IJIRR

International Journal of Information Research and Review
Vol. 04, Issue, 07, pp.4281-4287, July, 2017

RESEARCH ARTICLE

MULTIVARIATE IMPUTATION OF XBRL DATA FOR FINANCIAL STATEMENT ANALYSIS

[1]Amos Baranes and *[2]Rimona Palas

[1]Peres Academic Center, 8 HaNeviim St., POB 328, Rehovot, Israel
[2]Accounting Department, School of Business, College of Law and Business, #24 Ben Gurion St., Ramat Gan, Israel

ARTICLE INFO

Article History:

Received 27th April, 2017
Received in revised form
29th May, 2017
Accepted 17th June, 2017
Published online 26th July, 2017

Keywords:

Accounting information,
Earnings prediction,
Investment strategy,
XBRL,
Multivariate imputation.

ABSTRACT

XBRL company filingsprovide immediate availability and easy accessibility, for both researchers and investors, for financial statement analysis. The objective of this study is to examine whether large scale XBRL data can be used to predict the direction of movement of earnings. The study analyzes companies' XBRL filings of quarterly data 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. The results classified the companies as ones that would realize an increase, or a decrease, in earnings. Although the final model indicated an ability to predict subsequent earnings changes on average about 67% of the time, (similar to those of previous studies based on COMPUSTAT), it based the models on about 23% of the entire sample examined, and could classify less than 10% of the entire sample. A Multivariate Imputation by Chained Equations (MICE) was implemented to fill in the missing data. This increased the number of useable observations by about 11%, and increased the number of observations in the final models by 150%. The models utilized 56% of the original companies (more than double) and classified 27% of the original companies (about triple), and still increased the accuracy of prediction to 68%. These results suggest that XBRL data with imputation can be used as a financial statement analysis tool.

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 *et al.,* 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. This study does not attempt to provide better or newer financial statement analysis tools or investment strategies than previous studies, but rather examine the use of a new database. 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 (Ball, 1968; 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 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 Larcker, 1992; Bernard *et al.,* 1997; Stober, 1992; Setiono and Strong, 1998; Bird *et al.,* 2001) has been mixed. 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 use as a tool for investment decisions. However, the main objective is to examine 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

****Corresponding author: Rimona Palas***
Accounting Department, School of Business, College of Law and Business, #24 Ben Gurion St., Ramat Gan, Israel.

that enable 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. Despite the fact that 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, although 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 (Ou and Penman, 1989), XBRL data is immediately available. 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 (Miguel, 1977; Kinney and Swanson, 1993; Tallapally *et al.,* 2011) and data found in other accounting databases (Rosenberg and M. Houglet, 1974; Yang *et al.,* 2003). However, while there is still not enough research regarding the reliability of XBRL data, studies up to date seem positive: (Boritz and No, 2013; Henselmann *et al.,* 2015).

Ref (Vasarhelyi *et al.,* 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 (Williams, 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 a significant 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.

MATERIALS 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. Using the NASDAQ company list all 6,726 tickers listed on one of the three major US stock exchanges (AMEX, NASDAQ, and NYSE) were found. The quarterly financial data was obtained using XBRL Analyst (created by FinDynamics); 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.

Data

Of all 6,726 tickers, only 4,380 of the companies that were traded on Q1, 2016 filed with the SEC financial statements in XBRL format. Since all firms were required to report using XBRL by June 15, 2011, 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 4,380 tickers listed on the different stock exchangesthe following tickers were removed: 365 tickers for non-common stocks;387 tickers for companies with IPO's between 2012 and 2017; and 25 tickers for companies with more than one ticker (the same CIK). The final sample included 3,603 companies (53.6% of all tickers listed) that were publicly traded 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. In the attempt to duplicate the (Ou and Penman, 1989) study as closely as possible 68 variables were extracted from the data. It should be noted that some of the variables had to be calculated from the original filing, whereas some variables were already calculated as part of the XBRL Analyst tool. This database contained 79,191 records. In order to calculate growth variables and drifts, additional records were eliminated, which left 58 variables and 60,498 records. Additional records were removed in three stages. In the first stage every company that had more than 35% of the variables missing (20 variables) was removed, this stage removed 9.44% of the records and left 54,787 records. In the second elimination stage, every variable which had more than 15% missing data points was eliminated. This left 38 explanatory variables for the entire sample. Once these two stages were implemented a third stage, the removal of outliers (for both variables and stock returns) was implemented. Removal of outliers is important because it can drastically bias/change the fit estimates and predictions. In order to identify the outliers Interquartile range (IQR) method (see (Barbato *et al.,* 2011)) was used. Based on this method the data is arranged by value (from the lowest to the highest value) and is divided into four quartiles. The lowest quartile values (under 25%) is Q_1 and the highest quartile (over 75%) is Q_3, the interquartile range (IQR) is the range between Q_1 and Q_3, and therefore covers 50% of the data. The lower limit is computed as Q_1 – 1.5xIQR, and the upper limit is Q_3 + 1.5xIQR, any data value beyond these limits was recognized as an outlier and eliminated. Once all discussed data was removed 36 variables remained.

METHODS

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 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's EPS for the next quarter will be 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 MODELS

In the first run all 36 variables were used, a list of the variables found significant in each model is presented in Table 2. The number of variables found significant in the different models range from 9 to 15 (an average of 11.75) for each model, the total number of variables found significant for all models is 21. (Ou and Penman, 1989) found between 16-18 variables, and (Bird *et al.,* 2001) found 12 to 18 variables. Five of the variables (Δ Net Profit Margin, ROA,Δ Days sales to Accounts Recv, Δ Quick Ratio, Operating Income to Total Assets, and Δ Equity to Fixed Assets) were common for all the models, eight variables were common to two of the four models, and the other seven variables were specific to only one model. Of the five prominent variables (variables which appear in all four models), only two (Δ Quick Ratio, and Operating Income to Total Assets) appear in the [10]model, and three (Δ Net Profit Margin, ROA, and Operating Income to Total Assets) appear in the (Bird *et al.,* 2001) models.

The Model Forecasts

The accuracy of the forecasts is 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 3) ranges between 66% - 70%, with an average of 67.02%. These results are similar to the results presented by Ou and Penman (1989) which averaged 67% and those of Bird *et al*. (2001) which ranged between 60-67%. However, it should be noted that a very small number of companies, 23.1% were utilized(an average of 833 companies) in determining the models, out of the entiresample(3,603 companies, see Table 1).Of the companies utilized, the models were only able to classify an average of 42%, that is less than 10% (348 companies) of the entire sample.

Data Imputation

The main problem with the models presented is their inability to model many of the companies (only 23%) so that even if the models are able to classify approximately 42% of the companies utilized they create relatively small portfolios that include less than 10% of the entire sample (average of 348 out of 3,603). One of the reasons that the models could not use more data is because the data was not available, (Williams, 2015) found that 67% of variables would be incalculable or would return erroneous results with the data directly extracted from XBRL filings. [24]could only match approximately 70% of XBRL filings data to COMPUSTAT filing data. An accounting element may not be extracTable froman XBRL company filing due to several reasons, among them: the preparer erroneously did not tag the accounting element, the preparer used the wrong tag for an accounting element, or the SEC'sprotocol for the preparation of XBRL company filings set forth in the EDGAR FilerManual did not permit or require a tag. According to (Rubin, 1996) in order to overcome this problem of complex incomplete data, multiple imputation is the best method to be employed. There are several approaches for imputing multivariate data, Multivariate Imputation by Chained Equations (MICE) is considered to be a better alternative in cases where no suiTable multivariate distribution can be found. MICE specifies the multivariate imputation model on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable. Starting from an initial imputation, MICE draws imputations by iterating over the conditional densities. For the purpose of this study the package of MICE in R was implemented, while the package provides five iterations for implementation, only the first one was used for the current analysis. Table 4 presents changes from the original data (data) to that of the data with imputation (full data). The number of observations increased by about 10%, however this small change allowed for the most important change, and that is the number of companies that are were utilized by the models, which increased by an average of 144%. This means that more than twice as many companies may be examined by the models and used in the classification for prediction purposes.

The models based on data with imputation

In the first run all 36 variables were used, a list of the variables found significant in each model is presented in Table 5. The number of variables found significant in the different models range from 13 to 18 (an average of 16) for each model, the total number of variables found significant for all models is 26, with the original data only 21 variables were found significant for all models.

Table 1. Descriptive Data for the Study Sample

		N	Frequency	Percent
Stock Exchange	AMEX	3,603	12	0.33%
	NASDAQ	3,603	1702	47.24%
	NYSE	3,603	1889	52.43%
Size (Revenues)	< $10,000,000	3,603	612	16.99%
	$10,000,000- $100,000,000	3,603	1062	29.48%
	$100,000,000-$500,000,000	3,603	998	27.70%
	$500,000,000-$1,000,000,000	3,603	358	9.94%
	$1,000,000,000-$10,000,000,000	3,603	515	14.29%
	$10,000,000,000-$100,000,000,000	3,603	57	1.58%
	>$100,000,000,000	3,603	1	0.03%
Industry (SIC Code)	Agriculture, Forestry and Fishing (01-09)	3,603	12	0.33%
	Mining (10-14)	3,603	181	5.02%
	Construction (15-17)	3,603	52	1.44%
	Manufacturing (20-39)	3,603	1329	36.89%
	Transportation, Communications, Electric, Gas and Sanitary Services (40-49)	3,603	310	8.60%
	Wholesale Trade (50-51)	3,603	104	2.89%
	Retail Trade (52-59)	3,603	204	5.66%
	Real Estate (60-67)	3,603	861	23.90%
	Services (70-89)	3,603	550	15.27%
	Public Administration (91-99)	3,603	0	0.00%

Table 2. Results of the Logistic Regressions for Predicting Q3 2015 through Q2 2016

Variables	Q3/2015	Q4/2015	Q1/2016	Q2/2016
Δ Net Profit Margin	-0.5920	-0.5030	-0.5220	-0.5232
ROA	-37.2309	-20.2390	-20.3294	-35.4367
Δ Days sales to Accounts Recv.	1.2420	1.4173	1.3564	1.2439
Δ Quick Ratio	-0.5556	-0.4951	-0.6006	-0.6457
Operating Income to Total Assets	14.1379	10.7761	10.4624	13.8259
Δ Equity to Fixed Assets	-1.7640	-1.5512	-1.3140	-1.7376
Δ Capital Expenditures to Total Assets	0.0719			0.0455
Δ Total Revenue	-0.7836	-0.7640		
Sales to Total Assets	-0.7746			-0.2316
Sales to Fixed Assets	0.0687			0.0226
Working Capital to Total Assets	0.3499			0.2717
Sales to Total Accounts Recv.		-0.0072	-0.0125	
Δ Sales to Total Assets			-0.7362	-0.8982
Δ Working Capital			0.2434	0.2114
Δ Capital Expenditures to Total Assets	0.1810			
Long Term Debt to Equity	-0.1107			
Equity to Fixed Assets	-0.0189			
Current Ratio	-0.0066			
ROCE		-7.1971		
Return on Operating Expenditures			-6.8534	
Δ Capital Expenditures to Total Assets				0.0701

Table 3. Accuracy and Portfolio size

	Q3/2015	Q4/2015	Q1/2016	Q2/2016	Average
Accuracy	66.09%	70.00%	66.15%	65.85%	67.02%
Number of companies used in model	836	826	853	816	832.75
Portfolio Size	329	336	374	354	348.25
Percentage of Portfolio size	39.35%	40.68%	43.85%	43.38%	41.81%

Table 4. Changes in Data due to Imputation

	Q3/ 2015		Q4/ 2015		Q1/ 2016		Q2/ 2016		Average
	Data	Full Data	Data	Full Data	Data	Full Data	Data	Full Data	
Total observations	23,403	25,895	23,631	26,160	23,760	26,329	23,917	26,537	
Change		11%		11%		11%		11%	10.78%
Observ. in final model	9,011	25,895	12,611	26,160	11,300	26,329	9,779	26,537	
Change		187%		107%		133%		171%	149.79%
# Variables in model	15	13	9	14	10	18	13	18	
Change		-13%		56%		80%		38%	40.17%
# Companies in model	836	2,214	826	1,972	853	2,016	816	1,920	
Change		165%		139%		136%		135%	143.80%

Table 5. Results of Predicting Q3 2015 through Q2 2016 Full Data

Variables	Q3/2015	Q4/2015	Q1/2016	Q2/2016
Δ Net Profit Margin	-0.4525	-0.4453	-0.4402	-0.4519
ROA	-15.5055	-16.8892	-18.5010	-19.5109
Δ Equity to Fixed Assets	-2.0725	-1.7947	-1.7543	-2.4537
Δ Days sales to Accounts Recv.	0.8071	0.8706	0.8763	0.9137
Δ Quick Ratio	-0.5683	-0.5909	-0.5485	-0.5687
Operating Income to Total Assets	7.6220	6.4203	8.5289	10.1506
Net Profit Margin	-1.6839	-1.5142	-1.4182	-0.8134
Return on Operating Expenditures	-4.2742	-4.2676	-3.8261	
Sales to Total Assets	-0.3197		-0.3104	-0.3690
Pretax Income to Sales	0.5457	0.7506	0.6266	
Quick Ratio	-0.0191	-0.0236	-0.0278	
Δ Capital Expenditures to Total Assets		-0.0509	-0.0444	-0.0521
Δ Sales to Total Assets	-0.7386	-0.6431		
Sales to Fixed Assets		-0.0178	-0.0153	
Days Sales Accounts Recv.		0.0002	0.0002	
EBITDA to Sales			-0.3173	-0.3096
Δ Total Assets			1.5068	1.9125
Δ Total Revenue			-0.8867	-0.9298
Δ Capital Expenditures to Total Assets	0.0765			
Δ Production			0.1760	
ROCE				-3.2999
Δ Total Depreciation				1.8295
Working Capital to Total Assets				-0.1158
Δ Operating Income to Total Assets				-0.0950
Δ Pretax Income to Sales				0.0966
Δ Production				0.1667

Table 6. Accuracy and Portfolio Size– Full Data

	Q3/2015	Q4/2015	Q1/2016	Q2/2016	Average
Accuracy	66.32%	72.96%	66.73%	66.57%	
Number of companies used in model	2,214	1,972	2,016	1,920	2,031
Portfolio Size	966	957	952	957	958
Percentage of Portfolio size	43.63%	48.53%	47.22%	49.84%	47.31%

Seven of the variables (Δ Net Profit Margin, ROA, Δ Equity to Fixed Assets, Δ Days sales to Accounts Recv, Δ Quick Ratio, Operating Income to Total Assets, and Net Profit Margin) were common for all the models, five variables were common to three models, six variables were common to two of the four models, and the other eight variables were specific to only one model. Of the seven variables, common to all four models, five were also common to the previous models (before imputation). The stability of the model can be measured by the fact that of the 26 variables in the model twelve variables (46%) were common to 3-4 of the models, compared to five out of 21 (24%) variables from the previous models (before imputation).

The Model Forecasts

The accuracy of the models (presented in Table 6) ranges between 66% - 73%, with an average of 68.15% compared to the models based on the original data which averaged 67.02%. However, the most important issue is the significant change is in the number of companies, the number of companies utilized in the models increased on average by 144% (see Table 4) and now the model utilized about 56% (an average of 2,030 companies) of the entire sample (3,603 companies, see Table 1), where with the previous models only 23.1% was used (an average of 833 companies) in determining the models. The models with the full data were also able to classify an average of 47% of the companies utilized, compared to the previous models which classified only 42%.Of the entire sample of 3,603 companies, more than 26% were classified by the models with the full data, as opposed to less than 10% classified by the previous models.

The implication of these results is that not only were the models with full data able to classify more companies, they were able to this without losing the ability to accurately classify the companies as increasing or decreasing in earnings.

Conclusions

The focus of this study has been to examine the use of the newly mandated accounting data format of XBRL on a large scale in previously researched earning prediction models (replicating the (Ou and Penman, 1989) study and the (Bird *et al.,* 2001) study). The use of XBRL allows not only easier and less costly 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, and especially small investors.

The findings of the study suggest that XBRL data can be used in a large scale financial statement analysis, for both research and investment, as viable data source. While the models developed with the original data provided a similar accuracy rate to that of previous studies ((Ou and Penman, 1989; Bird *et al.,* 2001) and others), they were only able to classify a relatively small portion of the companies (less than 10%). Multivariate Imputation by Chained Equations (MICE) was employed to complete the data. The method was able not only to provide more robust models which were able to classify a much larger number of companies (more than 26% of the original companies), but to do so at a higher accuracy rate. This study contributes to previous research by expanding the scope of XBRL filings data used to all company filings and to by enhancing the original data by multivariate imputation. The attempt of the study is not to examine the validity of the prediction models presented, but to see if XBRL data filings may be used in this type of financial statement analysis. The main limitation of this study 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. There are several possible extensions of this study among them developing other methods of populating missing components and implementing more advanced methodologies for the ratio analysis. The passage of time, which will allow higher quality filings, will also enhance the use of the XBRL data.

REFERENCES

Alam P. and C. a. Brown, "Disaggregated earnings and the prediction of ROE and stock prices: a case of the banking industry," *Rev. Account. Financ.,* vol. 5, no. 4, pp. 443–463, 2006.

Ball, P. Ray; Brown, "An Empirical Evaluation of Accounting Income Numbers," *J. Account. Res., vol.* 6, no. 2, pp. 159–178, 1968.

Ball, R. and L. Shivakumar, "How much new information is there in earnings?," *J. Account. Res.,* vol. 46, no. 5, pp. 975–1016, 2008.

Baranes, A. and R. Palas, "The Prediction of Earnings Movements Using Accounting Data: Using XBRL," *Int. J. Account. Res.,* vol. 5, no. 1, 2017.

Barbato, G., E. M. Barini, G. Genta, and R. Levi, "Features and performance of some outlier detection methods," *J. Appl. Stat.,* vol. 38, no. 10, pp. 2133–2149, 2011.

Beaver, W. H. "The Information Content of Annual Earnings Announcements," *J. Account. Res.,* vol. 6, no. 1–2, p. 67, 1968.

Bernard, V. L.and J. K. Thomas, "Evidence that stock prices do not fully reflect the implications of current earnings for future earnings," *J. Account. Econ.,* vol. 13, no. 4, pp. 305–340, 1990.

Bernard, V., J. Thomas, and J. Wahlen, "Accounting-Based Stock Price Anomalies: Separating Market Inefficiencies from Risk," *Contemp. Account. Res.,* vol. 14, no. 2, pp. 89–136, 1997.

Bird, R., R. Gerlach, and A. D. Hall, "'The prediction of earnings movements using accounting data: An update and extension of Ou and Penman' -- a response," *J. Asset Manag.,* vol. 2, no. 2, pp. 180–195, 2001.

Bloomfield, R. J., R. Libby, and M. W. Nelson, "Do Investors Overrely on Old Elements of the Earnings Time Series?," Contemp. *Account. Res.,* vol. 20, no. 1, pp. 1–31, 2003.

Boritz, J. E. and W. G. No, "Electronic copy available at: http://ssrn.com/abstract=1433358," pp. 0–51, 2013.

Burnham, K. P. and D. R. Anderson, "Multimodel Inference\rUnderstanding AIC and BIC in Model Selection," *Sociol. Methods Res.,* vol. 33, no. 2, pp. 261–304, 2004.

Chychyla, R.and A. Kogan, "Using XBRL to Conduct a Large-Scale Study of Discrepancies between the Accounting Numbers in Compustat and SEC 10-K Filings," *J. Inf. Syst.,* vol. 29, no. 1, pp. 37–72, 2015.

D'Souza, J. M., K. Ramesh, and M. Shen, "The interdependence between institutional ownership and information dissemination by data aggregators," Accounting Review, vol. 85, no. 1. pp. 159–193, 2010.

Debreceny, R. S., S. M. Farewell, M. Piechocki, C. Felden, A. Gräning, and A. D'Eri, "Flex or break? Extensions in XBRL disclosures to the SEC," *Account. Horizons,* vol. 25, no. 4, pp. 631–657, 2011.

Du, H., M. a. Vasarhelyi, and X. Zheng, "XBRL Mandate: Thousands of Filing Errors and So What?," *J. Inf. Syst.,* vol. 27, no. 1, pp. 61–78, 2013.

Finger, C. A. "The ability of earnings to predict future earnigns and cash flow.," *J. Account. Res.,* vol. 32, no. 2, pp. 210–223, 1994.

Foster, G., C. Olsen, and T. Shevlin, "Earnings Releases, Anomalies, and the Behavior of Security Returns," *Source Account. Rev.,* vol. 59, no. 4, pp. 574–603, 1984.

Henselmann, K., D. Ditter, and E. Scherr, "Irregularities in Accounting Numbers and Earnings Management – A Novel Approach Based on SEC XBRL Filings," *J. Emerg. Technol. Account.,* vol. 12, no. 1, pp. 117–151, 2015.

Holthausen R. W.and D. F. Larcker, "The prediction of stock returns using financial statement information," *J. Account. Econ.,* vol. 15, no. 2–3, pp. 373–411, 1992.

Kinney, M. R.and E. P. Swanson, "The Accuracy and Adequacy of Tax Data in COMPUSTAT.," *J. Am. Tax. Assoc.,* vol. 15, no. 1, p. 121, 1993.

Miguel, J. G. S. "The Reliability of R&D Data in COMPUSTAT and 10-K Reports," Account. Rev., vol. 52, no. 3, pp. 638–641, 1977.

Ou J. A.and S. H. Penman, "Financial statement analysis and the prediction of stock returns," *J. Account. Econ.,* vol. 11, no. 4, pp. 295–329, 1989.

Ou, J. A. "The Information Content of Nonearnings Accounting Numbers as Earnings Predictors," *J. Account. Res.,* vol. 28, no. 1, pp. 144–163, 1990.

Penman, S. H. and X. J. Zhang, "Accounting conservatism, the quality of earnings, and stock returns," *Account. Rev.,* vol. 77, no. 2, pp. 237–264, 2002.

Rosenberg, B. and M. Houglet, "Error rates in crsp and compustat data bases and their implications," *J. Finance,* vol. 29, no. 4, pp. 1303–1310, 1974.

Rubin, D. B. "Multiple Imputation after 18+ Years," Journal of the American Statistical Association, vol. 91, no. 434. pp. 473–489, 1996.

Setiono, B. and N. Strong, "Predicting stock returns using financial statement information," *J. Bus. Financ. Account.,* vol. 25, no. 5–6, pp. 631–657, 1998.

Stober, T. L. "Summary financial statement measures and analysts' forecasts of earnings," *J. Account. Econ.,* vol. 15, no. 2–3, pp. 347–372, 1992.

Tallapally, P., M. S. Luehlfing, and M. Motha, "The Partnership Of EDGAR Online And XBRL - Should Compustat Care?," *Rev. Bus. Inf. Syst.,* vol. 15, pp. 39–46, 2011.

Vasarhelyi, M. A., D. Y. Chan, and J. P. Krahel, "Consequences of XBRL Standardization on Financial Statement Data," *J. Inf. Syst.,* vol. 26, no. 1, pp. 155–167, 2012.

Williams, K. L. “The Prediction of Future Earnings Using Financial Statement Information: Are XBRL Company Filings up to the Task?,” 2015.

Yang, D. C., M. a. Vasarhelyi, and C. Liu, “A note on the using of accounting databases,” *Ind. Manag. Data Syst.*, vol. 103, no. 3, pp. 204–210, 2003.

****