

## Analysis of Use of XBRL Based Accounting Data in Financial Research

PUSTYLNICK, I., TEMCHENKO, O., GUBARKOV, S.

School of Business

Far Eastern Federal University

Sukhanova St., 8, Vladivostok

RUSSIAN FEDERATION

igorpustylnick@hotmail.com

*Abstract:* - Since Extensible Business Reporting Language (XBRL) has become a language of mandatory statements submission in the U.S., it opened new opportunities for electronic parsing and auditing of these statements. We show the example of automatic extraction of the statement data and propose the improvements to the taxonomy based on Generally Acceptable Accounting Principles (GAAP) of the USA, which may be able to improve the extraction process.

*Key-words:* XBRL Parsing; financial statements; data extraction; data verification; XBRL data; M-Score; Z-Score

### 1 Introduction

Extensible Business Reporting Language (XBRL) was introduced at the beginning of the 21<sup>st</sup> century as a robust language of financial reporting [1]. The XML based language was soon accepted as a primary language of auditing of the corporate financial results [2]. The acceptance of the language as a tool of financial reporting soon followed. Since 2011 U.S. SEC uses XBRL as its language of choice of reporting of financial results by the U.S. business enterprises and the companies performing significant business activities at the territory of the USA.

Considering the size of the U.S. economy and the number of the potential reports produced by the organizations, it may be quite natural to assume that the large number of the financial reports can provide a viable research base for the use of XBRL and the validity of the data it contains.

U.S. based academic researchers working on various financial and accounting topics have been using the data, coming from the financial aggregators, such as YAHOO Finance and COMPUSTAT among others. The accuracy of the data coming from the aggregators have been questioned for its accuracy since the 70s of the 20<sup>th</sup> century [3]. With the appearance of a larger volume of XBRL reports the researchers have been comparing XBRL based financial data with the financial figures coming from COMPUSTAT. They found a large number of discrepancies

between the figures obtained from these sources by comparing data visually [4] or using specially written software [5]. The latter research reported findings on the large number of discrepancies between XBRL based data and COMPUSTAT provided financial feed.

The study by Chychyla and Kogan, mentioned in the previous paragraph, specifically compared the data coming from the aggregator with the data coming directly from the company in the form of year-end filing (known in the U.S. as 10-K filings). The following statement may never be directly proven, but the data coming directly from the company's year-end report must be considered as the only source of unequivocally trustworthy information about the company's financials.

Looking at the XBRL based reports from this angle, it becomes possible to suggest that the data in XBRL format, available free of charge from the financial regulators, such as U.S. SEC can become an alternative to the COMPUSTAT based financial feed, presently used by a large number of researches all over the world.

In this research we took the financial data available from the U.S. SEC for all companies submitting 10-K reports and examined it for suitability for the particular financial research. We based our research on the existing findings, which are based predominantly on the U.S. based data due to the availability and volume of statements.

Further in this paper we will discuss the past research we based upon, followed by the formulation of the research criteria we have used. The criteria will be followed by the discussion of the finding and our own analysis of the results obtained. We conclude this paper with the our own suggestions on the state of XBRL and its potential in financial research, followed by the limitations discovered during the research and suggestions on the future research efforts, which can be undertaken in this direction.

## 2 Problem Formulation

### 2.1 State and Usability of XBRL

At the time when XBRL was announced as the financial reporting tool of choice in the USA several groups of researchers undertook the efforts of evaluating the validity of the financial reports submitted to U.S. SEC.

XBRL as a language is fairly simple and is an extension of widely used Extended Markup Language (XML) in the area of financial reporting resulted in creating own XML Schemas (XSD files) with tags defined for the financial reporting [6].

The group of researchers led by R. Debrecey attempted to extract accounting variables from the XBRL based statements [7]. The group used the help of a large number of research assistants but managed to evaluate a large number of financial statements and accounting variables. As a result, the researchers showed the probabilities of extraction of each variable.

Various countries have accepted various approaches to the creation of XBRL statements and to the definition of the taxonomies. Now, there are two major types of taxonomies (open or extensible) and closed [8]. Closed taxonomies, used in Italy, Ireland, U.K., etc., do not represent a lot of interest in XBRL research. The contents of the XBRL based research is predetermined and fits the existing XML schema with no exceptions.

The research on the open taxonomies showed much higher volatility of the XBRL based financial reports [9]. The mentioned group of researchers found that the companies using opening taxonomies (all U.S. SEC filing companies) attempt substituting known XBRL financial tags

defined in the U.S. GAAP taxonomy with the tags, specific to the company, which produces such substitution. As a result, the financial statements with such substitutions could not be parsed as they lacked the known U.S. GAAP taxonomy tags.

The research on the extension tags produced the evaluation results, which showed that uniformity of the statements does not exist any longer. However, it has never been sufficiently tied to the previous findings on the availability of the financial data, reported in [7]. In this study we intend to close this gap by examining how many statements can be deemed usable in financial research.

### 2.2 Description of Particular Financial Problem

Validation of XBRL as a potential tool of financial research required taking any problem, results of which were significantly well formalized based on the results of the prior research in the financial field. Based on the results of validation of availability of XBRL it was expected that the data extracted from such statements could also yield significant financial results.

The problem of detection of the financial statements manipulations was researched in [10]. The research, performed on manually extracted data, can be considered free of the potential computer errors. It allowed creating a finite set of the accounting variables, which was used in detecting manipulations with financial statements.

Validation of the financial statements was performed using Z-Score of Altman [11], which was described by the following formula:

$$Z = 1.2X1 + 1.4X2 + 3.3X3 + 0.6X4 + 1.0X5 \quad (1)$$

$$X1 = \frac{\text{Working Capital}}{\text{Total Assets}} \quad (2)$$

$$X2 = \frac{\text{Retained Earnings}}{\text{Total Assets}} \quad (3)$$

$$X3 = \frac{\text{EBIT}}{\text{Total Assets}} \quad (4)$$

$$X4 = \frac{\text{Market Value of Equity}}{\text{Book Value of Debt}} \quad (5)$$

$$X5 = \frac{\text{Revenue}}{\text{Total Assets}} \quad (6)$$

The ratios, included in the formula, contain financial variables, which can be easily found in the corporate financial statements. All variables, except for the Working Capital are present in the Balance Sheet or Statement of Income directly. Working Capital represents the difference between Current Assets and Current Liabilities, which are present in the Balance Sheet. We intended to find these variables by finding and parsing their XBRL tags and produce the values for each company under review.

At the time when the research effort, described in [10] was starting, the approach of manipulation detection, used by Beneish [12] was very popular in the practical detection of financial manipulations. The indicator of manipulation M-Score was calculated using the following formula:

$$M = -4.84 + 0.920 * DSRI + 0.528 * GMI + 0.404 * AQI + 0.892 * SGI + 0.115 * DEPI - 0.172 * SGAI + 4.679 * TATAI - 0.327 * LEVI \quad (7)$$

The indicator variables used in (7) are based on the ratios used financial analysts. In order to detect the ongoing manipulations Beneish created a formula, which uses values of the ratios in the adjacent years.

$$DSR = \frac{\text{Accounts Receivable}}{\text{Revenue} * 365} \quad (8)$$

$$GM = \frac{\text{Revenue} - COGS}{COGS} \quad (9)$$

$$AQ = \frac{\text{Assets} - PP\&E}{\text{Assets}} \quad (10)$$

$$SG = \text{Revenue} \quad (11)$$

$$DEP = \frac{\text{Depreciation}}{\text{Depreciation} + PP\&E} \quad (12)$$

$$SGA = \frac{\text{Operating Expenses}}{\text{Revenue}} \quad (13)$$

$$TATA = \frac{\text{Working Capital} - \text{Depreciation}}{\text{Total Assets}} \quad (14)$$

$$LEV = \frac{\text{L.T. Debt} + \text{Current Liabilities}}{\text{Total Assets}} \quad (15)$$

Similarly to the variables of the formula of Z-Score, the variables of the M-Score formula contain the variables present in the balance sheet and the income statement. For the value of PP&E we use the value of the Net Property Plant and Equipment, which is always present on the corporate balance sheet.

The values of the indicators, present in the formula (7) are calculated using the following equation:

$$XI = \frac{X_i}{X_{i-1}} \quad (16)$$

In formula (16) subscripts represent the values of the variable in the current year (i) and in the preceding year (i-1). The appearance of the letter 'I' in the names of indicators in M-Score calculation shows that their values were calculated using formula (16).

Analyzing all accounting variables used in the formulas (1) and (7) we compiled the list of the variables, which need to be obtained by parsing XBRL. The names and the origin statements of the variables are presented in Table 1.

**Table 1. List of Parsed Variables**

Variable	Statement
Total Assets	Balance Sheet
Total Liabilities	Balance Sheet
Current Assets	Balance Sheet
Current Liabilities	Balance Sheet
Accounts Receivable	Balance Sheet
Retained Earnings	Balance Sheet
Market Value Shares	Balance Sheet
Book Value of Debt*	Balance Sheet
Revenue	Income Statement
COGS**	Income Statement
EBIT***	Income Statement
PP&E	Balance Sheet
Depreciation****	Income Statement
Long Term Debt	Balance Sheet

\* In the balance sheet sometimes replaced by total liabilities

\*\* Companies may not report COGS at all in the major statements

\*\*\* EBIT is a term originated in managerial accounting. Operating Income is commonly used although the values of the two can differ significantly

\*\*\*\* Depreciation expense is commonly reported in the income statement. It can also be found in the Cash Flow Statement.

Parsing of the XBRL-based financial statements pursue two separate goals:

- Finding the variables and obtaining the values of these variables is the first and the most important task. By doing this, we attempt to verify how many statements in XBRL format can be used to produce a feed, containing the values of the accounting variables, which could be used in financial research of the specified kind.
- Based on the values of variables of M and Z we attempted to verify whether the accounting data for the variables under review exhibits the previously predicted behavior.

Using the research, presented in [13], we evaluated the values of M and Z for the certain groups of companies. We also used the algorithm of detection of traces of earnings management in the parsed XBRL based statements. The algorithm compared the rate of growth of the values of variables P and R presented by the following formulas.

$$P = 0.367 * Y_1 + 0.980 * Y_2 \quad (17)$$

$$Y_1 = \frac{\text{Shareholders Equity}}{\text{Total Assets}} \quad (18)$$

$$Y_2 = \frac{\text{Revenue}}{\text{Total Assets}} \quad (19)$$

$$R = 0.150 * X_1 + 0.924 * X_2 \quad (20)$$

$$X_1 = \frac{\text{Working Capital}}{\text{Total Assets}} \quad (21)$$

$$X_2 = \frac{\text{Operating Income}}{\text{Total Assets}} \quad (22)$$

The rate of growth of variables P and R is calculated using the following formula:

$$\Delta X = \frac{|X_i - X_{i-1}|}{X_i} \quad (23)$$

Here subscript (i) represents a current year and the subscript (i-1) represents the year preceding current. The mentioned research [13] established the values for the variables M, Z and  $\Delta P-\Delta R$  for various groups of companies. The parsed values of the variables were used in the described formulas (1), (7), (17) and (20) to verify the XBRL data.

The whole research was exploratory in nature. Prior to evaluation of the parsed values it was not possible to expect anything from them. Since the parsing was performed against the financial

statements of the random companies we did not expect that there will be a large number of companies exhibiting deviations in the financial behavior. Table 2 represents the values, which average company is expected to have for the indicators M, Z and  $\Delta P-\Delta R$ .

**Table 2. Expected Values of Indicators**

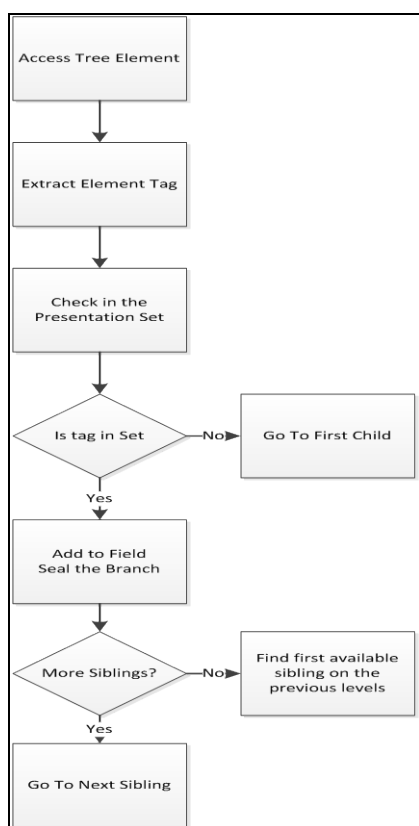
Indicator	Min. Value	Max. Value
Z-Score	1.0	3.0
M-Score	-2.5	-1.5
$\Delta P-\Delta R$	-0.3	0.3

The values of M-Score in Table 2 show that we expected some of the companies to exhibit behavior, which can be classified as fraudulent, i.e. exceeding the cutoff value of -1.78 specified in [12]. It is also possible to have a number of companies with the statements, obtained for parsing, which may have values of Z-Score close to 1.0, which puts them in the pre-bankruptcy state according to [11].

### 2.3 Evaluation of Sample

When the statement is parsed, we obtain values for all variables, specified earlier. The following algorithm is used to obtain the values of the variables under review.

#### Fig. 1. Extraction from XBRL



The parsing uses the same approach, which was identified in [7]. Especially designed program reads the taxonomy and forms the trees, which start from the element, representing a desired financial variable, and traverse them further attempting to obtain the values of the elements placed deeper in the tree. The depth traversal stops when the value is found and the program continues with traversal for the elements of the same level. As a result, the program must receive the value of the desired element as a sum of all values of its descendants.

The tests show that the program does not find an element only if it does not exist in the parsed statements. For all elements which exist in the statement under their own tags or descendent tags, the value is successfully obtained. These findings allowed us to suggest that by taking the percent of the statements in the full set having a particular variable with non-zero value will represent the percent of usable statements with respect to this variable.

U.S. GAAP does not require companies to show the accounting variables of the financial statements, which do not have non-zero values. At the same time, various variables, which are required to be present in the financial statements can be legitimately amiss. This means that the total

percentage of the usable statements can be lower than even the lowest frequency of occurrence of any individual variable.

Out of all variables, which are listed in Table 1, all but one variable are considered mandatory. The company may be legitimately free of any debt at the end of any particular year. This condition will let us accept the statements with zero debt as potentially legitimate.

After the statements are parsed and evaluated, the next step of the evaluation of the XBRL based financial feed is calculating values of  $M$ ,  $Z$ , and  $\Delta P - \Delta R$  and evaluating them against the expected values. We expected that in a rather large sample of XBRL based statements, the calculated values will be in the predicted range. Following the research in [13], we evaluated these values against three other samples, namely:

- Sample of companies, which were convicted with financial statement fraud by U.S. SEC,
- Companies, which had very high Z-Score and, accordingly to [14] they do not have a reason to commit financial statement fraud.
- Companies with the country with emerging economy, where financial markets are functioning in fluctuating fashion [15].

Based on the cited sources we had a clear understanding of what is expected from each of the sample. We also expected that the sample of financial data, parsed from the statements in XBRL format can be evaluated against other samples and will behave in the manner which was described earlier.

### 3 Problem Analysis

#### 3.1 Initial Data Evaluation

##### 3.1.1 Analysis of sample

For the analysis, we took all statements from 10-K corporate reports, submitted to U.S. SEC in XBRL format in the years 2012-2016. Examination of the underlying taxonomies revealed that they do not differ sufficiently and all statements can be processed using the same taxonomy published in 2012. The results of parsing are compiled in Table 3.

**Table 3. Percentages of Usable Data per Variable and Year.**

Variable	Total	2011	2012	2013	2014	2015
Assets	92.74%	92.37%	92.34%	92.56%	93.63%	92.68%
Curr. Ass.	93.91%	93.75%	93.58%	94.02%	94.30%	93.74%
Cash	93.54%	93.19%	93.32%	93.55%	94.10%	93.51%
Acc. Rec.	60.96%	59.35%	61.28%	59.95%	63.28%	60.68%
Cap. Ass.	81.08%	81.26%	80.37%	79.82%	82.92%	80.96%
Deprec.	39.92%	43.23%	38.95%	39.08%	40.06%	38.20%
Liabilities	96.85%	96.65%	96.69%	96.78%	97.09%	96.95%
Curr. liab.	78.49%	76.01%	77.77%	78.60%	80.07%	79.40%
EBIT	96.09%	97.19%	96.07%	96.00%	96.11%	95.07%
Debt	45.62%	41.13%	45.37%	44.46%	50.58%	47.05%
Shares	76.20%	77.45%	75.64%	75/02%	76.52%	76.41%
Ret. Earn.	85.67%	83.56%	84.30%	83.14%	88.77%	89.13%
Revenue	79.71%	77.72%	79.69%	79.46%	81.82%	80.32%
COGS	72.89%	72.03%	73.70%	72.88%	74.66%	71.64%

Simple visual examination of data in the Table 1 shows that variables, which we required for our research could be divided into three groups based on the frequency of their appearance in the XBRL presentation of the financial statements parsed for this research.

The first group are the variables, such as Total Assets, Total Liabilities, Current Assets, Cash and EBIT which appear in over 90% of all statements. It is possible to make a conclusion that these variables are the same for the majority of companies and there is no need to omit or substitute them. The second group is the largest of the three. It is comprised from the accounting variables, which appear frequently in the financial statements of the companies under review. The lower frequency of their discovery can be explained by two reasons:

- Few of the companies do not report these variables in the financial statements. Some of the statements checked visually to confirm the findings were quite rudimentary
- There is another group of the companies, which constantly uses extensions for the variables in question. Since extensions are not tied to the main GAAP taxonomy, it is quite hard to distinguish whether the variable (XML tag) used in the XBRL based report can be substituted with the tag, existing in the taxonomy. This finding confirms the results reported in [9].

The third group of accounting variables includes variables like Total Debt and Depreciation, which appear in less than 50% of all statements under review. The reason for such low frequency of appearance of the variables lies may be the nature of the business (the company does not carry debt and/or depreciating assets). On the other hand, these variables can be even more prone to the overly aggressive use of extensions.

Some researchers consider using Fuzzy Logic and OWL to parse XBRL statements [16]. In this approach they will look for the “soft” attributes, such as tag description to determine the purpose of this tag and its attribution in the major financial statements.

The other approach, which can be equally useful for the increase of ability to extract the variables is the use of the calculation formulas to determine if there are missed variables, which are included in the calculation of more common variables, such as Assets, Liabilities, etc.

Neither of the two specified approaches guarantees that attribution of the variable to certain type is certain. We can potentially increase the frequencies of appearance of certain variables, removing Type II errors. However, statistical approach can produce a sufficient number of false positives, resulting in the wrong attribution of the variables and subsequently increasing a number of potential Type I errors in variable recognition.

Further examination of the values of frequencies and their distributions was reflected in the tables, provided in appendices A1 and A2. We used Fisher distribution to identify whether the frequencies of the appearance of variables in the XBRL based financial statements have similar distributions by the year of appearance.

The comparison of the ratios of standard deviations with the values of Fisher distribution show that variables appear and disappear on the financial statements independently of each other. In terms of our research it means that more frequent appearance of one variable in the statements of given year cannot guarantee more frequent appearance of any other variable in the statements of the same year.

The sample is rather small (5 years for each variable) but it can verify that the variables behave absolutely independently from each other. Based on the examination of such sample we can conclude that the appearance/disappearance of the variable in XBRL based financial reports can be attributed to one of the three causes: (1) the variable is legitimately not reported by the company; (2) the variable is reported with the use of the alternative tag, which is a part of a different tree within the same taxonomy; (3) The variable is reported using the extension tag, which is not tied to the primary GAAP taxonomy by any means.

Another interesting statistic to be mentioned. The total number of usable statements was 6472, which represents 28.9% of all statements in XBRL format, processed for years 2011-2015. Despite the fact, that some of the statements can be legitimately discarded as not reporting the full complement of the variables, which need to be reported, the number is still very low.

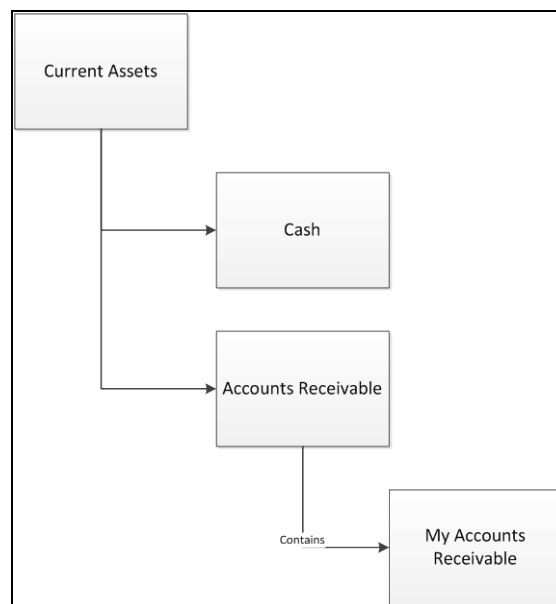
**3.1.2 Implication for Parsing**

One of the approaches, commonly used in the U.S. XBRL community is using fuzzy logic and OWL in parsing XBRL based statements. Positive results in parsing the documents using such approach are described by [17]. This approach is based on the assumption that XBRL is used for visual reporting of the statements in the documents submitted to the regulator or for the public display by the organization. In this case, the “unknown” tag, included by the creator of the XBRL statement will be offset by the meaningful description of the data, submitted for display.

We haven’t spent enough time with this approach, but it can be supposed that any decision-making algorithm is prone to the statistical errors and can only supply a certain probability of the correctness of data.

Instead, it is possible to apply an absolutely deterministic algorithm based on object containment pattern [18].

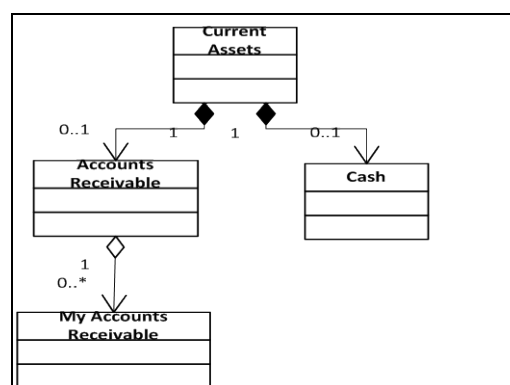
**Fig. 2 Containment Pattern**



Considering, that for accounts receivable the company uses a custom tag, we can propose that the custom tag will be always included inside the regular accounts receivable tag instead of substituting it. If the company uses such approach, it will be able to report its own custom account receivable information without introducing a confusion during parsing.

The implications for the developers of the schema (taxonomy) are summarized in Fig. 3

**Fig. 3. Object Relationships between Tags**



Object relationships between elements of the schema are very close to the classic containment. The tags, which are part of the GAAP taxonomy, are always present in the same form, which is

denoted by containment relationship. The tag, which may be introduced by a particular company has an aggregation relationship with the GAAP tag. This will ensure that GAAP tag (Accounts Receivable) is *always* present when the tag, representing My Accounts Receivable is present. Such approach will guarantee that custom tags introduced by the companies are not omitted or misread.

### 3.2 Evaluation of Parsed Data

M-Score by Beneish requires the data from two adjacent years, the calculation of Z-Score usually requires data recorded for the current year. In order to be consistent with the numbers of entries for both indicators, we omitted Z-Score calculations for 2011. The calculations for Z-Score are presented in Table 4.

**Table 4. Z-Score Statistics for Samples**

Var	N	Mean	St. Dev	Q1	Median	Q3
Fraud	124	0.323	4.745	0.329	0.97	1.896
Clean	36	3.452	1.166	2.79	3.346	3.882
All	5083	1.1941	2.4899	0.4451	1.4717	2.4882
2012	1201	1.307	2.5103	0.5563	1.5591	2.5623
2013	1382	1.1891	2.5704	0.4813	1.4953	2.4781
2014	1196	1.3221	2.2215	0.5483	1.5049	2.4686
2015	1226	0.8817	2.595	0.1234	1.2364	2.3186

All samples, presented in Table 4, were subject to Kolmogorov-Smirnov normality test. All samples appear normal with at least 99% significance. T-Value comparison of the means showed that means of all XBRL samples are different from the mean of the clean sample. The values of XBRL data do not exhibit any out-of-usual behavior, showing that there is a fair number of companies with strong performance, which have relatively high M-Score.

During the extraction, the results revealed another problem, which was not anticipated at the beginning. There was a mismatch in the presentation of precision for certain variables. Although this fact does not affect the display of the financial statements, it is clearly affecting the calculations as they appear to be shifted based on increased, decreased and/or omitted precision attribute.

Calculations for M-Score are presented in Table 5.

**Table 5. M-Score Calculation for All Samples**

Var	N	Mean	St. Dev	Q1	Median	Q3
Fraud	123	-1.682	1.358	-2.452	-1.781	-0.929
Clean	36	-1.239	1.038	-1.87	-1.38	-0.579
All	1480	-1.486	1.332	-2.385	-1.625	-0.726
2012	322	-1.520	1.333	-2.433	-1.672	-0.798
2013	422	-1.442	1.479	-2.322	-1.501	-0.707
2014	377	-1.438	1.205	-2.353	-1.616	-0.673
2015	331	-1.584	1.293	-2.492	-1.702	-0.764

All samples, presented in Table 5, were subject to Kolmogorov-Smirnov normality test. All samples appear normal with at least 99% significance. T-Value comparison of the means showed that means of all XBRL samples are different from the mean of the clean sample. The values of XBRL data do not exhibit any out-of-usual behavior, showing that there is a fair number of companies with strong performance, which have relatively high M-Score.

During the extraction, the results revealed another problem, which was not anticipated at the beginning. There was a mismatch in the presentation of precision for certain variables. Although this fact does not affect the display of the financial statements, it is clearly affecting the calculations as they appear to be shifted based on increased, decreased and/or omitted precision attribute.

Table 6 represents the percentages of usable entries for each indicator, considering that all variables included must have non-null values. The “No Outlays” columns represents the values, which do not have any out-of-norm values, such as  $|Z| > 10$ ,  $|\Delta P - \Delta R| > 5$  or zero values for any indicator in M-Score formula. The percentages of “No-Outlays” are based on the total number of usable values of 5441.

**Table 6. Percentages of Usable Indicators**

Var	Total	2012	2013	2014	2015	No Outlays
M	12.39%	12.1%	12.8%	15.6%	10.4%	27.20%
Z	49.61%	49.4%	48.5%	50.6%	49.8%	93.42%

The yield of usable statements appears to be rather low at 5441 out of 16439 in the years of 2012-2015 or 1360 per year on average. However, the similar research on earnings management performed by [19] used close to 15000 entries in 10 years or 1500 per year on average. Hence, XBRL feed yield fits



the needs of a research of the mentioned magnitude.

The data obtained from XBRL represents the random sample of data. Calculated values for Z-Score, presented in Table 3 are clearly in so-called 'grey zone' ( $1 \leq Z \leq 3$ ), where the majority of the U.S. Companies operate. The values of M-Score appear to be lower than the values for the companies with exceptional performance and higher than the ones for the revenue manipulators. This is also consistent with the expectations for a moderately well-performing random company. Therefore, it is possible to say, that if there is a sufficient volume of data, extracted from XBRL via the previously described algorithm, it can be used in financial research as a random data (similar to the one obtained from Yahoo or COMPUSTAT).

Evaluation of the same data against the sample from the country with the emerging economy is presented in Table 7.

**Table 7. M-Score for All Announced Samples**

Var	N	Mean	St. Dev	Q1	Median	Q3
Clean	98	-1.83	1.137	-2.46	-1.949	-1.13
Fraud	32	-1.26	0.940	-1.87	-1.380	-0.58
Random	1082	-1.51	1.255	-2.37	-1.621	-0.80
Emerg.	35	-1.44	2.178	-2.18	-1.009	-0.08

Evaluation of the random (XBRL) sample against fraud and clean sample shows that the mean and the median appear to be in the middle position. The values obtained for the random (XBRL) sample show that they fit in the predicted threshold, listed in Table 2. The analysis of mean and median of the sample, based on the values extracted from XBRL show that they appear to be in the middle between the values of the fraud and the clean sample. It can be explained by the fact that some of the data in the random sample may be from the companies, which recorded fraudulent data and some of the entries are close to the ones of the 'clean' sample.

We applied similar analysis to the values of  $\Delta P$ - $\Delta R$ , which were calculated for all samples and compiled in Table 8.

**Table 8. Values of  $\Delta P$ - $\Delta R$  for All Samples**

Var	N	Mean	St. Dev	Q1	Median	Q3
Clean	98	0.078	0.844	-0.20	-0.045	0.280
Fraud	32	-0.13	1.230	-0.20	-0.027	0.172
Random	1082	-0.00	0.469	-0.06	-0.001	0.054
Emerg.	35	-0.07	0.906	-0.26	-0.067	0.246

$\Delta P$ - $\Delta R$  indicator can have positive and negative values, which equally predict the existence of earnings management in the statement of the company under review. Therefore it is more important to examine standard deviation and the values of Q1 and Q3 to estimate how many computed values of the indicator are over the predicted threshold.

It can be observed that the values of the indicator for the random sample are closer to 0, which indicates the absence of traces of earnings management (or lower percentage of occurrence) than in the fraud and the emerging samples. Overall, the sample of data from XBRL based financial statements behaves as a sample of companies picked at random. Such behavior shows that this sample can be considered as a random sample of the financial statements and can be used in the financial and accounting research.

## 4 Conclusions and Directions of Research

In this research we examined the sample of financial data obtained from the XBRL based financial statements. It can be concluded that it is possible to obtain this data by employing simple parsing of XML, contained in XBRL document. We applied this data in a problem, which required a certain number of financial variables. The problem has not been tailored to the availability of the data in XBRL based statements.

The random selection of the problem shows that XBRL based data can be applied to any research in the accounting and financial field. The yield of the data, obtained by using direct parsing of XML, will be the same or similar to the one obtained in this research.

The variables required in this research were divided into three categories stable with occurrence of over 90%, less stable with occurrence of 60-90% and random with the appearance of under 60%. Of course, the use of the stable variables will give a better yield of financial data within the same sample of XBRL based statements. The use of the less stable variables will have a smaller yield of data and the use of the variables randomly

appearing can further reduce the number of useful statements.

When constructing the feed out of the financial data parsed from XBRL based statements there are two solutions: keep all statements and give zero values to the omitted data or keep only statements, which have required information. If the feed is constructed from all statements, the task of filtering will shift to the user of the feed but the number of useful (for a particular) statements will remain the same.

The paper shows the ways of improvement of the yield of usable data from the field. Use of the containment pattern, described in the research will definitely increase the usability and the understanding of data, but it has to be mandated by the regulators, such as U.S. SEC, who have ownership of the taxonomy and the rules of XBRL based submissions.

There are several ways to somewhat improve the yield of data coming from the XBRL based statements. The U.S. GAAP taxonomy has several statement trees, created for the companies in the various lines of business (real estate, commercial, financial, etc.). It is possible to read the line of business of the company in XBRL based report and extract the data by using a tree in the taxonomy, corresponding to it. This approach does not guarantee good results but it may be useful if the XML tags are selected for use based on the line of business. We will consider such improvements in the future research

Overall, it can be stated that XBRL based financial feed is a valid alternative to COMPUSTAT, Yahoo and Google feed and can be used in the financial research if needed. At the moment, there are no programs, which can generate such feed. The future research may be able to formalize the algorithms and the patterns used in this research and present them for the use of the community at large.

### References

[1] R. Debreceeny and G. Gray, "The production and use of semantically rich accounting reports on the Internet: XML and XBRL," *International Journal of Accounting Information Systems*, vol. 2, no. 1, pp. 47 -74,

2001.

[2] R. Pinsker, "XBRL awareness in auditing: a sleeping giant?," *Managerial Auditing*, vol. 18, no. 9, pp. 732-726, 2003.

[3] J. San Miguel, "The Reliability of R&D Data in COMPUSTAT and 10-K Reports," *The Accounting Review*, vol. 52, no. 3, pp. 638 - 641, 1977.

[4] E. Boritz and W. No, "The Quality of Interactive Data: XBRL Versus Compustat, Yahoo Finance, and Google Finance," 2013. [Online]. Available: <https://ssrn.com/abstract=2253638>. [Accessed 2017].

[5] R. Chychyla and A. Kogan, "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*, vol. 29, no. 1, pp. 37-72, 2015.

[6] K. Kernan, "XBRL Around the World," *Journal of Accountancy*, vol. 206, no. 4, pp. 62-66, 2008.

[7] R. Debreceeny, A. D'Eri, C. Felden, S. Farewell and M. Piechocki, "Feeding the Information Value Chain: Deriving Analytical Ratios from XBRL filings to the SEC," in *22nd XBRL International Conference*, Kansas City, KS, 2010.

[8] C. Hoffman, *Financial Reporting Using XBRL: IFRS and US GAAP Edition*, New York, NY: Lulu Publishing, 2006.

[9] R. S. Debreceeny, S. M. Farewell, M. Piechocki, C. Felden, A. Graening and A. D'Eri, "Flex or Break? Extensions in XBRL Disclosures to the SEC," *Accounting Horizons*, vol. 25, no. 4, pp. 631-657, 2011.

[10] I. Pustylnick, "An Algorithm of Detection of Revenue and Retained Earnings Manipulation," *Accounting and Taxation*, vol. 4, no. 2, pp. 95-105, 2012.

[11] E. Altman, "Financial Ratios, Discriminant

Analysis and the Prediction of Corporate Bankruptcy," *Journal of Finance*, 1968.

- [12] M. Beneish, "Detecting GAAP violation: implications for assessing earnings management among firms with extreme financial performance," *Journal of Accounting and Public Policy*, vol. 16, no. 3, pp. 271-309, 1997.
- [13] I. Pustyl'nik, "COMPARISON OF LIQUIDITY BASED AND FINANCIAL PERFORMANCE BASED INDICATORS IN FINANCIAL ANALYSIS," *Oeconomia Copernicana*, no. 1, pp. 83-97, 2017.
- [14] M. J. Lenard and P. Alam, "An Historical Perspective on Fraud Detection: From Bankruptcy Models to Most Effective Indicators of Fraud in Recent Incidents," *Journal of Forensic & Investigative Accounting*, vol. 1, no. 1, 2009.
- [15] H. Shin, "The Second Phase of Global Liquidity and Its Impact on Emerging Economies," in *Volatile Capital Flows in Korea*, New York, NY, Palgrave MacMillan, 2014, pp. 247-257.
- [16] S. Mendez-Nuniez and G. Trivino, "Combining Semantic Web technologies and Computational Theory of Perceptions for text generation in financial analysis," in *International Conference on Fuzzy Systems (IEEE)*, 2010.
- [17] C. Hoffman and M. Rodriguez, "Digitizing Financial Reports \_ Issues and Insights: A Viewpoint," *International Journal of Digital Accounting Research*, vol. 13, pp. 73-98, 2013.
- [18] C. Blilie, "Patterns in Scientific Software: an introduction," *Computing in Science & Engine*, vol. 4, pp. 48-53, 2002.
- [19] I. Jansen, S. Ramnath and T. Yohn, "A Diagnostic for Earnings Management Using Changes in Asset Turnover and Profit Margin," *Contemporary Accounting Research*, vol. 29, pp. 221-251, 2012.
- [20] E. Pustyl'nik, *Statistical Methods of analysis and processing of observations*, Moscow: Nauka Publishing, 1968.

## A1 Values of Fisher Distribution F for Accounting Variables

The table in this appendix contains  $\sigma^2/\sigma^2$  values for all accounting variables compiled in Table 1.

	Assets	Curr. Ass.	Cash	Acc. Rec.	Cap. Ass.	Deprec.	Liabilities	Curr. liab.	EBIT	Debt	Shares	Ret. Earn.	Revenue
<b>Curr. Ass.</b>	3.48												
<b>Cash</b>	2.31	1.50											
<b>Acc. Rec.</b>	8.16	28.41	18.88										
<b>Cap. Ass.</b>	4.92	17.13	11.38	1.66									
<b>Deprec.</b>	13.88	48.31	32.10	1.70	2.82								
<b>Liabilities</b>	8.22	2.36	3.56	67.13	40.47	114.15							
<b>Curr. liab.</b>	8.85	30.80	20.47	1.08	1.80	1.57	72.78						
<b>EBIT</b>	2.01	7.01	4.66	4.06	2.44	6.90	16.55	4.40					
<b>Debt</b>	42.92	149.36	99.25	5.26	8.72	3.09	352.92	4.85	21.32				
<b>Shares</b>	3.04	10.59	7.04	2.68	1.62	4.56	25.02	2.91	1.51	14.11			
<b>Ret. Earn.</b>	30.52	106.21	70.57	3.74	6.20	2.20	250.96	3.45	15.16	1.41	10.03		
<b>Revenue</b>	7.85	27.30	18.14	1.04	1.59	1.77	64.51	1.13	3.90	5.47	2.58	3.89	
<b>COGS</b>	5.39	18.76	12.46	1.51	1.10	2.58	44.32	1.64	2.68	7.96	1.77	5.66	1.46

All samples have the same size of 5. The cut-off value, designating same/not same distribution for both samples is F=5.1 [20]

## A2 Correlation Statistics for All Accounting Variables

The table in this appendix contains values of correlation between frequencies of appearance of all variables in Table 1.

	Assets	Curr. Ass.	Cash	Acc. Rec.	Cap. Ass.	Deprec.	Liabilities	Curr. liab.	EBIT	Debt	Shares	Ret. Earn.	Revenue
<b>Curr. Ass.</b>	0.873**												
<b>Cash</b>	0.969***	0.890**											
<b>Acc. Rec.</b>	0.850*	0.573	0.848*										
<b>Cap. Ass.</b>	0.838*	0.601	0.681	0.711									
<b>Deprec.</b>	-0.110	0.019	-0.300	-0.334	0.302								
<b>Liabilities</b>	0.910**	0.719	0.915**	0.774	0.684	-0.403							
<b>Curr. liab.</b>	0.756	0.608	0.857*	0.737	0.367	-0.720	0.913**						
<b>EBIT</b>	-0.191	0.002	-0.324	-0.298	0.137	0.926**	-0.554	-0.764					
<b>Debt</b>	0.858*	0.604	0.900**	0.926**	0.600	-0.562	0.921**	0.932**	-0.594				
<b>Shares</b>	0.126	-0.052	-0.119	-0.056	0.616	0.734	0.030	-0.365	0.463	-0.173			
<b>Ret. Earn.</b>	0.701	0.346	0.650	0.664	0.660	-0.391	0.887**	0.762	-0.629	0.813*	0.262		
<b>Revenue</b>	0.840*	0.612	0.902**	0.913**	0.543	-0.607	0.907**	0.948**	-0.615	0.997***	-0.252	0.775	
<b>COGS</b>	0.636	0.563	0.674	0.824*	0.466	-0.126	0.387	0.421	0.095	0.622	-0.268	0.128	0.634

\*\*\* corresponds to 99% significance, \*\* to 95% significance, \* to 90% significance