Data Mining at the Texas State Auditor's Office

Thomas J. Winn, Jr.
State Auditor's Office, Austin, Texas

Abstract

In general, data mining is the iterative and interactive process of finding new and potentially useful knowledge from large quantities of data.[1] The author's use of that term encompasses sophisticated statistical techniques, as well as many simpler data analysis methods. The mission of the Texas State Auditor's Office is to actively provide government leaders with useful information that improves accountability.[2] This expository presentation describes some data mining work at the State Auditor's Office in support of auditing and investigations activities. Particular attention will be given to fraud detection and risk assessment.

The Texas State Auditor's Office[3]

The State Auditor's Office (SAO) is the independent auditor for Texas State government. Operating within the legislative branch, the SAO provides government leaders and citizens with independent, objective, and reliable information about the operations of state agencies and higher education institutions. The goal of the SAO is to actively assist government leaders to create and maintain strong accountability systems that ensure efficient, effective operation of state agencies and universities. The SAO performs work at the direction of the Legislative Audit Committee, a permanent standing joint committee of the State Legislature that is co-chaired by the Lieutenant Governor and the Speaker of the House of Representatives. At the direction of the Legislative Audit Committee, the State Auditor would conduct an audit or investigation of any entity receiving funds from the state. The SAO focuses its efforts on the highest financial, business, and service risk. Our risk assessment process includes an ongoing statewide analysis of data collected from a variety of sources.

The SAO conducts financial audits, compliance audits, economy and efficiency audits, effectiveness audits, special audits, and investigations. Many State agencies, and other organizations, have internal auditors that conduct auditing activities that are similar to those at the SAO, but an important distinction is that the SAO does not "own" the data that are examined using the analytical procedures it uses. Since the data belong to other entities, before any analysis can be performed the data must be extracted and converted into a format that will be usable by our auditors. And then the reliability of the data is verified.

---

[1] Thomas J. Winn Jr., "Fraud Detection -- A Primer for SAS Programmers", Proceedings of the 31st Annual SAS Users Group International Conference, (March 2006), Paper 080-31.
[2] State Auditor's Office Web site, http://www.sao.state.tx.us
[3] Texas Government Code, Chapter 321.

The State of Texas has 214 agencies, boards, commissions, schools, and higher-education institutions. Many of these organizations have their own internal record-keeping systems, which they maintain on their own computing platforms. The data stored in an agency's internal system are kept in non-standard, agency-specific data files. Since most of the data reside on completely separate computing platforms, and since the data are stored in a variety of formats, then accessing the information cannot be accomplished by just "pointing and clicking." There do exist a few centrally administered statewide databases (pertaining to personnel, payroll, property, and financial accounting), but these also reside on computers outside of the SAO. Moreover, many of the commonly used files from the statewide systems are very large. For example, an extract file containing information regarding state financial accounting transactions for only one year includes more than 55 million records, and has a record length of 1300 bytes.

Introduction to Data Mining at the SAO

In data mining, techniques are used to uncover hidden patterns and subtle relationships among data in databases, and also to infer decision rules that may be useful in predicting future outcomes.[4] Some data miners have access to heavy equipment, while others have only "picks and shovels" for their use. At the SAO, systems analysts use various components of the SAS System to perform data extraction, data analysis, and data manipulation to support the audits and reviews that are conducted by our office. Until recently, we have not had access to SAS® Enterprise Miner™ software. Nevertheless, data mining is done at the SAO. SAO auditors and investigators use Microsoft Excel, Microsoft Access, ACL, and in-house-developed Statistical Toolbox[5] software to perform their data analysis work. And SAO systems analysts use Base SAS® and SAS/STAT® software as their primary data analysis toolset.

Experienced data miners with access to state-of-the-art analytical tools like SAS® Enterprise Miner™ might expect that an audit office would find such tools as decision trees, cluster analysis, and artificial neural networks to be indispensable for risk assessment and fraud detection. However, software acquisition decisions are frequently made differently in the public sector than elsewhere. In the absence of such advanced software, the SAO has accomplished those important tasks using other approaches. Of course, someone could posit that risk assessment or fraud detection might have been accomplished more effectively, or more efficiently, using Enterprise Miner or another data mining product solution, but as is often the case in other contexts, we all accommodate ourselves to whatever tools are readily available to us.

[4] David Hand, Heikki Mannila, and Padhraic Smyth, Principles of Data Mining, (Cambridge, Massachusetts:The MIT Press, 2001), pp. 1-4.
[5] State Auditor's Office, http://www.sao.state.tx.us/Resources/tools/toolbox.html

Risk Assessment at the SAO[6]

Auditors know that they can't examine *everything* about an entity, and so they use risk assessment to help them to focus their audits on the most important areas. Risks are undesirable events that could happen.[7] Basically, risk assessment involves supplying answers to the following questions: What could go wrong? What would be affected? What would be the magnitude of the impact? What is the likelihood that it might occur? And finally, which risks are most important? Risk entails two components: exposure and uncertainty[8] or, stated more precisely, the magnitude of a potential adverse effect associated with a particular hazard, and the probability that the problem will occur.[9] The literature concerning risk spans many fields, including actuarial science, engineering, environmental science, public health, and others. Auditors have their own way of thinking about risk. Risk assessment is an ongoing process.

The SAO's risk model characterizes financial, business, and service risks as being low, medium, or high, based upon a subjective assessment of pertinent information gathered together from many sources (audit reports, fieldwork, performance measures, communications, anecdotal information, articles in newspapers, etc.). Key risk areas include such items as:

Financial Risk –

- Funds are not spent as the Legislature intended,
- Anticipated revenues are not collected,
- Assets are not protected or used appropriately,
- Workforce is not appropriately sized and qualified;

Service Risk –

- Citizens are not provided the services mandated by the Legislature;

Business Risk –

- Information needed to manage day to day operations is not accurate/available,
- Investments in technology are not managed and used correctly,
- Management does not address risks that affect its ability to reach desired outcomes.

A scoring process is used, in which various risk factors and special circumstances are taken into account. Some risks arise due to external factors such as: technological advances, economic changes, changes in laws, natural catastrophes, or changes to major suppliers. Other risks are associated with internal factors, for example: downsizing, change of leadership, high staff turnover, or redesign of operating processes.

---

[6] Babette Laibovitz, "Risk Assessment at the State Auditor's Office", State Auditor's Office, April 24, 2006.

[7] 5BeAnActuary Web site, http://www.beanactuary.org/about/whatis.cfm

[8] Glyn A Holton, "Defining Risk", Financial Analysts Journal, Vol. 60, No. 6, November/December 2004, pp. 19-25.

[9] Wikipedia webpage for "Risk Assessment", http://en.wikipedia.org/w/index.php?title=Risk_assessment

Annually, a proposed audit plan is developed, based upon statutory requirements, and the results from the risk assessment process that is carried out at a statewide level. The proposed audit plan is reviewed by management, and then is submitted to the Legislative Audit Committee for approval.

After the audit plan is adopted, and the goals and objectives of each audit are determined, then the auditors continue focusing on risk. They do this to determine which controls would be the most important to test, based upon the areas that are vulnerable to the most important risks. This process of thinking about risks and controls begins at the planning stage of each audit, and it continues throughout the project as new information becomes available.

Data Mining by SAO Auditors

An audit is an evaluation of an organization, system, process, project, or product.[10] There are various types of evaluations which auditors perform, including financial audits, performance audits, and attestation engagements. However, regardless of the type of audit, at the S.A.O. each time an audit is undertaken, there are specific objectives which are based upon an assessment of the potential risks. So the scope of each of the audits we conduct is narrowly defined beforehand. Therefore, while the risk assessment process uses data mining with a fairly wide focus, auditing sometimes uses narrowly focused data mining in order to achieve the audit objectives.[11] Both of those approaches contribute to accomplishing the mission of the SAO.

For auditors, data mining is generally used as an exploratory tool, pointing toward areas of large data sources whose details require further evaluation using traditional audit procedures. Auditors use elementary descriptive statistics to establish norms and to identify outliers in data; they also analyze trends, fluctuations, and ratios; they perform between-period comparisons; and they use least-squares regression for testing reasonableness, identifying exceptions, and making forecasts. Auditors sometimes analyze the digits in accounts payable data using Benford's Law (and related distributions pertaining to digital analysis), to search for possible irregularities. When performing audits in which personal data (perhaps pertaining to employees or to clients) are within the audit scope, auditors will check Social Security Numbers to see if they appear to be valid, and they also will compare the SSNs with those contained in the Social Security Administration's death master file. And occasionally auditors also look for possible matches in different data files, by comparing names and addresses using fuzzy logic.

[10] http://en.wikipedia.org/wiki/Audit

[11] Ron Franke, "Data Mining Efforts at the Texas State Auditor's Office", presented to the National Intergovernmental Audit Forum, November 18, 2003.

Recent audit literature contains examples of the use of data mining to identify high risk transactions and control weaknesses, in the context of continuous auditing[12] ; however, while the S.A.O. does employ tools for monitoring certain statewide data, these efforts do not necessarily lead to audits.

Auditors recognize that there is a chance that their audit procedures will not lead them to issue a correct opinion. Statisticians recognize two types of errors: type I errors are false positives (negative instances erroneously reported as positives), and type II errors are false negatives (positive instances erroneously reported as negatives). Audit risk is the risk that the auditor will give an inappropriate audit opinion. In general, this would happen if a material error occurred in the accounting records that was not detected either by the internal controls or by the audit procedures (which would be a type II error). To reduce the likelihood of material misstatements, auditors typically increase the size of their samples, but this can lead to overauditing. An approach that is used at the SAO for auditing the combined financial statements of all state agencies is to estimate the probability of materially misstated accounts at three different levels of audit effort, for the largest state agencies, and also on a statewide basis, and then to perform a Monte Carlo simulation to help determine the probability of audit risk for each combination of agency and audit effort, in order to determine the most efficient allocation of audit work for the annual statewide financial audit, beforehand. This approach has been used for many years at the SAO. For most audits, SAO auditors use a sampling decision tree, to ensure that their samples will be large enough to be sufficient, and yet small enough to be cost-effective.

Fraud

Recent highly publicized corporate scandals have resulted in strict new laws, policies, and standards that affect most companies, as well as many other organizations. The Sarbanes-Oxley Act of 2002 specifies new requirements regarding the financial management of publicly traded companies. It requires companies to implement antifraud programs and controls. Also in 2002, the American Institute of Certified Public Accountants (AICPA) issued Statement on Auditing Standards No. 99[13], which specifies requirements for auditors concerning fraud considerations in financial statement audits.[14]

---

[12] Jennifer Moore, Karina Barton, and Joseph O'Donnell, "Continuous Auditing, XBRL, and Data Mining", New York State Society of Certified Public Accountants, Technology Assurance Committee (June 15, 2004), http://www.nysscpa.org/committees/emergingtech/auditing2004.ppt ; and
David Coderre, "Continuous Auditing: Implications for Assurance, Monitoring, and Risk Assessment", The Institute of Internal Auditors, Global Technology Audit Guide Volume 3 (2005); and
J. Donald Warren Jr. and Xenia Ley Parker, Continuous Auditing: Potential for Internal Auditors, The Institute of Internal Auditors (2003)..

[13] AICPA, Consideration of Fraud in a Financial Statement Audit – SAS No.99, (New York: American Institute of Certified Public Accountants, 2002).

[14] Michael Ramos, "Auditor's Responsibility for Fraud Detection", http://www.aicpa.org/pubs/jofa/jan2003/ramos.htm.

The Association of Certified Fraud Examiners estimated that, in 2006, occupational fraud and abuse will cost the U.S. economy about $652 billion, which represents a loss of about 5 percent of revenues.[15] Fraud is a significant and growing financial risk that threatens the profitability and reputation of companies around the world. Fraud can occur anywhere, and it has become a major concern of many organizations.

> *Definition of Fraud:*
> *"A knowing misrepresentation of the truth or concealment of a material fact to induce another to act to his or her detriment."* [16]

Fraud is an intentional act meant to induce another person to part with something of value, or to surrender a legal right. It is a deliberate misrepresentation or concealment of information in order to deceive or mislead. Fraud can range from minor employee theft and unproductive behavior to misappropriation of assets and fraudulent financial reporting. In different situational contexts, fraud can take somewhat different forms – for example, bribery, embezzlement, securities fraud, health care fraud, money-laundering scams, insurance fraud, software piracy, internet fraud, telemarketing fraud, mortgage foreclosure scams, and identity theft -- these all have their own special characteristics. There are at least as many types of fraud as there are types of people who commit it. But in each instance, fraud involves deception. Someone knowingly lies in order to obtain an unlawful benefit, or an unfair advantage.

Some examples of fraud include:

- any dishonest or fraudulent act;
- forgery or alteration of a check, bank draft, or financial document;
- misappropriation of assets;
- deliberate impropriety in the handling or reporting of money or financial transactions;
- wrongfully using influence in a business transaction to receive a benefit (such as bribery, kickbacks, and bid-rigging);
- profiteering as a result of insider information;
- disclosing insider information to another person in order for them to secure unlawful gain.

The Association of Certified Fraud Examiners has identified a fraud taxonomy that includes more than 50 different types of occupational fraud alone,

---

[15] Association of Certified Fraud Examiners, 2006 Report to the Nation on Occupational Fraud and Abuse, http://www.acfe.com/documents/2006-rttn.pdf.
[16] Bryan A. Garner (Editor), Black's law Dictionary, 7th edition., (St. Paul, Minnesota: West Group, 1999), p. 670.

considering fraud where people use their occupation for personal enrichment by misusing or misapplying their employers' resources. [17]

Abuse is behavior that is deficient or improper when compared with behavior that a prudent person would consider reasonable and necessary business practice given the facts and circumstances.[18] When abuse occurs, there is no violation of any law, regulation, or contract provision, but such instances are harmful, and they need to be minimized.

Determining the difference between improper acts and fraud is a matter for decision by the judicial system. But the detection of acts which may turn out to have been fraudulent is something that all administrators, private and public, need to be involved in. Perhaps the best way to prevent fraud is for many people in every organization to become vigilant in detecting and reporting possibly fraudulent activities. The Texas Government Code (Section 321.022) requires state agencies and universities to report suspected fraud or unlawful conduct to the State Auditor's Office. The SAO coordinates investigations concerning allegations of impropriety and illegal acts affecting state resources.

Importance of Internal Controls

An internal control system is a framework of processes and procedures for planning, organizing, directing, and controlling program operations, and for measuring, reporting, and monitoring program performance.[19] Internal controls ensure the reliability of the operational data, and provide assurance for attainment of the organization's goals and objectives. Internal controls include requirements which pertain to the recording, processing, and reporting of financial data, and the safeguarding of an organization's assets (cash, inventory, equipment, etc.).

Here are a few important principles concerning internal controls:

- Duties need to be distributed among employees so that no one person will have too much control in any financial area. No single employee should be able to handle any transaction from start to finish. Whenever several people handle different aspects concerning transactions, cross-checking of data occurs, mistakes and irregularities are reduced, and the chance that a fraud may go unnoticed decreases.
- Authorization procedures must be established and observed, to ensure the propriety and validity of certain activities and transactions. There should be written policies and procedures governing who may authorize transactions. Authorizations must be obtained *before*, and *not after* the fact.

[17] Association of Certified Fraud Examiners, op.cit.
[18] United States General Accounting Office, Government Auditing Standards, 2003 Revision, GAO-03-673G, (Washington, DC: General Accounting Office, 2003), p. 106.
[19] Ibid., p. 23.

- Documents must be properly designed and used. Retain all written records. Forms should be pre-numbered sequentially, and carefully accounted for. Spoiled documents should be voided, and retained.
- Reconciliation procedures should be followed, in which related sets of data are compared by an independent party, in order to identify and examine any differences. There also should be regular inventories of physical assets.
- Internal controls also include provisions regarding physical security of assets and records. Access to records should be based upon need, and this must be reviewed regularly.

There need to be periodic reviews of the internal controls, to make sure that they are working. Fraud can occur whenever there are weaknesses in internal controls.

There usually are three conditions present when fraud occurs: incentive (pressure), opportunity, and rationalization (finding a justification for committing a fraudulent act).[20] Furthermore, there is a possibility that management or certain other key employees may have the ability to override at least some internal controls. They generally will take steps to conceal fraudulent acts, which will make it difficult to detect.

Fraud Detection at the SAO

Consideration of the possibility of fraud and abuse is now a part of *every* audit performed by the Texas State Auditor's Office. Here is a partial listing of some possible fraud symptoms:[21]

General

- Altered documents
- General Ledger out of balance
- Liquid paper and erasures on original documents
- Provision of copies when originals are expected, especially a refusal or systematic inability to provide original documents
- Missing documents or data
- Customer complaints
- Documents damaged by water, fire, and/or theft

Cash

- Overages/shortages in cash drawer/petty cash
- Lack of a systematic reconciliation of the cash drawer
- Inappropriate or overly frequent use of postal or money orders

[20] Statement on Auditing Standards #99 - Consideration of Fraud in a Financial Statement Audit, AICPA, [Note -- the "fraud triangle" was first described by criminologist Donald R. Cressey, Other People's Money: A Study in the Social Psychology of Embezzlement (The Free Press, Glencoe, IL, 1953)].

[21] State Auditor's Office, "Project Procedures Manual – Fraud" (an online resource on the SAO intranet).

Cash in Bank

- Control of checks different from deposit
- Deposits in transit increasing in amount and/or number
- Excessive voids or refunds
- Lack of systematic reconciliation of bank account or reconciling items that remain uncleared for long periods of time

Accounts Receivable

- Adjustments to receivables or entries without formal approval (especially manual, non-cash debit/credit adjustments)
- Increase in past due accounts or write-offs of late charges
- No collections on past due or written-off accounts
- Customer invoices billed out of sequence
- Gaps in invoice numbers
- Excessive sale voids or refunds (especially for specific employee, location, or customer)
- Undercharges for sales of goods or services

Inventory/Assets

- Shortages or adjustments to inventory
- Increased scrap or increased or premature surplus
- Duplicate payments on purchases
- Delivery to an off-site location
- Lack of systematic physical verification of inventory/assets
- Inventory/assets not found in location expected or recorded in inventory or asset management system
- Excessive or frequent use of loss leaders or takeaways

Purchases

- Support for payments not cancelled or marked paid
- Deviation from specifications on delivered goods or services
- Goods purchased in excess of needs. Payments made in currency when checks are expected
- Awards made to contractors with poor track records
- Activation of a dormant account, followed by a payment

Payroll

- Second endorsements on payroll checks
- Missing employees (i.e., employees not listed in directory)
- Employees with inactive or bogus social security numbers or multiple employees with the same social security number.
- Employees receiving payroll check (or direct deposit of pay, if applicable), but payroll file does not indicate withholding for taxes, health insurance, etc.
- Employees whose address is a Post Office Box

- Specimen signature different from endorsement
- Inflated hours
- Large gap between employee qualifications and job duties
- Frequent use of "must hires"
- Missing human resources data
- Small group of staff always get overtime
- Theft or disappearance of unclaimed checks
- Inappropriately high leave accruals

Accounts Payable

- Invoices not on preprinted or letterhead forms
- Adjustments to payables or entries without formal approval (especially manual, non-cash debit/credit adjustments)
- Invoices created on outdated forms
- Payees with common names and/or addresses
- Vendor's address or phone number the same as an employee's address or phone number
- Duplicate payments or over-payments to vendors
- Payments made to vendors that are not on the approved vendor list
- Vendors with missing information in the master file or list, such as missing phone numbers, missing addresses
- Vendors whose address is a Post Office Box

The presence of fraud symptoms would not necessarily indicate the existence of fraud. But, at the SAO, tips and observed symptoms are followed-up on carefully. The SAO has developed and implemented special procedures for auditors to use in detecting fraud and abuse in the following high-risk areas: payroll, contracting and procurement, credit cards, revenues and receipts, assets management, and construction.

Wherever possible, computer programs are used to uncover the existence of pertinent fraud symptoms in automated information systems. Very few of those computer programs would be regarded as advanced analytical tools, and none of them would be located on a map of sophisticated data mining techniques. But they get the job done! The author recently published a paper which described some basic methods for detecting possibly fraudulent activities using only Base SAS software.[22]

One might infer from the preceding discussion that finding fraud is just a matter of looking for instances of certain well-known symptoms in the data. However, it must be pointed out that, since fraud schemes are dynamic and adaptive, there is no magical software tool for fraud detection which would automatically detect all of the fraudulent transactions, and that would continue to be useful into the future. Most fraud schemes intentionally exploit weaknesses in an

[22]Winn, op.cit.

organization's internal controls, and that is the reason it is difficult for auditors to find fraud. We believe that advanced analytical tools may contribute to the fraud detection process, but that there is no substitute for creative, and sometimes tedious, investigative effort by auditors. SAO auditors inform our data analysts about any updates in fraud symptoms that we should be looking for.

Conclusion

Data mining is the iterative and interactive process of finding new and potentially useful knowledge from large quantities of data. The author's use of that term encompasses sophisticated statistical techniques, as well as many simpler data analysis methods. SAO systems analysts and auditors use a variety of analytical tools for risk assessment, fraud detection, and other types of data analysis involving large databases; however, until recently those tools did not include decision trees, cluster analysis, or artificial neural networks. This paper provided a general explanation of some of the data mining work at the SAO. It will be interesting to observe any future improvements in continuous monitoring, risk assessment, and fraud detection at the SAO, as a result of using SAS® Enterprise Miner™ software.

Author Information

Tom Winn, Ph.D.
Senior Systems Analyst
Texas State Auditor's Office
P.O. Box 12067
Austin, TX 78711-2067

Telephone: 512 / 936-9735
E-Mail: twinn@sao.state.tx.us