Development of a Data Mining Software for the Application of Cluster Analysis of Companies and Industries listed on NASDAQ

Jan L. Schroeder | 12176855

Abstract — The literature review of over 50 journal papers focusing on financial ratio analysis in the broadest sense identifies the following gap: no researcher applied a cluster analysis to gain understanding of cross-industry financial performance patterns measured with financial ratios and based on a data set that is statistically significant. As a result, this paper describes the development of a data mining software to extract financial data of 3282 companies listed on NASDAQ, the second largest stock exchange in the world with regard to its market capitalisation. The non-hierarchical cluster algorithm k-means is used to cluster industries with similar performance characteristics. Major conclusions are: the amount of inventory influences industries' performance measured by current ratio the most. Service industries are found to be characterised by low current ratios. Banks and investment industries represent the highest gross margins. ROI and ROE seem to be unsuitable for cross-industry performance comparisons.

Keywords — Financial Ratio Analysis, Accounting Research, Cluster Analysis, k-means, DBSCAN, NASDAQ

I. Introduction

The following work contributes to the field of data analysis, accounting research and industry performance classification.

An extensive literature review based on over 50 journal papers is conducted to acknowledge the gap in literature and to derive the eight widely cited financial ratios and several best practises. As identified, the literature lacks information, analysis and comprehension of cross-industry financial performance characteristics. Furthermore, no research is identified that analyses large-scale data sets consisting of financial information about companies listed on US stock exchanges, such as NASDAQ or NYSE. Especially, the lack of access to large-scale data sets is mentioned several times in the literature as a limitation to research.

To fill the gap in the literature, a data mining software is developed to extract income statements and balance sheets of 3282 companies listed on the second largest stock exchange in the world with regard to its market capitalisation, NASDAQ. The companies are grouped into 119 industries by using the NASDAQ industry classification tree so that the arithmetic mean of each ratio per industry is derived. Thus, the financial performance of every industry is characterised by eight average financial ratios. Ten clusters for each ratio are identified by applying the non-hierarchical cluster algorithm k-means, i.e. all industries per financial ratio are grouped in terms of their performance similarity. Then, the arithmetic mean and standard deviation is calculated for every cluster for every financial ratio. The two clusters with the highest and the two clusters with the lowest average performance are selected for each ratio to draw conclusions and to identify patterns.

The conclusions and outcomes of this research generate an added value due to several reasons: group industry average performance indicators can serve companies as performance targets or can be used for the analysis of competitors' performance (Maricica & Georgeta 2012). According to Gupta and Huefner (1972), 'reference to a group average may also enable a firm to evaluate its own ratios against several industries having similar characteristics'. The evaluation of governmental proposals, such as investment tax credits as a stimulant to the economy might consider outcomes found in this research. Comprehension about the behaviour of turnover ratios might be useful in assessing the effects of changes in underlying economic characteristics. In addition, 'an industry that anticipates increased vertical integration might employ capital-output relationships to estimate the additional investment necessary to maintain its level of sales' (Rushinek & Rushinek 1987). An underdeveloped country might use a similar approach to the one applied in this research to estimate the capital needed to support a given industry.

The paper is divided into 5 sections: introduction, literature review, research method, results and discussions and conclusions.

II. Literature Review

Over 50 journal papers dealing with financial ratios in the broadest sense are critically reviewed to clearly highlight the gap in the literature. The relevance of my selected approaches is supported by the findings in the literature.

The purpose and the area of application of financial ratio analysis are found to vary significantly, i.e. it is combined with various statistical analysis approaches in various industries within different countries.

The discussion is structured as follows: a general review is given before research papers focused on the following regions and countries are examined: China, Europe (cross-country analysis), Greek, Iran, Malaysia, Romania, Taiwan, Turkey, UK, United Arab Emirates and USA. A final summary precisely points out the gap in the literature.

General

As indicated by Horrigan (1968), financial ratios emerged between 1900 and 1919 and evolved over 100 years to one of the most impactful financial measurement approaches worldwide (Maricica & Georgeta 2012).

Altman (1968) was the first researcher found in the literature who extended financial ratio analysis with discriminant analysis to predict corporate bankruptcy with an accuracy of 94% two years before the bankruptcy occurs. Financial ratio analysis can be used for business credit evaluation and assessment of solvency, or predicting corporate bond ratings (Beaver 1968; Gupta & Huefner 1972; Rushinek & Rushinek 1987).

Analysts might make judgements based on clustered multi-industry samples, e.g. they might compare a company's financial ratio with an economy-wide or stock-market wide ratio (Sudarsanam & Taffler 1995). Therefore, the results of my work might be used for this purpose.

Sim et al. (2011) clustered stocks that are characterised by homogenous financial ratio values across years to study afterwards their price movements, i.e. they calculated the correlation between financial ratios and price movements. The method of clustering financial ratios is similar to the selected approach in my paper. However, TRICLUSTER and Cross-Graph Quasi-Bicliques are used as clustering approaches. Both approaches are suitable to conduct a cluster analysis within a multidimensional space. Thus, future research can extend my one-dimensional approach with k-means, so that an analysis of an eight dimensional space can be conducted. The research method by Sim et al. (2011) consists of three main phases which are similar to my selected research method, i.e. starting with data preparation, continue with data mining so that afterwards data analysis can be conducted. Financial figures of stocks were downloaded from Compustat for the years 2000 to 2006. However, Compustat only offered limited data sets, so that the performance evaluation only included 9 Chemical industries. Thirty-two financial ratios were used based on the formulae from Investopedia (Pinkasovitch 2009).

Furthermore, financial ratios are used by partners in supply chains in revenue-sharing and savings-sharing contracts and in the design of supply chain networks to handle demand uncertainty (Longinidis & Georgiadis 2011). Mixed-Integer Linear Programming based on financial ratios was used to develop a model that integrated financial aspects with design decisions for the supply chain. Twelve ratios classified into the following groups were selected: liquidity ratios, asset management ratios, solvency ratios and profitability ratios. However, no activity ratios are used. Another critical point is that their discussions only consider the calculated ratios for one company.

Niemann, Schmidt & Neukirchen (2008) developed a methodology for measuring and managing the heterogeneity in financial ratios which can be used for corporate ratings, for example by Standards & Poor. Their motivation for the development was that ‘industry sectors can differ substantially in terms of balance sheet structure’, i.e. the main problem is that ‘companies with the same rating coincides with substantially different value of financial ratios across different subgroups, vice versa’. Niemann, Schmidt & Neukirchen (2008) do not mention any data supporting their statement. The analysis of financial ratio heterogeneity considered 9 industries and 9 financial ratios from the period 1998 to 2002. This is one of two research projects found that used data samples consisting of more than 2000 companies from the Compustat/GlobalVantage database. As future research outlook, they mentioned the importance of industry clustering in terms of performance, i.e. my paper can be seen as an extension to their work.

McIvor et al. (2004) outlined the importance of ratio analysis in the corporate acquisition process. The goal of their paper was the development of a fuzzy approach to support analysts by identifying take-over targets, i.e. companies worth to acquire. McIvor et al. (2004) stated the importance of computerised expert systems for measuring financial ratios of companies worldwide. Thus, a further motivation for my project is given. The analysis of financial profiles in the acquisition process includes ratios associated with liquidity, profitability, gearing and efficiency. My selected approach does not consider gearing and efficiency ratios because McIvor et al. (2004) clearly stated that such ratios are not suitable for industry comparisons. Twelve ratios were implemented in the proposed system to measure the level of acquisition potential. However, the paper does not deal with the issue of data mining, i.e. no suggestion is given where the necessary financial data should be gained from. Obtaining the necessary data to start the proposed analysis is crucial. Thus, a further motivation for the development of a data mining software for financial data is given.

Another motivation for my research is given by Pinches et al. (1975) explaining the importance of gaining an understanding of the performance of industries so that firms can adjust their financial ratios towards the industry average. Pinches et al. (1975) derived 48 financial ratios to identify classifications in the set of ratios with factor analysis. Based on this work, the financial ratios selected for my work were chosen.

Gutierrez & Carmona (1988) applied fuzzy set theory in the context of financial ratio analysis with respect to liquidity measurement. Their developed model is used for decision making within single companies and not focused on entire industries. The data source necessary to use the model is not presented and discussed.

A further possible use case for my results is found in the research of Lewellen (2004) because strong evidence was found that market returns are predictable by using financial ratios. Thus, the results generated in my work can be used to predict stock returns.

As the literature review identified over 10 researchers from different countries worldwide dealing with the prediction of business and industry failure, future research can analyse the extracted data for my project by starting with the approach suggested by Sharma & Mahajan (1980).

Martikainen et al. (1995) discovered in 10 financial ratios distribution irregularities affecting the results concerning the classification of firms and industries. Due to this reason, a time-series analysis was not conducted in my work because the time-series instability can only be removed by transforming the ratios. This might be done in future research.

As stated by Wang & Lee (2008), it is impractical to consider more than 10 financial ratios for a cross-industry analysis. Thus, my work is based on 8 financial ratios resulting from the comparison of all financial ratios used in the identified literature, i.e. all 8 financial ratios are included in the majority of all found performance measuring approaches for various industries.

China

A principal component analysis based on financial ratios is used by Li & Zhang (2011) encompassing the real estate industry in China. The analysis only considers 12 real estate companies listed on the Chinese stock market and includes 10 financial ratios while using SPSS. It is not transparent what ratio formulae are used. In addition, the data mining and data analysis approach are not presented. For the purpose of guiding benchmarking activities, pricing decisions and regulatory monitoring, Avkiran (2011) investigated the financial performance of 19 Chinese banks by applying Data Envelopment Analysis (DEA) based on 13 financial ratios. No conclusions are presented to comprehend performance characteristics of industry clusters. Seng & Lai (2010) are the only researchers found in the literature who developed a similar data mining approach to the approach developed in my project. However, their system is focused on the Chinese stock market and not on NASDAQ. In addition, they implemented an all-encompassing Business Valuation Model allowing selecting between different valuation strategies.

A financial ratio analysis in combination with the soft set theory to predict business failure in the Chinese market is a further example of the variety of applications (Xu et al. 2014). Nine ratios are considered to forecast businesses' performance with vector machines and neural networks in SPSS and MATLAB. However, no clustering approach was used for the 240 firms listed on the Chinese stock exchange. Xu et al. (2014) do not describe how they gained and cleaned the data set. This is the most crucial part for data-driven research projects, as the quality of the results highly depends on the quality of the gained data.

Europe

Gallizo, Gargallo & Salvador (2008) applied Bayesian analysis to derive relationships between 12 financial ratios. The outcome of their research was also considered during the discussion of the results in my work. According to Gallizo, Gargallo & Salvador (2008), more highly leveraged firms tend to increase their profitability and efficiency, although their liquidity declines. Data for the European manufacturing industry was gained from the AMADEUS database, so that income statements and balance sheet for 395 firms were analysed. However, neither a prediction approach nor cluster analysis was conducted.

The importance of financial analysis of diversification and similarity with financial ratios is also stated by Cinca, Molinero & Larraz (2005). Their research was supported by the European Regional Development Fund to evaluate the performance of 11 European countries with regards to their manufacturing industries by using 15 financial ratios. Financial data for 14 years was obtained from the BACH database containing financial information about all industries in 11 European countries. The hierarchical cluster algorithm by Ward's was used in SPSS to develop a dendrogram visualising the clusters based on the calculated financial ratios. The results present three options to organise manufacturing companies: a Latin one, a Scandinavian one, and a Germanic one. Considering those findings supported the analysis of the clusters generated in my work. However, the comparison of Ward's method with DBSCAN clearly gave the indication that DBSCAN shows better performance in noisy data sets. Thus, DBSCAN was selected as the counterpart of k-means.

The importance and potential of cross-country measurements based on financial ratios were also indicated by Ou & Penman (1989) and Yli-Olli & Virtanen (1989), resulting in a further motivation for my work. Gallizo, Jiménez & Salvador (2003) utilised financial ratio analysis in combination with the partial adjustment model to draw the conclusion that debt ratios show the least sensitivity to different types of shocks. Considering this conclusion, debt ratios were not included in my research. The researchers present another application of financial ratios, i.e. determining the sensitivity of a country to external financial shocks and comparing the level of sensitivity for countries. Ten financial ratios were derived for 7 European countries: Austria, Denmark, France, Germany, Italy, Netherlands and Spain. Future research projects can combine my outcomes with the results of Gallizo, Jiménez & Salvador (2003) to gain further insights in terms of a cross-country perspective.

Gallizo, Jiménez & Salvador (2002) also draw the conclusion that the sensitivity of industries can be determined by applying the Bayesian hierarchical model. This approach can be used by future research based on the data sets generated in my work.

Greek

DEA combined with financial ratios analysis to evaluate the performance of 23 Greek manufacturing sectors was used by Halkos & Tzeremes (2012). The data was provided by ICAP Business Information Services, an online service offering financial data of Greek companies. Over 4000 companies were analysed to develop rankings with respect to 6 financial ratios. This is the only work found in the literature that considers more companies than my work. However, no similar classifications approaches were applied. Halkos & Salamouris (2004) also used the DEA approach in combination with 7 financial ratios to measure the efficiency of Greek commercial banks from 1997 to 1990. However, the data set was limited to 17 banks and no clustering was conducted.

Dimitras et al. (1999) objective was the prediction of business success and failure for 40 Greek firms from 13 industries by applying a rough set approach instead of discriminant analysis used by Deakin (1976). Twenty-nine ratios were considered.

Iran

Shaverdi, Heshmati & Ramezani (2014) used financial ratios to measure the performance of the Iranian petrochemical sector. The fuzzy set theory was combined with 15 ratios classified into growth, profitability, activity, financial leverage and liquidity. Shaverdi, Heshmati & Ramezani (2014) count as the rare researchers found that defined all ratios so that calculations are transparent and reproducible. However, no information is given about the data mining and data cleaning process.

Malaysia

The performance of the Malaysian industry was measured by Zulkifli (2010). Major objectives were the identification of correlations between financial ratios and the analysis of Malaysian financial trends from 2000 to 2008. As the analysed data set only consisted of 40 companies, a statistical significance is not found.

Romania

Maricica & Georgeta (2012) analysed the potential of financial ratios to predict business failure of Romanian companies. Financial data for 63 companies listed on the Bucharest Stock Exchange were manually collected. The lack of sufficient data was mentioned as a major problem. T-tests were used to test the significance of the difference in means for companies classified as 'likely to fail' and 'not likely to fail'. A clear approach to cluster companies into the two groups is not presented.

Taiwan

The research of Huang et al. (2008) evaluates the financial performance of 660 enterprises listed on the Taiwanese stock exchange while using 21 financial ratios. However, no in-depth clustering was conducted, as enterprises were only associated with two groups: 'fine' and 'risk'. In contrast to the approach selected for my work, a back-propagation neural network was used as clustering approach.

Kung & Wen (2007) analysed 20 venture capital enterprises in Taiwan based on the financial data from 2001 and 2003 to find significant financial ratios affecting the financial performance most. Again, the manually extracted data set from the stock exchange is too small to gain conclusions that are statistical valid. Wang (2009) combined grey relation analysis with fuzzy multi-criteria group decision to evaluate the financial performance of Taiwanese container lines. Wang (2008) also evaluated the financial performance of 3 airlines in Taiwan while considering the outcomes of Feng & Wang (2000). The objective was similar to the objective of my project: partition financial ratios into several clusters to find representative characteristics. The approach used by Wang (2009) only included one industry and the small number of 3 analysed container lines and 3 airlines is not suitable to draw general conclusions.

Turkey

Boyacioglu, Kara & Baykan (2009) describe another common application for financial ratio analysis applied in Turkey: prediction of bank failures. The prediction of failures can be comprehended as classification problem, i.e. healthy banks and non-healthy ones have to be clustered into two groups. Twenty financial ratios from six groups are used: capital adequacy, asset quality, management quality, earnings, liquidity and sensitivity to market risk. The lack of profitability and activities ratios is seen as critical. Boyacioglu, Kara & Baykan (2009) consider artificial neural networks and multivariate statistical methods, such as multivariate discriminant analysis and k-means, which emphasises the applicability of k-means for performance clustering. Multi-layer perceptron and learning vector quantisation can be applied in future research based on my extracted raw data, as both methods are stated as most successful with regard to the prediction of financial failure. Instead of KNIME, SPSS and MATLAB were used. The lack of access to large-scale data sets is mentioned so that Boyacioglu, Kara & Baykan (2009) only considered the Turkish bank sector from 1997 to 2003.

The literature contains several examples in which financial ratios are combined with discriminant analysis used for the prediction of business failure and bankruptcy risk, not only focused on banks (Altman, Haldeman & Narayanan 1977). Especially the combination of ratio analysis and artificial neural networks is getting popular (Huang et al. 2008).

A two-step performance measuring approach was used by Delen, Kuzey & Uyar (2013), consisting of exploratory factor analysis and four different decision tree algorithms. The objective of the research was the development of a prediction model for financial ratios of manufacturing industries. The analysis considered 31 ratios and was based on 2245 Turkish publicly traded companies obtained from FINNET. The problem was stated, that large parts of the data set were corrupted and not suitable for the analysis. Their findings affected the analysis of my data sets because they find out that EBIT-to-Equity and Net Profit Margin impact companies' performances the most. Both ratios are therefore considered in my work.

Further performance evaluations with 13 financial ratios of 15 Turkish cement firms listed on the Istanbul Stock Exchange was conducted by Ertuğrul & Karakaşoğlu (2009). However, no significant characteristics of the industries' performance are described as the sample data set is not statistical significant. The same ratios were used in combination with a fuzzy analytic hierarchy process and the TOPSIS method to evaluate the performance of the Turkish automotive industry. The popularity of financial performance measurement within the Turkish industry sector is also presented by (Mercan et al. 2003), who applied DEA based on 12 financial ratios. In addition, twenty-five financial ratios were used by Öcal et al. (2007) to determine trends in the Turkish construction industry. Financial data for 1997-2001 of 28 companies were manually extracted from the Istanbul Stock Exchange. No research dealt so far with the performance classification problem of Turkish industries.

UK

Sudarsanam & Taffler (1995) explained that financial ratios are used by managers to control and measure the growth of a company. They considered 24 ratios to analyse six manufacturing industries containing 500 companies in the UK. A time-series analysis based on the financial data from 1981 to 1986 was conducted. The financial data was gained from EXSTAT, a computerised database of UK financial information. They stated the problem of having data sets that are inconsistent. Similar to my approach, Sudarsanam & Taffler (1995) used the industry classification from the London Stock Exchange.

United Arab Emirates

The prediction of bankruptcy with financial ratio analysis gained also popularity in the United Arab Emirates (Al-Kassar & Soileau 2014). However, only data for six companies were used operating in the oil and manufacturing industry. The results cannot be seen as statistically significant because the sample set is too small. In addition, no activity ratios were considered.

USA

Cowen & Hoffer (1982) conducted a cluster analysis for 72 companies of the US oil industry using Dunn and Bradstreet ratios and Compustat data from the years 1966 – 1975. Cowen & Hoffer (1982) also stated the problem of missing data points in the provided files. Fourteen ratios for factor analysis and cluster analysis within SPSS were used to address the question of industry subgroups. The importance of industry clustering in terms of financial performance is mentioned, as the results can be used for capital budgeting, portfolio and stock investment analysis, and profitability studies (Lev & Sunder 1979).

Gupta and Huefner (1972) stated that industry characteristics could be determined from cluster analysis of financial ratios. Their motivation was to figure out if inter-industry differences tended to disappear as more highly aggregated ratios were used. This lead to the evaluation of financial ratios at a macro level for broad industry classes to find correlations between ratios and basic industry attributes. The study only focused on 20 US manufacturing industries, which cannot be seen as statistical significant. Another critical aspect is the small number of 4 financial ratios used. An early-stage hierarchical cluster analysis approach was applied. The so called 'minimum method' developed by Johnson outlines the disadvantage of grouping all data points in too many clusters if aborted to early. If aborted to late, only one cluster is found containing all data points. Thus, this cluster analysis approach is not suitable for the implementation in software. The study presents two major problems: the question about the validity of data and the problem of classifying appropriate industries. Both problems are solved in my work.

Watkins (2000) and Zeller, Stanko & Cleverley (1996) deal with classification patterns of US hospital financial ratios.

Their research gives information about a national database of hospital ratios, called Financial Analysis Service (FAS). FAS produces average financial ratios based on the financial data of hospitals. Twenty-one ratios were used, tailored to the financial characteristics of hospitals. All financial ratios used in my work are also included in those 21 ratios. Factor analysis was used to describe the variability between the ratios for the years 1990 to 1994. As the analysed data set is older than 20 years, the results might be obsolete today.

In addition, the stability of financial ratio groups in the US hospital industry was analysed by Chu et al. (1991). The motivation was to cluster hospitals presenting similar financial ratios. Data were gained from the Compustat database. However, as a lack of cluster analysis algorithms implemented in software at this time, no characteristic patterns were discovered.

Summary

Summarising the literature review, only five researchers used a cluster analysis approach in combination with financial ratios. However, no one applied a cluster approach to group industries with similar performance expressed by a financial ratio based on a data sample that is statistically significant. Therefore, the literature lacks a profound knowledge about the comprehension of cross-industry financial performance characteristics. Furthermore, no research is found in the literature analysing companies and industries listed on US stock exchanges, such as NASDAQ or NYSE. Especially, the lack of access to large-scale data sets is mentioned several times as a limitation to research.

The consideration of the major points described above leads to the motivation of the development of a software to extract financial data of all 3282 NASDAQ listed companies operating in 119 industries. To fill the gap in the literature in terms of the analysis of cross-industry performance evaluations, a non-hierarchical cluster algorithm is used so that potential patterns can be identified and data turned into information.

III. Research Method

This section is divided into 6 sub-sections to describe the applied research methods. The first sub-section states the financial ratios used and their formulae. An explanation about the data source and the development of the software for this project follows. The cluster analysis algorithm k-means is presented and the data analysis environment set up in KNIME. Limitations are presented at the end.

A. *Financial Ratio Analysis*

The development of financial ratios for analysing accounting statements derived from the evolution of accounting in the United States (Horrigan 1965). Financial ratios describe the financial performance of a company or industry and are calculated by using the firm's income statement, balance sheet and cash flow statement (Chen & Shimerda 1981).

The following eight financial ratios are selected with respect to the outcomes of the literature review, i.e. ratios are used that are mentioned most often in similar research projects. The formulae of the ratios are derived from Pinkasovitch (2009) and from Ross, Westerfield & Jordan (2008). Ratios from the most commonly used classifications are considered: liquidity, activity and profitability.

1) Current Ratio: An indication of a firm's ability to service its current obligations.

$$Current\ Ratio = \frac{Current\ Assets}{Current\ Liabilities}$$

2) Gross Margin: The gross margin indicates the amount of funds available to pay the firm's expenses other than its cost of sales.

$$Gross\ Margin = \frac{Gross\ Income}{Revenue}$$

3) *Net Margin:* The net margin tells the percentage of sales that remains for the shareholders: either dividends or retained earnings.

$$Net\ Margin = \frac{Net\ Income\ after\ Tax}{Revenue}$$

4) Return On Investment: The return on investment measures the effectiveness of employing the company's resources. A low ratio may indicate that the assets grow faster than sales which can result in high debt. Limitation: a heavily depreciated plant or a large amount of intangible assets causes distortion of the value.

$$ROI = \frac{EBIT}{Assets}$$

5) *Return On Equity:* The higher the return on equity, the more effective the management.

$$ROE = \frac{Net\ Income\ after\ Tax}{Equity}$$

6) *Accounts Receivable Turnover:* The higher the turnover of receivable, the shorter the time between sale and cash collection. Limitation: seasonal fluctuations are not considered.

$$Account\ Rec.Turnover = \frac{Revenue}{Acc.Rec.}$$

7) Collection Period: The average time in days that receivables are outstanding.

$$Collection\ Period = \frac{365}{Acc.Rec.Turnover}$$

8) *Asset Turnover:* A measure of how effectively assets have been used in generating revenue or how hard resources are working. Limitations: organisations with new assets are disadvantaged, as their assets have higher values than those in older companies.

$$Asset\ Turnover = \frac{Revenue}{Assets}$$

B. *Data Source*

The Australian Stock Exchange does not offer income statements and balance sheets in the form other stock exchanges world-wide do. Thus, an analysis of Australian publicly listed companies is not applicable.

As the literature also lacks a performance analysis of companies and industries listed on NASDAQ, the second largest stock exchange is used as a data source. The NASDAQ homepage provides a downloadable list of 3282 companies. The CSV file contains:

- The NASDAQ symbols, e.g. MSFT for Microsoft.
- For each company the URL to access all relevant financial data: www.nasdaq.com/symbol/msft
- The industry each company belongs to, e.g. technology.

The income statement and balance sheet for each company are accessible via attaching the following query information to the previously mentioned URL:

- financials?query=income-statement
- financials?query=balance-sheet

Thus, the financial information for Microsoft is accessible via:

- www.nasdaq.com/symbol/msft/financials?query=income-statement
- www.nasdaq.com/symbol/msft/financials?query=balance-sheet

The data for income statements and balance sheets cannot be downloaded and only be accessed by using a web client supporting the Hypertext Transfer Protocol, such as Firefox. The restriction of accessibility leads to the development of a data mining software, named web crawler.

C. Data Extraction - Web Crawlers

Two web crawlers consisting of 4210 lines of code were developed to autonomously extract the income statements and balance sheets for the 3282 companies listed on NASDAQ. Both crawlers are based on PHP 5.3 and use the library *simplehtmldom 1.5* to transfer and extract the necessary data points.

Analysing the NASDAQ Document Object Model tree resulted in the findings that HTML table-tags, such as td, are used to structure the data for the income statements and balance sheets. The selected library allows the iteration through all table-tags. Pattern recognition functions based on regular expressions are developed to identify, extract, clean and store the relevant data points in a MySQL database. After cleaning the extracted data, 29% are identified as corrupted. Thus, the list of companies reduced to 2331. The formulae for the financial ratios were implemented in a separate extension. Two further algorithms were developed that allows grouping all companies into related industries and calculating the average financial ratios and the standard deviation for each of the 119 industries.

All data points were exported into a CSV file for the cluster analysis with k-means.

D. Theoretical Data Analysis - Cluster Analysis with k-means

A cluster is defined as a group of data points with values similar to each other. The values within a cluster are homogenous, whereas clusters amongst each other are heterogeneous (Kaufman & Rousseeuw 2009).

A cluster analysis is a pattern discovery procedure which is one of the techniques covered by the umbrella term of 'data mining'. The procedure starts from the point of view, that no patterns are known at the beginning of its application. A simple example for the application of a cluster analysis is market segmentation to identify customer segments, so that specific advertising campaigns can be used to target different segments. Every cluster analysis requires a set of data points in which patterns can be identified. A method is required to measure the degree of similarity between the data points.

Two common approaches exist to find clusters: hierarchical and non-hierarchical. A hierarchical approach does not require the specification of the numbers of clusters to be found in the data set. In contrast, non-hierarchical approaches require the definition of numbers of clusters to be found. The former identifies the number of clusters on its own, whereas the latter only identifies the specified number of clusters. Hierarchical approaches are characterised by one major disadvantage: once a data point is early placed in a cluster, it gets never allocated to a different, even better cluster. Non-hierarchical cluster approaches allow iterations until every data point belongs to the cluster it best fits to.

The hierarchical cluster algorithm Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was compared with the non-hierarchical algorithm k-means. After implementing DBSCAN and k-means in KNIME (see next sub-section) and setting up a test environment with different sample data sets containing noise, DBSCAN's advantage of successfully identifying clusters in noisy environments is acknowledged. After applying DBSCAN to detect industry clusters, it was found that the accuracy strongly depends on the parametrisation. Grouping industries with a similar financial ratio with a low value range, e.g. from 0 to 10, is not possible as the majority of data points are determined as noise. Thus, k-means is used for this research.

The k-means algorithm expects two inputs:

1. Set of data points: $x_1 \ldots x_n$
2. *k* as the number of clusters to be found

The algorithm starts by placing k centroids $c_1 \ldots c_k$ at random locations in the space of the data set. Two iterations are implemented. The first iteration runs through each data point x_i to find the nearest centroid c_j. The minimal distance between x_i and c_j as the cluster centre is calculated by using the Euclidian distance:

$$d(x_i, c_j) = \sqrt{\sum_{i=1}^{d} (q_i - p_i)^2}\,,$$
$$p = (p_1, p_2, \ldots p_n), q = (q_1, q_2, \ldots q_n), d = \text{dimension}$$

Thus, the point x_i is assigned to the nearest cluster *j*. The second iteration runs through each cluster $j=1 \ldots k$. The position of each centroid c_j is recomputed by calculating the arithmetic mean of all *n* data points x_i assigned to this cluster in the previous iteration:

$$c_j(a) = \frac{1}{n} \sum_{x_i \to c_j} x_i(a)\,, \text{for } a = 1..d$$

The algorithm repeats the two previous mentioned steps until no point changes its cluster membership anymore. One of the reasons for k-means popularity is its high speed, expressed in O-notation:

$$O(\#iterations * \#clusters * \#instances * \#dimensions)$$

After the clusters are constructed with k-means, an internal validation analysis is conducted to evaluate the results in terms of the information intrinsic to the data alone. An external validation analysis follows to outline differences and relationships with other variables that were not used to build the clusters (Rendón et al. 2011).

E. Applied Data Analysis - KNIME

The Konstanz Information Miner KNIME provides a data analysis platform with already implemented machine learning algorithms and data mining applications. The graphical user interface allows assembling a data pipe for data processing, modelling, analysis and visualisation. The KNIME setup is presented in figure 1. The CSV files containing the average financial ratios for each industry are important into KNIME via the file reader node. The subsequent k-means (k=10) node applies the cluster algorithm. Node 12 writes the results in a CSV file. Node 3, 4, 5 and 6 are used for visualisation.

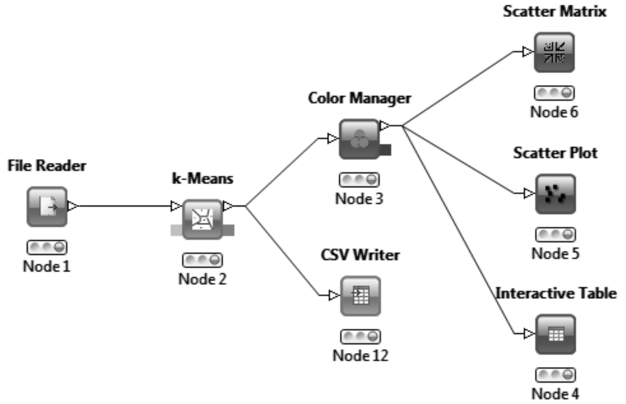

Figure 1, KNIME Setup

F. Limitations

The following limitations for this research are given:

1. Growth ratios focused on asset, operating profit, sale and shareholder's equity are not considered because a time-series analysis is required. Martikainen et al. (1995) outlines the instability of ratios with respect to time-series analysis and the need for an up-front transformation of the ratios.
2. Economic up- and down trends are not considered while drawing conclusions.
3. Performance measurement is conducted in terms of financial performs, i.e. non-financial indicators, such as number of patents filed per year per industry, are not considered.
4. Market structures are not considered in the discussions, i.e. monopoly and oligopoly influences are excluded in drawing conclusions.

IV. Results & Discussion

The following section presents a summary of major results and related discussions to gain insights and develop conclusions from the generated data set. Even if 10 clusters for each financial ratio are identified, only 4 clusters are presented and discussed. Two clusters containing industries with the highest performance and two clusters containing industries with the lowest performance are outlined for each financial ratio.

A. Current Ratio

Table 1 represents the summary of clusters for the KPI current ratio. The first two clusters are characterised by an average current ratio of approximately 4, i.e. companies in those industries utilise 4 times more current assets than their amount of current liabilities. Industries in the two best performing clusters such as distributors, apparel, precious metals or agricultural chemicals have one characteristic in common: high inventory, i.e. high inventory as a contributing factor to current assets results in higher current ratios as compared to the clusters with the lowest performance. Industries represented by a current ratio approximately equal to 1 are characterised by utilising the same amount of current assets and current liabilities. Thus, in the two worst performing clusters in terms of the current ratio, several service industries can be found, such as Power Generation, Oil/Gas Transmission, Air Freight/Delivery, Environmental or Television services. A common characteristic of service industries is low to no inventory, i.e. the major contributing factor to current assets is only accounts receivable. As the structure of current liabilities is similar in all industries in all four clusters, the amount of inventory is the major factor leading to the difference between the best and worst performing clusters.

Table 1, Current Ratio Clusters

Industries	Companies	Average	Standard Deviation
Clusters with Highest Performance			
Consumer Specialties Wholesale Distributors	4 2	4.55	1.88
Agricultural Chemicals Apparel Electrical Products Medical Specialities Medical/Dental Instruments Ordnance And Accessories Precious Metals	13 12 14 17 14 3 39	3.71	3.45
Clusters with Lowest Performance			
Air Freight/Delivery Services Aluminum Beverages (Production/Distribution) Coal Mining Computer Manufacturing Consumer Electronics/Appliances Environmental Services Hotels/Resorts Integrated oil Companies Movies/Entertainment Oil & Gas Production Other Pharmaceuticals Railroads	12 2 14 10 4 7 6 17 23 7 25 5 11	1.28	0.71
Consumer Electronics/Video Chains Diversified Commercial Services Electric Utilities: Central Oil/Gas Transmission Power Generation Publishing Real Estate Investment Trusts Retail: Computer Software & Peripheral Television Services	2 16 59 12 19 4 177 2 7	0.98	0.73

B. Gross Margin

The major industries included in the cluster with the highest gross margin are financial and investment services, such as banks, investment management and savings institutions (see table 2). The first cluster has a much higher average gross margin than the second best performing cluster by also having a lower standard deviation while both clusters contain service industries. One explanation for the difference is that financial services generate much more revenue than services such as advertising or television due to the nature of their business models. In general, service industries do not have major cost of goods sold, because expenses for labour is the only contributing factor to direct costs of the production of their services. In contrast, manufacturing industries have to include expenses for production material which results in higher cost of goods sold and, thus, in a lower gross margin. Therefore, industries requiring larger amounts of raw material, such as Automotive Aftermarket or Steel/Iron Ore outline a significant lower gross margin of 17.10 percent. The cluster with the lowest gross margin also has the highest standard deviation amongst all, i.e. higher statistical outliers by fewer companies in comparison to the other clusters. Thus, a single contributing factor explaining the worst performance cannot be derived.

Table 2, Gross Margin Clusters

Industries	Companies	Average (%)	Standard Deviation
Clusters with Highest Performance			
Banks	2	91.88	10.77
Building operators	3		
Commercial Banks	23		
Finance: Consumer Services	32		
Investment Bankers/Brokers/Service	18		
Investment Managers	24		
Major Banks	58		
Miscellaneous	2		
Savings Institutions	6		
Specialty Insurers	10		
Advertising	8	69.35	23.98
Computer peripheral equipment	3		
Computer Software: Prepackaged Software	35		
Software: Programming, Data Processing	11		
Diversified Commercial Services	16		
Diversified Financial Services	2		
Hospital/Nursing Management	15		
Major Pharmaceuticals	24		
Newspapers/Magazines	10		
Oil & Gas Production 1	25		
Real Estate	26		
Real Estate Investment Trusts	177		
Services-Misc. Amusement & Recreation	11		
Television Services	7		
Clusters with Lowest Performance			
Automotive Aftermarket	16	17.10	19.26
Engineering & Construction	7		
Integrated oil Companies	23		
Meat/Poultry/Fish	5		
Medical Specialities	17		
Military/Government/Technical	18		
Steel/Iron Ore	18		
Aluminum	2	-0.94	32.65
Building Products	5		
Retail: Computer Software & Equipment	2		

C. Net Margin

The best performing cluster in terms of the net margin consists of two industries: Oil/Gas Production and Real Estate Investment Trusts (see table 3). Both industries are also included in the cluster with the second best performance with respect to the gross margin. The comparison of the income statements of companies operating in the Oil & Gas Production industry and in Real Estate Investment Trusts gives the following explanation why both industries outline similar net margins, even if their business models (asset structure, customer segments, source of revenue, etc.) completely differ from each other: Oil and Gas Production companies are characterised by higher revenues due to the enormous quantities produced, but also by higher operating expenses due to their equipment necessary for operating oil and gas plants. In contrast, companies operating in Real Estate Investment Trusts outline less revenue, but also less cost of goods sold and operating expenses resulting from their activities only focused on collective investment. However, the high standard deviation of 172.12 represents a large fluctuation within both industries, i.e. the aforementioned explanation can be seen as general, but not necessarily as statistical valid.

The industries Aerospace, Coal Mining and Railroads are not included in the two best performing clusters for the gross margin, but associated with the second best performing cluster in terms of the net margin. Those industries present lower gross margins compared to the service industries because higher cost of goods sold occur due to the focus on production and manufacturing. With respect to the second cluster, service industries also represent highly profitable industries. The high profitability might be a result of low capital and operational expenditure, as production facilities or raw materials are not required for revenue generation. The standard deviation for this cluster is 10 times lower compared to the other three clusters, i.e. the explanation for the industries performance in this cluster can be seen as generally applicable.

The clusters with the lowest performance in terms of the net margin are characterised by having high standard deviations (102.60 and 218.87). Thus, exact reasons explaining the negative net margin of both clusters cannot be given as high fluctuations in the net margins would require the analysis of every single industry in isolation.

Table 3, Net Margin Clusters

Industries	Companies	Average (%)	Standard Deviation
Clusters with Highest Performance			
Oil & Gas Production	25	37.03	172.12
Real Estate Investment Trusts	177		
Aerospace	9	14.25	12.44
Banks	2		
Biotechnology: Electromedical Apparatus	2		
Business Services	32		
Coal Mining	10		
Commercial Banks	23		
Diversified Commercial Services	16		
Investment Managers	24		
Major Banks	58		
Property-Casualty Insurers	51		
Railroads	11		
Savings Institutions	6		
Television Services	7		
Water Supply	7		
Clusters with Lowest Performance			
Building Products	5	-20.23	102.60
Computer Communications Equipment	4		
Electronic Components	8		
Finance: Consumer Services	32		
Industrial Machinery/Components	63		
Biotechnology: Physical & Biological Research	3	-44.19	218.87
Major Chemicals	44		
Precious Metals	39		

D. Return on Investment

The best performing cluster in terms of ROI includes Computer Manufacturing and Integrated Oil Companies (see table 4). The average ROI for this cluster is more than 4 times higher than the average ROI of the second best performing cluster. Thus, on average, companies in those industries generate 4 times more operating profit (Earnings Before Interest & Tax) than companies in the second cluster while assuming similar asset structures. However, a heavily depreciated plant or a large amount of intangible assets causes a distortion of the value. Significant is the amount of industries focused on computer equipment and software found in the cluster with the second lowest performance in terms of ROI. The majority of those industries in the four clusters are also found in the best and worst performing clusters for measuring asset turnover. The only difference between asset turnover and

ROI is that asset turnover does not consider cost of goods sold and operating expenses. This leads to the conclusion that the industries grouped in the same ROI and asset turnover clusters with regard to the performance level (high/low), are characterised by the same proportion of cost of goods sold and operating expenses in relation to their revenue.

The comparison of the number of industries per cluster shows that the first (best performance) and fourth cluster (worst performance) contain significantly less industries than the second and third cluster. This leads to the conclusion that 115 industries represent an average ROI between 0 and 2, i.e. no significant pattern can be determined.

Table 4, Return on Investment Clusters

Industries	Companies	Average	Standard Deviation
Clusters with Highest Performance			
Computer Manufacturing	4	9.08	27.51
Integrated Oil Companies	23		
Auto Manufacturing	10	2.09	3.31
Building operators	3		
Commercial Banks	23		
Department/Specialty Retail Stores	11		
Major Banks	58		
Clusters with Lowest Performance			
Computer Communications Equipment	4	-0.02	0.17
Computer peripheral equipment	3		
Computer Software: Prepackaged Software	35		
Miscellaneous manufacturing industries	4		
Recreational Products/Toys	5		
Transportation Services	5		
Aluminum	2	-0.18	0.26
Catalog/Specialty Distribution	4		

E. *Return on Equity*

The highest average ROE is presented by a cluster only including one industry: Integrated Oil Companies (see table 5). Companies in this industry are on average more than 3 times more efficient in utilising shareholders' equity than companies associated with industries in the second best performing cluster. The balance sheets of companies within the Integrated Oil industry contain significant more shareholders' equity than injected into companies associated with industries in the second best performing cluster. The multiplicative effect of investment might be a potential reason for the high ROE. However, the standard deviation (62.26) for the industry hinders a general conclusion.

Only 5 out of 119 industries are allocated to the two clusters showing the highest average ROE. In addition, only 3 industries are included in the two clusters having the lowest ROE. Thus, more than 90 industries outline a ROE between -2 and 3, i.e. no significant performance differences between industries are identified. This pattern is similar to the one find in the ROI clusters.

Table 5, Return on Equity Clusters

Industries	Companies	Average	Standard Deviation
Clusters with Highest Performance			
Integrated Oil Companies	23	11.53	62.26
Clothing/Shoe/Accessory Stores	22	3.21	10.22
Computer Manufacturing	4		
Medical/Nursing Services	9		
Precious Metals	39		
Clusters with Lowest Performance			
Diversified Commercial Services	16	-1.65	6.09
Oil Refining/Marketing	13		
Farming/Seeds/Milling	9	-6.44	19.84

F. *Collection Period*

Industries allocated to the two best performing clusters in terms of collection period are on average able to collect outstanding payments between 15 and 30 days (see table 6). Major industries in the two clusters are focused on retail and business-to-customer, such as clothing stores, restaurants, speciality retail, food chains or building material stores. The reason for this characteristic might lay in the fact, that such stores allow their customers payments with cash, debit or credit card within 30 days. The fourth cluster consists of three financial services, i.e. a high collection period might be characteristic for this type of industry. In contrast to the clusters focused on ROI and ROE, the amount of industries is more equally spread over all four clusters.

Table 6, Collection Period Clusters

Industries	Companies	Average (Days)	Standard Deviation
Clusters with Highest Performance			
Catalog/Specialty Distribution	4	13.37	13.18
Clothing/Shoe/Accessory Stores	22		
Consumer Electronics/Video Chains	2		
Department/Specialty Retail Stores	11		
Food Chains	11		
Integrated Oil Companies	23		
Other Pharmaceuticals	5		
Other Specialty Stores	17		
Air Freight/Delivery Services	12	27.27	31.54
Aluminum	2		
Automotive Aftermarket	16		
Oil Refining/Marketing	13		
Restaurants	14		
RETAIL: Building Materials	5		
Clusters with Lowest Performance			
Accident & Health Insurance	8	112.45	106.36
Advertising	8		
Biotechnology: Biological Research	3		
Biotechnology: Electrotherapeutic Apparatus	2		
Broadcasting	10		
Building operators	3		
Business Services	32		
Software: Programming, Data Processing	11		
Industrial Machinery/Components	63		
Life Insurance	30		
Diversified Financial Services	2	256.42	707.85
Finance: Consumer Services	32		
Investment Bankers/Brokers/Service	18		
Oil & Gas Production	125		
Property-Casualty Insurers	51		

G. Asset Turnover

Comparing the number of industries within the clusters for asset turnover clearly shows a significant difference between best performing and worst performing clusters (see table 7). Both best performing clusters contain together 6 industries, whereas the two lowest performing clusters contain 43 industries by having a low standard deviation. In contrast, the standard deviation for the 6 industries is 40 to 50 times higher. Many industries are characterised by low asset turnovers and only few industries by high turnovers, i.e. 43 industries in the last two clusters only generate revenue approximately equal to their assets. Major industries in those clusters are service industries. Thus, asset turnover might not be a suitable key performance measure to compare service industries.

Table 7, Asset Turnover Clusters

Industries	Companies	Average	Standard Deviation
Clusters with Highest Performance			
Auto Manufacturing	10	40.89	49.48
Computer Manufacturing	4		
Integrated oil Companies	23		
Department/Specialty Retail Stores	11	24.61	38.83
Other Pharmaceuticals	5		
Retail: Computer Software & Peripheral Equipment	2		
Clusters with Lowest Performance			
Broadcasting	10	1.26	1.06
Business Services	32		
Computer Software: Prepackaged Software	35		
Diversified Commercial Services	16		
Diversified Financial Services	2		
EDP Services	32		
Electrical Products	14		
Engineering & Construction	7		
Forest Products	7		
Home Furnishings	10		
Homebuilding	26		
Hotels/Resorts	17		
Investment Bankers/Brokers/Service	18		
Investment Managers	24		
Miscellaneous manufacturing industries	4		
Office Equipment/Supplies/Services	9		
Oil/Gas Transmission	12		
Oilfield Services/Equipment	16		
Ordnance And Accessories	3		
Other Consumer Services	28		
Other Specialty Stores	17		
Plastic Products	6		
Professional Services	17		
Recreational Products/Toys	5		
Specialty Chemicals	5		
Textiles	4		
Water Supply	7		
Banks	2	0.57	0.39
Biotechnology: Biological Research	3		
Biotechnology: Laboratory Analytical Instruments	4		
Building Products	5		
Computer Communications Equipment	4		
Computer peripheral equipment	3		
Software: Programming, Data Processing	11		
Fluid Controls	7		
Marine Transportation	49		
Mining & Quarrying of Nonmetallic Minerals	11		
Miscellaneous	2		
Movies/Entertainment	7		
Newspapers/Magazines	10		
Publishing	4		
Real Estate	26		
Real Estate Investment Trusts	177		
Savings Institutions	6		
Services-Misc. Amusement & Recreation	11		
Transportation Services	5		

V. Conclusion

An extensive literature review considering more than 50 journal papers identified the gap in literature that no researcher conducted a cluster analysis to gain understanding of cross-industry financial performance patterns measured with financial ratios and based on a data set that is statistically significant. As the lack of having access to large-scale financial data is mentioned in the literature several times, a data mining software was developed to extract the income statements and balance sheets for the years 2010 to 2014 from 3282 companies listed on the NASDAQ stock exchange in the USA. It was found that 29% of the extracted data was corrupted due to inconsistencies of the NASDAQ information. Based on the extracted and cleaned data points, eight financial ratios are calculated for every company for the year 2014. The companies are grouped into 119 industries by using the NASDAQ industry classification tree so that the arithmetic mean of each ratio per industry is derived. Thus, the financial performance of every industry is characterised by eight average financial ratios. Ten clusters for each ratio are identified by applying the non-hierarchical cluster algorithm k-means, i.e. all industries per financial ratio are grouped with regard to their performance similarity. The arithmetic mean and standard deviation is calculated for every cluster for every financial ratio. Following, the two clusters with the highest and the two clusters with the lowest average performance are selected for each ratio to draw conclusions and to identify patterns. The major findings are listed below:

1. All identified clusters are unique in terms of the types of included industries, i.e. two or more identical clusters with the same industries do not exist.
2. No industry is found that is included in all best performing clusters. In contrast, no industry is included in all worst performing clusters.
3. The best performing industry of all analysed industries is Integrated Oil. This industry is included in the best performing clusters in terms of ROI, ROE, Collection Period and Asset Turnover. However, it is also allocated to the clusters having the second lowest average current ratio and gross margin.
4. The number of industries per high and low performing clusters significantly varies per ratio, e.g. 43 industries are included in the two low performing asset turnover clusters, and whereas only 3 industries are included in the two low performing ROE clusters.
5. Analysing the clusters for current ratio leads to the following conclusion: the amount of inventory is the factor influencing this financial ratio the most, i.e. an industry characterised by a high average inventory has a high average current ratio, whereas industries having low to no inventory are characterised by low current ratios. Service industries are found to be typical for having low current ratios.
6. In general, service industries tend to have the highest gross margin amongst all 119 industries. More precisely, banks and investment industries are found to have the significantly highest gross margin combined with a very low standard deviation.
7. Three out of four net margin clusters are characterised by very high standard deviations making the development of general conclusions not possible.
8. The majority of industries listed in the four ROI clusters are also found in the clusters for asset turnover.

Thus, the industries grouped in the same ROI and asset turnover clusters with regards to the performance level (high/low), are characterised by the same proportion of cost of goods sold and operating expenses in relation to their revenue.

9. Retail and business-to-customer industries tend to have lower collection periods than business-to-business industries.
10. Forty-three industries are characterised by low asset turnovers and only 6 industries by high turnovers, i.e. the majority of industries generate revenue equal to their amount of assets.

The comparison of the hierarchical-cluster algorithm DBSCAN with the non-hierarchical cluster algorithm k-means leads to the finding that DBSCAN is unsuitable for the detection of clusters in financial ratios.

If the results of this work are used in future research, it has to be considered that a snapshot of the performance of 119 industries in the financial year 2014 is analysed, i.e. economic up- and down-trends are not considered in the discussions. This research only focuses on financial performance measurement, i.e. non-financial measures, such as the number of filed patents per year per industry are not considered. Furthermore, monopoly and oligopoly influences are excluded in drawing conclusions.

Having such as an extensive pool of financial data facilitates multiple future research projects. Distribution functions of financial ratios per industry might be derived for 2010 to 2014 so that the calculation of correlations between performance distributions in industries is possible. A comprehension of the correlation between industries’ performances over time might be derived. Drawing conclusions from the literature review, the following approaches might be worth to pursue: TRICLUSTER and Cross-Graph Quasi-Bicliques might be used to conduct a cluster analysis within a multidimensional space, i.e. considering each financial ratio as a separate dimension so that an eight dimensional space is evaluated to gain understanding of clusters including several financial ratios. Multi-layer perceptron and learning vector quantisation might be applied in future research based to predict companies’ financial failure based on the data extracted in this work. Furthermore, the developed data mining software might be extended to extract financial data from Yahoo finance or google finance service so that more companies are included in the evaluation.

Finally, this work builds a profound basis for my future PhD thesis.

References

Al-Kassar, T.A. & Soileau, J.S. 2014, 'Financial performance evaluation and bankruptcy prediction (failure)', *Arab Economic and Business Journal*, vol. 9, no. 2, pp. 147-55.

Altman, E.I. 1968, 'Financial ratios, discriminant analysis and the prediction of corporate bankruptcy', *The journal of finance*, vol. 23, no. 4, pp. 589-609.

Altman, E.I., Haldeman, R.G. & Narayanan, P. 1977, 'ZETA TM analysis A new model to identify bankruptcy risk of corporations', *Journal of banking & finance*, vol. 1, no. 1, pp. 29-54.

Avkiran, N.K. 2011, 'Association of DEA super-efficiency estimates with financial ratios: Investigating the case for Chinese banks', *Omega*, vol. 39, no. 3, pp. 323-34.

Beaver, W.H. 1968, 'Market prices, financial ratios, and the prediction of failure', *Journal of accounting research*, pp. 179-92.

Boyacioglu, M.A., Kara, Y. & Baykan, Ö.K. 2009, 'Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey', *Expert Systems with Applications*, vol. 36, no. 2, pp. 3355-66.

Chen, K.H. & Shimerda, T.A. 1981, 'An Empirical Analysis of Useful Financial Ratios', *Financial Management*, vol. 10, no. 1, pp. 51-60.

Chu, D.K., Zollinger, T.W., Kelly, A.S. & Saywell, R.M. 1991, 'An empirical analysis of cash flow, working capital, and the stability of financial ratio groups in the hospital industry', *Journal of Accounting and Public Policy*, vol. 10, no. 1, pp. 39-58.

Cinca, C.S., Molinero, C.M. & Larraz, J.G. 2005, 'Country and size effects in financial ratios: A European perspective', *Global Finance Journal*, vol. 16, no. 1, pp. 26-47.

Cowen, S.S. & Hoffer, J.A. 1982, 'Usefulness of financial ratios in a single industry', *Journal of Business Research*, vol. 10, no. 1, pp. 103-18.

Deakin, E.B. 1976, 'Distributions of financial accounting ratios: some empirical evidence', *Accounting Review*, pp. 90-6.

Delen, D., Kuzey, C. & Uyar, A. 2013, 'Measuring firm performance using financial ratios: A decision tree approach', *Expert Systems with Applications*, vol. 40, no. 10, pp. 3970-83.

Dimitras, A., Slowinski, R., Susmaga, R. & Zopounidis, C. 1999, 'Business failure prediction using rough sets', *European Journal of Operational Research*, vol. 114, no. 2, pp. 263-80.

Ertuğrul, İ. & Karakaşoğlu, N. 2009, 'Performance evaluation of Turkish cement firms with fuzzy analytic hierarchy process and TOPSIS methods', *Expert Systems with Applications*, vol. 36, no. 1, pp. 702-15.

Feng, C.-M. & Wang, R.-T. 2000, 'Performance evaluation for airlines including the consideration of financial ratios', *Journal of Air Transport Management*, vol. 6, no. 3, pp. 133-42.

Gallizo, J.L., Gargallo, P. & Salvador, M. 2008, 'Multivariate partial adjustment of financial ratios: a Bayesian hierarchical approach', *Journal of applied econometrics*, vol. 23, no. 1, pp. 43-64.

Gallizo, J.L., Jiménez, F. & Salvador, M. 2002, 'Adjusting financial ratios: a Bayesian analysis of the Spanish manufacturing sector', *Omega*, vol. 30, no. 3, pp. 185-95.

Gallizo, J.L., Jiménez, F. & Salvador, M. 2003, 'Evaluating the effects of financial ratio adjustment in European financial statements', *European Accounting Review*, vol. 12, no. 2, pp. 357-77.

Gupta, M.C. & Huefner, R.J. 1972, 'A cluster analysis study of financial ratios and industry characteristics', *Journal of Accounting Research*, pp. 77-95.

Gutierrez, I. & Carmona, S. 1988, 'A fuzzy set approach to financial ratio analysis', *European Journal of Operational Research*, vol. 36, no. 1, pp. 78-84.

Halkos, G.E. & Salamouris, D.S. 2004, 'Efficiency measurement of the Greek commercial banks with the use of financial ratios: a data envelopment analysis approach', *Management Accounting Research*, vol. 15, no. 2, pp. 201-24.

Halkos, G.E. & Tzeremes, N.G. 2012, 'Industry performance evaluation with the use of financial ratios: An application of bootstrapped DEA', *Expert Systems with Applications*, vol. 39, no. 5, pp. 5872-80.

Horrigan, J.O. 1965, 'Some Empirical Bases of Financial Ratio Analysis', *The Accounting Review*, vol. 40, no. 3, pp. 558-68.

Horrigan, J.O. 1968, 'A short history of financial ratio analysis', *Accounting Review*, pp. 284-94.

Huang, S.-M., Tsai, C.-F., Yen, D.C. & Cheng, Y.-L. 2008, 'A hybrid financial analysis model for business failure prediction', *Expert Systems with Applications*, vol. 35, no. 3, pp. 1034-40.

Kaufman, L. & Rousseeuw, P.J. 2009, *Finding groups in data: an introduction to cluster analysis*, vol. 344, John Wiley & Sons.

Kung, C.-Y. & Wen, K.-L. 2007, 'Applying grey relational analysis and grey decision-making to evaluate the relationship between company attributes and its financial performance—a case study of venture capital enterprises in Taiwan', *Decision Support Systems*, vol. 43, no. 3, pp. 842-52.

Lev, B. & Sunder, S. 1979, 'Methodological issues in the use of financial ratios', *Journal of Accounting and Economics*, vol. 1, no. 3, pp. 187-210.

Lewellen, J. 2004, 'Predicting returns with financial ratios', *Journal of Financial Economics*, vol. 74, no. 2, pp. 209-35.

Li, Y. & Zhang, Q. 2011, 'The application of principal component analysis on financial analysis in real estate listed company', *Procedia Engineering*, vol. 15, pp. 4499-503.

Longinidis, P. & Georgiadis, M.C. 2011, 'Integration of financial statement analysis in the optimal design of supply chain networks under demand uncertainty', *International journal of production economics*, vol. 129, no. 2, pp. 262-76.

Maricica, M. & Georgeta, V. 2012, 'Business failure risk analysis using financial ratios', *Procedia-Social and Behavioral Sciences*, vol. 62, pp. 728-32.

Martikainen, T., Perttunen, J., Yli-Olli, P. & Gunasekaran, A. 1995, 'Financial ratio distribution irregularities: implications for ratio classification', *European Journal of Operational Research*, vol. 80, no. 1, pp. 34-44.

McIvor, R.T., McCloskey, A.G., Humphreys, P.K. & Maguire, L.P. 2004, 'Using a fuzzy approach to support financial analysis in the corporate acquisition process', *Expert Systems with Applications*, vol. 27, no. 4, pp. 533-47.

Mercan, M., Reisman, A., Yolalan, R. & Emel, A.B. 2003, 'The effect of scale and mode of ownership on the financial performance of the Turkish banking sector: results of a DEA-based analysis', *Socio-Economic Planning Sciences*, vol. 37, no. 3, pp. 185-202.

Niemann, M., Schmidt, J.H. & Neukirchen, M. 2008, 'Improving performance of corporate rating prediction models by reducing financial ratio heterogeneity', *Journal of Banking & Finance*, vol. 32, no. 3, pp. 434-46.

Öcal, M.E., Oral, E.L., Erdis, E. & Vural, G. 2007, 'Industry financial ratios—application of factor analysis in Turkish construction industry', *Building and Environment*, vol. 42, no. 1, pp. 385-92.

Ou, J.A. & Penman, S.H. 1989, 'Financial statement analysis and the prediction of stock returns', *Journal of accounting and economics*, vol. 11, no. 4, pp. 295-329.

Pinches, G.E., Eubank, A.A., Mingo, K.A. & Caruthers, J.K. 1975, 'The hierarchical classification of financial ratios', *Journal of Business Research*, vol. 3, no. 4, pp. 295-310.

Pinkasovitch, A. 2009, 'Investopedia', Investopedia.

Rendón, E., Abundez, I., Arizmendi, A. & Quiroz, E. 2011, 'Internal versus external cluster validation indexes', *International Journal of computers and communications*, vol. 5, no. 1, pp. 27-34.

Ross, S.A., Westerfield, R. & Jordan, B.D. 2008, *Fundamentals of corporate finance*, Tata McGraw-Hill Education.

Rushinek, A. & Rushinek, S.F. 1987, 'Using financial ratios to predict insolvency', *Journal of business research*, vol. 15, no. 1, pp. 93-100.

Seng, J.-L. & Lai, J.T. 2010, 'An Intelligent information segmentation approach to extract financial data for business valuation', *Expert Systems with Applications*, vol. 37, no. 9, pp. 6515-30.

Sharma, S. & Mahajan, V. 1980, 'Early warning indicators of business failure', *The Journal of Marketing*, pp. 80-9.

Shaverdi, M., Heshmati, M.R. & Ramezani, I. 2014, 'Application of Fuzzy AHP Approach for Financial Performance Evaluation of Iranian Petrochemical Sector', *Procedia Computer Science*, vol. 31, pp. 995-1004.

Sim, K., Liu, G., Gopalkrishnan, V. & Li, J. 2011, 'A case study on financial ratios via cross-graph quasi-bicliques', *Information Sciences*, vol. 181, no. 1, pp. 201-16.

Sudarsanam, P.S. & Taffler, R. 1995, 'Financial ratio proportionality and inter-temporal stability: An empirical analysis', *Journal of banking & finance*, vol. 19, no. 1, pp. 45-60.

Wang, Y.-J. 2008, 'Applying FMCDM to evaluate financial performance of domestic airlines in Taiwan', *Expert Systems with Applications*, vol. 34, no. 3, pp. 1837-45.

Wang, Y.-J. 2009, 'Combining grey relation analysis with FMCGDM to evaluate financial performance of Taiwan container lines', *Expert Systems with Applications*, vol. 36, no. 2, pp. 2424-32.

Wang, Y.-J. & Lee, H.-S. 2008, 'A clustering method to identify representative financial ratios', *Information Sciences*, vol. 178, no. 4, pp. 1087-97.

Watkins, A.L. 2000, 'Hospital financial ratio classification patterns revisited: Upon considering nonfinancial information', *Journal of Accounting and Public Policy*, vol. 19, no. 1, pp. 73-95.

Xu, W., Xiao, Z., Dang, X., Yang, D. & Yang, X. 2014, 'Financial ratio selection for business failure prediction using soft set theory', *Knowledge-Based Systems*, vol. 63, pp. 59-67.

Yli-Olli, P. & Virtanen, I. 1989, 'On the long-term stability and cross-country invariance of financial ratio patterns', *European Journal of Operational Research*, vol. 39, no. 1, pp. 40-53.

Zeller, T.L., Stanko, B.B. & Cleverley, W.O. 1996, 'A revised classification pattern of hospital financial ratios', *Journal of Accounting and Public Policy*, vol. 15, no. 2, pp. 161-81.

Zulkifli, N.A. 2010, 'Industry Financial Ratios-Application Of Factor Analysis In Malaysian Industrial Sector', Universiti Sains Malaysia.