# A machine learning approach to risk disclosure reporting

Max Cardoso de Resende * Alexandre Rezende Ferreira †

**XXIII Encontro de Economia da Região Sul – ANPEC SUL 2020**

**Abstract**

It is widely recognized that corporate annual reports play a key role in financial markets. Given the debate on risk analysis, this paper applies a machine learning statistical technique called Latent Dirichlet Allocation (LDA) in order to classify companies risks reported on 10-k SEC Form from 2006 to 2017 and applies a predictive logit model to assess the idiosyncratic risks of individual firms and relate it to firm-specific characteristics, such as market equity, total assets, among others. Among several results, it was verified that non-diversifiable risks, such as tax, competition, insurance, intellectual property and government behave similar throughout all the industries, whereas Financial Statements concerns appear to be temporary. Moreover, market equity, total assets and the firm's age are predictive of all risks. and firms for which the risk is captured are smaller on average, present lower market equity and total assets besides been younger and slightly less profitable when compared to traditional firms.

**Keywords** Idiosyncratic risks; Topic modeling; LDA; 10-k SEC Form.

**JEL Classification**: C58; G17; G32.

*Corresponding author. Department of Economics, Federal University of Santa Catarina; e-mail: max.resende@ufsc.br.
†Department of Economics, Federal University of Santa Catarina; e-mail: alexandrerezende@yahoo.com.br.

# 1 Introduction

It is a core concept in finance that undiversifiable risks are systematic and originate from macroeconomic variables, which implies in investors demanding higher expected returns given their inability to hedge or assessing market risk. On the other hand, diversified or idiosyncratic risks are related to a particular company and its core business. According to Ang *et al.* (2006) and Elton *et al.* (2009), there is empirical evidence of a negative correlation between diversifiable risks and returns to investors, but the sources of those risks are not well understood, and it is common to use market returns as a proxy for systematic risks, and the variance of the residuals of a factor model as a measure for idiosyncratic risk.

The Capital Asset Pricing Model(CAPM) is the reference model to measure the market risk of financial securities. Its derivation is the result of Markowitz Optimal Portfolio Model (Rubinstein, 2002), which postulates that risk can be reduced, but not eliminated, without changing expected portfolio return through diversification. Despite its straight forward application, several tests cast doubt on the CAPM model, especially when defying market return as a proxy for systematic risk (Harvey *et al.* , 2016; Hou *et al.* , 2017). Previous theoretical work on idiosyncratic risk shows how under certain conditions, there is a positive correlation between idiosyncratic risk and expected returns (Merton, 1987; Hirshleifer, 1988), while empirical evidence has shown a negative relationship between them instead (Ang *et al.* , 2006; Brandt *et al.* , 2009). This is called "the idiosyncratic volatility puzzle" and many researchers have tried to explain it, such as: Barberis & Huang (2008), Fu (2009), Han & Lesmond (2011), Jiang *et al.* (2009) and Wong (2011). Thus, Hou & Loh (2016) provides a review of all these results and asses by how much each of then contribute to explain of the puzzle, concluding that lottery preference and market friction are the best explanations.

It is widely recognized that corporate annual reports play a key role in financial markets, especially, because it disclosures several information about a firm. Therefore, according to Huang & Li (2011) and Loughran & McDonald (2016) several algorithms have been developed for labeling, extract and quantify textual information reported in these documents. More recently, with the increase in computational efficiency, applying machine learning methodologies has become an important issue to enhance financial strategies. Given the debate on risk mapping and the use of computational techniques applied to the financial market, this paper goes beyond the traditional proxies and contribute to this literature by looking into the risks companies report in their 10-K Form submitted to the SEC (U.S. Securities and Exchange Commission) from 2006 [1] to 2017, in order to classify then into systematic or idiosyncratic based on a machine learning statistical technique called Latent Dirichlet Allocation (LDA). Afterwards, this research applies a logit regression for idiosyncratic risk prediction of individual firms and relate it to firm-specific characteristics for a sample of 2415 companies to understand what drives then.

Thus, this study provides evidence that market (systematic) risks, such as tax, competition, insurance, intellectual property and government behave similar throughout all the industries, while employee and environment matters were presented as the least relevant risks and financial results effects sens to be temporally on the firm's risk profile. Moreover, market equity, total assets and firm's age are the predictive factors of all types of risks, and the firms for which the risk is captured are smaller on average, present lower market equity and total assets, besides been younger and slightly less profitable firms and presents similar book-to-market ratios and investment rates when compared to traditional firms. These findings suggest that even if big and consolidate firms disclosure their risks poorly, investors still places a higher risk-aversion on younger and less popular firms as suggested by the classical financial theory.

The paper proceeds as follows. Section 2 presents previous literature dealing with machine learning methods to extract relevant information from companies annual financial reports; Section 3 briefly describes the machine learning statistical technique; Section 4 describes the 10-K SEC Form; Section 5 describes

[1] The year that SEC started requiring disclosure of Risk Factors.

the operational perspective of identifying the most relevant risks and how they behave throughout all the industries; Section 6 assess how balance sheet variables correlate with the presence of each risk extracted from the LDA algorithm through a logit model; and, Section 7 brings the final remarks.

## 2 Literature review

Assessing and managing risk to quantity textual information in financial statements applying machine learning methods delivers a more robust capability to traders and researchers to detect meaningful information from data and to evaluate complex correlations in a way that risk disclosures reflect subsequent market measures of risk.

Li (2008) examines the relation between public companies annual reports (10-K Form) readability with firm performance and earnings persistence using the Gunning Fog index[2] and the length of the document. The empirical evidence suggests that annual reports of firms with poor performance are more difficult to be interpret, whereas the profits of firms with annual reports that are easier to read are more persistent. In the matter of corporate finance, these findings can be interpreted as the principal-agent problem, where managers may be opportunistically structuring the annual reports to hide adverse information from investors.

Kravet & Muslu (2013) investigates the information content in the risk factor section of SEC Form 10-k and its impact on stock market using machine learning - UNIX Perl code - for more than 5000 firms from 1994 to 2007. They have concluded that risk factor disclosures that are lengthy or more less firm-specific experience negative capital market consequences, such as higher cost of capital, greater stock price volatility, weaker market responses, and declines in analysts' ability to assess fundamental risks. Moreover, Loughran & McDonald (2016) in a survey of the literature provides details on textual analysis algorithms to analyze information content, earnings quality, market efficiency, and assess risk factors. The authors show that traditional concept readability and Naıve Bayesian methodology is ineffective.

Campbell *et al.* (2014) identified a list of words that repeatedly appeared in firms' risk factor sections using LDA. In general, 30% of the key words relates to financial risk, litigation risk, tax risk, while 70% accounts for either "other – systematic" or "other – idiosyncratic" risk factors. They concluded that by the percentage of key words associated with different types of risk in the 10-K Form Risk Section, the type of exposure a firm faces determines whether they devote a greater portion of their report toward describing that risk type and that managers provide useful risk factor disclosures and investors incorporate this information into their portfolios.

Other studies, such as Das & Chen (2007), Oh & Sheng (2011) and Nann *et al.* (2013) focused on stock prices prediction movements using stock micro blog postings, author profile, and inter-day stock price posts disclosure by Yahoo Finance based on data mining algorithms (Naïve Bayes, Support Vector Machine and Weka data mining software). The overall results suggests that market activity has strong correlation with information disclosure and it is a good prediction for market volatility. The authors have also concluded that portfolios constructed with firms predicted by the text-based model are shown to produce positive average stock return.

Overall, the articles presented here provides evidence of informational value in annual and quarterly risk factor disclosures by showing how to extract relevant features using machine learning techniques from then and by presenting a strong correlation between the linguistic features of annual reports, risk analysis and firm performance. However, they lack clear evidence on how the risk classification sustain across industries and on how they correlate with key performance indicators from the firms.

[2] A statistical formula that measures readability developed by Robert Gunning as a function of the number of words per sentence and the number of syllables per word.

## 3 Latent Dirichlet allocation approach

In text mining, there is a collections of documents (unlabeled data), such as blog posts, news articles or reports, that can be divided into natural groups so that in order to understand them separately. That is, topic modeling is a method for unsupervised classification of such documents to discover information, new features and be useful for categorization, similar to clustering on numeric data, which finds natural groups of items (Kumar & Ravi, 2016). In short, the flow that happens in this type of modeling is a reduction of the analyzed texts; the application of unsupervised machine learning that generates topics, where each topic incorporates a number of words; and, finally classify the analyzed documents according to the topics generated, indicating which topic fits best. There are many models to perform this type of analysis, besides the LDA[3], such as, Latent Semantic Analysis (LSA / LSI), Probabilistic Latent Semantic Analysis (pLSA) and the Non Negative Matrix Factorization (NNM).

Latent Dirichlet Allocation[4] is a mathematical algorithm for fitting a topic model proposed by Blei *et al.* (2003), where documents are defined as a mixture of various topics, so instead of belonging to a specific topic, it lies in the simplex formed by all the topics. And topics are represented by a set of terms, which can be a word or a set of words in a particular order. It is important that a set of words can be set to allow for both "shares" and "common shares" to be representative of a topic without any prior on which is more relevant. For this paper purpose, the documents are each of the risks that managers report in their Form 10-K, and since managers are reporting risks, the topics are the kinds of risks reported – as an example: political risk, environmental risk; employee risk.

One fundamental assumption of the model is that the probability that a word appears in the topics is given by a sparse Dirichlet distribution which lead to extreme concentrations of mass in certain values. That is, a word that has a high probability of appearing in one topic, has almost zero probability of appearing in another topic. Thus, given these assumptions about the document formation process, and a prior distribution of documents into topics and words into topics, LDA uses Bayesian updating to extract what are these representative words for each topic and their respective probability.

## 4 Form 10-k

The Form 10-K is an annual report required by the U.S. Securities and Exchange, that gives a comprehensive summary of public company's financial performance that should be reported 90 days after the end of each fiscal year. The SEC define risks as the reasons why the stock would be volatile and requires companies to list in their Form 10-K the risk factors they are subject to. As an example, in 2016, Apple Inc. (AAPL) wrote as a risk factor in her Form 10-K:

> **The Company's success depends largely on the continued service and availability of key personnel.** Much of the Company's future success depends on the continued availability and service of key personnel, including its Chief Executive Officer, executive team and other highly skilled employees. Experienced personnel in the technology industry are in high demand and competition for their talents is intense, especially in Silicon Valley, where most of the Company's key personnel are located.

[3]This paper uses Python package scikit-learn and the English stop words dictionary to do all the steps mentioned here.
[4]The Dirichlet in LDA comes from the Dirichlet distribution used in this model.

This is one out of 27 risks the company reported that year. This part of the document accounts for 11% of the text in the Form 10-K. There is some early evidence that this risk is informative, correlating with measures of past and future volatility (Campbell *et al.* (2014)). Although there is some criticism that these risks are generic, there is pressure from the SEC to make them more specific. And there is evidence that analysts forecasts are more precise, the more the risks are informative (Hope *et al.* (2016)).

To make use of those risk factors listed, topic modelling was applied to discover relevant topics in a collection of documents. In this case, the documents are each of the risks that managers report in their firm's Form 10-K from 2006 to 2017. And since managers are reporting risks, the topics are the kinds of risks reported. As an example: political risk, environmental risk. Constructive means that it assumes a model for the data generating process: Therefore, LDA assumes that there exists certain words that are representative of a topic, even if we have no prior knowledge of what are those words, and assess if they have a high probability of appearing on one topic and low probability of appearing in another. So, given a prior distribution of documents and words into topics, LDA uses Bayesian updating to extract what are these representative words for each topic. For example:

> clinical, fda, approval, candidates, trials, clinical trials, product candidates, drug, research, marketing

This is the group of words that compose one of the topics which LDA can identified. By inspecting those words and a random sample of documents that are classified into this topic, one may infer that what is represented here is the risk of an unsuccessful drug test. Thus, it may be assigned a meaning of Clinical Trials, for example..

# 5 Risk factor disclosure: LDA model fitting

Annual Form 10-K filings are downloaded from the SEC's Electronic Data Gathering and Retrieval (EDGAR) database, and are then processed to generate appropriate counting measures that objectively quantify firms' risk disclosures. The files downloaded from the SEC database need to be processed to extract the individual risk factors. These files contain all the information the companies sent as part of the form, including the actual text of it, all the exhibits that accompany it and a header with information about the submission (firm CIK, submission date, end of fiscal year, filenames contained in it and more). This article follows Campbell *et al.* (2014) in the way they identify the Risk Factors section.

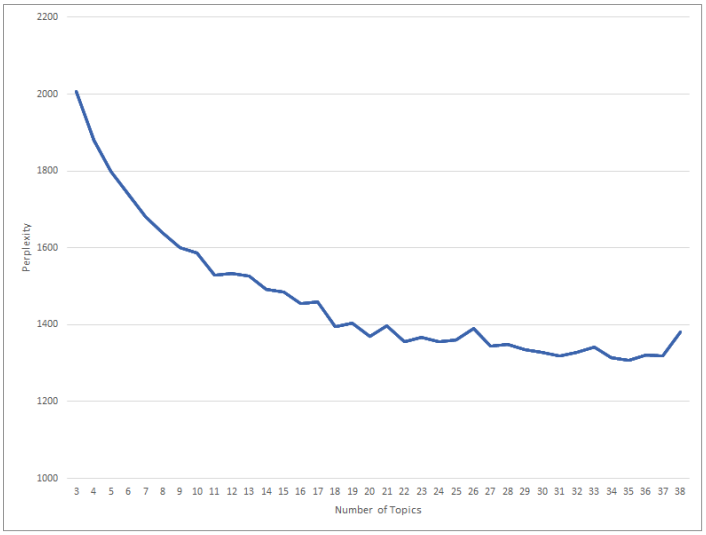
From an operational perspective, after extracting the risks from the Form 10-K, it was created a matrix of document-terms where each document is a row, each term is a column, and in each cell is the number of times the term appears in the documents. Them, all the stop words were extract from the documents, which are articles, conjunctions and similar words that are not informative or have a meaning by itself, like "the", "a", "and". Next, it was defined that a word is any sequence of 3 letters or more, regardless of case, excluding numbers, and as a term any sequence of 1 or 2 words. A restriction had to be imposed to the size of document-term matrix to a level that can be processed with the available computational resources. That is, a term can have a maximum document frequency of 0.1 and minimum 0.001, and within this range it was used the 5000 most frequent terms. This is in line with the restrictions imposed by Hansen *et al.* (2017), in which 9000 terms are used, but for a smaller document sample. Lastly, having the terms been identified, the algorithm now counts the terms in each document and creates a document-term matrix where each document is a row, each term is a column, and in each cell is the number of times the term appears in the document. This is the input LDA uses to identify topics.

Moreover, to build interpretable topic models, LDA needs a prior on the number of topics that comprise the documents available. To give some guidance on that, there is a standard measure of goodness-of-fit used in information theory and language modeling, perplexity score (PP), which is defined as (Blei *et al.* , 2003):

$$PP = exp\left[-\frac{\sum_d \sum_v x_{d,v} log\left(\sum_k \beta_k^v \theta_d^k\right)}{\sum_d N_d}\right] \tag{1}$$

where $x_{d,v}$ is the count of term $v$ in document $d$, $\beta_k^v$ is the probability of term $v$ for topic $k$, $\theta_d^k$ is the share of document $d$ in topic $k$ and $N_d$ is the total number of terms in document $d$. Figure 1 plots the perplexities scores from fitting a LDA model with 3 up to 38 topics:

Figure 1: Perplexity results for LDA model for different number of topics.




According to Blei *et al.* (2003) and Blei (2012), lower perplexity score indicates better generalization performance, however this number is just a guidance and the interpretability of the topics should prevail over using the number that minimizes perplexity. For this research purpose, it was chosen 18 topics, once a higher number of topics seem to provide only small increments in the goodness-of-fit.

To extract the risks from Section 1A of the 10-k Statements, the section was split into different risks by each subtitle and processed using LDA. The Top Words column in the Table 1 below, are the ones technique classified as being representative of each of the eighteen risks. Basically, for each risk, it was given a Title and an Abbreviation used to refer to those risks in different tables and figure henceforth. The sample of 10-Ks used ranges from 2006, when the section was introduced, to 2017. To avoid any forward looking bias, the first year alone was used to identify the risks and then this classification was used in the remaining of the sample. Results are reported in Table 1.

Table 1: List of risks in the Risk Factors section of the 10-K Forms

| Title | Abb | Top Words - LDA Classification |
|---|---|---|
| Employee Retention | EMP | personnel, employees, retain, key, attract, qualified, executive, officer, year ended, chief |
| Environmental | ENV | environmental, claims, liability, liabilities, damages, property, legal, litigation, incur, laws regulations |
| Competition | COM | technologies, develop, compete, technology, competition, software, new products, greater, marketing, acceptance |
| Oil and Gas Production | OIL | gas, natural, oil, properties, natural gas, prices, production, reserves, energy, construction |
| Tax & Losses | TAX | income, tax, loans, rates, loan, losses, rate, net, investment, real |
| Quarterly Results | RES | million, year, approximately, net, quarter, ended, fiscal, period, total, agreement |
| Supply chain | SC | manufacturing, contracts, supply, suppliers, contract, production, demand, materials, manufacturers, inventory |
| Clinical Trials | CT | clinical, fda, approval, candidates, trials, clinical trials, product candidates, drug, research, marketing |
| Political | POL | government, state, federal, health, laws regulations, regulation, comply, compliance, act, applicable |
| International | INT | foreign, united, united states, international, currency, countries, exchange, political, fluctuations, accounts |
| Debt & Financing | DEB | debt, credit, financing, indebtedness, facility, funds, pay, obligations, dividends, distributions |
| IT | IT | systems, demand, growth, internet, distribution, decline, information, network, data, security |
| Reporting & Controls | REP | internal, reporting, controls, financial reporting, trading, price common, accounting, market price, fluctuations, internal control |
| Acquisitions | ACQ | acquisitions, acquisition, acquired, growth, businesses, strategy, successfully, strategic, manage, acquire |
| Governance | GOV | directors, board, stockholders, board directors, provisions, preferred, rights, shares, preferred stock, partner |
| Insurance | INS | insurance, coverage, liability, claims, losses, competition, product liability, consumer, institutions, policies |
| Financial Statements | STM | shares, shares common, statements, interests, form, corporation, ownership, shareholders, public, units |
| Intellectual Property | IP | rights, property, intellectual, intellectual property, patent, patents, proprietary, license, technology, parties |

LDA provide as output for each document its position in the simplex formed by the topics. So for each firm's risk yearly reported there is a vector of numbers between 0 and 1 for each of the 18 topics (risks). Then, to move from this mixed participation to a binary participation, in which each risk is of a particular kind, it was collected the maximum number from these vectors representing the position of risks in the simplex. Then, a cutoff was imposed at the 7th decile of the distribution of those collected numbers, or 0.6059. So, given the position in the simplex of a particular risk, if the value for a particular topic is greater than 0.6059, it can be assumed assume that the firm faces that source of risk, whereas if there is not any value greater than 0.6059, no particular risk was mentioned.

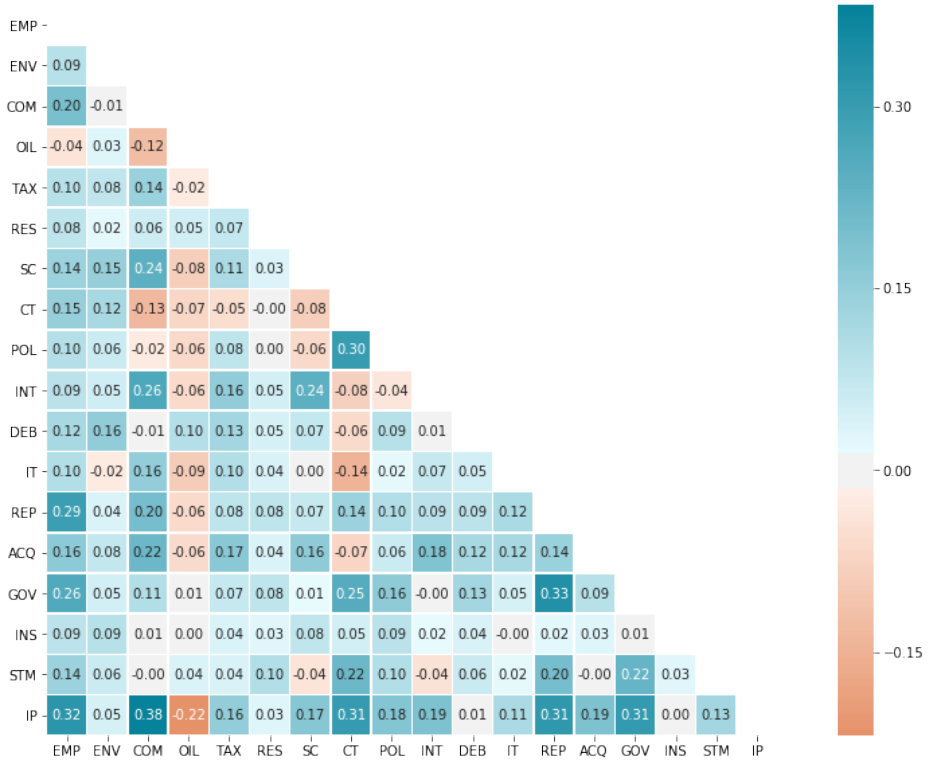
This threshold comes with a trade-off: it allows the researcher/trader to clearly say that a particular risk was mentioned by the company, given that it devoted some time to describe it almost uniquely. On the other hand, it ignores intersections between risks. For example, if the firm mentioned that international competition is relevant, it might not be captured neither by Competition or International individually. Also,

it makes it hard to claim the absence of a risk, because falling bellow the threshold just means that the company didn't devote enough attention to it.

The risks found by the fitted model are listed in the first column of Table 1 which, also, reports the Top 10 representative words returned by the LDA model, ordered from highest to lowest coefficient – it was attributed to them a title based on the word list and an abbreviation (Abb) that will be used in all empirical exercises to come. Moreover, it is interesting how all of them have an unexpected component. For example, Environmental risk can be exemplified by the chance that an accident in an offshore platform might spill oil and Government risk by the chance of new regulations and the unforeseen costs of complying with them.

Figure 2 and Table 5 (in Appendix) reports all the correlations between the risks, considering their mutual appearance in the same firm, during the same year. Overall, these risks have low correlations between each other, ranging from 0.38 to -0.22, but when looking at pairwise results some interesting features emerge. The analysis of Intellectual Property (IP) correlations are important to treat it as a strategic business asset: first, a 0.32 positive correlation with Employee related risk is consistent with the idea that the firm needs intellectual capital to produce innovations, so the risk of not attracting or retaining key employees is relevant (Klein, 2009); second, IP is 0.38 correlated with Competition showing how these firms see these properties as key to keep their competitiveness and increase markup (Acemoglu & Akcigit, 2012); and, third, it displays high correlations with Clinical Trials related risks, Reporting and Governance, whereas, is negatively correlated with Oil, suggesting that this protection is not relevant, probably because the oil production process is already well known.

Figure 2: Heatmap of risk correlations

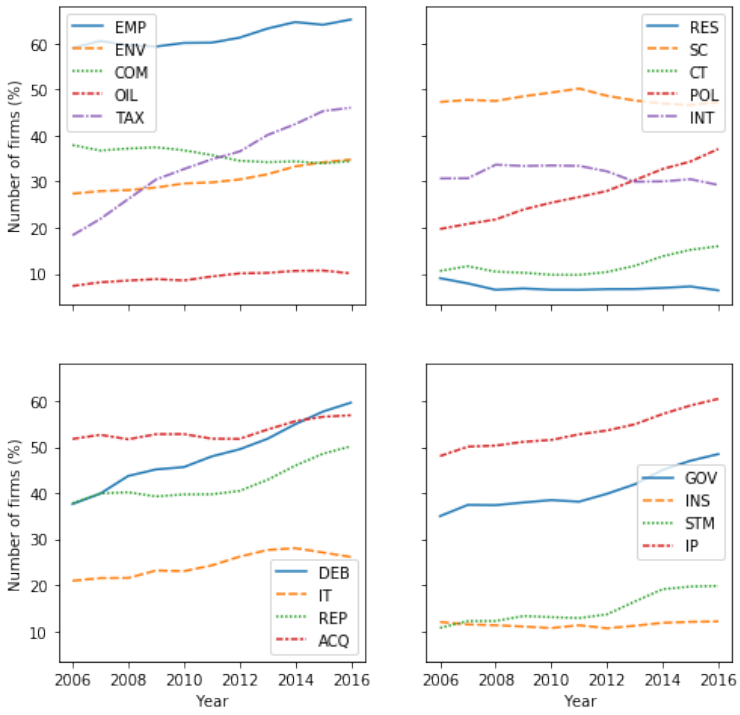


Moreover, Clinical trials are positively correlated with both Political (0.30) and IP risks (0.31), reinforcing how government regulations are relevant on risk analysis despite not incurring in greater fear of failure in treatment testing or lawsuits, (Kunreuther, 2002) and by bringing evidence that firms, indeed, value the protection of their discoveries, respectively. As well, Governance and Reporting are positively correlated, suggesting that firms that worry about their governance practices are aware of how they disclose financial statements (0.33).

And, as for oil risks, the results have reported negatively correlations with almost all other risks, with the exceptions of Results, Debt, and Financial Statements, which might be a feature of the sample period, when the oil price has presented severed volatility (Laurini *et al.* , 2020). On the other hand, the outcomes suggest a contradictory fact by reporting only a small positively correlation with environment aspects (0.03), because Refineries are generally considered a major source of pollutants in areas where they are located and are regulated by a number of environmental laws related to air, land and water (Muralidhar, 2010).

Figure 3 and Table 6 - in Appendix - reports the dynamics of reporting the risks and a raw count of how many firms did report a risk each year used for this paper analysis. It can be noticed that there is an increase in reporting Debt risks – which are mainly related to changes in capital structure through a combination of equities and liabilities and distribution of dividends across shareholders – ranging from 38% in 2006 to 60% of the sample in 2016. Indeed, this high level of disclosure is consistent with the rise in outstanding corporate debt across all firms, which can explain why firms are reporting more risks related to their access to credit. Besides, there is an increase in Tax related risk disclosure from 18% in 2006 to 46% in 2016. Other minor increases were observed in risk related to financial statements, which went from 11% to 20% and on reporting and internal controls ricks, which had a 11% increase in its disclosure index.

Figure 3: Percentage of firms that report the risk each year



Furthermore, this research assess how concentrated those risks are in a particular industry. In order to do that, first, it was defined a 2 digit SIC code[5] to identify which specific industry the firm belongs to and, second, two approaches of market concentration measure $Concentration = \sum_{i=1}^{N} s_i^2$ were applied (Brown & Kapadia, 2007): (i) Equal Weighted (EW), where $s_i$ is the number of firms reporting the risk in industry $i$, divided by the total number of firms reporting that risk; (ii) Value Weighted (VW) concentration measure, where each firm is weighted based on its size (or market capitalization), such that $s_i$ becomes the total market capitalization of all firms reporting the risk in industry $i$ , divided by the total market capitalization of all firms reporting that risk. Both measures range from $\frac{1}{N}$ to 1, where $N$ is the number of different industries that at least one firm reports the risk.

Intuitively, EW approach considers the risk contribution from each industry component is made equal, regardless of how small or large the firm is. These measure is very different from the cap-value weighted method where each firm is weighted based on its size (or market capitalization). Roughly speaking, both index is influenced more by the biggest firms in a industry as it is the sum of the squared of all the firms, whereas VW can express better the effects of the largest firms in the industry, which results in the performance of few dominating the rest and increasing concentration risk, where the resulting measure is similar to a minimum-variance portfolio subject to a diversification constraint on the weights of its components (Brown & Kapadia, 2007; Campbell *et al.* , 2014).

Overall, the results from Table 2 suggests that Clinical Trials is potentially threatening for the pharmaceutical industry, and Oil exploration and production risks is also highly concentrated in the miming industry which raises, specially, due to changes in global economic conditions and volatility in oil and natural gas prices. However, undiversifiable risks, such as tax, competition, insurance, intellectual property and government behave similar throughout all the industries, while employee and environment matters were presented as the least relevant risks. One possible explanation why these last two risks are not well documented in the 10-k Form may be the result of the degree of establishment and reliance on US labor and environmental laws.

Table 2: Concentration of risks per industry

| Risk | Concentration | | Risk | Concentration | |
|---|---|---|---|---|---|
| | EW | VW | | EW | VW |
| Employee | 0.070 | 0.088 | Competition | 0.124 | 0.139 |
| Environmental | 0.062 | 0.071 | Tax | 0.079 | 0.145 |
| Oil | 0.242 | 0.267 | Supply chain | 0.076 | 0.105 |
| Results | 0.079 | 0.090 | Government | 0.099 | 0.098 |
| Clinical Trials | 0.442 | 0.577 | Reporting | 0.085 | 0.137 |
| International | 0.095 | 0.092 | Insurance | 0.064 | 0.127 |
| Debt & Financing | 0.048 | 0.056 | Intellectual Property | 0.117 | 0.148 |
| IT | 0.115 | 0.136 | Governance | 0.089 | 0.090 |
| Acquisitions | 0.072 | 0.061 | Financial Statements | 0.078 | 0.141 |

[5]The Standard Industrial Classification (SIC) are two-digit code that categorize the industries that companies belong to based on their business activities.

Assessing if debt risk is driven mostly by idiosyncratic firm characteristics or by systematic factors is an important issue for the assessment of financial stability. Results demonstrated that this source of risk is the least concentrated in an industry, and should be interpreted as a systematic factor suggesting that macroeconomic conditions have an important role in explaining the evolution of credit risk. However, default probabilities are also influenced by several firm-specific characteristics (Bonfim, 2009), which highlights the need to taking simultaneously into account micro data as well as macroeconomic information.

Lastly, Table 3 presents the probability that a company will report next period's risks, given that in the current period it did not and vice-versa. The results suggests that the probability of going from non-reporting to reporting is greater for all risks except for employee risk. It seems to indicate that as soon as the company hits a point where attracting and retaining talent become a concern, it keeps that way for longer. On the other hand, Clinical Trial and Intellectual Property have very low probabilities of transition, indicating that they might be more related to the company non-variant characteristics than to a particular situation or point in time. The one more likely to exit are Results and Financial Statements, showing that concerns with negative results might be temporary. Following Brown & Kapadia (2007), the analyses suggests that the willingness of managers to supply equity to risky new information varies over time and that this variation explains observed trends in idiosyncratic risk disclosure.

Table 3: Risk change probabilities

| | Reporting the risk | Unreport the risk |
|---|---|---|
| EMP | 8.39% | 5.80% |
| ENV | 4.66% | 12.04% |
| COM | 4.03% | 6.99% |
| OIL | 1.79% | 19.65% |
| TAX | 5.61% | 19.08% |
| RES | 2.40% | 30.72% |
| SC | 6.35% | 7.28% |
| CT | 0.58% | 4.36% |
| POL | 3.94% | 15.25% |
| INT | 5.05% | 11.19% |
| DEB | 5.85% | 10.78% |
| IT | 5.16% | 17.59% |
| REP | 6.84% | 10.23% |
| ACQ | 11.32% | 10.21% |
| GOV | 3.02% | 5.25% |
| INS | 1.57% | 13.16% |
| STM | 3.69% | 21.07% |
| IP | 2.57% | 3.99% |

# 6 Prediction Modelling for risks

The final exercise proposed on this paper aims to analyze how balance sheet variables correlate with the presence of each risk and then to produce a reliable model for the risk profile of each firm who filled the 10-K SEC Form back in 2006 – which comprises information on 2415 American firms. Therefore, this study proposes a more sophisticated event study methodology by applying a logit[6] regression for r idiosyncratic risk prediction regarding several dimensions of the firm's financial situation:

$$\text{logit}(Risk_{i,t}) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta' X$$

$$\text{logit}(P(Risk_{i,t}) = 1) = \alpha + \beta_1 * ME_{i,t} + \beta_2 * TA_{i,t} + \beta_3 * IA_{i,t} + \beta_4 * Prof_{i,t} + \beta_5 * Age_{i,t} + \beta_6 * BM_{i,t}$$

where ME is the market equity of the firm calculated as current stock price times shares outstanding. TA is total assets which controls for firm size, IA is investment in assets calculated as the yearly change in assets, Prof and BM are, respectively, the firm's profitability and book to market value ratio – these variables are all related to Fama & French (2015) – and, Age is years for which hte companies have been listed on NYSE (Jovanovic & Rousseau, 2001; Wu *et al.* , 2010).

Table 4 and Table 7 – in Appendix – shows how firm-specific characteristics correlate with the presence of risk. In general, firms for which the risk is captured are smaller on average, having lower market equity and total assets. Also, they are are younger, slightly less profitable and presents similar book-to-market ratios and investment rates, when compared to traditional firms. This suggests that even if big and consolidate firms report their risks poorly, investors still places a higher risk-aversion on less popular firms as suggested by the classical finance theory.

Further, the results obtained show that the risk profile of traded firms is significantly influenced by market equity, total assets and the age of listing. Equity markets can be volatile, once share prices rise and fall in response to market conditions, company-specific events, among other political and economic developments and this may explain why higher valued firms face lower financial related risks (Bonfim, 2009) – DEB, STM and INS – but still feels threat by human capital risks – EMP and IP (Klein, 2009). As for total assets display a negative effect on the majority of risks, addressing that bigger companies face a lower level of risk, probably because promotes stability operating cash flows, have access to risk management expertise, or that have economies of scale in hedging costs, are more likely to offers better guarantees for loans than smaller firms (Jorge & Augusto, 2011).

It is worth noting that newly listed firms tend to show higher idiosyncratic risk probabilities because younger firms typically have weaker fundamentals (Jovanovic & Rousseau, 2001), especially those linked to financial performance (RES, DEB, REP and STM), as argued by Wu *et al.* (2010) – suggesting that an increasing proportion of newly listed firms could lead to an upward trend in idiosyncratic risk – and to mitigate agency costs they have to provide more information about their business activities. That is, the information disclosure revealed by the firm is often considered difficult to understand and and may fail to promote stakeholder engagement, making investors require a higher return because of possible estimation

[6] Logistic Regression is the technique most often used in the market for risk analysis and can be stated in terms of probability (P) (Brooks, 2019): $\pi = P(Y = y|X = x) = \frac{e^{\alpha+\beta_i X_i}}{1+e^{\alpha+\beta_i X_I}}$, where $\pi = P(.)$ is the probability of the outcome of interest – in this research context, the risk extracted from the LDA procedure – equals to 0 or 1, $\alpha$ is the intercept and $\beta$ is a vector of regression coefficients which are usually estimated via maximum likelihood estimation using numerical methods, and $X_i$ is a set of $n$ independent variables.

errors resulting from the information. Also, the results provide some evidence that older firms are more reluctant to make environmentally responsible investment decisions as expressed by Fisher-Vanden & Thorburn (2011) and appear to conflict with risk measures and firm value maximization. Importantly, this finding has important implications for the success of voluntary environmental and greenhouse gas emissions reduction programs indicating that environmental risk tends to increase in the U.S.

Table 4: Logit estimation of the probability of risks risk given by the LDA approach

| Risk | ME | TA | IA | Prof | Age | BM | Pseudo $R^2$ |
|---|---|---|---|---|---|---|---|
| EMP | $0.290^{***}$ | $-1.015^{***}$ | $-0.058^{**}$ | $-0.007$ | $-0.474^{***}$ | $-0.055^{***}$ | 0.067 |
| | (10.16) | (−10.87) | (−2.37) | (−0.44) | (−20.44) | (−4.70) | |
| ENV | $-0.090^{***}$ | 0.004 | $-0.006$ | $-0.028$ | $0.069^{**}$ | 0.010 | 0.001 |
| | (−5.74) | (0.23) | (−0.44) | (−1.18) | (2.68) | (1.00) | |
| COM | $0.200^{***}$ | $-0.517^{***}$ | $-0.002$ | $-0.010$ | $-0.120^{***}$ | $-0.040^{**}$ | 0.010 |
| | (3.27) | (−3.60) | (−0.10) | (−0.62) | (−5.38) | (−2.31) | |
| OIL | $-0.219^{***}$ | $0.136^{**}$ | $0.037^{*}$ | $-0.001$ | $0.205^{***}$ | $0.113^{*}$ | 0.011 |
| | (−3.49) | (2.53) | (2.16) | (−0.14) | (12.05) | (1.84) | |
| TAX | 0.028 | $-0.001$ | 0.003 | 0.003 | 0.009 | 0.009 | 0.000 |
| | (1.51) | (−0.04) | (0.29) | (0.33) | (0.80) | (0.42) | |
| RES | $-0.560^{**}$ | $-0.151$ | $-0.701^{*}$ | 0.001 | $-0.171^{***}$ | $0.096^{*}$ | 0.013 |
| | (−3.02) | (−1.02) | (−1.94) | (0.08) | (−4.30) | (2.14) | |
| SC | $-0.031$ | $-0.099^{*}$ | $-0.199$ | $-0.003$ | $0.164^{***}$ | $0.066^{**}$ | 0.006 |
| | (−1.00) | (−2.05) | (−0.68) | (−0.36) | (7.81) | (2.38) | |
| CT | $0.737^{***}$ | $-2.665^{***}$ | $-0.001$ | $-0.015$ | $-0.810^{***}$ | $-0.418^{***}$ | 0.081 |
| | (4.96) | (−5.85) | (−0.17) | (−0.36) | (−15.20) | (−3.92) | |
| POL | $0.059^{***}$ | $-0.023$ | 0.013 | $-0.012$ | $-0.300^{***}$ | $-0.131^{***}$ | 0.014 |
| | (4.40) | (−1.69) | (1.29) | (−0.58) | (−13.37) | (−3.30) | |
| INT | $0.201^{***}$ | $-0.314^{***}$ | $-0.003$ | 0.006 | $0.136^{***}$ | $-0.008$ | 0.005 |
| | (6.53) | (−10.34) | (−0.22) | (0.55) | (11.48) | (−0.65) | |
| DEB | $-0.422^{***}$ | $0.132^{***}$ | 0.002 | $-0.016$ | $-0.135^{***}$ | 0.004 | 0.012 |
| | (−5.11) | (4.86) | (0.20) | (−1.57) | (−9.31) | (0.30) | |
| IT | $0.047^{**}$ | $-0.035^{***}$ | $-0.056$ | 0.005 | $-0.201^{***}$ | 0.018 | 0.006 |
| | (2.93) | (−3.68) | (−1.21) | (0.34) | (−8.36) | (1.26) | |
| REP | 0.016 | $-0.871^{***}$ | 0.000 | $-0.051$ | $-0.659^{***}$ | 0.014 | 0.076 |
| | (0.14) | (−4.40) | (0.01) | (−1.05) | (−48.42) | (0.92) | |
| ACQ | $0.065^{***}$ | $-0.171^{***}$ | 0.008 | 0.027 | $-0.055^{**}$ | 0.013 | 0.003 |
| | (4.97) | (−9.08) | (0.68) | (1.33) | (−2.69) | (1.54) | |
| GOV | $-0.044$ | $-0.525^{***}$ | 0.006 | $-0.006$ | $-0.905^{***}$ | $-0.127^{***}$ | 0.102 |
| | (−0.54) | (−6.31) | (0.51) | (−0.40) | (−23.24) | (−4.49) | |
| INS | $-0.051$ | $-0.139^{**}$ | $-0.004$ | $-0.035^{**}$ | 0.007 | 0.013 | 0.002 |
| | (−0.92) | (−2.93) | (−0.38) | (−2.61) | (0.33) | (1.09) | |
| STM | $-0.131^{**}$ | $-0.109^{*}$ | 0.003 | $-0.011$ | $-0.766^{***}$ | $-0.078^{***}$ | 0.052 |
| | (−2.28) | (−2.03) | (0.22) | (−0.31) | (−14.52) | (−3.32) | |
| IP | $0.202^{**}$ | $-0.327^{**}$ | $-0.042^{*}$ | $-0.026$ | $-0.589^{***}$ | $-0.379^{***}$ | 0.068 |
| | (2.57) | (−2.75) | (−1.91) | (−1.19) | (−31.27) | (−4.59) | |

Note: *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively; Standard errors clustered by year. T-stats in parenthesis.

# 7 Final remarks

According to Huang & Li (2011) and Loughran & McDonald (2016) several algorithms have been developed for labeling, extract and quantify textual information reported in corporate financial reports. Under this matter, this research looks into the risk section from Form 10-K SEC in order to extract and classify the key companies risks into systematic and idiosyncratic based on text mining algorithm called Latent Dirichlet Allocation (LDA) and applies a logistic regression underlying the Fama and French 5 Factor Asset Pricing Model and the importance of time listing to identify which firm-quality factors are most significant to determine its risk profile.

Overall, the dominance of statements of general risk management policy and a lack of coherence in the risk narratives implies that a risk information gap exists and consequently stakeholders are unable to adequately assess the risk profile of a company. In particular, the results provide enough evidence that tax, competition, insurance, intellectual property and government risks behave similarly throughout all the industries, that is, those risks were not concentrated into specific industries to claim that they were idiosyncratic. The least concentrated risk in an industry is the defaulting risk (DEB), where a low concentration would imply that it is a systematic factor. However, default probabilities are influenced by several firm-specific characteristics. which makes it also linked to both micro fundamentals and macroeconomic information.

Moreover, market equity, total assets and firm's age are the predictive factors of risks, and the firms for which the risk is captured are smaller on average, present lower market equity and total assets, besides been younger and slightly less profitable when compared to traditional firms, although presents similar book-to-market ratios and investment rates. This suggests that even if big and consolidate firms disclosure their risks poorly, investors still places a higher risk-aversion on less popular firms as suggested by the classical financial theory. However, investing in securities involves risk of loss that shareholders should be prepared to bear and portfolio diversification is a way to mitigate investor's exposition to firm-specific risk.

Additional work remains to be done to validate these results, where the classification of risks into systematic and idiosyncratic is highly debatable and serves as a new guidance tool of exploration into this matter. The intuition behind it was to verify if the risks extracted through LDA were concentrated into particular industries to claim that they were idiosyncratic, otherwise low concentration imply that it is a systematic risk. One direction to go is to explore the dynamic links between firm-specific variables and macroeconomics developments to evaluate how firms are willing to disclosure information thought different business cycles, hence, how investors would sort portfolios.

# References


Acemoglu, Daron, & Akcigit, Ufuk. 2012. Intellectual property rights policy, competition and innovation. *Journal of the European Economic Association*, **10**(1), 1–42.

Ang, Andrew, Hodrick, Robert J, Xing, Yuhang, & Zhang, Xiaoyan. 2006. The cross-section of volatility and expected returns. *The Journal of Finance*, **61**(1), 259–299.

Barberis, Nicholas, & Huang, Ming. 2008. Stocks as lotteries: The implications of probability weighting for security prices. *American Economic Review*, **98**(5), 2066–2100.

Blei, David M. 2012. Probabilistic topic models. *Communications of the ACM*, **55**(4), 77–84.

Blei, David M, Ng, Andrew Y, & Jordan, Michael I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, **3**(Jan), 993–1022.

Bonfim, Diana. 2009. Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics. *Journal of Banking & Finance*, **33**(2), 281–299.

Brandt, Michael W, Brav, Alon, Graham, John R, & Kumar, Alok. 2009. The idiosyncratic volatility puzzle: Time trend or speculative episodes? *The Review of Financial Studies*, **23**(2), 863–899.

Brooks, Chris. 2019. *Introductory econometrics for finance*. Cambridge university press.

Brown, Gregory, & Kapadia, Nishad. 2007. Firm-specific risk and equity market development. *Journal of Financial Economics*, **84**(2), 358–388.

Campbell, John L, Chen, Hsinchun, Dhaliwal, Dan S, Lu, Hsin-min, & Steele, Logan B. 2014. The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies*, **19**(1), 396–455.

Das, Sanjiv R, & Chen, Mike Y. 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management science*, **53**(9), 1375–1388.

Elton, Edwin J, Gruber, Martin J, Brown, Stephen J, & Goetzmann, William N. 2009. *Modern portfolio theory and investment analysis*. John Wiley & Sons.

Fama, Eugene F, & French, Kenneth R. 2015. A five-factor asset pricing model. *Journal of financial economics*, **116**(1), 1–22.

Fisher-Vanden, Karen, & Thorburn, Karin S. 2011. Voluntary corporate environmental initiatives and shareholder wealth. *Journal of Environmental Economics and management*, **62**(3), 430–445.

Fu, Fangjian. 2009. Idiosyncratic risk and the cross-section of expected stock returns. *Journal of financial Economics*, **91**(1), 24–37.

Han, Yufeng, & Lesmond, David. 2011. Liquidity biases and the pricing of cross-sectional idiosyncratic volatility. *The Review of Financial Studies*, **24**(5), 1590–1629.

Hansen, Stephen, McMahon, Michael, & Prat, Andrea. 2017. Transparency and deliberation within the FOMC: a computational linguistics approach. *The Quarterly Journal of Economics*, **133**(2), 801–870.

Harvey, Campbell R, Liu, Yan, & Zhu, Heqing. 2016. ... and the cross-section of expected returns. *The Review of Financial Studies*, **29**(1), 5–68.

Hirshleifer, David. 1988. Residual risk, trading costs, and commodity futures risk premia. *The Review of Financial Studies*, **1**(2), 173–193.

Hope, Ole-Kristian, Hu, Danqi, & Lu, Hai. 2016. The benefits of specific risk-factor disclosures. *Review of Accounting Studies*, **21**(4), 1005–1045.

Hou, Kewei, & Loh, Roger K. 2016. Have we solved the idiosyncratic volatility puzzle? *Journal of Financial Economics*, **121**(1), 167–194.

Hou, Kewei, Xue, Chen, & Zhang, Lu. 2017. *Replicating anomalies*. Tech. rept. National Bureau of Economic Research.

Huang, Ke-Wei, & Li, Zhuolun. 2011. A multilabel text classification algorithm for labeling risk factors in SEC form 10-K. *ACM Transactions on Management Information Systems (TMIS)*, **2**(3), 18.

Jiang, George J, Xu, Danielle, & Yao, Tong. 2009. The information content of idiosyncratic volatility. *Journal of Financial and Quantitative Analysis*, **44**(1), 1–28.

Jorge, Maria João da Silva, & Augusto, Mário António Gomes. 2011. Financial risk exposures and risk management: evidence from european nonfinancial firms. *RAM. Revista de Administração Mackenzie*, **12**(5), 65–97.

Jovanovic, Boyan, & Rousseau, Peter L. 2001. Why wait? A century of life before IPO. *American Economic Review*, **91**(2), 336–341.

Klein, David A. 2009. *The strategic management of intellectual capital*. Routledge.

Kravet, Todd, & Muslu, Volkan. 2013. Textual risk disclosures and investors' risk perceptions. *Review of Accounting Studies*, **18**(4), 1088–1122.

Kumar, B Shravan, & Ravi, Vadlamani. 2016. A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, **114**, 128–147.

Kunreuther, Howard. 2002. Risk analysis and risk management in an uncertain world 1. *Risk Analysis: An International Journal*, **22**(4), 655–664.

Laurini, Márcio Poletti, Mauad, Roberto Baltieri, & Aiube, Fernando Antônio Lucena. 2020. The impact of co-jumps in the oil sector. *Research in International Business and Finance*, **52**, 101197.

Li, Feng. 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics*, **45**(2-3), 221–247.

Loughran, Tim, & McDonald, Bill. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, **54**(4), 1187–1230.

Merton, Robert C. 1987. A simple model of capital market equilibrium with incomplete information. *The journal of finance*, **42**(3), 483–510.

Muralidhar, K. 2010. Enterprise risk management in the Middle East oil industry. *International Journal of Energy Sector Management*.

Nann, Stefan, Krauss, Jonas, & Schoder, Detlef. 2013. Predictive analytics on public data-the case of stock markets.

Oh, Chong, & Sheng, Olivia. 2011. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. *Pages 1–19 of: Icis*. Citeseer.

Rubinstein, Mark. 2002. Markowitz's "Portfolio Selection": A Fifty-Year Retrospective. *The Journal of finance*, **57**(3), 1041–1045.

Wong, Peter. 2011. *Earnings Shocks and the Idiosyncratic Volatility Anomaly in the Cross-Section of Stock Returns*. Tech. rept. Working Paper. The Ohio State University.

Wu, Yanhui, Gaunt, Clive, & Gray, Stephen. 2010. A comparison of alternative bankruptcy prediction models. *Journal of Contemporary Accounting & Economics*, **6**(1), 34–45.

# Appendix

Table 5: Risk correlations Matrix

| Risk | EMP | ENV | COM | OIL | TAX | RES | SC | CT | POL | INT | DEB | IT | REP | ACQ | GOV | INS | STM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ENV | 0.09 | | | | | | | | | | | | | | | | |
| COM | 0.20 | -0.01 | | | | | | | | | | | | | | | |
| OIL | -0.04 | 0.03 | -0.12 | | | | | | | | | | | | | | |
| TAX | 0.10 | 0.08 | 0.14 | -0.02 | | | | | | | | | | | | | |
| RES | 0.08 | 0.02 | 0.06 | 0.05 | 0.07 | | | | | | | | | | | | |
| SC | 0.14 | 0.15 | 0.24 | -0.08 | 0.11 | 0.03 | | | | | | | | | | | |
| CT | 0.15 | 0.12 | -0.13 | -0.07 | -0.05 | -0.00 | -0.08 | | | | | | | | | | |
| POL | 0.10 | 0.06 | -0.02 | -0.06 | 0.08 | 0.00 | -0.06 | 0.30 | | | | | | | | | |
| INT | 0.09 | 0.05 | 0.26 | -0.06 | 0.16 | 0.05 | 0.24 | -0.08 | -0.04 | | | | | | | | |
| DEB | 0.12 | 0.16 | -0.01 | 0.10 | 0.13 | 0.05 | 0.07 | -0.06 | 0.09 | 0.01 | | | | | | | |
| IT | 0.10 | -0.02 | 0.16 | -0.09 | 0.10 | 0.04 | 0.00 | -0.14 | 0.02 | 0.07 | 0.05 | | | | | | |
| REP | 0.29 | 0.04 | 0.20 | -0.06 | 0.08 | 0.08 | 0.07 | 0.14 | 0.10 | 0.09 | 0.09 | 0.12 | | | | | |
| ACQ | 0.16 | 0.08 | 0.22 | -0.06 | 0.17 | 0.04 | 0.16 | -0.07 | 0.06 | 0.18 | 0.12 | 0.12 | 0.14 | | | | |
| GOV | 0.26 | 0.05 | 0.11 | 0.01 | 0.07 | 0.08 | 0.01 | 0.25 | 0.16 | -0.00 | 0.13 | 0.05 | 0.33 | 0.09 | | | |
| INS | 0.09 | 0.09 | 0.01 | 0.00 | 0.04 | 0.03 | 0.08 | 0.05 | 0.09 | 0.02 | 0.04 | -0.00 | 0.02 | 0.03 | 0.01 | | |
| STM | 0.14 | 0.06 | -0.00 | 0.04 | 0.04 | 0.10 | -0.04 | 0.22 | 0.10 | -0.04 | 0.06 | 0.02 | 0.20 | -0.00 | 0.22 | 0.03 | |
| IP | 0.32 | 0.05 | 0.38 | -0.22 | 0.16 | 0.03 | 0.17 | 0.31 | 0.18 | 0.19 | 0.01 | 0.11 | 0.31 | 0.19 | 0.31 | 0.00 | 0.13 |

Table 6: Risks reporting by year: This table shows the percentage of firms that report the risk each year

| | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EMP | 59.05% | 60.55% | 59.62% | 59.33% | 60.11% | 60.17% | 61.26% | 63.23% | 64.65% | 64.05% | 65.21% | 61.57% |
| ENV | 27.37% | 27.94% | 28.16% | 28.72% | 29.60% | 29.83% | 30.45% | 31.59% | 33.30% | 34.19% | 34.79% | 30.54% |
| COM | 37.93% | 36.76% | 37.17% | 37.46% | 36.81% | 35.76% | 34.53% | 34.23% | 34.39% | 33.98% | 34.43% | 35.77% |
| OIL | 7.33% | 8.13% | 8.51% | 8.83% | 8.53% | 9.39% | 10.06% | 10.18% | 10.61% | 10.68% | 10.08% | 9.30% |
| TAX | 18.34% | 21.90% | 26.20% | 30.45% | 32.72% | 34.85% | 36.52% | 40.11% | 42.48% | 45.32% | 46.06% | 34.09% |
| RES | 9.03% | 7.89% | 6.51% | 6.80% | 6.51% | 6.49% | 6.65% | 6.67% | 6.91% | 7.22% | 6.38% | 7.01% |
| SC | 47.29% | 47.75% | 47.52% | 48.52% | 49.34% | 50.22% | 48.63% | 47.61% | 46.96% | 46.63% | 47.24% | 47.97% |
| CT | 10.60% | 11.63% | 10.43% | 10.22% | 9.76% | 9.74% | 10.37% | 11.72% | 13.78% | 15.19% | 15.96% | 11.76% |
| POL | 19.71% | 20.81% | 21.78% | 23.94% | 25.42% | 26.67% | 27.97% | 30.32% | 32.74% | 34.37% | 37.07% | 27.35% |
| INT | 30.68% | 30.72% | 33.67% | 33.40% | 33.51% | 33.42% | 32.23% | 29.97% | 30.04% | 30.52% | 29.27% | 31.58% |
| DEB | 37.64% | 39.90% | 43.72% | 45.14% | 45.65% | 47.97% | 49.51% | 51.78% | 54.96% | 57.71% | 59.64% | 48.51% |
| IT | 20.95% | 21.54% | 21.57% | 23.18% | 23.04% | 24.29% | 26.20% | 27.64% | 28.04% | 27.10% | 26.13% | 24.52% |
| REP | 37.81% | 39.94% | 40.18% | 39.27% | 39.71% | 39.74% | 40.47% | 42.87% | 45.96% | 48.56% | 50.16% | 42.24% |
| ACQ | 51.76% | 52.66% | 51.69% | 52.79% | 52.81% | 51.82% | 51.77% | 53.80% | 55.61% | 56.57% | 56.91% | 53.47% |
| GOV | 34.99% | 37.44% | 37.38% | 37.96% | 38.48% | 38.14% | 39.85% | 41.90% | 45.04% | 47.02% | 48.47% | 40.61% |
| INS | 12.01% | 11.47% | 11.31% | 11.02% | 10.69% | 11.30% | 10.64% | 11.19% | 11.83% | 12.04% | 12.13% | 11.42% |
| STM | 10.72% | 12.20% | 12.22% | 13.30% | 13.06% | 12.86% | 13.65% | 16.45% | 19.13% | 19.70% | 19.84% | 14.83% |
| IP | 48.03% | 50.08% | 50.31% | 51.14% | 51.54% | 52.77% | 53.59% | 54.94% | 57.17% | 59.02% | 60.47% | 53.55% |
| Total | 2415 | 2484 | 2397 | 2368 | 2274 | 2310 | 2256 | 2279 | 2300 | 2284 | 2193 | 2323.64 |

Table 7: Summary statistics for the firms in 2006 grouped by risk.

| Risk | | ME | TA | IA | Prof | Age | BM | Count |
|---|---|---|---|---|---|---|---|---|
| EMP | Risk | 2074.11 | 1324.40 | 0.28 | 0.02 | 13.91 | 0.39 | 1426 |
| | | 8628.93 | 4668.03 | 0.99 | 2.61 | 11.95 | 1.50 | |
| | No Risk | 7142.30 | 6328.37 | 0.95 | 0.76 | 23.82 | 0.48 | 989 |
| | | 27042.65 | 29669.99 | 19.19 | 13.35 | 19.24 | 0.42 | |
| ENV | Risk | 4098.64 | 3300.49 | 0.20 | 0.13 | 19.52 | 0.45 | 661 |
| | | 14139.90 | 12686.12 | 0.58 | 1.68 | 18.09 | 0.41 | |
| | No Risk | 4168.88 | 3401.21 | 0.69 | 0.40 | 17.39 | 0.41 | 1754 |
| | | 20148.80 | 21485.01 | 14.43 | 10.25 | 15.27 | 1.36 | |
| COM | Risk | 2894.51 | 1666.78 | 0.25 | 0.05 | 15.55 | 0.41 | 916 |
| | | 11750.16 | 6427.46 | 0.77 | 2.90 | 13.53 | 0.56 | |
| | No Risk | 4916.64 | 4416.66 | 0.74 | 0.49 | 19.45 | 0.43 | 1499 |
| | | 21846.93 | 24145.18 | 15.61 | 10.91 | 17.35 | 1.44 | |
| OIL | Risk | 4545.66 | 4749.95 | 0.43 | 0.85 | 22.55 | 0.50 | 177 |
| | | 11895.97 | 14418.24 | 0.86 | 8.29 | 19.64 | 0.29 | |
| | No Risk | 4118.33 | 3264.79 | 0.57 | 0.28 | 17.61 | 0.42 | 2238 |
| | | 19130.60 | 19816.65 | 12.78 | 8.82 | 15.75 | 1.23 | |
| TAX | Risk | 5106.31 | 3773.71 | 0.20 | 0.51 | 17.41 | 0.47 | 443 |
| | | 17231.44 | 15917.05 | 0.48 | 5.72 | 15.39 | 0.44 | |
| | No Risk | 3934.75 | 3283.77 | 0.64 | 0.28 | 18.09 | 0.41 | 1972 |
| | | 19004.89 | 20188.74 | 13.62 | 9.33 | 16.27 | 1.29 | |
| RES | Risk | 1892.60 | 1070.72 | 0.26 | 0.62 | 14.70 | 0.47 | 218 |
| | | 5002.20 | 2946.64 | 0.78 | 7.55 | 12.97 | 0.47 | |
| | No Risk | 4373.61 | 3602.15 | 0.59 | 0.30 | 18.29 | 0.42 | 2197 |
| | | 19522.80 | 20381.87 | 12.90 | 8.89 | 16.36 | 1.23 | |
| SC | Risk | 3071.73 | 2060.93 | 0.25 | 0.09 | 17.98 | 0.47 | 1142 |
| | | 11436.81 | 6240.04 | 0.73 | 2.61 | 15.97 | 0.39 | |
| | No Risk | 5116.65 | 4551.27 | 0.84 | 0.53 | 17.96 | 0.38 | 1273 |
| | | 23322.20 | 26111.41 | 16.93 | 11.83 | 16.25 | 1.58 | |
| CT | Risk | 2145.08 | 949.97 | 0.44 | -0.25 | 11.03 | 0.24 | 256 |
| | | 10032.73 | 4586.43 | 1.76 | 3.02 | 8.77 | 0.27 | |
| | No Risk | 4387.34 | 3661.03 | 0.57 | 0.39 | 18.79 | 0.45 | 2159 |
| | | 19455.22 | 20516.64 | 13.00 | 9.22 | 16.58 | 1.25 | |
| POL | Risk | 3630.26 | 2523.57 | 0.30 | -0.07 | 14.43 | 0.38 | 476 |
| | | 10168.06 | 8643.82 | 1.29 | 3.97 | 14.38 | 0.39 | |
| | No Risk | 4277.16 | 3582.32 | 0.62 | 0.42 | 18.84 | 0.43 | 1939 |
| | | 20245.70 | 21303.02 | 13.72 | 9.59 | 16.40 | 1.31 | |
| INT | Risk | 4297.29 | 2846.03 | 1.27 | 0.21 | 18.33 | 0.43 | 741 |
| | | 14435.45 | 9701.71 | 22.17 | 1.88 | 15.57 | 0.38 | |
| | No Risk | 4084.30 | 3607.19 | 0.24 | 0.38 | 17.81 | 0.42 | 1674 |
| | | 20299.71 | 22480.25 | 0.93 | 10.47 | 16.35 | 1.40 | |
| DEB | Risk | 2191.88 | 2095.43 | 0.24 | 0.30 | 16.42 | 0.39 | 909 |

Continued on next page

Table 7 – continued from previous page

| Risk | | ME | TA | IA | Prof | Age | BM | Count |
|---|---|---|---|---|---|---|---|---|
| | | 5039.50 | 4659.36 | 0.63 | 5.04 | 15.54 | 1.83 | |
| | No Risk | 5331.34 | 4145.15 | 0.75 | 0.34 | 18.91 | 0.44 | 1506 |
| | | 23270.56 | 24363.32 | 15.57 | 10.41 | 16.39 | 0.47 | |
| IT | Risk | 3678.08 | 2659.36 | 0.61 | 0.08 | 14.55 | 0.43 | 506 |
| | | 13496.22 | 9385.30 | 9.52 | 6.12 | 13.39 | 0.51 | |
| | No Risk | 4274.65 | 3562.97 | 0.54 | 0.39 | 18.88 | 0.42 | 1909 |
| | | 19846.19 | 21361.05 | 12.95 | 9.36 | 16.65 | 1.30 | |
| REP | Risk | 1812.41 | 1211.47 | 0.51 | -0.08 | 13.16 | 0.43 | 913 |
| | | 8159.40 | 5322.42 | 7.11 | 2.65 | 10.74 | 0.56 | |
| | No Risk | 5570.36 | 4687.93 | 0.59 | 0.57 | 20.89 | 0.42 | 1502 |
| | | 22721.22 | 24250.53 | 14.59 | 10.93 | 18.02 | 1.43 | |
| ACQ | Risk | 3282.81 | 2481.89 | 0.25 | 0.18 | 16.59 | 0.42 | 1250 |
| | | 11658.48 | 9190.01 | 0.73 | 4.16 | 15.01 | 1.55 | |
| | No Risk | 5079.75 | 4330.45 | 0.89 | 0.48 | 19.45 | 0.43 | 1165 |
| | | 24026.08 | 26344.80 | 17.70 | 11.88 | 17.11 | 0.57 | |
| GOV | Risk | 1617.88 | 1098.33 | 0.98 | -0.02 | 11.97 | 0.39 | 845 |
| | | 6214.26 | 4570.71 | 19.44 | 1.13 | 10.63 | 0.54 | |
| | No Risk | 5512.30 | 4598.25 | 0.33 | 0.51 | 21.20 | 0.44 | 1570 |
| | | 22618.70 | 23830.95 | 5.43 | 10.85 | 17.57 | 1.41 | |
| INS | Risk | 3200.89 | 1936.95 | 0.20 | 0.18 | 16.29 | 0.39 | 290 |
| | | 11753.03 | 5545.05 | 0.46 | 2.00 | 13.90 | 0.76 | |
| | No Risk | 4279.13 | 3569.71 | 0.61 | 0.34 | 18.20 | 0.43 | 2125 |
| | | 19448.81 | 20651.49 | 13.12 | 9.33 | 16.38 | 1.23 | |
| STM | Risk | 3483.05 | 2532.31 | 2.66 | 0.05 | 12.49 | 0.35 | 259 |
| | | 17741.17 | 17243.68 | 35.10 | 8.33 | 13.28 | 0.81 | |
| | No Risk | 4229.73 | 3474.71 | 0.30 | 0.36 | 18.63 | 0.43 | 2156 |
| | | 18807.85 | 19725.48 | 4.64 | 8.83 | 16.30 | 1.22 | |
| IP | Risk | 2550.15 | 1389.19 | 0.29 | 0.29 | 13.69 | 0.37 | 1160 |
| | | 10554.54 | 5394.92 | 1.08 | 12.10 | 11.86 | 0.50 | |
| | No Risk | 5628.08 | 5207.88 | 0.80 | 0.36 | 21.93 | 0.48 | 1255 |
| | | 23774.79 | 26383.49 | 17.04 | 3.61 | 18.36 | 1.57 | |

Note: ME and TA are expressed in millions of U.S. dollars.