RIMONA PALAS
College of Business and Law, Israel

AMOS BARANES
Peres Academic Center, Israel

THE PREDICTION OF EARNINGS MOVEMENT USING MANDATED XBRL DATA – INDUSTRY ANALYSIS

Abstract:

The immediate availability and easy accessibility of the XBRL filings will allow researchers and investors, especially small investors, to implement investment strategies based on this information. The objective of this study is to examine whether previous studies, predicting the direction of movement of earnings, are still relevant when using the newly SEC required, XBRL database (standardized financial reporting system).

The study analyzes NYSE companies' XBRL quarterly data, from 2011 to 2015, using a two-step Logit regression model. The model is then used to arrive at the probability of the directional movement of earnings between current quarter and subsequent quarter, adjusted for a drift. Additional models are created by dividing the sample into industry membership (based on SIC codes).

The results classified the companies as ones that would realize an increase in earnings or a decrease in earnings. The final model indicated a significant ability to predict subsequent earnings changes. The predictions appear to be correct on average about 70.7% of the time (higher than those of previous studies based on COMPUSTAT).

The industry based models, although do not increase the accuracy of the model (an average of 68%) do increase the portfolio size. In other words, the model, based on industry, is able to classify more companies with a higher probability.

These results suggest that XBRL data can be used as a means for forecasting movements in earnings, and creating a profitable investment strategy.

Keywords:

accounting information, earnings prediction, investment strategy, XBRL, industry analysis

JEL Classification: F37

INTRODUCTION

The ability to predict earnings based on past performance has been recognized as a measure of earnings quality (Penman and Zhang 2002) and while Ball and Shivakumar (2008) conclude that earnings announcements provide only a modest amount of new information to the share market, Bloomfield, Libby and Nelson (2003) show that investors over rely on old earnings performance when predicting future earnings performance.

These studies highlight the necessity to develop a tool to better predict future earnings and help develop various investment strategies.

While most of the previous studies use data available on COMPUSTAT, this study uses data directly reported by companies in XBRL format, which is freely available immediately after reporting to the SEC.

Many research papers have concentrated on the importance of earnings announcements and forecasts in the determination of investment decisions. While earlier research has only been able to show relatively low informativeness of earnings (Ball and Brown 1968; Beaver, 1968; Foster, Olsen and Shevlin 1984; and Bernard and Thomas 1990) later studies were able to show the incremental information content of specific components of the financial statements. For example, Finger (1994) shows that earnings provide information for future earnings and cash flows. Ou (1990) and Ou and Penman (1989) predict sign changes in the earnings per share using forecasting models developed from various income statement and balance sheet components. Shroff (1999) assesses the predictive ability of a "composite" model, which forecasts as a function of current earnings and current security prices, against three univariate benchmarks: a random walk model, a random walk with drift model, and a first order autoregressive/moving average (ARIMA) model. The findings indicate that the composite model obtains significantly lower forecast errors relative to the benchmark models. Roychowdhury and Sletten (2012) find that earnings informativeness is higher in bad-news periods than in good-new periods. Alam and Brown (2006) were able to show that disaggregated earnings data were better able to predict next period's earnings in the banking industry.

Ou and Penman (1989) were the first researchers to focus on the usefulness of accounting information to predict the direction of movement of earnings relative to trend adjusted current earnings. The study is important because it evaluates whether accounting information can consequently be used as the basis for a profitable investment strategy. Given investors' reliance on earnings this could be a valuable tool for a profitable investment strategy. The authors found that financial statement analysis can provide a measure that is an indicator of future earnings which in turn is used as a successful investment strategy. However, the evidence from subsequent studies (Holthausen and Larker 1992; Bernard et al. 1997; Stober 1992; Setiono and Strong 1998; and Bird, Gerlach and Hall 2001) has been mixed.

Holthausen and Larker (1992) suggest that a better strategy, yielding higher returns would be one that is based on an economic concept, such as the industry the

company is traded in. Lee (1985) proposes that statistical inferences from financial data depend on the ability to normalize the information and indicates that a major control factor is industry membership.

One objective of this study is to repeat the original Ou and Penman (1989) study over a more recent time period, and based on industry membership, examining its validity as a viable tool for investment decisions. However, the main objective is to examine the validity of the methodology using, not the original COMPUSAT database, but the XBRL database. XBRL (eXtensible Business Reporting Language) is a freely available and global standard for exchanging business information. XBRL allows the expression of semantic meaning commonly required in business reporting. One use of XBRL is to define and exchange financial information, such as financial statements.

The SEC has created the XBRL U.S. GAAP Financial Reporting Taxonomy. This taxonomy is a collection of accounting data concepts and rules that enables companies to present their financial reports electronically. The SEC's deployment was launched in 2008 in phases, and all public U.S. GAAP companies were required to file their financial reports using the XBRL reporting technology starting from June 15, 2011.

While COMPUSTAT has been a popular source of financial information for both academics and practitioners, it is costly while XBRL filings are freely available. XBRL filings also have a time advantage, while they are published concurrently with the related PDF versions, it takes an average of 14 weekdays from the time a company files with the SEC for that data to appear in COMPUSTAT (D'Souza, Ramesh, and Shen. 2010). In addition, the reliability of COMPUSTAT has also been questioned, prior studies have shown that COMPUSTAT data may differ from the original corporate financial data (San Miguel 1977; Kinney and Swanson 1993; Tallapally, Luehlfing, and Motha 2011) and data found in other accounting databases (Rosenberg and Houglet 1974; Yang, Vasarhelyi and Liu 2003).

On the other hand, while there is still not enough research regarding the reliability of XBRL data, studies up to date seem positive: Boritz and No (2013) find that when examining the quality of interactive data XBRL tagged information it is the most complete and most accurate source of company data compared with COMPUSTAT, Yahoo Finance and Google Finance; Chychyla and Kogan (2015), although did not attempt to compare COMPUSTAT and XBRL 10-K reports, found that COPUSTAT significantly alters numbers reported on the 10-K filings; and Heselmann, Henselmann, Ditter and Scherr (2015) suggest that XBRL analysis is a useful tool in assessing irregularities in accounting data.

Vasarhelyi, Chan and Krahel (2012) made suggestions for new research opportunities as a result of the evolving XBRL technology. Their suggestion was to examine whether findings from prior research that relied on private vendor databases (such as COMPUSTAT), if replicated, will still hold using XBRL database. This paper is an attempt to follow their suggestion, and examine the ability of earnings to indicate future earnings.

Recently there have been studies attempting to assess the usefulness of XBRL filing data in predicting future earnings (Williamsm 2015, Baranes and Palas, 2017) however, their database was limited as were the results. The current study is an attempt to utilize the XBRL database in financial analysis, prediction of future earnings, on a much larger scale which is more representative of the market. The XBRL data, filed by all NYSE traded companies, is used to replicate the same methodology used by Ou and Penman (1989).

These studies suggest that not only can XBRL data be an important tool for researchers, but may also be a more efficient tool for investors, given the timeliness of the data, and especially smaller investors, given the low cost.

The paper is organized as follows, the second section reviews academic literature evaluating Ou and Penman (1989) and subsequent studies and examining research conducted on the validity of XBRL as a means for data. The third section outlines the method employed and the data used. Section four presents and discusses the results for the model (based on the total sample) developed to forecast future movements in earnings, in terms of accuracy and as a basis for profitable investment strategy. The last section concludes the paper.

ACADEMIC RESEARCH

In this section will be presented a review of relevant literature on two issues: an evaluation of the Ou and Penman (1989) study and subsequent studies and an examination of the validity of XBRL as a means for data comparison, including current literature on earnings prediction. The two issues will be examined separately.

Evaluation of the Ou and Penman (1989) and Consequent Studies

Ou and Penman (1989)

Ou and Penman (1989) is considered a foundation paper in accounting research literature (cited 124 times according to PROQUEST) because they were the first to focus on the usefulness of accounting information to predict the direction of the movement of earnings relative to trend adjusted current earnings.

Using an extensive financial statement analysis (68 accounting variables) the study modeled the direction of movements (increase/ decrease) in earnings per share (EPS) one year out. The sample was obtained from the 1984 COMPUSTAT annual report files and the study was conducted in several stages. In the first stage a chi-squared test was applied to a univariate LOGIT estimation and conducted for 68 accounting variables using annual report data over the period 1965-1972 and then again over the period 1973-1977. In both periods 34 (50%) of the coefficients estimated had p-values less than 0.10. In the second stage a multivariate model was used, on the variables found in the first stage, using a step-wise procedure, deleting descriptors not significant at the 0.10 level with all other descriptors included. In this stage, stage two, additional descriptors were dropped resulting in a model with 16 explanatory variables

(for the 1965-1972 period) and 18 variables (for the 1973-1977 period). The results of both time periods were then used to forecast the probability of a company's EPS lying above its trend-adjusted EPS in each of the years from 1973-1983. The companies were classified with a probability above 0.5 (the test was then repeated with p>0.6) as one that would realize an increase in EPS or a company with a probability below 0.5 (the test was then repeated with p<0.4) as one that would realize a decrease in EPS.

Although the two models only had 6 descriptors which appear in both time periods, many of the descriptors captured similar operating characteristics. For example, inventories, sales and deflated earnings appear in more than one descriptor. The two models classified the firms consistently 78.7% of the time (for a classification of above or below 0.5).

The results of the final models' indicated a significant ability of the descriptors to jointly describe subsequent earnings changes. The χ_1^2 values from the 2X2 contingency table are highly significant and the predictions appear to be correct about 60% of the time for a probability cutoff of (0.5, 0.5) and 66% of the time for a (0.6, 0.4) cutoff.

Ou and Penman (1989) continued to develop a trading strategy based on these predictions. Stocks were assigned long and short investment positions based on their probability. They purchased an equally weighted portfolio of all stocks whose estimated probability was in access of 0.6 (long position), and sold an equally weighted portfolio of all stocks whose probability was below 0.4 (short position). This strategy realized a return of 8.3% over a one year holding period, an incremental 5.7% in the second year, and 5.5% in the third year.

Replication of Ou and Penman (1989)

There have been many replications of the Ou and Penman (1989) study over different time periods, different countries, different industries, in comparison with analysts' predictions, and with additional methodologies, with mixed results.

Holthausen and Larker (1992) reexamined Ou and Penman (1989) using a different time period (1978-1988), including Over-the-Counter firms, and using only 60 of the original 68 ratios. The study estimated four different logit model (two exchanges: NYSE/AMEX and OTC, and two time periods: 1973-1977 and 1978-1982) which retained 15 ratios (the original Ou and Penman 1989 study had 18 ratios). The correlation in the probability scores between the 1973-1977 model and the 1978-1982 model for NYSE/AMEX (OTC) firms was 0.70 (0.58). The predictive ability of their models were qualitatively similar (to Ou and Penman, 1989), using a cut-off of 0.5 the overall accuracy is 60.1% (compared to 60%) and using cut-offs of 0.4 and 0.6 had an overall predictive accuracy of 65.0% (compared to 67%). However, the profitability of the trading strategy realized little value added over the period of their study; that is the Ou and Penman (1989) strategy worked well in the 1978-1982 period (a common period for both studies) regardless of exchange with an excess return varying from 6.9% to 10.3% (8.0% to 11.4% on OTC firms). However, the strategy performed poorly in the 1983-1988 period, where returns were negative (ranging from -4% to -5%) regardless of the exchange.

Bernard et. al (1997) replicated the Ou and Penman (1989) study using the same logit model to make predictions for the same years (1973-1977 and 1978-1983) and re-estimate logit model (using their approach over a previous estimation period) to produce probabilities for earnings increase for the 1984-1988 and 1989-1992 periods. The mean profitability of their investment strategies produced excess return of 4.74% in the first year and 1.24% in the second year.

Stober (1992) compared the Ou and Penman (1989) model prediction ability to that of analysts' forecasts of earnings. Using the same time period as Ou and Penman (1989) they found that the model accurately predicts the signs of one-year-ahead EPS 46% of the time, analysts' forecasts are correct about 54% of the time but a combined model correctly predicted the sign 78% of the time.

Setiono and Strong (1998) examined the Ou and Penman (1989) model using a UK sample over a period from 1980 to 1988 and found that a portfolio based on the forecasted probabilities realized abnormal returns.

Bird et al. (2001) extended the Ou and Penman (1989) model by covering a later time period (the years 1983-1997) and by encompassing the UK and Australian markets in addition to the US market. Their results found 12 variables (compared to Ou and Penman (1989) 18) and using a cut-off of 0.5 showed an accuracy of 57.5%- 62% (compared to 60%) and using cut-offs of 0.4 and 0.6 had an average predictive accuracy of 60.5%-66.5% (compared to 67%) depending on the country examined. Similar to the finding of Holthausen and Larker (1992) for the period 1983-1988, their investment strategy, based on the Ou and Penman (1989) model yielded negative returns.

Earnings Prediction Based on Industry Analysis

There is economic sense, and research evidence, that industry membership affects the way financial data represents the company's financial and operating activities. In practice, auditors compare client performance indicators as ratios industry standards to help evaluate risk (Winograd, Gerson and Berlin, 2000). Lee (1985) finds that controlling for industry is meaningful in financial statement comparisons, and Cairney and Fletcher (2009) suggest that companies grouped based on SIC codes display homogenous characteristics.

Examining the impact of industry on the ability to predict earnings is largely unaddressed in the literature. Jordan, Clark and Donald (2009) examined specific industries by applying a simple regression analysis to each of 25 of the variables used by Ou and Penman (1989) in order to explain variations in the E/P ratios of publicly traded oil and gas firms during the years 2005-2006. Their results showed that three independent variables were significant in relation to the E/P ratio when examined individually and remain statistically significant when combined in a multiple regression model. The model was able to explain almost 62% of the variation in firms' E/P ratios.

Alam and Brown (2006) examined the ability of disaggregated earnings to predict ROE in the banking industry. The results show that the mean adjusted R-square significantly increased from 0.576 to 0.623 with the progressive disaggregation of earnings during the years 1979-1996. The results also demonstrate that

disaggregated components are better able to predict next period earnings than aggregated earnings.

All of these studies suggest that while there might be validity to using financial information to predicting earnings a more finely tuned and timely tool is necessary.

Validity of XBRL

Extensible Business Reporting Language (XBRL) is a business and financial reporting technology that is being implemented to enhance internal and external reporting, electronic filing, and sharing of information.

Beginning in 2009 the SEC requires that all publicly traded companies must submit financial reports in a standardized structure using XBRL to the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system under a three-year phase-in schedule. In the first phase, as of June 15, 2009 large accelerated filers that have a worldwide public common equity float above $5 billion as of the end of the second fiscal quarter of the most recently completed fiscal year, and who prepare their financial statements according to U.S. GAAP (Generally Accepted Accounting Principles), are subject to XBRL quarterly filings. In the second phase, as of June 15, 2010 all other large accelerated filers are required to comply. In the last phase, which started on June 15, 2011, all remaining filers, including smaller reporting companies, are required to file XBRL quarterly reports as an exhibit to the traditional filings (SEC 2009).

The novelty of the XBRL structured financial reports is that the reporting content is marked up with standardized elements (XBRL tags) from a publicized list of pre-defined items (XBRL taxonomy). For example, the 2013 U.S. GAAP taxonomy contain approximately 19,000 XBRL tags that allow the user to easily extract the desired information for analysis purposes.

Literature suggests that there are several advantages of using SEC XBRL filings both for the adopting companies as well as the capital markets and research:

The XBRL structure enables unique identification and reliable extraction of accounting numbers from the financial reports – additional information comes tagged and there are no distortions due to the use of different display formats (Henselmann et al, 2015).

There is no deviation from the expected digit distribution due to differences between varying database providers (Henselmann et al. 2015).

XBRL has the potential to streamline internal accounting practices leading to cost savings and improved efficiency and effectiveness in the accounting and finance function as well as enhanced internal control leading to cost savings and improved efficiency (Amrhein, Farewell and Pinsker 2010).

The aim of the SEC XBRL mandate is to decrease information asymmetry by improving the information processing capability of regulatory filings (SEC 2009). XBRL-structured SEC filings are expected to improve data gathering and analyses by reducing manual data entries, and bringing all filings to a "common ground". Although

early research has found inconsistencies, errors, or unnecessary extensions in the XBRL filings (Debreceny et al. 2011; Du, Vasarhelyi and Zheng 2013) more recent studies found XBRL data to be not only with less errors than other forms of data, but to also provide higher quality information.

Boritz and No (2013) compared XBRL data filed with the SEC with the data provided by three data aggregators: COMPUSTAT, Google Finance, and Yahoo Finance. Their results find a significant rate of omission of more than 50% in the financial statement items provided by the aggregators compared with the interactive SEC XBRL data. For items that are not omitted they find between 5-8% mismatches, with approximately 56% differences being greater than conventional materiality. The implications of their study is that XBRL information is the most complete and most accurate source of company data.

Chychyla and Kogan (2015) found that the values reported in COMPUSTAT significantly differ from the values reported in XBRL SEC filings. Although they do not attempt to compare COMPUSTAT and XBRL SEC filings they find that COMPUSTAT significantly alters numbers reported, specifically 17 (out of 30) variables reported by COMPUSTAT are differ from values reported by XBRL SEC filings. They were able to demonstrate how XBRL data can be utilized in an automated large-scale fashion to extract and process commonly used accounting numbers.

Liu and O'Farrel (2013) examine the ability of XBRL data in terms of improving transparency and quality of financial accounting information as proxiedby forecast accuracy. Their results found a significant improvement in analyst forecast accuracy since XBRL mandates.

Henselmann et. al (2015) state that the XBRL data may provide the SEC and investors a simple measure to flag financial reports carrying higher probability of human interaction. Their study, which was based on XBRL 10-K filings submitted to the SEC between July 2009 and March 2013, measured a firm-year-level of abnormal digit frequency and explored its association with earnings quality. Their findings are consistent with the underlying assumption that higher manipulation of earnings is reflected in higher irregularities in the frequency of digits in accounting numbers reported in the financial reports, which may indicate lower earnings quality.

Although XBRL data and its study is still at the early stage these studies suggest that XBRL data is a useful and accurate tool for financial statement analysis and may be used to predict the direction of future movement in earnings.

Replication of the Ou and Penman (1989) study using XBRL
As stated, the study of XBRL is still in the early stages, however, there have been a couple of attempts to use XBRL data in replicating Ou and Penman (1989) and predict changes in earnings.

Williams (2015) investigated whether XBRL company filings, filed in the years 2007-2009 may be used in models attempting to predict future earnings. The study examined whether 70 accounting concepts, extracted from S&P 500 companies XBRL filings, provided adequate data needed to create earnings prediction models (Ou and

Penman, 1989; Abarbanell and Bushee, 1998) and what modifications would make XBRL much more useful. The findings of the study were that XBRL filings, during the investigated period, could not be used to create earnings prediction models; however, adjusting the data, by populating any missing accounting concepts, was shown to be more useful.

Baranes and Palas (2017), reexamined the usefulness of XBRL S&P 500 company filings in the prediction of future earnings, in the years 2011-2015. Their results show that these filings, without any modifications, were not only useful in predicting future earnings changes, based on the Ou and Penman (1989) model, but provided better predictions than previous models using COMPUSTAT data.

These recent studies suggest that XBRL filings may be used in to predict future earnings changes.

The current study attempts to expand the above mentioned research by widening the data size and using a large scale data base of XBRL filings to examine its usefulness in predicting future earnings changes.

DATA AND METHOD
XBRL
XBRL uses meta information to describe data items and link them together through various relationships. In order for the data to be compared across companies the same taxonomy must be used by all filers. Therefore, the SEC has created the XBRL U.S. GAAP Financial Reporting Taxonomy. This taxonomy defines common rules on how to present standard accounting information in XBRL filings. For companies that wish to file information that is not standard (company specific filings) may do so through extensions. Extensions are an important part of XBRL filings that provide additional reporting flexibility, however Debreceny et al. (2011) found that 40 percent of all extensions were unnecessary because the corresponding elements exist in the U.S. GAAP Financial Reporting Taxonomy.

The quarterly financial data was obtained using XBRL Analyst; an Excel plugin that allows users to access the company's XBRL tagged data from its XBRL SEC filing via the XBRL US database. Using this software not only allows for easy access and analysis of the data but also for the calculation of any missing balances. For example, the balance reported in each XBRL filing for total liabilities is not available on the original XBRL filing but is extracted and calculated on the XBRL Analyst. An accounting element may not be available due several reasons: the preparer erroneously marked the information, the SEC's taxonomy did not permit or require this balance.

Data
The sample is of all companies traded on the NYSE on Q1, 2016 who filed with the SEC financial statements in XBRL format. Since all of these firms were required to

report using XBRL by June 15, 2011 (see validity of XBRL), this ensured that the longest time frame could be used for the analysis. The data is from quarterly filings from 1st quarter of 2011 to 3nd quarter of 2016 (23 quarters).

Of the 2,785 tickers listed on the NYSE 394 tickers were for companies that had more than one security and were therefore eliminated from the sample. 538 companies did not have complete financial information in XBRL format. 197 companies were financial institutions (SIC codes 6000-6500, excluding SIC code 6324), because their disclosure and presentations standards differ from other types of companies, and similar to other studies (Williams, 2015), they were eliminated from the sample.

The final sample included 1,656 companies (59.5% of all tickers listed) that were traded on the NYSE on Q1, 2016. The final sample is compatible with previous research using XBRL, Williams (2015) sample included 296 companies (59.2%), and Baranes and Palas (2017) sample included 343 companies (68.6%) of the total population of S&P 500 companies. Table 1 lists descriptive data for these companies.

Table 1 - descriptive data for the study sample

		N	Frequency	Percent
Size (Revenues)	< $100,000,000	1,656	435	26.25
	$100,000,000-$500,000,000	1,656	492	29.69
	$500,000,000-$1,000,000,000	1,656	258	15.57
	$1,000,000,000-$10,000,000,000	1,656	411	24.80
	$10,000,000,000-$100,000,000,000	1,656	59	3.56
	>$100,000,000,000	1,656	1	0.06
Industry (SIC Code)	Agriculture, Forestry and Fishing (01-09)	1,656	3	0.18
	Mining (10-14)	1,656	142	8.57
	Construction (15-17)	1,656	38	2.29
	Manufacturing (20-39)	1,656	609	36.75
	Transportation, Communications, Electric, Gas and Sanitary Services (40-49)	1,656	207	12.49
	Wholesale Trade (50-51)	1,656	54	3.28
	Retail Trade (52-59)	1,656	113	6.82
	Real Estate (60-67)	1,656	231	13.94
	Services (70-89)	1,656	259	15.63
	Public Administration (91-99)	1,656	0	0/00

In the attempt to duplicate the Ou and Penman (1989) study as closely as possible 58 variables were at first used from the original 68 variables. The only variables not included in the first run were those who were not available for a large number of companies (200 or more).

Method

Similar to the Ou and Penman (1989) method, a two-step approach was used to develop the model. In the first step a logistic regression univariate model was used to evaluate the significance of each explanatory variable. Only variables which were found to be associated significantly (at a 10% level) with the direction of earnings per share, above the drift, were maintained. The drift term was estimated as the mean earnings per share change over the four prior quarters to the estimated quarter (see Ou and Penman, 1989).

In the second step, a stepwise logistic regression model was then used to determine the variables to be included in the final model. A two-ways (backward and forward) process of adding and removing variables to minimize the Akaike Information Criterion (AIC) measure of goodness of fit was used and implemented with the R software version 3.2.2. As discussed in (Burnham and Anderson, 2004) the AIC measure has several advantages over the Bayesian Information Criterion (BIC). The first part of the process (backwards) involved a cycle of including all the remaining variables in a single regression, and then progressively removing those that did not prove significant based on the AIC measure of goodness. The same process was repeated (forward) by starting with one variable, measuring the AIC and then adding another variable. A variable was considered insignificant if the total AIC score of the model increased by adding another variable.

A different model was developed for each of the quarters for which a forecast was made, using quarterly data from the previous three years of observations – for example, the forecast period for Q3, 2015, is Q2, 2013 to Q2, 2015. This approach deviates from the method used by Ou and Penman (1989), who used the same model to arrive at a probability of the directional movement in EPS for all subsequent periods. The method adopted was the one used by Bird et al. (2001), who developed a different model for each of the periods the forecasts were made.

The logistic models, were then used to provide a forecast of the probability that the company of it EPS for the next quarter being above its current EPS. Based on these probabilities the stock can be classified. A company stock is assigned to a 'long' position (EPS are expected to increase) if the probability is greater than 0.6, and to a 'short' position (EPS are expected to decrease) if the probability is less 0.4.

The same methodology used for the total sample was then used for the industry analysis. The sample was divided into ten industries based on SIC codes (Table 1). Two industries, Agriculture, Forestry and Fishing (SIC codes 01-09), and Public Administration (SIC codes 91-99) were eliminated because of the small number of companies.

MODELS FOR THE TOTAL SAMPLE

On the first run all 58 variables were used, a list of the variables found significant in each model is presented in Table 2.

Table 2: Results of the logistic regressions for predicting Q2 2015 through Q1 2016

Variables	Q3/2015	Q4/2015	Q1/2016	Q2/2016
Change in Net Profit Margin	-0.00542	-0.00545	-0.00345	-0.02269
Change in Working Capital / Total Assets	-0.00301	-0.00309	-0.00265	
Change in Long Term Debt / Equity		-0.00013	-0.00013	-0.00015
Change in Pretax Income / Sales			-0.00776	-0.00110
Change in EBITDA / Sales			-0.00035	-0.00343
Net Profit Margin	-0.00104			
Sales / Total Accounts Receivables		0.00022		
Change in Capital Expenditure / Total Assets				-0.00006
Accuracy %	0.867	0.667	0.526	0.769
Portfolio Size	15	18	19	38
Number of companies used in model	1,507	1,367	1,149	1,429
Percentage of Portfolio size	1.00%	1.32%	1.65%	2.66%

The number of variables found significant in the different models range from 3 to 5 for each model, the total number of variables found significant for all models is 8. Ou and Penman (1989) found between 16-18 variables, Bird et al. (2001) found 12 to 18 variables, and Baranes and Palas (2017) found 3-9 variables. Only one of the variables (Change in Net Profit Margin) was common for all the models, two variables (Change in Working Capital to Total assets and Change in Long Term Debt to Equity) was common to three of the four models, two variables (Change in Pretax Income to Sales and Change in EBITDA to Sales) was common to two of the models, and the other three variables were specific to only one model.

While the variable common to all models, Change in Net Profit Margin, does not appear in either Ou and Penman (1989) model or Bird et al. (2001) models, the other prominent variables (variables which appear in more than one model), appear in some of the models described.

The Model Forecasts

The accuracy of the forecasts are judged on the basis of the percentage of companies classified as 'long' that actually experienced an increase in EPS and those classified as 'short' that actually experience a decrease in EPS. The accuracy of the models (presented in Table 2) ranges between 53% - 87%, with an average of 70.7%. These results are better than those presented by Ou and Penman (1989) which averaged 67% and those of Bird et al. (2001) which ranged between 60-67%.

Although the models present an impressive accuracy rate, not achieved by previous studies, the use of XBRL company filings data as forecasting tool should still be approached cautiously. The first issue has to do with the portfolio size, that is the number of companies that were assigned a probability position (long, short). The number of companies used in each model is different, based on the availability of their information for each period. The percentage of companies which were assigned a probability position ranges between 1.00% and 2.66%. This means that more than 97% of the companies could not be classified based on the prediction models. Ou and Penman (1989), using COMPUSTAT data, were able to classify approximately 50% of the companies within the same probability range (between 0.6 and 0.4).

An attempt was made to reduce the number of original variables by discarding those variables with less than 10,000 observations (maximum observations = 22*1,656= 36,432). This attempt did not yield larger portfolios or better accuracy.

MODELS BASED ON INDUSTRY

For the Industry analysis a list of significant variables was found for each industry and is presented in Table 3.

Table 3: The significant variables in each industry

	Total Sample	Mining (10-14)	Construction (15-17)	Manufacturing (20-39)	Transportation, Communications, Electric, Gas and Sanitary Services (40-49)	Wholesale Trade (50-51)	Retail Trade (52-59)	Real Estate (60-67)	Services (70-89)
Number of variables in each industry	8	9	8	11	9	7	12	12	12
ROA		X	X		X		X		
Change Pre Tax Income/Sales	X			X			X	X	X
Change Net Profit Margin	X			X		X			
Change Dividends per share					X				X
Depreciation / PP&E							X	X	
Change Inventory							X		X
Net Profit Margin	X							X	X
Change Days in Accounts Recv		X							
Change in Inventory Turnover			X						
Change Research and Development Expense				X					
Change Production				X					
Change Total Debt							X		
Change Working Capital							X		
Change in Operating Income/ Total Assets								X	
Current Ratio									X
Pre Taxe Income/Sales									
Sales / Total Assets									
Change in Working Capital/ Total Assets	X								
Change in Capital Expenditure to Total Assets	X								
Accuracy %	0.707	0.720	0.694	0.643	0.698	0.463	0.778	0.789	0.659
Average Portfolio Size	22.5	35	12.8	16	26.5	4.8	35.3	25.8	7.8
Number of companies used in model	1363	93	20.3	217.3	75.3	41	75.3	136	11.3
Percentage of Portfolio size	1.65%	37.63%	63.05%	7.36%	35.19%	11.71%	46.88%	18.97%	69.03%

The total number of variables found significant in the different industries range from 7 to 12 for each industry, the total number of variables found significant for the entire sample is 8. The variables which were found significant in three to four quarters for each industry, are presented in Table 3. Only two of the variables (ROA and Change in Pretax Income/Sales) are common to four industries, and only Change in Pretax Income/Sales is also common to the total sample.

These results suggest that different industries have different characteristics which are represented in a different model for each industry.

The Model Forecasts

The average accuracy of the quarterly models for each industry (presented in Table 3) range between 46% and 79% with an average of 68%. While these results are not as good as for the entire sample (an average of 70.7%) these are still better than those presented by Ou and Penman (1989) and those of Bird et al. (2001).

Analysis by industry does present a large improvement in the portfolio size created by each model. Percentage of companies which were assigned a probability position ranges between 7.36% and 69.03%, which is much larger than the largest portfolio in the total sample (2.66%). This means that models created per industry can classify a larger number of companies and predict for them the direction of earnings.

CONCLUSIONS

The focus of this study has been to examine previously researched earning prediction models utilizing the newly mandated accounting data format of XBRL (replicating the Ou and Penman,1989 study and the Bird et al. 2001 study). The use of XBRL allows not only easier access to the data but also the ability to adjust the models almost immediately as current information is posted, thus providing a much more relevant tool for investors.

The findings of the study suggest that XBRL data can be used in a large scale financial statement analysis, in both research and investment, as viable data source. The models developed provided a higher accuracy rate than that of previous studies (Ou and Penman,1989; Bird et al. 2001, and others). Further analysis of the data, disaggregation based on industry membership, further improved the results.

This study contributes to previous research by expanding the uses of XBRL findings to predict future earnings which may create a profitable investment strategy. However, there are still limitations which need to be further explored.

One limitation is the relatively short time period data (from 2011) of the SEC XBRL mandate. The short time period not only limits the amount of data available but may also cause other problems such as inconsistencies, errors, or unnecessary extensions in the XBRL filings (Debreceny et al. 2011; Du et al. 2013). However, given that there are indications that XBRL quality increases over time (Du et al. 2013), the methodology may be tested again in the future.

Another limitation is the portfolio size of the final prediction model, the investment portfolio, based on the total sample, contains less than 3% of companies examined. Although this is not necessarily a problem for investors because, as suggested by Ou and Penman (1989), a smaller portfolio may yield higher returns, this might suggest that the ability of XBRL data to predict future earnings may be limited. This limitation was alleviated when the investment portfolios were industry based.

There are several possible extensions of this study, using XBRL data, among them increasing the data size, developing methods of populating missing components and implementing more advance methodologies for the ratio analysis. A very useful extension is the utilization of the prediction model to create an investment portfolio.

Appendix 1 **– Fluctuations of leading market indexes over time (Yahoo Finance)**
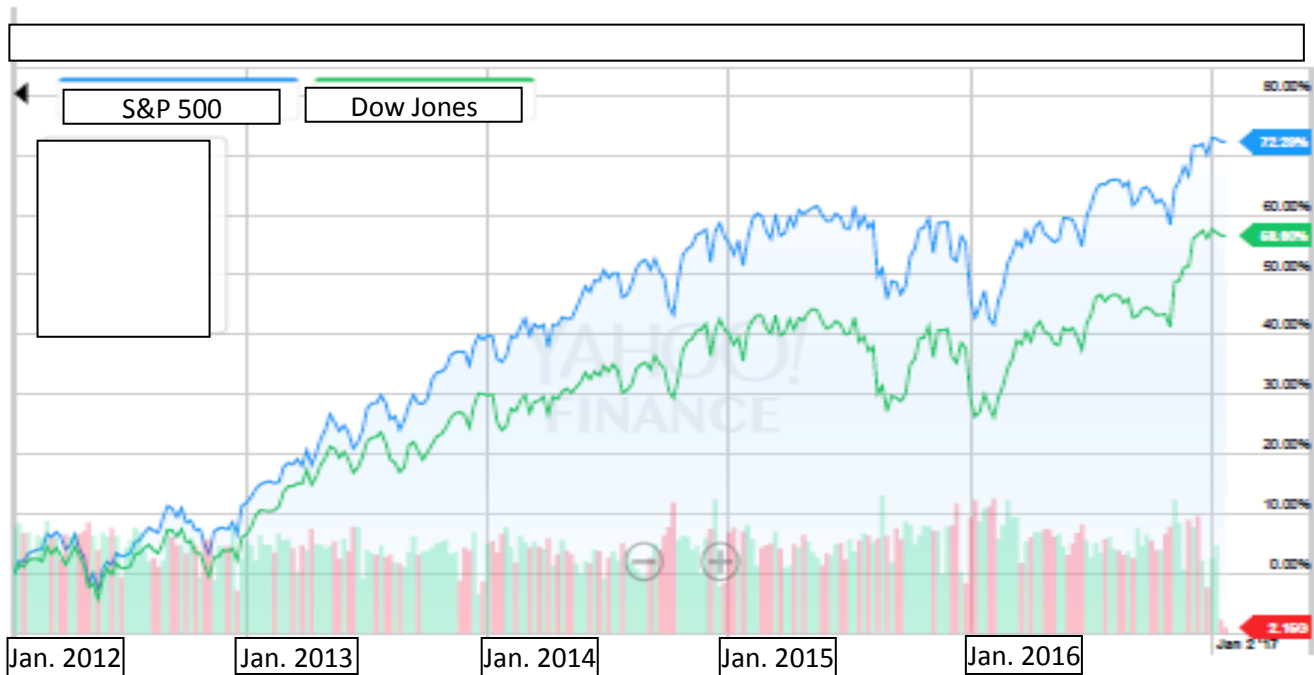
References

Abarbanell, J. S. and B. J. Bushee. 1998. "Abnormal Returns to a Fundamental Analysis Strategy." *The Accounting Review* 73 (1): 19-45.

Alam, P. and C. A. Brown. 2006. "Disaggregated Earnings and the Prediction of ROE and Stock Prices: A Case of the Banking Industry." *Review of Accounting & Finance* 5 (4): 443-463.

Amrhein, D. G., S. Farewell, and R. Pinsker. 2009. "REA and XBRL GL: Synergies for the 21st Century Business Reporting System." *International Journal of Digital Accounting Research* 9: 127-152.

Ball, R. and P. Brown. 1968. "An Empirical Evaluation of Accounting Income Numbers." *Journal of Accounting Research* 6, no. 2: 159-178.

Ball, R. and L. Shivakumar. 2008. "How Much New Information Is There in Earnings?." *Journal of Accounting Research* 46, no. 5: 975-1016.

BEAVER, W. H. 1968. "The Information Content of Annual Earnings Announcements." *Journal of Accounting Research* 6, no. 3: 67-92.

Baranes, A. and R. Palas, 2017." The Prediction of Earnings Movements Using Accounting Data: Using XBRL", *International Journal of Accounting Research* 4, no. 2 (pending).

Bernard, V.L. and Thomas, J.K. 1990, "Evidence That Stock Prices Do Not Fully Reflect the Implications of Current Earnings for Future Earnings", *Journal of Accounting and Economics*, vol. 13, no. 4, pp. 305.

Bernard, V. L., and J. K. Thomas. 1990. "EVIDENCE THAT STOCK PRICES DO NOT FULLY REFLECT THE IMPLICATIONS OF CURRENT EARNINGS FOR FUTURE EARNINGS." *Journal Of Accounting & Economics* 13, no. 4: 305-340.

Bird, R., R. Gerlach, and A.D. Hall. 2001. "The prediction of earnings movements using accounting data: An update and extension of Ou and Penman." *Journal Of Asset Management* 2, no. 2: 180.

BLOOMFIELD, R. J., R. LIBBY, and M. W. NELSON. 2003. "Do Investors Overrely on Old Elements of the Earnings Time Series?." *Contemporary Accounting Research* 20, no. 1: 1-31.

Boritz, J. E., and W.G No. 2008. "The SEC's XBRL Voluntary Filing Program on EDGAR: A Case for Quality Assurance." *Current Issues In Auditing* 2, no. 2: A36-A50.

Boritz, J. E., and W.G No. 2013 "The Quality of Interactive Data: XBRL Versus COMPUSTAT, Yahoo Finance, and Google Finance". Available at SSRN: http://ssrn.com/abstract=2253638 or http://dx.doi.org/10.2139/ssrn.2253638

Burnham, K. P. and D. R. Anderson. 2004. "Multimodel Inference: Understanding AIC and BIC in Model Selection." *Sociological Methods and Research* 33 (2): 261-304.

Cairney, T and L.B. Fletcher. 2009. "Are NAICS Industries more Homogeneous than SIC Industries?" *Academy of Accounting and Financial Studies Journal* 13 (3): 27-44.

Chychyla, R. and A. Kogan. 2015. "Using XBRL to Conduct a Large-Scale Study of Discrepancies between the Accounting Numbers in Compustat and SEC 10-K Filings." *Journal of Information Systems* 29 (1): 37.

D'Souza, Julia,M., K. Ramesh, and Min Shen. 2010. "The Interdependence between Institutional Ownership and Information Dissemination by Data Aggregators." *The Accounting Review* 85 (1): 159-193.

Debreceny, R. S., S. M. Farewell, M. Piechocki, C. Felden, A. Gräning, and A. d'Eri. 2011. "Flex Or Break? Extensions in XBRL Disclosures to the SEC." *Accounting Horizons* 25 (4): 631-657.

Du, H., M. A. Vasarhelyi, and X. Zheng. 2013. "XBRL Mandate: Thousands of Filing Errors and so what?" *Journal of Information Systems* 27 (1): 61.

Finger, C. A. 1994. "The Ability of Earnings to Predict Future Earnings and Cash Flow." *Journal of Accounting Research* 32 (2): 210.

Foster, G., C. Olsen, and T. Shevlin. 1984. "Earnings Releases, Anomalies, and the Behavior of Security Returns." *The Accounting Review* 59 (4): 574.

Henselmann, K., D. Ditter, and E. Scherr. 2015. "Irregularities in Accounting Numbers and Earnings Management - A Novel Approach Based on SEC XBRL Filings." *Journal of Emerging Technologies in Accounting* 12 (1): 117.

Holthausen, R. W. and D. F. Larcker. 1992. "The Prediction of Stock Returns using Financial Statement Information." *Journal of Accounting & Economics* 15 (2): 373.

Jordan, C. E., S. J. Clark, and Mallory Donald. 2009. "USING FINANCIAL STATEMENT ANALYSIS TO EXPLAIN THE VARIATION IN FIRMS' EARNINGS-PRICE RATIOS." *Academy Of Accounting & Financial Studies Journal* 13, no. 1: 91-101.

Kinney, M. R. and E. P. Swanson. 1993. "The Accuracy and Adequacy of Tax Data in COMPUSTAT." *The Journal of the American Taxation Association* 15 (1): 121.

Lee, Chi-Wen Jevons. 1985. "Stochastic Properties of Cross-Sectional Financial Data." *Journal Of Accounting Research* 23, no. 1: 213-227

Liu, C. and G. O'Farrell. 2013. "The Role of Accounting Values in the Relation between XBRL and Forecast Accuracy." *International Journal of Accounting and Information Management* 21 (4): 297-313.

Ou, J. A. 1990. "The Information Content of Nonearnings Accounting Numbers as Earnings Predictors." *Journal of Accounting Research* 28 (1): 144.

Ou, J. A. and S. H. Penman. 1989. "Financial Statement Analysis and the Prediction of Stock Returns." *Journal of Accounting & Economics* 11 (4): 295.

Ou, J. A., S. H. Penman, and D. F. Larcker. 1989. "Accounting Measurement, Price-Earnings Ratio, and the Information Content of Security Prices; Discussion." *Journal of Accounting Research* 27: 111.

Penman, S. H. 1992. "Return to Fundamentals." *Journal of Accounting, Auditing & Finance* 7 (4): 465.

Penman, S. H. and X. Zhang. 2002. "Accounting Conservatism, the Quality of Earnings, and Stock Returns." *The Accounting Review* 77 (2): 237-264.

Rosenberg, B. and M. Houglet. 1974. "ERROR RATES IN CRSP AND COMPUSTAT DATA BASES AND THEIR IMPLICATIONS." *The Journal of Finance* 29 (4): 1303.

Roychowdhury, S. and E. Sletten. 2012. "Voluntary Disclosure Incentives and Earnings Informativeness." *The Accounting Review* 87 (5): 1679-1708.

San Miguel, J.G. 1977. "THE RELIABILITY OF R&D DATA IN COMPUSTAT AND 10-K REPORTS." *The Accounting Review* 52 (3): 638.

Securities and Exchange Commission (SEC). 2009. Interactive Data to Improve Financial Reporting. Release Nos. 33-9002, 34-59324; 39-2461; IC-28609. Washington, DC: SEC.

Securities and Exchange Commission (SEC). 2011. Response of the Division of Corporation Finance; Response of the Office of the Chief Accountant. Division of Corporation Finance No-Action Letter.

Setiono, B. and N. Strong. 1998. "Predicting Stock Returns using Financial Statement Information." *Journal of Business Finance & Accounting* 25 (5): 631-657.

Shroff, P. K. 1999. "The Variability of Earnings and Non-Earnings Information and Earnings Prediction." *Journal of Business Finance & Accounting* 26 (7): 863-882.

Stober, T. L. 1992. "Summary Financial Statement Measures and Analysts' Forecasts of Earnings." *Journal of Accounting & Economics* 15 (2): 347.

Tallapally, P., M. S. Luehlfing, and M. Motha. 2011. "The Partnership of EDGAR Online and XBRL - should Compustat Care?" *The Review of Business Information Systems* 15 (4): 39-46.

Vasarhelyi, M. A., D. Y. Chan, and J. P. Krahel. 2012. "Consequences of XBRL Standardization on Financial Statement Data." *Journal of Information Systems* 26 (1): 155-167.

Wahlen, J. M. and M. M. Wieland. 2011. "Can Financial Statement Analysis Beat Consensus Analysts' Recommendations?" *Review of Accounting Studies* 16 (1): 89-115.

Williams, K. L., 2015. "The prediction of future earnings using financial statement information: Are XBRL company filings up to the task?" PHD dissertation, Patterson School of Accountancy The University of Mississippi.

Winograd, B.N., J.S. Gerson and B.L. Berlin. 2000. "Audit Practices of PricewaterhouseCoopers." *Auditing: A Journal of Practice & Theory*, 19 (Fall): 175-182.

Yang, D. C., M. A. Vasarhelyi, and C. Liu. 2003. "A Note on the using of Accounting Databases." *Industrial Management & Data Systems* 103 (3): 204-210.