INTELLIGENT SYSTEMS IN ACCOUNTING, FINANCE AND MANAGEMENT
Intell. Sys. Acc. Fin. Mgmt. **15**, 41–56 (2007)
Published online in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/isaf.284

A GENETIC ALGORITHM APPROACH TO DETECTING TEMPORAL PATTERNS INDICATIVE OF FINANCIAL STATEMENT FRAUD

BETHANY HOOGS,* THOMAS KIEHL, CHRISTINA LACOMB AND DENIZ SENTURK
GE Global Research Center, Niskayuna, NY 12309, USA

SUMMARY

This study presents a genetic algorithm approach to detecting financial statement fraud. The study uses a sample comprising a target class of 51 companies accused by the Securities and Exchange Commission of improperly recognizing revenue and a peer class of 339 companies matched on industry and size (revenue). Variables include 76 comparative metrics, based on specific financial metrics and ratios that capture company performance in the context of historical and industry performance, and nine company characteristics. Time-based patterns detected by the genetic algorithm accurately classify 63% of the target class companies and 95% of the peer class companies. Copyright © 2007 John Wiley & Sons, Ltd.

1. INTRODUCTION

Financial statement fraud is a significant risk for external stakeholders, such as investors and commercial lenders in companies. Whereas financial decline is a well-researched phenomenon with commercially available credit scores and prediction tools, research on detecting financial statement fraud is comparatively new. This is especially true for research in fraud detection using publicly available data accessible to external stakeholders. However, the risk of loss to stakeholders is potentially much greater for financial statement fraud than for financial decline. The discovery of fraud often comes as a surprise effecting immediate and severe impacts on the value of the company (Karpoff and Lott, 1993; D'Agostino and Williams, 2002). Stakeholders that maintain large portfolios, such as commercial lenders, need automated solutions for detecting financial statement fraud to mitigate risk of loss. In order to be practical and acceptable to stakeholders, the detection techniques need to produce low levels of 'false alarms', be robust to real-world conditions, which include missing data, and ideally should be transparent, so that stakeholders can understand the basis on which classifications are made.

Financial statement fraud detection is challenging for several reasons. It is a rare event, with an estimated probability around 2% (Persons, 1995). As a result, sample sizes for training models are necessarily small. Existing research in detecting fraud using public indicators shows that there are no 'magic bullet' variables that conclusively identify fraud. Schilit (2002), for example, advocates a large number of indicators with varying applicability to different methods of fraud. Furthermore, evidence of fraud is intentionally disguised by the perpetrators.

* Correspondence to: B. Hoogs, GE Global Research Center, One Research Circle, Niskayuna, NY 12309, USA. E-mail: hoogsbk@research.ge.com

This study presents a genetic algorithm approach to detecting patterns in publicly available financial data that are characteristic of fraudulent financial reporting (Kiehl *et al.*, 2005). Patterns generated by this algorithm accurately classified 63% of alleged fraud companies and 95% of the non-fraud companies in our sample. Genetic algorithms have many strengths well suited to the problem of fraud detection. They provide the capability to learn class boundaries that are non-linear functions of multiple variables, allowing solutions that cannot be achieved by linear methods. They also provide efficient exploration of large search spaces (Mitchell, 1996), allowing use of a wide range of metrics and ratios used by financial analysts in assessing financial health. Furthermore, genetic algorithms perform explicit feature selection during learning. This is a significant benefit in fraud detection, as there is little consensus among experts about variables that consistently indicate fraudulent behaviour. The flexibility in design of genetic algorithms allows the incorporation of logic to handle missing data and the definition of output formats using problem-domain language to enhance user understanding of the results.

Owing to the criminal nature of financial statement fraud, the assertion of fraud for a company not committing fraud can have serious implications. When used by investors or commercial lenders, this type of false alarm from the model (i.e. false positive) will likely result in a lost investment opportunity. It will also reduce the model consumer's belief in the model and may eventually cause rejection of the model if the model 'cries wolf' more often than is acceptable. False positive rates in some existing fraud-detection models (e.g. Persons, 1995; Fanning and Cogger, 1998) are higher when correctly classifying a majority of fraud cases than might be acceptable to investors. Other models did not report false positive rates (Kwon and Feroz, 1996), or did not correctly classify a majority of the fraud companies in hold-out test samples (Kaminski *et al.*, 2004). Some of these models assume that the cost of misclassifying a fraud company (i.e. a false negative) is higher than the cost of a false positive (Persons, 1995; Kaminski *et al.*, 2004). This may be the case in situations such as audits, when failure to detect fraud may result in litigation against the auditor (Kaminski *et al.*, 2004). If the model is to be used to make unalterable investment or lending decisions, then the consumer of the model may consider a false negative less costly than a false positive. In that case, models are needed that can achieve moderate to high fraud-classification rates while keeping false positives at a minimum.

Most fraud-detection models take the form of logistic regression models (e.g. Persons, 1995; Lee *et al.*, 1999) and neural networks (Kwon and Feroz, 1996; Fanning and Cogger, 1998). Logistic regression and neural network models can be impractical in applications with high dimensionality and limited sample availability, requiring parsimonious variable selection that discards much of the available information. Logistic regression models and neural networks also have limited ability to handle missing values in training data, an endemic problem when using publicly available financial data. Techniques, such as list-wise deletion, that exclude all observations that have missing values in the independent variables radically reduce the number of available observations from which to construct the model (Myrtveit and Stensrud, 2001). Although advanced techniques are available to estimate missing entries, these techniques require the model consumers to accept the use of invented data in the model's development.

Another limitation of logistic regression models and neural networks is the inability to utilize time-varying independent variables unless time, itself, is an important predictor, as is the case in time-series models. In applications for which time is not a contributor to the dependent variable, such as predicting whether a company will be the target of a Securities and Exchange Commission (SEC) investigation, the independent variables used in statistical models are limited to static variables for a specific point in time. Some statistical fraud-detection models (e.g. Fanning and Cogger,

DOI: 10.1002/isaf

1998; Lee *et al.*, 1999) include independent variables that represent slopes, or changes in measures for a set of prior periods. However, slope measures are still restricted to a very limited number of metrics and time periods—typically one metric and two time periods. Other models (Persons, 1995; Kwon and Feroz, 1996; Kaminski *et al.*, 2004) utilized data from a single fiscal period.

Neural nets have a further limitation of not providing transparent results, offering little insight into the classification process. Given the limitations of neural networks and logistic regression models for applications characterized by high dimensionality of time-dependent variables, new techniques capable of capturing patterns across multiple metrics and across time are warranted.

The genetic algorithm presented here takes advantage of expanded information not exploited in existing research, including comparative views of financial metrics and ratios, and the relationships between these comparative metrics over time. The comparative metrics capture current company performance within the context of historical and industry performance. The patterns produced by our genetic algorithm comprise combinations of the comparative metrics across multiple fiscal periods, thus capturing multi-quarter interactions of context-driven performance metrics. The algorithm selects pattern variables from a set of 85 comparative metrics and company characteristics, covering a wide range of financial health indicators. Because the patterns consider multiple fiscal periods, there are multiple opportunities to detect indicators of fraud, making the patterns robust to occasional missing values in relevant metrics. Combinations of patterns in which each pattern captures the same type of behaviour as the other patterns, but uses different metrics, can also mitigate the impact of metrics that have missing values for specific subsets of the population. For example, if certain companies rarely report inventory, behaviour similar to inventory increases captured in the rest of the population can be captured for this subset using a related metric such as current assets. Finally, the patterns easily translate to financial domain terminology, offering complete transparency into classification logic.

In performing this study, we used Accounting and Auditing Enforcement Releases (AAERs) published by the SEC to define our sample of alleged fraudulent companies, following established practice in existing research (e.g. Fanning and Cogger, 1998; Lee *et al.*, 1999; Kaminski *et al.*, 2004). We diverged from the matched-pairs approach used in many studies (e.g. Persons, 1995; Kaminski *et al.*, 2004) and instead included multiple firms per alleged fraud sample, selected from the same industry and in the same size range. Lee *et al.* (1999) point out that the matched-pairs approach can overstate the importance of explanatory variables (see also Zmijewski (1984)). The use of multiple peer firms per fraud sample in this study reduces this bias. Our sampling approach more closely approximates the infrequency of fraud and, yet, still controls for financial metric variations due to size and industry.

In this study, we refer to the sample of companies drawn from the AAERs as fraud companies, consistent with established practice, but we recognize that many companies accused in the AAERs reached a settlement with the SEC, without admitting or denying the allegations. The patterns identified by our genetic algorithm detect companies that exhibit financial behaviours characteristic of companies accused of fraud by the SEC. We make no assertions as to whether companies matched by these patterns have actually committed fraud.

We referred to existing research to identify the set of financial metrics and ratios for which we generated industry and history comparative variables. These variables comprise 'exceptional anomaly scores' that capture the performance of the company, as indicated by specific metrics and ratios, in the context of historical and industry performance.

Our study is structured as follows. Section 2 discusses related research. Section 3 identifies and describes the variables that we used in performing the study. Section 4 describes the selection of

DOI: 10.1002/isaf

sample companies. Section 5 provides statistical analysis of company and data characteristics in the sample. Section 6 presents the genetic algorithm approach and results. Section 7 presents our conclusions.

2. RELATED WORK

There are few empirical studies of the ability to detect fraud using publicly available data. Most models designed to detect fraud rely on qualitative assessments, such as might be made by an auditor. For example, Bell and Carcello (1999) found that the factor ‘management lied to the auditor or was overly evasive when responding to audit inquiries’ was significantly related to the occurrence of fraud. However, outside investors typically cannot obtain this type of information and must rely, instead, on data in publicly available reports. We reviewed extant work in empirical models designed to detect financial statement fraud using publicly available data. These generally fell into two categories: regression/discriminant models and neural networks.

Persons (1995) estimated stepwise-logistic models for detection of fraud in the first year of fraud and the preceding year. The sample included 103 firms for the fraud-year sample and 100 firms for the preceding-year sample, identified by SEC enforcement actions, and an equal number of non-fraud firms matched on industry and time period. The fraud-year model correctly classified 47% of fraud firms, while misclassifying 14% of non-fraud firms, assuming a relative error cost of 30:1 of misclassifying fraud versus misclassifying non-fraud. The preceding-year model correctly classified 64% of fraud firms, while misclassifying 21% of non-fraud firms at the same relative error cost. Persons concluded that financial leverage, capital turnover, asset composition and firm size are significant factors in fraud detection.

Kwon and Feroz (1996) compared the results of a multilayered perceptron neural network model with a logit model trained on the 5 years up to and including the SEC investigation year for 35 firms identified by SEC AAERs and 35 control firms matched on industry, size and time period. Variables included profitability, sensitivity, difficult to audit, and going-concern ratios, and chief executive officer (CEO), chief financial officer (CFO) and auditor turnover events. Average classification accuracy of the neural network model was 88%, compared with 47% average classification accuracy using logit. They concluded that a multilayered perceptron can be powerful for detection problems with statistical uncertainty and that non-financial information increases predictive ability over using financial information alone.

Fanning and Cogger (1998) compared preliminary results of a neural network approach with those of stepwise logistic regression, linear discriminant analysis and quadratic discriminant analysis for detecting fraud in the first year of fraudulent filing charged by the SEC. They used a matched-pairs sample of 102 fraud companies identified in SEC AAERs and 102 non-fraud companies matched on industry, fiscal year end and company size. The neural net model included the following variables: percentage of outside directors on the board, having a non-Big Six auditor, growth, ratios of accounts receivable to sales, net plant property and equipment to assets, debt to equity, and trend variables for greater than 10% increase in accounts receivable and gross margin. It accurately classified 69% of the fraud companies, while misclassifying 20% of the non-fraud companies in their training data, and accurately identified 66% of the fraud companies, while misclassifying 41% of the non-fraud companies in their hold-out sample. They concluded that the neural network approach is effective in detecting fraud using publicly available information.

Lee *et al.* (1999) examined the difference between earnings and operating cash flow as an indicator of fraud in the year prior to the discovery of the fraud. Their sample included 56 fraud firms identified by SEC enforcement releases and the Wall Street Journal Index and 564 non-fraud firms

DOI: 10.1002/isaf

selected from the same industries and time periods as the fraud firms. Variables in their logistic regression model included earnings, earnings minus cash flow, leverage, sales growth, market return, retained earnings, market value, age, and indicator variables for auditor change, new securities issued, NYSE listing, AMEX listing and 3000 Standard Industry Classification (SIC) code group. The model correctly classified 73% of fraud firms, while misclassifying 10% of non-fraud firms at the 10% probability cut-off, and correctly classified 61% of fraud firms, while misclassifying 4% of non-fraud firms at the 20% probability cut-off. They concluded that there is a significant relationship between fraud and the earnings minus cash flow, and that earnings exceeding cash flow can be used to signal potential fraudulent reporting.

Peasnell *et al.* (2000) designed a cross-sectional model to estimate abnormal accruals and evaluated performance of the new model and cross-sectional versions of two established accruals models. They tested the ability of the three models to detect earnings management by artificially inducing income-increasing changes to working capital accounts of sample companies. All three models detected earnings management with close to 100% accuracy when the amount of simulated earnings management exceeded 5% of lagged total assets. The new model had higher performance in detecting induced expense manipulation than the established models, and the established models had higher performance on induced bad debt and revenue manipulation. They concluded that the most appropriate type of abnormal accruals model to use in detecting earnings management depends on the expected type of earnings manipulation, and that using the three models together offers the best opportunity to detect earnings manipulation.

Kaminski *et al.* (2004) performed an exploratory study examining the ability of financial ratios to detect fraud in the 7-year time range from 3 years prior to the first year of fraudulent filing to 3 years after. Their matched-pairs sample included 79 fraud firms identified in SEC AAERs and 79 non-fraud firms matched on size, time period and industry. Results of discriminant analysis using cross-validation correctly classified between 2% and 42% of the fraud firms, while misclassifying between 10% and 16% of the non-fraud firms. They conclude that financial ratios have limited ability to detect fraudulent financial reporting.

Our interest is in models to detect warning signs for financial statement fraud during the allegedly fraudulent fiscal periods, and possibly the prior year, which may exhibit motive for fraud. Early detection of fraud provides the best opportunity to mitigate losses on investments in potentially fraudulent companies. Three of the studies that we reviewed achieved fraud classification accuracies greater than 50% when trained on data during or prior to the alleged fraud period: Persons (1995), Fanning and Cogger (1998), and Lee *et al.* (1999). Models trained on data from later fiscal periods may rely on indicators that occur late, or subsequent to the fraudulent time period, reducing the model's usefulness in mitigating loss. Furthermore, models trained on time periods subsequent to the fraudulent quarters may detect indicators that are the result of discovery of the fraud, in essence detecting the detection of fraud by others, rather than detecting the fraud itself. Alternatively, these models may be detecting the signs of financial decline which often follow closely behind fraud, especially for those methods of fraud which cannot be sustained for an extended period of time.

The genetic algorithm presented in this study detected patterns that correctly classified 63% of the alleged fraud companies, while misclassifying 5% of the non-fraud companies. These rates are comparable to those published by Lee *et al.* (1999) at the 20% probability cut-off. To understand the comparative capability of the two models better, we implemented an approximation of the Lee *et al.* model to test on our sample set of companies. Classification rates of the approximated model on our sample were considerably lower than the published classification rates, correctly classifying 43–44% of the alleged fraud companies and misclassifying 22–26% of the non-fraud companies. In

DOI: 10.1002/isaf

Section 6 we discuss differences between the original model and our approximation, and potential reasons for the differences in results.

3. VARIABLES

Most of the variables suggested in extant work are financial metrics or ratios providing an isolated financial value or relationship at a point in time. We wished to examine the discriminatory power of financial metrics and relationships within the context of history and industry norms. Using the variables recommended by extant work as guidance, we defined comparative metrics, called exceptional anomaly scores, that capture company performance with respect to the company's history and with respect to the company's industry.

Senturk *et al.* (2004) established techniques to calculate exceptional anomaly scores for financial filing metrics that quantify the deviation of each metric from normal as defined by the company's past (called a *z*-within score) and the company's peers in the industry (called a *z*-between score). The exceptional anomaly scores are a variation of *z*-scores that are stable when used for small sample sizes, such as eight peers or 3 years of history. In general, an 'exceptional' technique may be defined as a technique for calculating a statistical value associated with a set of data and a target value, such that the target value is excluded from the calculation of the exceptional measurement. By using an exceptional technique, we prevent the particular target value within a group from skewing the measurements used to characterize that group. The exceptional anomaly scores highlight anomalous situations in which the company is reporting results that are significantly different from their past and/or significantly different from their peers in the industry.

The variables in our study comprise *z*-within and *z*-between exceptional anomaly scores for the financial metrics and ratios suggested in fraud-detection literature. Senturk and LaComb, in 'An application of quality function deployment to identify fraudulent financial statements' (unpublished working paper, 2004), ranked publicly available indicators recommended by various fraud-detection authors, using the number of authors recommending each indicator, author emphasis, and an assessment of each author's level of expertise in fraud detection as ranking criteria. This ranking highlights frequently recommended publicly available indicators, such as 'Cash from operations decreasing or not correlated with earnings' and 'Raw financials or ratios unusual for business (vertical analysis)'.

Work by Persons (1995), Kwon and Feroz (1996), Fanning and Cogger (1998), Lee *et al.* (1999), Peasnell *et al.* (2000), and Kaminski *et al.* (2004) suggest several financial metrics and ratios, including profitability, sales growth, accounts receivable/assets, accounts receivable/sales, accounts receivable growth, inventory/sales, plant property and equipment/assets, liabilities/assets, revenue/assets, current assets/assets, difference between earnings and operating cash flow, and size based on total assets. In an examination of several indexes defined by Beneish (1999), Wells (2001) suggests additional ratios, including asset quality and days' sales in receivables.

Several studies (Kwon and Feroz, 1996; Fanning and Cogger, 1998; Lee *et al.*, 1999) suggest event and company characteristic data, such as auditor changes, management changes and having the same person in the CEO and CFO role, in addition to financial data.

We calculated *z*-within and *z*-between scores for the financial metrics[1] and ratios presented in Table I. We also defined variables capturing age, size and profitability characteristics of

[1] Supplied by the Mergent Global Company Data Feed (www.mergent.com).

DOI: 10.1002/isaf

Table I. Financial metrics and ratios and characteristic variables

Variable	Description
AGE	Number of years since first filing available from data provider
AR	Accounts Receivable
AR_ADJ	AR/TOTA
AR_GROWTH	(AR − AR_PRIOR)/ABS(AR_PRIOR), where AR_PRIOR is the AR value in the prior fiscal year/quarter
AR_TOTCA	AR/TOTCA
AR_WO_TOTR	–AR_ZW – TOTR_ZW, for z-within, similar for z-between
ASSET_Q	1 – ((TOTCA + PPEN)/TOTA)
CCE	Cash and Cash Equivalents
CCE_ADJ	CCE/TOTA
CFFF	Cash Flow from Financing
CFFI	Cash Flow from Investing
CFFO	Cash Flow from Operations
CFFO_ADJ	CFFO/TOTA
CFFO_WO_NI	CFFO – NI
CFFO_WO_NI_TOTR	(CFFO – NI)/TOTR
CFFO_WO_TOTR	TOTR_ZW − CFFO_ZW, for z-within, similar for z-between
COG	Cost of Goods Sold
DAYS_SALES_OUTS	Days Sales Outstanding: [(QUARTER × 90) × AR]/TOTR
NI	Net Income
NI_ADJ	NI/TOTA
NI_TOTR	Net Profit Margin: NI/TOTR
OPEXP	Operating Expenses
OPEXP_ADJ	OPEXP/TOTA
OPINC_TOTR	Gross Profit Margin: (Earnings before Taxes + Other Income)/TOTR
PPEN	Plant Property and Equipment Net
PPEN_ADJ	PPEN/TOTA
PROFIT_CFFO_2YRS, PROFIT_CFFO_3YRS	# prior 4th quarters where Cash From Operations (CFFO) > 0 over prior (2 or 3) years/# prior 4th quarters where CFFO is not missing over prior (2 or 3) years
PROFIT_NI_2YRS, PROFIT_NI_3YRS	# prior 4th quarters where Net Income (NI) > 0 over prior (2 or 3) years/# prior 4th quarters where NI is not missing over prior (2 or 3) years
PROFIT_OPINC_2YRS, PROFIT_OPINC_3YRS	# prior 4th quarters where Operating Income (OPINC) > 0 over prior (2 or 3) years/# prior 4th quarters where OPINC is not missing over prior (2 or 3) years
SIZE_ASSETS	The average of the 4th quarter Total Asset (TOTA) values for the prior 3 years
SIZE_REVENUE	The average of the 4th quarter Total Revenue (TOTR) values for the prior 3 years
TOTA	Total Assets
TOTA_GROWTH	(TOTA − TOTA_PRIOR)/ABS(TOTA_PRIOR), where TOTA_PRIOR is the TOTA value in the prior fiscal year/quarter
TOTCA	Total Current Assets
TOTE	Total Equity
TOTE_ADJ	TOTE/TOTA
TOTL	Total Liabilities
TOTL_ADJ	TOTL/TOTA
TOTL_ADJ_INTAN	TOTL/(TOTA − Gross Intangibles)
TOTR	Total Revenue
TOTR_ADJ	TOTR/TOTA
TOTR_GROWTH	(TOTR − TOTR_PRIOR)/ABS(TOTR_PRIOR), where TOTR_PRIOR is the TOTR value in the prior fiscal year/quarter
WC_ADJ	WORKING_CAPITAL/TOTA
WORKING_CAPITAL	TOTCL − TOTCA

companies. These characteristic variables are also listed in Table I. For the nine characteristic variables (AGE, PROFIT_CFFO_2YRS, PROFIT_CFFO_3YRS, PROFIT_NI_2YRS, PROFIT_NI_3YRS, PROFIT_OPINC_2YRS, PROFIT_OPINC_3YRS, SIZE_ASSETS, SIZE_REVENUE) we included only the single raw value. For the 38 financial metrics and ratios, we included both the z-within and z-between values for each variable. This resulted in 85 variables. We evaluated auditor change events as a potential variable, but found insufficient numbers of events within the relevant time periods to warrant inclusion.

Other financial metric and ratio variables had missing exceptional anomaly score rates greater than 35% in our dataset. Metrics that are populated this infrequently are unlikely to contribute to successful discriminatory patterns, as they will restrict the applicability of the patterns to less than 65% of the data. Exceptional anomaly scores for the following metrics were excluded for this reason: accounts payable, debt, days' sales in inventory, inventory, and interest expense.

4. SAMPLE SELECTION

We reviewed AAERs published by the SEC between May 2002 and March 2004, to identify companies accused of fraudulent financial reporting. We restricted our review to those AAERs that mention violations of SEC Rule 10b-5 of the 1934 Securities Exchange Act. As Fanning and Cogger (1998) point out, the use of SEC enforcement releases results in a subsample of companies perpetrating financial statement fraud which may reduce the ability to generalize study results, as the subsample excludes non-public companies and fraudulent companies that have evaded detection. The 249 AAERs that we reviewed uniquely identified 122 companies accused by the SEC of employing improper accounting techniques. The Mergent Global Company Data Feed contained financial metric data for 101 of these companies.

We categorized the alleged fraud method(s) for 98 of the 101 companies based on our interpretation of the fraudulent acts alleged by the SEC in the relevant AAERs. We used categories based on the seven fraud techniques ('Shenanigans') described by Schilit (2002). The three companies excluded from the categorization were charged by the SEC with misappropriation of assets, which does not specifically align with any of the Shenanigans. The categorization showed that 29 companies employed multiple methods of fraud, and the remainder employed a single method. The frequency of methods employed were as follows: 67 companies used improper revenue recognition techniques (Shenanigans 1 and 2), six companies improperly used one-time gains to boost income (Shenanigan 3), 22 companies shifted expenses to a different time period (Shenanigan 4), 26 companies improperly omitted or reduced liabilities (Shenanigan 5), and three companies improperly shifted profits to a later time period (Shenanigans 6 and 7).

For our initial model we defined a target class comprising the 67 companies accused by the SEC of improper revenue recognition. Schilit (2002) recommends looking for different warning signs for different methods of fraud, and revenue recognition appears to be the most prevalent method in use. Analysis of fraud cases from 1987 to 1997 (Beasley *et al.*, 1999) found that over half involved premature or fictitious revenue recognition. Our categorization identified improper revenue recognition in 67% of the sample. Therefore, a specific model for detecting this method provides the highest impact in an overall fraud detection approach. This revenue recognition model can be augmented in the future with other specific models for additional fraud methods.

Analysis of missing data in the target class led to the removal of 16 additional companies, reducing the number of companies to 51. Our algorithm is robust to missing data, and we asserted

DOI: 10.1002/isaf

a liberal minimum data requirement that all target class companies must have at least 50% of one fiscal quarter of z-betweens and z-withins populated within the alleged fraud period. Available data for the 16 companies removed from the target class did not meet this minimum requirement, in many cases having 100% missing z-betweens or z-withins during the alleged fraud period.

The time periods included for target class companies comprise all fiscal quarters starting 1 year prior to the first fiscal quarter alleged to be fraudulent by the SEC, and ending on the last fiscal quarter of the alleged fraudulent period. Although it is understood that the fraud period alleged by the SEC may not be fully inclusive of all periods for which the company has falsified financial statements, it does represent those periods for which the SEC believed there was sufficient evidence to charge the company with fraud. It is during this period that we expect to identify anomalous financial results with respect to history and/or industry. The year prior to fraud was included to allow detection of multi-quarter patterns potentially starting with non-fraudulent conditions that subsequently descend into fraud.

The final target class dataset included 538 company-quarter observations for the 51 alleged fraudulent companies, ranging from 1991 to 2003, with the majority of the observations between 1997 and 2000. The duration of alleged frauds in the target class ranged between 1 and 20 quarters, with the majority lasting six quarters, and impacting two annual filings. The majority of the alleged fraudulent periods spanned the years 1998–2000.

We included in our sample a peer class consisting of up to eight peer companies for each target company. These are the same peers that were used in calculating the z-between measures. The peers were selected such that they are closest in size to the target company, based on total revenue, and are in the same SIC code. In ideal circumstances, four of the eight companies chosen are slightly smaller than the target company and the remaining four are slightly larger. In cases where a balanced set of peers could not be identified, such as when the target company is the largest or smallest company in the SIC, the closest set of peers was selected even though they may not necessarily be balanced in relation to the target company. The resultant peer class comprised 339 companies in the SIC codes and size ranges represented in the target class.

It is important to note that the peer class companies were not examined with respect to SEC enforcement actions, other than excluding companies already in the target class. Thus, our peer class may include companies charged with fraud by the SEC outside of the time range of the AAERs that we reviewed. As with any sampling of public companies, the peer class may also include companies committing financial statement fraud that has not been detected or for which a formal investigation has not been announced by the SEC.

We utilized a peer class that is considerably larger than the target class. This provides a means by which we can mitigate the effect of overfitting the target class and ensure that the results can be generalized more accurately to a wider population than would be the case if a matched-pairs dataset were used (Weiss and Hirsh, 2000).

The time periods included for peer class companies comprise all fiscal quarters starting with the first quarter of 1998. We selected 1998 as a cut-off date to maximize coverage of the alleged fraud periods in our target class, but also reduce coverage of years prior to the issuance of Statement of Accounting Standards No. 82 (AICPA, 1997), and Staff Accounting Bulletin 101 (SEC, 1999). These standards likely had a significant impact on the accounting practices of many companies, and we believe that a high representation in the peer class of fiscal quarters prior to their issuance would reduce detection of discriminating factors that are effective in the current operating environment.

An alternative candidate cut-off date of 2003 would restrict the peer class time periods to fiscal quarters subsequent to the issuance of Statement of Accounting Standards No. 99 (AICPA, 2002),

DOI: 10.1002/isaf

which became effective for audits of financial statements for periods beginning on or after December 15, 2002. However, restriction of the peer-class time periods to 2003 and beyond would have resulted in a calendar period too far outside the time range of the target class, and would likely convolute the differences related to fraud with those related to time.

The peer class dataset included 7241 company-quarter observations for the 339 companies, ranging from 1998 to 2004, with the majority of the observations between 1998 and 2003.

5. DESCRIPTIVE STATISTICS

We examined defining characteristics, including age, size, profitability and SIC category, in our target and peer classes. We also analysed the levels of missing data, as this is pervasive in our sample and may be a convoluting factor in the discrimination between target and peer classes.

Table II shows descriptive statistics for company characteristics in the target and peer classes. For the purposes of our analysis, age is calculated as the number of years between the current observation and the first filing offered by the data provider. For descriptions and formulas for the other variables presented, see Table I.

Age, size based on total revenue and total assets, and all three profitability measures are statistically different between the target and peer classes. Companies within the peer class are older than target companies, make less revenue, and have fewer assets than companies in the target class. Peer companies are also less likely to be profitable in terms of net income and operating income, but more likely to be profitable in terms of cash from operations.

Table III shows the distribution of SIC codes among target and peer class companies. Since the Business Services (73) industry makes up a large proportion of our target companies, an equal representation was given in the peer class through the peer selection process. By ensuring equal

Table II. Descriptive statistics for company characteristics in the target and peer classes

Variable	*N*	Mean	StdDev	Missing	Minimum	Maximum
Target companies (N = 51)						
AGE	538	5.25	1.82	0	1	11
SIZE_ASSETS	538	2,335,323,071	6,938,791,825	0	1,372,379	38,095,666,667
SIZE_REVENUE	538	1,787,197,251	5,344,994,521	0	448,500	34,087,666,667
PROFIT_NI_2YRS	538	0.62	0.41	0	0	1
PROFIT_NI_3YRS	538	0.61	0.38	0	0	1
PROFIT_OPINC_2YRS	538	0.64	0.42	0	0	1
PROFIT_OPINC_3YRS	538	0.63	0.39	0	0	1
PROFIT_CFFO_2YRS	538	0.58	0.41	0	0	1
PROFIT_CFFO_3YRS	538	0.59	0.38	0	0	1
Peer companies (N = 339)						
AGE	7,241	6.63	2.30	0	0	12
SIZE_ASSETS	7,214	1,683,549,791	9,049,913,679	27	0	192,303,666,667
SIZE_REVENUE	7,241	1,203,706,996	4,751,041,914	0	0	61,037,333,333
PROFIT_NI_2YRS	7,232	0.54	0.44	9	0	1
PROFIT_NI_3YRS	7,236	0.55	0.40	5	0	1
PROFIT_OPINC_2YRS	7,241	0.51	0.44	0	0	1
PROFIT_OPINC_3YRS	7,241	0.52	0.41	0	0	1
PROFIT_CFFO_2YRS	7,237	0.63	0.42	4	0	1
PROFIT_CFFO_3YRS	7,237	0.63	0.40	4	0	1

Table III. Distribution of SIC codes among target and peer class companies

Target companies (N = 51)			Peer companies (N = 339)		
SIC	Frequency	Percent	SIC	Frequency	Percent
13	1	1.96	13	6	1.77
23	2	3.92	23	16	4.72
28	1	1.96	28	8	2.36
34	1	1.96	34	5	1.47
35	6	11.76	35	39	11.5
36	5	9.8	36	33	9.73
38	4	7.84	38	28	8.26
47	1	1.96	47	7	2.06
48	2	3.92	48	8	2.36
49	2	3.92	49	14	4.13
50	2	3.92	50	15	4.42
51	1	1.96	51	5	1.47
56	1	1.96	56	6	1.77
73	21	41.18	73	141	41.59
87	1	1.96	87	8	2.36

Table IV. Data population rates for the target and peer companies

Data population rate (%)							
Target companies (N = 51)				Peer companies (N = 339)			
Populated	z-within	z-between	Overall	Populated	z-within	z-between	Overall
<5	0.0	0.0	0.0	<5	1.2	1.2	0.0
<15	7.8	2.0	0.0	<15	3.5	2.4	0.3
<25	21.6	2.0	2.0	<25	7.1	3.2	0.3
<35	23.5	3.9	2.0	<35	15.0	3.2	0.9
<45	33.3	5.9	5.9	<45	22.7	3.2	4.7
<55	41.2	7.8	19.6	<55	26.8	3.8	10.6
<65	47.1	7.8	29.4	<65	31.9	5.3	22.7
<75	49.0	7.8	41.2	<75	36.9	8.6	33.3
<85	58.8	11.8	49.0	<85	46.0	22.4	41.3
<95	78.4	33.3	76.5	<95	59.6	53.4	67.8
≤100	100.0	100.0	100.0	≤100	100.0	100.0	100.0

distribution of the industry segments between the classes, we can mitigate the impact of sampling bias due to industry representation within the alleged fraudulent companies.

Table IV shows the data population rates for the target and peer companies for each of three views: considering only z-withins, considering only z-betweens, and considering all exceptional anomaly scores. Overall, data are more frequently missing in the target class than in the peer class, due primarily to the frequency of missing z-within scores, which require 3 years of prior values for calculation. The difference in data population rates is due to the different time periods represented in the target and peer companies. The financial periods represented by target companies are more likely to be farther in the past than the peer companies, for which the fiscal time period is 1998–2004. Higher rates of missing data within the target population make the task of discriminating allegedly fraudulent companies from their peers more challenging, since there are fewer opportunities for detection.

DOI: 10.1002/isaf

6. METHOD AND RESULTS

We implemented a genetic algorithm to detect patterns that discriminate allegedly fraudulent companies from their peers (Kiehl *et al.*, 2005). The algorithm considers combinations of variables and the interactions of variables across time when generating candidate patterns. Genetic algorithms, genetic programming and various other flavours of evolutionary computation have long been used in rule induction and classification tasks. These applications include evolutionary computation approaches to tuning fuzzy rule sets (Herrara *et al.*, 1995) or neural nets (Montana and Davis, 1989). Genetic programming is a natural fit for free-form evolution of regular expressions. Our method has much in common with the genetic programming application that evolved PROSITE expressions to discover motifs in proteins (Koza *et al*., 1999).

Although our system has some similarities to genetic programming approaches, the representation is explicitly constrained to a set number of clauses in a predetermined relationship to each other. In general, these types of system can be described as 'genetics based learning systems' (Goldberg, 1989). Our system follows the 'Pitt' approach as described by Goldberg (1989), where each individual in the population is a self-contained classifier. This is in contrast to the Michigan approach that uses the population in its entirety as the classifier.

Little work has been done in the relationship of fraud factors across time. Extant work includes the time dimension only as individual time-based variables, such as the Growth variable used by Fanning and Cogger (1998) capturing growth rate of sales over 3 years. The inclusion of a time dimension allows discovery of temporal sequences of indicators, such as increasing accounts receivable followed by growing divergence between net income and cash flow from operations, which may provide greater discriminatory power than single-time-period combinations of variables. It also affords greater freedom in the selection of time periods for target companies, allowing the inclusion of all allegedly fraudulent periods for a company, from a single quarter to multiple years. Conversely, the static time periods used by other models potentially exclude relevant data.

The genetic algorithm evolves patterns comprising combinations of multiple metrics across multiple quarters, guided by a fitness function that rewards patterns that discriminate well between the target and peer companies. Each company is represented by a data structure containing all of the metrics for all of the relevant quarters, in time-ordered sequence. For the alleged fraud companies in our sample, these structures contain the quarterly metric values for the alleged fraud period and the preceding four quarters. For the peer companies, these structures contain the quarterly metric values from 1998 to 2004. The patterns evolved by the genetic algorithm consist of multiple phrases, each of the form '*n* out *m* quarters of metric *i* are *operator* than *threshold*' (Kiehl *et al.*, 2005). The algorithm evaluates the patterns against the company data structures in a sliding-window approach starting at the earliest quarter available, through the last, with a window width of *m*. The parameter *m* is a predefined maximum number of quarters allowed in a pattern phrase. A match between the pattern and the company data structure at any time during the quarters represented in the data structure constitutes a match, resulting in the company being labelled as potentially fraudulent by the algorithm. Thus, the detection is performed at the company level rather than at the quarter level.

Each execution of the genetic algorithm results in a single pattern that has the best discriminatory ability achieved over multiple generations of evolution by the algorithm. Owing to the stochastic nature of genetic algorithms, each execution will most likely result in a different solution. Multiple patterns from multiple executions of the genetic algorithm against the same dataset can be combined to achieve potentially greater coverage of the dataset than is provided by a single pattern. In these combinations, a match against any single pattern is considered a match for the combination.

DOI: 10.1002/isaf

We provided the genetic algorithm all 85 *z*-within, *z*-between and company characteristic variables. A strength of genetic algorithms is efficient parallel exploration of large search spaces (Mitchell, 1996). We took advantage of this strength, allowing the algorithm to select any combination of up to four variables across as many as five quarters in defining candidate patterns. The genetic algorithm employs an elitist steady-state approach, with uniform crossover, tournament selection, $P(\text{crossover}) = 0.5$ and $P(\text{mutation}) = 0.05$ (Kiehl *et al.*, 2005).

We conducted multiple experiments executing the genetic algorithm on the sample, deriving hundreds of candidate solutions. During execution, our algorithm randomly removes a predefined number of target and peer companies as a hold-out sample, performs a predefined number of runs on the remaining training set, and tests the resulting pattern from each run against the hold-out sample. We used hold-out sample sizes of approximately 30%. The output patterns are characterized by the number of matches against the training set target companies, training set peer companies, hold-out target companies, and hold-out peer companies. We then analysed combinations of candidate patterns meeting threshold classification rates to identify pattern combinations that maximize the overall target classification (true positive) rate while maintaining a low overall peer misclassification (false positive) rate.

Through these experiments we achieved a set of three patterns that correctly classifies 63% of the target companies, while misclassifying 5% of the peers. Table V shows classification accuracy, as a percentage of target and peer companies correctly classified, of the individual patterns and of the combined pattern set. The individual patterns correctly classify between 20% and 39% of the target companies, while misclassifying 3% or fewer peers. We note that the classification accuracy for target companies was significantly lower on the hold-out sample, although peer classification rates remained more stable. This is a common problem when working with small datasets, and may reduce the ability of the patterns to generalize to other datasets. There is considerable overlap in the companies that the individual patterns match, but there is enough distinction among the patterns that they can successfully be combined to provide greater overall coverage of the alleged fraud sample, while still maintaining a low peer misclassification rate.

In addition to providing good classification results, the patterns are easily explainable. For example, pattern 1 asserts the following concurrent conditions: at least one out of two quarters of CFFF_Z-Between < –0.4; at least one out of three quarters of NI_Z-Within > 2; at least one out of two quarters of CFFO_WO_NI_Z-Between < –1.5; at least one out of two quarters of TOTCA_Z-Between > 0.9. Each of the phrases of the pattern must be true concurrently to cause a match. Understanding that the thresholds in the asserted conditions are exceptional anomaly scores representing a concept similar to *z*-scores, this rule can be described very straightforwardly: CFFF is higher than company's peers, NI is growing significantly, CFFO minus NI is significantly worse than peers, and TOTCA is higher than peers.

Table V. Classification accuracy

	Classification accuracy (%)					
	Overall		Training set		Hold-out set	
	Target	Peer	Target	Peer	Target	Peer
Pattern 1	39	97	44	97	27	98
Pattern 2	37	97	42	97	27	96
Pattern 3	27	99	32	98	13	100
Combined	63	95				

DOI: 10.1002/isaf

The patterns are also effective even when constituent metrics have missing data within the relevant time period. For example, one target company successfully matched by pattern 1 had 38% missing values for three of the metrics used in pattern 1 and 57% missing values for the fourth metric during the alleged fraud period and preceding year.

The combined performance of the patterns is comparable to the classification rates published by Lee *et al.* (1999) at the 20% probability cut-off. To understand the comparative capability of the two models better, we implemented an approximation of the Lee *et al.* model to test on our sample set of companies. Our approximated model included the same raw financial variables in the Lee *et al.* model, with the exception of market return, retained earnings, auditor change, and new equity securities, as we were unable to obtain these variables for the companies in our dataset. The approximated model correctly classified 43% of the allegedly fraudulent companies and misclassified 21% of the peers at the 20% probability cut-off. At the 40% probability cut-off, the approximated model achieved 22% correct classification of alleged fraud companies and misclassified 5% of the peers. Potential contributors to the difference between published results and the results against our data include:

1. *Time periods*. We executed the model on all fiscal quarters spanning 1 year prior to the alleged fraud period through the fraud period and considered a company to be classified as fraudulent if any of these quarters were classified as fraudulent by the model, whereas Lee *et al.* used a single observation of the year prior to fraud discovery.
2. *Variables*. We were unable to include the market return, retained earnings, auditor change, and new securities issued variables in the approximated model (only the new securities variable was significant in their model).
3. *Missing data*. Five of the 51 alleged fraud companies and 53 of the 339 peers could not be scored at all during the relevant period due to missing data.

We also tested the approximated model using a single observation of the final annual filing during the fraud period to attempt to mitigate differences due to time periods, achieving 44% classification of the alleged fraud companies and 26% misclassification of the peers at the 20% probability cut-off.

Oversampling of a rare event, such as is the case with studies that utilize a matched pair design (1:1 sampling) or a 2:1 oversampling design (2:1 sampling), can lead to choice-based sample bias (Zmijewski, 1984). This is because the proportion of the rare event in the sample is considerably higher than would be expected in the population. Although the proportion of target and peer companies in our sample is not representative of the true rate, we have included as many peer firms as possible in our sample while still representing companies of similar size and industry to those in our target sample. This reduces the risk of artificially inflated classification accuracy. However, classification accuracy may be artificially inflated for any study where the proportion of target companies in the sample does not reflect the expected rate in the population.

Each company in our sample needed to have at least 3 years of history available for the calculation of the exceptional anomaly scores. This may result in survivorship bias, the tendency for failed companies to be excluded from the sample because they did not survive at least 3 years. This may also result in a somewhat inflated classification accuracy of the target companies.

Lastly, there may be bias due to the difference in time periods used in the study for target and peer companies. As mentioned previously, the time periods included for peer class companies comprise all fiscal quarters starting with the first quarter of 1998 in order to mitigate anticipated changes to the reporting behaviour of non-fraudulent companies in response to the Statement of Accounting Standards No. 82 (AICPA, 1997) and Staff Accounting Bulletin 101 (SEC, 1999).

DOI: 10.1002/isaf

7. CONCLUSIONS

Based on the results presented in this study, we conclude that exceptional anomaly scores are valuable metrics for characterizing corporate financial behaviour, and that patterns considering the interactions of exceptional anomaly scores over time are effective in detecting potentially fraudulent behaviour. We further conclude that genetic algorithms are a successful technique for detecting discriminatory patterns in challenging domains characterized by high dimensionality and pervasive missing values. The patterns generated by the genetic algorithm are easily translated to domain-appropriate language and, therefore, easily understood by external stakeholders. Furthermore, the patterns are capable of identifying potentially fraudulent behaviour despite occasional missing values, and provide low false-positive rates, making them practical for use by external stakeholders.

The research presented in this study, which focused on improper revenue recognition, can be extended to other methods of financial statement fraud. The techniques presented here can also be extended to incorporate company events and other qualitative indicators that may provide additional discriminatory power. Additional research can also test the stability of learned fraud detection patterns over time by testing patterns on a sample selected from a later time-frame than the training sample.

REFERENCES

AICPA. 1997. *Consideration of Fraud in a Financial Statement Audit*. Statement of Auditing Standards No. 82. American Institute of Certified Public Accountants: New York, NY.

AICPA. 2002. *Considerations of Fraud in a Financial Statement Audit*. Statement of Auditing Standards No. 99. American Institute of Certified Public Accountants: New York, NY.

Beasley M, Carcello J, Hermanson D. 1999. *Fraudulent Financial Reporting: 1987–1997—An Analysis of U.S. Public Companies*. Committee of Sponsoring Organizations of the Treadway Commission (COSO): New York, NY.

Bell T, Carcello J. 2000. A decision aid for accessing the likelihood of fraudulent financial reporting. *Auditing: A Journal of Practice & Theory* **19**(Spring): 169–184.

Beneish M. 1999. The detection of earnings manipulation. *Financial Analysts Journal* (Sep.–Oct.): 1–11.

D'Agostino D, Williams O. 2002. *Financial Statement Restatements: Trends, Market Impacts, Regulatory Responses, and Remaining Challenges*. United States General Accounting Office (GAO): Washington, DC.

Fanning K, Cogger K. 1998. Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance and Management* **7**: 21–41.

Goldberg DE. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley: Reading, MA.

Herrara F, Lozano M, Verdegay JL. 1995. Tuning fuzzy logic controllers by genetic algorithms. *International Journal of Approximate Reasoning* **12**: 299–315.

Kaminski K, Wetzel T, Guan L. 2004. Can financial ratios detect fraudulent financial reporting? *Managerial Auditing Journal* **19**(1):15–28.

Karpoff J, Lott J. 1993. The reputational penalty firms bear from committing criminal fraud. *Journal of Law and Economics* **36**(2):757–802.

Kiehl T, Hoogs B, LaComb C, Senturk D. 2005. Evolving multi-variate time-series patterns for the discrimination of fraudulent financial filings. In *Proceedings (Late-Breaking Papers) of the Genetic and Evolutionary Computing Conference*, Washington, DC.

Koza JR, Bennett III FH, Andre D, Keane MA. 1999. *Genetic Programming III, Darwinian Invention and Problem Solving*. Morgan Kaufmann: San Francisco, CA.

Kwon T, Feroz E. 1996. A multilayered perceptron approach to prediction of the SEC's investigation targets. *IEEE Transactions on Neural Networks* **7**(5): 1286–1290.

Lee T, Ingram R, Howard T. 1999. The difference between earnings and operating cash flow as an indicator of financial reporting fraud. *Contemporary Accounting Research* **16**(4): 749–786.

DOI: 10.1002/isaf

Mitchell M. 1996. *An Introduction to Genetic Algorithms*. Massachusetts Institute of Technology: Cambridge, MA.

Montana DJ, Davis L. 1989. Training feedforward neural networks using genetic algorithms. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, Detroit, MI.

Myrtveit I, Stensrud E. 2001. Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods. *IEEE Transactions on Software Engineering* **27**(11): 999–1013.

Peasnell K, Pope P, Young S. 2000. Detecting earnings management using cross-sectional abnormal accruals models. *Accounting and Business Research* **30**(4): 313–326.

Persons O. 1995. Using financial statement data to identify factors associated with fraudulent financial reporting. *Journal of Applied Business Research* **11**(3): 38–46.

Schilit H. 2002. *Financial Shenanigans: How to Detect Accounting Gimmicks & Fraud in Financial Reports*, second edition. McGraw-Hill: New York, NY.

SEC. 1999. *Revenue Recognition in Financial Statements*. SEC Staff Accounting Bulletin: No. 101. Securities and Exchange Commission: New York, NY.

Senturk D, LaComb C, Doganaksoy M, Neagu R. 2004. Financial anomaly detection: a six sigma approach to detecting misleading financials and financial decline. In *ASA Joint Statistical Meetings*, Toronto, Canada.

Weiss G, Hirsh H. 2000. Learning to predict extremely rare events. In *Papers from the AAAI Workshop on Learning from Imbalanced Data Sets. Technical Report WS-00-05*. AAAI Press: Menlo Park, CA; 64–68.

Wells J. 2001. Irrational ratios – financial statements and ZZZZ Best fraud case. *Journal of Accountancy* (August): 80–83.

Zmijewski M. 1984. Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research* **22**(Supplement): 59–86.

DOI: 10.1002/isaf