ESSAYS ON ACCOUNTING DATA DIFFERENCES AND AUDIT LEARNING

BY ROMAN CHYCHYLA

A dissertation submitted to the
Graduate School—Newark
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
Graduate Program in Management

Written under the direction of
Dr. Alexander Kogan
and approved by

Dr. Alexander Kogan

Dr. Glenn Shafer

Dr. Miklos Vasarhelyi

Dr. Vladimir Vovk

Newark, New Jersey

May, 2014

ABSTRACT OF THE DISSERTATION

Essays on Accounting Data Differences and Audit Learning

by Roman Chychyla

Dissertation Director: Dr. Alexander Kogan

The dissertation comprises of three essays that 1) compare accounting numbers in Capital IQ's Compustat North America Fundamentals Annual, the most popular accounting database in accounting research, to the original numbers in corporate reports, 2) study the effects of Compustat's data standardization procedures on accounting-based bankruptcy prediction models, and 3) develop a framework to enhance the performance of analytical learning models in a multi-period auditing setting.

In the first essay, we conduct the first large-scale comparison of Compustat and 10-K data. Specifically, we compare 30 accounting line items of approximately 5,000 companies for the period from October 1, 2011, to September 30, 2012. We find that the values reported in Compustat significantly differ from the values reported in 10-K filings. We also find that the amount and magnitude of the original data alterations introduced by Compustat depend on the type of the accounting item and company characteristics such as industry and size.

Numbers that appear in Compustat are standardized – adjusted to fit fixed variable definitions – to ensure "...consistent and comparable data across companies,

industries and business cycles..." However, there has been no evidence in the academic literature that Compustat's standardized numbers provide more benefits than the original numbers in financial statements. In the second essay, we examine the effects of Compustat's data standardization using Altman's 1968 and Ohlson's 1980 bankruptcy prediction models as examples. We find that Compustat's data standardization not only yields no improvements for bankruptcy prediction models, but also has a significant negative impact on the predictive accuracy of Altman's model (up to 8.56%)

There are several challenges in applying analytical models to the auditing problem of identifying irregular transactions. We argue that because of these challenges standard statistical models may not be well-suited for auditing and have to be modified to achieve better performance. In the third essay, we propose a framework to boost the performance of analytical learning models in auditing. The results of framework's testing on the real data show a significant increase of performance of the tested models.

Acknowledgements

This dissertation would not have been written without the help, guidance, and support of wonderful people that I got lucky to get to know in my life.

First and foremost, I would like to thank Professor Alexander Kogan, who has become much more than a mentor to me over the past few years I spent at Rutgers. He always finds time for others, and I am constantly amazed by his brilliance, encouragement, and dedication.

Also I would like to express my gratitude to Professor Glenn Shafer for his supervision, caring, and patience, and Professor Vladimir Vovk, who has been always there to help and provide suggestions. These two people brought me to academia, and for that I am forever in their debt.

I am most grateful to Professor Miklos Vasarhelyi for his continuous help, advice, strong belief in his students, and all the jokes that never fail to make me smile.

I would like to thank Professor Dan Palmon for his excellent work as a Department Chair and enormous assistance I have received from him. Also, to Professor Peter Gillett for an irreplaceable experience of being his TA.

I thank Professor Michael Alles, Professor Carolyn Levine, and Professor Valentin Dimitrov for, among other things, the invaluable comments they gave me. I also appreciate all the help I have received from the Rutgers Community.

Finally, I would like to thank my wife Khrystyna. Words cannot describe the amount of love and support she bestows upon me. I thank my parents, who made it all possible.

Table of Contents

List of Tables

List of Figures

Introduction

The single most important part of any empirical study is data. Empirical research studies data to understand various phenomena and contribute to our knowledge of the world. Data is also at the very heart of accounting. Accounting systems capture, store, and process business data, and report the resulting information to the interested parties. The quality of information outputted by accounting systems and the accounting research depends on the inputted data and the way it is processed. Therefore, its is important to study both the properties of data and the procedures that transform the data into useful information. The first part of this dissertation consists of two related essays that study the type of data provided by Compustat North America Fundamentals, the most popular database in empirical accounting research. The second part develops a framework that utilizes distributional properties of transactional data to enhance processing performance of analytical models in a multi-period auditing setting.

The first essay is presented in Chapter 1. It studies the amount and magnitude of discrepancies in Compustat North America Fundamentals, an accounting database that is frequently used for both research and decision-making. It has been documented that information found in Compustat database differs from both the information found in other accounting databases and the information disclosed in corporate financial filings (San Miguel 1977; Rosenberg and Houglet 1974; Yang, Vasarhelyi, and Liu 2003; Tallapally, Luehlfing, and Motha 2011; 2012; Boritz and No 2013). However, previous studies that compare numbers in Compustat to numbers in the original corporate reports share a major limitation – they analyze samples of small sizes that may not fully reflect the "true" amount and magnitude of

data alterations introduced by Compustat. The challenge of utilizing larger samples is that it results in a very high cost, if the number comparison is carried out manually. We overcome this limitation by automating the number comparison procedures with the help of the recently introduced eXtensible Business Reporting Language (XBRL) reporting technology.

We conduct the first large-scale comparison of Compustat and 10-K data. Specifically, we compare 30 accounting line items of approximately 5,000 U.S. companies for the period from October 1, 2011, to September 30, 2012. We find that the values reported in Compustat significantly differ from the values reported in 10-K filings. We also find that the amount and magnitude of the original data alterations introduced by Compustat depend on the type of the accounting item and company characteristics such as industry and size.

Chapter 2 of this dissertation is the natural extension of Chapter 1. It presents an essay that examines the effects of data standardization procedures implemented in Compustat database. Compustat's data standardization is the process of adjusting original numbers reported in companies' reports to match Compustat's fixed variable definitions, and is the main driver of differences between Compustat numbers and original numbers reported by companies. Compustat argues that "*[s]tandardized data ensures that you have consistent and comparable data across companies, industries and business cycles, and offers a solid foundation for your rigorous analysis.*" However, there has been no evidence to support this statement in academic literature. The original numbers in financial reports often are not constrained by fixed definitions under current Generally Accepted Accounting Principles (GAAP) standards. Moreover, the developer of GAAP, Financial Accounting Standards Board (FASB), in Statement of Financial Accounting Concepts No. 2, discourages the use of fixed definitions of accounting items saying that "*[t]hat kind of uniformity may even adversely affect comparability of information if it conceals real differences between enterprises.*"

The study introduced in Chapter 2 is the first to examine the effects of Compustat's data standardization using bankruptcy prediction models as examples. Specifically, we study whether using Compustat's standardized data as opposed to original 10-K data improves Altman's 1968 and Ohlson's 1980 bankruptcy prediction models. We find that Compustat's data standardization not only yields no improvements for bankruptcy prediction models, but also has a significant negative impact on the predictive accuracy of Altman's model (up to 8.56%).

Chapter 3 is an essay that discusses challenges in applying analytical learning models in a multi-period audit to identify irregular transactions, and develops a framework for analytical models to overcome these challenges. The advantage of using analytical models (as opposed to manual procedures) in auditing is that they are able to process large populations of transactional data (as opposed to samples) with a relatively low cost. In addition, analytical models are more effective in identifying data patterns than humans are. The problem associated with applying analytical models, that has not received much attention in the literature, is that most of them are not designed to operate in an auditing setting where the number of irregular transactions is low relative to the number all transactions (the problem of unbalanced data), and the incremental statistical learning in each audit period is limited to a small portion of transactions chosen to be investigated (the problem of one-sided feedback).

In Chapter 3, we introduce a framework for analytical learning models that changes the way the models learn and predict. The framework exchanges the immediate gain from investigating the most suspicious and important transactions in return for more accurate statistical model by spending audit resources to learn more about the underlying distribution of the transactional data. A more accurate statistical model may yield more benefits in the future. The proposed framework is tested on real-world data. The results show a significant boost in performance of analytical models under the proposed framework.

Chapter 1

Using XBRL to conduct a large-scale study of discrepancies between the accounting numbers in Compustat and SEC 10-K filings

1.1 Introduction

Compustat is a popular source of financial information for both academics and practitioners. It has been maintained by Standard and Poor's[1] company since 1962. Many accounting empirical studies are based on Compustat data. However, it has been questioned how reliable the data found in Compustat is (and by extension in other accounting databases). Prior studies have shown that Compustat data may differ from the real-world financial data (San Miguel 1977; Kinney and Swanson 1993; Tallapally, Luehlfing, and Motha 2011; 2012; Boritz and No 2013) and data found in other accounting databases (Rosenberg and Houglet 1974; Yang, Vasarhelyi, and Liu 2003).

Compustat relies on companies' original reporting documents (such as 10-Qs and 10-Ks) to populate its fundamentals data set. However, the original companies' data is standardized per Compustat's variable definitions to ensure "consistent and comparable data across companies, industries and time periods without reporting biases or data discrepancies".[2] This fact generates two important questions: 1) should Compustat standardize the original data reported by companies, and 2) what is the extent of this standardization. In this Chapter we address the second question, and leave the

1. Standard & Poor's is a division of The McGraw-Hill Companies.

2. www.compustat.com, as of February 16, 2014.

first question for the future research. In addition to data standardization, Compustat numbers may differ from the original numbers reported by companies due to typos, missing values, not-up-to date values, etc. It is important to estimate the amount and magnitude of differences between Compustat data and the original data, since these differences may affect the results of accounting studies that utilize Compustat.

The best way to assess the extent of data alteration in Compustat is to compare Compustat numbers to the numbers in the original reports. Unfortunately, this is a very costly procedure, if done manually. Prior studies that contrasted Compustat and original reports' numbers had an important limitation of using small samples in their analyses. Small samples may not fully represent the population, and omit uncommon types of observations (that are likely to result in discrepancies between Compustat and original reports). In this study, we utilize XBRL (eXtensible Business Reporting Language) reporting technology to automatically extract accounting numbers from XBRL 10-K financial reports and compare them to Compustat numbers. The SEC has mandated the use of XBRL reporting by all U.S. GAAP filers starting from June 2011. To our best knowledge, this is the first study to conduct such a large-scale data comparison.

Specifically, we study the amount and magnitude of discrepancies between Compustat North America Fundamentals Annual and 10-K reports of U.S. companies for the period from October 1, 2011 to September 30, 2012. Our analysis includes more than 5,000 companies, and 30 accounting items that are often used in empirical accounting research. We limit our analysis to 10-K reports only since they are audited, and to one 10-K report per company to make our sample unbiased and study whether company characteristics are related to the amount and magnitude of discrepancies.

Although we use the XBRL technology to extract data from XBRL 10-K reports, our objective is not to compare Compustat and XBRL 10-Ks, but rather to compare Compustat and traditional 10-Ks. XBRL 10-K reports sometimes contain errors that make them different from plain-text 10-Ks. We develop automated and manual

procedures to eliminate XBRL-related errors.

We find that Compustat significantly alters numbers reported in the 10-K filings. In particular, we find that Compustat values of 17 out of 30 analyzed variables significantly differ from values reported in the 10-K filings. We also find that the type of statement where variable is reported and company characteristics such as industry and size are related to the amount and magnitude of discrepancies. Specifically, the amount of discrepancies is significantly affected by 1) type of financial statement (except for the telecommunications industry), 2) industry regardless of financial statement type, and 3) revenue size for the financial industry or balance sheet items.

Our contributions to the literature are as follows. Firstly, we contribute to the studies of Rosenberg and Houglet (1974); Bennin (1980); Kinney and Swanson (1993); Yang, Vasarhelyi, and Liu (2003); Boritz and No (2013), and others that analyze differences between Compustat and other accounting data sources, by conducting the first large-scale comparison of Compustat North America Fundamentals Annual and 10-K data. This comparison provides a more accurate and comprehensive statistics of discrepancies between Compustat and 10-K data. Secondly, we contribute to the existing XBRL literature by demonstrating how XBRL data can be utilized in an automated fashion to extract and process commonly used accounting numbers. We develop a methodology for an automated large-scale comparison of Compustat and XBRL data. The methodology comprises of several steps: data extraction, data merging, concept mapping, difference calculation, automated error detection, and discrepancy analysis.

The Chapter is organized as follows. Section 1.2 reviews academic literature relevant to this study. Section 1.3 discusses types of data alterations in Compustat. Section 1.4 presents Compustat data comparison methodology. Section 1.5 reports the main findings of this study. Finally, Section 1.6 concludes the Chapter.

1.2 Relevant literature

In this section, we review relevant literature that studies the impact of discrepancies in accounting databases. Since there are two main causes of discrepancies – errors and different data definitions – we consider the respective literature separately.

1.2.1 Erroneous data

Rosenberg and Houglet (1974) is the first paper to consider the quality of data found in Compustat. The authors match and merge Compustat and CRSP data on monthly price relatives for 844 industrial sector companies from January 1963 to June 1968, and for 97 utilities sector companies from March 1962 to June 1968. In total, they compare 41,296 monthly price relatives between Compustat and CRSP data sets. Out of these, they find 1,202 (2.91%) to be erroneous including 294 (0.71%) that differ by more than 5%, and 125 (0.3%) that differ by more than 20%. The authors compare 34 discrepancies between utilities price relatives to the original published sources; 4 (12%) discrepancies were due to CRSP errors, and 30 (88%) were due to Compustat errors suggesting that CRSP is a more reliable database for monthly price relatives. The study finds these errors to change the data distribution, with higher moments being affected more significantly.

Bennin (1980) did a follow-up study of Rosenberg and Houglet (1974) using updated Compustat and CRSP data, and for a longer time period. He compares monthly relative prices of both industrials and utilities for the period from January 1962 through July 1978. Out of 187,460 prices, he finds 471 (0.25%) to differ by more than 5%. This is a huge drop of error rate compared to Rosenberg and Houglet (1974) – 0.25% versus 0.71%. The study suggests that Compustat had corrected a number of errors after Rosenberg and Houglet (1974) study.

Beedles and Simkowitz (1978) replicate the study of McEnally (1974) that investigates the return behavior of high-risk common stocks extracted from the CRSP

database. After correcting the CRSP database errors, Beedles and Simkowitz find that the results changed significantly due to changes in higher moments of data distribution.

San Miguel (1977) is the first study to compare Compustat data to 10-K reports. Specifically, he compares Research and Development (R&D) expense data in Compustat with the original 10-K reports for a sample of 256 companies that reported R&D expense in 1972.[3] He finds 78 (30%) discrepancies between the two datasets. At least 52 (66.67%) discrepancies were due to errors in Compustat. Out of those 15% were rather significant - the discrepancies amounted to more than 100% of the Net Income values of the respective companies. The author also finds that there was insufficient information in the 10-K reports to classify the remaining 26 discrepancies. He attributes part of the errors to the new rules for the 10-K Form R&D disclosure requirements that became effective in 1972, and to the complexity of some 10-K reports that Compustat personnel had to deal with.

Kinney and Swanson (1993) examine the accuracy of tax data in Compustat. The authors randomly select 100 Compustat companies from fiscal 1985 listings. They compare 19 tax variables as reported by Compustat and the original financial reports for the years of 1986-1988. The error rates for the considered variables ranged from 0.76% to 11.65% with the error magnitudes being substantial. The authors find the tax error rates to be: 1) high for utility companies, 2) low for balance sheet items, and 3) high for cash flow statement items. They also find a significant number of missing values for some variables whereas these values are reported in the financial statements.

Kern and Morris (1994) study the data differences between Compustat and Value Line databases. The authors compare Sales and Total Assets data for the years of 1971-1990. They find that there is no statistically significant difference with respect

3. In 1972, 1,357 Compustat firms reported R&D expense. A sample used in San Miguel (1977) represents 19% of those.

to Total Assets. However, they find significant differences between Sales values. The number of discrepancies between Sales ranged from 31.4% to 33.9% across different years. The years of 1985-1990 had the highest number of material discrepancies (137 in Total Assets and 378 in Sales). The authors examined the annual reports for those years to identify the sources of the discrepancies. Many discrepancies (40.9% for Total Assets and 15.1% for Sales) were due to Value Line reporting data in currency other than U.S. dollars. The largest source of discrepancies for Sales was the difference in data definitions used in the databases. For example, Value Line include only the income items in Sales that are related to the company's major line of business, while Compustat reports Sales for the entire consolidated entity; this resulted in 47.6% of all discrepancies. But the authors also find some unexplainable differences (at least 1.5% of all discrepancies).

Yang, Vasarhelyi, and Liu (2003) compare Compustat and Value Line values of seven frequently used accounting variables for the years of 1976-1981.[4] Out of 10,353 observations (of 1,479 companies), 1,284 (12.5%) were discrepancies larger than 1%. To identify the cause of discrepancies, the authors draw a subsample of 200 companies and compare the Compustat and Value Line 1981 data to the original 1981 financial statements data. Out of 1,400 observations, they find 320 mismatches. Out of those mismatches, 185 (57.81%) were explainable discrepancies (i.e., discrepancies due to different data definitions, currency and industry factors), and 135 were unexplainable that were either due to errors or undisclosed coding rules. Compustat amounted to more unexplained discrepancies – 99 (73% of all unexplained differences) – than Value Line. The authors also replicate a part of the Rosenberg and Houglet (1974) study by computing and comparing data distribution moments of Compustat and Value Line databases. At least for some variables (e.g., Current Assets) the differences were

4. Yang, Vasarhelyi, and Liu (2003) examined The Accounting Review, The Journal of Accounting Research and The Journal of Accounting and Economics for the 1976-1981 period to find the most frequently used variables in accounting research. They identified the following variables: Total Assets, Net Sales, Inventories, Net Income, Current Liabilities, Depreciation, Depletion, and Amortization, and Gross Plant.

found to be statistically significant. The authors conclude that accounting databases of data aggregators contain a certain level of erroneous data and do not always agree on data definitions and industry classifications.

Tallapally, Luehlfing, and Motha (2011) compare EDGAR Online[5] and Compustat values of the Cost of Goods Sold (COGS) item for a subset of DOW 30 companies for the fiscal year of 2009. Out of 26 companies considered, there was only one match in COGS between Compustat and EDGAR Online. The average magnitude of remaining 25 discrepancies is 14.23% with Compustat reporting, overall, lower numbers. The authors were not able to reconcile the numbers, but assume that these differences are due to data definitions used in Compustat.

Tallapally, Luehlfing, and Motha (2012) is a very similar study to the previous one with the same authors (Tallapally, Luehlfing, and Motha 2011). It compares Compustat and 10-K XBRL (eXtensible Business Reporting Language) data for the same year (2009) and for almost the same set of companies[6] but with regards to Sales/Revenue item instead of COGS. The data extracted from 10-K XBRL should be similar (most likely the same) to the data extracted from EDGAR Online.[7] The authors find differences in Sales/Revenues between Compustat and 10-K XBRL filings, but not as many and not as large as in the study that compared COGS values. Namely, the authors find 6 discrepancies (22% of all observations) with the average magnitude equal to 5.19%. As in the previous study, the authors do not reconcile the discrepancies.

Boritz and No (2013) manually compare financial items reported in 150 XBRL 10-K filings of 75 companies to the corresponding items provided by three data aggregators: Compustat, Yahoo Finance, and Google Finance. The study finds that

5. Not to be confused with SEC's EDGAR. EDGAR Online is a public for-profit company that provides SEC fillings information extracted from SEC's EDGAR.

6. In this study (2012), the authors added additional company to their sample, making total number of companies equal to 27.

7. EDGAR Online extracts information directly from SEC's XBRL filings.

around 50% of financial items that are reported in XBRL 10-Ks are not present in the aggregators' data sets, and that the percentage of mismatches between XBRL and aggregators' data ranges from 4.8% to 8% with 56% of all differences being material. Out of three financial statements – Balance Sheet, Income Statement, and Cash Flow Statement – Balance Sheet was associated with the smallest percentage of mismatches. In addition, Compustat data resulted in the largest percentage of mismatches (44.3%) and the lowest number of omissions (50.9%).

The most related studies to ours are Tallapally, Luehlfing, and Motha (2011); Tallapally, Luehlfing, and Motha (2012), and Boritz and No (2013). Therefore, we would like point out some key differences with them. Firstly, this study focuses on the accounting data provided by Compustat. The objective is to compare Compustat numbers to the numbers found in the 10-K filings. Although, we utilize XBRL 10-K filings to extract the numbers, we do not aim to compare Compustat and XBRL data since XBRL data also may differ from plain-text 10-K data. In our study, we try to remove all XBRL 10-K values that differ from 10-K values (using both manual and automated procedures described in §1.4). Secondly, we conduct a large-scale comparison that involves more than 5,000 companies (as opposed to 75 in Boritz and No (2013) and 27 in Tallapally, Luehlfing, and Motha (2012)) by automating the comparison procedure. Thirdly, we study not only the amount of discrepancies between the data sets, but also the magnitude of discrepancies and their effect on non-discrepancy observations. Fourthly, we develop and present a methodology for an automated (as opposed to manual) XBRL and Compustat data comparison. This methodology is critical for our study since such a large-scale study would not be possible without it. Finally, some of our findings differ from the above mentioned studies (e.g., we do not find significant differences in Total Liabilities between Compustat and 10-K data as Boritz and No (2013) do).

1.2.2 Data definitions and comparability

Stone (1968) is one of the first studies to caution researchers about the use of digital accounting databases and Compustat, in particular. The study recognizes several shortcomings of Compustat in terms of data comparability. First, it argues, the Compustat annual data inherits SEC 10-K filings weaknesses. Companies may not be directly comparable based only on the 10-K reports since they may use different accounting, e.g., FIFO versus LIFO, or straight line depreciation versus accelerated depreciation, etc. Second, the data definitions used to adjust 10-K data to make it more comparable may fail because of the previous point and may significantly alter the accounting used by a company. Moreover, it creates the appearance of uniformity of data, and a user of Compustat data may draw wrong conclusions assuming items to be comparable. Also, users not fully aware of data definitions are likely to run into problems when creating their own financial ratios or proxies. Last but not least, the study argues that Compustat does not provide qualitative disclosures that may be essential to understand the financial position of a company.

Thies and Revsine (1977) examine the implications of Compustat definitions and policies for the Capital Expenditures item on the empirical inflation accounting research, and specifically for the purpose of asset layering. Empirical inflation accounting studies often require to transform conventional accounting numbers to their inflation-adjusted estimates. This is generally done by arraying fixed assets in layers with respect to their acquisition (cost and dates). The study argues that Compustat item Capital Expenditures cannot be used effectively for the purpose of fixed asset layering since: 1) Compustat definition of the item does not include fixed assets acquired through merger or acquisition, and 2) Compustat report capital expenditures net of retirements (if this net figure is shown in a financial report). To test the adequacy of Compustat Capital Expenditures data, the authors impute fixed assets retirements from the Compustat data for the years 1960 through 1974 for S&P 425 industrial firms. They use a simple criterion: check how many imputed retirements

are negative.[8] The proportion of negative values ranged from 7.5% to 30.7% across different years, indicating that Compustat data is too distorted to be used for the purpose of fixed asset layering. The authors suggest to use SEC's 10-K data directly instead of Compustat data, even though it is less accessible and convenient.

Collins and O'Connor (1978) criticize Eskew (1975) study due to its failure to adjust data across different databases: Compustat, Moody's Industrial Reports, and companies' annual reports. They argue that data definitions in Compustat differ from the ones in Moody's and annual reports. The researchers should take special care of data definitions when matching and merging data from several databases, since it can lead to false conclusions. Collins and O'Connor replicate Eskew (1975) study by addressing shortcomings of the latter. The results obtained differ significantly from the original ones.

Guenther and Rosman (1994) study the differences between SIC codes assigned to companies by Compustat and CRSP databases and their effect on accounting research. Using a sample of 1,810 companies, the authors find 1989 Compustat and CRSP databases disagree: on 71% of all companies at the 4-digit SIC level, on 54% of companies at the 3-digit level, on 38% of companies at the 2-digit level, and on 22% of companies at the 1-digit level. The authors also examine the homogeneity of Compustat and CRSP. They find that the Compustat SIC code classification yields both higher correlation of intra-industry monthly stock returns and lower variances of intra-industry financial ratios than CRSP SIC classification. To test the impact of the differences in SIC classification between two databases, the authors replicate the study of Freeman and Tse (1992) using both Compustat and CRSP SIC codes. The results obtained using Compustat codes are similar to the results of Freeman and Tse since they also use Compustat. However, the results obtained by utilizing CRSP codes were significantly different from the original ones.

Kahle and Walkling (1996) do a follow-up study of Guenther and Rosman (1994)

8. Clearly, fixed assets retirements cannot be negative

by examining the differences in SIC classification of Compustat and CRSP databases for approximately 10,000 firms over the years of 1974-1993. They find similar level of discrepancies as Guenther and Rosman: there is 79% disagreement at the 4-digit SIC level, and 35% disagreement at the 2-digit SIC level. To further analyze the impact of these differences, the authors choose six financial characteristics[9], draw a random sample of firms for each characteristic, and conduct simulations to measure the power and specification of Compustat and CRSP SIC classifications based on these random samples. The authors find that Compustat's SIC classification is better at detecting abnormal performance[10] than CRSP's SIC classification.

1.3 Data alterations in Compustat

Compustat North America Fundamentals Annual data items may differ from the original accounting data items disclosed by companies in their annual financial reports. We identify four reasons for having such differences:

1. Compustat transformed original value to match Compustat's standard definition of the variable,
2. Compustat's value is erroneous (due to typos, rounding, etc.),
3. Compustat's value is not up to date,
4. Compustat does not provide a value for the data item (i.e., missing data).

We found data standardization to be the main source of discrepancies between Compustat and 10-K data. Compustat argues that data standardization results in

9. Specifically, the authors choose: Operating Return on Sales, Operating Return on Assets, Leverage, Asset Turnover, Payout ratio, and Market-to-Book ratio.

10. Abnormal performance in Kahle and Walkling (1996) is simulated by adding errors to the variables to increase the difference between sample and control firms' characteristics.

more "consistent and comparable data" that provides "solid foundation for your rigorous analysis".[11] However, some studies have expressed skepticism on whether such alterations improve data comparability (e.g. Stone 1968). All data in Compustat is standardized by default, and in most cases footnotes and data codes do not indicate whether the values have being altered.[12] However, Compustat does provide transparency and analyst notes data that explain adjustments that have been made to obtain standardized items. In this study, we do not rely on Compustat's transparency and analyst notes data to avoid potential data provider bias, and use XBRL 10-K filings as an independent source of original financial data provided by companies to compare values between Compustat and 10-K, and reconcile discrepancies between the two data sets.

An example of data standardization is Compustat reporting Total Assets of American Water Works Company at the end of December 2011 to be $13,809,643,000, while the value of Total Assets in the 10-K being $14,776,391,000. The difference is the value of Contributions in Aid of Construction of $966,748,000.[13] It is not obvious whether this adjustment enhances data comparability - by not including Contributions in Aid of Construction item in Total Assets Compustat underreports the amount of resources that the company has.

Erroneous data is another reason why Compustat numbers may differ from the original numbers reported by companies. Unlike standardized data, errors are unintentional alterations of the original or standardized data. Errors may occur due to input typos, use of wrong accounts and balances, data misinterpretation, data

11. www.compustat.com, as of February 16, 2014

12. A very small number of alterations is reported through footnotes. For example, footnote "JE" indicates that the reported value differs form the reported amount by deferred taxes.

13. American Water Works Company in its 10-K describes Contributions in Aid Of Constructions as follows: "Regulated utility subsidiaries may receive advances and contributions from customers, home builders and real estate developers to fund construction necessary to extend service to new areas . . . Advances that are no longer refundable are reclassified to contributions in aid of construction. Contributions in aid of construction are permanent collections of plant assets or cash for a particular construction project."

rounding, data calculations, etc. The danger of having errors in the data is that they are hard to identify – they may appear to be standardized numbers. For example, Compustat reports the Gross Profit of Transwitch Corp for the year of 2011 to be $19,932,000, while the company reports the value of Gross Profit to be $17,932,000. The difference of 2 million U.S. dollars might be due to Compustat's adjustment, although we failed to reconcile this discrepancy. It may also be due to input typo since the two numbers differ only in one digit.

Accounting researchers are well-aware of the existence of erroneous data in popular data sets. Sometimes they assume that errors in the data will result in some extreme values. Therefore, they use outlier detection techniques to find extreme values, and either delete them or transform them (using techniques such as winsorising). However, this approach is not ideal since erroneous values do not have to be extreme, and extreme values do not have to be erroneous. In many cases, extreme values that are not erroneous should be included in a study sample since they represent possible states of the object studied. Moreover, deleting extreme observations is very likely to affect the output of most empirical models due to their sensitivity to outliers, and this may drastically change the results of the study.

A different type of data discrepancy between Compustat and companies' financial reports is when Compustat does not update accounting numbers due to amendments. Compustat uses amended filings (i.e., 10-K/A and 8-K/A) to update North America Fundamentals Annual data, but does not use restated numbers from the subsequent 10-K or 8-K forms to update its annual data.[14] Unfortunately, in some cases, Compustat fails to updates its data items after amendments in a timely manner. For instance, on March 15, 2012, ADA-ES reported its Net Loss to be $19,851,000; however, on October 19, 2012, the company issued a 10-K/A and restated its Net Loss

14. According to Standard & Poor's "Compustat Understanding the Data", March 2014, document: "When a company files an amended source, such as a 10-Q/A or 10-K/A, S&P updates this data and treats it as a new source for that period. The amended source overrides the original source...Quarterly income statement data is restated by S&P and Annual data is not."

to be $22,819,000. As of February 22, 2013, Compustat has not updated the originally reported number with the restated one. The resulting difference of $2,968,000 amounts to more than 10% of the Net Loss.

Data that are not updated should be considered separately from erroneous data since there were no errors in recording the original (not restated) numbers from financial reports. On the other hand, the fact the numbers were restated indicates that there is a material difference between the original and updated numbers. Hence, using accounting data that has not been updated may result in significantly different results of an empirical study.

Another reason why Compustat data may differ from is the absence of data values for some variables in Compustat database. For example, Compustat did not report the Total Assets of Airwave Labs ($1,139,182) at the end of 2012, although Compustat did report company's Total Liabilities and Stockholder's Equity. Accounting researchers usually drop observations with missing data, or in rare cases substitute missing data with estimated values (e.g., industry averages). Clearly, in some cases a variable is not applicable to a company (e.g., Sales), or may not be reported in financial reports (e.g., non-GAAP measures). However, if an accounting number is reported in a financial statement, and there is a corresponding Compustat variable for that number, then it should be present in Compustat. Missing data does not allow researchers to study the whole population, and similarly to data alterations, may change the results of empirical studies.

1.4 Comparison of Compustat annual data with 10-K numbers using XBRL

The best way to assess the amount and magnitude of data alterations in Compustat is to compare the numerical financial data reported in Compustat to the original data reported in companies' financial reports. In particular, the annual accounting

numbers in Compustat can be compared to the accounting numbers found in the 10-K financial reports.

However, a large-scale comparison of accounting data items reported in Compustat and corporate 10-K filings would be a very costly procedure, if done manually. It was the only option not so long ago. However, since the SEC mandated the use of the XBRL reporting technology in the SEC filings for all U.S. GAAP filers for the fiscal years ending on or after June 15th, 2011, it has become possible to automate this process.

1.4.1 XBRL as a means for data comparison

In our study, we take advantage of the newly available XBRL financial reporting to compare the annual data items reported in Compustat North America Fundamentals. In essence, XBRL is a formal language for communicating business information. Data items are described using meta-information, and are linked together through various relationships. Information carried by XBRL is both human- and computer-readable. A special software can render an XBRL document and present it to an end-user as an electronic document humans are used to dealing with. However, the real advantage of XBRL is that it allows computer software to parse XBRL documents, and find, extract, and present information in an automated fashion. This drastically reduces the cost of manual labor needed to process such documents. It also allows to process a large amount of documents within a matter of seconds.

The SEC has realized the benefits of XBRL reporting, and in April 2005 the SEC adopted *XBRL Voluntary Filing Program* (SEC 2005). This program enabled U.S. GAAP filers to voluntarily prepare financial statements using XBRL. The objective of the program was to determine the usefulness of XBRL as a format for reporting financial information by analyzing volunteers' feedback comments. In 2009, the SEC adopted *Interactive Data to Improve Financial Reporting* final rules (2009) that mandated companies that prepare their financial statements in accordance with the

U.S. GAAP to file supplemental XBRL documents for 10-Q, 10-K, and 8-K reports in addition to plain-text ones. This mandate was implemented in several phases. In the first phase, domestic and foreign large accelerated filers with worldwide public common equity float above $5 billion and the fiscal periods ending on or after June 15, 2009, were required to file with XBRL. In the last phase, all the U.S. GAAP companies were required to file using XBRL for the fiscal years ending on, or after June 15, 2011.

One of the most fundamental uses of corporate accounting data is for financial data comparison. Users of financial statements need to compare financial positions of various companies for their decision-making. Since XBRL is just a language that can be used for business data reporting, it does not have built-in capabilities for multi-document data comparison. Therefore, the SEC has created the XBRL U.S. GAAP Financial Reporting Taxonomy. This taxonomy is a collection of common accounting data concepts, definitions, types, and relations that is meant to accommodate most of companies' financial reporting needs. Since most accounting items in financial reports is fairly standard, the XBRL U.S. GAAP Financial Reporting Taxonomy defines common rules how to present this standard information in XBRL filings. Therefore, in theory, standard accounting numbers can be easily compared across different filings using the XBRL U.S. GAAP Financial Reporting Taxonomy.

Accounting information that is not standard, i.e., company- and filing-specific information, is represented in XBRL documents through so-called extensions. Extensions are an important part of XBRL filings that provide additional reporting flexibility. However, the cost of utilizing extensions is the reduced comparability since extensions are not necessarily created in the same manner by different filers. Before using an extension concept in a filing, the filer should search the XBRL U.S. GAAP Financial Reporting Taxonomy for a possible match. If no suitable match is found, the filer is allowed to create an extension. Debreceny et al. (2011) examined 67 XBRL filings from the period of April 15,2009 to June 2010, and found that more

than 40% of all extensions were unnecessary because the corresponding elements exist in the U.S. GAAP Financial Reporting Taxonomy.

As mentioned above, all public U.S. GAAP companies are required to file their financial reports using the XBRL reporting technology starting from June 15, 2011, with standard accounting data items being formally reported in a similar manner due to the use of common XBRL U.S. GAAP Financial Reporting Taxonomy. Given that accounting variables found in Compustat are fairly standard, it becomes clear that XBRL filings can be utilized to automatically extract standard accounting numbers from corporate financial reports and compare them to numbers found in Compustat.

1.4.2 Comparison methodology

In our study, we compare annual accounting data of 5014 company-unique XBRL filings (10-Ks) to the appropriate Compustat data for the period from October 1, 2011, to September 30, 2012.[15] There are several reasons why we consider this particular period. First of all, to better assess the data quality in Compustat we analyze only one annual filing per company. This ensures that we do not introduce additional bias and noise to the data and cover all possible companies and industries. Secondly, we analyze filings for the recent 2011-2012 period to capture the current data quality state of the Compustat data set. Moreover, a more recent period makes more sense from the XBRL perspective since annual XBRL filings are a recent addition to corporate 10-K filings, and recent XBRL filings are likely to be more accurate than the older ones (e.g., Du, Vasarhelyi, and Zheng 2013). Finally, we downloaded the data on February 22, 2013 which introduces a 145-day lag between the latest date in the considered period (September 30, 2012) and the data download date. This ensures that most of the late filings are captured in our data, and that Compustat has had sufficient amount of time to update missing and restated (as reported in companies'

15. We only consider annual (and quarterly) reports since these are verified by audit firms.

amendments) data items.

Our comparison methodology comprises of six steps:

1. Extracting data from Compustat.
2. Extracting data from XBRL 10-K filings.
3. Merging Compustat and XBRL data.
4. Creating mappings between Compustat variables and XBRL reporting concepts.
5. Calculating differences between Compustat variables and the associated XBRL reporting concepts.
6. Analyzing discrepancies between Compustat and XBRL 10-K filings.

Below, we explain each step in more details.

Extracting data from Compustat

One of the main advantages of using Compustat is that the data extraction process is fairly simple. This feature is particularly attractive to researchers that rely on accounting numbers in their studies. However, some users of Compustat may not realize that some accounting numbers reported by companies have been altered to fit Compustat's definitions of variables. It is important to understand the data alterations and the effects they may have on the output of a research that relies on the Compustat data.

We extracted all the available variables from Compustat North America Fundamentals Annual through Wharton Research Data Services (WRDS) interface for the U.S. companies for the period from October 1, 2011, to September 30, 2012. If there was more than one observation for a company, we kept only the latest one.

Dollar values in XBRL filings are measured in $1 denominations. However, in the Compustat data set, different variables use different units of measurement

(numéraires). For example, the Net Income variable is measured in millions of U.S. dollars ($1,000,000), while Earnings per Share (EPS) is measured in U.S. dollars ($1). We used variable definitions in Compustat Manual to transform all Compustat values to be measured in U.S. dollars ($1).

Extracting data from XBRL 10-K filings

We downloaded all the domestic XBRL 10-K filings (including amendments) from the SEC's Electronic Data-Gathering, Analysis, and Retrieval (EDGAR) system's File Transfer Protocol (FTP) server with the reported fiscal year being in the one-year range from October 1, 2011, to September 30, 2012.[16] The filings were downloaded on February 22, 2013. Hence, most of the late filings and filing amendments are likely to be included in the data set. For each company, we kept only the latest 10-K filing, and, where applicable, all of its amendments (i.e., 10-K/As). We kept only the latest company filing to ensure that the reported amount and magnitude of discrepancies are not biased by certain types of companies and data alterations. And the reason to have both amendments and the original filings is to verify whether Compustat updates original numbers to restated ones in a timely fashion.

As mentioned previously, XBRL 10-K filings use the XBRL U.S. GAAP Financial Reporting Taxonomy, a collection of common financial and reporting concepts, definitions, types and relationships. Although, filers can define their own XBRL data concepts (extensions), SEC encourages them to use standard concepts found in the XBRL U.S. GAAP Financial Reporting Taxonomy whenever possible. By doing this, filers in their XBRL fillings would report similar accounting concepts using similar XBRL U.S. GAAP elements, which would greatly enhance data comparability across different companies and filings.

In our study, we compare 30 common accounting concepts from Balance Sheet, Income Statement, and Statement of Cash Flows between Compustat and 10-K data

16. SEC's EDGAR's FTP server can be reached at ftp://ftp.sec.gov/edgar/.

sets.[17] All of these concepts are present in the XBRL U.S. GAAP Financial Reporting Taxonomy. We do not consider extensions. In addition, we only consider numbers that were tagged in XBRL 10-K filings individually (detailed tagging) as opposed to being a part of a text block (block tagging) to ensure accurate data extraction. We also only consider numbers measured in U.S. dollars (or dollars per share) to avoid us introducing bias of currency conversion.

Unfortunately, extracting data from XBRL documents is not a simple process. The main problem is that filers do not utilize XBRL reporting technology in a consistent manner. The current SEC's implementation of the XBRL U.S. GAAP Financial Reporting taxonomy allows a certain amount of flexibility in reporting. In addition, we and other researchers (2013) find that many XBRL filings contain errors. These findings should be viewed in the context of 24-month liability provision for the first-time filers.[18] The end result is that even though filers use the same XBRL U.S. GAAP Financial Reporting Taxonomy, different filers describe similar standard accounting items differently in their XBRL reports. As a consequence, it is quite a challenge to extract accurately the information of interest from many filings in an automated fashion.

A particularly problematic current XBRL reporting practice is the use of XBRL contexts. An XBRL *context* provides additional information about the concept reported, usually including such information as the concept's associated period of report, entity, dimension, etc. Although a useful instrument, companies tend to use it differently, thus reducing the benefit of XBRL reporting standardization. For example, the XBRL U.S. GAAP Financial Reporting Taxonomy defines "LegalEntityAxis" dimension that identifies the specific entity that a concept is related to. For an XBRL concept associated with a parent company, different fillings may provide: 1) no value for the XBRL "LegalEntityAxis" dimension, 2) "ParentCompanyMember" value for

17. See Table A.2 for the full list of 30 accounting concepts analyzed.

18. The limited liability provision was set to expire on October 31, 2014.

the "LegalEntityAxis", or 3) a custom member definition for the "LegalEntityAxis" dimension. This example shows two major issues. The first one is that it is not clear what context should a concept of interest have. In other words, there can be many instances of the same concept, each having its own context; the problem is to pick the correct one. The other issue is the concepts' entity attribution – in many cases, a filing reports accounting information for many entities, and it may be problematic to attribute each concept to its corresponding entity since each entity may have its own value for the "LegalEntityAxis" dimension that is usually an extension (i.e., not defined in the XBRL U.S. GAAP Financial Reporting Taxonomy). For example, Huntsman Corporation and Huntsman International LLC are manufacturers of chemical goods that jointly file annual and quarterly filings. However, each company has different values for similar accounts and a different unique identifier with both SEC and Compustat. Since these companies share a common XBRL filing it is important to attribute numbers to the correct entities.

Theoretically, if one filing is associated with several U.S. GAAP filers that have Central Index Keys (CIKs) assigned to them, the XBRL filing may include information about those filers through the means of special "Document and Entity Information" (DEI) XBRL concepts. However, in practice, this is rarely the case, and not much information is revealed about the entities other than the main one in the XBRL filing. In the case of the above-mentioned Huntsman Corporation and Huntsman International LLC, only the entity information of Huntsman Corporation is disclosed through DEI XBRL Concepts in their joint 2011 10-K filing.

The issues described above make the processes of extraction of relevant information and its attribution to various entities very complicated and perhaps even ambiguous. It is true that those issues are irrelevant when a user of financial statements is concerned with only one filing – an XBRL software can render the filing in a comprehensible format. However, the same functionality is provided by a plain-text

filing without the cost of utilizing the XBRL reporting technology. The main advantage of XBRL is that computer software can automatically process the embedded tags in the XBRL document without any need for user intervention. This is especially important when a software processes many XBRL documents and compares the data within those documents. Hence, the value of XBRL decreases when filers become too frivolous in the way they utilize XBRL.

To overcome the problem of attributing a filing's reporting concepts to entities associated with that filing, we need to identify the correspondence between the values of the XBRL "EntityLegalAxis" dimension embedded in the concept contexts and the actual entities. In this study, we utilize the following method:

1. Use the EDGAR index file to extract information about entities, such as legal names, Central Index Keys (CIKs), addresses, etc.[19]

2. Whenever possible extract entity context information from "Document and Entity Information" (DEI) XBRL concepts embedded in the XBRL instance document.

3. Analyze common entity member values such as "ParentCompanyMember", "SuccessorMember", etc. in the XBRL instance document.

4. Extract textual descriptions (labels) for all possible values of the "EntityLegalAxis" dimension found in the XBRL label linkbase document.

5. Match entities' information found in the EDGAR index files with the values found in the XBRL filing.

The last step involves comparing the values of CIKs (whenever possible), or the descriptions (labels) of the XBRL entity member values with the entities' legal names.

19. Every 10-K filing in EDGAR is associated with an index file that contain basic information about the filing and filing entities.

Since the entity names are often not exact, e.g. "Microsoft Corporation" and "Microsoft Corp.", we use the Jaccard text similarity measure (Jaccard 1901) to compare labels and EDGAR legal names. The Jaccard similarity measure is defined as follows. Formally, given two pieces of text, let T_1 be the set of all words in the first text, and T_2 be the set of all words in the second one. The Jaccard similarity measure between these two pieces of text is defined as

$$J(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}. \tag{1.1}$$

That is, the Jaccard similarity measure is equal to the number of shared words divided by the number of all words. We consider two pieces of text to be similar, only if the similarity measure between them is 0.3 or greater.

Finally, by means of their contexts, XBRL concepts may be defined in other dimensions than the entity dimension, such as geographical area. Since we want to compare XBRL data items with Compustat data items that are usually general in nature, we do not extract the concepts that are defined in dimensions other than the entity dimension. The only exceptions are dimensions that indicate a restatement, type of financial report, or components of partner capital.

For the purpose of our study, we created our own applications that would process XBRL 10-K filings due to unavailability of such software in the public domain. However, recently a number of tools have appeared that are able to process a large number of XBRL filings. Many of these tools are still in the development stage, and may not be fully functional. However, potentially they can be used by academics and practitioners to obtain the original data as reported by companies.

To ensure that our applications reliably process XBRL data, we adopt a software verification technique called software walk-through. This peer review technique is one of the standard procedures defined in the IEEE Standard for Software Reviews and Audits (IEEE 2008). This procedure has been empirically shown to outperform

other common software validation procedures such as functional and structural testing (Basili and Selby 1987) and computer-based testing (Myers 1978).

Merging Compustat and XBRL 10-K data

Data items in Compustat and XBRL 10-K observations should be merged based on the same entity and reporting period. Matching data by reporting periods does not present a particular problem. However, matching by entities does.

Each company in the EDGAR database is identified by the SEC's unique identifier – Central Index Key (CIK). In addition, EDGAR's XBRL filing document name convention requires to identify the main entity through its Ticker code, a unique identifier for stock market public traders. Compustat has its own unique identifier called Gvkey. However, Compustat does provide values for both CIK and Ticker.

Central Index Key is the best unique entity identifier to match data items from both data sets since EDGAR may not provide Ticker codes for all entities associated with a filing. However, for many observations, the value of the CIK variable in Compustat is missing. In those cases, we use either 1) the combination of Ticker code, company name and address, or 2) the company name and address to match the entities across data sets. As before, since the company name and address may not be represented uniquely (e.g. "One Green Avenue" versus "1 Green Ave."), we use the Jaccard text similarity measure (described in §1.4.2) to match the appropriate text fields.[20]

Overall, we extracted 7,466 XBRL 10-K company-unique observations, and 7,375 Compustat company-unique observations for the period from October 1, 2011, to September 30, 2012. Out of those, we were able to merge 5,014 observations. Most of the companies that we were unable to match were either 1) too small to be present

20. Using textual similarity techniques to match Compustat and 10-K companies adds 12 more companies to our analysis. We have repeated the discrepancy analysis without those 12 companies. The results did not change.

in Compustat 2) not present in the EDGAR database, or 3) did not file 10-K with XBRL.[21]

Creating mappings between Compustat variables and XBRL reporting concepts

To calculate differences between Compustat and XBRL 10-K data sets, each Compustat item has to be matched to the appropriate XBRL reporting concept. It is important to note that this is not a one-to-one match. For example, Compustat's variable "Depreciation and Amortization" can be represented in the XBRL U.S. GAAP Financial Reporting Taxonomy by means of either "Depreciation," "DepreciationAndAmortization," or "DepreciationDepletionAndAmortization" reporting concept. That is, depending on the filing either of those three XBRL concepts can match Compustat's Depreciation and Amortization variable.

Creating a mapping between Compustat variables and XBRL manually would be very costly and, most likely, inefficient. For this study, we developed an algorithm that creates mappings between variables in the data sets in an automated fashion. The algorithm is implemented under the assumption that Compustat variables and the corresponding 10-K XBRL concepts should have similar values in most cases when the values in Compustat are not altered. In other words, we assume the number of discrepancies between the two data sets to be rather small for the unaltered data. For each Compustat variable, the algorithm identifies a set of XBRL U.S. GAAP concepts that may match the variable using the information from the merged data set. In this study we compare 30 common accounting items that are present in the XBRL U.S. GAAP Financial Reporting Taxonomy, and therefore the algorithm does not utilize XBRL extensions.

---

21. The SEC's mandate to file using XBRL as of June 15, 2011, required to file all financial reports starting with the first 10-Q. That is, if for some company the first filing after June 15, 2011, was not 10-Q, but 10-K, the company was allowed not to file this 10-K using XBRL. For more details see item 601, Regulation S-K.

The mapping algorithm works as follows. For each Compustat variable, it identifies all observations from the merged data set where that Compustat variable has a non-zero value.[22] Then, for each of those observations, the algorithm finds all the U.S. GAAP XBRL concepts whose values are equal to the ones of the Compustat variable (up to a rounding error), if there are any. XBRL concepts that match the Compustat variable in at least 10% of all cases are considered to be *mapping candidates* for that Compustat variable. Ideally, a Compustat variable should match at least one of its mapping candidates in all the observations. Finally, the algorithm assesses how well the mapping candidates fit a Compustat variable, by calculating a measure that we call *mapping coverage*. Mapping coverage is defined as the number of observations where Compustat variable was matched by at least one of the XBRL mapping candidates divided by the number of all observations with non-zero values of that Compustat variable. For example a mapping coverage of 0.5 means that the obtained variable mapping explains at least 50% of Compustat values for that variable. Most variables that we used in our analysis have relatively high values of mapping coverage, but some do not due to large amounts of Compustat data alterations.

To sum up, for each Compustat variable, we use the sample of merged observations to find a set of standard XBRL concepts that often match (up to rounding error) that Compustat variable. These standard XBRL concepts (mapping candidates) are then used to calculate differences between the values of that Compustat variable and the values reported in XBRL 10-K reports.

Calculating differences between Compustat variables and the associated XBRL reporting concepts

Once the data sets have been merged and the mappings between Compustat and XBRL 10-K items have been established, it is possible to calculate the data differences

22. Although, Compustat data allows variables to have no value, we found that sometimes zero values may also indicate no value. We decided to excluded zero values since they may yield unreliable mappings.

between the two data sets.

For each Compustat variable, we compare values of that variable to values of the appropriate XBRL concepts as defined by the mappings that we have previously established. If a value of a Compustat variable matches the value of at least one matching XBRL concept, then we assume that there is no difference between the Compustat value and the XBRL 10-K value. In general, when we compare values of Compustat and XBRL 10-K filings, three cases are possible:

1. There is no difference between values.

2. There is a difference between values.

3. Either Compustat or XBRL value is missing.

It is not clear why the Compustat data set would have missing values apart from the case where the values are not applicable or not reported by a filer. From the XBRL perspective, if Compustat reports a value and XBRL 10-K filing does not report matching concepts, then this is an artifact of either Compustat reporting calculated values, shortcomings of the variable mapping process, or errors in the XBRL filing. Unfortunately, we found that there are a number of issues with XBRL 10-K reports that result in missing values. Some typical problems are:

- Accounting number is not reported in a 10-K filing.[23]

- Accounting number is not reported in an XBRL filing.

- Accounting number is reported using a wrong XBRL U.S. GAAP concept.

- A standard accounting number is reported through the use of extension.

23. In some cases, 10-K filings do not explicitly provide values of accounting items. For example, many U.S. GAAP companies (approximately 30%) do not directly report values of Total Liabilities in their financial reports – Total Liabilities are often reported together with Stockholder's equity as one number. An advantage of using Compustat, is that Compustat calculates many accounting numbers that are not directly reported in financial statements. However, verification of those numbers may not be an easy task due to the complex nature of calculations of some accounting items.

- Accounting number refers to a wrong period.

In many cases, there is a difference between the values reported in Compustat and XBRL 10-K. We express this difference as a relative difference. We utilize the relative difference measure as opposed to the absolute one since it better reflects the magnitude of differences and allows to compare the extent of differences between different Compustat variables.

In cases where there is just one matching XBRL concept with a non-zero value, the calculation of the relative difference between Compustat and XBRL data items is straightforward. However, there may be cases where there are more than one matching XBRL concepts with different non-zero values. In such cases, we choose the XBRL concept whose value differs from the Compustat value the least, and calculate the relative difference between that XBRL concept and the Compustat value.

Formally, if c is the value of the Compustat variable, and $x_1, x_2, \ldots, x_n$ are the values of the matching XBRL concepts, we choose the value x from $x_1, x_2, \ldots, x_n$ such that

$$x = \underset{i=1\ldots n}{\operatorname{argmin}} x_i - c. \tag{1.2}$$

The chosen value of x is then used to calculate the relative difference between 10-K and Compustat values:

$$\frac{x - c}{c} \tag{1.3}$$

This is a generalization of the regular relative difference – i.e., if there is only one matching XBRL value, than expressions (1.2)-(1.3) are equivalent to the regular relative difference.

Analyzing discrepancies between Compustat and XBRL 10-K filings

The aim of this study is to compare numbers in Compustat to the original numbers in 10-K filings. We utilize XBRL technology to extract numbers from XBRL 10-K reports and compare them to Compustat numbers. There is a risk that some

discrepancies obtained by this method are the products of XBRL reporting. To address this issue, we examine the causes of each discrepancy through both automated and manual procedures.

Ideally, the numbers as reported in XBRL 10-K filing should precisely match the numbers as reported in the corresponding plain-text 10-K filings. However, XBRL 10-K reports may (and do) contain errors.[24] For example, values may be reported without minus signs making them positive when they should be negative, or values may be reported in thousands of dollars instead of dollars. These will result in discrepancies. In addition, XBRL reports may use extensions or non-conventional dimensions to describe the data items. The algorithm we utilize to extract and match XBRL data may not be able to capture such XBRL items. If this happens, then in the best case it will result in a missing XBRL value in our matched data. In the worst case, it will result in a difference in the matched data set since the algorithm could have extracted other XBRL items that could sometimes match a Compustat variable (i.e., mapping candidates, see §1.4.2).

Given that discrepancies between data sets may exist due to XBRL-related errors, it is necessary to eliminate all such discrepancies from our analysis in order to assess the true amount and magnitude of data alterations in Compustat. Fortunately, it is possible to identify sources of most discrepancies in an automated fashion. For example, a simple check against a Compustat value may help identify if an XBRL value has a wrong sign. In addition to simple checks that compare data values against each other, we employ other automated procedures described below that target specific sources of discrepancies.

To find out whether the extraction algorithm was unable to extract the correct value from XBRL 10-K due to the value being reported in a non-standard manner or

24. For this study, the most important errors are errors that result in wrong values of wrong XBRL concepts. These may affect the results of our study, and are the ones that we try to eliminate. However, there may be other errors in XBRL data that do not affect the results of our study. For more information on 10-K XBRL quality, please see Hoffman (2013) and Boritz and No (2008).

with an error (e.g. extension, wrong tag name, non-conventional dimensions, etc.), we employ an algorithm that automatically scans the XBRL 10-K filing to see if any other XBRL concept matches the Compustat value. If there is such a concept, the algorithm reports the reason for the XBRL concept being not extracted or matched. For example, a standard accounting item may be reported through the use of XBRL extensions, or the item may be attributed to a different entity.

As mentioned previously, Compustat sometimes fails to update numbers after a 10-K/A with restated numbers has been filed. If the original 10-K filing was filed using the XBRL reporting technology, it is possible to check whether a discrepancy between Compustat and XBRL filing is due to the Compustat value being not up to date by comparing the Compustat value to the matching XBRL values in the original, pre-amendment 10-K filing.

Finally and most importantly, Compustat often alters original numbers reported in corporate filings to standardize values according to Compustat's definitions. It would be very costly to identify such alterations using manual procedures. Therefore, we developed an algorithm that tries to explain a difference between Compustat and XBRL 10-K filing by searching the XBRL filing for concepts whose values could explain the difference (up to a rounding error). Essentially, the algorithm attempts to find the adjustments made to the original 10-K number by Compustat. The algorithm first tries to find an XBRL concept in the XBRL 10-K filing that if subtracted or added to the matching XBRL value results in no discrepancy with the Compustat value. If no such item is found, the algorithm tries to find a combination of two XBRL concepts and combination of subtractions/additions that would result in no discrepancy. Finally, if no such combination was found, the algorithm will search for a combination of three XBRL concepts and their respective combination of subtractions/additions that would result in no discrepancy. If no such ternary combination is found, the algorithm stops.

It should be noted that although the algorithm described above is very powerful

at explaining Compustat data standardization, it may yield a spurious combination of complementing XBRL items - i.e., a combination that by chance results in adjustments that yield no discrepancies between Compustat and XBRL items. Therefore, it is imperative to identify such spurious combinations. For each variable, and each automatically found combination, we manually scan the adjustment and assess whether it appears suspicious or not. For example, it is very likely that Compustat may adjust the value of Cost of Goods Sold by adding the value of Labor and Related Expense item; however, it is highly unlikely that Compustat will adjust the value of Cost of Goods Sold by adding the value of Current Assets. The latter incident must be investigated manually by comparing the Compustat's value to the plain-text 10-K's value. Although we do apply our best judgment when deciding whether an automatically-found adjustment should be investigated manually or not, this is still an ambiguous process, and hence is a limitation of our study. On the other hand, we are interested only in those spurious combinations that help us eliminate XBRL errors, and we have found that the number of XBRL errors among these is very low (around 0% to 3%, depending on the variable).[25]

The automated procedures described below are able to explain most of discrepancies. However, some discrepancies still require manual checks in order to be explained. In those cases, we compare values found in Compustat and XBRL 10-Ks to the values in the corresponding plain-text 10-K reports. It should be noted that the nature of discrepancies that require manual checks is such that it may be impossible to identify the exact reasons for having these discrepancies. However, we can usually tell whether the discrepancy is due to adjustment/error in Compustat, XBRL 10-K report, or both. Overall, we have manually compared 1,800 discrepancy items (around 1.5% of all items) using original 10-K filings.

---

25. Most XBRL errors are usually found by applying other checks described in the section that targets XBRL errors specifically.

1.5 Analysis and results

For our analysis of differences between Compustat and 10-K numbers, we have chosen 30 top-level Compustat variables that are commonly reported by financial entities and frequently used in accounting academic literature. For the full list of variables, their description and descriptive statistics, please see Table A.2. Half of the analyzed variables are Balance Sheet variables, 11 variables are Income Statement variables, and 4 variables are Cash Flow Statement variables.

For all 30 Compustat variables we have performed the comparison procedure described in §1.4, i.e., we have merged the Compustat and XBRL 10-K data, created mappings between data sets, calculated differences and analyzed discrepancies. Most variable values reported in Compustat have been successfully matched to the appropriate XBRL 10-K values. However, some have not – this is mostly due to filers not reporting numbers that can be calculated from other 10-K numbers, or due to the use of XBRL extension concepts. It is interesting to note from Table A.2 that in some cases Compustat did not report values that were present in the 10-Ks (e.g., Compustat did not report Total Assets values for 42 companies in our sample). Although, the number of missing observations is not very large, this is a cause for concern.

As mentioned in §1.4.2, we were able to reconcile most of the discrepancies between Compustat and 10-K data. This indicates that the main reason for the differences between Compustat and 10-K reports are the data standardization procedures implemented by Compustat.

1.5.1 Discrepancy analysis by variables

Table A.3 reports the amount and magnitude of discrepancies for each Compustat variable. As mentioned before, we define a discrepancy as a difference between Compustat and XBRL 10-K numbers that exceeds a rounding error. Many of these

discrepancies would be considered "material" from the audit perspective. To measure the amount of "material" discrepancies we adopt a rule-of-thumb definition of material discrepancies – a Balance Sheet discrepancy is "material" if it exceeds 0.5% of Total Assets, an Income Statement discrepancy is "material" if it exceeds 5% of Net Income, and a Cash Flow Statement discrepancy is "material" if it exceeds 5% of reporting periods' change in Cash amount. These definitions of materiality are commonly used in practice and were utilized in Boritz and No (2013). In addition, we consider an Earnings per Share discrepancies to be material if the absolute difference between data set values exceeds 5 cents.[26]

Although out sample size is quite large, it does not cover the whole population of available observations in Compustat. Assuming that our sample is representative of the population, for each variable, we estimate the minimum percentage of "material" discrepancies between Compustat and 10-K in the population (with 99% probability) based on the amount of material discrepancy observations in our sample. This estimate is a lower bound of a 99% confidence interval of a one-tailed binomial test that compares the number of material discrepancies to the number of all matched observations for a given variable. The estimates are provided in Table A.3.

In Table A.3, we report two types of discrepancy statistics with mean, median, and standard deviation information. The first one summarizes difference information about the observations that resulted in discrepancies, i.e., this is descriptive statistics of discrepancy observations. Specifically, for each discrepancy observation we calculate the *absolute* relative difference between 10-K and Compustat values, and report mean, median, and standard deviation of all such differences for a given accounting variable.

The second type of descriptive statistics summarizes difference information about all matched Compustat and XBRL 10-K observations. These statistics are meant

26. Some audit practices set the amount of materiality to the level that would change the value of EPS by 1 cent. We decided to take a more conservative (with respect to deciding what a discrepancy is) approach and define a discrepancy in EPS to be material if it exceeds 5 cents.

to show the effects of discrepancy observations on the whole population of observations. We calculate (non-absolute) relative differences between all matched 10-K and Compustat values. The mean value indicates the average relative difference between all Compustat and XBRL 10-K matched observations with positive values indicating that 10-K values on average are greater than Compustat values, and negative values indicating vice versa. For each variable, we test whether there is a significant difference between Compustat and XBRL 10-K numbers by conducting non-parametric Wilcoxon's signed-rank test.[27] The results of these tests show that values of 17 (out of 30) variables significantly differ across Compustat and XBRL 10-K data sets (with the significance level of at least 95%). In addition to the mean values, we also report the median values of relative differences between Compustat and XBRL matched population, but since median is the "middle value", these values are equal to 0 unless the amount of discrepancy observations for a variable exceeds the amount of non-discrepancy observations.

The results in Table A.3 indicate that there are significant differences between Compustat and 10-K numbers. In addition, we would like to point out a couple of interesting observations. Firstly, the amount of discrepancies, their magnitude, and their overall effect on all matched observations differ by variable. High-level variables that have simple definitions (e.g. Total Assets = all assets, Total Liabilities = all liabilities, Net Income = all revenues - all expenses) tend to have less discrepancies than variables that have more complex definitions (e.g. Cost of Goods Sold, or Gross Profit).[28][29] Since the components of more complex variables are more likely to differ

27. We do not use absolute relative differences between 10-K and Compustat for the Wilcoxon signed-rank tests, but instead use singed (non-absolute) relative differences. This may result in differences' effects being statistically underestimated since opposite sign values may cancel out each other. Using absolute relative differences would invalidate Wilcoxon signed-rank tests (as well as t-tests) and overestimate the effecst of the differences.

28. A smaller number of discrepancies does not necessarily result in smaller magnitude or overall effect on all matched observations.

29. A significant amount of discrepancies in Stockholder's Equity variables is caused by Compustat including temporary, minority, and other equity often not included by companies in their reported

across companies, industries and time, such variables tend to be more heavily adjusted by Compustat to fit Compustat's variable standard definitions.

Secondly, Compustat provides two variables of Retained Earnings item that we include in our analysis – Retained Earnings (RE) and Retained Earnings Unadjusted (REUNA). The difference is that the former is the Compustat-adjusted (standardized) version of Retained Earnings, while the latter should represent numbers as reported in 10-Ks. From Table A.3, it follows that the values of REUNA do differ from XBRL 10-K values in 1.36% of all cases (with 0.76% being material). However, more striking is the fact that RE values differ from XBRL 10-K values in 72.21% of all cases (with 37.94% being material). Such a large difference between REUNA and RE variables cannot be an artifact of the XBRL reporting or due to limitations of our study (as discussed in §1.4.2) since REUNA values have matched XBRL 10-K values in 98.64% of all cases. This example indicates that Compustat heavily adjusts original numbers reported in 10-Ks.

Finally, as already mentioned, for most variables, relative differences between Compustat and XBRL 10-K numbers are statistically significant. This means that the distributions of accounting variables in Compustat and 10-Ks are significantly different which may have a profound effect on the results of accounting studies that utilize the Compustat data set. This, in particular, is true for studies that utilize linear regression models since these models are very sensitive to changes in data distribution (Klein and Rossin 1999).

Our results agree with the results of Tallapally, Luehlfing, and Motha (2011); Tallapally, Luehlfing, and Motha (2012), that compared Cost of Goods Sold and Sales values between Compustat and XBRL 10-K filings for a sample of 27 companies, in a sense that the Cost of Goods Sold item tends to have more discrepancies and with larger magnitudes than the Sales item. Some of our results differ slightly from the results of Boritz and No (2013). For example, we do not find Total Liabilities to be

---

Stockholder's Equity.

significantly different across the data sets (both in number and magnitude), but we do find higher rate of discrepancies for the Receivables item (17% versus 7.3%). The differences are most likely due to us studying a much larger sample and have only one observation per company in our sample.

1.5.2 Discrepancy analysis by industry, size, and XBRL adoption phase

Company characteristics such as industry and size affect what is reported in 10-K statements and how it is reported. The same items will often have different contexts for different companies – e.g., things that comprise Cost of Goods Sold item of a retailing company are very different from components of Cost of Goods Sold item of a manufacturing company. Therefore, it is important to study the amount and magnitude of discrepancies with regard to different characteristics of companies. In our study, we look at three company characteristics: industry, size and XBRL adoption phase.

There are several ways to define company industry (e.g., using SIC or NAICS codes). In this study, we utilize Fama/French 12-industry classification system (Fama and French 1997) since it has been used extensively in accounting literature. In addition, there are only 12 industry groups (as opposed to 75 groups defined by 2-digit SIC codes) that facilitates better delivery and comprehension of results. Please refer to Table A.1 for the list and description of Fama/French 12 industries.

As has been mentioned previously, in 2009 the SEC adopted *Interactive Data to Improve Financial Reporting* final rules that mandated U.S. GAAP companies to report their financial statements with XBRL. The mandate was implemented in three phases. Phase one with the deadline in June 2009 required all large accelerated filers with worldwide public common equity float above $5 billion to adopt XBRL. Phase 2 with the deadline in June 2010 required all other large accelerated filers to adopt

XBRL. Finally, Phase 3 with the deadline in June 2011 required all U.S. GAAP companies to file with XBRL. We use the XBRL adoption phase as a characteristic of a company – XBRL adoption phase can be thought as one measure of company size.[30] In addition, we use company revenues as yet another, more refined, measure of size.

We provide discrepancy information by industry and XBRL adoption phase in Tables A.4 and A.5. Table A.4 reports the number of observations, percentage of discrepancy observations, and the median relative value of discrepancies for each Compustat variable and Fama/French industry.[31] In Table A.5, we report discrepancy statistics by industry and by XBRL adoption phase. These statistics include information about discrepancy counts as well as mean and median relative discrepancies.

Results in Tables A.4 and A.5 suggest that there is a difference in amounts and magnitude of discrepancies for different variables, industries, and sizes. To test it formally, we utilize analysis of deviance of discrepancy observations. Analysis of deviance is similar to analysis of variance (ANOVA) and analysis of covariance (ANCOVA), but can be used for generalized linear models (instead of linear regressions). It allows to test whether there are significant differences between the values of a response variable for different groups (categorical variables) while controlling for other continuous variables.

We utilize analysis of deviance in two ways. First, we test whether the propensity of an observation to be a discrepancy is related to company characteristics. In other words, we assume that an observation can have two states – either being a discrepancy or not – and test whether the value of that state depends on company

30. XBRL adoption phase has also been shown to reduce the number of errors (Du, Vasarhelyi, and Zheng 2013; Boritz and No 2013) and extensions (Debreceny et al. 2011) in XBRL filings. However, these effects are less relevant to the analysis in our study since we do not consider extension elements and attempt to remove all XBRL-related errors from the sample.

31. We report median values instead of means because the former are less biased statistics, i.e. they are not influenced by extreme values of discrepancies as much as the mean values.

characteristics. Since the response variable in this case is a state of an observation, this type of an analysis is an observation-level analysis. Second, we test whether company characteristics are related to the number of discrepancies for that company. For each company in the matched data set, we calculate the ratio of the number of discrepancy observations to the number of all matched observations for that company. The resulting ratio is a number between 0 and 1, with 0 indicating that there are no discrepancy observations for that company, and 1 indicating that all matched observations are discrepancies. Then, we test whether such company ratios depend on company characteristics. This kind of analysis is a company-level analysis since discrepancy ratios are defined at the company level. Because response variables' values are not unbounded real numbers – in the first case it is a binary variable, and in the second case it is a number between 0 and 1 – we cannot assume that they are normally distributed and utilize ANCOVA. However, we can use generalized linear models (logistic regressions) and analysis of deviance instead.

Analysis of deviance results are presented in Table A.6. In addition to studying how company characteristics are related to discrepancies, we also study how these characteristics are related to material discrepancies. Hence, Table A.6 shows results both when we use discrepancy notion for our response variables and when we use material discrepancy notion. In addition to company characteristics, in our analysis, we include the type of statement where the accounting variables are reported since Tables A.3 and A.4 suggest that the amount and magnitude of discrepancies also depend on the type of a variable.

Panel A of the table contains the results of deviance analyses that study the effects of revenue (size), industry, phase, variable statement, and various interactions between industry, phase, and variable statement. These results are fairly consistent and show that effects of all characteristics and their interactions are statistically significant except for the XBRL adoption phase in all cases, and revenue in one case. Also, the interaction terms are less significant (or not statistically significant

in some cases) when only material discrepancies are considered. Although the main effect of XBRL adoption phase is not statistically significant, its interactions with other characteristics are. The insignificance of the main effect can be explained by XBRL adoption phase being related to revenue since they both represent the size of a company.[32] To avoid this effect, we conduct another set of analyses with the XBRL adoption phase variable removed. The results are reported in Panel B of Table A.6, and show that all characteristics and their interactions are statistically significant, including revenue. Moreover, the significance of company revenue has increased drastically.

The interaction effect between industry and statement type categorical variables is significant according to Panel B in Table A.6. This indicates that the effects of independent variables may depend on the level of the other independent variables. To understand this relationship, we analyze so-called "simple main effects" – we study the effects of independent variables when the level of other categorical variables is fixed. In other words, we partition observations into groups by one categorical variable (either by industry or statement type), and then, for each group, run the analysis of deviance to study the effect of the other independent variables.

The results for simple main effects analysis are reported in Table A.7.[33] Panel A of this Table describes the effects of revenue and statement type for different industries. In particular, we observe that the amount of discrepancies significantly depends on the statement type except for the case of the telecommunications industry. Also, the effect of revenue size on discrepancy amounts is only significant for the financial industry. Panel B reports the effects of revenue size and industry type for each type

32. This was confirmed by the results of an additional ANOVA test (not reported in the study) that showed the differences between mean values of revenues across different XBRL adoption phases being 99.9% significant. Phase 1 companies had on average the largest revenues, and phase 3 companies the smallest.

33. Table A.7 contains results when all discrepancies (both material and not) are included in the dependent variables. We do not report separate results when only material discrepancies are considered since they are qualitatively the same.

of considered financial statements. The results indicate that industry significantly affects the amount of discrepancies for all financial statements. In addition, only for Balance Sheet variables, the amount of discrepancies depends on the revenue size.

Overall, the results show that the type of variable, industry, size, and XBRL adoption phase are related to the amount and magnitude of discrepancies. The results are slightly different from Boritz and No (2013) in a sense that we do find industry to be a significant factor (with significance levels of 99.9% in all cases) as related to the number of discrepancies.

1.6 Summary

In this study, we conduct the first large-scale comparison of Compustat and 10-K accounting numbers to study the amount and magnitude of data alterations in the Compustat North America Fundamentals Annual data set. Specifically, we compare 30 accounting items commonly used in the accounting literature for more than 5,000 of domestic U.S. GAAP companies for the period from October 1, 2011, to September 30, 2012. This large-scale comparison has become possible due to SEC-mandated XBRL financial reporting for all U.S. GAAP filers as of June 2011, since XBRL allows to extract data from XBRL 10-K filings in an automated fashion

We have developed and presented a methodology to compare Compustat numbers to XBRL numbers. This methodology addresses issues of data extraction, merging, variable mapping, difference calculation, automated error detection, and discrepancy analysis. It not only provides insights about data alterations in the Compustat data set, but also is a useful tool for accounting researchers to validate their data.

We find that Compustat significantly alters numbers reported in the 10-K filings. Specifically, we find that Compustat values of 17 (out of 30) variables significantly differ from values reported in 10-K filings. Variables that usually do not have significant number of discrepancies are variables that have fairly simple definitions (e.g.

Total Assets, Total Liabilities or Net Income). Variables that have more complex definitions (e.g. Cost of Goods Sold or Gross Profit) are more likely to differ across the data sets.

We also find that the type of statement where variable is reported and company characteristics such as industry, size, and XBRL adoption phase are related to the amount and magnitude of discrepancies. Specifically, we show that the amount of discrepancies is significantly affected by 1) type of financial statement (except for the telecommunications industry), 2) industry regardless of financial statement type, and 3) revenue size for the financial industry or balance sheet items.

Our findings suggest that data alterations in Compustat are non-trivial, and may potentially influence results of accounting studies that utilize the Compustat data set. Compustat itself argues that its data standardization practices improve data comparability across companies. More research and discussion are needed on this matter.

Chapter 2

Does Compustat data standardization improve bankruptcy prediction models?

2.1 Introduction

In Chapter 1 of this dissertation, we show that there are significant differences between accounting numbers in Compustat and the original 10-K reports. Most of these differences are caused by data standardization practices implemented in Compustat, i.e., accounting numbers are adjusted to match Compustat standard definitions of accounting concepts. According to S&P Capital IQ's website, "*[s]tandardized data ensures that you have consistent and comparable data across companies, industries and business cycles, and offers a solid foundation for your rigorous analysis.*"[1] However, the benefits of Compustat data standardization have been rarely questioned in accounting research community. o the best of our knowledge, there has been no academic research that measures the impact of Compustat's data standardization.

Stone (1968) is one of the few and most likely the first study to warn researchers about potential problems with Compustat's standardized reporting. Specifically, Stone (1968) argues that data standardization "*creates the appearance of absolute uniformity*" while it is "*in reality, a uniformity of classification only, and not a uniformity of accounting methods.*" This concern is also echoed by Financial Accounting Standards Board (FASB) in Statement of Financial Accounting Concepts No. 2 that warns against using standardized charts of accounts – "*[t]hat kind of uniformity may*

1. http://www.compustat.com, as of January 13, 2014.

even adversely affect comparability of information if it conceals real differences between enterprises.".[2] In addition, FASB cautions against "over-improving" data comparability since "*[i]mproving comparability may destroy or weaken relevance or reliability if, to secure comparability between two measures, one of them has to be obtained by a method yielding less relevant or less reliable information.*"[3] In fact, FASB's Generally Accepted Accounting Principles (GAAP) provide firms with a certain amount of discretion when reporting accounting numbers.

In this Chapter, we examine whether data standardization practices implemented in Compustat result in enhanced data, and as a consequence, improved empirical models that rely on the data provided in Compustat. Specifically, we study whether two popular models of financial distress, Altman's Z Score (Altman 1968) and Ohlson's O Score (Ohlson 1980), are enhanced by standardized numbers provided in Compustat as compared to the original numbers of 10-K reports.

We choose Atlman's and Ohlson's bankruptcy prediction models for a number of reasons. Firstly, bankruptcy prediction is an important problem for market participants that is directly related to their decisions. Bankruptcy prediction models are not only used to predict bankruptcies, but also to measure the overall financial health of firms. Secondly, Altman's and Ohlson's models are often used in both contemporary research and practice despite them being more than 30 years old. Thirdly, both models rely on fundamental financial accounting ratios and numbers (e.g., measures of liquidity, leverage, profitability, debt financing, etc.) that are important characteristics of business performance. Fourthly, the models are easy to replicate with little ambiguity with respect to sample and variable selection. Fifthly, in order to be accurate, these models require a significant level of comparability and consistency of accounting numbers across firms and time. The objective of data standardization is to enhance comparability and consistency, and thus improve the accuracy of the

2. Statement of Financial Accounting Concepts No. 2, ¶116.

3. See footnote 2.

models. Hence these models are valid test subjects to study the consequences of data standardization. Finally, it is possible to statistically infer whether either of these models perform better if Compustat's standardized data is used instead of the original 10-K data.

We find that Compustat data standardization not only yields no improvements for these models, but also has a negative impact on Altman's model predictive ability. Specifically we show that using Compustat standardized numbers instead of original 10-K numbers yields 1) significantly different outputs of both Altman's and Ohlson's models, 2) significantly worse predictive accuracy of Altman's model (up to 8.56%), and 3) no significant improvement for Ohlson's model.

In this essay, we first test whether the choice of input, original 10-K or standardized Compustat data, results in significant differences in models' outputs. Then, we replicate Ohlson's and Altman's studies using recent cases of bankruptcy events, and compare how well the models discriminate between bankrupt and non-bankrupt observations when different data sources are used. Finally, we study how the choice of data affects the predictive ability of the cross-validated models, i.e., models that are estimated and tested on different observation sets. In addition, we perform a number of robustness checks to validate our findings.

Our contributions to the literature are as follows. Firstly, we first empirically study the effects of data standardization in Compustat, and empirically show that numbers in the original 10-K reports are at least as good, or better measures of financial health of a firm as standardized numbers in Compustat are. An important implication of this finding is that accounting researchers need to re-evaluate the benefits of using standardized data, and perhaps consider alternative sources of data such as SEC's eXtensible Business Reporting Language (XBRL) data. Secondly, we contribute to the literature that studies discrepancies between Compustat and other data sets (including 10-K reports) by demonstrating that significant differences across data sets may lead to significant differences in performances of empirical models derived

from these data sets. Thirdly, we show that bankruptcy prediction models that use accounting data may be enhanced by using 10-K data instead of Compustat's data.

This Chapter is organized as follows. In Section 2.2, we describe data and methodology we use to conduct the study. In Section 2.3, we present results of our multi-step analysis, and perform a number of robustness checks to validate the main results. Section 2.4 summarizes the Chapter. Brief overviews of Altman's 1968 and Ohlson's 1980 models as well as supplemental tables are provided in Appendix B.

2.2 Data and methodology

2.2.1 Model comparison methodology

To study the effects of Compustat's data standardization on Altman's 1968 and Ohlson's 1980 models, we first test if using Compustat data results in different scores and models as compared to the original 10-K numbers.[4] We create two matched data samples - one based on Compustat numbers, and the other based on the original 10-K numbers. We use these data samples to perform three comparison procedures for each model (see Figure 2.1). The procedures are meant to test for any differences resulting from using Compustat standardized versus original 10-K data with regards to three dimensions: output, explanatory power, and predictive ability.

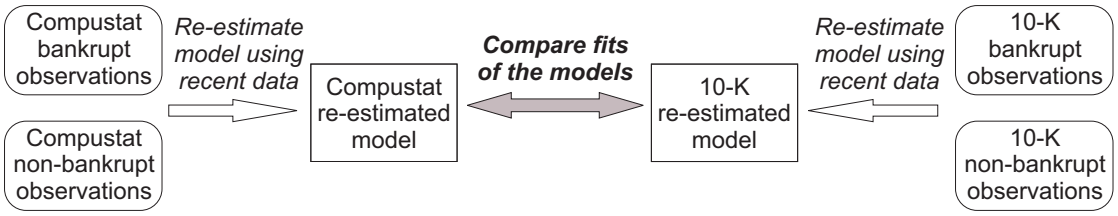
The first procedure is a simple test of whether standardized Compustat data yield significantly different model scores as compared to 10-K data. We apply the original models to our matched sample of Compustat and 10-K numbers to calculate two sets of scores for each model. We test these two sets of scores for any statistically significant difference.

The second procedure is meant to test the explanatory power of models constructed using Compustat and 10-K numbers. For this test, we use the recent cases

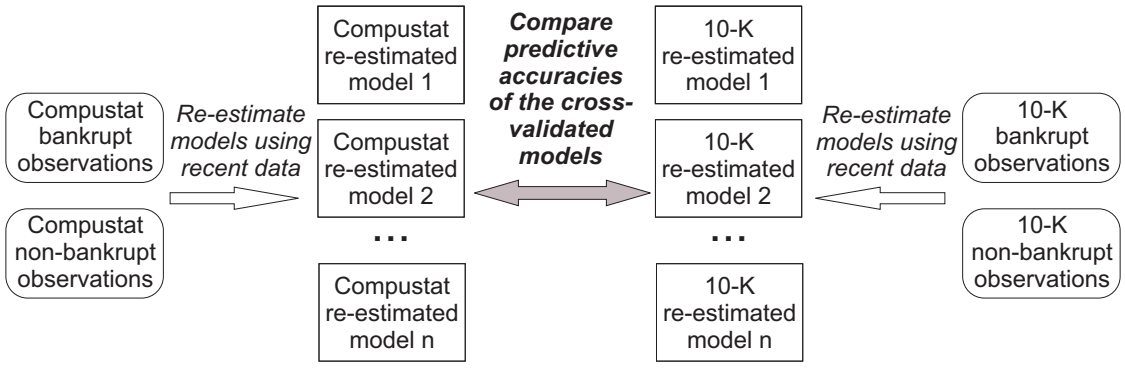
4. Brief summaries of Altman's and Ohlson's models are provided in the Appendix.

Figure 2.1: Comparison methodology.

(a) Test for differences in output.



(b) Test for difference in explanatory power.



(c) Test for difference in predictive ability.

of bankruptcies (from 2009 - 2013) to replicate the studies of Altman (Altman 1968) and Ohlson (Ohlson 1980) by re-estimating the models with recent data. For the re-estimated Altman and Ohlson models, we then compare accuracies and fit of the model built using original 10-K numbers to the model built using Compustat standardized data. We use the same data sample that was used to re-estimate a model to calculate the accuracy and fit of the model.

Finally, we compare the predictive accuracy of the models built using Compustat standardized numbers and 10-K original numbers. The procedure described in the previous paragraph cannot be used to draw inference about predictive abilities of the models, since the same data is used to both estimate and test the models (Joy and Tollefson 1975; Mensah 1984). We use a popular method of comparing predictive accuracy of classification models, k-fold stratified cross validated paired t test (see Dietterich (1998) for more details), to compare the predictive ability of the models. This test uses both resampling and cross-validation methods to create several instances of the models whose predictive accuracy rates are then compared by a t test. We also impose stratified cross-validation since observations used to estimate the Ohlson model are highly unbalanced (i.e., the number of bankrupt observations is much smaller than the number of non-bankrupt ones).

2.2.2 Data extraction and sample selection

We use the methodology developed in Chapter 1 of this dissertation to retrieve the original accounting numbers from 10-K reports and match them to the appropriate Compustat Fundamentals Annual numbers.[5] This methodology leverages eXtensible Business Reporting Language (XBRL), a formal language of communicating business information that was mandated by the U.S. Securities and Exchange Commission

5. We use Compustat Fundamentals Annual data as opposed to other versions of Compustat data, such as Compustat Point-in-Time, since this dataset is used most frequently in research, including cases when Altman's and Ohlson's models are utilized.

(SEC) to be used in preparing quarterly and annual reports by all U.S. GAAP companies as of June 2011.

The usage of XBRL 10-K reports limits our samples to the years of 2009 - 2013. Furthermore, not all accounting numbers can be directly extracted from XBRL reports due to various XBRL-related reasons.[6] In addition, Compustat data set does not include all public companies and may have missing values needed to calculate financial ratios for Altman's and Ohlson's models. Also, Altman's model requires market data that we extract from CRSP database. Although, Ohlson's model does not require market data, Ohlson (1980) considers only companies that are traded on stock exchanges or over-the-counter markets. Hence, we exclude company-year observations that could not be matched to CRSP market data. Finally, following Ohlson (1980), we do not include companies classified as utilities, transportation companies, and financial service companies.[7]

Since we study the effects of Compustat's data standardization, it is important to remove other causes of number discrepancies between Compustat and 10-K data. The methodology developed in Chapter 1 allows us to find XBRL data errors, Compustat data errors, and reconcile discrepancies between 10-K and Compustat data, i.e., identify adjustments made by Compustat to obtain standardized data. We remove all erroneous data (both XBRL- and Compustat-related) and data that cannot be reconciled from our samples. As a consequence, any discrepancies in our data samples are most likely due to data standardization. The only exception is the sample of bankrupt observations described in the following section. We used the original, not restated 10-K forms to extract 10-K data related to bankrupt observations. For some of these observations, Compustat has updated the numbers with the restated ones. Hence, Compustat numbers would be not only standardized in those cases, but also

6. See Section 1.4 for more details.

7. Altman (1968) considers only manufacturing companies. We repeat our analysis using manufacturing companies alone as a robustness check.

restated. We expect this to give a slight advantage to Compustat data over 10-K data since the restated numbers should be more accurate.

Overall, our sample contains 5,015 observations with accounting ratios needed for Altman's model, and 3,449 observations that can be used in Ohlson's model. The number of observations for Ohlson's model is smaller since Ohlson's model requires values for Net Income for both the current and the previous fiscal years. Tables B.1 and B.2 report descriptive statistics for these samples.

Compustat database is biased toward large companies. As a result the average Total Assets in the Altman sample are around 4 billion U.S. dollars and the median Total Assets are 706 million U.S. dollars. Companies in the Ohlson's sample are slightly larger with the mean assets of about 4.7B$ and the median assets of approximately 787M$.

For both models, there is a significant difference in Total Liabilities between 10-K and Compustat data. This appears to be unexpected given the findings of prior literature (Boritz and No 2013; Chychyla and Kogan 2013). However, after manually analyzing these differences, we found that they are caused by us using non-restated 10-K numbers for bankrupt observations that are restated in Compustat Fundamentals Annual database.

For Altman's sample, there is a significant difference in Retained Earnings and Earnings before Interest and Tax between Compustat and 10-K data. For Ohlson's sample, there is a significant difference in Operating Income before Depreciation between the two data sets. We discuss how those differences impact the accounting ratios scores of the bankruptcy models in the results section.

Selecting bankrupt and non-bankrupt observations

To assess the explanatory power and predictive ability of the models, we need to re-estimate the original models with company-year observations that are known to be either bankrupt or non-bankrupt.

The reason behind this is two-fold. Firstly, the original models are more than 30 years old and may require calibration. Secondly and more importantly, the original models have been estimated on different data than ours – Altman (1968) used data from Moody's Manual, and Ohlson (1980) used a combination of 10-K reports and Compustat data.[8] Since we want to test the impact of data sources on the models, we have to be consistent with respect to the data we use to estimate and train the models. E.g., in the case of Compustat, we use Compustat data to estimate the model, and Compustat data to test the model.[9]

We adopt Ohlson's (1980) definition of bankruptcy, that of any indication of bankruptcy proceedings. We also use his measure of time to bankruptcy which is the time from the date a pre-bankruptcy 10-K report was filed to the date of the bankruptcy event.

Public companies are required to notify shareholders about the bankruptcy or receivership events by filing an 8-K form that includes a special Item 1.03, Bankruptcy or Receivership, that describes the event. To create a sample of bankrupt observations, we first identify all 8-K forms that include bankruptcy or receivership item that were filed after January 1, 2009. We then manually read the selected 8-K forms to confirm a bankruptcy event, establish the type of the bankruptcy (e.g., Chapter 11, 7, etc.), and record the date of the bankruptcy event. For all bankruptcy events, we identify related 10-K forms with filing dates preceding the bankruptcy event. We consider only pre-bankruptcy original 10-K reports filed after January 1, 2009 since, as mentioned previously, XBRL 10-K reports we use to extract 10-K data became available in 2009, and it is important to avoid any out-of-the-sample year-specific biases that may affect financial positions of companies. In addition, as already mentioned, we exclude certain industries and require a presence of matching

8. According to Collins and O'Connor (1978), data definitions in Moody's Manual differ from the ones of Compustat.

9. We do compare the predictive accuracies of the original (not re-estimated) models on our sample as robustness check.

data in both Compustat and CRSP data sets. We extract accounting numbers from the 10-K reports meeting these criteria (either manually, or if possible using XBRL), and use them to create 10-K version of the bankrupt observations. The matching Compustat company-year observations provide standardized data for those bankrupt observations.

We limit our analysis only to bankrupt observations whose time period from the release of the 10-K form to the bankruptcy event is not greater than 2 years (730 days). This would make our study consistent with both Altman (1968) and Ohlson (1980). Our final sample of bankrupt observations consists of 146 company-year observations with 85 10-K forms released within one year prior to bankruptcy events, and 61 10-Ks released between one and two years prior to bankruptcy events. Chapter 11 filings amount to 128 bankruptcy observations, Chapter 7 filings amount to 9 bankruptcy observations, and other types of bankruptcy filings amount to the remaining 9 observations. It is worth noting that 45 companies in our bankrupt sample are associated with two observations, i.e., we include two 10-Ks of the same company in our bankrupt sample if the time lag between the earlier 10-K and the bankruptcy event is not greater than 2 years. This bankrupt sample is used in re-estimation of both Altman's and Ohlson's models. The descriptive statistics for the bankrupt sample are provided in Table B.3.

Not surprisingly, on average bankrupt company-year observations have smaller assets compared to the whole samples of matched observations – the mean Total Assets equals to 1043.5M$, and the median to 363.828M$. Interestingly, the average Total Liabilities are of the same size as the average Total Assets – 1043.5M$. In contrast, the average Total Liabilities for the whole matched samples are a little above half of the Total Assets (see Tables B.1 and B.2). This effect is exploited in Ohlson's model by one of the model's ratios – Total Liabilities over Total Assets.

Similarly to the whole matched samples, there are significant differences in Total Liabilities, Retained Earnings, Operating Income after Depreciation, and Earnings

before Interest and Taxes between 10-K and Compustat data. In addition, there is a significant difference in Total Assets. Similarly to the difference in Total Liabilities, the difference in Total Assets is caused by us using the original, non-restated 10-K numbers for bankrupt observations while Compustat Fundamentals Annual dataset comprises of restated standardized numbers. As mentioned previously, we believe that restated numbers should give a slight edge to Compustat numbers over 10-K numbers due to increased data reliability.

We also create a sample of "non-bankrupt" observations. We consider a company-year observation to be non-bankrupt if the company 1) was not identified as bankrupt in the previous step, and 2) filed a 10-K form on or after June 1, 2012. The latter requirement is meant to provide additional assurance that the company has not incurred bankruptcy-related event. While we use the same sample of bankrupt observations for both Altman's and Ohlson's model, we use different samples of non-bankrupt firms following approaches used in Altman (1968) and Ohlson (1980). We explain the sample selection procedures of non-bankrupt companies for each model separately below.

Non-bankrupt sample used to re-estimate Altman's 1968 model

The Altman's 1968 Z Score model uses Multiple Discriminant Analysis (MDA) to create a linear discriminant model that tries to best separate two group of observations, bankrupt and non-bankrupt. The linear discriminant is based on the financial ratios that Altman empirically found to yield the best predictions of bankruptcy. Altman (1968) matches all bankrupt company-year observations with similar in terms of industry and size non-bankrupt observations. Altman (1968) uses the matched sample of bankrupt and non-bankrupt observations to estimate his MDA model.

We follow Altman (1968) by assigning each bankrupt company-year observation in our sample to a non-bankrupt observation on a stratified random basis with respect

to company size, industry, and reporting fiscal year.[10] This gives us a sample of 292 observations that comprises of 146 bankrupt observations, and matching 146 non-bankrupt observations. The descriptive statistics of non-bankrupt observations are reported in Table B.4.

Our non-bankrupt sample consists of larger companies than the non-bankrupt one, but not as large as the ones in the sample with all matched observations (see Table B.1). This is the result of our random stratified matching procedure and a larger likelihood for smaller companies to experience a bankruptcy event. In fact, the non-bankrupt observations used in Altman (1968) are also larger in terms of company size than the respective bankrupt observations.

There is only one variable with statistically significant differences between Compustat and 10-K data – Retained earnings. Earnings before Interest and Tax is not statistically significantly different across two sources of data for this sample (as opposed to us finding significant differences in all other considered samples).

Non-bankrupt sample used to re-estimate Ohlson's 1980 model

Unlike Altman's Z Score, Ohlson's 1980 model uses logistic regression to estimate a model for predicting financial distress. The logistic model has slightly weaker mathematical assumptions than MDA model making it preferable to MDA in many applications. The output of Ohlson's logistic model are O scores. Unlike Altman's Z scores, O scores can be converted to probabilities of firms experiencing financial distress within a certain period of time. In other words, O scores have precise probabilistic meanings that are easy to interpret.

Ohlson (1980) uses the whole population of non-bankrupt companies as opposed to a matched sample. Although this is a more realistic approach, it also results in a

10. We performed matching in the following way. First, for every company-year observation, we identified all non-bankrupt company-year observations in the same industry and the same fiscal year as the bankrupt observation. Then we chose five observations whose Total Assets were the closest to one of the bankrupt observation. Finally, we randomly choose an observation out of those five.

highly unbalanced data with number of non-bankrupt observations being significantly greater than the number of bankrupt observations. Unbalanced data is known to have a significant negative impact on performance of statistical models estimated using such data (He and Garcia 2009). A trivial model that predicts everything to be non-bankrupt may have as good accuracy as a statistical model based on an unbalanced data. In fact, Ohlson's 1980 model that predicts bankruptcy within one year achieves accuracy of 96.12%. Olson compares this accuracy to the 91.15% accuracy achieved by a trivial model (Ohlson 1980, p. 120). However, 91.15% accuracy of the trivial model is a typo – the correct number is 95.15% (calculated as 2,058/(105+2,058)×100%). Hence, Ohlson's model resulted in less than 1% accuracy improvement over the trivial one.

We follow Ohlson (1980) by first identifying all observations in our matched sample that can be classified as non-bankrupt according to the definition in §2.2.2. We then randomly keep only one observation per company. Our non-bankrupt sample consists of 2,525 observations, and the descriptive statistics of this sample is reported in Table B.5.

Our non-bankrupt sample for Ohlson's model consists of larger-size companies than the ones in the bankrupt sample, but not as large as in the whole matched sample. The mean and median Total Assets are approximately 2.9B$ and 360M$. As in the case of the whole matched sample, we find significant differences in Total Liabilities and Operating Income after Depreciation between 10-K and Compustat data sets.

2.3 Results

2.3.1 Differences in scores

First, we test whether using Compustat numbers instead of 10-K numbers would yield any significant differences in output scores of the original Altman's and Ohlson's

model. The descriptive statistics for the samples used for these tests, and the outcomes of these tests are given in Tables B.1 (for Altman's model) and B.2 (for Ohlson's model).

Altman's model

For the Altman's original 1968 model, we find that the choice of data source (10-K or Compustat) significantly impacts the scores produced by the model (with 99% confidence level). On average, Altman's Z scores produced by the 10-K model are larger by 0.009. Altman's Z score is a continuous number that is hard to interpret. For that reason, Altman (1968) defines three discrimination zones by creating two cutoff score values, 1.81 and 2.99. Scores below 1.81 indicate bankrupt zone, between 1.81 and 2.99 indicate "gray" zone, and above 2.99 indicate non-bankrupt zone. We test whether there are any differences in Altman's discrimination zones between the two sources of data. We code each score as either 1, 2, or 3 based on its value with 1 indicating bankrupt zone, 2 indicating "gray" zone, and 3 indicating non-bankrupt zone. We then use Wilcoxon signed-rank test to compare the codes between the two sources of data.[11] We find that there are significant differences in discrimination zones yielded by 10-K and Compustat data (see Table B.1). Note that the test involving discrimination zones is much stronger than the test involving Z scores since different Z scores may result in the same discrimination zones.

Out of five ratios used in the original Altman's model, three, X_2 (Retained Earnings/Total Assets), X_3 (Earnings before Interest and Tax/Total Assets), and X_4 (Market Value of Equity/Total Liabilities), are significantly different across 10-K and Compustat datasets. Difference in X_4 is caused by difference in Total Liabilities due to us using original 10-K numbers for bankrupt observations that are not restated as

11. In this case, Wilcoxon's signed-rank test is more appropriate than the regular t test. Wilcoxon's signed-rank test does not require variables to be numerical, but only ordinal (i.e., given two different zones of discrimination, the test only requires to know what zone is "greater", but not *how much* one zone is "greater" than the other).

explained above. Removing all observations in our sample with non-zero differences in Total Liabilities between 10-K and Compustat datasets does not change the results.

Differences in X_2 and X_3 are due to significant differences in Retained Earnings and Earnings before Interest and Tax (EBIT) across the datasets. In fact, substituting Compustat values of Retained Earnings and EBIT with 10-K values of these variables would result in no significant difference between Z scores. However, substituting only one of the Compustat variables with the matching 10-K variable would not make the difference in Z scores and discrimination zones statistically insignificant.

Ohlson's model

We also find significant differences between O scores of Ohlson's 1980 model generated using 10-K and Compustat data. Ohlson (1980) constructs three logistic models to predict bankruptcy: 1) within one year of the 10-K filing date, 2) in the second year after the 10-K filing date, and 3) within two years of the 10-K filing date. In Table B.2 we label these models "Model 1", "Model 2", and "Model 3", respectively. We calculate O scores for all three models, and compare them. "Raw" O scores do not carry much meaning; hence we convert them first to probabilities of companies experiencing bankruptcy events in the future, and then to predicted classes with 1 indicating predicted future bankruptcy, and 0 otherwise. We compare the resulting bankruptcy classifications. For all models, we find significant differences in O scores and predicted classes between 10-K and Compustat data.

Differences in Total Liabilities and Operating Income after Depreciation (OIADP) between 10-K and Compustat data yield significant differences in the respective versions of TLTA and FUTL variables in Ohlson's model (see Table B.2). As in the case of Altman's model, removing all observations from the sample that result in non-zero differences in Total Liabilities does not change the results. However, replacing Compustat values of OIADP with the matching 10-K values yields no significant differences in O scores and predicted classes. This means that differences in OIADP

values drive differences in output of Ohlson's model between 10-K and Compustat datasets.

2.3.2 Differences in explanatory power

As detailed in the previous section, for both original Altman's and Ohlson's models, we find statistically significant differences in models' output if Compustat standardized data is used instead of the original 10-K data. These results suggest that the effects of Compustat data standardization are non-trivial, and cannot be ignored. We examine whether these effects are positive or negative with regards to bankruptcy prediction in this and the next section.

In this section, we compare the explanatory power of the Altman's and Ohlson's models derived using 10-K and Compustat data. Both Altman's and Ohlson's models were derived from different data sets than ours more than 30 years ago. Hence, it is essential to re-estimate both models using our data. We use the recent cases of bankruptcies to re-estimate the model and compare their accuracy and overall fit. In other words, we compare how much variance with respect to bankruptcy predictions are explained by a model based on standardized Compustat numbers as opposed to a model based on the original 10-K numbers.

For both Altman's and Ohlson's model, we estimate three types of predictive models – one that predicts bankruptcy in the first year after the release of the 10-K form, one that predicts bankruptcy in the second year after the release of the 10-K form, and one that predicts bankruptcy within two years after the 10-K form was released. To estimate these three types of models, we accordingly create three groups of bankrupt observations based on the amount of time between the release of 10-K form and the bankruptcy event. Then we merge those three groups of bankrupt observations with the non-bankrupt ones to obtain the samples used to re-estimate the models. For Altman's model, observations in each bankrupt group are merged with the same number of matched non-bankrupt observations. For Ohlson's model,

observations in each bankrupt group are merged with all non-bankrupt observations (i.e., non-bankrupt observations are the same for all bankrupt groups).

Accuracy and fit metrics of the re-estimated models are reported in Tables 2.1 and 2.2 for Altman's and Ohlson's models respectively. In those tables, *Accuracy* is the overall accuracy of a model, and *Trivial Accuracy* is the accuracy of a model that predicts everything to be non-bankrupt, i.e., it is the percentage of non-bankrupt observations in a sample. Trivial model accuracies are used as baselines for the predictive abilities of re-estimated models. We test whether a model accuracy is significantly different from the accuracy of the trivial model by performing a one-sided binomial test with null hypothesis being that accuracies are the same, and alternative being that model's accuracy is greater than the trivial one. P-values of those tests are reported as *P-value (Acc. > Trivial)*. We also measure *Cohen's Kappa* between the predicted observation classes (i.e., bankrupt or non-bankrupt) and the actual ones. Cohen's Kappa is a measure of agreement between two classifications. Its values are between 0 and 1, with 0 value meaning total disagreement and 1 meaning total agreement. Values of Cohen's Kappa are adjusted for both sample class bias (i.e., number of bankrupt versus non-bankrupt observations in a sample) and model class bias (i.e., the overall propensity of a model to assign a particular class to random observation). For Altman's model we report the P-value of the F likelihood ratio test as $\mathbb{P}(> F)$. This test indicates the significance of the discriminatory power of Altman's model. Similarly, for Ohlson's model we report P-value of the χ^2 likelihood ratio test as $\mathbb{P}(> \chi^2)$. We also report the value of pseudo R^2 for the Ohlson's model.

Altman's model

We find that all re-estimated Altman's models yield significantly better accuracies on the fitted samples than the respective trivial models that predict everything to be non-bankrupt. For all three types of models that predict bankrupt events in the first, second, and first two years after the release of 10-K statements, 10-K data based

Metrics	*10-K models*			*Compustat models*		
	first year	*second year*	*within two years*	*first year*	*second year*	*within two years*
Accuracy	84.71%	75.41%	78.42%	79.41%	72.95%	76.03%
Trivial Accuracy	50.00%	50.00%	50.00%	50.00%	50.00%	50.00%
P-value (Acc. > Trivial)	0.000	0.000	0.000	0.000	0.000	0.000
Cohen's Kappa	0.694	0.508	0.568	0.588	0.459	0.521
$\mathbb{P}(>\mathrm{F})$	0.000	0.000	0.000	0.000	0.000	0.000

Table 2.1: Accuracy and fit statistics of fitted Altman's models.

models have higher accuracies than similar Compustat data based models. These results also hold with respect to Cohen's Kappa measure.

The largest difference between 10-K and Compustat models is in the case of bankruptcy prediction within the first year of 10-K release – 10-K model achieved 84.71% accuracy that is 5.29% higher than 79.11% accuracy of the respective Compustat model. Both 10-K and Compustat are less accurate in the case of bankruptcy prediction in the second year after the 10-K release, their respective accuracies are 75.41% and 72.95%. This is not surprising given longer time horizon. Accuracies of the models that try to predict bankruptcy within two years are somewhere in the middle.

The differences in the accuracies of fitted models are mainly driven by differences in values of the Earnings before Interest and Tax (EBIT) variable. Replacing Compustat values of EBIT with the appropriate 10-K values would make Compustat models to be on par with 10-K models. This result may seem to suggest that standardized EBIT variable may negatively impact the prediction performance of some statistical models. However, it would be a mistake to draw any substantive conclusions based on the outcomes of fitted models alone since these models are estimated on the same sample they are tested on. For this reason, we test the predictive ability of the cross-validated models in the following section.

Metrics	*10-K models*			*Compustat models*		
	first year	*second year*	*within two years*	*first year*	*second year*	*within two years*
Accuracy	96.70%	97.64%	94.61%	96.70%	97.64%	94.57%
Trivial Accuracy	96.74%	97.64%	94.53%	96.74%	97.64%	94.53%
P-value (Acc. > Trivial)	0.572	0.533	0.454	0.572	0.533	0.488
Cohen's Kappa	0.097	0.000	0.121	0.097	0.000	0.120
$\mathbb{P}(> \chi^2)$	0.000	0.000	0.000	0.000	0.000	0.000
Pseudo R^2	0.264	0.134	0.235	0.258	0.146	0.240

Table 2.2: Accuracy and fit statistics of fitted Ohlson's models.

Ohlson's model

For Ohlson's fitted models we find that 1) the accuracies of Compustat and 10-K models are very similar, and that 2) the accuracies of all models are not statistically different from the respective trivial models that predict all observations to be non-bankrupt. The latter finding is indicated by large P-values of the one-sided binomial tests reported as *P-value (Acc. > Trivial)* in Table 2.2. This is the result of using very unbalanced samples to estimate logistic regression models – in the best scenario the percentage of bankrupt observations is only 5.47%. This fact is also captured by low values of Cohen's Kappa despite the high values of model accuracies – Cohen's measure of agreement is adjusted by class biases both in a sample and a model. In fact, models that predict bankruptcy in the second year after the release of 10-K are identical to the trivial models. Note that unlike in the case of Altman's model, the overall best accuracy is achieved when predicting bankruptcy in the second year after the 10-K release since the ratio of non-bankrupt to bankrupt observations is the largest in this case – 0.9764; for Altman's model, this ratio is always 0.5 regardless of the time horizon since we keep only those non-bankrupt observations that match the bankrupt ones.

As already discussed in Section 2.2.2, unbalanced data samples tend to result in weak statistical models (2009), and the fitted models estimated in Ohlson's 1980

study yielded marginally better accuracies than their respective non-trivial models. Interestingly, model that predicts bankruptcy in the second year (Model 2) in Ohlson (1980) is not statistically more accurate than the trivial model as measured by the binomial test. In our case such 10-K and Compustat models are equivalent to their trivial models.

2.3.3 Differences in predictive ability

In the previous section, we tested whether Compustat standardized numbers increase the explanatory power of Altman's and Ohlson's models. However, these tests do no tell us whether Compustat's data standardization has any effects on predictive ability of the models since we used the same sample to estimate the models and make predictions. To assess the predictive ability of the re-estimated models we utilize a common cross-validation technique, k-fold cross-validated paired t test (see Dietterich (1998) for details). Specifically, we perform a 10-fold cross-validated comparison 10 times and compare the differences in predictive accuracies using a t test. In addition to t test, we utilize a more powerful Wilcoxon signed-rank test as a robustness check. In addition, we employ the stratified sampling technique to randomly select folds for cross-validation. This allows us to more accurately compare models when the sample is unbalanced which is the case for the Ohlson's model (only 5.4% of the sample are bankrupt observations).

The results of the cross-validated comparison of predicative accuracy of both Altman's and Ohlson's model are summarized in Table 2.3. For both Altman's and Ohlson's models, Table 2.3 reports mean prediction accuracies of cross-validated models that predict bankruptcy in the first year, second year, and within two years of the release of 10-K form. It also reports prediction accuracies of trivial models, models that predict everything to be non-bankrupt. As mentioned previously, trivial models are very accurate for samples used to estimate Ohlson's model since those samples are highly unbalanced. In addition, Table 2.3 reports differences in accuracy means

Mean prediction accuracy	*Altman's model*			*Ohlson's model*		
	first year	*second year*	*within two years*	*first year*	*second year*	*within two years*
10-K	80.87%	75.01%	77.10%	96.82%	97.66%	94.70%
Compustat	72.31%	71.08%	71.57%	96.83%	97.66%	94.71%
Trivial	50.00%	50.00%	50.00%	96.74%	97.64%	94.53%
Difference	8.56%	3.93%	5.53%	−0.01 %	0.00%	−0.01 %
P-value	0.000	0.001	0.000	0.158	0.566	0.320

Table 2.3: Predictive accuracies of cross-validated Altman's and Ohlson's models.

of 10-K and Compustat based models, and the p-values of the mean difference t tests. The p-values of similar Wilcoxon signed-rank tests were of the same magnitudes as the ones reported in Table 2.3.

Altman's model

In all cases, we find that predictive accuracies of Altman's 10-K data based models significantly outperform similar models based on standardized Compustat data. The largest gap between the models' mean predictive accuracies is in the case of bankruptcy prediction within one year – the models based on the original 10-K data are on average 80.87% accurate in their predictions, while the models that use Compustat standardized data are on average 72.31% accurate. The difference of more than 8% is economically non-trivial suggesting that standardized data may not only fail to enhance prediction models, but actually worsen them.

The models that predict bankruptcy in the second year after the release of 10-K form are less accurate than the models that predict bankruptcy within one year for obvious reasons. Mean accuracies of the models that predict bankruptcy within two years are somewhere in between. Yet, in all cases 10-K data based models are superior. This seems to agree with our previous finding that 10-K data yield higher explanatory power of Altman's model than standardized Compustat data.

We empirically find that replacing Compustat values of Earnings before Interest

and Tax (EBIT) with 10-K values of the same variable would enhance Compustat models, and make all the differences in mean predictive accuracies of cross-validated models insignificant. Hence, in case of predictive accuracy of Altman's model, EBIT is the main driver of models' performance differences.

Ohlson's model

Results for Ohlson's model are very similar to the ones from the previous section. Specifically, we find no statistically significant differences between 10-K and Compustat models. Although the accuracies of those models are fairly high (94.7% and above), they are not better than accuracies of the trivial models that predict everything to be non-bankrupt. As discussed previously, such poor performance is probably due to highly unbalanced data set. For this reason, in the next section we perform a robustness check with the Ohlson's model being estimated on the matched sample we use to re-estimate Altman's model.

2.3.4 Drivers of differences

Overall, our findings indicate that Compustat data standardization does not improve bankruptcy prediction models. Moreover, it has a significantly negative impact on the predictive accuracy of Altman's model. We also find that outputs of the models differ significantly if Compustat data is used instead of the original 10-K data. In this section we report common adjustments made by Compustat to variables that drive differences between 10-K and Compustat data.

In Section 2.3.1 we find that differences in Retained Earnings and Earnings Before Interest and Tax (EBIT) cause the differences in output of Altman's model, and differences in Operating Income after Depreciation (OIADP) cause the differences in output of Ohlson's model. Significant differences in predictive accuracies of Altman's model are caused by differences in EBIT variable.

In case of Retained Earnings, among 4,056 discrepancies, around 85% are caused by inclusion of accumulated other comprehensive income (loss), 4% are caused by inclusion of equity attributable to non-controlling interest, and 3% are caused by including foreign currency translation adjustment in Compustat data.

Differences in OIADP are driven by many items: out of all 1,980 discrepancies, around 9% are related to restructuring charges, 9% to impairment of assets and goodwill, and 8% to business acquisitions, 2.2% to asset disposition, etc. A few other studies find similar items to drive differences between GAAP and pro forma earnings (e.g., Bradshaw and Sloan 2002, p. 44; Bowen, Davis, and Matsumoto 2005, p. 1021; Elliott 2006, p.121).

Differences in EBIT are similar to the ones in OIADP. In addition, items like other operating income, severance cost, advertising expense, litigation costs, etc. contribute to the differences in EBIT. Interestingly, Compustat does not fully include non-operating income into its definition of EBIT. According to Compustat's manual, EBIT is equal to "*Sales - Net (SALE) minus Cost of Goods Sold (COGS) minus Selling, General & Administrative Expense (XSGA) minus Depreciation/Amortization (DP)*". EBIT is a non-GAAP measure, but it is often used in literature and practice. We used its common definition as a sum of Operating Income and Non-operating Income (e.g., see Bodie, Kane, and Marcus (2008)) for the 10-K version of the variable.[12] As a robustness check, we apply similar definition to Compustat's data and re-evaluate the predictive ability of Compustat based models; the results do not change (see Section 2.3.5).

12. We also exclude Interest Expense from Non-operating Income if it was a part of it.

Mean prediction accuracy	*Altman's model*			*Ohlson's model*		
	first year	*second year*	*within two years*	*first year*	*second year*	*within two years*
10-K	76.47%	74.59%	75.68%	82.18%	82.79%	77.76%
Compustat	75.29%	72.13%	73.97%	82.41%	83.02%	78.14%
Trivial	50.00%	50.00%	50.00%	96.74%	97.64%	94.53%
Difference	1.08%	2.46%	1.71%	−0.23 %	−0.23 %	−0.38 %
P-value	0.790	0.614	0.549	0.758	0.753	0.640

Table 2.4: Predictive accuracies of the original Altman's and Ohlson's models.

2.3.5 Robustness checks

Our main results indicate that Compustat standardized data yield no improvements for original Altman's 1968 and Ohlson's 1980 bankruptcy prediction models. Moreover, in the case of Altman's model, Compustat standardized data seem to have a negative impact on the model performance. We perform additional robustness check detailed below to confirm our findings.

Assessing predictive accuracies of original (not re-estimated) Altman's and Ohlson's model

To asses both the explanatory power and predictive ability, we re-estimate both Altman's and Ohlson's model on the recent 2009-2013 data using recent cases of bankruptcies. As mentioned in Section 2.2.2, we do this because the original models 1) are more than 30 years old and likely need to be updated, and 2) were estimated using different data sources than ours that does not allow us to rigorously test whether Compustat data standardization improves models or not. However, as a robustness check, we asses and compare the predictive accuracies of the original models on our 10-K and Compustat samples. Note, that we do not need to cross-validate the models because the original models were estimated on different samples. However, we do test for significant differences in accuracies using binomial test.

The results are reported in Table 2.4. The results tend to agree with our main

Mean prediction accuracy	*first year*	*second year*	*within two years*
10-K	78.44%	79.33%	79.57%
Compustat	78.75%	78.91%	79.85%
Trivial	50.00%	50.00%	50.00%
Difference	−0.31 %	0.41%	−0.28 %
P-value	0.401	0.511	0.284

Table 2.5: Predictive accuracies of cross-validated Ohlson's model estimated on the matched sample.

finding that Altman's model predicts better if 10-K numbers are used instead of Compustat standardized numbers, although the differences in accuracies are not as large as in the case of cross-validated models and not statistically significant. Interestingly, original Altman's model yields higher accuracies than cross-validated re-estimated model if applied to Compustat data set, but not better than fitted model. Atlman's 10-K cross-validated re-estimated models achieve higher accuracies than the original Altman's model.

For Ohlson's model, 10-K models perform slightly (less than 0.4%) worse; however both Compustat and 10-K models are much worse than trivial models that predict everything to be non-bankrupt.

Re-estimating Ohlson's model on the matched sample

We find that Ohlson's re-estimated models do not achieve better accuracies than trivial models that predict everything to be non-bankrupt. The reason for that is the highly unbalanced sample we use to replicate Ohlson's model. This may be also be the reason why we find no significant differences between Compustat and 10-K Ohlson's models. Therefore, we re-estimate Ohlson's model on the matched sample we use to re-estimate Altman's model to check whether it will result in any significant difference between Compustat and 10-K versions of the model.

The predictive accuracies of the re-estimated cross-validated Ohlson's models are reported in Table 2.5. As previously, we find no statistically significant differences between 10-K and Compustat based models. However, the models are better than their trivial counterparts. The Ohlson's model also seems to yield better accuracies than Altman's model on the matched sample (except in the case of bankruptcy prediction within the first year of 10-K release).

Re-estimating Altman's model on the sample with manufacturers only

Altman (1968) developed Z score model for manufacturing firms only. Score models for private and non-manufacturing companies were developed later (see Altman (2000)). In all our samples we have included firms other than manufacturers (although we did exclude utilities, transportation companies, and financial service companies following Ohlson (1980)). This may have had a negative impact on the predictive ability of Altman's model. We exclude non-manufacturing companies from our matched sample and re-estimate Altman's model on the resulting sample.

The reduced sample contains 41 bankrupt observations one year prior to bankruptcy, and 32 bankrupt observations two years prior to bankruptcy. The number of non-bankrupt observations matches the number of the bankrupt ones. Because of the smaller sample size we perform 5-fold cross validated comparison (as opposed to 10-fold) 10 times to assess whether difference in accuracies between 10-K and Compustat models are significant.[13] The results are reported in Table 2.6.

The results are consistent with our previous findings. In all cases, we find that Atlman's models based on 10-K numbers outperform respective models based on Compustat standardized numbers. It seems that including non-manufacturing companies did not have any negative impact on the models' accuracies since excluding them did not yield any improvement. Moreover, the models estimated on the reduced sample are weaker.

13. The results still hold if we use 10-fold cross validation.

Mean prediction accuracy	*first year*	*second year*	*within two years*
10-K	70.00%	67.33%	68.33%
Compustat	65.50%	64.50%	64.67%
Trivial	50.00%	50.00%	50.00%
Difference	4.50%	2.83%	3.66%
P-value	0.000	0.008	0.007

Table 2.6: Predictive accuracies of cross-validated Altman's models estimated on the matched sample of manufacturing companies.

We also re-estimate Ohlson's model on the sample that comprises of manufacturers only. The results do not change.

Recalculating Compustat's Earnings Before Interest and Tax for Altman's model

To re-estimate values of the Altman's model, we need values of Earnings before Interest and Tax that is a non-GAAP measure. As mentioned previously, we defined EBIT as a sum of Operating and Non-Operating Income with interest being removed. According to Compustat's online manual, the definition of EBIT in Compustat is "*Sales - Net (SALE) minus Cost of Goods Sold (COGS) minus Selling, General & Administrative Expense (XSGA) minus Depreciation/Amortization (DP)*". To test whether different definitions of non-GAAP EBIT measure drive our results, we recalculate Compustat's EBIT according to our definition of EBIT, and use this variable to re-estimate Compustat's version of Altman's model.[14] We compare this alternative Compustat's model to the 10-K model in terms of predictive ability. The results are reported in Table 2.7.

The results do not change if Compustat's definition of EBIT is matched to the one we use in our 10-K sample. Note, that mean prediction accuracies of 10-K models

14. We calculate this alternative value of Compustat's EBIT as OIADP+NOPI-XINT.

Mean prediction accuracy	*first year*	*second year*	*within two years*
10-K	80.44%	74.83%	77.11%
Compustat	71.25%	72.08%	70.68%
Trivial	50.00%	50.00%	50.00%
Difference	9.19%	2.75%	6.43%
P-value	0.000	0.006	0.000

Table 2.7: Predictive accuracies of cross-validated Altman's models with an alternative definition of Compustat's EBIT variable.

are slightly different than in Table 2.3 since the cross-validation technique we employ involves random selection of samples used to estimate and test models.

2.4 Summary

S&P Capital IQ's Compustat reports standardized accounting numbers that are different from the original numbers in 10-K reports. This essay is the first study to empirically measure the benefits of Compustat's data standardization by examining whether Compustat's data improves two popular bankruptcy prediction models, Altman's 1968 Z score and Ohlson's 1980 O score.

Our main finding is that Compustat's data standardization not only yields no significant improvements for Altman's and Ohlson's model, but also has significantly negative impact on Altman's predictive accuracy. This result is supported by several robustness checks. We also find that using Compustat's standardized data instead of the original 10-K data results in significantly different outputs of both Ohlson's and Altman's models.

Our findings suggest that Compustat's standardized data may not be better than the original 10-K data for the purposes of bankruptcy prediction. However, more research is needed to evaluate the effects of Compustat data standardization on other important accounting models.

Chapter 3

Exploration and exploitation in deciding what to audit

3.1 Introduction

In this Chapter, we consider the problem of identifying irregular transactions from a set of observed transactions in a multi-period auditing setting. A transaction is *irregular*, if it is either fraudulent or erroneous. Since there is a cost to investigating a transaction by the internal audit, it would be prohibitively expensive to investigate each and every transaction to find the irregular ones. Therefore, an audit team may investigate only a portion of transactions. A question, then, arises: what transactions to investigate? The traditional audit approach is to choose a random sample from the population of all transactions. However, this approach tends to ignore the information that is known about transactions. A better approach is to ask internal audit to look into transactions that are identified as *suspicious*. Such reduction of the problem may be effective only if one has a good method of identifying suspicious transactions. In a perfect case each irregular transaction would be marked as suspicious and each suspicious transaction would be irregular with a probability close to one.

In the era of technology and computers, it seems natural to apply analytical models to identify suspicious transactions. It is cheap, fast, and, possibly, accurate. Clearly, this approach cannot be worse than the traditional audit approach of choosing a random sample from the population of all transactions. The question is how good it is. We believe that the answer depends not only on *what* analytical models are utilized, but also on *how* they are utilized. In this essay, we find that some popular analytical

models (e.g., logistic regression) may not perform that well in the auditing setting if applied traditionally. This finding is due to the peculiarities of the auditing setting itself which will be discussed below. However, by changing the way the analytical models are used we were able to achieve much better performance.

In this essay, we create a framework for analyitical models that can be used to identify suspicious transactions. We argue that most standard analytical models may not be well-suited for auditing and have to be modified in order to achieve better performance.

In order for a statistical model to be effective it should be able to *learn* from its past predictions. Just as people gain experience, a statistical model updates itself over time by taking into account how accurate it was with its previous predictions. Such information about the past predictions and their successes and failures is often called a *feedback*.

In the auditing problem of identifying irregular transactions, a model receives feedback only from the past transactions that were identified by it as suspicious and were investigated. In other words, the model uses only the information from one part of previous transactions to update itself, the part which was found to be suspicious. This is called the problem of *one-sided feedback*, and it may introduce significant difficulties for a statistical model, or even make it useless. The pitfall here is that the model will be biased towards certain types of irregular transactions seen so far and may fail to recognize irregular transactions of other types. Since the audit data is usually unbalanced, i.e., the number of irregular transactions is relatively small, such bias may be very significant.

In this essay, we try to mitigate the one-sided feedback problem by building a framework on top of the statistical model that changes the way the model learns and predicts. The idea behind this framework is to separate the prediction and investigation decision problems. The framework trades offs the immediate gain from investigating the most suspicious and important transactions for developing a more

accurate statistical model by spending audit resources to learn more about the underlying distribution of the transactional data. A more accurate statistical model may yield more benefits in the future. The proposed framework is tested on the real-world data. The results show a significant boost in performance for some statistical models.

The remainder of this Chapter is laid out as follows. In Section 3.2, we formalize a typical setting of auditing transactions when analytical models are used to decide which transactions to investigate. We introduce the framework to increase performance of analytical models in Section 3.3. Section 3.4 demonstrates how this framework can be applied to some statistical models. We test one implementation of the framework on the real-world data and discuss the results in Section 3.5. Finally, we summarize the Chapter in Section 3.6.

3.2 Auditing transactions with analytical models

Let us briefly discuss a setting where transactions are audited with the help of analytical models. We assume that audit is conducted periodically where periods can be rather short (e.g., days, hours, etc.). In each period, business generates transactions that are to be audited at the end of that period. Due to limited audit staff and time, internal audit can investigate only a certain number of transactions. Thus, the main problem is to choose which transactions to investigate.

In the traditional audit approach, the choice of which transactions to investigate is random. However, it might not be the most effective approach: it does not take into account any known information about the transactions (e.g. amount, date and time, product or service, payee, payer, etc...). Instead, we consider a more modern setting where a set of analytical models is used to choose which transactions to investigate. Such models may analyze the information about the current transactions and use the information about previous transactions and audit investigations to *predict* whether a particular transactions is irregular or not. Moreover, once a new transaction has been

investigated and its true nature (i.e., irregular or non-irregular) has been revealed, a model may *learn* from this information and *update* its beliefs to make better predictions in the future. Such models are known as *statistical models*. Examples of these are regression models, decision trees, neural networks, support vector machines, and many more. We will assume that in such an audit setting at least one statistical model is utilized.

Transactions are represented by a set of attributes that are believed to reasonably describe them. For each transaction one can observe the values of its attributes. Each transaction may result in a certain amount loss if being irregular. An example of such loss may be a dollar amount of a check, a cost of restating financial statements due to errors in accounts' balances, etc. We assume that a business can estimate an amount of a loss that can result if the transaction is indeed irregular. This is a reasonable assumption: based on the past history business can produce an estimate of the loss that may result if the transaction is irregular. Given a transaction, a set of analytical models decides which transactions to investigate based on its attributes and available past information. Of course, the number of such decisions depends on the available resources of the internal audit for that particular period. The internal audit investigates each chosen transaction, and finds it as either irregular or not. If the transaction was found to be irregular, we assume that the business bears no loss associated with it, since it has been handled in a timely manner. We assume that transactions are independent in a sense that an investigation of a particular transaction would result in revealing only its true nature, i.e., not the nature of other transactions. Hence, investigation of an irregular transaction will prevent the loss associated only with this particular transaction. Finally, learning models utilize all this information to update their beliefs.

Let us formally describe this setting. In each period t, transactions are generated by various business activities. Let $x_t^j \in \mathbb{X}$ be the jth transaction in period t with $\mathbb{X}$ indicating the set of transactional attributes, and let N_t be the total number of

An audit setting

for period $t = 1, 2, \ldots$ **do**

1. Business generates transactions $X_t = \{x_t^1, x_t^2, \ldots, x_t^{N_t}\}$, where $x_t^j \in \mathbb{X}, j = 1 \ldots N_t$.
2. Business estimates potential losses $L_t = \{l_t^1, l_t^2, \ldots, l_t^{N_t}\}$, where $l_t^j \in \mathbb{R}, j = 1 \ldots N_t$.
3. Auditor announces audit capacity $c_t \in \mathbb{R}$.
4. Analytical models choose transactions to investigate $I_t = \{x_t^{i_1}, x_t^{i_2}, \ldots\ x_t^{i_{k_t}}\} \subseteq X_t$, such that $k_t \leq c_t$.
5. Auditor investigates transactions I_t and reveals their true nature $Y_t = \{y_t^{i_1}, y_t^{i_2}, \ldots\ y_t^{i_{k_t}}\}$, where $y^{i_j} \in \{0, 1\}, j = 1 \ldots k_t$.
6. Statistical (learning) models store information about X_t, L_t, I_t, and Y_t for the future decisions.

end for

Figure 3.1: A formal representation of a setting for auditing transactions with analytical models.

transactions in that period. Therefore, in period t business produces a set $X_t = \{x_t^1, x_t^2, \ldots, x_t^{N_t}\}$ of transactions. For simplicity, we will assume that this and other related sets are ordered.[1] For each transaction x_t^j, business produces an estimate $l_t^j \in \mathbb{R}$ of the potential loss that may occur if transaction x_t^j is irregular. Hence, for a set of transactions X_t there is a corresponding set of estimated losses $L_t = \{l_t^1, l_t^2, \ldots, l_t^{N_t}\}$. At the end of each period, the internal audit decides on the number of transactions that may be investigated in that period. We will call this number an *audit capacity*, and denote it as c_t. Once all transactions has been observed and the audit capacity for the period has been determined, a set of analytical model decides which transactions to investigate based on the available current and past information. In other words, analytical models choose a subset of transactions $I_t = \{x_t^{i_1}, x_t^{i_2}, \ldots\ x_t^{i_{k_t}}\}$ from the set X_t of all transactions in that period, such that the number of chosen transactions

1. One can always create a one-to-one mapping between related sets to achieve the same effect.

k_t does not exceed the available audit capacity c_t in that period. After that, the internal audit investigates the chosen transactions I_t, and for each such transaction $x_t^{i_j}$ reveals its true nature $y_t^{i_j} \in \{0, 1\}$, with $y_t^{i_j} = 1$ indicating that transaction $x_t^{i_j}$ is irregular, and $y_t^{i_j} = 0$ that it is not. Therefore, for the set $I_t = \{x_t^{i_1}, x_t^{i_2}, \dots x_t^{i_{k_t}}\}$ of transactions chosen to be investigated, the audit team generates a set of their true labels $Y_t = \{y_t^{i_1}, y_t^{i_2}, \dots y_t^{i_{k_t}}\}$. This setting is summarized in Figure 3.1.

The benefits from auditing transactions depends not only on whether the audit team was able to identify irregular transactions, but also on what irregular transactions were identified. For example, identifying an irregular transaction that may result in a loss of \$100 may not be as good as identifying an irregular transaction that may result in \$1000. Therefore, we assume that for every transaction x_t^j, a certain *utility* is derived when the transaction is investigated. Let $u(\cdot)$ be a utility function that for every transaction x_t^j gives its utility $u(x_t^j)$. A straightforward example of such utility function is a function that for every transaction gives the dollar amount of loss associated with the transaction, i.e., $u(x_t^j) = y_t^j l_t^j$. Note that if x_t^j is not irregular, the latter function will be equal to 0 since the transaction does not result in a loss. The objective is to maximize the total utility across all periods.

The *audit utility for period t* is the sum of utilities of transactions that were investigated in period t, i.e., it is equal to

$$U_t(I_t) = \sum_{j=1}^{k_t} u(x_t^{i_j}). \tag{3.1}$$

Here $U_t(\cdot)$ is the set function that for every possible set $I_t \subseteq X_t$ of transactions chosen to be investigated, outputs the corresponding audit utility. Then, the *total audit utility for the first T periods* is equal to

$$\mathcal{U}_T(I_1, I_2, \dots, I_T) = \sum_{t=1}^{T} U_t(I_t) = \sum_{t=1}^{T} \sum_{j=1}^{k_t} u(x_t^{i_j}). \tag{3.2}$$

Now we can define the *problem of auditing transactions* as

$$\begin{aligned} \text{maximize} \quad & \mathcal{U}_T(I_1, I_2, \ldots, I_T) \\ \text{subject to} \quad & I_t \subseteq X_t,\ t = 1, \ldots, T \\ & |I_t| \leq c_t,\ t = 1, \ldots, T. \end{aligned} \tag{3.3}$$

3.3 The exploration and exploitation framework for improving analytical models

In previous section we discussed a general setting for auditing transactions with analytical models. Figure 3.1 list six steps that have to be taken in each period in such a setting. Steps 1-3, and 4 are somewhat idiosyncratic with respect to a business and an audit team. In this essay, we concentrate more on step 4 – analytical models deciding which transactions to investigate – and on the related step 6 – statistical models updating themselves with the new information.

Let us discuss what qualities are desirable for a set of analytical models used to decide which transactions to investigate. Firstly, statistical models should be able to learn from previous information, and update themselves to achieve better performance in the future. Secondly, all of the available audit capacity in each period should be used effectively. Thirdly, transactions that may result in higher losses should have a priority when deciding whether they are suspicious or not. Finally, the analytical models should be able to learn more about the underlying distribution of transactional attributes in order to find new types of irregularities, or change their bias towards the known ones. Achieving these qualities is not an easy task due to one-sided feedback presence in the auditing setting, i.e., only the true nature of investigated transactions is revealed; the analytical models learn nothing about transactions that were not found suspicious.

The reader may have noticed that the described above qualities of analytical models are not well aligned with each other. There is a conflict between the need

to investigate the transactions that may result in the highest losses and the need to learn more about the intrinsic distribution of the transactional attributes. Therefore, there has to be a trade-off between these two objectives. This is known as the *exploration/exploitation trade-off*. This problem was primarily studied for multi-armed bandit problems (Berry and Fristedt 1985; Robbins 1952; Auer, Cesa-Bianchi, and Fischer 2002). Techniques to achieve a good balance of exploration and exploitation were also applied in other areas such as reinforcement learning (Sutton and Barto 1998) and evolutionary programming (Holland 1992). In our setting, exploration refers to forcing a statistical model to mark some transaction as suspicious (even if the model does not find them suspicious at all) in order to learn more about the underlying distribution of the transactional attributes and, thus, avoid potential bias. Too much exploration may lead to model yielding too many *false suspicious* transactions (suspicious transactions that are not irregular). The other side of the coin is exploitation, which refers to allowing the model to choose those transactions to investigate that are expected to yield the highest loss. The goal is to strike a fine balance between exploration and exploitation.

A high-level representation of the proposed framework is shown in Figure 3.2. An input to analytical models is the information about the current transactions (i.e., transactions, their attributes and estimated losses in the current period) as well as the information about past transactions (transactions, attributes, estimated losses, investigated transactions and their true nature, etc. in the previous periods). An output from the analytical models are transactions that have to be investigated. These transactions are then investigated by the audit team that reveals their true nature. The information about investigated transactions and their true nature is a valuable feedback used by analytical models in the later periods. Analytical models themselves are divided into three categories: preprocessing models, exploitation models, and exploration models.

Preprocessing models are utilized before exploitation and exploration models.

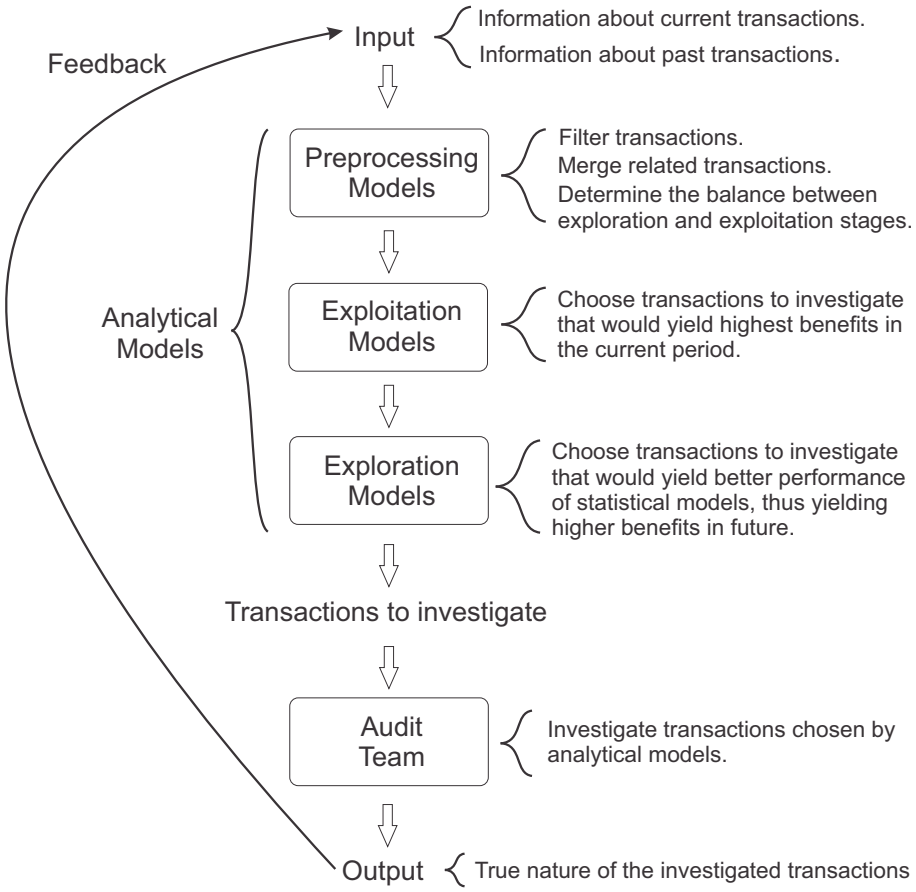


Figure 3.2: The exploration and exploitation framework for improving analytical models.

This models are used to filter transactions (e.g. remove all transactions with an estimated loss being less than $100), merge transactions (e.g., merge split payments into one transaction), and apply other user-defined rules. In other words, these models transform an initial input to facilitate the process of deciding which transactions to investigate. Preprocessing models may also define how many investigation decisions should be allocated for exploration and exploitation based on the available audit capacity and other information (e.g., expected losses for current transactions, accuracy of the statistical models employed, internal audit preferences, etc.).

Exploitation models are used after the preprocessing models, but before the exploration models. The objective of the exploration models is to choose transactions to investigate so that to obtain the highest benefits in the current period (e.g. maximize

the total prevented loss in this period). Therefore, exploration models try to achieve a short-term goal. Given a set of transactions, exploitation models make a portion of decisions which transactions to investigate. The second part of such decisions are made by exploration models.

Exploration models are the last analytical models to be utilized. Their objective is to choose transactions to investigate that will benefit the learning of the analytical models the most. Some transactions are more valuable than others from the models' learning point of view. Investigating such transactions and discovering their true nature may increase the accuracy and performance of the statistical models, thus potentially increasing the benefits in the future. Exploration models try to choose such transactions to be investigated. In this sense, exploration models are forward-looking.

In the following section, we show one way how to define exploration and exploitation models based on some popular statistical models, and how to utilize these models in conjunction in the auditing setting.

3.4 How to build exploration and exploitation models

In this section, we show how to build exploration and exploitation models for the proposed framework. Note that the method described here is one of the many possible ways to do it.

For simplicity, we will utilize only one statistical model to build one exploration and one exploitation model. In principle, many different statistical models may be utilized. The proposed method can be applied to a large number of statistical models. The only requirement for the statistical model is its ability for each transaction x_t^j to produce a *probability estimate* p_t^j, a number between 0 and 1, of the transaction been irregular. If this probability is close to 1, we interpret it as the statistical model deeming the transaction to be irregular almost certainly. If it is close to 0, then

statistical model regards the transaction to be non-irregular almost certainly.

Formally, we assume that in period t, there is a statistical model s_t such that for each transaction x_t^j from the set $X_t = \{x_t^1, x_t^2, \ldots, x_t^{N_t}\}$ of all transactions in period t, generates a probability estimate $p_t^j \in [0, 1]$, i.e., $s_t(x_t^j) = p_t^j$. The reason why statistical model s_t depends on t is that it is updated every period with new information available from the previous period(s). Statistical model s_t will be used as an underlying model to generate exploration and exploitation models for period t.

For this example, we will assume that the objective is to maximize the total prevented loss across all periods. In other words, the utility function is defined as

$$u(x_t^j) = y_t^j l_t^j, \tag{3.4}$$

where l_t^j is the estimate of loss for transaction x_t^j, and y_t^j is the true nature of x_t^j, with $y_t^j = 1$ indicating that x_t^j is irregular, and $y_t^j = 0$ meaning that it is not.

The objective function (3.2), then, is equal to

$$\mathcal{U}_T(I_1, I_2, \ldots, I_T) = \sum_{t=1}^{T} \sum_{j=1}^{k_t} y_t^j l_t^j. \tag{3.5}$$

3.4.1 Trade-off between exploration and exploitation

Exploration and exploitation models have different goals that usually contradict each other. Exploitation aims to choose transactions to investigate that will allow to reap the highest benefits in the current period (e.g. minimize loss in the current period). Exploration, on the other hand, tries to choose transactions to investigate that will increase performance of the statistical models, thus, allowing to get more benefits in the future. Exploration often means sacrificing today to gain more benefit tomorrow.

Since, in period t, the number of transactions that can be investigated is limited to audit capacity c_t, we have to decide how many investigation decisions to allocate for exploration and exploitation. We will assume that the number of investigated

transactions in each period is equal to the audit capacity, i.e., $k_t = c_t$. Let ϵ_t, $\epsilon_t \leq c_t$, be the number of decisions allocated for exploration. Let us cal this number the *exploration capacity* for period t. Then the number of decisions allocated for exploitation is equal to $c_t - \epsilon_t$. Therefore, in each period t, we have to decide on the value of ϵ_t.

We propose a natural way to dynamically balance exploration and exploitation in each period. The idea is the following: in period t, assess the accuracy of statistical model s_{t-1} in period $t-1$; if the accuracy is high, concentrate on exploitation in period t, and if the accuracy is low, concentrate on exploration in period t to improve the accuracy of the model in the future periods. We consider the accuracy of the statistical model only in the most recent $(t-1)$ period, since it reflects the current accuracy of the statistical model, i.e., the one we are likely to get in the current (t) period.

Let $P_{t-1} = \{p_{t-1}^{i_1}, p_{t-1}^{i_2}, \dots p_{t-1}^{i_{c_{t-1}}}\}$ be the probability estimates of the transactions $I_{t-1} = \{x_{t-1}^{i_1}, x_{t-1}^{i_2}, \dots x_{t-1}^{i_{c_{t-1}}}\}$ investigated in the period $t-1$, and let $Y_{t-1} = \{y_{t-1}^{i_1}, y_{t-1}^{i_2}, \dots y_{t-1}^{i_{c_{t-1}}}\}$ be the corresponding true labels of these transactions. These probability estimates P_{t-1} and true labels Y_{t-1} can be considered as two distinct probability distributions and, hence, we can compute a distance between them. Such distance would indicate how close the probability estimates yielded by the statistical model are to the true labels of these transactions, and, therefore, how accurate they are. A good measure to calculate such distance is the cross-entropy. In our case it is equal to

$$D_{t-1}(Y,P) = -\sum_{j=1}^{c_{t-1}} \left\{ y_{t-1}^{i_j} \log_2 p_{t-1}^{i_j} + (1 - y_{t-1}^{i_j}) \log_2 (1 - p_{t-1}^{i_j}) \right\} \tag{3.6}$$

with $0 \cdot \log_2 0$ defined as 0.

The value of $D_{t-1}(Y,P)$ is an absolute measure of the statistical model accuracy and can take any value greater or equal than 0. It may be hard to interpret

when choosing the exploration capacity ϵ_t. We need a benchmark value to compare $D_{t-1}(Y, P)$ to. Consider a case when the statistical model has no information about the previous transactions. In this case the best probability estimate that the model can yield for a transaction being irregular is 0.5 (a fifty-fifty chance). The value of the cross-entropy for this random-guessing model, therefore, would be

$$\begin{aligned} D_{t-1}(Y, 0.5) &= -\sum_{j=1}^{c_{t-1}} \left\{ y_{t-1}^{i_j} \log_2 0.5 + (1 - y_{t-1}^{i_j}) \log_2 0.5 \right\} \\ &= -\sum_{i=1}^{c_{t-1}} \log_2 0.5 = -\sum_{i=1}^{c_{t-1}} -1 = c_{t-1}. \end{aligned} \tag{3.7}$$

That is $D_{t-1}(Y, 0.5)$ is equal to the audit capacity in period $t-1$. Dividing $D_{t-1}(Y, P)$ by $D_{t-1}(Y, 0.5) = c_{t-1}$ yields a relative measure of the statistical model accuracy with respect to the accuracy of the random-guessing model. If this fraction is less than 1, then the statistical model does a better job than simply random guessing. Otherwise, its performance is alarmingly poor. In the latter case, we would require a lot of exploration to calibrate the statistical model and increase its future performance.

Theoretically, there is no upper bound on the possible values of $D^{t-1}(y, p)$, i.e., it may be arbitrarily large. However, for our purposes, we may limit its highest value to be twice the value of c_{t-1}. To put it simply, if the statistical model's performance is two times worse than the random guessing, it is bad enough to impose the maximum exploration on the model. This way, our relative measure of the statistical model's accuracy is equal to

$$\min \left(\frac{D_{t-1}(Y, P)}{c_{t-1}}, 2 \right), \tag{3.8}$$

and its normalized version (the one that lies between 0 and 1) would be

$$R_{t-1}(Y, P) = \min \left(\frac{D_{t-1}(Y, P)}{2c_{t-1}}, 1 \right). \tag{3.9}$$

Determining the exploration capacity

Input: exploration rate ρ

1. Compute cross-entropy $D_{t-1}(Y, P)$ as in (3.6)
2. Compute measure of accuracy $R_{t-1}(Y, P)$ as in (3.9)
3. Compute the exploration capacity $\epsilon_t = \lfloor \rho R^{t-1}(y, p) c_t \rfloor$

Output: exploration capacity ϵ_t

Figure 3.3: A way to calculate the exploration capacity ϵ_t for period t to determine the balance between exploration and exploitation in that period.

Then, we can define the exploration capacity as

$$\epsilon_t = \lfloor \rho R_{t-1}(Y, P) c_t \rfloor, \tag{3.10}$$

where $\lfloor \cdot \rfloor$ is the floor function, and $\rho \in [0, 1]$ is the *exploration rate* parameter. The exploration rate parameter is set by the audit team, and defines the exploration capacity's sensitivity to the statistical model accuracy. The greater it is, the more transactions would be reserved for exploration. In the worst case scenario, when $R_{t-1}(Y, P) = 1$, the exploration rate equals the fraction of the audit capacity reserved for exploration. For example, if $R_{t-1}(Y, P) = 1$ and $\rho = 0.95$, then 95% of the audit capacity will be used for exploration in period t. For a brief summary of how the exploration capacity is calculated see Figure 3.3.

3.4.2 Exploitation model

The objective of exploitation model is to reap the highest benefits in the current period. In our case, we may assume that this objective is to minimize the total expected loss in the current period.

Given the value of the exploration capacity ϵ_t (see previous section), the number of investigation decisions in the exploitation stage is equal to $c_t - \epsilon_t$. Therefore we need to choose $c_t - \epsilon_t$ transactions from the set $X_t = \{x_t^1, x_t^2, \ldots, x_t^{N_t}\}$ of all

Exploitation Model

1. Define $E_t = \{e_t^1, e_t^2, \ldots, e_t^{N_t}\}$ where $e_t^j = x_t^j \cdot p_t^j$.
2. Define $\bar{E}_t = \{e_t^{j_1}, e_t^{j_2}, \ldots, e_t^{j_{N_t}}\}$ where $e_t^{j_1} \geq e_t^{j_2} \geq \cdots \geq e_t^{j_{N_t}}$.
3. Choose transactions $x_t^{j_1}, x_t^{j_2} \ldots, x_t^{j_{c_t-\epsilon_t}}$ to investigate.

Output: transactions $x_t^{j_1}, x_t^{j_2} \ldots, x_t^{j_{c_t-\epsilon_t}}$

Figure 3.4: The procedure of choosing transactions for exploitation.

transactions in period t. For every transaction $x_t^j \in X_t$, the statistical model generates a probability estimate p_t^j of this transaction being irregular. Therefore, the statistical model generates a set $P_t = \{p_t^1, p_t^2, \ldots, p_t^{N_t}\}$ of probability estimates in period t.

The value $e_t^j = p_t^j \cdot l_t^j$ is the expected loss associated with the transaction x_t under the utilized statistical model. Let $E_t = \{e_t^1, e_t^2, \ldots, e_t^{N_t}\}$ be the set of expected losses that corresponds to the set of transactions X_t. Let $\bar{E}_t = \{e_t^{j_1}, e_t^{j_2}, \ldots, e_t^{j_{N_t}}\}$ be the ordered set of expected losses such that $e_t^{j_1} \geq e_t^{j_2} \geq \cdots \geq e_t^{j_{N_t}}$. Then, indexes j_1, j_2, $\ldots$, $j_{c_t-\epsilon_t}$ correspond to $c_t - \epsilon_t$ transactions for which the expected losses are the highest. Hence, the exploitation model will choose transactions $x_t^{j_1}, x_t^{j_2} \ldots, x_t^{j_{c_t-\epsilon_t}}$ to investigate. A summary of the process is given in Figure 3.4.

3.4.3 Exploration model

The purpose of the exploration is to learn more about the underlying distribution of the transactional attributes by investigating transactions possibly other than the ones with the highest expected losses. Therefore, the exploration model does not decide which transaction to investigate based on the expected loss of transactions.

The exploration model has to choose ϵ_t transactions out of $m_t = N_t - c_t + \epsilon_t$ transactions (because the exploitation model has already chosen $c_t - \epsilon_t$ transactions). In this essay, we propose to randomly choose transactions to investigate out the remaining ones. However, the chance of a particular transaction being chosen is proportional to how uncertain the statistical model is about its true nature. In other words, the less certain the statistical model is about a transaction, the greater the

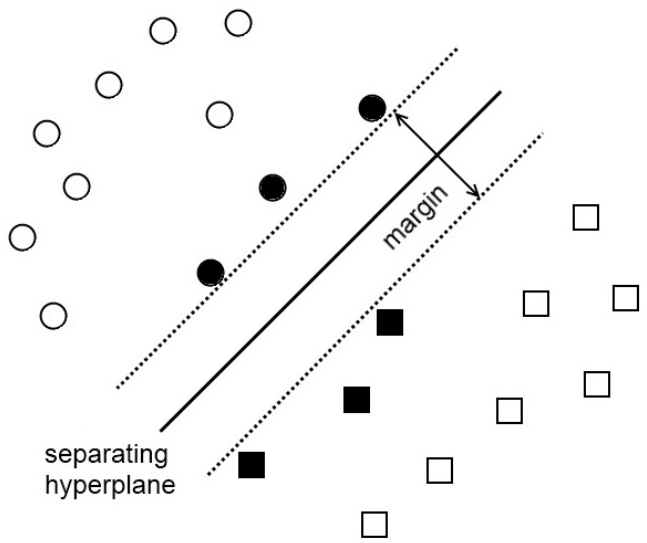


Figure 3.5: Illustration of an idea to choose transactions to investigate based on the statistical model's uncertainty about them.

chance it will be investigated. Transactions that the statistical model is least certain about may be very useful in exploring the underlying distribution of transactional attributes. By learning from such observations the statistical model may enhance its accuracy of predicting irregular transactions.

This idea is somewhat similar to the one in Lewis and Gale (1994). However, in our case the choice of transactions is non-deterministic. Such randomization may be more beneficial in the auditing problem where the number of irregular transactions is relatively small.

To illustrate the above point, consider the case of a margin statistical model (such as support vector machines) as in Figure 3.5. This margin statistical model tries to separate observations of different classes by dividing the space of attributes with a hyperplane. The hyperplane creates a margin as in Figure 3.5. The statistical model assumes an observation to belong to the one class (circles) if it is on the one side of the margin, and to belong to the other class (rectangles) if it is on the other side. The larger the distance from the observation to the margin, the higher is the probability of it belonging to the respective class. Therefore, the model is least certain about the

observations that are very close to the margin (black circles and triangles in Figure 3.5). Each time the model learns from new observations the margin is re-estimated. Learning from the observations close to the margin may change it more drastically than learning from the observations far from the margin. Therefore, such observations may provide the most valuable "experience" to the model.

Let $X_t^m = \{x_t^{q_1}, x_t^{q_2}, \ldots, x_t^{q_{m_t}}\}$ be the set of transactions in period t inputted to the exploration model (all transactions except those decided to be investigated in the exploitation stage). For each transaction $x_t^j \in X_t^m$, we may define the *measure of* the statistical model's *uncertainty* about the transaction's true nature by calculating how close its probability estimate p_t^j is to the value of 0.5[2]:

$$u_t^j = |p_t^j - 0.5|. \tag{3.11}$$

The number u_t^j is between 0 and 0.5, with 0 indicating the highest uncertainty and 0.5 indicating the highest confidence about the true nature of transaction by the statistical model.

Let us randomly generate m_t numbers drawn from the uniform distribution on the interval $[0, 1]$. Let $\mathcal{R}_t = \{r_t^{q_1}, r_t^{q_2}, \ldots, r_t^{q_{m_t}}\}$ be the set of those numbers. The product $g_t^j = r_t^j \cdot u_t^j$ is a random number between 0 and 0.5. The smaller is the u_t^j, the higher is the probability of g_t^j being close to 0 (rather than to 0.5). Choosing ϵ_t transactions for which such products are the smallest is equivalent to randomly choosing ϵ_t transactions out of a pool of m_t transactions, in a way that the chance of choosing a particular transaction is proportional to the degree of the statistical model's uncertainty about it. Let $G_t = \{g_t^{q_1}, g_t^{q_2}, \ldots, g_t^{q_{m_t}}\}$ be the set of all such products, and $\bar{G}_t = \{g_t^{w_1}, g_t^{w_2}, \ldots, g_t^{w_{m_t}}\}$ be the ordered set of those products, where $g_t^{w_1} \leq g_t^{w_2} \leq \cdots \leq g_t^{w_{m_t}}$. Then, the exploration model chooses transactions

2. In our case, the probability estimate of 0.5 represents the statistical model's total uncertainty about the true nature of a transaction.

Exploration Model

1. For each $x_t^j \in X_t^m$, compute the measure of uncertainty $u_t^j = |p_t^j - 0.5|$.
2. For each $x_t^j \in X_t^m$, generate a random number $r_t^j \in [0,1] \sim \mathcal{U}(0,1)$.
3. Define $G_t = \{g_t^{q_1}, g_t^{q_2}, \ldots, g_t^{q_{m_t}}\}$, where $g_t^{q_i} = u_t^{q_i} \cdot r_t^{q_i}$.
4. Define $\bar{G}_t = \{g_t^{w_1}, g_t^{w_2}, \ldots, g_t^{w_{m_t}}\}$, where $g_t^{w_1} \leq g_t^{w_2} \leq \cdots \leq g_t^{w_{m_t}}$.

Choose transactions $x_t^{w_1}, x_t^{w_2}, \ldots, x_t^{w_\epsilon}$ to investigate.

Output: transactions $x_t^{w_1}, x_t^{w_2}, \ldots, x_t^{w_\epsilon}$ to investigate.

Figure 3.6: The procedure of choosing transactions for exploration.

$x_t^{w_1}, x_t^{w_2}, \ldots, x_t^{w_\epsilon}$ to investigate. A summary for exploration is provided in Figure 3.6.

3.5 Empirical testing

In this section, we test the described above implementation of the proposed exploration/exploitation auditing framework on the two real-world data sets and evaluate the results. The data sets used to demonstrate the performance of the framework are: credit card data of a large multinational bank and census data of the U.S. Census Bureau. The first data set is of a business nature and, arguably, resembles a real audit data in out setup. The second data set is a well-studied data utilized in many studies (e.g. Kohavi 1996; Cohen and Singer 1999) and in the Data Mining and Knowledge Discovery (KDD) Cup competition.

3.5.1 Measures for comparison

In order to compare the performance of the unmodified statistical model (which we will call *normal model* for short) with the derived exploration/exploitation models we need a benchmark. Each set of analytical models in each period outputs transactions to be investigated, i.e., it outputs $I_1, I_2, \ldots, I_T$. A benchmark measure should take these as an input and produce a number as an output that we can use to compare

the models.

Since in Section 3.4 we assumed that the objective is to maximize the total prevented loss, we will use the objective function (3.5) in our comparison:

$$\mathcal{U}_T(I_1, I_2, \ldots, I_T) = \sum_{t=1}^{T} \sum_{j=1}^{k_t} y_t^j l_t^j. \tag{3.12}$$

This measure is an absolute measure of the prevented loss. We also would like to have a relative measure of the prevented loss. The *relative prevented loss in period t* is equal to

$$P_t(I_t) = \frac{\sum_{d=1}^{k_t} y_t^{i_d} l_t^{i_d}}{\sum_{n=1}^{N_t} y_t^n l_t^n}. \tag{3.13}$$

In the above equation, the nominator is equal to the prevented loss, and the denominator is equal to the total loss in period t. We, then, define the *mean relative prevented loss (MRLP) for the first T periods* as

$$MRLP_T = \frac{1}{T} \sum_{t=1}^{T} P_t. \tag{3.14}$$

We will also use MRLP in our testing. It is a better measure to compare the accuracy of the analytical models than (3.12). It gives a fairer comparison of the models' performance across all periods.

3.5.2 Statistical models

In our testing, we use two statistical models: logistic regression, and support vector machines (SVM) with a linear kernel. Logistic regression is a widely used statistical model that yields probabilities of observations belonging to particular classes. SVM is a popular classification model that originated in machine learning literature. Standard versions of SVM output only class predictions. However, Wu, Lin, and Weng (2004); Platt (2000); Lin, Lin, and Weng (2003) develop methods for SVM to yield

probability estimates for observations. We use LIBSVM library (Chang and Lin 2011) implementation of SVM outputs probability estimates.

For each statistical model, we construct the appropriate exploration and exploitation models as described in Section 3.4. For each exploration and exploitation model, we report results for different values of the exploration rate; namely, for $\rho = 0.25$, $\rho = 0.5$, $\rho = 0.75$ and $\rho = 1$.

3.5.3 Multinational bank credit card data

We first test the exploration and exploitation auditing framework using a credit card data of a large multinational bank.

Data description

The credit card data we used in this study contains information about the bank's credit cards opened in 2011 and their status as of the first quarter of 2012. Each observation indicates a rather general information about the credit card account and its owner. The list of variables is provided in Appendix C.

To simulate an irregular transaction, we assume that an observation is irregular if the credit card was canceled by the bank. The bank may cancel credit card because of various reasons including fraud. We also assume that each observation carries a loss if being irregular with the value of the loss being equal to the amount of the credit limit associated with the account. Credit limit is an intuitive estimate for the loss associated with a credit card. We randomly selected 500,000 observations in a way that 1% (5,000) of these are irregular. The small percentage of irregular observations is used to emulate the small number of irregular transactions, and therefore creates the unbalanced data set problem.

We partitioned all 500,000 observations into 500 periods with 1,000 observations each. We set the number of transactions that can be investigated (audit capacity) in

	Normal model	Exploration/exploitation models			
		$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 1$
Logistic Regression					
Total prevented loss	938823	2038096	2347275	2295943	2274050
Difference	0	1099273	1408452	1357120	1335227
Linear SVM					
Total prevented loss	2132940	2272404	2215005	2162068	2156688
Difference	0	139464	82065	29128	23748

Table 3.1: Credit card data testing results as measured by the total prevented loss. The difference row indicates the difference in the prevented loss between the exploration/exploitation models and the normal model. Higher values are better.

each period to 100 (10% of all transactions).

Testing results

The results of the exploration and exploitation framework testing are reported in Tables 3.1 and 3.2.

The results show that the logistic regression model is improved a lot if the exploration and exploitation technique is utilized. The mean relative prevented loss increased from 11.56% to 24.58% when the exploration and exploitation model was used with the exploration coefficient parameter of $\rho = 0.75$. This translates to more then 110% better performance of the exploration and exploitation logistic regression model as compared to the normal logistic regression model. Also the exploration and exploitation logistic regression model demonstrated the best results across all models tested for the credit card data set.

The results for the SVM model are not as impressive as the ones for the logistic regression model. The normal SVM model yields better results than the normal logistic regression model, but the exploration and exploitation framework was able to improve the normal SVM model only by a little – in the best case by 8% (for $\rho = 0.25$).

	Normal model	Exploration/exploitation models			
		$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 1$
Logistic Regression					
MRPL	11.56%	21.35%	23.37%	24.58%	23.90%
Difference	0%	9.79%	11.81%	13.02%	12.34%
Linear SVM					
MRPL	15.89%	17.20%	16.62%	16.26%	16.24%
Difference	0%	1.31%	0.73%	0.37%	0.35%

Table 3.2: Credit card data testing results as measured by the Mean Relative Prevented Loss (MRPL) in percentage. The difference row indicates the difference in MRLP between the exploration/exploitation models and the normal model. Higher values are better.

It appears that the choice of the exploration coefficient parameter, ρ, is important. Moreover, the optimal value seems to be model-dependent. For the logistic regression, the optimal exploration coefficient value is 0.75, while for the SVM the optimal value is 0.25 for the credit card data set.

It is interesting to see how the performance of the exploration and exploitation model changes over time. Figure C.1 in Appendix C shows the period performance differences, as measured by the relative prevented loss measure, between the logistic regression exploration and exploitation model (with $\rho = 0.5$) and logistic regression normal model. Negative values indicate a better performance of the normal model, while positive values indicate a better performance of the exploration and exploitation model.

To have a clearer picture what happens on average, we fitted the difference points in Figure C.1 to a quadratic polynomial curve in Figure C.1. From Figure C.1 it follows that at first the sacrifice of investigation decisions for exploration yields a worse performance of the exploration and exploitation model compared to the normal model. However, over time the exploration model performance becomes significantly better on average. Hence, the exploration and exploitation model yields a high value in the future for a price of slightly worse performance in the present.

Also note that the fitted curve in Figure C.1 is concave (as opposed to convex) meaning that the marginal benefit of the model decreases over time and less exploration is required in the later periods. This indicates that exploration capacity should be allocated dynamically based on some performance criteria (in our implementation of the framework, it is based on the accuracy of the underlying statistical model, see §3.4.1). At some time the difference in performance begin to decrease as the exploration and exploitation model reaches its potential and the normal model gains more experience. At this point, it may be also beneficial to decrease the value of the exploration coefficient of the exploration and the exploitation model.

3.5.4 Census data

The second test data we utilize in this study is large a census data set obtained from the University of California, Irvin (UCI) Machine Learning repository (Asuncion and Newman 2007).

Data description

The census data is extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. It was used in dozens of studies (Kohavi 1996; Cohen and Singer 1999) and in the Data Mining and Knowledge Discovery (KDD) Cup competition of 1999. We use the KDD version of the data. Out of 42 fields in the data, we kept only 12 since most data variables are categorical and, thus, would yield too many dummy variables in the model if used in full. The list of the used variables can be found in Appendix C.

Each observation in the data is treated as a transaction. We used the variable "Education" as a class label in our testing. If the value of the variable is "Doctorate degree" or "Master's degree", then the transaction is considered to be irregular. Otherwise, it is not irregular. For the loss variable we used the variable "Age". This way the loss may be partially correlated with the true label of the transaction that

	Normal model	Exploration/exploitation models			
		$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 1$
Logistic Regression					
Total prevented loss	266675	238314	315292	310742	305592
Difference	0	46937	48617	44047	38917
Linear SVM					
Total prevented loss	260974	288969	288380	273592	266439
Difference	0	27995	27406	12618	5465

Table 3.3: Census data testing results as measured by the total prevented loss. The difference row indicates the difference in the prevented loss between the exploration/exploitation models and the normal model. Higher values are better.

would probably be the case in the real-world setting.

We consider 200 auditing periods. For each period, we randomly choose 1000 observations from the data set (without repeating). In each period the audit capacity is set to 100 (10% of the number transactions in each period). The total number of irregular transactions is equal to 7881 (3.94% of all transactions) which results in the highly unbalanced data set.

Testing results

The results of the framework testing on the census data is presented in Tables 3.3 and 3.4.

Similarly to the credit card data testing, the results show that the logistic regression model yields the best results. For logistic regression, the exploration and exploitation models significantly outperform the normal one. The total prevented loss for the exploration/exploitation models with exploration rate $\rho = 0.5$ is 18.23 % greater than the total prevented loss of the normal one. The model performance is increased by 57.3% as measured by the MRPL. It is interesting to observe, that the results tend to be better as we increase the exploration rate at first ($\rho = 0.5$ yields better results than $\rho = 0.25$), and then decrease as we increase the exploration rate

	Normal model	Exploration/exploitation models			
		$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 1$
Logistic Regression					
MRPL	22.53%	35.44%	35.93%	34.67%	33.24%
Difference	0%	12.91%	13.4%	12.14%	10.71%
Linear SVM					
MRPL	20.92%	28.68%	28.76%	25.44%	22.47%
Difference	0	7.76%	7.84%	4.52%	1.55%

Table 3.4: Census data testing results as measured by the Mean Relative Prevented Loss (MRPL) in percentage. The difference row indicates the difference in MRLP between the exploration/exploitation models and the normal model. Higher values are better.

even more ($\rho = 0.5$ is better than $\rho = 0.75$, and $\rho = 0.75$ is better than $\rho = 1$). This suggests, that there is a sweet point for the exploration parameter somewhere between $\rho = 0.25$ and $\rho = 0.75$, that blends the optimal amount of exploration and exploitation.

SVM model is the second best model. Again, the exploration and exploitation models are better than the normal one. The exploration and exploitation model performs the best with ρ being equal to 0.25 or 0.5. The testing shows that the exploration and exploitation SVM models can outperform the normal one by 37.59% in terms of model accuracy. The total prevented loss increased by 10.73% compared to the normal model. As in the case with the logistic regression model, the SVM model performance is decreased if the exploration rate ρ is too large.

3.6 Summary

In this Chapter, we consider the auditing problem of identifying irregular transactions from a set of observed transactions. Specifically, we consider a multi-period setting where a set of analytical models is used to identify suspicious transactions which would be later investigated by the audit staff.

Although, some statistical models are able to learn from the past history of transactions, they may not be able to learn effectively due to the nature of the auditing setting. In particular, a model only learns from the past transactions that were identified by it as suspicious and were investigated. This may result in model biasing towards certain types of irregular transactions that have been previously investigated. Therefore, it might be unable to identify the other types of irregularities. This is known as the problem of one-sided feedback. This problem may be even more pronounced in the auditing setting where the number of irregular transactions is significantly less than the number of non-irregular transaction.

In this essay, we develop a framework, that boosts the performance of analytical models in the auditing setting. The framework utilizes the exploration and exploitation technique, and separates the prediction and investigation decisions to learn more about the transactional attributes distribution while pursuing the main goal (e.g., maximization of the prevented loss).

We demonstrate how to build simple exploration and exploitation models from a class of statistical models that outputs probability predictions (e.g. logistic regression). We also show how to dynamically adjust the degree of exploration and exploitation based on past accuracy of the analytical models.

We test the framework on large-scale, real-world data with two popular statistical models: logistic regression, and support vector machines. The results show that he performance of the statistical models can be drastically improved if the exploration and exploitation framework is used.

Conclusions

This dissertation studies the properties of accounting data and their implications from two different perspectives. The first part of the dissertation examines the differences between Compustat North America Fundamentals, the most frequently used financial database in accounting research, and annual financial reports filed by U.S. companies, and empirically studies the effects of these differences on accounting-based bankruptcy prediction models. The second part of the dissertation discusses the unique characteristics of transactional data and how these characteristics aggravate the problem of applying analytical learning models in a multi-period audit setting, and develops a solution to address this problem.

Chapter 1 is the first essay, in which we utilize the XBRL reporting technology to conduct the first large-scale comparison of numbers found in Compustat North America Fundamentals Annual and numbers as reported in the original financial reports filed by domestic U.S. GAAP companies with the SEC. We develop a comparison methodology that allows automated data extraction from XBRL 10-K reports, mapping XBRL data to the appropriate Compustat variables, and reconciliation of any identified discrepancies between the two sources of data. We apply this methodology to compare 30 popular accounting line items between Compustat and 10-K reports of more than 5,000 companies with filing period end dates ranging from October 1, 2011, to September 30, 2012. The results show that 17 out of 30 compared accounting items significantly differ across the data sets. The differences are mostly due to Compustat's standardization practices that involve adjustments of the original numbers to fit Compustat's standardized definitions of variables. Accounting items with more complex definitions (e.g., Cost of Goods Sold) tend to differ more than the

accounting items with less complex definitions (e.g., Total Assets). In addition, we show that company's characteristics such as industry and size, and the type of financial statement where the numbers are reported affect the amount and magnitude of discrepancies. These results show that the differences between Compustat and 10-K filings are non-trivial, and are likely to affect outcomes of the studies that utilize Compustat as opposed to 10-K data.

Chapter 2 extends the study of Chapter 1 by examining the effects of using standardized Compustat data as opposed to the original 10-K data for the purposes of bankruptcy prediction. We consider two popular accounting-based bankruptcy prediction models, Altman's 1968 and Ohlson's 1980 models. For each model, we compare the output, explanatory power, and predictive ability between two versions of the model - one based on Compustat data and the other based on the original 10-K data. We find that there is a significant difference in outputs of both Altman's and Ohlson's models if Compustat data is utilized instead of the original 10-K data. We also find that Altman's model based on 10-K data has a significantly better predictive accuracy (up to 8.56%) than the corresponding Compustat-based model. The results suggest that Compustat's standardization practices may have a negative effect on the performance of bankruptcy prediction models.

Finally, Chapter 3 considers the problem of applying analytical learning models for the purpose of identifying irregular transactions in a multi-period auditing setting. Transactional data presents two major challenges for using analytical models: 1) the data is highly unbalanced – the number of irregular transactions is usually a small fraction of all transactions that reduces the likelihood of detection of irregularities, and 2) in each auditing period, due to constrained audit resources, only a limited number of transactions can be investigated that does not provide an optimal learning experience for analytical models that utilize historical data for calibration. To mitigate these effects, we propose a framework for analytical models that is based on the concepts of data exploration and exploitation (Berry and Fristedt 1985; Robbins 1952;

Auer, Cesa-Bianchi, and Fischer 2002). The framework exchanges the immediate gain from investigating the most suspicious and important transactions (exploitation) in return for more accurate statistical model by spending audit resources to learn more about the underlying distribution of the transactional data (exploration). We test the framework on two real-world data sets. The results show significant increase in performance of analytical models when the framework is utilized.

Bibliography

Altman, E. I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance* 23 (4): 589–609.

Altman, E. I. 2000. Predicting financial distress of companies: revisiting the z-score and zeta models. *Stern School of Business, New York University:*9–12.

Asuncion, A., and D. Newman. 2007. *UCI machine learning repository.* University of California, Irvine, School of Information and Computer Sciences. http://www.ics.uci.edu/$%5Csim$mlearn/%7BMLR%7Depository.html.

Auer, P., N. Cesa-Bianchi, and P. Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47 (2-3): 235–256.

Basili, V. R., and R. W. Selby. 1987. Comparing the effectiveness of software testing strategies. *Software Engineering, IEEE Transactions on,* no. 12:1278–1296.

Beedles, W. L., and M. A. Simkowitz. 1978. A Note on Skewness and Data Errors. *The Journal of Finance* 33 (1): pages.

Bennin, R. 1980. Error Rates in CRSP and COMPUSTAT: A Second Look. *The Journal of Finance* 35 (5): pages.

Berry, D. A., and B. Fristedt. 1985. *Bandit problems: sequential allocation of experiments.* Chapman / Hall, London.

Bodie, Z., A. Kane, and A. Marcus. 2008. *Essentials of investments.* 7th ed. Irwin/McGraw-Hill.

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 (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 (Apr.).

Bowen, R. M., A. K. Davis, and D. A. Matsumoto. 2005. Emphasis on pro forma versus gaap earnings in quarterly press releases: determinants, sec intervention, and market reactions. *The Accounting Review* 80 (4): 1011–1038.

Bradshaw, M. T., and R. G. Sloan. 2002. Gaap versus the street: an empirical assessment of two alternative definitions of earnings. *Journal of Accounting Research* 40 (1): 41–66.

Chang, C.-C., and C.-J. Lin. 2011. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm, *ACM Transactions on Intelligent Systems and Technology* 2 (3): 27:1–27:27.

Chychyla, R., and A. Kogan. 2013. Using xbrl to conduct a large-scale study of discrepancies between the accounting numbers in compustat and sec 10-k filings. Available at SSRN: http://ssrn.com/abstract=2304473 or http://dx.doi.org/10.2139/ssrn.2304473.

Cohen, W. W., and Y. Singer. 1999. A simple, fast, and effective rule learner. In *Proceedings of the sixteenth national conference on artificial intelligence and the eleventh innovative applications of artificial intelligence conference innovative applications of artificial intelligence,* 335–342. AAAI '99/IAAI '99. Orlando, Florida, United States: American Association for Artificial Intelligence. ISBN: 0-262-51106-1. http://dl.acm.org/citation.cfm?id=315149.315320.

Collins, D. W., and M. C. O'Connor. 1978. An Examination of the Association between Accounting and Share Price Data in the Extractive Petroleum Industry: A Comment and Extension. *The Accounting Review* 53 (1): 228–239.

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.

Dietterich, T. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* 10 (7): 1895–1923.

Du, H., M. A. Vasarhelyi, and X. Zheng. 2013. XBRL mandate: Thousands of filing errors and so what? *Journal of Information Systems* 27 (1): 61–78.

Elliott, W. B. 2006. Are investors influenced by pro forma emphasis and reconciliations in earnings announcements? [In English]. *The Accounting Review* 81 (1): pages. ISSN: 00014826.

Eskew, R. K. 1975. An Examination of the Association between Accounting and Share Price Data in the Extractive Petroleum Industry. *The Accounting Review* 50 (2): 316–324.

Fama, E. F., and K. R. French. 1997. Industry costs of equity. *Journal of Financial Economics* 43 (2): 153–193.

Freeman, R., and S. Tse. 1992. An earnings prediction approach to examining intercompany information transfers. *Journal of Accounting and Economics* 15 (4): 509–523.

Guenther, D. A., and A. J. Rosman. 1994. Differences between COMPUSTAT and CRSP SIC codes and related effects on research. *Journal of Accounting and Economics* 18 (1): 115–128.

He, H., and E. A. Garcia. 2009. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on* 21 (9): 1263–1284.

Hoffman, C. 2013. *Analysis of SEC XBRL Financial Filings.* xbrl.squarespace.com. Available at http://www.xbrlsite.com/2013/Library/AnalysisOfSECXBRLFilings-2013-07-03.pdf, July.

Holland, J. H. 1992. *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.* (First edition 1975, Ann Arbor: University of Michigan Press.) MIT Press.

Institute of Electrical and Electronics Engineers. 2008. *IEEE Standard for Software Reviews and Audits.*

Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles* 37:547–579.

Joy, O. M., and J. O. Tollefson. 1975. On the financial applications of discriminant analysis [in English]. *The Journal of Financial and Quantitative Analysis* 10 (5): pages. ISSN: 00221090.

Kahle, K. M., and R. A. Walkling. 1996. The Impact of Industry Classifications on Financial Research. *Journal of Financial and Quantitative Analysis* 31 (03): 309–335.

Kern, B. B., and M. H. Morris. 1994. Differences in the COMPUSTAT and Expanded Value Line Databases and the Potential Impact on Empirical Research. *The Accounting Review* 69 (1): pages.

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

Klein, B., and D. Rossin. 1999. Data quality in linear regression models: effect of errors in test data and errors in training data on predictive accuracy. *Informing Science* 2 (2).

Kohavi, R. 1996. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the second international conference on knowledge discovery and data mining,* 202–207. AAAI Press.

Lewis, D. D., and W. A. Gale. 1994. A sequential algorithm for training text classifiers. In *Sigir,* ed. W. B. Croft and C. J. van Rijsbergen, 3–12. ACM/Springer. ISBN: 3-540-19889-X.

Lin, H.-T., C.-J. Lin, and R. Weng. 2003. *A note on Platt's probabilistic outputs for support vector machines.* (Technical Report). National Taiwan University.

McEnally, R. W. 1974. A Note on the Return Behavior of High Risk Common Stocks. *The Journal of Finance* 29 (1): 199–202.

Mensah, Y. M. 1984. An examination of the stationarity of multivariate bankruptcy prediction models: a methodological study [in English]. *Journal of Accounting Research* 22 (1): pages. ISSN: 00218456.

Myers, G. J. 1978. A controlled experiment in program testing and code walkthroughs/inspections. *Communications of the ACM* 21 (9): 760–768.

Ohlson, J. A. 1980. Financial ratios and the probabilistic prediction of bankruptcy [in English]. *Journal of Accounting Research* 18 (1): pages.

Platt, J. 2000. Probabilities for sv machines. In *Advances in large margin classifiers,* ed. A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, 61–74. Cambridge, MA: MIT Press.

Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58 (5): 527–535.

Rosenberg, B., and M. Houglet. 1974. Error Rates in CRSP and Compustat Data Bases and their Implications. *The Journal of Finance* 29 (4): pages.

San Miguel, J. G. 1977. The Reliability of R&D Data in COMPUSTAT and 10-K Reports. *The Accounting Review* 52 (3): pages.

Securities and Exchange Commission (SEC). 2005. *XBRL Voluntary Financial Reporting Program on the EDGAR System.,* Feb.

Securities and Exchange Commission (SEC). 2009. *Interactive Data to Improve Financial Reporting,* Jan.

Stone, D. 1968. Caveats in Computer-Aided Financial Analysis. *Financial Analysts Journal* 24 (1): 149–153.

Sutton, R. S., and A. G. Barto. 1998. *Reinforcement learning: an introduction.* Cambridge, MA: MIT Press. citeseer.ist.psu.edu/sutton98reinforcement.html.

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

Tallapally, P., M. S. Luehlfing, and M. Motha. 2012. Data differences - XBRL versus Compustat. *Journal of Technology Research* 3.

Thies, J. B., and L. Revsine. 1977. Capital Expenditures Data for Inflation Accounting Studies. *The Accounting Review* 52 (1): 216–221.

Wu, T.-F., C.-J. Lin, and R. C. Weng. 2004. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research* 5 (Aug.): 975–1005.

Yang, D., M. Vasarhelyi, and C. Liu. 2003. A note on the using of accounting databases. *Industrial Management & Data Systems* 103 (3): 204–210.

Appendix A
Compustat and 10-K data comparison tables

Table A.1: Fama/French 12 industry classification

Industry Variable	*Variable Description*	*Examples of Industries Included*
NoDur	Consumer non-durables	Food, Tobacco, Textiles, Apparel, Leather, Toys
Durbl	Consumer durables	Cars, Television Sets, Furniture, Household Appliances
Manuf	Manufacturing	Machinery, Trucks, Planes, Office Furniture, Paper, Printing
Enrgy	Energy	Oil, Gas, Coal Extraction and Products
Chems	Chemicals	Chemicals and Allied Products
BusEq	Business equipment	Computers, Software, and Electronic Equipment
Telcm	Telecommunications	Telephone and Television Transmission
Utils	Utilities	Water, Gas, Electricity Utilities
Shops	Shops	Wholesale, Retail, and Some Services (Laundries, Repair Shops)
Hlth	Health	Healthcare, Medical Equipment, and Drugs
Money	Money	Finance

Table A.2: Descriptive statistics of Compustat variables

Statement	*Variable*	*Variable Description*	Observation counts			Descriptive Statistics of Matched Data		
			Available	*Missing*	*Matched (Perc.)*	*Mean*	*St. Dev.*	*Median*
Balance Sheet	ACT	Current Assets	3,946	0	3,902 (99.11%)	1,073.53	4,237.43	132.31
	CH	Cash	4,848	0	4,791 (99.07%)	374.64	2,899.25	26.70
	RECTR	Receivables (Trade)	3,849	0	3,414 (88.93%)	536.88	6,365.74	42.93
	INVT	Inventories	3,374	0	2,435 (72.26%)	330.71	1,317.23	36.44
	PPENT	Property, Plant and Equipment	4,582	0	4,381 (95.84%)	1,213.73	6,164.46	40.05
	DPACT	Depreciation, Depletion and Amortization	3,978	0	2,122 (54.61%)	1,621.78	7,357.35	231.16
	GDWL	Goodwill	2,659	0	2,541 (95.71%)	991.02	4,162.30	80.75
	AT	Total Assets	4,958	42	4,932 (99.74%)	8,581.13	84,465.52	509.03
	LCT	Current Liabilities	3,978	0	3,880 (97.76%)	737.45	3,153.04	59.06
	AP	Accounts Payable (Trade)	4,862	0	4,591 (94.66%)	1,883.37	30,076.04	23.54
	DLTT	Long-Term Debt	3,586	0	2,726 (76.19%)	2,173.00	12,316.16	238.59
	LT	Liabilities	4,965	0	3,540 (71.60%)	8,411.89	93,457.13	255.79
	RE	Retained Earnings	4,860	0	4,462 (92.29%)	802.39	8,133.42	9.45
	REUNA	Retained Earnings (Unadjusted)	4,714	0	4,329 (92.46%)	876.40	8,526.66	8.51
	TEQ	Stockholders' Equity	4,980	0	4,856 (97.80%)	1,652.20	8,795.99	145.95

Available – observations present in Compustat; *Missing* – observations present in XBRL 10-K filings, but not in Compustat.

Matched – observations present both in Compustat and XBRL 10-K filings excluding erroneous XBRL observations.

Mean, St. Dev., Median – values of descriptive statistics based on Compustat values of the matched observations.

Mean and median values are measured in millions of U.S. dollars, except for per share numbers that are measured in U.S. dollars.

(continued on the next page)

Table A.2 : (continued from the previous page)

Statement	*Variable*	*Variable Description*	Observation counts			Descriptive Statistics of Matched Data		
			Available	*Missing*	*Matched (Perc.)*	*Mean*	*St. Dev.*	*Median*
Income Statement	REVT	Revenue	4,679	0	3,619 (77.73%)	3,489.00	15,996.30	359.45
	SALE	Sales/Turnover	4,679	0	2,410 (51.58%)	3,499.89	14,993.02	402.63
	IDIT	Interest and Related Income	1,973	0	1,600 (82.09%)	6.51	39.61	0.26
	COGS	Cost of Goods Sold	4,684	4	2,621 (56.11%)	2,129.08	10,312.62	164.34
	XAD	Advertising Expense	1,948	0	1,012 (52.16%)	124.25	497.74	5.15
	XINT	Interest and Related Expense	1,862	0	1,605 (86.85%)	147.38	688.97	16.10
	XRD	Research and Development Expense	1,861	0	1,601 (86.78%)	147.58	689.81	16.10
	GP	Gross Profit (Loss)	4,708	0	2,028 (43.08%)	783.94	3,418.41	97.46
	NI	Net Income (Loss)	4,972	34	4,920 (99.29%)	198.94	1,387.93	5.83
	EPSPI	Earnings Per Share (Basic)	4,657	0	4,435 (95.66%)	0.33	45.01	0.39
	EPSFI	Earnings Per Share (Diluted)	4,658	0	4,436 (95.67%)	0.33	45.00	0.39
CF Statement	OANCF	Operating Activities	4,967	0	4,910 (99.25%)	424.75	2,703.35	21.57
	FINCF	Financing Activities	4,882	0	4,826 (99.12%)	-225.77	9,147.81	0.01
	IVNCF	Investing Activities	4,793	0	4,727 (99.16%)	-186.48	9,372.46	-18.44
	CHECH	Cash and Cash Equivalents - Increase (Decrease)	4,909	58	4,841 (98.90%)	22.05	685.63	0.03

Available – observations present in Compustat; *Missing* – observations present in XBRL 10-K filings, but not in Compustat.

Matched – observations present both in Compustat and XBRL 10-K filings excluding erroneous XBRL observations.

Mean, St. Dev., Median – values of descriptive statistics based on Compustat values of the matched observations.

Mean and median values are measured in millions of U.S. dollars, except for per share numbers that are measured in U.S. dollars.

Table A.3: Difference statistics of Compustat and XBRL 10-K numbers

Statement	*Variable*	*Variable Description*	Observation counts				Discrepancy statistics			Difference between 10-K and Compustat		
				Discrepancies								
			Matched	*All (Perc.)*	*Material (Perc.)*	*Material Pop. Estim.*	*Mean*	*Median*	*St. Dev.*	*Mean*	*Median*	*St. Dev.*
BS	ACT	Current Assets	3,902	13 (0.33%)	7 (0.18%)	0.06%	8.27%	1.41%	16.13	-0.01%	0.00%	1.01
	CH	Cash	4,791	40 (0.83%)	18 (0.38%)	0.20%	86.65%	9.75%	239.18	0.61%	0.00%	22.98
	RECTR	Receivables (Trade)	3,414	580 (16.99%)	414 (12.13%)	10.86%	32.98%	18.75%	42.27	-3.05%***	0.00%	21.88
	INVT	Inventories	2,435	115 (4.72%)	100 (4.11%)	3.23%	35.04%	26.19%	28.35	-1.65%***	0.00%	9.64
	PPENT	Property, Plant and Equipment	4,381	213 (4.86%)	184 (4.20%)	3.53%	44.21%	30.19%	38.95	-2.11%***	0.00%	12.81
	DPACT	Depreciation, Depletion and Amortization	2,122	98 (4.62%)	81 (3.82%)	2.91%	44.70%	30.78%	39.77	-1.94%***	0.00%	12.68
	GDWL	Goodwill	2,541	18 (0.71%)	2 (0.08%)	0.01%	7.27%	0.68%	15.23	0.03%	0.00%	1.39
	AT	Total Assets	4,932	17 (0.34%)	12 (0.24%)	0.11%	164.69%	1.70%	647.02	0.54%	0.00%	38.10
	LCT	Current Liabilities	3,880	20 (0.52%)	10 (0.26%)	0.11%	3,273.67%	0.90%	14,605.82	16.84%	0.00%	1,048.76
	AP	Accounts Payable (Trade)	4,591	383 (8.34%)	312 (6.80%)	5.96%	622.75%	46.93%	6,199.09	48.48%**	0.00%	1,796.72
	DLTT	Long-Term Debt	2,726	628 (23.04%)	448 (16.43%)	14.81%	617.82%	11.53%	13,815.95	134.17%***	0.00%	6,632.49
	LT	Total Liabilities	3,540	50 (1.41%)	27 (0.76%)	0.46%	1,374.34%	0.88%	9,669.96	19.31%**	0.00%	1,149.35
	RE	Retained Earnings	4,462	3,222 (72.21%)	1,693 (37.94%)	36.25%	24.13%	2.94%	165.02	-0.83%***	0.00%	141.71
	REUNA	Retained Earnings (Unadjusted)	4,329	59 (1.36%)	33 (0.76%)	0.49%	106.82%	0.84%	357.43	-0.58%**	0.00%	43.21
	TEQ	Stockholders' Equity	4,856	213 (4.39%)	158 (3.25%)	2.69%	243.70%	6.01%	3,123.11	8.39%***	0.00%	654.56

Matched – observations present both in Compustat and XBRL 10-K filings excluding erroneous XBRL observations.
All (Perc.) – number (and percentage of matched observations) of discrepancy observations.
Material (Perc.) – number (and percentage of matched observations) of material discrepancy observations.
Material Pop. Estim – a 99% probability estimate of the minimum amount of material discrepancies in the *population* of *all* observations (yielded by the binomial test).
Discrepancy statistics – descriptive statistics of *absolute* relative value differences between Compustat and XBRL 10-K observations with discrepancies.
Difference between 10-K and Compustat – descriptive statistics of (non-absolute) relative value differences between all matched 10-K and Compustat observations.
***, **, * indicate significance of Wilcoxon's signed-rank test at 99%, 95%, and 90% levels respectively.

(continued on the next page)

Table A.3 : (continued from the previous page)

Statement	*Variable*	*Variable Description*	Observation counts				Discrepancy statistics			Difference between 10-K and Compustat		
				Discrepancies								
			Matched	*All (Perc.)*	*Material (Perc.)*	*Material Pop. Estim.*	*Mean*	*Median*	*St. Dev.*	*Mean*	*Median*	*St. Dev.*
IS	REVT	Revenue	3,619	345 (9.53%)	209 (5.78%)	4.91%	7.44%	0.80%	22.91	0.00%	0.00%	7.43
	SALE	Sales/Turnover (Net)	2,410	122 (5.06%)	85 (3.53%)	2.71%	14.42%	1.57%	27.42	-0.18%***	0.00%	6.95
	IDIT	Interest and Related Income	1,600	36 (2.25%)	8 (0.50%)	0.18%	1,895.10%	95.90%	9,367.97	32.16%	0.00%	1,414.48
	COGS	Cost of Goods Sold	2,621	2,229 (85.04%)	1,989 (75.89%)	73.89%	24.55%	5.78%	350.61	8.56%***	3.10%	324.00
	XAD	Advertising Expense	1,012	24 (2.37%)	12 (1.19%)	0.54%	2,475.41%	47.74%	9,888.73	58.35%**	0.00%	1,538.40
	XINT	Interest and Related Expense	1,605	55 (3.43%)	34 (2.12%)	1.37%	42.92%	8.04%	81.48	0.67%	0.00%	16.92
	XRD	Research and Development Expense	1,601	51 (3.19%)	31 (1.94%)	1.22%	42.07%	6.78%	84.12	0.75%	0.00%	16.64
	GP	Gross Profit (Loss)	2,028	1,756 (86.59%)	1,595 (78.65%)	76.45%	23.14%	8.28%	106.29	-6.28%***	-6.70%	101.03
	NI	Net Income (Loss)	4,920	45 (0.91%)	15 (0.30%)	0.15%	252.44%	0.90%	1,630.73	2.19%	0.00%	156.09
	EPSPI	Earnings Per Share (Basic)	4,435	132 (2.98%)	100 (2.25%)	1.77%	1,321.48%	36.11%	6,758.45	37.80%***	0.00%	1,183.24
	EPSFI	Earnings Per Share (Diluted)	4,436	133 (3.00%)	100 (2.25%)	1.77%	1,292.78%	33.53%	6,735.06	37.22%***	0.00%	1,182.72
CF	OANCF	Operating Activities	4,910	134 (2.73%)	80 (1.63%)	1.24%	43.63%	2.27%	222.85	0.05%**	0.00%	37.38
	FINCF	Financing Activities	4,826	40 (0.83%)	23 (0.48%)	0.28%	81.94%	4.59%	329.64	0.26%	0.00%	30.56
	IVNCF	Investing Activities	4,727	82 (1.73%)	49 (1.04%)	0.72%	119.20%	4.92%	642.18	-0.94%	0.00%	85.52
	CHECH	Cash and Cash Equivalents - Increase (Decrease)	4,841	119 (2.46%)	42 (0.87%)	0.59%	24.96%	1.93%	93.45	-0.36%	0.00%	15.10

Matched – observations present both in Compustat and XBRL 10-K filings excluding erroneous XBRL observations.
All (Perc.) – number (and percentage of matched observations) of discrepancy observations.
Material (Perc.) – number (and percentage of matched observations) of material discrepancy observations.
Material Pop. Estim – a 99% probability estimate of the minimum amount of material discrepancies in the *population* of *all* observations (yielded by the binomial test).
Discrepancy statistics – descriptive statistics of *absolute* relative value differences between Compustat and XBRL 10-K observations with discrepancies.
Difference between 10-K and Compustat – descriptive statistics of (non-absolute) relative value differences between all matched 10-K and Compustat observations.
***, **, * indicate significance of Wilcoxon's signed-rank test at 99%, 95%, and 90% levels respectively.

Table A.4: Discrepancy statistics by industries

Variable	*Statistics Description*	Industry statistics											
		NoDur	*Durbl*	*Manuf*	*Enrgy*	*Chems*	*BusEq*	*Telcm*	*Utils*	*Shops*	*Hlth*	*Money*	*Other*
ACT	Observation count	225	109	381	228	126	745	132	200	415	555	137	649
	Discrepancy percentage	0.00%	0.00%	0.52%	0.00%	0.79%	0.27%	0.00%	0.00%	0.24%	0.18%	0.73%	0.77%
	Median discrepancy	–	–	0.75%	–	4.67%	6.67%	–	–	0.00%	0.64%	0.44%	1.41%
CH	Observation count	222	112	375	227	125	741	132	185	408	551	1040	673
	Discrepancy percentage	0.00%	0.00%	0.27%	0.00%	1.60%	0.27%	0.00%	0.00%	0.25%	1.09%	1.92%	1.19%
	Median discrepancy	–	–	17.84%	–	0.07%	1.00%	–	–	9.94%	0.03%	24.68%	1.25%
RECTR	Observation count	206	101	353	184	116	684	124	171	335	407	191	542
	Discrepancy percentage	12.62%	18.81%	12.18%	28.26%	12.93%	10.96%	16.13%	45.03%	15.52%	8.85%	46.07%	14.21%
	Median discrepancy	5.28%	11.41%	6.92%	11.32%	4.49%	13.13%	7.80%	34.65%	15.19%	10.16%	77.12%	18.30%
INVT	Observation count	195	105	362	97	110	501	53	86	336	323	26	241
	Discrepancy percentage	1.54%	4.76%	8.29%	3.09%	4.55%	5.19%	5.66%	8.14%	0.60%	1.86%	7.69%	9.54%
	Median discrepancy	34.84%	17.27%	23.91%	61.62%	7.73%	28.69%	30.55%	24.04%	53.00%	6.39%	28.32%	56.74%
PPENT	Observation count	214	108	378	184	122	725	125	198	399	526	780	622
	Discrepancy percentage	5.14%	11.11%	5.03%	13.59%	3.28%	1.79%	2.40%	7.58%	3.51%	1.52%	3.72%	9.65%
	Median discrepancy	17.16%	36.69%	5.92%	94.21%	15.05%	28.01%	3.37%	7.70%	23.34%	5.65%	61.53%	59.39%
DPACT	Observation count	123	53	233	144	72	385	64	164	223	206	136	319
	Discrepancy percentage	1.63%	5.66%	2.58%	16.67%	0.00%	1.30%	1.56%	10.37%	5.83%	1.94%	2.94%	5.96%
	Median discrepancy	10.41%	5.84%	22.42%	96.67%	–	25.21%	14.07%	2.96%	8.11%	1.71%	90.30%	61.62%
GDWL	Observation count	134	67	259	66	71	477	100	33	253	223	462	396
	Discrepancy percentage	0.75%	0.00%	0.77%	0.00%	0.00%	0.42%	0.00%	3.03%	0.00%	1.35%	1.73%	0.25%
	Median discrepancy	0.10%	–	4.65%	–	–	2.47%	–	0.01%	–	17.83%	0.62%	0.71%
AT	Observation count	227	115	388	233	127	752	131	202	421	558	1083	695
	Discrepancy percentage	0.00%	0.00%	0.26%	0.00%	0.79%	0.27%	0.00%	1.49%	0.24%	0.36%	0.28%	0.58%
	Median discrepancy	–	–	0.34%	–	17.55%	1.50%	–	7.00%	0.01%	2.19%	0.35%	10.26%
LCT	Observation count	226	108	377	225	125	741	131	200	416	552	140	639
	Discrepancy percentage	0.00%	0.93%	0.27%	0.00%	0.00%	0.67%	0.00%	0.00%	0.24%	0.72%	0.00%	1.25%
	Median discrepancy	–	66.23%	0.15%	–	–	1.66%	–	–	0.24%	0.01%	–	2.50%
AP	Observation count	221	108	376	217	124	736	128	180	413	542	888	658
	Discrepancy percentage	8.14%	4.63%	4.26%	22.12%	12.10%	3.94%	11.72%	5.56%	8.23%	6.64%	10.92%	9.12%
	Median discrepancy	24.61%	37.58%	49.21%	34.44%	53.16%	78.22%	76.10%	26.64%	19.89%	121.69%	52.45%	69.40%

Observation count – number of matched observations in an industry.
Discrepancy percentage – percentage of discrepancy observations of matched observations in an industry.
Median discrepancy – value of the median *absolute* relative discrepancy between Compustat and 10-K in an industry.
For information about industry variables, please see Table A.1.

(continued on the next page)

Table A.4 : (continued from the previous page)

Variable	*Statistics Description*	Industry Statistics											
		NoDur	*Durbl*	*Manuf*	*Enrgy*	*Chems*	*BusEq*	*Telcm*	*Utils*	*Shops*	*Hlth*	*Money*	*Other*
DLTT	Observation count	153	72	275	162	89	335	103	194	295	278	360	410
	Discrepancy percentage	13.07%	6.94%	14.55%	12.96%	14.61%	18.51%	24.27%	15.98%	20.34%	17.63%	62.78%	18.54%
	Median discrepancy	11.03%	1.04%	9.32%	1.74%	9.01%	20.61%	3.59%	2.42%	7.66%	11.25%	17.12%	7.16%
LT	Observation count	146	71	232	140	84	532	101	57	254	393	1038	492
	Discrepancy percentage	0.00%	1.41%	1.29%	1.43%	0.00%	0.38%	0.00%	5.26%	0.79%	1.53%	1.83%	2.44%
	Median discrepancy	–	3.60%	0.51%	12.53%	–	1.31%	–	1.51%	0.13%	8.23%	0.41%	1.60%
RE	Observation count	210	110	376	196	115	717	125	150	382	514	956	611
	Discrepancy percentage	80.00%	78.18%	77.66%	51.53%	76.52%	74.20%	70.40%	79.33%	67.80%	60.12%	82.53%	63.99%
	Median discrepancy	5.44%	7.60%	7.95%	2.63%	10.43%	1.46%	1.86%	2.41%	1.62%	0.36%	4.71%	2.11%
REUNA	Observation count	210	110	376	194	114	715	127	25	382	514	955	607
	Discrepancy percentage	0.95%	0.00%	1.06%	0.52%	1.75%	0.98%	1.57%	0.00%	0.52%	1.75%	2.41%	1.15%
	Median discrepancy	1.67%	–	0.00%	0.01%	29.58%	1.67%	61.09%	–	6.18%	0.07%	1.41%	1.65%
TEQ	Observation count	225	115	387	217	124	749	134	172	414	561	1070	688
	Discrepancy percentage	1.33%	3.48%	2.84%	5.07%	5.65%	2.94%	8.96%	24.42%	4.11%	4.63%	2.99%	3.78%
	Median discrepancy	4.67%	24.40%	0.76%	22.10%	19.20%	8.19%	27.78%	2.19%	9.62%	18.39%	5.12%	9.54%
REVT	Observation count	212	110	380	151	118	686	115	121	396	433	406	491
	Discrepancy percentage	7.55%	3.64%	3.95%	16.56%	5.08%	3.35%	4.35%	4.13%	5.05%	4.39%	46.06%	4.07%
	Median discrepancy	1.82%	0.19%	0.26%	2.28%	0.84%	0.80%	3.03%	0.73%	0.56%	1.47%	0.74%	0.56%
SALE	Observation count	175	98	348	51	100	539	68	29	332	315	51	304
	Discrepancy percentage	8.57%	4.08%	4.31%	9.80%	5.00%	3.15%	5.88%	0.00%	5.42%	5.40%	13.73%	4.93%
	Median discrepancy	2.17%	11.73%	0.26%	5.14%	0.43%	1.92%	1.89%	–	1.43%	2.34%	6.61%	0.68%
IDIT	Observation count	92	42	179	87	55	299	49	48	148	257	51	293
	Discrepancy percentage	0.00%	4.76%	1.12%	1.15%	0.00%	3.01%	2.04%	0.00%	2.70%	1.95%	13.73%	1.71%
	Median discrepancy	–	166.21%	135.37%	0.30%	–	73.33%	125.98%	–	200.00%	19.28%	68.01%	187.53%
COGS	Observation count	189	106	361	34	109	663	84	25	323	337	63	327
	Discrepancy percentage	88.89%	96.23%	94.74%	73.53%	90.83%	91.40%	58.33%	84.00%	70.59%	89.61%	87.30%	70.95%
	Median discrepancy	4.56%	4.18%	4.30%	11.65%	4.79%	7.61%	12.41%	17.21%	2.75%	12.27%	30.32%	6.94%
XAD	Observation count	76	26	53	2	19	155	34	0	137	56	349	105
	Discrepancy percentage	3.95%	0.00%	0.00%	0.00%	0.00%	0.00%	2.94%	–	4.38%	3.57%	2.29%	3.81%
	Median discrepancy	110.25%	–	–	–	–	–	56.14%	–	31.55%	1.95%	22.36%	515.03%

Observation count – number of matched observations in an industry.
Discrepancy percentage – percentage of discrepancy observations of matched observations in an industry.
Median discrepancy – value of the median *absolute* relative discrepancy between Compustat and 10-K in an industry.
For information about industry variables, please see Table A.1.

(continued on the next page)

Table A.4 : (continued from the previous page)

Variable	*Statistics Description*	Industry Statistics											
		NoDur	*Durbl*	*Manuf*	*Enrgy*	*Chems*	*BusEq*	*Telcm*	*Utils*	*Shops*	*Hlth*	*Money*	*Other*
XINT	Observation count	55	64	190	9	75	581	20	3	25	452	22	109
	Discrepancy percentage	1.82%	0.00%	2.11%	0.00%	2.67%	1.89%	0.00%	0.00%	0.00%	7.30%	0.00%	3.67%
	Median discrepancy	97.21%	–	9.77%	–	35.48%	2.68%	–	–	–	11.33%	–	15.43%
XRD	Observation count	55	64	190	9	74	579	20	3	25	451	22	109
	Discrepancy percentage	1.82%	0.00%	2.11%	0.00%	1.35%	1.55%	0.00%	0.00%	0.00%	7.10%	0.00%	3.67%
	Median discrepancy	97.21%	–	9.77%	–	5.57%	1.66%	–	–	–	10.52%	–	15.43%
GP	Observation count	158	93	298	18	86	544	23	25	280	237	36	230
	Discrepancy percentage	91.14%	95.70%	95.64%	100.00%	88.37%	91.91%	52.17%	96.00%	69.64%	88.19%	91.67%	74.35%
	Median discrepancy	7.54%	10.29%	10.77%	40.77%	8.47%	7.30%	13.25%	73.15%	5.59%	7.89%	24.57%	12.18%
NI	Observation count	227	116	390	231	124	750	130	199	424	561	1072	696
	Discrepancy percentage	0.88%	0.86%	0.00%	2.16%	1.61%	0.27%	0.77%	1.51%	0.47%	0.89%	1.31%	1.15%
	Median discrepancy	1.01%	0.01%	–	0.60%	7.56%	3.70%	1.60%	0.11%	1.10%	4.53%	2.04%	1.18%
EPSPI	Observation count	203	101	366	202	117	697	116	107	376	528	1016	606
	Discrepancy percentage	1.97%	0.99%	1.91%	2.97%	7.69%	2.73%	4.31%	0.93%	2.13%	4.36%	2.36%	4.13%
	Median discrepancy	167.19%	34.78%	36.36%	66.40%	46.74%	60.18%	90.00%	747.37%	26.52%	116.55%	9.41%	33.33%
EPSFI	Observation count	203	101	366	202	117	697	116	107	376	528	1017	606
	Discrepancy percentage	1.97%	0.99%	1.91%	2.97%	7.69%	2.73%	4.31%	0.93%	2.13%	4.36%	2.46%	4.13%
	Median discrepancy	63.62%	14.49%	21.12%	66.40%	46.74%	60.18%	90.00%	747.37%	26.52%	116.55%	9.92%	33.33%
OANCF	Observation count	227	116	387	231	126	748	132	199	420	560	1071	693
	Discrepancy percentage	2.20%	0.00%	3.62%	3.03%	3.97%	2.81%	8.33%	0.50%	3.10%	2.14%	1.77%	3.75%
	Median discrepancy	0.57%	–	2.43%	1.11%	1.87%	1.62%	8.46%	2.29%	1.59%	10.15%	2.81%	4.46%
FINCF	Observation count	223	113	382	225	124	729	133	199	417	546	1055	680
	Discrepancy percentage	0.45%	0.00%	0.79%	0.44%	0.81%	0.82%	5.26%	0.50%	0.24%	0.55%	0.47%	1.62%
	Median discrepancy	0.01%	–	0.85%	156.82%	72.52%	0.73%	11.86%	1.15%	0.20%	34.01%	5.38%	10.99%
IVNCF	Observation count	217	112	378	223	123	716	128	197	403	526	1048	656
	Discrepancy percentage	1.38%	0.00%	1.32%	3.14%	4.07%	1.26%	7.81%	0.51%	1.24%	1.52%	1.15%	2.59%
	Median discrepancy	11.71%	–	158.80%	1.11%	0.73%	2.74%	17.63%	0.10%	6.75%	1.70%	5.05%	4.62%
CHECH	Observation count	224	115	379	225	124	747	130	190	414	554	1060	679
	Discrepancy percentage	6.25%	2.61%	3.69%	1.33%	2.42%	2.68%	6.92%	0.00%	2.66%	1.26%	1.13%	3.39%
	Median discrepancy	2.41%	23.27%	2.07%	0.19%	8.22%	1.39%	10.09%	–	3.01%	0.72%	0.62%	1.93%

Observation count – number of matched observations in an industry.
Discrepancy percentage – percentage of discrepancy observations of matched observations in an industry.
Median discrepancy – value of the median *absolute* relative discrepancy between Compustat and 10-K in an industry.
For information about industry variables, please see Table A.1.

Table A.5: Discrepancy statistics by industry and XBRL adoption phase

Industry	Phase I Filers					Phase II Filers					Phase III Filers				
		Discrepancies (as %)		Discrepancy Stats			Discrepancies (as %)		Discrepancy Stats			Discrepancies (as %)		Discrepancy Stats	
	Count	*All*	*Material*	*Mean*	*Median*	*Count*	*All*	*Material*	*Mean*	*Median*	*Count*	*All*	*Material*	*Mean*	*Median*
NoDur	443	12.87%	11.29%	18.64%	7.55%	1,057	11.26%	10.12%	14.76%	5.49%	3,973	11.55%	9.11%	47.96%	5.84%
Durbl	150	16.00%	14.67%	36.06%	12.19%	437	11.90%	10.76%	10.17%	5.97%	2,254	12.07%	10.20%	24.47%	6.52%
Manuf	1,001	11.99%	10.69%	53.71%	10.36%	2,267	11.65%	9.70%	13.71%	6.13%	6,507	12.36%	10.27%	30.00%	6.40%
Enrgy	761	11.17%	7.75%	22.18%	6.38%	606	10.73%	6.93%	26.71%	9.87%	3,247	7.61%	5.45%	269.44%	13.19%
Chems	333	12.91%	11.41%	16.06%	8.03%	601	11.98%	10.32%	23.08%	7.11%	2,201	11.86%	9.45%	460.46%	6.52%
BusEq	1,253	12.21%	9.50%	19.58%	6.92%	3,052	11.34%	8.52%	19.30%	6.70%	14,360	10.85%	8.37%	26.37%	5.94%
Telcm	402	11.19%	9.70%	24.05%	8.46%	533	10.88%	6.94%	27.02%	7.40%	2,073	8.97%	5.98%	103.03%	5.58%
Utils	927	9.71%	6.04%	14.66%	5.91%	1,306	12.56%	8.27%	26.85%	7.29%	1,436	8.98%	5.50%	35.72%	7.17%
Shops	837	9.32%	7.29%	12.98%	3.39%	2,042	10.28%	7.25%	13.04%	3.19%	6,963	9.91%	7.24%	57.11%	4.59%
Hlth	837	11.35%	9.20%	29.24%	7.15%	1,356	11.36%	7.96%	19.91%	6.53%	10,851	8.81%	6.02%	652.87%	6.67%
Money	1,241	14.59%	9.11%	25.41%	9.12%	2,827	11.78%	7.18%	388.03%	6.67%	13,533	9.13%	4.60%	118.35%	6.66%
Other	929	9.90%	7.75%	25.71%	10.98%	2,732	8.75%	5.60%	22.71%	5.26%	11,165	9.28%	6.57%	239.36%	7.41%

Count – number of matched observations for specific industry and XBRL adoption phase.

All – percentage of all discrepancy observations for specific industry and phase.

Material – percentage of material discrepancy observations for specific industry and phase.

Mean, Median – mean and median statistics of *absolute* relative value differences between Compustat and 10-K of discrepancy observations.

Table A.6: Analysis of deviance of discrepancy observations

Panel A: Analysis of deviance with both Phase and Revenue variables included

Term	Observation level analysis						Company level analysis					
	All discrepancies			Material discrepancies			All discrepancies			Material discrepancies		
	χ^2	*df*	$\mathbb{P}(>\chi^2)$	χ^2	*df*	$\mathbb{P}(>\chi^2)$	χ^2	*df*	$\mathbb{P}(>\chi^2)$	χ^2	*df*	$\mathbb{P}(>\chi^2)$
Revenue	4.36	1	0.037*	11.02	1	0.001***	2.44	1	0.118	5.15	1	0.023*
Industry	63.53	11	0.000***	54.25	11	0.000***	66.57	11	0.000***	49.18	11	0.000***
Phase	1.16	2	0.561	0.00	2	1.000	1.09	2	0.579	0.00	2	1.000
Statement	952.29	2	0.000***	971.23	2	0.000***	719.19	2	0.000***	712.13	2	0.000***
Industry×Phase	42.94	22	0.005**	44.20	22	0.003**	43.95	22	0.004**	44.40	22	0.003**
Industry×Statement	443.78	22	0.000***	358.55	22	0.000***	318.23	22	0.000***	262.87	22	0.000***
Phase×Statement	13.87	4	0.008**	10.44	4	0.033*	9.99	4	0.041*	6.37	4	0.173
Industry×Phase×Statement	76.41	44	0.002**	67.77	44	0.012*	67.21	44	0.014**	51.97	44	0.191

Panel B: Analysis of deviance with Phase variable excluded

Term	Observation level analysis						Company level analysis					
	All discrepancies			Material discrepancies			All discrepancies			Material discrepancies		
	χ^2	*df*	$\mathbb{P}(>\chi^2)$	χ^2	*df*	$\mathbb{P}(>\chi^2)$	χ^2	*df*	$\mathbb{P}(>\chi^2)$	χ^2	*df*	$\mathbb{P}(>\chi^2)$
Revenue	10.56	1	0.001**	22.97	1	0.000***	6.69	1	0.010**	11.32	1	0.001***
Industry	87.91	11	0.000***	97.52	11	0.000***	97.07	11	0.000***	100.24	11	0.000***
Statement	1641.67	2	0.000***	1559.81	2	0.000***	1516.50	2	0.000***	1390.22	2	0.000***
Industry×Statement	834.94	22	0.000***	569.67	22	0.000***	633.98	22	0.000***	426.36	22	0.000***

***, **, * indicate significance of tests at 99.9%, 99%, and 95% levels respectively.
In all models, type III of sum of squares was utilized

Table A.7: Analysis of deviance for simple main effects

Panel A: Analysis of deviance within industry groups

Industry	Observation level analysis				Company level analysis			
	Revenue		Statement		Revenue		Statement	
	χ^2	$\mathbb{P}(>\chi^2)$	χ^2	$\mathbb{P}(>\chi^2)$	χ^2	$\mathbb{P}(>\chi^2)$	χ^2	$\mathbb{P}(>\chi^2)$
NoDur	0.17	0.676	257.29	0.000***	0.05	0.826	254.23	0.000***
Durbl	0.23	0.635	172.66	0.000***	0.30	0.581	266.91	0.000***
Manuf	1.84	0.175	468.74	0.000***	0.36	0.546	512.82	0.000***
Enrgy	1.24	0.266	76.88	0.000***	0.57	0.451	50.41	0.000***
Chems	0.16	0.685	127.40	0.000***	0.00	0.994	135.44	0.000***
BusEq	0.18	0.670	827.61	0.000***	0.15	0.696	750.39	0.000***
Telcm	1.21	0.272	5.11	0.078*	1.33	0.250	3.82	0.148
Utils	0.51	0.474	187.15	0.000***	0.24	0.623	123.87	0.000***
Shops	3.64	0.056*	334.19	0.000***	1.96	0.161	341.73	0.000***
Hlth	0.08	0.774	459.01	0.000***	0.06	0.804	373.31	0.000***
Money	16.00	0.000***	764.24	0.000***	11.77	0.001***	533.46	0.000***
Other	0.00	0.993	245.35	0.000***	0.18	0.672	201.33	0.000***

Panel B: Analysis of deviance within financial statement type groups

Statement	Observation level analysis				Company level analysis			
	Revenue		Industry		Revenue		Industry	
	χ^2	$\mathbb{P}(>\chi^2)$	χ^2	$\mathbb{P}(>\chi^2)$	χ^2	$\mathbb{P}(>\chi^2)$	χ^2	$\mathbb{P}(>\chi^2)$
Balance Sheet	12.02	0.001***	327.62	0.000***	19.07	0.000***	498.23	0.000***
Income Statement	0.67	0.413	502.56	0.000***	1.80	0.180	621.15	0.000***
Cash Flow Statement	0.19	0.661	93.36	0.000***	0.15	0.694	52.08	0.000***

***, **, * indicate significance of tests at 99.9%, 99%, and 95% levels respectively.
In all models, type III of sum of squares was utilized

Appendix B

Compustat standardization and bankruptcy prediction models and tables

Altman's 1968 model

The original Altman's Z score (Altman 1968) is a multiple discriminant function of five financial ratios that were empirically found to be good predictors of bankruptcy for manufacturing companies. The function is defined as:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 0.999X_5,$$

where

X_1 = Working Capital / Total Assets,
X_2 = Retained Earnings / Total Assets,
X_3 = Earnings Before Interest and Taxes / Total Assets,
X_4 = Market Value of Equity / Book Value of Total Debt,
X_5 = Sales / Total Assets.

In the original study, the measure was used to define three zones of discrimination based on the values of Z score. Specifically, firms with scores lower than 1.81 fell into "bankrupt" zone, firms with scores higher than 2.99 fell into "non-bankrupt" zone, and firms with scores between 1.81 and 2.99 fell into "zone of ignorance".

Ohlson's 1980 model

Ohlson's O score (Ohlson 1980) is a logistic regression model that predicts corporate bankruptcy. In fact, Ohlson (1980) builds three types of models: Model 1 to predict bankruptcy within the first year of 10-K release, Model 2 to predict bankruptcy in the second year of 10-K release, and Model 3 to predict bankruptcy within two years of 10-K release. The models are defined as follows:

Model 1 :

$$-1.32 - 0.407 \cdot SIZE + 6.03 \cdot TLTA - 1.43 \cdot WCTA + 0.0757 \cdot CLCA - 2.37 \cdot NITA$$
$$- 1.83 \cdot FUTL + 0.285 \cdot INTWO - 1.72 \cdot OENEG - 0.521 \cdot CHIN,$$

Model 2 :

$$1.84 - 0.519 \cdot SIZE + 4.76 \cdot TLTA - 1.71 \cdot WCTA - 0.2970 \cdot CLCA - 2.74 \cdot NITA$$
$$- 2.18 \cdot FUTL - 0.780 \cdot INTWO - 1.98 \cdot OENEG + 0.4218 \cdot CHIN,$$

Model 3 :

$$1.13 - 0.478 \cdot SIZE + 5.29 \cdot TLTA - 0.99 \cdot WCTA + 0.0620 \cdot CLCA - 4.62 \cdot NITA$$
$$- 2.25 \cdot FUTL - 0.521 \cdot INTWO - 1.91 \cdot OENEG + 0.212 \cdot CHIN.$$

where

$SIZE$ = log(Total Assets / GNP price-level index),
$TLTA$ = Total Liabilities / Total Assets,
$WCTA$ = Working Capital / Total Assets,
$CLCA$ = Current Liabilities / Current Assets,
$OENEG$ = 1 if Total Liabilities exceed Total Assets, 0 otherwise,
$NITA$ = Net Income / Total Assets,
$FUTL$ = Operating Income / Total Liabilities,
$OENEG$ = 1 if Net Income was negative for the last two years, 0 otherwise,
$CHIN$ = $(\text{NI}_t - \text{NI}_{t-1})/(|\text{NI}_t| + |\text{NI}_{t-1}|)$, where NI_t and NI_{t-1} are the current and previous Net Incomes respectively.

The output of Ohlson's model, O score, can be converted to probability of company experiencing bankruptcy event in the future.

Table B.1: Original Altman's model. Descriptive statistics of all matched accounting variables, ratios, and Z scores (sample size n=5,015).

Variable	*Variable Description*	10-K data			Compustat data			Difference between 10-K and Compustat		
		Mean	*Median*	*St. Dev.*	*Mean*	*Median*	*St. Dev.*	*Mean*	*Median*	*St. Dev.*
AT	Total Assets	4003.897	705.991	12 209.882	4003.896	705.991	12 209.880	0.001	0.000	1.075
LT	Total Liabilities	2344.390	315.777	7458.889	2344.120	315.777	7458.695	0.269***	0.000	11.529
ACT	Total Current Assets	1512.429	294.809	4628.785	1512.407	294.809	4628.791	0.022	0.000	1.300
LCT	Total Current Liabilities	913.135	130.823	3114.309	913.120	130.823	3114.311	0.015	0.000	1.500
RE	Retained Earnings	935.450	46.733	6131.564	816.519	38.959	5868.113	118.931***	0.000	773.830
SALE	Sales	3674.849	632.787	11 182.025	3627.582	629.685	11 006.760	47.267	0.000	1223.591
EBIT	Earnings Before Interest and Tax	428.733	47.948	1785.721	429.429	50.831	1699.630	−0.696***	0.247	553.021
X_1	(ACT-LCT)/AT	0.269	0.255	0.361	0.269	0.255	0.361	0.000	0.000	0.016
X_2	RE/AT	−0.821	0.115	4.137	−0.835	0.104	4.137	0.013***	0.000	0.090
X_3	EBIT/AT	0.011	0.079	0.318	0.015	0.076	0.296	−0.004***	0.001	0.133
X_4	MVALUE/LT	5.408	2.372	13.843	5.407	2.371	13.843	0.001***	0.000	0.043
X_5	SALE/AT	1.066	0.867	0.851	1.063	0.865	0.833	0.004	0.000	0.090
Z	Original Altman's Z Score	3.519	3.181	10.286	3.510	3.139	10.286	0.009***	0.009	0.469
Zone	Original Altman's Zone of Discrimination	2.263	3.000	0.857	2.254	3.000	0.858	0.008***	0.000	0.182

All variable values and their differences are measured in billions of U.S. dollars. Ratios and Z scores are not scaled.

MVALUE denotes market value of equity. This value is obtained from CRSP dataset and is independent of Compustat and 10-K data sets.

***, **, * indicate statistical significance of mean difference at 99%, 95%, and 90% levels, respectively, as measured by Wilcoxon's signed-rank test.

Table B.2: Original Ohlson's model. Descriptive statistics of all matched accounting variables, ratios, and O scores (sample size n=3,449).

Variable	*Variable Description*	10-K data			Compustat data			Difference between 10-K and Compustat		
		Mean	*Median*	*St. Dev.*	*Mean*	*Median*	*St. Dev.*	*Mean*	*Median*	*St. Dev.*
AT	Total Assets	4689.856	786.644	13 903.811	4689.854	786.644	13 903.809	0.002	0.000	1.288
LT	Total Liabilities	2783.117	364.638	8464.851	2782.742	364.638	8464.625	0.375***	0.000	13.799
ACT	Total Current Assets	1741.348	297.776	5416.940	1741.311	297.776	5416.950	0.036	0.000	1.535
LCT	Total Current Liabilities	1068.544	144.994	3543.294	1068.521	144.994	3543.298	0.023	0.000	1.791
OIADP	Operating Income after Depreciation	467.802	41.484	1995.980	524.276	54.743	1998.451	−56.474***	−0.504	684.197
NI_t	Net Income	297.643	18.198	1448.904	297.620	18.198	1448.918	0.022	0.000	1.909
NI_{t-1}	Net Income for the previous period	280.824	22.554	1261.797	280.814	22.440	1261.793	0.011	0.000	1.970
SIZE	log(AT/GNP price-level index)	15.011	15.452	2.715	15.011	15.452	2.715	0.000	0.000	0.009
TLTA	LT/AT	1.859	0.520	39.274	1.859	0.520	39.274	0.000**	0.000	0.017
WCTA	(ACT-LCT)/AT	−1.004	0.203	39.227	−1.004	0.203	39.227	0.000	0.000	0.025
CLCA	LCT/ACT	4.897	0.523	74.211	4.897	0.523	74.211	0.000	0.000	0.047
OENEG	1 if LT > AT, 0 otherwise	0.100	0.000	0.300	0.100	0.000	0.300	0.000	0.000	0.024
NITA	NI_t/AT	−0.936	0.036	28.608	−0.936	0.036	28.608	0.000	0.000	0.007
FUTL	OIADP/LT	−0.283	0.109	4.224	−0.232	0.123	3.881	−0.051***	−0.002	0.617
INTWO	1 if $NI_t < 0$ and $NI_{t-1} < 0$, 0 otherwise	0.269	0.000	0.444	0.270	0.000	0.444	−0.001	0.000	0.024
CHIN	$(NI_t\text{-}NI_{t-1})/(\lvert NI_t\rvert+\lvert NI_{t-1}\rvert)$	−0.023	0.021	0.498	−0.023	0.021	0.497	0.000	0.000	0.021
O_1	Ohlson's Model 1 Score	8.237	−4.920	319.110	8.145	−4.947	319.096	0.093***	0.004	1.134
$CLASS_1$	Ohlson's Model 1 Class	0.138	0.000	0.345	0.133	0.000	0.340	0.005***	0.000	0.081
O_2	Ohlson's Model 2 Score	5.922	−4.741	271.378	5.811	−4.784	271.354	0.111***	0.005	1.346
$CLASS_2$	Ohlson's Model 2 Class	0.126	0.000	0.332	0.121	0.000	0.327	0.004***	0.000	0.074
O_3	Ohlson's Model 3 Score	9.709	−4.162	308.840	9.594	−4.192	308.817	0.114***	0.005	1.390
$CLASS_3$	Ohlson's Model 3 Class	0.167	0.000	0.373	0.162	0.000	0.369	0.005***	0.000	0.088

All variable values and their differences are measured in billions of U.S. dollars. Ratios and O scores are not scaled.

GNP price-level index is calculated as (Nominal GNP/Real GNP)*100. GNP values are obtained from FRED, Federal Reserve Economic Data, from the Federal Reserve Bank of St. Louis, and is independent of Compustat and 10-K data sets.

***, **, * indicate statistical significance of mean difference at 99%, 95%, and 90% levels, respectively, as measured by Wilcoxon's signed-rank test.

Table B.3: Descriptive statistics of bankrupt observations (sample size n=146).

Variable	*Variable Description*	10-K data			Compustat data			Difference between 10-K and Compustat		
		Mean	*Median*	*St. Dev.*	*Mean*	*Median*	*St. Dev.*	*Mean*	*Median*	*St. Dev.*
AT	Total Assets	1043.476	363.828	2113.252	1043.443	363.828	2112.913	0.033*	0.000	6.324
LT	Total Liabilities	1043.501	261.931	2570.754	1034.961	261.931	2546.720	8.541**	0.000	67.235
ACT	Total Current Assets	327.741	102.759	699.775	326.867	102.759	699.771	0.874	0.000	7.488
LCT	Total Current Liabilities	330.899	94.562	693.205	330.346	94.562	693.172	0.553	0.000	8.780
RE	Retained Earnings	−530.044	−232.775	1682.962	−574.919	−243.747	1627.287	44.874**	0.000	237.290
SALE	Sales	953.160	200.095	2110.238	947.715	200.096	2109.277	5.445	0.000	79.271
OIADP	Operating Income after Depreciation	−118.419	−38.800	227.397	−5.837	−12.193	133.978	−112.581***	−3.738	280.805
EBIT	Earnings Before Interest and Tax	−123.302	−42.610	264.093	−5.837	−12.193	133.978	−117.465***	−3.938	311.183
NI_t	Net Income	−174.946	−60.432	348.924	−175.474	−59.506	349.643	0.528	0.000	9.362
NI_{t-1}	Net Income for the previous period	−94.042	−21.822	229.083	−93.764	−22.744	228.331	−0.278	0.000	6.756
X_1	(ACT-LCT)/AT	−0.142	0.043	1.487	−0.151	0.042	1.486	0.009	0.000	0.093
X_2	RE/AT	−3.530	−0.845	10.807	−3.537	−0.845	10.806	0.007*	0.000	0.081
X_3	EBIT/AT	−0.432	−0.192	0.755	−0.318	−0.062	0.722	−0.115***	−0.025	0.265
X_4	MVALUE/LT	1.204	0.247	3.800	1.185	0.247	3.794	0.019	0.000	0.248
X_5	SALE/AT	1.009	0.709	1.090	1.021	0.709	1.078	−0.012	0.000	0.165
SIZE	log(AT/GNP price-level index)	14.407	14.747	1.821	14.411	14.747	1.819	−0.004	0.000	0.042
TLTA	LT/AT	1.072	0.849	1.572	1.074	0.855	1.570	−0.002	0.000	0.085
WCTA	(ACT-LCT)/AT	−0.142	0.043	1.487	−0.151	0.042	1.486	0.009	0.000	0.093
CLCA	LCT/ACT	1.936	0.893	3.368	1.952	0.905	3.365	−0.016	0.000	0.179
OENEG	1 if LT > AT, 0 otherwise	0.356	0.000	0.481	0.356	0.000	0.481	0.000	0.000	0.117
NITA	NI_t/AT	−0.491	−0.260	0.821	−0.491	−0.260	0.819	−0.001	0.000	0.035
FUTL	OIADP/LT	−0.647	−0.240	1.428	−0.522	−0.078	1.407	−0.125***	−0.023	0.245
INTWO	1 if $NI_t < 0$ and $NI_{t-1} < 0$, 0 otherwise	0.753	1.000	0.433	0.760	1.000	0.428	−0.007	0.000	0.083
CHIN	$(NI_t\text{-}NI_{t-1})/(\lvert NI_t\rvert+\lvert NI_{t-1}\rvert)$	−0.214	−0.196	0.568	−0.212	−0.198	0.566	−0.002	0.000	0.037

All variable values and their differences are measured in billions of U.S. dollars. Ratios are not scaled.
MVALUE denotes market value of equity. This value is obtained from CRSP dataset and is independent of Compustat and 10-K data sets.
GNP price-level index is calculated as (Nominal GNP/Real GNP)*100. GNP values are obtained from FRED, Federal Reserve Economic Data, from the Federal Reserve Bank of St. Louis, and is independent of Compustat and 10-K data sets.
***, **, * indicate statistical significance of mean difference at 99%, 95%, and 90% levels, respectively, as measured by Wilcoxon's signed-rank test.

Table B.4: Altman's model. Descriptive statistics of matched non-bankrupt observations (sample size n=146).

Variable	*Variable Description*	10-K data			Compustat data			Difference between 10-K and Compustat		
		Mean	*Median*	*St. Dev.*	*Mean*	*Median*	*St. Dev.*	*Mean*	*Median*	*St. Dev.*
AT	Total Assets	2539.985	610.014	6630.507	2540.037	610.014	6630.491	−0.052	0.000	0.673
LT	Total Liabilities	1415.503	211.452	3409.618	1414.770	211.452	3408.912	0.733	0.000	14.791
ACT	Total Current Assets	941.515	269.726	2333.134	941.514	269.726	2333.134	0.001	0.000	0.010
LCT	Total Current Liabilities	590.684	107.416	1771.342	590.692	107.416	1771.339	−0.007	0.000	0.087
RE	Retained Earnings	574.152	50.856	2412.042	549.967	52.216	2440.207	24.184***	0.037	135.888
SALE	Sales	3703.519	479.438	12 957.675	3695.081	479.438	12 955.554	8.439	0.000	173.684
EBIT	Earnings Before Interest and Tax	239.119	35.831	922.459	247.460	34.142	899.257	−8.341	0.044	171.940
X_1	(ACT-LCT)/AT	0.250	0.253	0.275	0.250	0.253	0.275	0.000	0.000	0.000
X_2	RE/AT	−0.934	0.168	6.387	−0.951	0.158	6.387	0.017***	0.000	0.113
X_3	EBIT/AT	−0.017	0.070	0.404	−0.022	0.072	0.401	0.005	0.000	0.186
X_4	MVALUE/LT	5.258	1.918	8.501	5.052	1.874	8.219	0.206	0.000	2.466
X_5	SALE/AT	1.105	0.758	1.207	1.091	0.758	1.112	0.014	0.000	0.186

All variable values and their differences are measured in billions of U.S. dollars. Ratios are not scaled.

MVALUE denotes market value of equity. This value is obtained from CRSP dataset and is independent of Compustat and 10-K data sets.

***, **, * indicate statistical significance of mean difference at 99%, 95%, and 90% levels, respectively, as measured by Wilcoxon's signed-rank test.

Table B.5: Ohlson's model. Descriptive statistics of non-bankrupt observations (sample size n=2,525).

Variable	*Variable Description*	10-K data			Compustat data			Difference between 10-K and Compustat		
		Mean	*Median*	*St. Dev.*	*Mean*	*Median*	*St. Dev.*	*Mean*	*Median*	*St. Dev.*
AT	Total Assets	2913.851	359.484	11 738.453	2913.851	359.484	11 738.453	0.000	0.000	0.022
LT	Total Liabilities	1730.724	136.783	7200.847	1730.708	136.783	7200.846	0.016***	0.000	0.322
ACT	Total Current Assets	1058.297	160.443	4066.631	1058.297	160.443	4066.631	0.000	0.000	0.053
LCT	Total Current Liabilities	643.582	73.004	2717.971	643.582	73.004	2717.971	0.000	0.000	0.027
OIADP	Operating Income after Depreciation	272.513	14.630	1507.214	306.634	19.386	1433.159	−34.121***	−0.016	656.316
NI_t	Net Income	174.908	5.935	1013.696	174.908	5.935	1013.695	0.000	0.000	0.026
NI_{t-1}	Net Income for the previous period	157.553	7.260	925.448	157.522	7.242	925.455	0.030	0.000	1.658
SIZE	log(AT/GNP price-level index)	14.335	14.668	2.710	14.335	14.668	2.710	0.000	0.000	0.000
TLTA	LT/AT	2.332	0.499	46.224	2.332	0.499	46.224	0.000	0.000	0.001
WCTA	(ACT-LCT)/AT	−1.448	0.222	46.170	−1.448	0.221	46.170	0.000	0.000	0.020
CLCA	LCT/ACT	6.448	0.508	87.298	6.447	0.508	87.298	0.001	0.000	0.036
OENEG	1 if LT > AT, 0 otherwise	0.107	0.000	0.309	0.107	0.000	0.309	0.000	0.000	0.000
NITA	NI_t/AT	−1.286	0.028	33.671	−1.286	0.028	33.671	0.000	0.000	0.000
FUTL	OIADP/LT	−0.432	0.094	4.944	−0.378	0.106	4.540	−0.055***	0.000	0.721
INTWO	1 if $NI_t < 0$ and $NI_{t-1} < 0$, 0 otherwise	0.304	0.000	0.460	0.304	0.000	0.460	0.000	0.000	0.020
CHIN	$(NI_t\text{-}NI_{t-1})/(\lvert NI_t\rvert+\lvert NI_{t-1}\rvert)$	−0.013	0.029	0.516	−0.012	0.029	0.516	0.000	0.000	0.024

All variable values and their differences are measured in billions of U.S. dollars. Ratios are not scaled.

GNP price-level index is calculated as (Nominal GNP/Real GNP)*100. GNP values are obtained from FRED, Federal Reserve Economic Data, from the Federal Reserve Bank of St. Louis, and is independent of Compustat and 10-K data sets.

***, **, * indicate statistical significance of mean difference at 99%, 95%, and 90% levels, respectively, as measured by Wilcoxon's signed-rank test.

Appendix C

Exploration and exploitation framework tables and figures

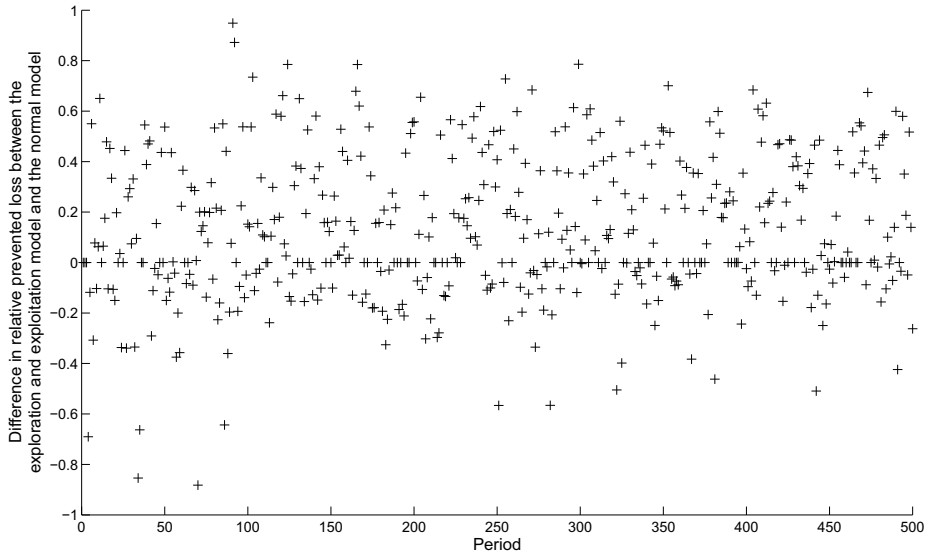
Table C.1: List and description of the variables in the credit card data set.

Variable	**Type**	**Description**
Age	Numerical	Age of a customer
Gender	Categorical	Gender of a customer
Income	Numerical	Value of declared income
BehaviorScore	Numerical	Banks internal behavior score
IsHouseOwner	Categorical	Indicates whether a customer is a house owner
AccAge	Numerical	Credit account age (in months)
NumPurachases	Numerical	Total number of purchases
NumCashWithdrawals	Numerical	Total number of cash withdrawals
NumLatePayments	Categorical	Total number of late payments
CreditLimit	Numerical	Credit limit
IsCanceled	Categorical	Indicates whether the account was canceled by the bank

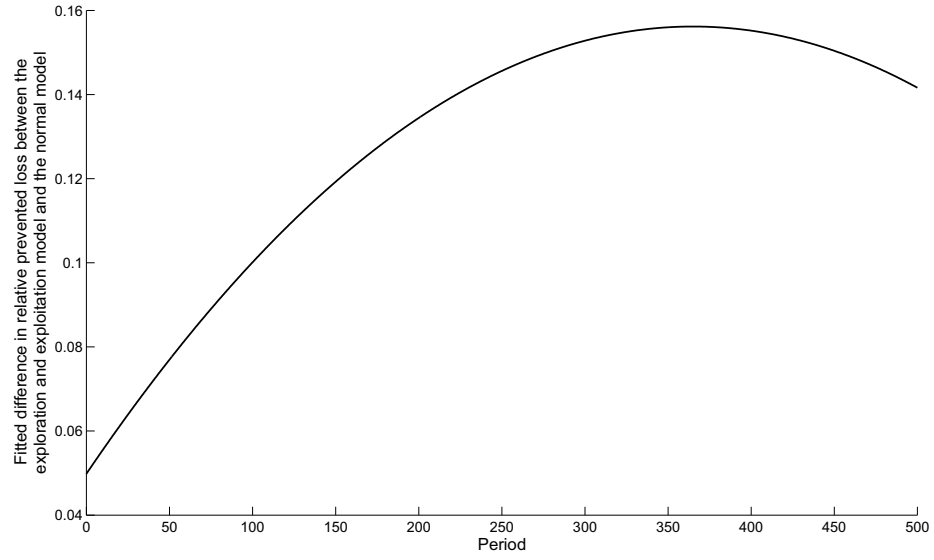
Table C.2: List and description of the variables in the census data set.

Variable	**Type**	**Description**
Age	Numerical	Age of a person
Education	Categorical	Level of education
ClassWorker	Categorical	Class of worker
Gender	Categorical	Gender of a person
MaritalStatus	Categorical	Marital status of a person
TaxFillerStatus	Categorical	Tax filer status
HouseHoldSummary	Categorical	Household summary variable
NumberPersWorkEmplyer	Numerical	Number of persons working for an employer
Citizenship	Categorical	Citizenship and origin (with respect to the U.S.)
OwnBussinesOrSelfEmployed	Categorical	Owns business or is self-employed
VeteranBenef	Categorical	Value of the veteran benefits
WeeksWorkedYear	Numerical	Weeks worked in year

Figure C.1: Period differences in relative prevented loss between the exploration and exploitation logistic model (with exploration coefficient $\rho = 0.5$) and the normal logistic model.Comparison methodology.



(a) Pointwise differences.



(b) Fitted differences.

Curriculum Vitae
Roman Chychyla

1987	Born in Chervonograd, L'vivska Oblast, Ukraine.
1994 - 2001	Attended Primary and Secondary School # 8 in Chervonograd, L'vivska Oblast, Ukraine.
2001 - 2004	Attended Chervonograd Gymnasium in Chervonograd, L'vivska Oblast, Ukraine.
2004 - 2009	Attended Ivan Franko National University of L'viv, L'viv, Ukraine.
2008	Bachelor of Mathematics, Ivan Franko National University of L'viv, L'viv, Ukraine.
2009	Master of Science in Statistics, Ivan Franko National University of L'viv, L'viv, Ukraine.
2009 - 2014	Attended Rutgers University, Newark, NJ.
2014	PhD in Accounting, Rutgers University, Newark, NJ.
2014 - present	Employed by the University of Miami, Coral Gables, FL - Visiting Assistant Professor of Accounting.