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, 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.



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.