. Then, to find effective and efficient methods and models, able to select companies for tax audit it is, in fact, an interesting and necessary task for the tax authorities.

To achieve the proposed goal, we have used data of Portuguese companies operating in the Construction Minerals sector during the years 2010 and 2011 including data contained in the Balance Sheet and Income Statement. In fact, we propose a new empirical approach that allows us to measure the capacity of some indicators that may reveal a necessity to audit some firms with harbingers of fraud and evasion. Specifically, we use the cluster analysis to treat information gathered and draw conclusions that can be useful in the future. Consistent with financial statements arguments, our findings reveal that it is possible to classify the firms more prone to tax evasion.

Section I introduces a theoretical framework that takes account of the existing literature and the empirical evidence on traditional explanations of fraud and tax evasion by using financial statements. Section II describes the data and variables used in our analysis. In Section III, we present the research design, with the hypothesis proposed, our methodology and results. Finally, in section IV, we present the conclusion.

>> LITERATURE REVIEW

The Construction Minerals (extraction and fabrication) is an economically strategic and cross-sector implementation in Portugal. The sector comprises the extraction and processing of stone for industrial and ornamental purposes. The values of the results presented in the years in study, the exercise of these activities have generated significant losses which, in a preliminary approach, may indicate that it is a sector where tax evasion is a common practice.

In this section, we present the concept of tax evasion and some of the most important literature in this field.

In Portugal, the detection of behaviors which could indicate tax evasion, based on disclosures in the financial statements and tax returns that firms are obliged to perform is not an easy task.

Tax evasion is shown in various ways, concealment of income, handling of stocks and results, the broad interpretation of tax laws, the absence of tax results to established accounting adjustments and in addition to the existence of agents who are not formally registered and are responsible for the parallel or informal economy[1].

Currently the Tax and Customs Authority (AT) has proven quite effective in obtaining, organizing and storing data from statements that taxpayers systematically send, which has allowed the development of an extremely comprehensive database of easy handling. However, not always this huge amount of data is transformed into knowledge that can be useful to the development of its activities, including those related to the detection of fraud and tax evasion.

Therefore, it becomes essential to develop processes for data mining based on mathematical and statistical models that make use of databases in order to identify possible correlations and/or systematic relationships between variables that can be validated through the application of detected patterns to new subsets of data.

The processes of data mining are extremely useful in exploratory analysis of scenarios where there is no predetermined notion about the behavior of the data (Kantardzic, 2011). These techniques are generally divided into two groups, the techniques of descriptive character and predictive techniques. The first one, describe the data set in a concise manner and let know their

[1] See, Engström and Holmlund (2009) for a comprehensive study about tax evasion in a small country as Portugal. These authors used Swedish income and expenditure data to examine the extent of underreporting of income among self-employed individuals

main general properties while the predictive techniques analyze the data in order to build models and try to predict the behavior of new data (Gupta and Gill, 2012).

In this sense, this research, although exploratory in nature, presents a methodology that aims to contribute to the development of signaling mechanisms of possible tax evaders by analyzing indicators that were extracted for database available.

In the international literature, as opposed to the Portugal situation, there are a significant number of authors dedicated in studying the fraudulent financial reporting, associated or not with tax evasion, through the analysis of financial statements issued by companies (see, for example, Mohd, et al., 2010 or Zhu and Gao, 2011 for a recent researches).

Several studies have used various statistical techniques that consider univariate and multivariate criteria and the regression analysis and the use of artificial neural networks have been the most popular methods (Rao et al, 2009). There are also authors who combine, simultaneously, various techniques in an attempt to determine what the most appropriated to button data is (Kirkos et al, 2007). Recently, Dhiman, et al., (2013) have used the cluster analysis and method of a decision tree as the best approach to obtain more robust and efficient results under uncertainty and risky context. Kallio and Back, (2009) have combined cluster analysis with neural networks.

In this study, given its exploratory nature, we chose the use of Cluster Analysis once, and apart from detecting any dependence between variables, it enables to catalogue data into groups with similar characteristics.

It is a statistical technique that automatically and without user's intervention (unsupervised classification), classifies the observations into different groups according to their similarities (statistical distance) without the need to consider the properties and characteristics of the previously known groups or a causal relationship among the used the variables.

Moreover, the cluster analysis is a tool for data analysis which classifies the various observations in groups with similar characteristics (Dhiman et al, 2013), so that the objects belonging to the same group are similar among them and significantly different from the objects placed in another group. The greater the statistical distance between the centres of the groups the larger are the differences among the elements that constitute each particular group.

Although the companies try to hide the best they can possible practices that indicate tax evasion, the various investigations have identified variables, built from the information contained in the financial statements, as capable of identifying fraudulent financial reporting. The financial performance indicators, generally known as ratios, have been widely used in models of bankruptcy prediction (we highlight the pionners Beaver, 1966 and Altman, 1968). The explanatory variables are generally financial indexes that represent the index of profitability, solvency and liquidity (Salmi and Martikainen, 1994).

More recently, Gupta and Gill (2012) have classified organizations as fraudulent or not fraudulent by using indexes or financial ratios related to profitability, liquidity and operational efficiency.

Lisowsky (2010) examined the use of abusive tax planning by companies, having for end used tax data obtained from the North American tax authority, concluding that the likelihood of a company using abusive tax planning is positively related to inconsistencies between accounting and tax returns, profitability and company size.

Wilson (2009) reached virtually the same conclusions and according to this author the probability of a company using abusive tax planning is positively related to the firm size, large differences between tax accounting and taxation, and also aggressive practices used in financial reporting.

Dyreng et al. (2010), describe the tax evasion as the ability to sustain an effective tax rate below the statutory rate. These authors shows that companies in the oil and gas industry, mining, insurance and real estate can sustain an effective tax rate of less than 20 percent of the legal rate during a period of ten years to features like company size, use of tax havens or high levels of tangible and intangible assets.

Some researchers (see, among others, Persons, 1995; Vanasco, 1998 or Spathis et al. 2002) suggest that managers can manipulate inventories, and in this way, costs of goods sold may not match the effective value of sales.

The use of debt ratios is an open question in previous literature (Persons, 1995). There are authors, among others, Chow and Rice (1992), who argue that high debt levels increase the likelihood of fraudulent financial reporting.

However, there are others authors who argue the contrary. Graham and Alan (2006) found empirical evidence that companies which adopt tax evasion practices have lower levels of indebtedness. This conclusion is concordant with the prospect that higher debt levels will allow companies to have lower accounting and tax result and thus they are discouraged to adopt strategies of tax evasion.

>> DATA AND SAMPLE CONSTITUTION

In this section we refer the methodology used for the establishment of the sample, the source of the database and the procedures adopted for the treatment of this data.

To develop the present study we have used AT sensitive data, relating to economic exercises of 2010 and 2011, of companies engaged in the extraction of marble and other carbonated rocks, granite extraction and ornamental rocks like limestone and chalk, extraction, extraction of gravel, sand and crushed stone, manufacture of marble and similar rocks and manufacture of granite and rocks.

The extraction and selection of data is perhaps the most important and decisive in the success of a scientific work. For this reason, to the composition of the sample, we begin by collecting and systematizing the information contained in the financial statements that were relevant in the light of the variables to be build.

The previous analysis of database led us to the exclusion of companies that had shown evidences of not being in activity (sales and services equal to zero). Then, were also deleted the companies with incomplete data, that is, those that for insufficient data did not certificate the construction of all variables in the analysis.

With respect to companies without sign of activity, it seems to us important, that AT should try to understand what the real reasons justifying the maintenance of companies without activity. For this purpose, we suggest that could be sent an inquiry to these companies to determine whether in fact they're ineffective or, on the contrary, it derives from evasion behaviors.

By choice of the authors, were still excluded firms with turnovers of less than € 100,000 .00, since we believe that, given the sector of activity, below this limit only operate companies that are of small size, although in a significant number. If they have been considered, they could cause bias in the data.

Regarding the data and taking into account this industry, it seems hardly credible that these companies present so few levels of business, that is, as this is an activity that involves large investments in equipment and companies would not survive with such low activity rates. For this type of companies, we believe that a methodology should be developed autonomously based on surveys to verify that the level of activity is consistent with the declared values.

Additionally, were also eliminated companies with turnover greater than € 10.000.000,00, these, although few in number, were classified as atypical observations and given their size they must be subject to ongoing monitoring.

Once, although related, we are dealing with two distinct sectors, mining and manufacturing is natural and expected that the financial indicators are different for these types of companies. For this motive, in order to obtain a more robust classification, we have two different samples depending on the economic activity code declared by companies.

Table 1 indicates the number of companies that were eliminated in the light of the criteria mentioned in the previous points as well as those which remain in the sample. The extractives sector presents, in 280 observations in each year and the sample in manufacturing sector consists of 696 companies in 2010 and 631 companies in 2011.

Table 1: Construction of the sample

Construction of the sample	**2010**	**2011**
Number of companies (initial)	1783	1757
Exclusion criteria		
No activity (Revenue equal to zero)	127	122
Whit incomplete data	356	335
With turnover < 100.000€	316	381
With turnover > 10.000.000€	8	8
Final sample by sector of activity		
Extraction	280	280
Manufacturing	696	631
Total of companies	976	911

Although the minimum values as well as the maximum values, resulting from the restrictions placed on the construction phase of the sample, it is important to recognize the sample variation of main economic indicators by sector of activity (tables 2 and 3).

Table 2: Main indicators of activity-Extraction Sector 2010-2011

Indicators	**Minimum**	**Media**	**Mean**	**Maximum**	**Standard Dev.**
			2010		
Total Assets	98.929,57	1.362.779,57	2.837.210,30	27.338.164,00	3.913.239,84
Turnover	100.617,00	646.474,13	1.305.089,95	9.040.863,89	1.607.403,51
EBITDA	-2.301.249,30	86.318,40	162.539,96	1.760.955,42	354.350,47

Indicators	Minimum	Media	Mean	Maximum	Standard Dev.
Net Income	-2.580.772,23	4.827,63	-13.618,63	970.272,84	274.792,30
Tax Loss	0,00	0,00	50.271,61	2.247.157,06	196.585,90
Taxable Income	0,00	10.603,40	58.486,60	1.354.640,98	150.774,61
2011					
Total Assets	79.130,93	1.267.051,38	2.885.603,94	26.322.983,23	4.111.115,89
Turnover	104.959,40	680.921,59	1.289.975,36	9.275.040,87	1.589.197,87
EBITDA	-1.017.299,28	53.905,26	140.976,24	4.199.749,72	382.439,81
Net Income	-1.400.018,72	3.656,83	-28.552,25	2.569.804,78	277.350,89
Tax Loss	0,00	0,00	70.192,66	1.761.190,03	207.746,12
Taxable Income	0,00	7.055,94	44.086,29	782.148,72	109.443,78

Table 3: Main Activity Indicators – Manufacturing Sector 2010-2011

Indicators	Minimum	Media	Mean	Maximum	Standard Dev.
2010					
Total Assets	45.462,57	510.382,36	1.092.231,27	18.706.723,00	2.703.538,89
Turnover	100.302,38	269.265,32	627.745,22	9.842.990,43	1.528.911,17
EBITDA	-309.012,04	20.576,65	55.060,87	1.797.435,46	224.985,93
Net Income	-958.526,94	2.427,33	-7.139,09	909.819,81	130.668,32
Tax Loss	0,00	0,00	19.063,23	897.352,92	91.673,76
Taxable Income	0,00	3.981,96	18.631,93	1.280.006,59	103.332,72
2011					
Total Assets	29.167,86	535.438,26	1.142.905,17	18.301.159,91	1.981.716,65
Turnover	100.138,47	264.757,49	617.690,96	8.626.794,39	986.152,11
EBITDA	-864.754,71	17.588,51	47.806,10	2.100.454,73	168.721,06
Net Income	-1.133.833,02	1.119,46	-13.820,65	893.231,28	98.667,33
Tax Loss	0,00	0,00	23.554,44	1.130.040,53	76.286,79
Taxable Income	0,00	2.936,59	15.128,49	1.172.992,39	62.858,14

Once constituted the sample, in the next section, we present the starting hypotheses and define the variables to be used in our model.

>> RESEARCH DESIGN

1. Hypothesis

In accordance with literature review and with the purpose of creating a instrument to classify and identify the operators about the risk of tax evasion, this study uses the methodology known as Cluster Analysis and a set of recognized ratios obtained through the financial statements of the companies (Table 4). Therefore, our study uses variables to capture the behavior of indicators of profitability, liquidity and activity regarding the effects of taxation.

To achieve this aim, we have considered ratios identified in prior literature, namely: Differences between the Accounting and Tax Result (DCF), Return on assets (ROA), Debt (END), Current ratio (LG), Inventory Turnover (RE), Weight of the Earnings Before interest, Tax, Depreciation and Amortization (EBITDA) on Total Assets, and Cost of Goods Sold against turnover (CMVN)).

In addition, two other variables were tested, Tax Return (RF) and Weight of Personnel Expense on Turnover (GPVN). The first one is a ratio commonly used by AT and the second one is a ratio for testing the operational efficiency based on labor, as sectors that use intensive labor.

Table 4: Constitution of the variables

Profitability indicators	
EBITDA	EBITDA / Total assets
ROA	Net Income / Total assets
RF	Tax Resul / Turnover (VN)
Fiscal Indicators	
DFC	[(Net Income)-(Tax Result)]/Turnover
Activity indicators	
RE	(Cost of goods sold and consumed materials) / [(beginning inventory+ ending inventory)/2]
GPVN	personnel expense / turnover
CMVN	Cost of goods sold and consumed materials /turnover
Liquidity indicators	
LG	Current assets/ Current Liabilities
END	Non-current liabilities /Total assets

Similar to others authors (see, among others, Fanning and Cogger, 1998; Altman and Sabato, 2007 and Gupta and Gill, 2012) we have used a set of indicators to measure the behavior of variables of profitability associated with the financial performance of companies.

In general we understand that companies committed to tax evasion practices tend to have low profits. This assertion is aligned with the behaviors of the indicators of profitability, where it is expected that fugitive tax practices generate lower return on assets, aiming at the reduction of taxes payable. Niskanen and Keloharju (2000) shown that in Finland there are companies that, to avoid taxes, declare low levels of income.

It will be expected that companies which present low levels of EBITDA and ROA (in proportion to the total assets), as a result of the use of under billing or overvaluation of their production costs. It will be expected that companies with low levels of EBITDA and ROA (in proportion to the total assets), as a result of the use of under-invoicing or overvaluation of their production costs, are much more prone to tax evasion.

In addition, we chose to test the variable tax profitability that confronts the taxable income (tax result and no accounting result) with the turnover.

As occurs in previous indicators, it is also expected that low profitability is bound to tax evasion practices. Therefore, in accordance with these points of view, we put the following hypotheses:

Hip.1. Companies with low level of EBITDA present a higher risk of tax evasion.
Hip.2. Companies with low level of ROA present a higher risk of tax evasion.
Hip.3. Companies with low level of RF present a higher risk of tax evasion.

The International literature (Wilson, 2009 or Lisowsky, 2010) point out that the difference between the accounting result and the tax result as an explanatory variable on detection of this phenomenon of tax evasion essentially associated with large companies, which have sufficient resources to adopt practices of abusive tax planning.

Although it go test this hypothesis, given that the Portuguese economy entrepreneurial consist mainly of micro, small and medium enterprises, we do not expect the difference between accounting income and taxable income is determinant in explaining the phenomenon, since this type of operators accounting is mainly intended to meet tax obligations, that is, accounting of the vast majority of our business is prepared to “trailer” of taxation.

Hip.4. Companies with large difference between the accounting result and the tax result have a higher risk of tax evasion**.**

Also the handling of inventories is considered by several authors as a variable associated with tax evasion mechanisms (Persons, 1995, or Kirkos et al., 2005).

Therefore, the inventory turnover may be an indicator with some capacity to identify potential evaders, since the manipulation of stocks can be directly linked with the omission of sales as to purchases.

Hip.5. Companies with low inventory turnover present a higher risk of tax evasion.

Other variable considered in this research, aims to relate the weight of personal expenditures and tax evasion. As we already said this is a sector with intensive labor and this factor is the main generator of added value.

Simultaneously, although they admit the possibility of not being accounted for the totality of spending on personnel, is also to consider the existence minimum monthly wage levels, which inhibit such manipulation of expenses beyond these limits. Accordingly, it is our expectation that companies which have a significant weight of personal expenditures vis-à-vis the turnover may be omitting income.

Hip.6. Companies with a high share of personnel expenses in turnover present higher risk of tax evasion.

The inventory manipulation and consequently the cost of goods sold and consumed, matter arises often in investigations of this topic (Fanning and Cogger, 1998 or Schilit, 2002). The initial expectation regarding the behavior of the weight of the cost of goods sold and materials consumed on the turnover is that the higher this variable the greater the likelihood of tax evasion.

This position is based on the consideration that the more reduced the margin of business, the lower shall be the taxable income.

However, during visits to six companies in the District of Vila Real, in the North of the country, engaged in the extraction of minerals, we concluded that in addition to a handicapped valorization of production, are not accounted for by-products and waste, which leads to an underestimation of inventories. Note that it is extremely difficult to determine the extractive capacity deployed in companies which can lead to the omission of the actual production and thus generate products for the parallel market. Moreover, not being revealed in the accounts, the production of goods, will be easier the omission of the invoicing.

Thus, in a second moment, the expectation on the behavior of this indicator is that firms in which the ratio of cost of good sold and materials consumed on turnover is low, have a higher risk of tax evasion.

Hip.7. Companies with low levels of costs of goods sold on the turnover have higher risk of tax evasion.

Concerning the liquidity analysis and once this variable will tend to analyze the ability of companies to honour its financial commitments, it will be our theory that possible tax evasion practices are related to lower liquidity values. That is, the smaller the value displayed in this indicator the greater the probability of being in the practice of evasive behaviors.

This happens, once a company that under billing, this is not on sales accounting and therefore also not reflected in the account customers/liquid funds making lower your current assets.

Hip.8. Companies with low levels of liquidity have a higher risk of tax evasion.

Finally, we use the level of indebtedness, as explanatory variable of the phenomenon of tax evasion (Wilson, 2009; Lisowsky, 2010). For example, Graham and Alan (2006) found evidence that firms adopting tax evasion have reduced debt levels.

This conclusion is concordant with the prospect that higher debt levels will allow companies to introduce smaller accounting and fiscal results and thus be discouraged to adopt strategies for tax evasion. Apart from this reason, purely financial, the Portuguese realit, still competes another factor, no less important, which is related to the preparation of financial statements and tax and its recipients, and in Portugal the main users of these statements are management tax and Financial Institutions.

Consequently, companies that have the need of outside capital will have greater need to develop statements that express a true and fair view of their financial and economic situation, resulting, in theory, that firms with higher debt levels will present demonstrations of real closer income derived.

Hip.9. Companies with low debt levels have a higher risk of tax evasion.

We intend, through these nine indicators, classify businesses in homogeneous groups according to their potential risk of tax evasion, using a classification model based on K-means technique analysis (IBM SPSS Statistics 19), approach that we will do in the next section.

2. Methodology

After construction of the variables it is necessary to study their behavior in order to identify any observations that, given the values they take, configure atypical situations that may demonstrate exceptional situations, declarative errors or possible deficiencies in the construction of the database

In this sense, we analyze the behavior of indicators constructed using the mean, median, maximum and minimum, standard deviation, skewness and curtoses, which allowed to detect the existence of extreme values. These observations may have great influence on the estimation of coefficients, which can affect decisively the quality of the model and, since it is a cluster analysis, significantly change the center of the groups that may be formed and consequently the classification of companies.

For example, in table 5, we can see the distribution of variables relating to the manufacturing sector in 2011, particularly the case of the current ratio (equal to 2373.96 maximum and average 6,042) and inventories turnover see if (maximum of 3389.81 and an average of 10.78), which have extreme values can influence the average of the indicators

Table 5: Distribution of the variables – Manufacturing 2011

Variable	**Minimum**	**Perc. 2**	**Median**	**Mean**	**Perc. 98**	Maxim
DFC	-0,2246	-0,0814	-0,0055	-0,0107	0,0185	0,1198
ROA	-1,1995	-0,4357	0,0021	-0,0393	0,0839	0,3409
RF	-1,6028	-0,6673	0,0094	-0,0624	0,1061	0,6271
End	0,0000	0,0000	0,1846	0,2623	1,0128	2,3534
LG	0,0461	0,3707	1,4490	6,042	9,3765	2373,96
RE	0,0102	0,1052	1,7174	10,78	39,696	3389,81
EBITDA	-1,1126	-0,3746	0,0390	0,0137	0,1708	0,5009
CMVN	0,0019	0,0385	0,3920	0,3907	0,7822	1,3932
GPVN	0,0000	0,0743	0,3434	0,3594	0,7675	1,2376

To resolve the issue, the extreme values were treated with the technique proposed by Löeffler and Posch (2011), the designated soundproofing variables. This method makes it possible to largely eliminate the extreme values without, however, reducing the number of companies. To this end we decided to completely soundproof the variables in 2%, i.e. replace the extreme values on the left side of the distribution by 2 percentile value and the extreme values on the right side of the distribution by the 98 Percentile. As can be seen in table 6, essentially in the variables where there was greater magnitude, the soundproofing allowed the easing of the burden of extreme values (outliers) with consequent change in average values. In the variables like, differences between the Accounting and Tax Result, Return on assets, indebtedness and EBITDA, the mentioned method allowed the Elimination of almost all "outliers".

The indicators corresponding to the Indebtedness, CMVN and GPVN feature a sampling distribution very close to a Normal distribution.

Then, since there are other variables in the data range to eliminate this effect and increase the accuracy of the model, we used log-transformed variables, according to the technique used by Rijken and Altman (2004). This technique allows more robust clusters since its center is not influenced by the data width.

Table 6: Distribution of the variables – 2011-Manufacturing After soundproofing (2%)

Variável	1º Q.	Mediana	Média	3º Q.	D. Padrão	Curtose	Assimetria
DFC	-0,0135	-0,0055	-0,0107	-0,0015	0,0171	6,5672	-2,3394
ROA	-0,0505	0,0021	-0,0366	0,0105	0,1000	5,3743	-2,2834
RF	-0,0824	0,0094	-0,0591	0,0258	0,1636	4,2220	-2,1275
End	0,0298	0,1846	0,2555	0,4026	0,2596	0,5748	1,1013
LG	1,0121	1,4490	2,1570	2,4972	1,8841	4,5248	2,1359
RE	0,7218	1,7174	4,0189	3,9453	6,8617	14,6151	3,6311
EBITDA	-0,0058	0,0390	0,0158	0,0701	0,1024	3,9413	-1,7793
CMVN	0,2555	0,3920	0,3884	0,5113	0,1779	-0,4774	0,1232
GPVN	0,2425	0,3434	0,3569	0,4574	0,1586	-0,1697	0,4507

The indicators corresponding to the Indebtedness, CMVN and GPVN feature a sampling distribution very close to a Normal distribution.

Then, since there are other variables in the data range to eliminate this effect and increase the accuracy of the model, we used log-transformed variables, according to the technique used by Rijken and Altman (2004). This technique allows more robust clusters since its center is not influenced by the data width.

To prepare the data to cluster analysis, we begin by examining the correlation between variables (see correlation tables) which allowed us to verify, as expected, that the variables of profitability were strongly correlated. For this reason were excluded variables corresponding to the return on assets and Tax Profitability, remaining in the model the corresponding variable to EBITDA because we concluded that this allows us to discriminate more effectively the behavior of observations.

As segmentation technique we have used the KMeans tool Analysis, with the number of groups attached to the priori. In the construction of the clusters we start with some preparatory building approaches for each year and various activity models with 6, 5, 4 and 3 clusters. Since the aim of this work is to sort/flag companies in homogeneous groups that allow us to highlight practices of tax evasion; we concluded that the ideal was to use only three groups.

This solution enables us to identify and group more effectively the companies in the various risk groups, once the groups are more heterogeneous among themselves.

Then we shall proceed to the analysis of the discriminatory capacity of each one of the variables. This step dictated the Elimination of variable corresponding to the difference between the Accounting and tax Result for not being in any of the sectors analyzed, statistically significant.

3. Results

As mentioned in the previous section, using the K-Means technique Analysis, based on six variables proposals, we have segmented the data available in order to obtain three homogeneous groups of observations for each sector of activity and economic exercise. This analysis resulted in the clusters which are defined in tables 7 and 8. Taking into account the behavior of previous hypotheses that we put, we named the clusters obtained as high risk, moderate risk, and Low Risk.

The first conclusion is that, as would be expected, the centers of clusters vary markedly between sectors of activity and maintain closely when evaluating centers its behavior over time.

In any of the sectors of activity, the largest number of companies is classified as group that features strong risks of tax evasion. With reference to the extraction sector, for the year of 2010, 156 companies were classified as belonging to the group that presents higher risk of tax evasion. In 2011, were 174 companies that belonging to this group.

The manufacturing sector follows the same trend and, in 2010, 371 companies belonging to the same group, value that passes the 270 in 2011. In opposition, with the exception of the manufacturing sector in 2010, the group with lower risk is one that presents fewer companies.

Concerning the Extraction sector (table 7), despite registering a shift for the Group at greatest risk, it turns out that the temporal analysis shows no significant changes in the location of each of the variables, with the exception of inventories turnover indicator and weight of EBITDA, always following the same trend among the distinct groups (or increase or decrease in all groups).

Table 7: Cluster Center-Extraction Sector

year	**2010**			**2011**		
Indicator	**High**	**Moderate**	**Low**	**High**	**Moderate**	**Low**
END	0,1686	0,2316	0,1397	0,1832	0,2551	0,1788
LG	0,7141	1,6675	0,9485	0,7935	1,8059	1,0117
RE	0,4982	0,5759	2,6560	0,5647	0,3817	2,8428
EBITDA	0,0694	0,0676	0,0923	0,0401	0,0709	0,1007
CMVN	0,1467	0,1425	0,2938	0,1628	0,1576	0,2907
GPVN	0,2737	0,2874	0,1820	0,2718	0,2741	0,1666
Number companies	156	68	56	174	57	49
% companies	55,71	24,29	20,00	62.14	20.36	17,50

For its part, in the manufacturing sector (table 8) we conclude that in 2011, unlike 2010, indebtedness and current ratio variables do not have values very distant between the Moderate and high risk groups. The same is not true in the ratio of inventories turnover which denotes a lower rotation, compared to 2010, in the high risk group and greater rotation in the Moderate risk group. However, with respect to the group having Low risk of tax evasion the location of the center of the group does not vary significantly over time.

It should be noted that in 2011 there is a rapprochement between the centers of the three groups, a situation largely justified by changes in the variables referred to in the previous point.

Thus, these contribute significantly to changing the centroid of each group and determine the visible movement of companies of High and low risk groups for the group that has moderate risk.

Table 8: Cluster Center-Manufacturing Sector

year	**2010**			**2011**		
Indicator	**High**	**Moderate**	**Low**	**High**	**Moderate**	**Low**
END	0,1559	0,3085	0,1417	0,2190	0,2163	0,1544
LG	0,7748	1,6471	0,9272	1,1029	0,9630	0,9819
RE	0,7745	0,9487	2,2636	0,4911	1,3240	2,7154
EBITDA	0,0273	0,0381	0,3618	0,2852	0,3364	0,3729
CMVN	0,3086	0,3108	0,3618	0,2852	0,3364	0,3329
GPVN	0,3148	0,3062	0,2310	0,3306	0,2940	0,2189

year	2010			2011		
Indicator	**High**	**Moderate**	**Low**	**High**	**Moderate**	**Low**
Number companies	371	148	177	270	267	94
% companies	53,30	21.26	25,43	42,79	42,31	14,90

To facilitate the reading of the values assigned to each variable in the definition of the location of the center of multiple clusters, we calculated the average amount that each indicator assumes when considered the observations of each of the clusters (tables 9 and 10). As an example we can mention that extractive industry companies that were classified in the highest risk group of tax evasion have on average in fiscal year 2010 indebtedness of 20% and weight of personal spending on turnover of 32.5%.

Table 9: Average values of the variables by risk group-Extraction Sector

year	2010			2011		
Indicator	**High**	**Moderate**	**Low**	**High**	**Moderate**	**Low**
END	0,2015	0,3042	0,1603	0,2201	0,3262	0,2145
LG	1,0857	5,4793	1,7153	1,283	6,0278	1,9992
RE	0,8121	1,0776	24,8959	0,9546	‘,599	31,9055
EBITDA	0,0592	0,064	0,0855	0,0202	0,0704	0,099
CMVN	0,1699	0,1626	0,366	0,1889	0,1881	0,3594
GPVN	0,3258	0,3475	0,2039	0,3293	0,3259	0,1855

Table 10: Average values of the variables by risk group-Manufacturing Sector

year	2010			2011		
Indicator	**High**	**Moderate**	**Low**	**High**	**Moderate**	**Low**
END	0,1889	0,3941	0,1752	0,2756	0,2761	0,185
LG	1,2211	5,3262	1,7982	2,7295	1,9646	1,9016
RE	1,3321	1,87	16,5173	0,6781	2,9501	25,9538
EBITDA	0,018	0,0318	0,068	0,0118	0,0122	0,0674
CMVN	0,3749	0,3717	0,4467	0,3429	0,4139	0,4621
GPVN	0,3816	0,3685	0,2676	0,4058	0,3512	0,2493

Then, it was necessary to conclude if groups formed by cluster analysis have indeed predictive capacity and discrimination ability between companies relating to the potential risk of tax evasion.

For this purpose we collect information relating to fiscal audits carried out by AT in 2009 to 2011 in companies belonging to sectors of activity under study. Subsequently, we classified companies, where there were significant adjustments in the light of the criteria determined by cluster analysis.

This analysis (table 11) has shown that 64% of companies would be classified in the highest risk group of tax evasion, 21% in the group with moderate risk of tax evasion and just 15% in the low-risk group of tax evasion. Therefore, it is concluded that this methodology can be used to flag potential evaders groups once the percentage of audited companies classified in the high-risk group is in fact high.

Table 11: Validation – classification of Audited Companies 2009-2011

Year	High	Moderate	Low
2009	21	6	5
2010	11	2	3
2011	2	3	0
Classification by groups	34	11	8
Classification by groups in %	64,15	20,75	15,10

As we already said the variables corresponding to the return on assets and Tax Profitability were excluded from the model because they are strongly correlated with EBITDA. For this reason, although we cannot accept or reject hypotheses 2 and 3, given the correlation between variables, it is expected that the behaviour of these variables are similar to the behavior of the EBITDA.

In turn, for any of the years and sector of activity, we did not obtain evidence to confirm the hypothesis 4. The difference between the Accounting Result and the Tax result declared by the companies features, among the distinct groups, average values very similar. So, we can't confirm that companies with large differences between the Accounting Result and the Tax result have a higher risk of Tax evasion.

Relating to the ratio of EBITDA in total assets is a variable that takes on the expected behavior, that is, companies with lower EBITDA are those that have a higher risk of tax evasion and we can confirm the hypothesis 1.

The variables corresponding to the activity ratios behave according to the hypotheses formulated. Companies with greater inventories turnover are those that present a lower risk of tax evasion. In this approach, we can confirm the hypothesis 5.

The variable corresponding to the weight of personnel costs to turnover presents a linear behavior being the higher the risk the higher the value of this variable, in line with our hypothesis 6.

With regard to the variable that reflects the weight of the cost of goods sold on the turnover behave opposite to the previous variable, i.e. companies that present higher values are those that show lower risk of tax evasion, which corroborates the hypothesis 7.

With regard to liquidity, the behavior assumed by the proposed variable does not allow to validate the hypothesis 8. Not always companies that have low liquidity were classified in the highest risk group of tax evasion.

The companies classified as belonging to the Group of moderate risk are those that exhibit higher levels of indebtedness, being this similar indicator in the companies that find themselves in groups of high risk and low risk. Thus, we haven't confirmed that companies that exhibit lower levels of indebtedness are evasive tax behaviour, and consequently we can rejects the hypothesis 9.

In short, in view of the above mentioned, we can see that the variables of profitability and activity, have the ability to sort/flag companies that present a higher risk of tax evasion.

>> CONCLUSION

Our research, although essentially exploratory, is unheard of in the panorama of scientific research, in particular in the context of taxation in Portugal, a small country in eurozone with the effects of a financial and foreign debt crisis.

In this work, has been tested and validated a statistical technique to extract data and dividing the population into homogeneous groups according to the level of risk for tax evasion.

We highlight the exposure of four variables with significant classificatory capacity to detect a possible tax evasion, by using the cluster analysis.

As our main contribution to the literature of this matter, we highlight the detection of four variables with significant classificatory capacity to predict tax evasion, the use of cluster analysis, classification of companies into homogenous groups face the risk of tax evasion, signaling the companies with the greatest potential for success in activities audit and validation of results that indicate a high percentage of hits.

Although the methodology used have suggested the existence of three groups of companies, the fact that an undertaking is not in the group that suggests higher risk of tax evasion is no guarantee that this will not occur. Actually, it is possible the existence of evaders in all the groups.

A possible interpretation is that the likelihood of tax evasion is not as high as we thought, although the companies can present evasive behavior via non-measurable aspects in the variables in question such as, for example, the existence of subsidiaries or the manipulation of transfer prices.

We can identify, as our major limitation, the non-inclusion of institutional characteristics and corporate governance factors, as well as type of business and organizational management.

As a suggestion for future studies we propose the compilation of a series of variables, through an international database, which through a methodology accepted and properly tested can detect in simplified form and automatically the companies- targets for the tax audit.

References

Altman, E. I. and Rijken H. A. (2004), "How Rating Agencies Achieve Rating Stability", Journal of Banking and Finance, No. 28.

Altman, E.I. (1968), "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy ", Journal of Finance, Vol. 23, Nº 4, pp. 589-609.

Beaver, W.H. (1966), "Financial Ratios as Predictors of Failures", Journal of Accounting Research, Vol. 5, pp. 71-102.

Chow, C. and Rice, S. (1982) 'Qualified audit opinions and share prices – an investigation', Auditing: A Journal of Practice and Theory, 1: 35–53.

Dhiman R., Vashisht S. and Sharma K. (2013), A Cluster Analysis and Decision Tree Hybrid Approach in Data Mining to Describing Tax Audit, International Journal of Computers & Technology, Vol. 4, No. 1.

Dyreng, S., M. Hanlon e E.L. Maydew (2010), "The Effects of Executives on Corporate Tax avoidance" The Accounting Review, 85 (4), pp 1163-1189 : doi: 10.2308/accr.2010.85.4.1163.

Engström, P. and Holmlund, B. (2009). Tax evasion and self-employment in a high-tax country: evidence from Sweden, Applied Economics (41) 2419-2430

Fanning K. and Cogger K., (1998), 'Neural Network Detection of Management Fraud Using Published Financial Data', International Journal of Intelligent Systems in Accounting, Finance & Management, Vol. 7, No. 1, pp. 21-24.

Graham, J. and Alan T., (2006), Tax Shelters and corporate debt policy. Journal of Financial Economics 81, pp.563-594. DOI: 10.1016/j.jfineco.2005.09.002

Gupta, R. and Gill, N. (2012), Data Mining Techniques – A Key for detection of financial statement fraud, International Journal of Computer Science and Information Security, Vol. 10, No. 3, pp. 49 – 57.

Kallio, K and Back, B (2009), the self-organizing map in selecting companies for tax audit. In: Proc. of the 32nd Annual Congress of the European Accounting Association, Tampere (2009).

Kantardzic M. (2011), Data Mining: Concepts, Models, Methods, and Algorithms, Wiley – IEEE Press.

Kirkos E., Spathis C. and Manolopoulos Y. (2007), Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications 32 (23) pp. 995–1003.

Kirkos E., Spathis C. and Manolopoulos Y. (2005), Detection of Fraudulent Financial Statements through the use of Data Mining Techniques, 2nd

International Conference on Enterprise Systems and Accounting 2005 July 11-12, Greece.

Lisowsky, Petro (2010), "Seeking Shelter: Empirically Modeling Tax Shelters Using Financial Statement Information", The Accounting Review, Vol. 85, (5), pp. 1693-1721.

Loeffler, G, e Posch, P.N. (2011),"Credit Risk Modeling using Excel and VBA", Wiley Finance

Mohd, J., Ahmad, N. and Saled, N.M. (2010). Fraudulent financial reporting and company characteristics: tax audit evidence, Journal of Financial Reporting and Accounting, 8 (2), pp.128 - 142

Niskanen, J., Keloharju, M. (2000), Earnings cosmetics in a tax-driven accounting environment: Evidence from Finnish public firms. The European Accounting Review 9(3), 443–452.

Persons O. (1995), Using financial statement data to identify factors associated with fraudulent financing reporting, Journal of Applied Business Research, vol. 11, pp. 38-46.

Rao G.S. , Apparao G., Singh A., Bhavani B. L., Eswar K., Rajaniet D. (2009), Financial Statement Fraud Detection by Data Mining, Int. J. of Advanced Networking and Applications, Vol. 1, pp. 159-163 (2009).

Schilit H., (2002), 'Financial Shenanigans: How to Detect Accounting Gimmicks and Fraud in Financial Reports', McGraw-Hill, New York, USA.

Spathis C., Doumpos M. and Zopounidis C. (2002), "Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariates statistical techniques," European Accounting Review, Vol.11, pp. 506-535.

Spathis C. (2002), "Detecting false financial statements using published, data: some evidence from Greece," Managerial Auditing Journal, vol. 17, no. 4, pp. 179-191.

Vanasco, R. R. (1998), Fraud auditing, Managerial Auditing Journal, 13(1), pp. 4–71.

Wilson, R. 2009. An examination of corporate tax shelter participants. The Accounting Review, Vol. 84, pp. 969–999.

Zhu, J. and Gao, S. (2011). Fraudulent Financial Reporting: Corporate Behavior of Chinese Listed Companies. Accounting in Asia.

STATISTICAL DATA

Tabela 12 - Correlations: extraction sector 2010

		LnDCFI	LnROAI	LnRFI	LnENDI	LnLGI	LnREI	LnPETAI	LnCMVN	LnGPVNI
LnDCFI	Pearson Correlation	1	0,081	-0,039	0,091	-0,034	0,041	0,025	-0,053	0,031
	Sig. (2-tailed)		0,176	0,521	0,130	0,574	0,491	0,673	0,377	0,606
	N	280	280	280	280	280	280	280	280	280
LnROAI	Pearson Correlation	0,081	1	0,865**	-0,057	0,184**	0,172**	0,807**	-0,057	-0,269**
	Sig. (2-tailed)	0,176		0,000	0,339	0,002	0,004	0,000	0,343	0,000
	N	280	280	280	280	280	280	280	280	280
LnRFI	Pearson Correlation	-0,039	0,865**	1	-0,08	0,181**	0,147*	0,709**	-0,097	-0,309**
	Sig. (2-tailed)	0,521	0,000		0,181	0,002	0,014	0,000	0,107	0,000
	N	280	280	280	280	280	280	280	280	280
LnENDI	Pearson Correlation	0,091	-0,057	-0,080	1	0,185**	-0,127*	-0,027	-0,124*	-0,045
	Sig. (2-tailed)	0,130	0,339	0,181		0,002	0,033	0,650	0,038	0,449
	N	280	280	280	280	280	280	280	280	280
LnLGI	Pearson Correlation	-0,034	0,184**	0,181**	0,185**	1	0,016	0,026	-0,037	0,032
	Sig. (2-tailed)	0,574	0,002	0,002	0,002		0,790	0,667	0,542	0,589
	N	280	280	280	280	280	280	280	280	280
LnREI	Pearson Correlation	0,041	0,172**	0,147*	-0,127*	0,016	1	0,153*	0,460**	-0,343**
	Sig. (2-tailed)	0,491	0,004	0,014	0,033	0,790		0,010	0,000	0,000
	N	280	280	280	280	280	280	280	280	280
LnPETAI	Pearson Correlation	0,025	0,807**	0,709**	-0,027	0,026	0,153*	1	-0,194**	-0,279**
	Sig. (2-tailed)	0,673	0,000	0,000	0,650	0,667	0,010		0,001	0,000
	N	280	280	280	280	280	280	280	280	280
LnCMVN	Pearson Correlation	-0,053	-0,057	-0,097	-0,124*	-0,037	0,460**	-0,194**	1	-0,378**
	Sig. (2-tailed)	0,377	0,343	0,107	0,038	0,542	0,000	0,001		0,000
	N	280	280	280	280	280	280	280	280	280
LnGPVNI	Pearson Correlation	0,031	-0,269**	-0,309	-0,045	0,032	0,343**	-0,279**	-0,378**	1
	Sig. (2-tailed)	0,606	0,000	0,000	0,449	0,589	0,000	0,000	0,000	
	N	280	280	280	280	280	280	280	280	280

** Correlation is significant at the 0.01 level (2-tailed); * Correlation is significant at the 0.05 level (2-tailed)

Tabela 13 - Correlations: extraction sector 2011

		LnDCFI	LnROAI	LnRFI	LnENDI	LnLGI	LnPETAI	LnCMVN	LnGPVNI
LnDCFI	Pearson Correlation	1	-0,082	-0,325**	0,000	-0,072	-0,174**	0,056	0,045
	Sig. (2-tailed)		0,17	0,000	0,994	0,229	0,003	0,354	0,449
	N	280	280	280	280	280	280	280	280
LnROAI	Pearson Correlation	-0,082	1	0,803**	-0,152**	0,259**	0,855**	0,041	-0,402**
	Sig. (2-tailed)	0,17		0,000	0,011	0,000	0,000	0,497	0,000
	N	280	280	280	280	280	280	280	280
LnRFI	Pearson Correlation	-0,325**	0,803**	1	-0,55	0,217**	0,731**	0,012	-0,403**
	Sig. (2-tailed)	0,000	0,000		0,360	0,000	0,000	0,837	0,000
	N	280	280	280	280	280	280	280	280

LnENDI	Pearson Correlation	0,000	-0,152	-0,055	1	0,163**	-0,092	-0,065	0,040
	Sig. (2-tailed)	0,994	0,011	0,360		0,006	0,124	0,280	0,501
	N	280	280	280	280	280	280	280	280
LnLGI	Pearson Correlation	-0,072	0,259**	0,217**	0,163**	1	0,203**	-0,020	-0,013
	Sig. (2-tailed)	0,229	0,000	0,000	0,006		0,001	0,733	0,826
	N	280	280	280	280	280	280	280	280
LnPETAI	Pearson Correlation	-0,174**	0,855**	0,731**	-0,092	0,203**	1	-0,084	-0,394**
	Sig. (2-tailed)	0,003	0,000	0,000	0,124	0,001		0,161	0,000
	N	280	280	280	280	280	280	280	280
LnCMVNI	Pearson Correlation	0,056	0,041	0,012	-0,065	-0,02	-0,084	1	-0,316**
	Sig. (2-tailed)	0,354	0,497	0,837	0,28	0,733	0,161		0,000
	N	280	280	280	280	280	280	280	280
LnGPVNI	Pearson Correlation	0,045	-0,402**	-0,403**	0,040	-0,013	-0,394**	-0,316**	1
	Sig. (2-tailed)	0,449	0,000	0,000	0,501	0,826	0,000	0,000	
	N	280	280	280	280	280	280	280	280

** Correlation is significant at the 0.01 level (2-tailed); * Correlation is significant at the 0.05 level (2-tailed)

Tabela 14 - Correlations: manufacturing industry 2010

		LnDCFI	LnROAI	LnRFI	LnENDI	LnLGI	LnREI	LnPETAI	LnCMVN	LnGPVNI
LnDCFI	Pearson Correlation	1	0,097*	0,007	0,070	0,058	-0,022	0,056	-0,002	0,042
	Sig. (2-tailed)		0,011	0,854	0,066	0,128	0,566	0,142	0,966	0,263
	N	696	696	696	696	696	696	696	696	696
LnROAI	Pearson Correlation	0,097*	1	0,897**	-0,140**	0,163**	0,157**	0,915**	-0,152**	-0,371**
	Sig. (2-tailed)	0,011		0,000	0,000	0,000	0,000	0,000	0,343	0,000
	N	696	696	696	696	696	696	696	696	696
LnRFI	Pearson Correlation	0,007	0,897**	1	-0,131**	0,152**	0,170**	0,822**	-0,149**	-0,403**
	Sig. (2-tailed)	0,854	0,000		0,001	0,000	0,000	0,000	0,000	0,000
	N	696	696	696	696	696	696	696	696	696
LnENDI	Pearson Correlation	0,070	-0,140**	-0,131**	1	0,323**	-0,104**	-0,083	-0,087*	0,097*
	Sig. (2-tailed)	0,066	0,000	0,001		0,000	0,006	0,029	0,022	0,011
	N	696	696	696	696	696	696	696	696	696
LnLGI	Pearson Correlation	0,058	0,163**	0,152**	0,323**	1	-0,031	0,097*	-0,016	-0,024
	Sig. (2-tailed)	0,128	0,000	0,000	0,000		0,413	0,010	0,679	0,532
	N	696	696	696	696	696	696	696	696	696
LnREI	Pearson Correlation	-0,022	0,157**	0,170**	-0,104*	-0,031	1	0,240**	0,264**	-0,403**
	Sig. (2-tailed)	0,566	0,000	0,000	0,006	0,413		0,000	0,000	0,000
	N	696	696	696	696	696	696	696	696	696
LnPETAI	Pearson Correlation	0,056	0,915**	0,822**	-0,083*	0,097*	0,240**	1	-0,216**	-0,419**
	Sig. (2-tailed)	0,142	0,000	0,000	0,029	0,010	0,000		0,000	0,000
	N	696	696	696	696	696	696	696	696	696
LnCMVNI	Pearson Correlation	-0,002	-0,152**	-0,149**	-0,087*	-0,016	0,264**	-0,216**	1	-0,427**
	Sig. (2-tailed)	0,966	0,000	0,000	0,022	0,679	0,000	0,000		0,000
	N	696	696	696	696	696	696	696	696	696
LnGPVNI	Pearson Correlation	0,042	-0,371**	-0,403**	0,097*	0,024	-0,403**	-0,419**	-0,427**	1
	Sig. (2-tailed)	0,263	0,000	0,000	0,011	0,532	0,000	0,000	0,000	
	N	696	696	696	696	696	696	696	696	696

** Correlation is significant at the 0.01 level (2-tailed); * Correlation is significant at the 0.05 level (2-tailed)

Tabela 15 - Correlations: manufacturing industry 2011

		LnDCFI	LnROAI	LnRFI	LnENDI	LnLGI	LnREI	LnPETAI	LnCMVN	LnGPVNI
LnDCFI	Pearson Correlation	1	0,223**	0,044	0,057	0,053	-0,067	-0,184**	0,009	0,030
	Sig. (2-tailed)		0,000	0,273	0,153	0,183	0,093	0,000	0,826	0,445
	N	631	631	631	631	631	631	631	631	631
LnROAI	Pearson Correlation	0,223*	1	0,862**	-0,216**	0,222**	0,079**	0,937**	-0,180**	-0,414**
	Sig. (2-tailed)	0,000		0,000	0,000	0,000	0,048	0,000	0,000	0,000
	N	631	631	631	631	631	631	631	631	631
LnRFI	Pearson Correlation	0,044	0,862**	1	-0,213**	0,183**	0,106**	0,828**	-0,194**	-0,447**
	Sig. (2-tailed)	0,273	0,000		0,000	0,000	0,007	0,000	0,000	0,000
	N	631	631	631	631	631	631	631	631	631
LnENDI	Pearson Correlation	0,057	-0,216**	-0,213**	1	0,242**	-0,098*	-0,162**	-0,044	0,123**
	Sig. (2-tailed)	0,153	0,000	0,000		0,000	0,014	0,000	0,269	0,002
	N	631	631	631	631	631	631	631	631	631
LnLGI	Pearson Correlation	0,053	0,222**	0,183**	0,242**	1	-0,057	0,149**	-0,005	-0,040
	Sig. (2-tailed)	0,183	0,000	0,000	0,000		0,152	0,000	0,892	0,319
	N	631	631	631	631	631	631	631	631	631
LnREI	Pearson Correlation	-0,067	0,079*	0,106**	-0,098*	-0,057	1	0,153**	0,281**	-0,361**
	Sig. (2-tailed)	0,093	0,048	0,007	0,014	0,152		0,000	0,000	0,000
	N	631	631	631	631	631	631	631	631	631
LnPETAI	Pearson Correlation	0,184**	0,937**	0,828**	-0,162**	0,149**	0,153**	1	-0,235**	-0,462**
	Sig. (2-tailed)	0,000	0,000	0,000	0,000	0,000	0,000		0,000	0,000
	N	631	631	631	631	631	631	631	631	631
LnCMVNI	Pearson Correlation	0,009	-0,180**	-0,194**	-0,044	-0,005	0,281**	-0,235**	1	-0,323**
	Sig. (2-tailed)	0,826	0,000	0,000	0,269	0,892	0,000	0,000		0,000
	N	631	631	631	631	631	631	631	631	631
LnGPVNI	Pearson Correlation	0,03	-0,414**	-0,447**	0,123**	-0,040	-0,361**	-0,462**	-0,323**	1
	Sig. (2-tailed)	0,445	0,000	0,000	0,002	0,319	0,000	0,000	0,000	
	N	631	631	631	631	631	631	631	631	631

** Correlation is significant at the 0.01 level (2-tailed); * Correlation is significant at the 0.05 level (2-tailed)

Tabela 16 - Anova -extraction sector – 2010

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
LnDCFI	0,000	2	0,001	277	0,240	0,787
LnROAI	0,020	2	0,004	277	5,280	0,006
LnRFI	0,071	2	0,021	277	3,455	0,033
LnENDI	0,146	2	0,027	277	5,457	0,005
LnLGI	21,593	2	0,080	277	269,980	0,000
LnREI	102,172	2	0,309	277	330,934	0,000
LnPETAI	0,012	2	0,008	277	1,557	0,213
LnCMVN	0,494	2	0,018	277	26,802	0,000
LnGPVNI	0,210	2	0,013	277	16,777	0,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Tabela 17 - Anova – extraction sector – 2011

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
LnDCFI	0,002	2	0,001	277	1,575	0,209
LnROAI	0,045	2	0,005	277	9,537	0,000
LnRFI	0,233	2	0,022	277	10,437	0,000
LnENDI	0,121	2	0,032	277	3,828	0,023
LnLGI	22,021	2	0,103	277	213,762	0,000
LnREI	109,809	2	0,293	277	374,861	0,000
LnPETAI	0,077	2	0,008	277	9,715	0,000
LnCMVNI	0,338	2	0,021	277	16,218	0,000
LnGPVNI	0,226	2	0,015	277	15,450	0,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Tabela 18 - Anova – manufacturing industry-2010

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
LnDCFI	0,000	2	0,000	693	0,455	0,635
LnROAI	0,076	2	0,005	693	14,061	0,000
LnRFI	0,207	2	0,015	693	12,938	0,000
LnENDI	1,228	2	0,030	693	40,896	0,000
LnLGI	40,776	2	0,090	693	451,279	0,000
LnREI	138,575	2	0,201	693	689,259	0,000
LnPETAI	0,137	2	0,007	693	18,884	0,000
LnCMVNI	0,183	2	0,015	693	12,044	0,000
LnGPVNI	0,446	2	0,012	693	37,514	0,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Tabela 19 - Anova – manufacturing sector – 2011

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
LnDCFI	0,000	2	0,000	628	1,474	0,230
LnROAI	0,055	2	0,008	628	7,259	0,001
LnRFI	0,142	2	0,018	628	7,837	0,000
LnENDI	0,185	2	0,037	628	5,070	0,007
LnLGI	0,273	2	0,214	628	1,280	0,279
LnREI	179,675	2	0,102	628	1767,275	0,000
LnPETAI	0,092	2	0,009	628	10,107	0,000
LnCMVNI	0,350	2	0,016	628	22,381	0,000
LnGPVNI	0,439	2	0,012	628	36,632	0,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Tabela 20 - Cluster Analysis 2010 – Extractives Sector

Initial Clusters Centers			
	Cluster		
	1	2	3
LnENDI	0,5882	0,0052	0,5399
LnLGI	0,3172	0,9278	2,4019
LnREI	1,2493	4,0561	0,1194
LnPETAI	0,2418	0,0210	-0,0865
LnCMVN	0,0193	0,3467	0,1268
LnGPVNI	0,3726	0,1293	0,2545

Iteration History ª			
	Cluster		
	1	2	3
1	0,819	0,594	0,852
2	0,079	0,280	0,043
3	0,074	0,250	0,013
4	0,039	0,099	0,031
5	0,035	0,089	0,025
6	0,041	0,062	0,053
7	0,026	0,040	0,030
8	0,007	0,020	0,008
9	0,003	0,000	0,007
10	0,000	0,000	0,000

ª Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is 0,000. the current iteration is 10. The minimum distance between initial centers is 2,400.

Final Clusters Centers			
	Cluster		
	1	2	3
LnENDI	0,1686	0,1397	0,2316
LnLGI	0,7141	0,9485	1,6675
LnREI	0,4982	2,6560	0,5759
LnPETAI	0,0694	0,0923	0,0676
LnCMVN	0,1467	0,2938	0,1425
LnGPVNI	0,2737	0,1820	0,2874

Number of Cases in each Cluster		
Cluster	1	156,000
	2	56,000
	3	68,000
Valid		280,000
Missing		0,000

Tabela 21 - Cluster Analysis 2011 – Extractives Sector

Initial Clusters Centers			
	Cluster		
	1	2	3
LnENDI	0,5914	0,5914	0,0000
LnLGI	1,6405	0,2980	2,5845
LnREI	4,4173	1,1257	0,0070
LnPETAI	0,1123	-0,1613	0,2910
LnCMVNI	0,5245	0,1803	0,0042
LnGPVNI	0,0374	0,5007	0,1578

Iteration History ª			
	Cluster		
	1	2	3
1	1,097	0,822	0,905
2	0,334	0,072	0,059
3	0,192	0,047	0,000
4	0,092	0,019	0,019
5	0,083	0,021	0,000
6	0,096	0,035	0,022
7	0,024	0,014	0,026
8	0,000	0,000	0,000

ª Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is 0,000. the current iteration is 8. The minimum distance between initial centers is 2,680.

Final Clusters Centers			
	Cluster		
	1	2	3
LnENDI	0,1788	0,1832	0,2551
LnLGI	1,0117	0,7935	1,8058
LnREI	2,8428	0,5647	0,3817
LnPETAI	0,1007	0,0401	0,0709
LnCMVN	0,2907	0,1628	0,1576
LnGPVNI	0,1666	0,2718	0,2741

Number of Cases in each Cluster		
Cluster	1	49,000
	2	174,000
	3	57,000
Valid		280,000
Missing		0,000

Tabela 22 - Cluster Analisys 2010 – Manufacturing Sector

Initial Clusters Centers			
	Cluster		
	1	2	3
LnENDI	0,0411	0,5537	0,0662
LnLGI	0,3372	2,2700	0,4748
LnREI	0,2905	1,5279	3,3157
LnPETAI	-0,1741	0,0727	0,0855
LnCMVNI	0,5238	0,4545	0,1089
LnGPVNI	0,3850	0,1368	0,1936

Iteration History ª			
	Cluster		
	1	2	3
1	0,782	0,852	0,874
2	0,022	0,098	0,119
3	0,015	0,075	0,069
4	0,009	0,038	0,036
5	0,006	0,036	0,025
6	0,006	0,036	0,018
7	0,007	0,020	0,009
8	0,005	0,025	0,014
9	0,006	0,010	0,016
10	0,004	0,011	0,013

ª Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is 0,013. the current iteration is 10. The minimum distance between initial centers is 2,460.

Final Clusters Centers			
	Cluster		
	1	2	3
LnENDI	0,1559	0,3085	0,1417
LnLGI	0,7748	1,6471	0,9272
LnREI	0,7745	0,9487	2,2636
LnPETAI	0,0273	0,0381	0,0727
LnCMVN	0,3086	0,3108	0,3618
LnGPVNI	0,3148	0,3062	0,2310

Number of Cases in each Cluster		
Cluster	1	371,000
	2	148,000
	3	177,000
Valid		696,000
Missing		0,000

Tabela 23 - Cluster Analysis 2011 – Manufacturing Sector

Initial Clusters Centers			
	Cluster		
	1	2	3
LnENDI	0,6995	0,0796	0,1087
LnLGI	2,3395	0,3153	2,0373
LnREI	0,1001	1,5816	3,6518
LnPETAI	0,0177	-0,3181	0,0482
LnCMVNI	0,0378	0,5779	0,4785
LnGPVNI	0,1415	0,4473	0,1192

Iteration History ª			
	Cluster		
	1	2	3
1	1,028	0,816	1,040
2	0,138	0,066	0,344
3	0,061	0,018	0,139
4	0,039	0,007	0,064
5	0,036	0,028	0,000
6	0,034	0,030	0,000
7	0,051	0,048	0,000
8	0,046	0,051	0,015
9	0,044	0,053	0,023
10	0,036	0,047	0,037

ª Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is 0,036. the current iteration is 10. The minimum distance between initial centers is 2,679.

Final Clusters Centers			
	Cluster		
	1	2	3
LnENDI	0,2190	0,2163	0,1544
LnLGI	1,1029	0,9630	0,9819
LnREI	0,4911	1,3240	2,7154
LnPETAI	0,0174	0,0095	0,0628
LnCMVN	0,2852	0,3364	0,3729
LnGPVNI	0,3306	0,2940	0,2189

Number of Cases in each Cluster		
Cluster	1	270,000
	2	267,000
	3	94,000
Valid		631,000
Missing		0,000