



FRAUD

M A G A Z I N E

A Publication of the Association of Certified Fraud Examiners

May/June 2005

Archive Issue: March/April 2004

[PRINTER FRIENDLY](#)

Search:

Archives:

Detecting Fraud in the Data Using AID, Part Two

By David H. Lindsay,
Ph.D., CFE, CPA,
CISA;
Paul Sheldon Foote,
Ph.D., Annhenrie
Campbell, Ph.D.,
CPA, CMA, CGA;
and David P. Reilly

In the Jan./Feb. 2004 issue we examined the basics of automatic intervention detection – an advanced computer-based tool. Here we describe the methodology of a research study conducted to explore the effectiveness of AID.

We developed a research study to evaluate whether automatic intervention detection could be used successfully to distinguish among companies with reported fraudulent data and those with no such reports. We selected eight companies which had been identified in the general media as having engaged in financial fraud. We downloaded newspaper articles from NEXIS outlining the specifics of each fraud. Table 1 lists the eight fraud firms and the type of fraud in which they were reported to have engaged.

Company	Nature of the Fraud
Cendant	Inflated earnings and improper use of reserves
Con Agra	Improper revenue recognition of fictitious sales
Enron	Failure to disclose liabilities and improper recognition of revenue
Grace	Improper use of reserves to facilitate income smoothing
McKesson	Premature recognition of sales revenue
Rite Aid	Recognition of fictitious vender credits
Sunbeam	Fictitious sales and improper use of reserves
Waste Management	Improper revenue recognition

Table 1

We pair-matched each fraud firm with a non-fraud firm classified within the same SIC code. If data availability permitted the identification of multiple non-fraud firms, we randomly selected two such pair-match firms (if available). The fraud firms and the pair-matched non-fraud firms are listed in Table 2.

Fraud Firms	Non-fraud Firms Pair-Match 1	Non-fraud Firms Pair-Match 2
--------------------	---	---

1	Cendant	Advance Tobacco Products	Competitive Technologies
2	Con Agra	Sara Lee	Classica
3	Enron	Mercury Air Group	World Fuel Service
4	Grace	Great Lakes Chemical	None
5	McKesson	Bergen Brunswig	None
6	Rite Aid	Drug Emporium	None
7	Sunbeam	Decorator Industries	None
8	Waste Management	Rich Coast	Wastemasters

Table 2

We obtained financial statements from the COMPU-STAT Annual Industrial File. Since this is an exploratory study, we downloaded data for all available data items. However, examination of the data on a firm-by-firm basis revealed that certain data item fields frequently were missing. We eliminated these fields because their data couldn't be compared across firms. The 45 remaining items are defined in Table 3.

COMPUSTAT DATA ITEMS TESTED

N1 CASH AND SHORT TERM INVEST.
 N2 RECEIVABLES
 N3 TOTAL CURRENT ASSETS
 N4 TOTAL CURRENT LIABILITIES
 N5 TOTAL ASSETS
 N6 NET PLANT AND EQUIPMENT
 N7 TOTAL LONG TERM DEBT
 N8 TANGIBLE COMMON EQUITY
 N9 NET SALES
 N10 OPERATING INCOME BEFORE DEPRECIATION.
 N11 DEPRECIATION AND AMORTIZATION EXPENSE
 N12 INTEREST EXPENSE
 N13 TOTAL INCOME TAXES
 N14 SPECIAL ITEMS
 N15 INCOME BEFORE SPECIAL ITEMS
 N16 AVAIL FOR COMMON AFTER ADJ.
 N17 COMMON SHARES OUTSTANDING
 N18 CUMULATIVE ADJUSTMENT FACTOR
 N19 CAPITAL EXPENDITURES
 N20 INVESTMENTS IN OTHERS
 N21 DEBT IN CURRENT LIABILITIES
 N22 DEF. TAXES & INV. CREDIT(BS)
 N23 RETAINED EARNINGS
 N24 TOTAL INVESTED CAPITAL
 N25 COST OF GOODS SOLD
 N26 DEBT DUE IN ONE YEAR
 N27 PRI EPS INCL. EXTRORD ITEMS
 N28 SHARES USED TO COMPUTE EPS
 N29 DILUTED EPS EX. EXTRAORDINARY ITEMS
 N30 PRIMARY EPS EX. EXTRAORDINARY ITEMS
 N31 COMMON EQUITY AS REPORTED
 N32 NON-OPERATING INCOME/EXPENSE
 N33 OTHER CURRENT ASSETS
 N34 OTHER ASSETS
 N35 ACCOUNTS PAYABLE
 N36 OTHER CURRENT LIABILITIES
 N37 DEFERRED TAXES (BAL. SHEET)
 N38 OTHER LIABILITIES

N39 DEBT (CONVERTIBLE)
 N40 DEBT (SUBORDINATED)
 N41 DEBT (NOTES)
 N42 DEBT (DEBENTURES)
 N43 DEBT (OTHER LONG-TERM)
 N44 CAPITALIZED LEASE OBLIGATION
 N45 COMMON STOCK
 Table 3

For each fraud firm and its pair-match, we downloaded 10 years of data. The 10-year period ended one year prior to the discovery of the fraud.

We counted the number of interventions at the last point in time and used that to rank the companies within each of the eight sets. Autobox™ correctly identified six of the eight fraudulent firms.

A tolerance threshold of 22 or more appears to indicate a good benchmark of when companies seem to be in a “red flag” zone as shown in Table 4.

Fraud Firms	Number of 45 B/S Items Found as Interventions in the Year before Fraud was Publicly Identified	
	Non-fraud Firms Pair-Match 1	Non-fraud Firms Pair-Match 2
Cendant 26	Advance Tobacco Products 3	Competitive Technologies 12
Con Agra 9	Sara Lee 15	Classica 10
Enron 22	Mercury Air Group 8	World Fuel Service 15
Grace 21	Great Lakes Chemical 11	
McKesson 29	Bergen Brunswig 16	
Rite Aid 29	Drug Emporium 10	
Sunbeam 2	Decorator Industries 7	
Waste Mgt 38	Rich Coast 4	Wastemasters 22

Table 4

(We didn't detect the frauds of Con Agra and Sunbeam. We would have detected Sunbeam if we had looked at the next to last point in time instead of the last point in time. It appears that “Chainsaw Al” Dunlop, CEO of Sunbeam, cleaned up his act before departing.)

To illustrate the identification of interventions we present graphs of three of the 45 items tested for Enron.

Outlier Detection Versus Plus/minus 3-Sima Charts

How else would you know that a point violated that model? In fact, the process of growing, understanding, finding, and examining outliers must be iterative. This isn't a new thought. Sir Frances Bacon, writing in

Novum Organum about 400 years ago said: “Errors of Nature, Sports and Monsters correct the understanding in regard to ordinary things, and reveal general forms. For whoever knows the ways of Nature will more easily notice her deviations; and, on the other hand, whoever knows her deviations will more accurately describe her ways.”

While no forecasting tool will be satisfactory in every circumstance, we believe automatic intervention detection based on Box-Jenkins modeling is superior to a number of other commonly applied approaches. Given accurate data, the central forecasting problem in AID is to develop a model that will reveal data anomalies that result from potentially fraudulent behavior while preventing

those same anomalies from distorting the model itself. The problem is made more challenging because data may not be completely accurate and anomalies may simply result from erroneous data collection and recording.

Some analysts think that they can remove outliers based on finding abnormal residuals to a simple fitted model or just by "eyeballing" the graph to look for unusual, extreme-appearing data points. If the outlier is outside of a particular probability limit (95 or 99), they then attempt to detect if there's something missing from the model. If not, they simply remove the outlier. (This procedure is illustrated in the next section.) The deletion or adjustment of the value so that there isn't an outlier effect is equivalent, statistically, to arbitrarily augmenting the model with a "0/1" variable where a "1" is used to denote a time point being considered and a "0" denotes a time point being disregarded. This manual adjustment, normally supported by visual or graphical analysis, often fails.

Additionally, this approach doesn't adjust for "inliers" whose effect on distorting a model is just as serious as "outliers." Inliers are defined as being "too normal or too close to the mean" and if ignored will bias the identification of the model and its parameters. Consider the time series 1,9,1,9,1,9,5,9 and how a simple model that identifies only extreme or excessive values might find nothing exceptional whereas a model more sensitive to patterns within the data would focus attention on the exceptional value of 5 at time period seven.

Further, to evaluate each and every unusual value separately is inefficient, misses the point of automatic intervention detection and doesn't allow any benefit from "data-scrubbing." Rather than using residuals, either visually or with a model, to identify outliers, automatic data-scrubbing or data-cleansing techniques instead can be applied to the underlying data. Such methods identify incorrect data using techniques such as comparison of similar data items according to given rules, identification of duplicate entries, and assurance of proper formatting. An individual, unusual value may, of course, represent an event of interest, a "pulse," rather than a problem with the data. A sequence of "unusual" values may arise at fixed intervals representing a "seasonal pulse." A sequence of values that may individually be within "bounds," might collectively represent a level shift, a new series of similar pulses that may or may not be permanent. A "time trend" may appear, at which point patterns in the data change. To complicate things a little bit more, there may be a local trend in the values. So, there are four types of "unusual values" that aren't errors in the data and that are important to the modeling process: 1) pulse; 2) seasonal pulse; 3) level shift; and 4) time trend. Therefore, while individual values may be within bounds, collectively they may be indicative of non-randomness or an intervention.

To assess an unusual value one needs to have a prediction. A prediction requires a model. Hopefully the model utilized isn't too simple, just simple enough.

Original Data may be Contaminated with Outliers

An original time series plot is a chronological or sequential representation of the readings. The mean is computed and the standard deviation is then used to place (+/-) 3 standard deviation limit lines. These are then superimposed on the actual data to assess what a reasonable spread or variation should be. Outlier points (points above or below 3 standard deviations) are immediately identified by sight and may be deleted from the next stage of the analysis.

The flaw in the above logic is obvious. The outliers will have distorted the initial computation of the standard deviation thus inflating it and masking other possible exceptions. Thus we need to simultaneously determine the process standard deviation and identify the outliers. This problem is exacerbated when you have autocorrelated data because this also has an effect on the standard deviation.

If the data is negatively autocorrelated i.e., high then low then high etc., the standard deviation is overstated. Similarly, if the data is positively autocorrelated i.e. slow drifts on either side of the mean, the standard deviation is understated.

Some would argue that the outliers can be identified via an “influential observation approach” or “Cook’s Distance approach.” Essentially this detection scheme focuses on the effect of the deletion of the outlier on the residual sum of squares or the resulting regression model. But this approach usually fails because, by definition, the outlier is an “unusual value” to its prediction and that prediction requires that a model already exist.

Autobox™ solves this problem by running computer-based experiments where an initial model is either identified or not before intervention detection is pursued. This is enhanced by also controlling for time trends and/or level shifts in distinctly different trials leading to the eventual global optimization.

Formal Presentation of How Outliers are Identified

Outliers can be represented as intervention variables of the forms: pulse, level shifts, seasonal pulses, and local time trends. The procedure for detecting the outlier variables is as follows. Develop the appropriate ARIMA model for the series. Test the hypothesis that there is an outlier via a series of regressions at each time period. Modify the residuals for any potential outlier and repeat the search until all possible outliers are discovered. These outliers can then be included as intervention variables in a multiple input Box-Jenkins model.

The noise model can be identified from the original series modified for the outliers. This option provides a more complete method for the development of a model to forecast a univariate time series. The basic premise is that a univariate time series may not be homogeneous and therefore the modeling procedure should account for this. By homogeneous, we mean that the underlying noise process of a univariate time series is random about a constant mean. If a series isn't homogeneous, then the process driving the series has undergone a change in structure and an ARIMA model isn't sufficient.

The Autobox™ heuristic that's in place checks the series for homogeneity and modifies the model if it finds any such changes in structure. The point is that it's necessary for the mean of the residuals to be close enough to zero so that it can be assumed to be zero for all intents and purposes. That requirement is necessary, but it isn't sufficient. The mean of the errors (residuals) must be near zero for all time slices or sections. This is a more stringent requirement for model adequacy and is at the heart of intervention detection. Note that some inferior forecasting programs use standardized residuals as the vehicle for identifying outliers. This is inadequate when the ARIMA model is non-null.

Summary and Conclusions

This study examined whether automatic intervention detection can be effectively used to distinguish companies with fraudulent reported data from those with no indication of fraud. Eight companies, identified in the general media as having engaged in financial statement fraud, were pair-matched with firms within the same SIC code that hadn't been identified by press as having engaged in financial statement fraud.¹ Financial statement data were obtained for the firms for a 10-year period ending one year prior to the discovery of the fraud. Hence, during the test period, the media hadn't yet publicized the fact that the firm was engaged in fraud. The number of interventions at the last point in time was used to rank the companies within the eight sets.

Intervention detection correctly identified six or seven of the fraudulent firms.² All fraud firms had a large number of interventions. These results are consistent with the supposition that automatic intervention detection can be effectively used to detect fraud firms.

Suggestions for Future Research

A critical next step is the application of intervention detection to test samples that reflect realistic base rates in which most firms aren't engaged in fraudulent reporting. This study used 45

COMPUSTAT data fields. Future studies may attempt to determine which of these fields are relevant to the detection of fraud and which are not. If the number of fields can be reduced, it will be possible to identify additional pair-matched firms.

Depending upon whether or not Sunbeam is included as one of the frauds detected, intervention detection was able to detect the fraud firm between 75 and 87.5 percent of the cases. This is true, even though we examined only 10 years of data. This suggests that the integrity of data is paramount. Future studies may fine tune the methodology by seeking to determine the incremental benefit of adding additional years of data.

David H. Lindsay, Ph.D., CFE, CPA, CISA, professor and chair of the Department of Accounting and Finance at California State University, Stanislaus. His email address is: DLindsay@csustan.edu.

Paul Sheldon Foote, Ph.D, is professor of accounting at California State University , Fullerton . His email address is: pfoote@fullerton.edu.

Annhenrie Campbell, Ph.D., CPA, CMA, CGFM, is a professor of accounting at California State University , Stanislaus. Her email address is: acampbel@athena.csustan.edu.

David P. Reilly is senior vice president at Automatic Forecasting Systems. His email address is: dave@autobox.com.

1 Wastemasters, a pair-matched firm to Waste Management, hadn't been identified in the general media as having engaged in financial statement fraud. However, Wastemasters subsequently was sued for activities related to its debentures.

2 Sunbeam would have been detected if we had looked at the next to last point in time instead of the last point in time.

References

- Albrecht, W.S. and C.C. Albrecht (2002) "Root Out Financial Deception," *Journal of Accountancy*, (193)4, pp. 30-34.
- Apostolou, B.A., J.M. Hassell, S.A. Webber and G.E. Sumners. 2001 "The relative importance of management fraud risk factors," *Behavioral Research in Accounting* (13) pp. 1-24.
- Box, G.E.P., and Tiao, G. (1975). "Intervention Analysis with Applications to Economic and Environmental Problems," *Journal of the American Statistical Association*, Vol 70, 70-79
- Busta, B., and R. Weinberg, (1998) "Using Benford's Law and neural networks as a review procedure," *Managerial Auditing Journal* (13)6, pp. 356-366.
- Coderre, D. (1999) "Computer assisted techniques for fraud detection," *The CPA Journal* (69)8, pp. 57-59.
- Davia, H.R. (2001) "Fraud 101: Techniques and Strategies for Detection" New York . John Wiley & Sons, Inc.
- Downing, D.J., and McLaughlin, S.B. (1986). "Detecting Shifts in Radical Growth Rates by the Use of Intervention Detection," *Engineering Physics and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge* .
- Glover, S.M., D. Prawitt, M.B. Romney (2000) "The software scene," *Internal Auditor* (57)4 pp. 49-57.
- Harvey , F. January 12, (2002) "A key role in detecting fraud patterns: neural networks," *Financial Times*. London . p. 3.
- Moyes, G.D. and I. Hasan. (1996) "An empirical analysis of fraud detection likelihood," *Managerial Auditing Journal* (11)3, pp. 41-46.
- Pincus, K. (1989) "The efficacy of a red flags questionnaire for assessing the possibility of fraud," *Accounting, Organizations and Society*, pp. 153-63.
- Reilly, D.P. (1980) "Experiences with an Automatic Box-Jenkins Modeling Algorithm," *Time Series Analysis*, ed. O.D. Anderson. (Amsterdam : North-Holland), pp. 493-508.
- Reilly, D.P. (1987). "Experiences with an Automatic Transfer Function Algorithm," *Computer Science and Statistics Proceedings of the 19th Symposium on the Interface*, ed. R.M. Heiberger, (Alexandria , VI: American Statistical Association), pp. 128-135.
- Rezaee, Z., A. Sharbatoghlie, R. Elam, P.L. McMickle. 2002. "Continuous auditing: Building automated auditing capability," *Auditing* (21)1, pp. 147-163.
- Wells, J.T. 2001. ". . . And nothing but the truth: uncovering fraudulent disclosures," *Journal of Accountancy* (192)7, pp. 47-52.

Glossary

ARIMA: This term identifies the model introduced by the statisticians Box and Jenkins in 1976 with the full name of Auto Regressive Integrated Moving Average model.

Autocorrelation: Rather than being independent of each other, each subsequent observation in a series is influenced by the value of the previous ones.

Box-Jenkins models: Developed by two statisticians, this type of model is used to study time series data and develop a basis for forecasting. Variations of the model allow for estimation of the pattern of a single variable (univariate) or of a number of related variables (multivariate).

Confidence interval: The range of possible values that are statistically most likely using a particular predictive model. The confidence interval is bounded by a probability limit.

Cook's distance: Also termed "influential observation" this refers to the measurement of the impact of a single observation on the resulting regression model.

Heuristic: A "rule of thumb" or relatively informal or intuitive decision rule

Noise model: A model of remaining unexplained data to be tested as part of a repetitive process of model-building.

Residual: The residual is the difference between an actual data point and the data point that a particular model would predict. Residuals from a model that closely fits a set of data will be smaller.

Time series data: Data collected as a sequence of periodic observations of one variable or set of variables is called time series data or longitudinal data. Longitudinal data should be contrasted with cross-sectional data in which a number of different variables are observed all at the same time.

Univariate time series: A sequence of periodic observations of a single variable.

Explanatory Web sites:

www.autobox.com

<http://searchcio.techtarget.com>

Web-based glossaries:

www.stats.gla.ac.uk/steps/glossary/index.html

<http://davidmlane.com/hyperstat/glossary.html>

[Home](#) | [About](#) | [Subscribe](#) | [Advertisers](#) | [Contributers](#) | [Archive](#)

All contents © 2004 Association of Certified Fraud Examiners.

Contact us at fraudmagazine@cfenet.com for more information