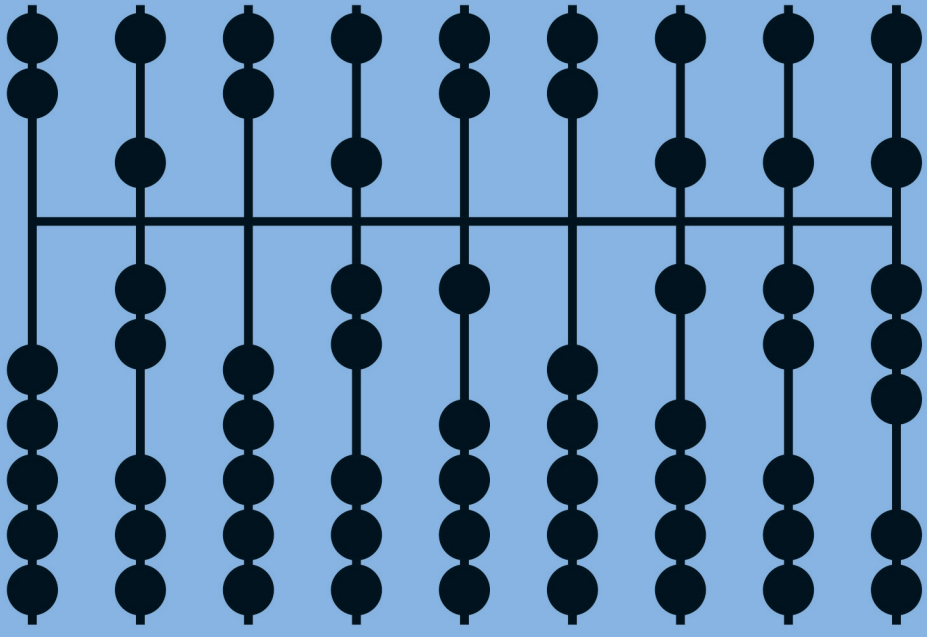
RUTGERS STUDIES IN ACCOUNTING ANALYTICS

Audit Analytics in the Financial Industry



EDITED BY
JUN DAI, MIKLOS A. VASARHELYI
AND ANN F. MEDINETS

Audit Analytics in the Financial Industry

This page intentionally left blank

RUTGERS STUDIES IN ACCOUNTING ANALYTICS

Audit Analytics in the Financial Industry

BY

JUN DAI

Southwestern University of Finance and Economics, China

MIKLOS A. VASARHELYI

Rutgers University, USA

and

ANN F. MEDINETS

Rutgers University, USA

United Kingdom – North America – Japan – India – Malaysia – China

Emerald Publishing Limited
Howard House, Wagon Lane, Bingley BD16 1WA, UK

First edition 2019

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-78743-086-0 (Print)
ISBN: 978-1-78743-085-3 (Online)
ISBN: 978-1-78743-173-7 (Epub)

ISOQAR certified
Management System,
awarded to Emerald
for adherence to
Environmental
standard
ISO 14001:2004.

Certificate Number 1985
ISO 14001

INVESTOR IN PEOPLE

Contents

Introduction: What is Audit Analytics?

Jun Dai and Miklos Vasarhelyi

The spate of accounting scandals and corporate failures since 2001 has brought unprecedented attention to the importance of corporate governance. The Enron scandal, revealed in October 2001, resulted in a loss of about $80 billion in market capitalization for investors (The Washington Post, 2002), and a year later, an audit team unearthed $3.8 billion in fraud at WorldCom (Pulliam & Solomon, 2002). Since then, both professional auditors and audit researchers have devoted significant effort to improving the capabilities of auditing, internal control, and continuous monitoring (Alles, Brennan, Kogan, & Vasarhelyi, 2006; Byrnes, 2015; Chan & Vasarhelyi, 2011; Jans, Alles, & Vasarhelyi, 2014; Vasarhelyi, Alles, & Williams, 2010).

"Big data" is receiving increased attention from accounting practitioners. Organizations have collected more data in 2 years than in the previous 2,000 years (Syed, Gillela, & Venugopal, 2013). For example, Walmart collects more than 1 million customer transactions every hour, and Facebook collects more than 200 gigabytes of data per night (Cao, Chychyla, & Stewart, 2015). In addition to data stored in traditional accounting systems, auditors are also able to acquire evidence from vast amounts of other complex data, such as non-financial data extracted from modern enterprise resource planning (ERP) systems or online databases, radio frequency identification trackers and networked sensors, social media, and even closed-circuit television videos in stores (Moffitt & Vasarhelyi, 2013). In addition, many countries now permit some of their government administrative information and data collected from their citizens and businesses to be open to the public, which provides auditors with even more data for monitoring and investigations (Dai & Li, 2016; O'Leary 2015; Schneider, Dai, Janvrin, Ajayi, & Raschke, 2015).

To extract and process data from a variety of sources to be used for identifying risks, collecting evidence, and ultimately supporting decisions, auditors are utilizing the emerging technology of audit analytics (AA). AA is defined as a science of

> discovering and analyzing patterns, identifying anomalies, and extracting other useful information in data underlying or related to the subject matter of an audit through analysis, modeling, and visualization for the purpose of planning or performing the audit. (AICPA, 2015)

The predecessor of AA is the analytical procedure, which has long been used as one of external auditors' techniques in the planning, substantive testing, and

completion phases of audits (AICPA, 2015). Since analytical procedures performed in the planning phase typically "use data aggregated at a high level" (AICPA, 2012), "the results of those analytical procedures provide only a broad initial indication about whether a material misstatement may exist" (AICPA, 2012). AA techniques can be applied to transaction-level data because such techniques generally maintain good performance even when used on huge and high-dimensionality data sets. As a result, AA can enhance the accuracy of risk assessment and improve the quality of planning.

Traditional analytical procedures usually rely heavily on sampling of audit-related data (AICPA, 2015). However, as large-scale ERP systems are rapidly growing in popularity among businesses, sufficient evidence can no longer be collected from only a sample of data. AA increases the tested population from limited samples (judgmental or statistical) to millions of transactions in full population testing, which enlarges the audit coverage from a small percent of overall transactions to the entire population (AICPA, 2015). Besides data recorded by a client firm's ERP system, auditors also have access to public data, such as social media postings (Moon, 2016), open government data (Dai & Li, 2016; Kozlowski, 2016), and weather data (Yoon, 2016). Emerging data analytics technologies have the capability to explore vast amounts of data in various structures and formats, which cannot be handled by traditional analytical procedures.

AA offers several advantages over traditional approaches. First, audit data analytics are more cost-effective in terms of evidence collection. On average, AA costs $0.01 compared to $4 for a standard audit of the same evidence.[1] In addition, many data analytics techniques are scalable in that they can generally maintain good performance when handling huge and high-dimensionality datasets (Alpaydin, 2010). Some AA techniques also have the ability to identify data patterns in an unsupervised learning paradigm in which the training data sets for building detection models do not need to contain class label information (Byrnes, 2015; Thiprungsri & Vasarhelyi, 2011).

Part One of this book presents two articles illustrating the process of applying AA to solving audit problems. Part Two contains four studies that use various AA techniques to discover fraud risks and potential frauds in the credit card sector. Part Three focuses on the insurance sector and uses two articles to show the application of clustering techniques in auditing. Part Four includes two chapters on how to employ AA in the transitory system for fraud/anomaly detection. Parts Five and Six illustrate the use of AA to assess risks in the lawsuit and payment processes.

Auditing researchers have been devoting significant efforts to integrating AA techniques into existing audit programs. AA can facilitate various stages of the audit process with simple or complex tests. Chapter 1 summarizes exploratory data analysis (EDA) techniques and the audit stages in which they could be employed for both internal and external audits. This research also conceptualizes the process of implementing EDA in audit procedures. Similarly, Chapter

[1] http://raw.rutgers.edu/node/89.html

2 provides guidance for auditors to apply these new technologies in actual audit work.

A variety of AA technologies can be employed to facilitate risk discovery, anomaly identification, and fraud detection. Chapters 3–5 explore the use of clustering methodologies to identify risky customer groups for a bank's credit card department. After grouping customers with similar characteristics and purchase/payment behaviors into clusters, the bank can manage each group differently and take actions for high-risk credit card holders.

Similar approaches are also employed to identify abnormal life insurance claims. Chapter 7 uses a simple *K*-means clustering model to group claims with similar characteristics and to flag unusually small clusters for further investigation. Chapter 8 explores the attributes to be used to identify outliers, and then uses clustering to assess whether life/disability insurance claim settlements are reasonable and whether the claims themselves are legitimate.

Decision tree is an AA technique that is easy to understand and can facilitate risk and error identification effectively by learning the characteristics and behavior patterns in the data. Chapter 6 shows the potential of Decision Trees for helping internal auditors to identify credit card delinquency, and Chapter 11 applies the Decision Tree methodology to the risk of lawsuits for credit card customers.

Fraud detection is another domain that can benefit from AA techniques. By analyzing transaction-level data, AA can capture unusual data flows and abnormal patterns. Chapters 9 and 10 illustrate how rule-based systems can facilitate fraud detection by incorporating expert knowledge into models. Chapter 9 illustrates the development and testing of a model to detect anomalous transactions in a bank's transitory accounts. Chapter 10 detects fraudulent transactions in the payment process for wire transfers by identifying potential fraud indicators, each of which is assigned a risk score based on perceived severity. Payments with total scores that exceed a threshold would be considered potentially fraudulent transactions that can be recommended for further investigation. Internal control is another important and complex area that could benefit from AA. In Chapter 12, two methods are presented. One of them is fuzzy logic which is used to create a generic risk model for assessing internal controls over payments and the other is the use of statistical tools to detect outliers and anomalies on the data.

The goal of this book is to provide insights for academics, auditors, and business professionals on potential applications of AA in the financial industry. Real-life data and audit problems are used to demonstrate how AA can facilitate the discovery of audit concerns that would be difficult or time consuming if traditional approaches were used.

References

AICPA. (2012). Understanding the entity and its environment and assessing the risks of material misstatement, AU-C Section 315. Retrieved from www.aicpa.org/Research/Standards/AuditAttest/DownloadableDocuments/AU-C-00315.pdf

AICPA. (2015). Audit analytics and continuous audit: Looking toward the future. Retrieved from https://www.aicpa.org/interestareas/frc/assuranceadvisoryservices/downloadabledocuments/auditanalytics_lookingtowardfuture.pdf

Alles, M. G., Brennan, G., Kogan, A., & Vasarhelyi, M. A. (2006). Continuous monitoring of business process controls: A pilot implementation of a continuous auditing system at Siemens. *International Journal of Accounting Information Systems*, *7*(2), 137–161.

Alpaydin, E. (2010). *Introduction to machine learning* (2nd ed.). Boston, MA: MIT Press.

Byrnes, P. E. (2015). *Developing automated applications for clustering and outlier detection: Data mining implications for auditing practice*. Doctoral dissertation, Rutgers University Graduate School, Newark, NJ.

Cao, M., Chychyla, R., & Stewart, T. (2015). Big data analytics in financial statement audits. *Accounting Horizons*, *29*(2), 423–429.

Chan, D. Y., & Vasarhelyi, M. A. (2011). Innovation and practice of continuous auditing. *International Journal of Accounting Information Systems*, *12*(2), 152–160.

Dai, J., & Li, Q. (2016). Designing audit apps for armchair auditors to analyze government procurement contracts. *Journal of Emerging Technologies in Accounting*, *13*(2), 71–88.

Jans, M., Alles, M., & Vasarhelyi, M. (2014). A field study on the use of process mining of event logs as an analytical procedure in auditing. *The Accounting Review*, *89*(5), 1751–1773.

Kozlowski, S. (2016). *A vision of an ENHanced ANalytic constituent environment: ENHANCE*. Doctoral dissertation, Rutgers University Graduate School, Newark, NJ.

Moffitt, K. C., & Vasarhelyi, M. A. (2013). AIS in an age of Big Data. *Journal of Information Systems*, *27*(2), 1–19.

Moon, D. (2016). *Crowdsourcing for social sciences researchers: Data gathering, teaching, learning and research dissemination from a single project*. London School of Economics and Political Science, LSE Impact Blog. Retrieved from https://blogs.lse.ac.uk/impactofsocialsciences/2016/10/03/crowdsourcing-for-social-sciences-researchers-data-gathering-teaching-learning-and-research-dissemination-from-a-single-project/

O'Leary, D.E. (2015). Armchair auditors: Crowdsourcing analysis of government expenditures. *Journal of Emerging Technologies in Accounting*, *12*(1), 71–91.

Pulliam, S., & Solomon, D. (2002). How three unlikely sleuths exposed fraud at WorldCom. *The Wall Street Journal*, October 30. Retrieved from https://www.wsj.com/articles/SB1035929943494003751

Schneider, G. P., Dai, J., Janvrin, D. J., Ajayi, K., & Raschke, R. L. (2015). Infer, predict, and assure: Accounting opportunities in data analytics. *Accounting Horizons*, *29*(3), 719–742.

Syed, A., Gillela, K., & Venugopal, C. (2013). The future revolution on Big Data. *International Journal of Advanced Research in Computer and Communication Engineering*, *2*(6), 2446–2451.

The Washington Post. (2002). Timeline of Enron's collapse. *The Washington Post*, February 25.

Thiprungsri, S., & Vasarhelyi, M. A. (2011). Cluster analysis for anomaly detection in accounting data: An audit approach. *International Journal of Digital Accounting Research*, *11*(17), 69–84.

Vasarhelyi, M. A., Alles, M. G., & Williams, K. T. (2010). *Continuous Assurance for the now economy. A thought leadership paper for the Institute of Chartered Accountants in Australia*. Queensland, Australia: Institute of Chartered Accountants.

Yoon, K. (2016). *Three essays on unorthodox audit evidence*. Doctoral dissertation, Rutgers University Graduate School, Newark, NJ.

Part I

Audit Analytics Procedures

This page intentionally left blank

Chapter 1

An Application of Exploratory Data Analysis in Auditing – Credit Card Retention Case*

Qi Liu

1. Introduction

Over the last several decades, operational risks in banking systems has attracted both regulatory and academic attention due to the devastating losses experienced by some banks. For example, Allied Irish Banks lost $750 million due to rogue trading,[1] and Prudential Insurance entered into a $4 billion class action settlement with regard to fraudulent sales practices over 13 years[2] (Muermann & Oktem, 2002).

An operational audit focuses on evaluating the efficiency, effectiveness, and economy of organizational activities to reduce operational risks and improve future performance (Lane, 1983). It plays an important role in ensuring that organizations achieve their strategies and objectives. This chapter demonstrates an application of exploratory data analysis (EDA) (Tukey, 1977) in a real operational audit setting and illustrates how internal auditors can benefit from this approach.

This case study in this chapter analyzes a data set of credit card annual fee discounts from an international bank in Brazil. In this case study, the EDA process is mainly applied to test three pre-defined audit objectives. The results of the EDA process are compared with the results of conventional audit procedures. The outcomes of this comparison demonstrate that EDA permits the auditor to obtain comprehensive findings easily with simple statistics and visualization techniques.

*This chapter is based on the third chapter of the author's dissertation (Liu, 2014).
[1]http://online.wsj.com/news/articles/SB10129910421902036 40
[2]http://caselaw.findlaw.com/us-3rd-circuit/1362355.html

Audit Analytics in the Financial Industry, 3–15

doi:10.1108/978-1-78743-085-320191001

The chapter begins with a description of the audit problems facing this bank and then discusses the data and specific methods used in this case. The results of both conventional audit procedures and the EDA process are then presented. Finally, implications and limitations of this case study are discussed.

2. The Audit Problem

2.1. Scenario

This study investigates the credit card division of a large international bank in Brazil. Most of the credit cards issued by this bank have annual fees. Clients who do not want to pay these fees may call the bank and ask for a fee cancelation or a fee reduction. In these circumstances, bank representatives negotiate with the clients about the fees. Based on clients' backgrounds, representatives can then offer appropriate discounts. During the discount negotiation process, bank representatives should follow the bank's policy; they cannot offer discounts higher than they are authorized to give, and they should give top priority to the benefit of the bank. In other words, they should offer the lowest discounts acceptable to the clients.

2.2. Audit Objectives

The bank suggested that the initial audit scope for this study is to identify the bank representatives whose behavior, in the course of the annual fee negotiations, may cause the loss of bank revenue. Risky behaviors include (1) offering higher discounts than allowed; (2) offering high discounts without making an effort to negotiate lower discounts; and (3) offering discounts without any client negotiation. Based on these behaviors, three audit objectives are developed:

(1) All bank representatives obeyed bank policy when offering discounts.
(2) Bank representatives offered the lowest possible discounts to retain clients.
(3) Bank representatives negotiated with clients for lower discounts before offering final discounts.

In addition to these issues, the audit scope is extended to discovering potential operational risks in the annual fee-offering process. Non-behavioral factors, such as lack of effective internal controls, can also lead to loss of revenue. Even though some cases are not directly related to current revenue losses, business process risks may cause future revenue loss.

To achieve these audit objectives, all related fields need to be thoroughly explored for irregularities, making this topic a suitable scenario for EDA. As in a traditional audit, the auditors must gain an understanding of the process and then identify the risks and problems related to this process and its associated internal control system before testing to determine whether any policies have been violated.

3. Methodology

3.1. Data

Two data sets are used in this case: the retention data and the account master data. The retention data includes information on customer phone calls made in January 2012. The data set consists of 195,694 records in total. Each record represents a customer's phone call and contains 162 fields.

The account master data is a large data set with 60,309,524 records and 504 fields. Each record represents a credit card account. All accounts opened in the bank from July 1980 to March 2012 are included in the data set. The fields in the account master data cover a wide variety of information relevant to the accounts and accounts holders: account information, such as account type and account status; demographic information, such as account holders' age and gender; and financial information, such as credit limits and late pay amounts. Account master data is updated by the bank on a continuous basis.

This case study uses eight attributes: call length, bank representative ID, supervisor ID, customer service center location, original fee, actual fee, sequence number of the account, and number of cards. Most of these attributes, such as call length, annual fee, and output annual fee, are necessary to test the original audit objectives. Other attributes are newly added during the EDA process, such as supervisor number and number of cards. The names, source database, and descriptions of these attributes are listed in Table 1.

Among these fields, call length, original fees, actual fees, and number of cards are continuous variables. Representative's ID, supervisor's ID, client's ID, account sequential number, and customer service center location are nominal variables. To protect clients' privacy, the account sequential numbers and clients IDs are

Table 1. Description of Attributes Included in This Study.

Attribute Name (Source Database)	Description
Call length (retention)	The duration of each call in seconds
Call location (retention)	The location of the customer service center
Agent number (retention)	ID of the bank representative answering the call
Supervisor number (retention)	ID of the representative's supervisor
Sequential number (retention and account master)	Sequence Number of an account
Annual fee (retention)	Original annual fees of a credit card
Output annual fee (retention)	Actual annual fees paid by each client
Number of cards (account master)	Number of cards associated with each account

encrypted in the data set. The encryption method preserves the integrity of the original data; each original value corresponds to a unique cipher text.

3.2. Data Preprocessing

Discounts offered by bank representatives play an important role in the process of analyzing loss of revenue. However, there is no field that directly reflects discounts in the raw retention data. Two existing fields that relate to discounts are original fees before negotiation and actual fees after negotiation. The difference represents the discount, which is needed to conduct EDA. Specifically, the discount is the difference between original fees and actual fees divided by the original fees. The formula used to calculate discounts is:

$$\text{Discount} = \frac{(\text{Original fee} - \text{Actual fee})}{\text{Original fee}} \times 100\%$$

EDA analyses may require account master data. Therefore, retention data and customer master data need to be joined so that related data elements can be matched. For example, while each client exists only once in the customer master data, each phone call to negotiate discounts creates another item in the retention data set. These many-to-one data sets can be joined based on this relationship. The joining process uses the account sequential number field as it exists in both data sets and is the unique identifier in the Visual Basic for Applications (VBA) data.

3.3. Applied EDA Techniques

In this case study, traditional EDA techniques, such as descriptive statistics, data transformation, and data visualization techniques are mainly used to explore the data. Descriptive statistics used in this study include frequency distribution, summary statistics (mean and standard deviation), and categorical summarization. Data transformation is achieved by the logarithm function. Applied data visualization techniques involve pie charts, bar charts, linear charts, and scatter plots.

4. Results and Discussion

4.1. Policy-violating Bank Representatives and Negative Discounts

4.1.1. Conventional Audit Procedures. To determine whether bank representatives are violating bank policy, the maximum discount that each bank representative is allowed to offer according to bank policy must be determined. The bank policy allows bank representatives to offer discounts up to 100% of the annuity to retain the customer, so the conventional audit procedure to test this audit objective is to check whether any bank representatives offered more than 100% discounts. Internal auditors can perform this test simply by applying a filter to select all of the records with discounts greater than 100%. In this case, this filter returned

Table 2. Descriptive Statistics of Discounts.

Field Name	Mean	Median	Minimum	Maximum	Std. Deviation
Discount	−2,326.04%	60%	−27,944,522.22%	100.00%	219,933.88%

no records, indicating that no bank representative violated bank policy. Thus, this audit objective is confirmed by a conventional audit procedure. Auditors can check this box on their checklist and move to the next one.

4.1.2. EDA Process. The first step of EDA is to display the distribution of related fields. As bank representatives' discount-offering behaviors are the main concern of the bank, the analysis begins with some descriptive statistics: mean, median, minimum value, maximum value, and standard deviation of the discounts offered by the representatives. The results are shown in Table 2.

According to these results, the maximum discount offered by the bank representatives is 100% of the annual fee. Using this number, the same conclusion can be drawn: No bank representatives offered more than 100% discount. Thus, no bank representative violated bank policy.

This table also shows that the minimum discount is a large negative value (−27,944,522.22%). The mean is also negative (−2,326.04%), which implies that negative discounts overwhelm positive discounts. In addition, the median discount amount is positive (60%) indicating that half of the discounts are larger than 60% and half of the discounts are smaller than 60%. These statistics imply the existence of a few extremely large negative discounts. The frequency distribution of discounts, shown in Fig. 1, also reveals that only 0.15% (286) discounts are negative.

According to the formula for discount, a negative discount means that the actual fee after negotiation is higher than the original annual fee. A negative discount, especially a large one, is counterintuitive.

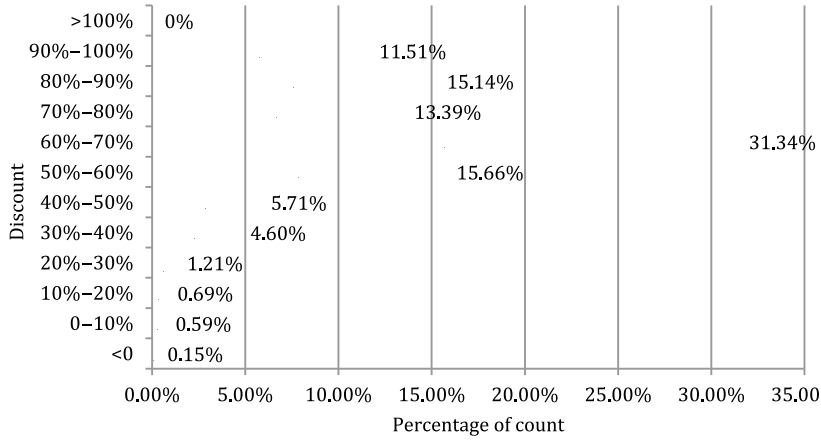


Fig. 1. Frequency Distribution of Discounts.

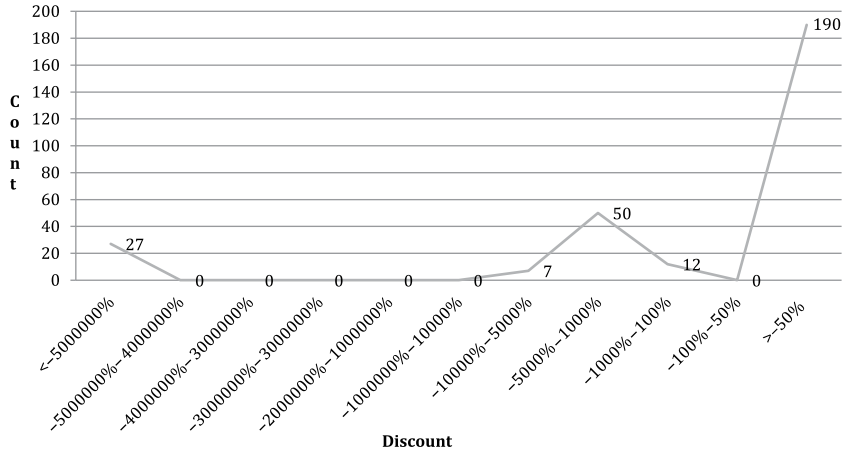


Fig. 2. Distribution of Negative Discounts.

A discussion with the bank's internal auditors yielded a potential explanation: In some cases, a group of people (e.g., a family) has the same credit card account in the form of primary cards and additional cards. If one of these customers called to negotiate the prices for the whole group, the actual fee may reflect the total actual fees of the group. As the actual fee for all cards may surpass the original fee for one card, the negative discounts may be due to group discounts offered to clients with more than one credit card.

To gain insight into negative discounts, the frequency distribution of negative discounts is calculated and displayed in Fig. 2.

Fig. 2 demonstrates a multimodal and discontinuous distribution of negative discounts featuring three separate clusters. The first cluster contains 27 records (10%) with extreme discounts (lower than −5,000,000%). The second cluster includes 69 data points (24%) associated with relatively significant discounts (between −10,000% and −100%). The third and largest cluster involves 190 records with small discounts (less than −50%). The negative discounts in this cluster may due to group discounts. However, this explanation cannot apply to the negative discounts in the other two clusters because of their extreme values. Therefore, these 96 records in the first and second clusters are considered as suspicious cases that may be attributable to errors or frauds.

Even though the remaining 190 cases have reasonable discounts, they are not necessarily group discounts. An easy verification for these data points is to determine whether a given client has multiple cards. The results of this verification are shown in Fig. 3.

Fig. 3 shows that 39 of these 190 clients (20.5%) have only one credit card, so their negative discounts cannot be explained by the effect of group discounts. Therefore, these 39 accounts are also considered suspicious.

As original fees and actual fees are the two determining factors in calculating discounts, the relationship between negative discounts and these two figures is

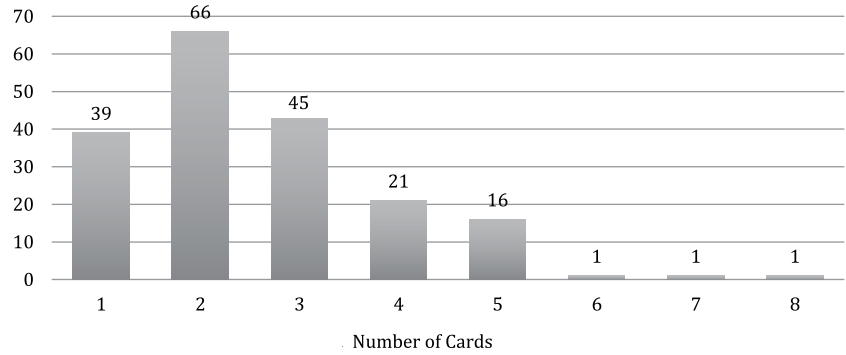


Fig. 3. Frequency Distribution of Number of Cards for the 190 Cases with Reasonable Negative Discounts.

examined to investigate the cause of this distribution. As the ranges of the variables are very wide, the values need to be transformed to another scale to display the data satisfactorily. Specifically, the values of original and actual fees are transformed to their logarithmic values. Due to the negative values, the logarithmic value of the absolute value of negative discounts is calculated. Scatterplots are then used to display the relationship between discounts and actual and original fees, shown in Fig. 4.

Fig. 4 reveals three clusters of negative discounts that are evenly distribute among the original fees. The same three clusters can also be observed in the scatterplot of discounts and actual fees. Hence, the new hypothesis is that these large negative discounts are caused by irregular actual fees.

As the number of extreme negative discounts is manageable, a substantive test is performed to investigate the specific reason for these extreme negative discounts. Among these 96 cases, 27 negative discounts are due to obvious input errors (e.g., dates are mistakenly input as the actual fees). The other 69 negative

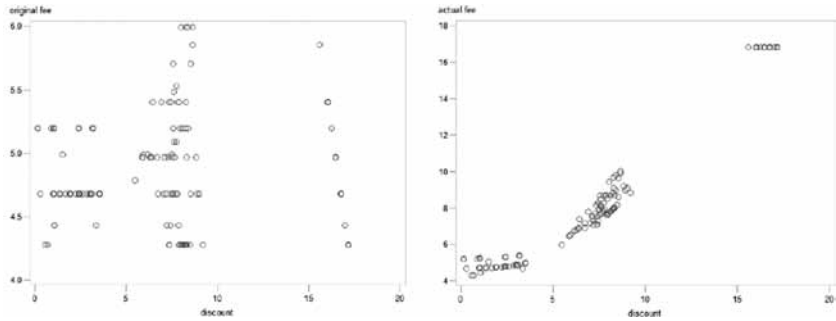


Fig. 4. Relationships between Negative Discounts and Original and Actual Fees.

discounts are caused by round, unreasonably large actual fees. These records may also include input errors, such as incorrect placement of a decimal point.

The analysis of extreme negative discounts points out some risks in the bank's internal control system. For example, the system should have a control to restrict the input format of each variable so that date format cannot be input into the actual fee field. By setting the upper and lower boundary of each field, the risk of unreasonable, extreme values can be moderated as well. These recommendations and the results of this analysis were reported to the bank, and a new audit objective was developed: Actual fees were recorded correctly.

Based on the EDA analysis, the 39 suspicious cases with reasonable negative discounts were reported to internal auditors for further investigation, and another new audit objective was proposed: Negative discounts have been offered to clients with multiple cards. Overall, using the EDA process to test this audit objective identified 135 abnormal cases, whereas no anomaly can be identified using conventional audit procedures. In addition to identifying these exceptional cases, EDA helped to generate two new audit objectives and suggest two new internal control functions.

4.2. Lazy and Inactive Bank Representatives

4.2.1. Conventional Audit Procedures. In addition to identifying representatives who violate policy, the bank also wants to identify representatives who make no effort to reduce the discount offered below 100% (i.e., "lazy" representatives). Internal auditors can use conventional audit procedures to calculate the ratio of 100% discounts to all discounts offered by each bank representative. The distribution of this ratio is shown in Fig. 5. Internal auditors can identify lazy representatives by setting a ratio threshold of acceptability. For example, if bank representatives who offer 100% discounts in more than half of their total phone calls are defined as lazy, Fig. 5 shows that internal auditors would identify 59 such representatives who have ratios greater than 0.5.

4.2.2. EDA Process. In the EDA process, the representatives who offered 100% discounts are first identified as they are the main concern of this audit

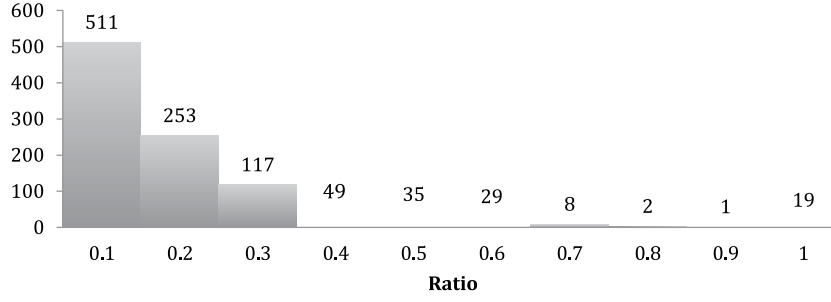


Fig. 5. Frequency Distribution of the Ratio of 100% Discounts to All Discounts Offered by Each Bank Representative.

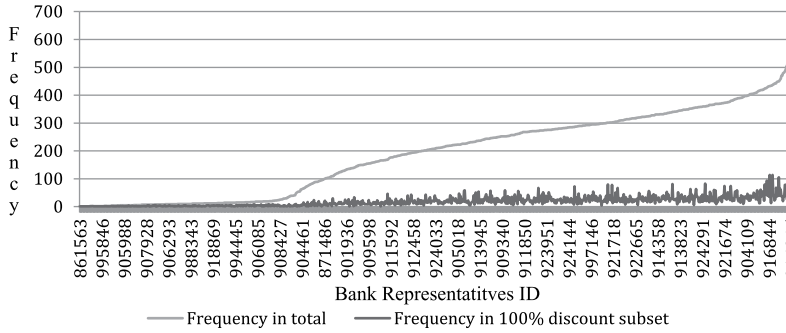


Fig. 6. Distribution of Bank Representatives Who Offered 100% Discounts in the Whole Retention Data and the 100% Discount Subset.

objective. Among all 1,151 representatives, 1,024 representatives offered 100% discount at least once. A comparison of these representatives' appearances in the full retention data set and the 100% discount subset is shown in Fig. 6.

Fig. 6 demonstrates that, in general, no bank representative offers unusually large numbers of 100% discounts; the number of 100% discounts is roughly proportional to the number of total discounts offered by a representative. However, the number of calls answered by these bank representatives varies significantly, with some representatives' call frequencies very close to zero. It is unlikely that a bank representative would answer very few calls during a month. To help detect these abnormal agents, descriptive statistics of the bank representatives' frequency distribution in the retention data are shown in Table 3.

Table 3 reveals that the 1,151 bank representatives answer an average of 170 calls. Some representatives only answered one call during the whole month, while others answered up to 623 calls. The representatives who answered only one or very few calls throughout the month are obviously anomalous. Statistically, anomalies can be defined by comparing the mean and standard deviation (Beckman & Cook, 1983). Therefore, the 403 representatives answering 22 or fewer calls (170–148) are considered suspicious.

According to the bank, a potential explanation for these representatives is that they are supervisors. It is reasonable that supervisors answer so few calls because they only deal with important and/or troublesome calls. Consequently, the hypothesis generated in this EDA process is that bank representatives who answered few phone calls are supervisors.

Table 3. Descriptive Statistics of the Frequency Distribution of Bank Representatives.

Mean	Std. Deviation	Minimum Value	Maximum Value	Count
170	148	1	623	1,151

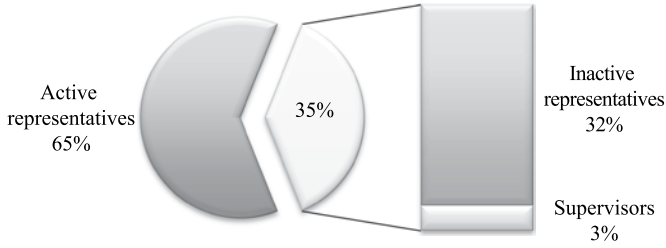


Fig. 7. Distribution of Bank Representatives.

After comparing these 403 representatives' IDs with supervisors' IDs, 33 of them are confirmed as supervisors. This leaves 370 still-suspicious representatives. Therefore, among the 1,151 bank representatives, 748 (65%) of them are active and 403 (35%) are inactive, among which 33 (3%) are supervisors and 370 (32%) are suspiciously inactive representatives. The distribution of bank representatives is shown in Fig. 7.

To locate the cause of this issue, the distributions of inactive and active representatives in different customer service centers are compared in Fig. 8, which reveals that 85.95% of the inactive representatives are concentrated in Sao Paulo, an amount disproportionate to the total number of representatives in that city. After reporting this finding to the bank's internal auditors, they suggested another potential explanation for these inactive bank representatives: Inactive bank representatives may be interns. Because Sao Paulo is the largest customer service center, it has more interns than the other customer centers. Therefore, the hypothesis created in this step is that non-supervisory inactive bank representatives are interns.

Because of limited access to other supporting audit evidence, a test of the hypothesis that non-supervisory inactive bank representatives are interns cannot be performed. Even so, the development of the hypothesis can still enhance the

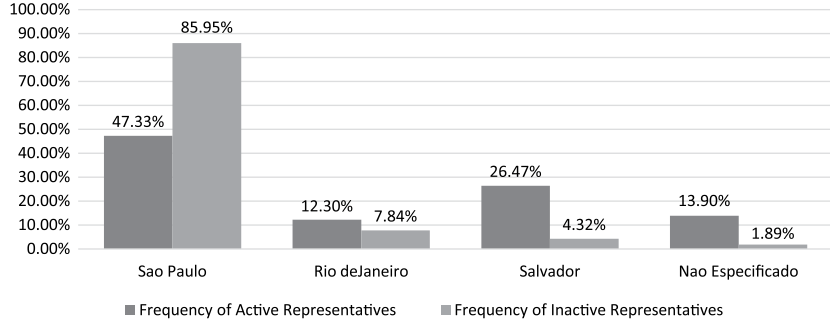


Fig. 8. Distributions of Active and Inactive Representatives in Different Customer Service Centers.

internal auditors' analysis. Once the hypothesis is confirmed with additional data, a new audit objective can be developed: All non-supervisory permanent bank representatives were active bank representatives.

The EDA process has revealed 370 inactive bank representatives, most of whom are not identified as suspicious when conventional audit procedures are used. Therefore, compared to conventional audit procedures, EDA allows internal auditors to obtain a more comprehensive and accurate set of anomalies. In addition, EDA discovers that a potential cause of these inactive representatives is related to the customer service center location.

4.3. Non-Negotiating Bank Representatives and Short Calls

4.3.1. Conventional Audit Procedures. The third type of representative of interest does not negotiate with clients but offers an immediate discount instead. As these calls should have relatively short durations, internal auditors can sort the call duration field to find unreasonably short calls (e.g., less than 60 seconds). This conventional audit procedure identifies 28,027 unreasonably short calls handled by 933 bank representatives.

4.3.2. EDA Process. In the EDA process, some descriptive statistics for the call duration field are calculated to display its distribution. The results are shown in Table 4.

According to the results, the shortest call lasts only 10 seconds and the longest call lasts 6,561 seconds (109 minutes, 21 seconds). This wide range impedes the display of the call duration frequency distribution. The average duration is 255 seconds and the median is 206 seconds, indicating that there are more short calls than long calls. In addition, 90% of the calls are less than 514 seconds, so the frequency distribution analysis focuses on the calls that last less than 600 seconds, shown in Fig. 9.

From the distribution in Fig. 9, two peaks are observed. One peak is between 2 and 3 minutes and the other is between 20 and 60 seconds. It is reasonable for a customer to negotiate a credit card annual fee discount with bank representatives for 2 to 3 minutes. However, it would seem impossible for a representative to finish a bona fide negotiation within 60 seconds. Therefore, the salient feature in this distribution is the abnormal peak between 20 and 60 seconds.

One possible hypothesis for these unreasonably short calls is that they were dropped, or accidentally disconnected due to network problems. In this case, they are ineffective phone calls, and no discount should have resulted.

As each call in the retention database is related to a discount, no discount should be associated with these calls. There are 28,027 calls in the retention data

Table 4. Descriptive Statistics of Call Duration.

Minimum	Maximum	Mean	Median	90th Percent	Count
10	6,561	255	206	514	195,694

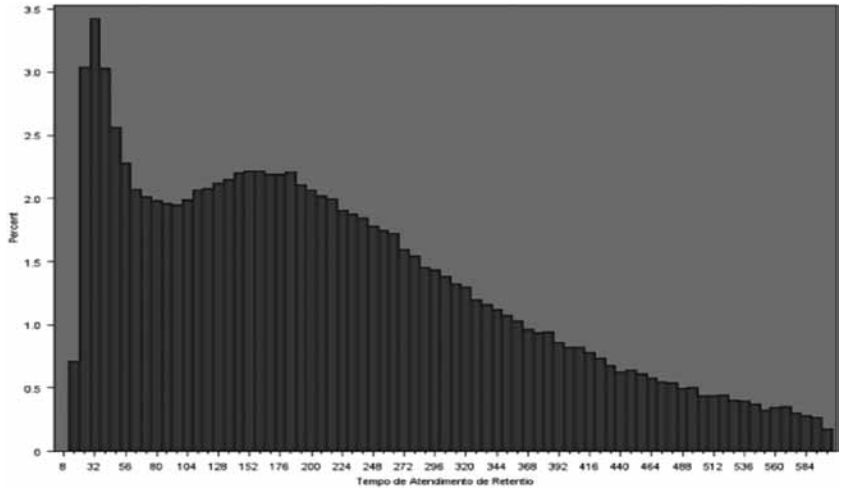


Fig. 9. Frequency Distribution of Call Duration Less Than 600 Seconds.

set under 60 seconds each. Only 121 short calls resulted in no discount; the other 27,906 short calls have non-zero discounts, so they are considered suspicious.

As was true with the previous audit objectives, it is not possible to identify the causes of these suspicious cases directly due to limited data access. Therefore, these findings are reported to the internal audit group for further investigation. After these suspicious cases are confirmed as irregularities, a new audit objective can be developed: All effective phone calls lasted more than one minute.

The findings from EDA are generally consistent with those from conventional audit procedures. Therefore, in addition to being utilized to explore hidden risk areas, EDA can also be used to confirm the results of conventional audit procedures as supplementary analysis or to replace conventional audit procedures as a standalone examination.

5. Conclusion

Using real data sets from an international bank in Brazil, this chapter provides an example of how internal auditors can apply EDA in an operational audit to assess internal control risks and detect fraud. This field study shows the risk assessment results of both conventional audit procedures and the EDA process. Comparing the two sets of results demonstrates the incremental contribution of the EDA process in uncovering risk areas that cannot be detected by conventional audit procedures.

Specifically, the data sets consist of information related to phone calls made by clients intending to negotiate their credit card annual fees. The original audit objectives relate to identifying bank representatives who may cause the bank to lose revenue. These representatives offer either higher discounts than allowed, the highest allowable discount without making an effort to negotiate a lower discount, or discounts without any negotiation. Conventional audit procedures

allow auditor to identify representatives who always offer maximum discounts and those who offer discounts without any negotiation. After applying an EDA process, hidden problems and additional abnormal cases are detected:

(1) 286 discounts are negative.
(2) 96 discounted annual fees were incorrectly input into the system.
(3) 39 group discounts were issued to customers having only one credit card.
(4) 370 bank representatives answered less than one call per day on average.
(5) 27,906 phone calls that resulted in discounts lasted less than one minute.

Based on these findings, four new audit objectives can be developed and added to the existing audit objectives to improve the audit quality:

(1) Actual fees were recorded correctly.
(2) Negative discounts have been offered to clients with multiple cards.
(3) All non-supervisory permanent bank representatives were active bank representatives.
(4) All effective phone calls lasted more than one minute.

One limitation of this field study is that a complete EDA processes could not be performed for the second and third audit objectives due to limited data access. In reality, internal auditors do not have this hindrance. Another limitation of this case study is its potential lack of generalizability. However, this issue affects only the results that are specific to the bank, not the general conclusion that EDA can identify abnormal cases and risk areas that conventional analytical procedures cannot detect.

References

Beckman, R. J., & Cook, R. D. (1983). Outlier…….s. *Technometrics*, *25*(2), 119–149.
Lane, D. C. (1983). The operational audit: A business appraisal approach to improved operations and profitability. *Journal of Operational Research Society*, *34*(10), 961–973.
Liu, Q. (2014). *The application of exploratory data analysis in auditing*. Doctoral dissertation, Rutgers University Graduate School, Newark, NJ.
Muermann, A., & Oktem, U. (2002). The near-miss management of operational risk. *The Journal of Risk Finance*, *4*(1), 25–36.
Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

This page intentionally left blank

Chapter 2

Audit Analytics: A Field Study of Credit Card After-sale Service Problem Detection at a Major Bank

Jun Dai, Paul Byrnes, Qi Liu and Miklos Vasarhelyi

1. Introduction

The Institute of Internal Auditors defines audit analytics as "the process of identifying, gathering, validating, analyzing, and interpreting various forms of data within an organization to further the purpose and mission of internal auditing" (Lambrechts, Lourens, Millar, & Sparks, 2011). Based on data analytics methods, audit analytics can drill down within the data to be audited, learn data patterns, and extract the potential business problems hidden within the data to facilitate internal auditing. Audit analytics can be applied throughout the audit process and can provide other value-added consulting activities (Lambrechts et al., 2011). Initially, audit analytics generally utilized simple analytical tools, such as descriptive statistics, linear regression, and probit analysis, to deal with internal auditing issues. However, advanced data analytical tools derived from artificial intelligence (AI), data mining, and machine learning are now being employed in the audit analytics domain. This shift is primarily attributable to the limited ability of traditional models to handle big data. In today's complex business environment, firms collect and generate extremely large amounts of data every day. Thus, it is desirable and often necessary for internal auditing techniques to have the capability to deal with large and high-dimensionality data sets. In addition, the growing demand of continuous monitoring and continuous auditing of this data requires high-efficiency approaches that are able to capture irregularities in real time. Because of this shift, the field of audit analytics is drawing increased attention from auditing researchers.

Given that audit analytics is an emerging set of techniques in the internal auditing profession, it is desirable to develop a general and systematic protocol for these processes. Such a protocol may provide relevant guidance for researchers and practitioners by offering a synopsis of procedures for applying audit

Audit Analytics in the Financial Industry, 17–33

doi:10.1108/978-1-78743-085-320191002

analytics tools to capture inefficient processes and to perform internal auditing activities, including detecting fraud and errors. Following this guidance, researchers and practitioners should be positioned to analyze internal auditing concerns more effectively and to discover potential problems hidden within the data being audited. Furthermore, this protocol should establish a foundation that can facilitate future research relative to the development of more detailed procedures or methods for dealing with internal auditing issues. Finally, the audit analytics protocol may help to extend the development of the audit analytics discipline so that students, researchers, and practitioners can more easily learn the approach of investigating internal audit issues through the lens of audit analytics.

In this chapter, a general audit analytics protocol is first developed for investigating internal auditing issues. The protocol contains eight steps: (1) identifying business scenarios; (2) defining audit concerns; (3) understanding audit data; (4) preparing the data; (5) selecting methods; (6) analyzing the data; (7) presenting and explaining the results; and (8) concluding the process. Next, this chapter presents a field study conducted to detect real-world credit card after-sale service problems at a major bank. The basic scenario involving credit card after-sale service entails customers calling the bank to request reductions of their card fees, and bank representatives offering discounts to retain the associated customers' accounts. During this process, problematic behaviors might occur. For example, some bank representatives may consistently offer abnormally high discounts to increase their own sales, even though such behavior may decrease the bank's profit. Such counterproductive behaviors should be detected so that actions can be taken to maintain the bank's profit margin moving forward. Therefore, the audit analytics process first considers auditing concerns from the bank's perspective, and then continues by extracting other potential auditing concerns behind the business process. In this field study, the protocol developed in the first step is strictly followed, and further detailed procedures and methods are then created specifically for the issues related to auditing credit card after-sale service.

The primary objectives of this chapter are to develop a general audit analytics protocol as guidance for researchers and auditors and to establish a foundation for developing detailed procedures to apply audit analytics relative to specific audit concerns. A secondary objective of this chapter is to evaluate the efficiency and adaptability of the general protocol through a field study. The results show that audit analytics procedures are an efficient way to identify audit-relevant information that cannot be detected easily by internal auditors using traditional audit protocols.

2. Related Work

Early audit analytics research mainly focuses on simple analytical methods to deal with auditing issues. This is particularly true in research prior to 2000. During that period, most audit analytics models are based on descriptive statistics, logistic regression, and probit analysis tools. For example, Persons (1995) utilizes stepwise-logistic models to examine how some financial factors are associated with fraudulent financial reporting. Chen and Leitch (1999) analyze the

relative productivity of four audit analytical methods, including Martingale, Census X-11, ARIMA, and a stepwise regression expectation model in simulated business and economic environments. Nigrini and Miller (2009) describe a new second-order test of Benford's Law for checking the authenticity and reliability of transaction-level accounting data. Kaminski, Wetzel, and Guan (2004) examine whether the financial ratios in fraudulent companies differ from those in non-fraudulent organizations, and their results provide empirical evidence that financial ratios have limited ability to detect and/or predict fraudulent financial reporting. Even though simple analytical methods have been widely studied for investigating auditing issues, these approaches have several limitations, such as their inability to analyze data sets that have high dimensionality and/or missing values, as well as limited ability in handling non-linear, separated data sets.

Some audit analytics studies focus on applying more complex techniques to auditing issues. This is attributable to the limitations of simple analytical methods and the increasing applicability of AI, data mining, and machine learning methods within the auditing domain. These more sophisticated analytical tools essentially fall into two categories: supervised learning and unsupervised learning techniques.

Supervised learning entails building models based on the use of a training data set that contains class label information. Commonly used supervised learning techniques include neural networks, expert systems, and support vector machines. Fanning and Cogger (1998) use a self-organizing artificial neural network with standard statistical tools to investigate whether publicly available attributes are useful in predicting fraudulent financial statements. Hornik and Ruf (1997) investigate the explanatory capabilities of expert systems in transferring knowledge to novice auditors. Pai, Hsu, and Wang (2011) introduce a support vector machine-based fraud warning model to reduce the risks of fraudulent financial statements. Kirkos et al. (2007) explore the effectiveness of some popular classification techniques, including decision trees, neural networks, and Bayesian belief networks, for detecting fraudulent financial statements and identifying the factors for fraud detection.

By contrast, unsupervised learning techniques, such as the self-organization map (SOM), hidden Markov model (HMM), and clustering, do not require training data sets with class label information. The SOM is used to classify and cluster input data, detect and derive hidden patterns in that data, and act as a filtering mechanism for additional layers. For example, Quah and Sriganesh (2008) use this method to decipher, filter, and analyze customer behavior for fraud detection. Srivastava, Kundu, Sural, and Majumdar (2008) use an HMM model to detect frauds by modeling the sequence of operations in credit card transaction processing. Thiprungsri (2010) examines the possibility of using clustering techniques in fraud detection and utilizes cluster-based outliers to evaluate group life insurance claims.

Supervised and unsupervised learning techniques are effective in accommodating data sets containing high dimensionality and missing values, and they can discriminate abnormal values from normal values via non-linear separation. Compared to supervised learning techniques, unsupervised learning approaches are better able to detect financial fraud in real time because they are scalable and capable of recognizing new forms of fraud. The main disadvantage of unsupervised learning techniques is that they are vulnerable to noise elements.

Although researchers have developed audit analytics methods for detecting financial fraud and errors and for verifying financial statements, few studies have focused on formulating general audit analytics protocols. In this chapter, a general audit analytics protocol will be developed. Then, the effectiveness of this protocol will be examined by applying it to credit card sales problems at a major bank.

3. Audit Analytics Protocol

This section analyzes the scope of internal auditing issues that can be analyzed by audit analytics methods, and creates a general audit analytics protocol for investigating internal auditing issues.

3.1. Scope of Internal Auditing Issues for Audit Analytics

Most audit analytics methods are derived from data analytics tools, such as statistical and data mining methods. These tools have good scalability and can produce accurate results on large data sets. For example, statistical approaches require at least 30 objects to achieve reliable results, and data mining has the capability of training on many thousands of records in constructing an accurate detection model. Commonly used data in audit analytics are transaction data sets, such as accounts receivable and accounts payable ledgers, and these files usually contain numerous records.

Audit analytics mainly uses data analytical tools to verify whether auditing data is meaningful for a chosen set of criteria. A primary objective of audit analytics is to detect abnormal data that does not adhere to some auditing expectation. This is done by comparing records and discriminating abnormal data from normal data. Audit analytics methods may also monitor incoming data and detect errors or fraud in real time to achieve the aims of continuous monitoring and continuous data auditing. However, audit analytics can only deal with auditing issues that are embedded within electronic data. For instance, audit analytics loses its power in field audits. In addition, audit analytics methods need to learn patterns from financial data to detect potential associated problems. Thus, the strength of audit analytics is mainly for data analyses other than simple data verification routines. For instance, audit analytics methods cannot be used to verify whether an entry in an account payable ledger is consistent with the corresponding value on the invoice.

3.2. A General Protocol for Audit Analytics

The general protocol for audit analytics developed in this study has eight steps: (1) identifying business scenarios; (2) defining audit concerns; (3) understanding auditing data; (4) preparing the data; (5) selecting methods; (6) analyzing the data; (7) presenting and explaining the results; and (8) concluding the process. A discussion of each step follows.

Identifying business scenarios: In this initial step in the audit analytics protocol, it is necessary to identify the scenarios, understand the business process

associated with internal auditing concerns, and decide whether audit analytics methods are applicable in a given case by referring to the scope of internal audit issues for audit analytics. If so, estimations must then be made relative to which auditing issues can be analyzed via audit analytics methods, which methods are likely to be most appropriate, and whether those methods can be implemented. For example, if a company sells accounting software to large firms through its salespeople, audit analytics can be applied to control and detect the behaviors of salespeople by analyzing their sales performance data stored in computers, such as individual sales amounts and total monthly sales amounts. In addition, by obtaining a better understanding of the firm's business scenarios, researchers and auditors are likely to discover the areas in which errors and abnormalities are most likely to occur, which can further assist in generating audit concerns. To complete this stage, it is also necessary to identify the relevant data sets of interest and to extract the data for actual analytical purposes.

Defining audit concerns: Based on a comprehensive understanding of the business scenarios, researchers and/or auditors define audit concerns in this step. Audit concerns are the potential problems existing in the business scenarios, which might decrease the efficiency of the company or market. For example, salespeople may give discounts to customers to entice them to purchase more products, but if these discounts exceed the maximum allowable amount established by the company, they will decrease the company's profit margin. As such, those abnormally high discounts are problems in the sales process and should be detected. Compared to the traditional internal auditing process, audit analytics can increase the breadth and depth of audit coverage due to its data processing and analysis capabilities. Therefore, the audit concerns defined by audit analytics can have a broader scope than traditional audit concerns. Furthermore, the audit concerns defined by audit analytics are often based on transaction data, which contains original data patterns that may be hidden through aggregation, rather than being based on the aggregated data used in traditional audits. Such audit concerns may be related to some potential data-driven indicators that can be used to identify existing or emerging risks or inefficient processes. These audit concerns may initially be obtained from the firm's managers. Then, the auditors and/or researchers can formulate additional audit concerns based on personal experiences and interests. After preliminary generation of audit concerns, researchers or auditors should examine each audit concern to determine whether it can be analyzed through audit analytics methods, and then define the final set of audit concerns.

Understanding the auditing data: Once the auditing concerns are defined, researchers and auditors need to scan the auditing data sets in an effort to understand the data structure, as well as the meanings and characteristics of the variables involved. There are two methods available for understanding auditing data: (1) scan the entire data set and understand each variable and (2) analyze the auditing concerns, identify the relevant attributes associated with each one, and investigate the pertinent variables from the data set in a detailed manner. The merit of the first method is that it can offer a "bird's-eye view" of the data set. Thus, researchers or auditors can consider each variable when analyzing the auditing

issues. However, this method is inefficient when the number of variables is large. By contrast, the second method is more efficient for dealing with data sets that contain many variables. However, there is a risk that relevant variables may ultimately be omitted from the analysis, creating information loss. Thus, careful considerations should be made prior to selecting a method for understanding a given data set. In today's complicated business environment, data sets used for auditing may have numerous variables, but often have only a few dimensions that are relevant or valuable for analyzing auditing concerns. In this case, the second method will likely be a more desirable and efficient way to understand auditing data.

Preparing the data: As the majority of audit analytics methods are data-driven, it is critical to prepare auditing data before implementing any auditing methods to ensure data integrity and validity. This step can be divided into three categories: data transformation, data cleaning, and data description.

(1) *Data transformation*: As audit data is often exported from a company's ERP system, the format of the data file may not be recognized by audit analytics software. Furthermore, the structure of the original data set may not be suitable for implementation of audit analytics techniques. Thus, it may be necessary to transform raw data sets into appropriate data formats.
(2) *Data cleaning:* Auditing data sets may contain certain values that are meaningless with regard to the corresponding variables. For instance, a negative sales price in a transaction data set might be an input error. Researchers or auditors should detect those meaningless values and either ask the company to correct them or exclude them from further analysis.
(3) *Data description*: Auditors or researchers need a clear understanding of each selected variable. The most common method for understanding an attribute is to formulate its descriptive statistics, such as mean, frequency, maximum, minimum, and standard deviation. In this way, researchers and auditors may not simply have an overview of the selected variables, but also discover abnormal values, such as extremely high maximum amounts or huge standard deviations.

Selecting methods: Based on audit concerns and available auditing data, researchers and auditors may then select the audit analytics methods to address specific audit concerns. The first part of this step entails a literature review to discover which methods have been used previously to deal with similar audit concerns. The second part of this step involves an analysis of those previous methods with the goal of selecting the appropriate approaches. During this procedure, the chosen methods may be tailored or otherwise improved for the specific audit concerns of interest. Usually, simple analytical tools are initially selected, such as distribution analysis, outlier detection, or graph analysis. Then, more complex analytical tools (e.g., clustering or a classification model) are utilized to identify hidden patterns in the auditing data and build models to detect abnormal data in real time. As various analytical tools may not be equally suitable for different types of data, auditors and/or researchers can select several analytical tools for a given audit concern, so that they can compare and combine results to ensure the reliability of the final results.

Analyzing the data: After deciding on the relevant methods, researchers and auditors can import the auditing data into the appropriate software and implement the chosen methods. In some cases, parameterization of those methods must be conducted to achieve optimal results. If the software has limited capability to meet the requirements of the auditing concerns, researchers and auditors should implement the methods by programing according to the specific requirements of the auditing concerns.

Presenting and explaining the results: This is a key step of the protocol. After analyzing the auditing data, researchers and auditors should explain and present the results. In many cases, the results of analytical methods have greater meaning outside of the business realm in the statistical domain. Researchers or auditors must be aware of this, so that they can explain and present the findings in terms that are applicable to the business context.

Concluding the process: The final step is to complete the process by summarizing the results and generating audit comments for the company. In this step, auditors may also refer to the results from other auditing activities.

4. Field Study Description

The field study in this chapter is undertaken at a major bank in South America. This bank is one of the 10 largest in the world based on market value. The bank's credit card after-sale service is the specific focus of the study. The credit card after-sale service process is a typical business process in most banks around the world, which makes the field study results more generalizable. After-sale service is a crucial procedure in the credit card business because the focus of credit card sale competition is not only on acquiring new customers, but also on retaining existing customers. Furthermore, this process requires extensive auditing to increase efficiency and reduce costs. In the field study, the audit analytics protocol discussed in Section 3 is applied in an effort to detect potential problems that may exist in the credit card after-sale service process.

The general credit card after-sale service scenario entails credit card customers calling the bank to request reductions in their card fees. In response, bank representatives may offer discounts to those customers in an effort to retain their accounts. The transaction data set used in this field study contains data related to credit card after-sale service events in January 2012, and includes relevant customer call and fee reduction data. The data set has 195,694 records in total, and each entry contains 162 variables.

5. Implementing the Audit Analytics Protocol

5.1. Identifying Business Scenarios

Discussions with the bank manager generated an overview of the credit card after-sale service process. Some customers who currently hold credit cards call the bank to ask for reductions of their card fees. This situation generally occurs when a customer is willing to cancel his/her account and leave the bank. If the bank loses this customer, the bank's profit will likely decrease because each credit card customer

often provides multiple sources of revenue to the bank, including card fees, interest on outstanding balances, and merchant income as a function of the customer's purchasing activity. Therefore, bank representatives often have an incentive to retain the customer by offering him/her a discount on card fees. However, excessive reduction of card fees will also reduce the bank's revenue. Thus, such extreme situations and other counterproductive behaviors should be identified in an effort to serve the best interests of the bank. A bank supervisor makes the decision whether to give discounts when the bank representatives cannot decide. The transaction data collected during the fee negotiation process is stored in an ERP system and is exported into a tab data file. A data dictionary explaining each variable in the data file is available. Therefore, it is feasible to apply audit analytics methods to the data files after transforming them into an appropriate data format.

5.2. Defining Audit Concern

After understanding the business process of the credit card after-sale service, audit concerns from the bank are used as a point of origin. Then, the business process is disaggregated to identify additional audit concerns. The main auditing concerns for the credit card after-sale service process include the following:

(1) Some discounts are meaningless (e.g., negative discounts). These are considered to be errors.
(2) Some bank representatives consistently offer high discounts. This behavior may increase individual sales performance, but will have a negative impact on the bank's bottom line.
(3) For each type of customer, there might be a certain range of optimal discounts that yields the most benefits for the bank. It is beneficial to estimate such optimal discount ranges to maximize the bank's utility.
(4) Some bank representatives offered unreasonably small numbers of discounts (less than 50) during January 2012. However, it is possible that these bank representatives are actually supervisors. The duty of supervisors is to provide discounts when the bank representatives are not able to do so. Thus, it may be reasonable to assume that supervisors only offer discounts to a few customers compared to bank representatives. Nevertheless, it is necessary to examine whether the bank representatives who offer only a few discounts are indeed supervisors.
(5) Some bank representatives offer discounts too readily. According to the duration of calls in the data set, some discounts are offered within 30 seconds of call origination, which appears too brief for bank representatives to make optimal decisions.
(6) Some call centers consistently give higher discounts than others do, regardless of the call length.
(7) The duration of calls may have some influence on the discounts offered. If a bank representative spends more time with a customer, and ultimately offers him/her a high discount, this process is considered inefficient. Therefore, it is necessary to discover the relationship between the call duration and the offered discounts to assess the efficiency of the fee reduction process.

(8) Some discount offers fail to retain customers, possibly due to the small amounts offered. Reviewing the discounts offered in unsuccessful retention cases can provide guidance for future retention activities.

5.3. Understanding the Auditing Data

The transaction data set in this study contains information about all calls relative to fee reduction requests made in January 2012. The data set contains a total of 162 variables that capture the details of each call, such as call duration, representative name, initial card fee, and revised card fee. Since there are more than one hundred variables in the data set, the second method for understanding the data is chosen. Specifically, the relevant variables are identified for analysis based on the auditing concerns. From this procedure, the following variables are identified for further analysis:

(1) Original price: the original card fee for the costumer before fee reduction.
(2) Actual price: the new card fee for the customer after fee reduction.
(3) Credit card name: there are 141 credit card types in the data set.
(4) Bank representative identification: there are 1,151 bank representatives in the data set.
(5) Supervisor identification: there are 267 supervisors in the data set.
(6) Call centers: there are 4 call centers in the data set.
(7) Call time: the duration of the call in seconds.

5.4. Preparing the Data

Before applying audit analytics methods to address auditing concerns, the selected variables are extracted from the data file, and the following data preparation tasks are performed:

(1) Data transformation: The data file is exported from an ERP system at the bank in tab format. Because ACL auditing software, Excel, and SAS are used to conduct the analyses, the tab data file is initially transferred into ACL and Excel formats. Discounts are another important variable for investigating auditing concerns, but these discounts are not directly given in the data set. However, they can be calculated as the difference between the original price and the corresponding actual price divided by the original price. Thus, the discount value is computed for each entry.
(2) Data cleaning: The data set contains some meaningless and missing values. For example, there are some negative discounts, implying that the original price is larger than the corresponding actual price. Furthermore, some records have missing discount information. Such cases are considered to be input or system errors, so they are excluded. Specifically, 286 out of 195,694 records (0.15%) have negative discounts, and 38,180 out of 195,694 records (7%) have missing values in discounts. All records containing meaningless and missing values are excluded, leaving 157,228 entries for further analysis.

Table 1. Descriptive Statistics of Selected Variables.

	Call Time	**Discount**
Max	4,693	1
Min	1	0
Mean	279.6824	0.613046
Std. dev.	213.548	0.1711

(3) Data description: Descriptive statistics for selected variables are obtained, including maximum value, minimum value, mean value, frequency, and standard deviation. The results are shown in Table 1.

The descriptive statistics show that call time has a large range (1–4,693 seconds). In addition, the maximum and minimum values are considered abnormal because it is impossible to provide a discount within 1 second, and a bank representative who spends almost 80 minutes with a customer demonstrates much lower efficiency in dealing with the fee reduction issue. Thus, further analysis of the duration of calls is necessary.

5.5. Selecting Methods

A literature review is conducted in accordance with the audit concerns discussed in Section 5.2, and the following methods are chosen to resolve the audit concerns:

(1) Use an outlier detection method to locate discounts with negative values.
(2) Use a frequency distribution analysis method to identify bank representatives who consistently offer high discounts.
(3) Use a clustering model to differentiate customer types, and use the three-sigma rule[1] to estimate optimal discount ranges.
(4) Use outlier detection to discover bank representatives who offer discounts less than 50 times during the month, and examine whether they are supervisors.
(5) Use a frequency distribution analysis method to analyze discounts made within 30 seconds, and identify bank representatives who often give discounts in this short timeframe.
(6) Use graph analysis to show the relationship between the duration of calls and the discounts negotiated by each call center, and compare their differences.
(7) Use regression analysis to compare the influence of the duration of calls on the discounts negotiated across call centers.
(8) Use frequency distribution to analyze discount values in unsuccessful retentions, and use ANOVA to examine whether there is a significant difference in discount values between successful retentions and unsuccessful retentions.

[1]The three-sigma rule essentially states that for a given population, nearly all of the data points reside within ± three standard deviations of the mean. For example, in a normal distribution, about 99.7% of the data will be within ± three standard deviations of the average.

5.6. Analyzing the Data

ACL auditing software, EXCEL, and SAS are used to analyze the auditing concerns, and none of the audit analytics methods applied in this field study require parameterization. In addition, all analyses can be completed by these three programs. Thus, no additional programing is necessary in this study.

6. Presenting and Explaining Results

6.1. Negative Discount Detection

Findings: Data analysis indicates that 286 out of 195,694 records (0.15%) have negative discounts, among which 27 (0.1%) negative discounts are due to obvious input errors (e.g., the value of the corresponding actual price is in date format). The remaining 259 negative discounts require further investigation.

6.2. High Discount Analysis

Some bank representatives regularly offer discounts that are very close to the maximum (more than 90%). This study defines these as high discounts.

Findings: Out of 157,228 records with positive discounts, 10,534 (6.7%) are identified as high discounts. These high discounts were given by 1,005 (87%) bank representatives. However, within that group, seven bank representatives are identified as having more than 45% of their discounts classified as high discounts.

Data visualization: Fig. 1 shows the seven bank representatives who offer high discounts for more than 45% of their total discount offers.

6.3. Optimal Discount Estimation

Customers are clustered into groups based on their behavior score, credit limit line, age, late payments, account age, and amount of additional assets. Each cluster represents a different type of customer, and the discount patterns are analyzed

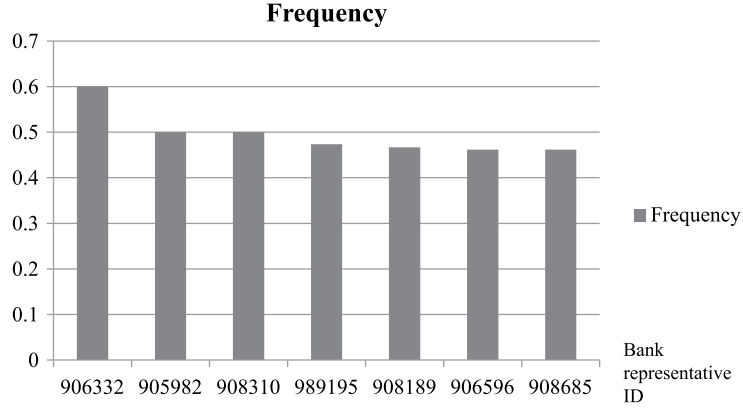


Fig. 1. Seven Bank Representatives Providing High Discounts for More than 45% of their Total Discount Offers.

within each customer type. Finally, the three-sigma rule is used to estimate the optimal discount range for each type of customers.

Findings: Seven customer clusters are generated based on the entire population, from which 146,413 successful retention records are selected to analyze the discount patterns within each customer type. The distribution of customer types and the mean, standard deviation, minimum, and maximum discount for each type are shown in Table 2.

Table 2 shows that the discounts offered to different types of customers differ only slightly. This indicates that bank representatives offer similar discounts to all customer types. Such behaviors can be considered ineffective. A more accurate and reasonable discount policy should be generated to differentiate the discounts offered to different types of customers. The estimated optimal discount ranges using the three-sigma rule are shown in Table 3. The one-sigma range (i.e., the range of one standard deviation away from the mean value) is the optimal discount range.

Table 2. Discount Distribution by Customer Type.

Analysis Variable: Discounts					
Customer type	**Frequency**	**Mean**	**Std. dev.**	**Minimum**	**Maximum**
1	7,880	0.700167	0.169475	0	1
2	28,228	0.635515	0.179615	0	1
3	39,433	0.60752	0.169302	0	1
4	14,429	0.706655	0.181224	0	1
5	12,954	0.734475	0.191367	0	1
6	34,584	0.684004	0.193415	0	1
7	8,905	0.624704	0.164819	0	1

Table 3. Optimal Discount Range Estimation Using the Three-Sigma Rule.

Customer Type	**1 σ Lower Bound**	**1 σ Upper Bound**	**2 σ Lower Bound**	**2 σ Upper Bound**	**3 σ Lower Bound**	**3 σ Upper Bound**
1	0.530693	0.869642	0.361218	1	0.191743	1
2	0.455901	0.81513	0.276286	0.994744	0.096672	1
3	0.438218	0.776822	0.268916	0.946124	0.099614	1
4	0.525431	0.887879	0.344206	1	0.162982	1
5	0.543108	0.925842	0.351741	1	0.160374	1
6	0.490589	0.877419	0.297174	1	0.103758	1
7	0.459885	0.789523	0.295066	0.954342	0.130248	1

6.4. Inactive Agents

Some bank representatives offer very few (less than 50) discounts during the entire month under analysis. These bank representatives are defined as inactive agents.

Findings: Using this definition, 436 agents are identified as inactive. Among them, only 48 are supervisors. Thus, the remaining 388 agents who are not supervisors are deemed suspicious.

Data visualization: Fig. 2 shows the percentage of active agents, inactive agents, and supervisors.

6.5. Short Call Analysis

Some calls are very short (i.e., less than 30 seconds), yet result in the bank representatives providing discounts to the customers. Those calls are considered to be suspicious because it is intuitively infeasible to complete a negotiation process adequately within 30 seconds.

Findings: Out of 157,228 records, 9,448 (6%) were identified as short calls. These calls were made by 770 (67%) bank representatives. Six bank representatives are identified as transacting more than 25% of their discounts in short calls.

Data visualization: Fig. 3 shows the proportion of short calls for the six bank representatives who offer more than 25% of their discounts in short calls.

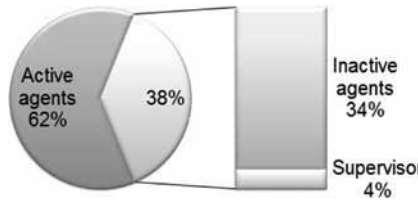


Fig. 2. Percentages of Active Agents, Inactive Agents, and Supervisors.

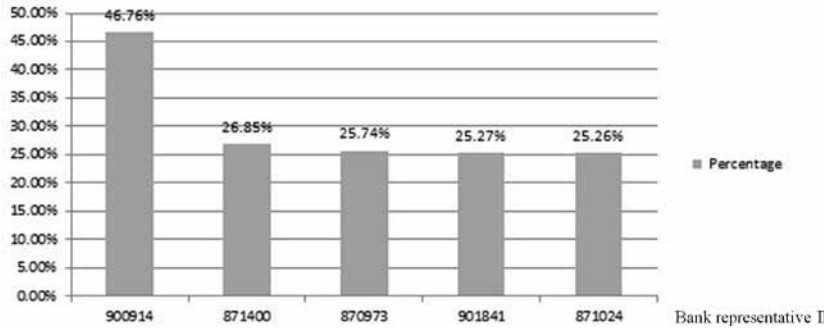


Fig. 3. Six Bank Representatives Offering More than 25% of Their Discounts in Short Calls.

6.6. Graphic Analysis of the Relationship between Call Duration and Discounts Offered by Call Centers

No matter how long the calls last, some call centers consistently offer higher discounts than others do. These call centers are considered to demonstrate low effectiveness.

Findings: For calls with duration of less than 10 minutes, the discounts offered in Call Center 4 are consistently higher than the discounts offered in other call centers. When the call time is less than two minutes, the discounts are significantly high and normally distributed. No specific pattern is recognized among calls that take more than 10 minutes.

Data visualization: Fig. 4 shows the graphic analysis of the relationship between the duration of calls and the discounts negotiated at each call center.

6.7. Regression Analysis

Regression analysis is conducted to explore the relationship between call duration and the discounts negotiated by each call center to investigate the differences in the effect of call duration on discounts among call centers. The data is normalized into standard scores before this analysis to adjust values measured on different scales to a notionally common scale. The regression results that express the relationship between the duration of calls and the discount offered at each of the four call centers are shown below:

Call Center 1: Discount = −0.6507 * call time + 2.3304
Call Center 2: Discount = −0.4825 * call time + 1.6906
Call Center 3: Discount = −0.3288 * call time + 2.2008
Call Center 4: Discount = −0.372 * call time + 1.742

These results show that call duration has a negative effect on discounts for each of the four call centers. The negative coefficients indicate that when the bank representatives spend longer time with customers, the discounts they offer are lower. The regression results also show that the duration of calls has much more influence on the discounts offered in Call Center 1 than it has in the other three centers.

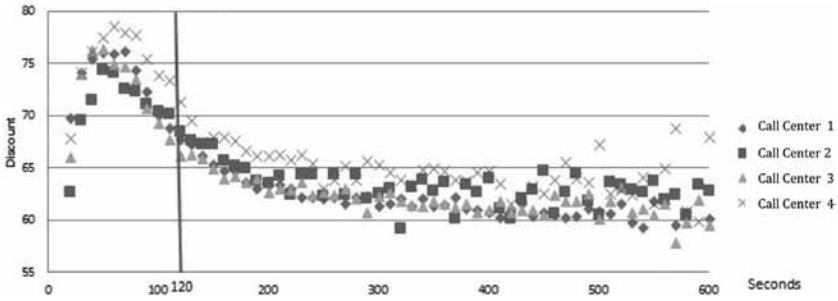


Fig. 4. Graphic Analysis of the Call Duration–Discount Relationship by Call Center.

6.8. Unsuccessful Retention Analysis

A possible reason for unsuccessful retentions might be that the discounts offered are too low to retain the customers. A frequency distribution analysis and ANOVA are conducted on the unsuccessful retentions to assess whether this is, in fact, the explanation for losing these customers.

Findings: Among the 150,177 records with account status information, 3,764 (2.5%) are identified as unsuccessful retentions. By analyzing the frequency distribution of the discounts (see Figure 5) offered in cases of unsuccessful retentions, the results show that those discounts have more relatively high values (greater than or equal to 50%) than relatively low values (less than 50%), which indicates that unsuccessful retentions are not directly attributable to low discounts.

The ANOVA results in Table 4 show that the *p*-value is significant, which indicates that the discounts offered in unsuccessful retentions are significantly different from those offered in successful retentions.

6.9. Recommendations

This bank has some problems related to its control over credit card fee reductions. A system of continuous monitoring of the discounts offered to customers can help the bank to control the behaviors of bank representatives. An accurate and reasonable discount policy should be generated as guidance for offering appropriate discounts to different types of customers. The bank also should investigate those inactive agents who are not supervisors, and determine the reasons for their inactive behavior. In addition, the bank should investigate the bank representatives who often provide discounts during calls of short duration. Finally, according

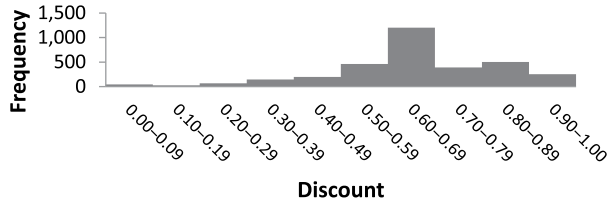


Fig. 5. Frequency Distribution of Discounts.

Table 4. Result of ANOVA on Discounts between Successful and Unsuccessful Retention.

Source	DF	Sum of Squares	Mean Square	*F* Value	*Pr*>*F*
Model	1	3.737848	3.737848	127.74	<0.0001
Error	150,175	4,394.377	0.029262		
Corrected total	150,176	4,398.115			

to an analysis of the relationship between call duration and discounts, the bank should improve the effectiveness of Call Center 4 relative to fee reductions.

7. Conclusion

Audit analytics is an emerging application of data analytics technology in the auditing profession, primarily because of its scalability in handling large data sets, adaptability for dynamic data sets, and performance with non-linear data. A general protocol for audit analytics can provide guidance for auditing professionals and enhance their understanding of the audit analytics process and tools. It also may provide a foundation for researchers to develop their own procedures or methods for investigating auditing issues, which would expand the utility of these analytical techniques.

In this chapter, a general audit analytics protocol is developed for investigating auditing issues. Following this protocol, a field study is conducted to detect credit card after-sale service problems at a major bank to explore the efficiency and adaptability of the audit analytics protocol. The results of the field study show that the protocol enables researchers or auditors to investigate audit concerns efficiently and effectively using a variety of analytical tools and uncover potential problems within the business process.

A limitation of field study research is that it often lacks generalizability. However, this issue only affects the selection of audit analytics methods and specific processes within the steps in the audit analytics protocol. It does not affect the conclusion that the general protocol provides an efficient way to conduct data analysis to resolve auditing concerns. In future work, the audit analytics protocol could be applied to more complex analytical methods in an effort to study its efficiency and adaptability.

References

Chen, Y., & Leitch, R. A. (1999). An analysis of the relative power characteristics of analytical procedures. *Auditing: A Journal of Practice and Theory*, *18*(2), 35–69.

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*, 21–24.

Hornik, S., & Ruf, B. M. (1997). Expert systems usage and knowledge acquisition: An empirical assessment of analogical reasoning in the evaluation of internal controls. *Journal of Information Systems*, *11*(2), 57–74.

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

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, *32*, 995–1003.

Lambrechts, A. J., Lourens, J., Millar, P., & Sparks, D. (2011). *Global technology audit guide (GTAG) 16: Data analysis technologies*. The Institute of Internal Auditors. Altamonte Springs, FL.

Nigrini, M., & Miller, S. J. (2009). Data diagnostics using second order tests of Benford's Law. *Auditing: A Journal of Practice and Theory*, *28*(2), 305–324.

Pai, P. F., Hsu, M. F., & Wang, M. C. (2011). A support vector machine-based model for detecting top management fraud. *Knowledge-Based Systems*, *24*(2), 314–321.

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

Quah, J. T. S., & Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. *Expert Systems with Applications*, *35*(4), 1721–1732.

Srivastava, A., Kundu, A., Sural, S., & Majumdar, A. K. (2008). Credit card fraud detection using hidden Markov model. *IEEE Transactions on Dependable and Secure Computing*, *5*(1), 37–48.

Thiprungsri, S. (2010). Cluster analysis for anomaly detection in accounting data. *Nineteenth annual strategic and emerging technologies research workshop*, San Francisco, CA.

This page intentionally left blank

Part II

Analytics in Credit Card Audits

This page intentionally left blank

Chapter 3

Automated Clustering: From Concept to Reality*

Paul Byrnes

1. Introduction

Fundamentally, clustering is a data mining technique used to find relevant patterns in data (Tan, Steinbach, & Kumar, 2006). For a particular data set, objects in a given partition are more like one another than they are like points in all other groupings. Within a cluster, there may be objects that are significantly different from the representative value, and these are identified as potential irregularities. Furthermore, if a partition contains an extremely small membership, this signals that all corresponding objects could be anomalous. Envisioning this in the domain of auditing business transaction data, identified irregularities would be investigated as a method to discover fraud.

In fact, audit standards promote the use of clustering procedures in the process of assessing fraud risk. For example, supplemental audit guidance AU-C 240 recommends the use of computer-assisted audit techniques like data mining to supplement auditors' fraud discovery activities. Furthermore, extant literature confirms the utility of clustering in an array of applications including but not limited to credit cards (Hao, Dayal, Sharma, Keim, & Janetzko, 2010), money laundering (Liu, Qian, Mao, & Zhu, 2011), and financial statements (Deng & Mei, 2009). For instance, Deng and Mei (2009) combine the use a self-organizing map and *K*-means clustering technique to analyze a set of 100 financial statements. Experimental findings show this procedure to be effective in the identification of financial statement irregularities.

Given the promise that clustering has demonstrated for detecting irregularities, it is logical that the accounting profession now advocates for this technique to be embedded within the audit process. However, clustering is currently a semi-automated procedure with multiple complex, manual decision points arising at intervals, so classically educated and trained auditors simply do not possess the requisite skill-set to use this data mining technique in a reliable, effective manner. However, if the process is fully automated, then the operational problems would

*This chapter is based on the third chapter of the author's dissertation (Byrnes, 2015).

Audit Analytics in the Financial Industry, 37–46

doi:10.1108/978-1-78743-085-320191003

vanish, and auditors would be empowered to incorporate clustering productively into their practice. In this chapter, a method for achieving this goal is discussed and implemented within the context of systematic partitioning of credit card customer data pertaining to a large international financial institution.

2. Background

Cluster analysis groups data points in a meaningful way, such that each object is comparable to items in the same cluster and different from objects assigned to other partitions (Tan et al., 2006). While this appears fairly intuitive and simple, several complexities exist in performing this operation effectively. In each instance of clustering, determining the algorithm to employ and the final model to represent the data are both non-trivial and historically manual tasks.

Many types of algorithms are available for performing the clustering function, and no single method is consistently preferred. In truth, the optimal algorithm is often contingent on the specific data being evaluated (Alpaydin, 2010). Examples of existing clustering algorithms include *K*-means, *K*-medoid, Expectation Maximization, and hierarchical approaches like Complete-Link and Ward's method. Of these, *K*-means is generally viewed as the most popular, and this is primarily attributable to its efficiency and relative simplicity (Garla, Chakraborty, & Gaeth, 2012). For instance, this method maintains linear time complexity, meaning that as a data set to be evaluated increases in volume, the time to cluster that data grows in a similar manner. This is in sharp contrast to the hierarchical methods that exhibit exponential time complexity, such that process time can quickly become intractable as data volume increases. While *K*-means demonstrates efficiency and other desirable characteristics, it is obviously not always appropriate because: (1) It assumes that the data follow a normal distribution; (2) It can have difficulty in addressing outliers; and (3) It sometimes produces empty clusters (Tan et al., 2006).

In truth, all data mining approaches have advantages and disadvantages, so an automated process for performing clustering must incorporate a meaningful combination of complementary methods. Then, for a given data set and suitable range for the number of clusters, models are constructed by the included algorithms within the context of a simulation routine. Once all solutions are built, a method for selecting the preferred model must be implemented. This is precisely where the next barrier to automated clustering arises.

Several measures have traditionally been adopted in assessing cluster quality, including objective metrics, such as the silhouette coefficient and Calinski–Harabasz measure, as well as relative approaches like "elbow" analysis in which quality is determined via recognition of the point(s) where diminishing returns occur relative to error change patterns. Often, a multi-faceted approach is utilized to identify the preferred model, and this process entails substantial manual decision making. Obviously, to automate this portion of the data mining operation, an objective measurement scheme would be necessary. However, it must first be empirically demonstrated that such a mechanism can serve as an effective proxy for model quality. In fact, Byrnes (2015) explores this issue via an extensive 10-fold cross-validation experiment, and finds that the silhouette coefficient can serve as an effective metric for use in automated clustering model selection.

3. Data

The data for this chapter are provided by a large, international banking institution, and it contains many attributes pertaining to the organization's credit card customers. The raw data set consists of 149,959 unique records. Relative to clustering, a primary initial challenge entails data preprocessing, including dimensionality reduction, problematic record elimination or transformation, and various issues concerning discretization, feature selection, and normalization.

The curse of dimensionality argues that, as the number of attributes increases beyond a certain point, the ability of data mining algorithms to produce meaningful results diminishes (Alpaydin, 2010). Therefore, it is advisable to include only truly useful features, while minimizing redundancy. In establishing relevant dimensions, individual creditworthiness indicators reported in prior research and other sources serve to offer fundamental guidance (e.g., Consumer Financial Protection Bureau, CFPB, 2012; Credit Score Decoder, CSD, 2013; Khandani, Kim, & Lo, 2010; MSUFCU, 2015; Shuai, Lai, Xu, & Zhou, 2013). An initial list of 10 dimensions is assembled based on this exploration, but further examination via descriptive statistics ultimately necessitates the elimination of certain variables. For example, the value of profitability represents the income a given customer presumably contributes to the institution. However, this dimension exclusively contains null entries within the current data set. Also, VIP_Code is an internal metric used by the bank in assessing the favorability of credit card clients, but 99.8% of the records show a value of "0" for this attribute. Therefore, it cannot provide adequate differentiation among customers. In finalizing dimensionality reduction, four attributes are maintained for clustering purposes. These include AccountAge, CreditLimit, AdditionalAssets, and LatePayments. These attributes are described later in Section 4.

Because the goal is to partition the customer base and perform outlier analysis in accordance with existing data, a conservative approach to record removal is taken. Objects are only deleted if they contain obvious errors or are unable to be appropriately preprocessed for other reasons. For example, LatePayments is computed by dividing actual late payments by account age in months. Therefore, the resulting attribute serves to standardize the measure for all customers, regardless of account age. However, some records reflect new accounts and thus possess an account age of zero. These records are discarded to avoid the problem of dividing by zero in deriving LatePayments. Following record elimination, the final data set includes 149,893 records and four dimensions.

4. Discretization, Feature Selection/Creation, and Normalization

Each attribute is examined to determine the preferred method of preprocessing. In doing so, characteristics such as data type are explicitly considered. Upon conclusion of this exercise, three dimensions are normalized and one variable is created and subsequently normalized.

Duration of credit history is an important variable in the determination of credit scores (CSD, 2013). In the data set, AccountAge refers to the length of time a given customer has maintained a credit card account, and this is a proxy

for credit history. AccountAge is a numeric attribute and is expressed in months. Therefore, normalization is the preferred approach to data preparation. In terms of preprocessing, normalization precludes numeric dimensions with larger inherent values from dominating attributes with innately smaller amounts (Han & Kamber, 2001). For example, if height in inches and weight in pounds are collectively used and provided equal weighting in a clustering operation involving people, results would be driven by the weight dimension. This is because it occupies a wider range, and would typically have a significantly larger value relative to height for each record. Fortunately, proper transformation can resolve this problem. Specifically, Shalabi, Shaaban, and Kassabeh (2006) offer a basis for normalizing values on a [0,1] scale as follows:

$$\textbf{Normalized Value} = \frac{\textbf{Actual Value} - \textbf{Minimum Value}}{\textbf{Maximum Value} - \textbf{Minimum Value}}$$

In this equation, actual value is a specific amount to be transformed, minimum value is the smallest amount in the range of the target dimension, and maximum value is the largest amount for the feature of interest. When the formula is applied to each cell of the AccountAge column, all associated amounts fall within the desired [0,1] scale.

Amount of credit and creditworthiness exhibit a direct relationship (Khandani et al., 2010), and this suggests that credit limit is important as a criterion is deciding creditworthiness. In the data set, the CreditLimit dimension is a numeric attribute that is measured in Brazilian Reais or Reals, and it indicates the maximum amount of credit a given customer has available. Given its numeric nature, it is normalized in the same manner as AccountAge.

In the provided data, AdditionalAssets pertains to the number of bank products a customer has in addition to a credit card. For instance, if a client holds a credit card, mortgage, and auto loan, then the number of additional assets pertaining to this customer is two. This dimension is an indicator of credit variety, which is an important metric of creditworthiness (Morgan, 2011). As with AccountAge and CreditLimit, AdditionalAssets is transformed using the formula from Shalabi et al. (2006).

Payment history is the main factor in determining credit scores, and late payment is the most significant variable within the payment history context (CSD, 2013). In the data set, LatePayments is created from other elements in the raw data to provide for a standardized representation. This is because a positive relationship is likely to exist between account age and number of late payments. To construct the late payments metric effectively, it is initially transformed so that values are made comparable among all records. Specifically, the LatePayments dimension is created by dividing the number of late payments by account age in months, thereby generating a measure of late payments per month for each record. Next, the data are examined to determine whether further transformation is necessary. In doing so, it is found that this variable exists on a range from zero to four. Consequently, to preprocess this attribute fully, normalization is done in a manner similar to the previous three dimensions.

5. Analysis and Results

To evaluate the models, simulation routines are initiated for five distinct algorithms, and the silhouette coefficient is the basis for assessing model quality such that the algorithm producing the largest value is preferred. The results of this process, presented in Table 1, suggest that the complete-link hierarchical method be used for clustering the data of interest.

For complementary insight, the entire simulation routine result set is plotted in Fig. 1.

For models with two through five clusters, the complete-link hierarchical method is strictly superior to the other four algorithms. Furthermore, it achieves peak performance at three clusters. For added clarity, a graph of silhouette coefficients for the complete-link method is considered in Fig. 2.

At three clusters, the maximum silhouette coefficient of 0.5817 is reached. Thus, a complete-link hierarchical three-cluster solution is selected for subsequent analyses. This model is generated and a set of visualizations and tables follow to note distinctive features of each partition. First, representative values for each profile are shown in Fig. 3.

Table 1. Algorithm Performance Rankings.

Algorithm Ranking	
Method	**Silhouette Coefficient**
Complete	0.5817
K-means	0.3936
Ward	0.3768
EM	0.3613
PAM	0.3383

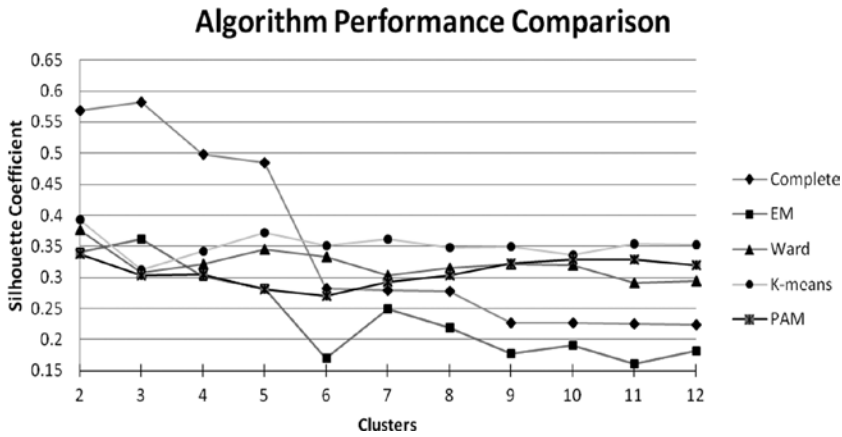


Fig. 1. Algorithm Performance Graph.

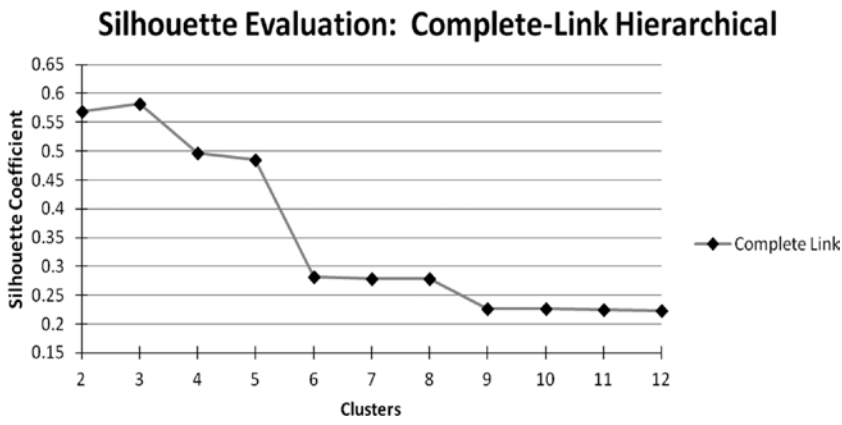


Fig. 2. Complete-link Hierarchical Method Performance Plot.

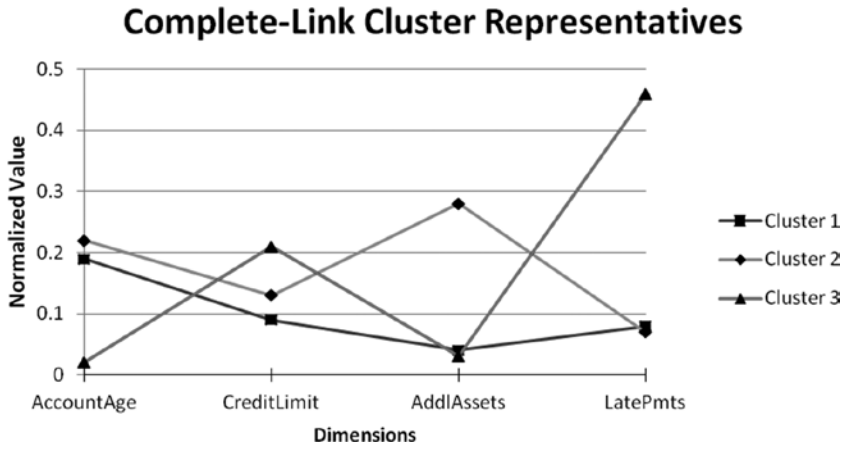


Fig. 3. Cluster Representative Values.

This view suggests that representative values for a given cluster are generally distinguishable from the others. In particular, Cluster 2 is most favorable relative to all dimensions and, thus, qualifies as the most creditworthy group of clients. In addition, Cluster 3 is inferior with respect to account age, late payments, and additional assets. Interestingly, while this partition corresponds to the least creditworthy customers, its average credit limit is actually higher than the other segments. Finally, Cluster 1 represents the intermediate category with representative values typically falling between the extremes.

Second, a three-dimensional principal component scatter plot is generated, in which a color coding scheme is used to distinguish all data points by partition (Fig. 4).

This image depicts all objects in three-dimensional space based on normalized principal component values. The vast majority of points is black in color and corresponds to the majority cluster. In fact, this group is classified as Cluster 1 by the complete-link model, and it contains 141,858 of the 149,893 observations

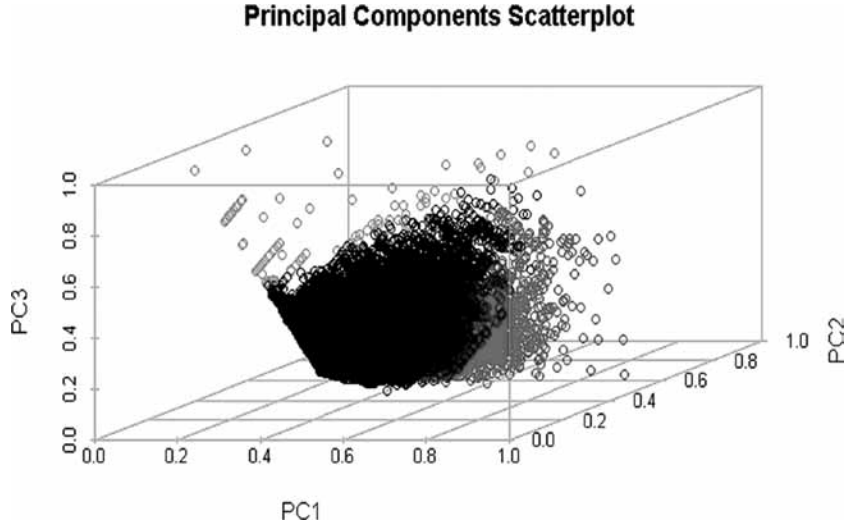


Fig. 4. 3D Plot of Normalized Principal Components.

in the data set (94.6% of total). Conversely, the minority group contains only 147 records (0.10% of total), and these appear as green elements in Fig. 4. Incidentally, these points comprise Cluster 3 in the complete-link model. Given the extremely small membership of this group, all associated records should be considered as potential outliers and should, therefore, be investigated individually. The remaining 7,888 records appear as red objects, and collectively form Cluster 2.

Finally, statistical testing is performed on a representative sample of the data using a non-parametric Kruskal–Wallis test. However, given that there is small membership in one of the three clusters, results might be taken with a degree of caution. Traditionally speaking, the minimally adequate group size for statistical testing purposes is 30 (Singleton & Straits, 1999). However, the authors have also suggested minimum recommended group size might often be much larger than this (e.g., 100 to 200). Results are shown in Table 2, and offer additional information arguing in favor of a three-cluster solution.

More specifically, all variables are shown to be highly significant, with each having a *p*-value below 0.001. This provides incremental evidence that each cluster is distinct from the others and thus describes a unique customer type.

To clarify existing customer group differences further, representative values are de-normalized and the corresponding results are displayed in Table 3.

Each profile clearly describes a unique customer type. For instance, Cluster 2 represents the most mature and creditworthy clients. In particular, mean account age is over 8 years, and of the total number of payments made by customers in this group, only 6.7% are late on average. However, this partition comprises just over 5% of the population.

By contrast, Cluster 3 refers to the least mature and least creditworthy individuals. Specifically, the mean account age is only nine months and nearly half of all payments submitted are late on average. Also, given the noted problems and issues with these customers, the average credit limit seems exorbitant, especially

Table 2. Kruskal–Wallis Test – Complete-Link 3 Cluster Model.

Field	Chi-squared	df	Significance Level
AccountAge	156.89	2	$p < 0.001$
CreditLimit	212.83	2	$p < 0.001$
AdditionalAssets	6,833.20	2	$p < 0.001$
LatePayments	92.70	2	$p < 0.001$

Based on a representative sample of 30,000 records.[1]

Table 3. De-normalized Representative Values for 3-Cluster Model.

Dimension	Cluster		
	1	2	3
AccountAge	83	98	9
CreditLimit	$7,462	$10,451	$16,741
AdditionalAssets	0.27	2.27	0.27
LatePayments (%)	6.95	6.66	42.36
Instances	141,858	7,888	147
Percent of total	94.64	5.26	0.10

when compared with the other profiles. Fortunately, this group contains only 147 clients (0.10% of total). Nevertheless, these customer accounts all deserve closer investigation and scrutiny to ensure that they are being managed and maintained in a fiscally responsible manner.

Cluster 1 falls between the extremes and, thus, corresponds to the intermediate portion of the customer base in terms of creditworthiness. Also, this profile contains the vast majority of credit card accounts. Overall, each customer group possesses a unique set of characteristics, and this suggests that each partition should be approached and managed differently. Note that the analysis in this chapter is substantially automated in the R programing environment. Consequently, significant processing and evaluation efficiencies were achieved, and many of the complexities of clustering were eliminated or at least mitigated. For convenience, the R code pertaining to this chapter is publicly available (Byrnes, 2015).

6. Conclusion

Results like those in Table 3 offer immediate utility to auditors relative to events such as risk assessment procedures and going concern issues. For example,

[1]Because the smallest cluster contains only 147 objects, a representative sample of 30,000 ensures that about 30 records from this group will be captured in the sample (30,000/149,893 × 147 = 29.42). Incidentally, tests on the full population were also performed and all results were found to be highly significant as well.

imagine that a comprehensive set of attributes and standardized procedures is established for describing the customer base. Some of these attributes might be industry-specific, and industry benchmarks could be established in these cases for comparison purposes. Once the pertinent data are formulated, it might be clustered and updated on a periodic basis (e.g., weekly or monthly) so that the auditor can perform trend analysis concerning this information. To facilitate efficiency, the pertinent customer and industry benchmark information could be reflected as numeric scores. Moving forward, the auditor could plot the customer base and benchmark scoring information over some desired time horizon, thus creating the ability to monitor and respond to changes in the organization's client structure. For instance, if the customer base score exhibits a significant decline and/or falls below the associated benchmark value, this would suggest substantially increasing risk. The auditor could incorporate this into the risk assessment results and then modify the associated audit plan and corresponding audit tests.

There is no question that customers are critical to organizational success. Auditors would certainly benefit from using information generated from pertinent customer base data and sharing the resulting findings with their client and associated stakeholders. Furthermore, clustering can theoretically assist in analyzing any relevant data stream as long as it can be represented in numeric terms. For example, clustering could be useful in conducting controls testing activities and detecting and mitigating the incidence of fraud, as well as assisting auditors in satisfying audit requirements relative to consideration of fraud in the financial statement audit context. It is hoped that this chapter will encourage auditors to move more actively toward adopting data mining technologies in their quests to provide information useful for stakeholders' decision making in the evolving real-time global economy.

References

Alpaydin, E. (2010). *Introduction to machine learning* (2nd ed.). Cambridge, MA: Massachusetts Institute of Technology.

Byrnes, P. E. (2015). *Developing automated applications for clustering and outlier detection: Data mining implications for auditing practice*. Doctoral dissertation, Rutgers University Graduate School, Newark, NJ.

Consumer Financial Protection Bureau (CFPB). (2012). Key dimensions and processes in the U.S. credit reporting system. Retrieved from http://files.consumerfinance.gov/f/201212_cfpb_credit-reporting-white-paper.pdf

Credit Score Decoder (CSD). (2013). Retrieved from http://creditscoredecoder.com/credit-score-calculated/

Deng, Q., & Mei, G. (2009). Combining self-organizing map and k-means clustering for detecting fraudulent financial statements. *IEEE international conference on granular computing,* August 17–19, 2009, Lushan Mountain, Nanchang, China (pp. 126–131).

Garla, S., Chakraborty, G., & Gaeth, G. (2012). *Comparison of* K-*means, normal mixtures, and probabilistic* D-*clustering for B2B segmentation using customers' perceptions*. Orlando, FL: SAS Global.

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. San Francisco, CA: Morgan-Kaufmann Publishers.

Hao, M., Dayal, U., Sharma, R., Keim, A., & Janetzko, H. (2010). *Visual analytics of large multi-dimensional data using variable binned scatter plots*. San Jose, CA: IS&T/SPIE Electronic Imaging conference proceedings.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit risk models via machine learning algorithms. Draft of May 9, 2010.

Liu, R., Qian, X., Mao, S., & Zhu, S. (2011). Research on anti-money laundering based on core decision tree algorithm. In *Control and decision conference* (CCDC), May 23–25, 2011, Mianyang, China (pp. 4322–4325).

Michigan State University Federal Credit Union (MSUFCU). (2015). What is a Credit Score?

Morgan, K. (2011). How to improve your credit score. Retrieved from http://voices.yahoo.com/how-improve-credit-score-8656533.html

Shalabi, L. A., Shaaban, Z., & Kassabeh, B. (2006). Data mining: A preprocessing engine. *Journal of Computer Science*, *2*(9), 735–739.

Shuai, L., Lai, H., Xu, C., & Zhou, Z. (2013). The discrimination method and empirical research of individual credit risk based on bilateral clustering. *Modern Economy*, *4*(7), 461–465.

Singleton, R. A., & Straits, B. C. (1999). *Approaches to social research* (3rd ed.). New York, NY: Oxford University Press.

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston, MA: Pearson Education.

Chapter 4

A Multi-faceted Outlier Detection Scheme for Use in Clustering*

Paul Byrnes

1. Introduction

Outlier detection is an important supplemental component for clustering. In related literature, outliers are essentially defined as irregularities or anomalies. For example, Hawkins (1980) perceives an outlier as "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism." Barnett and Lewis (1994) view the outlier as "an observation which appears to be inconsistent with the remainder of that set of data." Outlier detection is a procedure for capturing objects that are notably different from the others (Zimek, Campello, & Sander, 2013). In this chapter, a multiple-measure outlier detection method is proposed, formulated, and implemented.

2. Preliminary Issues in Outlier Detection

Initially, two outlier detection issues must be contemplated. First, one assertion in clustering-based anomaly detection is that normal objects within a given group or partition reside closer to the cluster representative (e.g., centroid), whereas exceptions exist farther from this value. The centroid is simply the average of all points in a cluster, and the arithmetic average of a partition could presumably be seen as a suitable reference point for determining whether a participating object is anomalous. However, the mean is a reliable measure of central tendency only when a normal distribution is at least approximated, although this criterion might often not be satisfied. When addressing non-normal distributions, the median is a preferred measure of central tendency. Furthermore, when the normality assumption holds, the median and mean are equally reliable metrics. Because applicability to

*This chapter is based on the fourth chapter of the author's dissertation (Byrnes, 2015).

Audit Analytics in the Financial Industry, 47–58

doi:10.1108/978-1-78743-085-320191004

a wider variety of distribution types is desirable, the median clearly serves as a more reliable measure of central tendency in the clustering context.

Second, while a multitude of measures are available for performing outlier detection the specific subset to consider initially is a function of the context in which outlier detection is to be conducted. Classification, nearest neighbor, clustering, and statistical based anomaly detection methods have all been discussed and explored in prior research (Chandola, Banerjee, & Kumar, 2009), and each approach is fundamentally different. In this chapter, outlier detection will obviously be restricted to clustering-based methods. In this environment, proximity measures are typically promoted, and these consist of various distance and similarity measures (Tan, Steinbach, & Kumar, 2006). Some popular distance measures are Minkowski, Manhattan, Euclidean, and Mahalanobis, while similarity measures include Simple Matching, Jaccard and Tanimoto Coefficients, and Cosine Similarity (Tan et al., 2006). These distance and similarity measures warrant further discussion.

3. Distance Measures for Outlier Detection

In considering distance measures, three are comparable in terms of equation structure. For example, the general formula for Minkowski distance is

$$\text{Minkowski distance} = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

where n is the number of observations, p the parameter, x_i the ith observation of x, and y_i the ith observation of y.

In this formula, p is a parameter that may be set as any integer value above zero. For example, when $p = 1$, Minkowski distance is simply Manhattan distance, and when $p = 2$, it becomes Euclidean distance. This parameter issue actually highlights a potential problem with Minkowski distance because prior intuition might often not exist concerning the appropriate setting for p in a given instance of outlier detection. This substantially limits its implementation utility in terms of ease of use.

Incidentally, Manhattan distance (i.e., city block distance) is the most simplistic of these distance measures. Furthermore, it is suboptimal in performing outlier detection on data sets with more than two dimensions or attributes. Given that this condition will frequently occur in practice, Manhattan distance is not given further consideration here.

Chandola et al. (2009) find that Euclidean distance is often used for outlier detection routines. Also, when attempting to locate outliers in n-dimensional space, Euclidean distance will be considerably more useful than Manhattan. The associated equation for Euclidean distance follows:

$$\begin{aligned} \text{Euclidean distance} &= d(p,q) = d(q,p) \\ &= \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \end{aligned}$$

where d is the distance between two vectors p and q, p a given customer record (vector), and q the median vector of a cluster (benchmark vector).

Mahalanobis distance is distinct from the Manhattan or Euclidean measures in that it incorporates the data's covariance matrix in calculating distances. In addition, it has demonstrated success in multi-variate outlier detection schemes (Starkweather, 2013). The associated formula follows:

$$\text{Mahalanobis distance } = \sqrt{(x - \mu)^T \, \mathrm{S}^{-1} \, (x - \mu)}$$

where x is an observation or object, μ the mean value (median in this chapter), and S the covariance matrix.

At this juncture, two distance measures, Euclidean and Mahalanobis, are legitimate candidates for performing outlier detection. Moving forward, other proximity measures are considered.

4. Similarity Measures for Outlier Detection

Similarity measures assess the degree of comparability between two objects. In this setting, Tan et al. (2006) discuss the Simple Matching, Jaccard, Cosine Similarity, and Tanimoto Coefficients. Two of these measures, Simple Matching and Jaccard Coefficients, are found to have limited value in anomaly detection because they both require binary data in deciding the extent to which two records are similar. Fortunately, Cosine Similarity and the Tanimoto Coefficient are not restricted to analyzing binary data. Consequently, these measures are explored further.

Cosine Similarity computes the cosine of the angle between two arrays x and y. In this chapter, x is a record (object) and y is the benchmark (median) vector. In calculating Cosine Similarity, a value of one indicates the angle between two vectors is zero degrees, suggesting the two objects are identical. Conversely, a value of zero means that the angle between two vectors is 90 degrees, so the two objects are completely dissimilar (Tan et al., 2006). The more dissimilar an object or record is relative to the benchmark vector, the more likely it is to be an outlier. The Cosine Similarity equation follows:

$$\text{Cosine Similarity } = \cos(x, y) \; = \frac{x \cdot y}{\| x \| \; \| y \|}$$

where x is a vector x (record or object), y a vector y (record or object), and y is the benchmark (median) vector.

The Tanimoto Coefficient is also referred to as the Extended Jaccard Coefficient. As with Cosine Similarity: (1) it may be employed with non-binary data (Tan et al., 2006); (2) a value of one indicates complete similarity, while zero suggests complete dissimilarity; and (3) the more dissimilar an object is relative to the benchmark vector, the more likely it is to be anomalous. The Tanimoto Coefficient formula is

$$\text{Tanimoto Coefficient } = \frac{x \cdot y}{\| x \|^2 + \| y \|^2 - x \cdot y}$$

where x is a vector x (record or object), y is a vector y (record or object), and y is the benchmark (median) vector.

At this point, four metrics are suitable for use in a clustering-based outlier detection method: (1) Euclidean distance; (2) Mahalanobis distance; (3) Cosine Similarity; and (4) Tanimoto Coefficient. Next, three additional factors merit attention.

5. Outlier Detection Method – Final Considerations

The first factor for consideration is that extant research demonstrates that two measures of a particular type (e.g., distance) will typically be more positively correlated than two metrics from different categories (Zimek et al., 2013). Additionally, distance measures are more reliable for high-density data, whereas similarity measures are more applicable to low density data (Tan et al., 2006). Consequently, ensembles have been proposed in the literature, whereby multiple measures are implemented in conjunction with a voting scheme initiated for detecting anomalies (Zimek et al., 2013). Given these findings, the Euclidean distance, Mahalanobis distance, Cosine Similarity and Tanimoto Coefficient delineated in the previous section are incorporated with an adjustable weighting system to facilitate outlier detection in data of varying levels of density. For example, if attempting to identify anomalies in a high density cluster, the weightings for Euclidean and Mahalanobis distances could be increased, while the weightings for Cosine Similarity and Tanimoto Coefficient could be decreased. In this way, adjustments can be readily made to enhance the effectiveness of outlier detection in a wide variety of circumstances.

However, in constructing such a system, an issue must first be addressed. Specifically, distance measures operate such that an object with the largest distance value exhibits the highest propensity for being anomalous. Conversely, similarity metrics support an opposite conclusion. To remedy this problem, similarity measures are transformed to dissimilarity measures by subtracting the associated formulas from one (Tan et al., 2006).

The second consideration is that a threshold value will provide a rough cut-off point relative to outlier status. In this regard, Chandola et al. (2009) discuss a procedure based on the box plot (Tukey, 1977). This procedure, referred to as the box plot rule, argues that an object more than 150% of the inter-quartile range below the 25th percentile or above the 75th percentile is anomalous. Acuna and Rodriguez (2011) elaborate on the box plot rule by differentiating between mild and severe outliers. They indicate that mild exceptions are those lying beyond 1.5 times but no more than 3 times the inter-quartile range below the 25th percentile or above the 75th percentile. An object falling more than 3 times the inter-quartile range below the 25th percentile or above the 75th percentile is a severe anomaly.

The third issue is that a mechanism is needed to rank identified exceptions in an aggregated manner so that the most egregious outliers will be investigated first. Issa (2013) finds that a large number of anomalies ("exceptional exceptions") are typically generated in practice such that examination of all outliers is simply

impractical. Furthermore, exceptions are often false positives that unnecessarily consume investigatory resources. In mitigating this problem in an audit context, a methodology based on a predictive ordered logistic regression model is developed and incorporated to review and prioritize control risk assessments so that investigation emphasis can be placed on the most problematic issues (Issa & Kogan, 2014). In another study, a suspicion score system is developed and implemented for prioritizing records relative to internal control rule violations in transactional data, thereby again confronting the issue of "exceptional exceptions" (Issa, Kogan, & Brown-Liburd, 2015). These procedures facilitate the productive use of scarce resources in examining problematic records. With this insight, and given that four separate measures are being adopted in this chapter, a scoring procedure is formulated for aggregating and ranking objects in each evaluated cluster.

Specifically, since the proximity measures do not all have identical scales, each measure is normalized on a (0,1] range. This both standardizes the metrics and ensures that a normalized measure for a given object or data point will not have a value of zero unless it is exactly equal to the median vector. The formula used to normalize distance and dissimilarity measures, discussed by Scandizzo (2005), is

$$\text{Normalized Value} = \frac{\text{Actual Value}}{\text{Maximum Value}}$$

Following normalization, an outlier score for a given object is computed as the sum of the four normalized measurements for that data point. With this approach, an outlier score may fall between zero and four, inclusive. Essentially, the object within a data set possessing the highest outlier score is most likely to be anomalous. The anomaly detection routine developed in this chapter is written in the R programing language, and is created such that individual weights can be readily manipulated for each measure in cases where differential weighting is warranted (see Appendix C in Byrnes, 2015).

6. Analysis and Results

In the previous chapter, "Automated Clustering: From Concept to Reality", a data set of credit card customers was evaluated. In this chapter, one of the resulting clusters (Cluster 3, which represents the credit card customers identified as the least creditworthy) is explored to demonstrate the implemented outlier detection method. Incidentally, the executed R program results in an output file that can be further analyzed in other software environments, such as Tableau, Clickview, and Excel.

The following set of images is produced using Tableau. It depicts all objects in Cluster 3 in all pair-wise combinations of the employed proximity measures.

Note that for a given proximity measure, the median vector value exists at the origin (lower-left of each plot). Objects close to this area are perceived to have an extremely low probability of being anomalous. As an object's distance or dissimilarity value increases, it resides further from the origin. In terms of visualization,

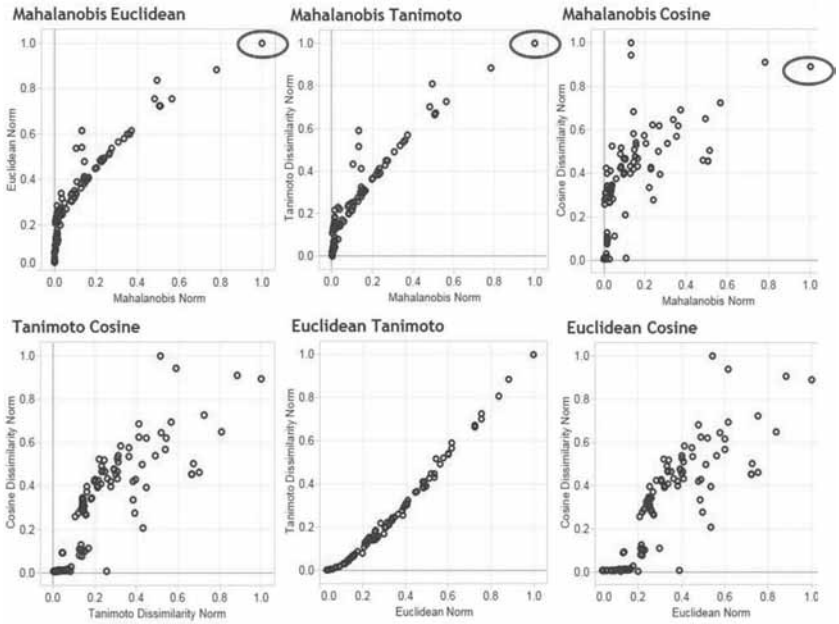


Fig. 1. Object Plots – Cluster 3.

data points appearing in the upper-right section of a given plot are viewed as having extremely high outlier potential in terms of both plotted measures. Also, objects in either the upper-left or lower-right sectors would exhibit a very high outlier potential in terms of one of the two graphed metrics, while simultaneously demonstrating a very low outlier potential in terms of the other measurement.

In examining the graphs, one object clearly emerges as the most prominent. Specifically, record 77090 is circled in red in the upper set of views, and is identified as the most anomalous according to three of the four proximity measures. Furthermore, it is assessed as the fourth most dissimilar record according to Cosine Similarity. As the initial set of images fails to offer detailed information, Fig. 2 shows an outlier dashboard that is constructed incorporating the outlier scoring system so as to allow for specific insights.

The top-left section of the dashboard shows all objects having outlier scores above two, and the associated records are shown in descending order, both graphically and numerically. From this view, two of the objects (i.e., 77090 and 157590) have extremely high outlier scores. Intuitively, these might be perceived as having "exceptional exceptions" status. The lower-left sector consists of a heat map showing the top 10 data points in terms of outlier scores. Furthermore, this heat map is structured in a manner such that the record appearing in the upper left region is considered most anomalous while the record appearing directly beneath it is viewed as second most interesting. By scanning the heat map from left to right in this manner, the record in the lower right area is ultimately identified as the 10th most anomalous. The right portion of the dashboard contains a box plot of all records in Cluster 3.

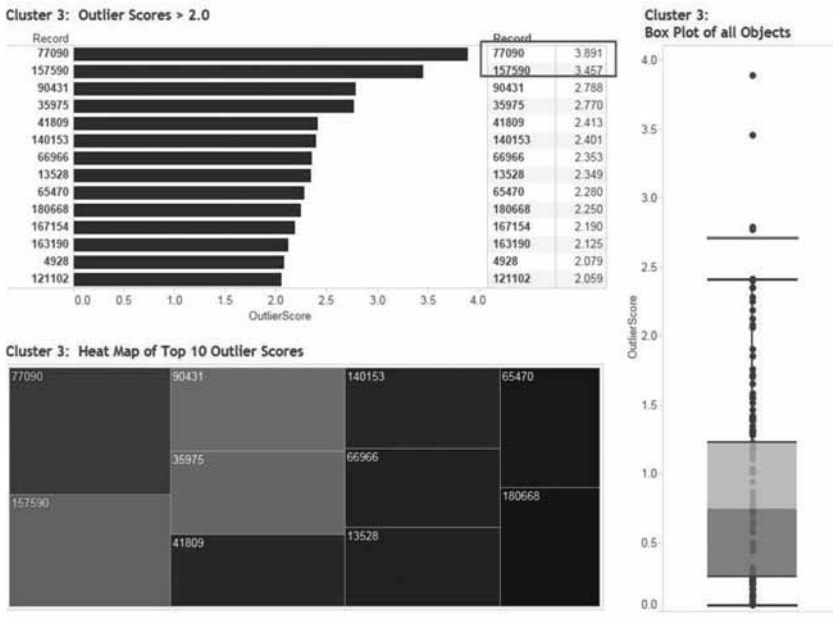


Fig. 2. Dashboard – Cluster 3.

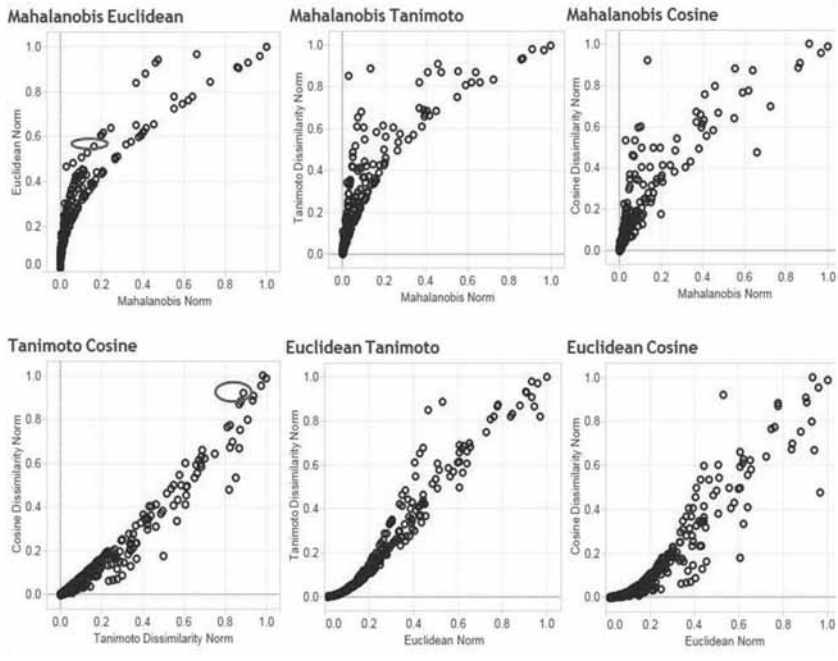


Fig. 3. Plots – Ratio Analysis.

The red horizontal line in the view is the cutoff threshold existing at 1.50 times the inter-quartile range. According to the box plot rule described earlier, any data point existing above this boundary is considered an outlier. In the current image, four objects (i.e., 35975, 90431, 157590, and 77090) satisfy this condition. While these items certainly warrant additional investigation, it may also be advisable to consider examining other data points with relatively high outlier scores.

Overall, this outlier detection method demonstrates the ability not only to identify potential exceptions, but also to prioritize them in a manner that facilitates productive use of valuable, scarce investigatory resources. Although the scheme is applied here in an effort to identify credit card customers who do not fit established profiles, it could theoretically be implemented in any clustering-based outlier detection activity for which the associated data can be represented numerically. For example, auditors could employ this method to locate anomalous transactions for the purpose of fraud discovery. In addition, the approach could be used during an audit for other activities including but not limited to audit planning, risk assessment, and analytical procedures. One example of such an application is discussed in the next section.

7. Outlier Detection – Auditing Context Example

Ratio analysis is a common analytical procedure in auditing and is often used as part of risk assessment, substantive testing, and final analytical procedures (Messier, Glover, & Prawitt, 2016). In a general sense, it entails evaluating four different types of measures, including liquidity, leverage, profitability, and activity ratios (Whittington & Pany, 2008). Historically, a univariate approach might be adopted, in which each ratio is considered and compared against the industry benchmark and/or historical values. While such a piecemeal technique has value, it fails to identify relationships exhibited via the combination or synthesis of examined ratios. For instance, in anomaly detection, objects identified as outliers in univariate space are often not found to be multi-variate outliers (Starkweather, 2013). This suggests that a multi-variate scheme should be considered, and could consist of treating the set of company ratios as an array to be compared with the industry benchmark vector of identical ratios (e.g., median).

To achieve this goal, various types of measures might be considered as previously discussed. For instance, two popular distance-based metrics include Mahalanobis and Euclidean distance. Furthermore, two common similarity measures are Cosine Similarity and the Tanimoto Coefficient (Tan et al., 2006). In deciding which measure(s) to incorporate, two issues must be confronted. First, no metric is strictly superior, and ensembles of multiple approaches have been shown to be particularly effective (Zimek et al., 2013). In fact, each measure possesses a set of strengths and weaknesses, and these should be contemplated before measure selection. Second, two metrics in a particular category tend to be more highly correlated than two measures of differing types (Zimek et al., 2013). Given these observations, an approach that employs a combination of measures should be explicitly considered.

In the following analysis example, the four measures discussed in Sections 3 and 4 are used individually and in the aggregate to perform multi-variate ratio

analysis of firms in the retail industry. The compiled data from Compustat consists of entities with Standard Industrial Classification codes between 5000 and 5999, and pertains to business operations during the 2013 fiscal year. Evaluated ratios include (1) current ratio; (2) long-term debt-to-equity ratio; (3) return on assets; and (4) inventory turnover. In preparing the data for analysis, ratios are scaled so that no single item will dominate outcomes. For example, consider the current ratio and inventory turnover. A typical current ratio might be about 2, whereas the usual inventory turnover value would be significantly higher (e.g., 7). In addition, turnover would tend to occupy a larger range relative to the current ratio. If these ratios are not transformed, inventory turnover would likely dominate the evaluation and results would thus be biased.

In computing distance and similarity measurements, the median firm is used as the benchmark or comparison array. For example, in calculating a given measurement for a particular company, the organization's ratio vector and the median firm's ratio vector are the computational inputs. In addition, after calculating similarities, results are converted to dissimilarities. In this way, a larger value for any given measure always indicates a greater absolute distance or difference from the median firm. Finally, for each entity, the four measurements are normalized on a (0,1] scale, and summed to arrive at a final outlier score, as previously described. In the multi-variate ratio analysis task that follows, company code 1106838 assumes the role of audit client.

To maximize efficiency, data processing is fully automated in R, and the resulting output file is used for subsequent visualization in Tableau. A pair-wise complete set of normalized distance/dissimilarity plots for all objects follows in the next set of images.

In each image, the median firm exists at the origin. Therefore, objects farther from this location are more distant and/or different from the industry benchmark. For example, the auditee is highlighted in the lower-left graph, and is identified as substantially different from the median firm in terms of both Tanimoto and Cosine dissimilarities. This client is again denoted in the upper-left plot. In this case, the object is not as far from the origin when only considering distance measures, and certainly not perceived as anomalous when using Mahalonobis distance as the exclusive criterion. While this viewpoint offers preliminary value, an aggregated representation of all measures yields a more specific insight.

In formulating this view, each metric is given equal weighting. For each object, the four measurements are summed during the data processing step to yield a final outlier score. Since each measure is normalized on a (0,1] scale, the maximum score for an object is four. An outlier score dashboard follows in Fig. 4.

Except for the box plot, image views are restricted to organizations with outlier scores of at least 2.0 to reduce clutter. A heat map is shown in the upper-left section of the dashboard, and the size of each rectangle corresponds to the relative magnitude of an outlier score. For instance, Company 30697 has the shape with the largest area and most intense color scheme. Consequently, it is viewed as having the highest probability of being anomalous. A comparable observation can be made from examining the lower-left image. Once again, Company 30697 is depicted as the most irregular object, although this finding is not as obvious.

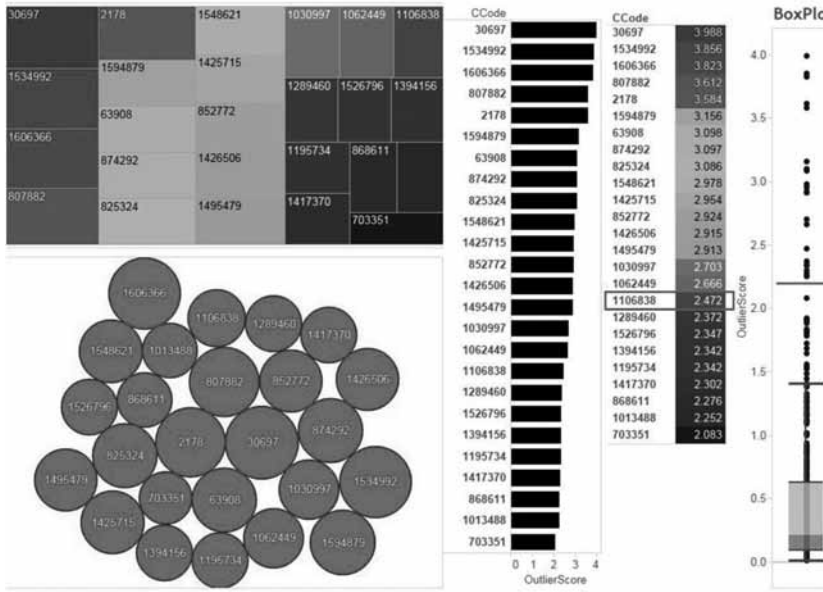


Fig. 4. Dashboard–Ratio Analysis.

In this analysis, the audit client maintains the 17th highest outlier score (i.e., 2.472), suggesting that, while several other entities are more likely to be problematic, the auditee is nevertheless worthy of further investigation. For example, the box plot image shows objects with outlier scores above approximately 1.4 fall beyond the normal range of the data distribution. Specifically, about 98% of points are expected to lie between the upper and lower horizontal lines of the box plot. A red horizontal line is positioned at three times the inter-quartile range above the 75th percentile, and points above that line are treated as severe outliers according to the box plot rule. This stipulates that records with scores above approximately 2.2 qualify as potentially serious outliers. Incidentally, the auditee satisfies this condition.

To provide contrast, ratios for the median firm, the entity with the lowest outlier score, the auditee, and the object with the highest outlier score are all presented in Table 1

In the data set, overall median and mean outlier scores are only 0.214 and 0.526, respectively. Although there are 16 entities with outlier scores greater than 2.472, the auditee is still substantially different from the typical firm in this industry. Furthermore, the client occupies the lower end of the distribution, suggesting that the organization has an unfavorable relative standing. In particular, its current ratio, long-term debt-to-equity ratio, and ROA values are all well below the associated industry benchmarks. Of particular concern, ROA is highly negative, demonstrating that the company suffered a substantial net loss in 2013. If the auditee is mature and well established, this is a significant problem. Conversely, the client's inventory turnover ratio is nearly four times larger than that of the

Table 1. Ratio Comparisons.

Company	Current Ratio	LT Debt-to-Equity	ROA	Inventory Turnover	Outlier Score
Median	1.66	0.29	0.04	7	0.000
791519	1.56	0.16	0.05	7	0.016
1106838	0.31	0.07	**−0.38***	26	2.472
30697	2.64	0.73	0.01	152	3.988

*Note: ROA for company 1106838 is negative because a net loss was reported for the period.

median firm. In isolation, this might be viewed as a positive signal. It could also at least partially explain the abnormally low current ratio. However, when taken in a multi-variate context, the company's performance is clearly substandard and suggestive of substantial risk. This would have a significant impact on risk assessment outcomes and subsequent formulation or revision of the audit plan and associated audit procedures. By contrast, the object having the lowest outlier score is quite comparable to the median firm, indicating that it is generally operating in alignment with what would be expected of firms in this industry during 2013.

8. Conclusion

The clustering-based outlier detection method proposed, developed, and implemented in this section should be applicable to a variety of accounting and auditing activities. A primary pre-condition for its usage is that the associated data can be represented meaningfully in numeric terms. In addition, determining the appropriate weightings to use for the individual proximity measures is an issue that both requires explicit consideration and warrants future research. Nevertheless, it is hoped that the information provided in this chapter may assist in moving the accounting profession toward the adoption of technology and automation in conducting future accounting and auditing tasks more productively.

References

Acuna, E., & Rodriguez, C. (2011). A meta analysis study of outlier detection methods in classification. Retrieved from https://www.researchgate.net/profile/Edgar_Acuna/publication/228728761_A_meta_analysis_study_of_outlier_detection_methods_in_classification/links/00b7d525e85fae9659000000.pdf

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. Hoboken, NJ: John Wiley.

Byrnes, P. E. (2015). *Developing automated applications for clustering and outlier detection: Data mining implications for auditing practice*. Doctoral dissertation, Rutgers University Graduate School, Newark, NJ.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 15–58.

Hawkins, D. (1980). *Identification of outliers*. London: Chapman and Hall.
Issa, H. (2013). *Exceptional exceptions*. Doctoral dissertation, Rutgers University Graduate School, Newark, NJ.
Issa, H., & Kogan, A. (2014). A predictive ordered logistic regression model as a tool for quality review of control risk assessments. *Journal of Information Systems*, *28*(2), 209–229.
Issa, H., Kogan, A., & Brown-Liburd, H. (2015). *Identifying and prioritizing irregularities using a rule-based model with a weighting system derived from experts' knowledge*. Unpublished working paper, Rutgers University, Newark, New Jersey.
Messier, W. F., Glover, S. M., & Prawitt, D. F. (2016). *Auditing and assurance services: A systematic approach*. New York, NY: McGraw-Hill Education.
Scandizzo, S. (2005). Risk mapping and key risk indicators in operational risk management. *Economic Notes*, *34*(2), 231–256.
Starkweather, J. (2013). Multivariate outlier detection with Mahalanobis distance. Retrieved from http://www.unt.edu/rss/class/Jon/Benchmarks/Moutlier_JDS_July2013.pdf
Tan, P., Steinbach, M., & Kumar, V. (2006). Introduction to data mining. Boston, MA: Pearson Education.
Tukey, J. (1977). *Exploratory data analysis*. Boston, MA: Addison-Wesley.
Whittington, O. R., & Pany, K. (2008). *Principles of auditing and other assurance services* (16th ed.). New York, NY: McGraw Hill/Irwin.
Zimek, A., Campello, R., & Sander, J. (2013). Ensembles for unsupervised outlier detection: Challenges and research questions. *SIGKDD Explorations Newsletter*, *15*(1), 11–22.

Chapter 5

Are Customers Offered Appropriate Discounts? An Exploratory Study of Using Clustering Techniques in Internal Auditing

Jun Dai, Paul Byrnes and Miklos Vasarhelyi

1. Introduction

Globalization has given rise to fierce competition in the marketplace. Consequently, possessing an ongoing competitive advantage is essential for most firms that seek to maintain or expand their position and optimize their profitability and profit growth. Clearly, many forms of competitive advantage exist, but one critical advantage pertains to the customer base. Therefore, an organization should be knowledgeable about its clients so that it can tailor its business activities to optimize the utility that it derives from its customers and the satisfaction that it provides to them. In particular, a firm's desirable customers are an invaluable asset. Therefore, firms should make specific efforts to retain these customers. On the other hand, an organization's less favorable customers are less likely to contribute to the firm's profitability and may even generate significant losses and other negative events, so these clients should be targeted with a different scheme. Initially, the firm may attempt to convert such clients to a more favorable profitability level. Alternatively, methods for controlling the losses brought about by persistently unprofitable customers would be in order. Therefore, a crucial mission of internal auditors should entail examining whether the current business process efficiently and consistently retains desirable customers while controlling losses incurred from less favorable customers. In doing so, a firm can improve its market position within its industry.

To accomplish this mission, segmenting the customer base is essential. Cluster analysis is a data analysis technique that can be appropriately employed for customer segmentation. This technique creates groups of target objects based on information found in the data that distinguishes the objects and the relationships among them (Tan et al., 2006). Instances within a given cluster are similar to one another and distinct from instances in other clusters. By clustering customers into groups, more favorable and less favorable customers can be distinguished.

Audit Analytics in the Financial Industry, 59–69

doi:10.1108/978-1-78743-085-320191005

After the customer base is partitioned, auditors can examine the efficiency and effectiveness of the company's current business activities in terms of providing appropriate services to different types of customers at appropriate prices.

This chapter investigates an internal auditing issue associated with the protocol for offering discounts to various credit card customers at a major bank in South America. The main objective is to examine whether discounts have been offered appropriately to different types of customers. To accomplish this goal, clustering techniques are used to group customers. After dimensionality reduction, data scaling, and normalization, clusters are then created based on pertinent customer attributes. A comparison of the performance of different clustering techniques recommends the Simple K-means algorithm to create clusters. Following the Simple K-means algorithm, seven clusters of credit card customers are identified. These groupings provide excellent segregation in terms of profitability to the company. Interestingly, descriptive statistics show that customers have not consistently been offered appropriately tailored discounts. Therefore, a more targeted and reasonable discount policy should be generated to differentiate the discounts offered to different types of customers.

Section 2 of this chapter provides a brief overview of related work. Section 3 defines the problem and describes the associated business scenarios. Details of the development of the cluster model are introduced in Section 4. Section 5 describes the experiments, evaluation metrics, parameterization, modeling, and results. Section 6 offers conclusions and discusses suggestions for future work on this topic.

2. Related Work

Clustering has been widely used in marketing research, especially in the areas of market segmentation, market structure analysis, and the study of consumer behavior (Thiprungsri and Vasarhelyi, 2011). Punj and Stewart (1983) conduct a comprehensive review of work prior to 1983, which essentially applies clustering techniques to marketing problems. Hosseini, Maleki, and Gholamian (2010) integrate the K-means algorithm into the current RFM (recency, frequency, and monetary) model and then classify customer product loyalty in the B2B (business-to-business) mode. Zhang, Huang, Tang, and Luo (2011) propose a clustering model for customer segmentation based on customer behaviors. Tuma, Decker, and Scholz (2011) review a number of journal articles published since 2000 in which cluster analysis is used empirically in a marketing research setting. The authors also discuss critical issues when applying cluster analysis to segment markets, and make suggestions for best practices and potential improvements.

Even though cluster analysis has become a common tool for marketing researchers, only a few studies have employed clustering methods to solve auditing problems. Thiprungsri and Vasarhelyi (2011) examine the possibility of using clustering technology to automate fraud filtering during an audit. Specifically, the authors use cluster analysis to help auditors evaluate group life insurance claims. Claims with similar characteristics are grouped together and clusters with small populations are flagged as potential fraud. Similarly, this chapter conducts an exploratory study using clustering techniques to deal with an internal audit issue.

3. Audit Problem

This study concerns the credit card division of a large banking institution based in South America. Currently, the organization lacks a robust policy for the way it responds to the fee-related inquiries of credit card customers. Specifically, when customers call to ask for fee reductions, representatives essentially comply with these requests. In addition, the historical range for granted discounts is all-inclusive, falling between 0% and 100%. Given inherent problems with the existing system, this study examines the utility of the current process for offering discounts. Other things being equal, a functional discount policy should provide customers who are more favorable to the bank with higher discounts compared to its less favorable customers. If the current process for offering discounts is ineffective, a new strategy for handling credit card customers' fees should be developed. To determine whether this is the case, relevant customer data review, preprocessing, and analysis are necessary.

4. Method

First, pertinent customer data is evaluated to identify the attributes and records that may be useful for clustering. Next, the data is preprocessed so that it can be suitably analyzed via clustering. This is achieved by determining data types and then performing categorization and normalization routines. Following this, the preprocessed data elements are actually clustered. In the clustering phase, K-means is primarily used and initial seed values and cluster sizes are considered in an effort to arrive at the best overall configuration.

4.1. Data Set Analysis

The bank provides historical data pertaining to its credit card clients. This original data set contains about 240 attributes and roughly 194,523 records. The first step is to select a record and its associated discount offering ID, which is a record identification attribute that allows the record to be attached to a clustered data point in the data set. Then, through preliminary discussions and reflection, eight attributes that appear potentially relevant for the clustering task are identified: (1) customer age; (2) behavior score; (3) VIP code; (4) credit limit; (5) late payments; (6) revenue value; (7) account age; and (8) quantity of additional assets.

Descriptive statistics are computed in an effort to validate the selected attributes to be clustered. In this process, two of these attributes are identified as not useful. The first is VIP code, which initially seemed relevant for generating customer groupings because it captures information about the level of customer desirability. However, statistics indicate that all but 34 records are assigned a VIP code of "0". Therefore, this dimension fails to provide adequate information for customer segmentation, so it will not facilitate the identification of customer types. The second deleted attribute is revenue value. This variable was assigned a "0" for all records, so it is clearly irrelevant for creating customer types.

Ultimately, six attributes are selected for clustering: (1) behavior score; (2) account age; (3) credit limit; (4) customer age; (5) quantity of additional assets; and (6) late payments.

Each of these attributes is anticipated to offer relevant data for establishing customer types. The first attribute, behavior score, is an internal metric that indicates customers' risks based on their purchase and payment behaviors. The range for the behavior scores falls between 1 and 12, where a higher value indicates better behavior patterns. Account age, the second attribute, denotes how long a customer has possessed the credit card. It is a loose indicator of both customer loyalty and desirability. Therefore, a customer with longer account age is generally viewed more positively. The third attribute, credit limit, is an indicator of several features including financial responsibility, disposable income, credit rating, and payment history. Thus, it would be a pertinent metric for clustering customer types. The fourth attribute, customer age, is a potential metric that is likely to maintain a positive relationship with favorability in terms of managing credit. However, there may be a ceiling beyond which customer age is expected to be negatively related to credit-worthiness. For example, an elderly individual might have an exceptionally high credit ranking and resulting customer favorability. However, a stock market decline or significant health problems may make this elderly customer more risky than younger customers are. The fifth attribute, quantity of additional assets, indicates how many other bank products a customer possesses besides the credit card. In general, the more additional assets a customer owns, the more valuable the customer should be. This suggests a positive relationship between quantity of additional assets and customer desirability.

The last attribute, late payments, refers to the number of late payments related to a credit card. Since it is an indicator of a customer's profitability to the bank, it is an excellent candidate for determining customer groupings. However, the primary challenge in using this attribute involves deciding on the most appropriate method for representation. It is unreasonable simply to use the number of late payments as a measure, because it has a positive correlation with account age. Customers who keep credit card accounts active for longer periods would tend to have more late payments. To avoid this kind of data distortion, this attribute is standardized by calculating the number of late payments per month. In this manner, the resulting values are comparable across all clients, regardless of account age.

4.2. Data Set Preprocessing

The first step in preprocessing involves logically categorizing attributes. Some of the dimensions chosen for clustering are naturally quantitative. However, they may provide information that is more meaningful for discovering customer types when they are treated as qualitative attributes. One such attribute is the behavior score. According to the bank's policy, each behavior score represents a certain risk level such that a higher value represents behaviors that are more favorable. Therefore, the behavior score should be treated as an ordinal attribute rather than a quantitative attribute. Based on this consideration, the values of the behavior scores from 12 to 1 are converted into types from A to L, respectively.

Type "A" indicates the most desirable behavior, whereas "L" indicates the least desirable behavior.

The other candidate for conversion from a quantitative to an ordinal variable is the customer age attribute. Some age-related policies that exist in practice are reviewed as guidance when establishing categories for this attribute. In particular, insurance companies tend to reduce the premiums of acceptable drivers after they attain the age of 25 in the United States. This could be perceived as an early maturity milestone, and it might suggest that such individuals are incrementally more responsible at roughly that point. In addition to this idea, the methodology in prior studies (Beirlant et al., 1992; Hirschman, 1979) is used to establish categories for the customer age attribute. Furthermore, the threshold age of 65 is when Brazilian citizens are able to obtain old-age pensions. This is also incorporated into the categorization scheme. Based on these considerations, five customer groupings are established. Type A includes customers 46–65 years old and has the lowest risk. Type B includes customers 36–45 years old and is the second least risky group. Type C includes customers over 65 years old and represents medium risk. Type D includes customers 25–35 years old and has the second highest risk. Type E includes customers 0–24 years old and is the most risky group.

In addition to categorization, all remaining quantitative attributes are normalized on a [0, 1] scale to facilitate clustering. There are two reasons for conducting normalization operations. First, some machine learning algorithms require normalized attribute representations to function correctly. Second, normalization helps prevent attributes with initially large ranges from dominating attributes with initially smaller ranges (Han & Kamber, 2001). Therefore, the remaining four attributes (account age, credit limit, quantity of additional assets, and late payments) are normalized using the Min–Max normalization (Shalabi, Shaaban, & Kasasbeh, 2006):

$$\text{Normalized Value} = \frac{V - \text{MinValue}}{\text{MaxValue} - \text{MinValue}}$$

where MaxValue and MinValue are, respectively, the maximum and minimum values for a given attribute, and V is the actual value for that attribute.

Finally, the transformed data set is examined and 356 records are eliminated from analysis because of missing, irrelevant, or incorrect values. The fully preprocessed data set contains 194,167 records, two categorized attributes, 4 normalized attributes, and 1 primary key attribute to facilitate record identification.

4.3. Clustering Model Selection

With the preprocessed data set, a key decision involves the choice of clustering algorithm to use. Because the data set used by internal auditing would often be expected to be extensive, computationally expensive approaches with exponential time complexity, such as hierarchical methods, are ruled out. In addition, the density-based spatial clustering of applications with noise (DBSCAN) method is not considered because it eliminates noise points. In this case, noise points would

include the most valuable customers and least valuable customers, which are meaningful and should be retained. By contrast, Simple K-Means is an attractive approach because of its relative efficiency. On the other hand, Simple K-Means tends to be ineffective on data that does not follow a Gaussian distribution. However, visualizing the data suggests that a nearly normal distribution might exist. Therefore, Simple K-Means is chosen as the clustering model for this study.

5. Experiment

In this section, Waikato Environment for Knowledge Analysis (WEKA), the open source software developed by the University of Weikato, New Zealand (Hall et al., 2009), is employed to create the Simple K-Means clustering model. This software offers functionality for many machine learning techniques, including various clustering methods. The following subsections present the steps for using WEKA to create the Simple K-Means clustering model and discuss the clustering results.

5.1. Evaluation Metric

The metric of within cluster sum of squared errors (SSE; Milligan & Cooper, 1985) is used to evaluate the performance of the clustering model. Specifically, SSE is the sum of the squared distance from each instance to its cluster centroid. In general, the lower this value is, the better the clustering results are.

5.2. Parameterization

Before using K-Means to cluster the data set, two of the model's parameters are investigated to obtain optimal overall performance. The parameters for initial seed and number of clusters vary with the data set and directly affect performance of the Simple K-Means clustering model. Thus, those parameters must be adjusted to optimize the performance of the algorithm.

Initial seed: The initial seed value in WEKA is used in generating a random number that is used to make the initial assignment of instances to clusters. In general, Simple K-Means is quite sensitive to how clusters are initially assigned. Thus, it is often necessary to try different initial seeds and evaluate the associated results.[1] In this experiment, some test runs are conducted on the entire data set using a variety of seeds ranging from 0 to 100 in increments of 10. The results, shown in Fig. 1, indicate that when the value of the seed equals 10, the SSE of the clustering model has the smallest value, which suggests that this model offers the best performance. Thus, the value of 10 is adopted as the optimum value for the initial seed.

Number of clusters: In Simple K-Means, the number of clusters is a parameter that must be predefined by the user. The clustering results rely significantly on the

[1] *K*-means clustering in WEKA: http://maya.cs.depaul.edu/classes/ect584/weka/k-means.html

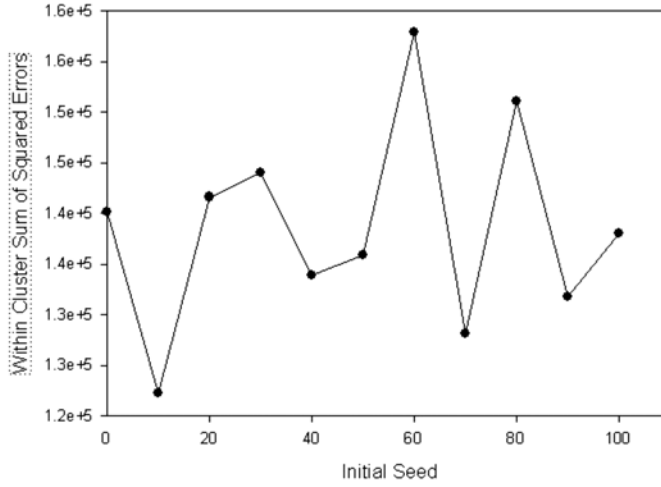


Fig. 1. Initial Seeds and Within Cluster Sum of Squared Errors.

choice of the number of clusters. Therefore, it helps to evaluate Simple K-Means by varying the number of clusters in an effort to identify the optimal value. To determine the number of clusters, the Simple K-Means algorithm is run for a variety of settings ranging from three through 10 clusters. To facilitate decision-making, a graph containing the number of clusters on the x-axis and SSE on the y-axis is constructed. Using the Simple K-Means clustering results, the SSE in each case is plotted for cluster numbers ranging from three through 10, and these points are connected via a series of line segments, as shown in Fig. 2. Next, the resulting graph is examined for "elbows" that suggest diminishing rates of SSE decline in moving from cluster n to cluster n+1. The point where the rate of SSE decline diminishes by the greatest amount will result in the most substantial "elbow," and the number of clusters associated with this point is considered optimal. Fig. 2 shows that such an "elbow" clearly exists at seven clusters. Therefore, optimal performance is achieved when the number of clusters is seven, so this is the number selected for subsequent cluster analysis.

5.3. Modeling

Following parameterization, a clustering model is created by applying Simple K-Means to a data set containing 194,167 records, 2 categorized attributes, and 4 normalized attributes. The initial seed to create the model is set at 10, and the number of clusters is established at 7.

5.4. Results

Table 1 shows the clustering results in terms of cluster centroids. The centroid for a categorical attribute is calculated as the mode of that attribute, while the

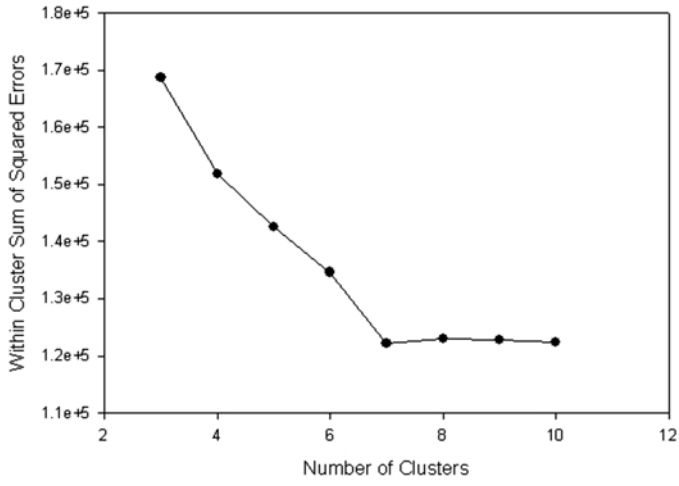


Fig. 2. Numbers of Clusters and Within Cluster Sum of Squared Errors.

Table 1. Clustering Results in Terms of Centroids of Clusters.

Cluster No.	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Frequency	40,180	24,567	31,738	21,833	14,716	31,827	29,306
Credit limit	0.0811	0.1156	0.0403	0.081	0.0276	0.0673	0.1229
Account age	0.2033	0.3042	0.081	0.1774	0.0666	0.1404	0.3493
Behavior score	A	B	D	A	D	C	A
Customer age	B	A	D	D	E	A	A
Addition assets	0.0669	0.0801	0.0581	0.0466	0.0262	0.0874	0.0808
Late payments	0.9296	0.962	0.88	0.9399	0.8796	0.9134	0.9586

cluster centroid for a numeric value is computed as the average of that attribute. The cluster centroids show that the most favorable (Cluster 6) and least favorable (Cluster 4) customer types are convincingly isolated. The second most favorable (Cluster 1) and second least favorable (Cluster 2) customer types are also well distinguished. In addition, the remaining three clusters in between have generally good segregation, although there appears to be a small amount of overlap. Overall, the clustering techniques are successful in identifying seven separate customer types in the analyzed data set.

Table 2 shows the descriptive statistics concerning discounts for customer types. In general, more favorable credit card customers should be considered for higher discounts relative to less favorable credit card customers. Table 2 shows that the most favorable customers (Cluster 6) are actually offered the highest average discount by bank representatives. Furthermore, the least favorable customers (Cluster 4) are given the lowest average discount according to historical records. These findings indicate that the bank representatives have been at least somewhat successful in differentiating extreme customers according to their experiences and/or internal policy mechanisms. However, a continued examination of Table 2 shows that the second least favorable customers (Cluster 2) and intermediate customers (Cluster 5) have been offered comparable discounts. Moreover, the second most favorable customers are offered even lower discounts than two non-extreme types of customers (Cluster 0 and Cluster 5). These findings indicate that the bank representatives are ineffective in distinguishing the customers within the intermediate segments. More important, the discount range for each customer group ranges from 0 to 100%, which strongly demonstrates that bank representatives do not apply objective policies and offer appropriately tailored discounts consistently according to customer type. In summary, the bank's current discount policy is not entirely effective or appropriate for all customer types. Therefore, a more accurate and reasonable discount policy should be formulated to structure the range of eligible discounts properly for different types of customers.

6. Conclusion

This chapter examines the efficiency and effectiveness of the discount policy for credit card customers at a major bank in Brazil. To achieve this objective, clustering techniques are used to determine customer types. After dimensionality reduction, data scaling and normalization, and experimentation, the Simple K-means algorithm is chosen to create clusters based on the six most important attributes. As a consequence, seven groups of credit card customers are identified, and these partitions provide excellent segregation of the most and second most favorable, as well as the least and second least favorable credit card customer groups.

Table 2. Discount Distribution by Customer Type.

Cluster	Frequency	Mean	Std. Dev.	Minimum	Maximum
Cluster 0	40,180	0.67432	0.192803	0	1
Cluster 1	24,567	0.662081	0.180136	0	1
Cluster 2	31,738	0.607236	0.165952	0	1
Cluster 3	21,833	0.683658	0.187253	0	1
Cluster 4	14,716	0.579616	0.157828	0	1
Cluster 5	31,827	0.629324	0.173054	0	1
Cluster 6	29,306	0.737986	0.194335	0	1

The remaining three clusters lying between the extremes also seem generally meaningful. The descriptive statistics on discounts by customer type show that, even though the most favorable and least favorable credit card customers have been offered the highest and lowest discounts on average, the customers within the intermediate clusters have not been consistently offered appropriately tailored discounts. Therefore, a more relevant and reasonable discount policy should be generated to differentiate the discounts offered to the various types of customers.

Future work should investigate discount patterns within each credit card customer cluster and construct a model to determine the optimal discount range for each cluster. Furthermore, a recommendation system could be designed to assist bank representatives in providing optimal discounts. Specifically, when a customer asks for a fee reduction, the bank representative would obtain an optimal discount range for this customer from the recommendation system. To achieve this, the representative would first enter the pertinent customer record information. Following submission of this data, the model would process and classify the record and return the optimal discount range information to the employee. In essence, the intention is for the envisioned recommendation system to provide useful and objective guidance to bank representatives in determining reasonable discounts for various customers to optimize the benefits for both the bank and its customer base.

References

Beirlant, J., Derveaux, V. De Meyer, A. M., Goovaerts, M. J., Labie, E., & Maenhoudt. B. (1992). Statistical risk evaluation applied to (Belgian) car insurance. *Insurance: Mathematics and Economics*, *10*(4), 289–302.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, *11*(1), 10–18.

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. Burlington, MA: Morgan Kaufmann Publishers.

Hirschman, E. C. (1979). Differences in consumer purchase behavior by credit card payment system. *Journal of Consumer Research*, *6*(1), 58–66.

Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, *37*(7), 5259–5264.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research*, *20*(2), 134–148.

Shalabi, L. A., Shaaban, Z., & Kasasbeh, B. (2006). Data mining: A preprocessing engine. *Journal of Computer Science*, *2*(9), 735–739.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*, New Delhi: Pearson Education.

Thiprungsri, S. (2010). Cluster analysis for anomaly detection in accounting data. In *Nineteenth annual strategic and emerging technologies research workshop*, San Francisco.

Thiprungsri, S., & Vasarhelyi, M. A. (2011). Cluster analysis for anomaly detection in accounting data: An audit approach. *The International Journal of Digital Accounting Research*, *11*(17), 69–84.

Tuma, M. N., Decker, R., & Scholz, S. (2011). A survey of the challenges and pitfalls of cluster analysis application in market segmentation. *International Journal of Market Research*, *53*(3), 391–414.

Zhang, T. J., Huang, X. H., Tang, J. F., & Luo, X. G. (2011). Case study on cluster analysis of the telecom customers based on consumers' behavior. In *IEEE 18th international conference on industrial engineering and engineering management (IEandEM)* (pp. 1358–1362). Changchun, China.

This page intentionally left blank

Chapter 6

Predicting Credit Card Delinquency: An Application of the Decision Tree Technique

Ting Sun and Miklos Vasarhelyi

1. Introduction

A credit card account is characterized as being delinquent when the cardholder fails to make a payment before the due date. Credit card issuers face a high risk of credit card delinquency. Roughly 1 out of 20 Americans with credit files[1] are at least 30 days late on a credit card or other non-mortgage bill payment (Ratcliffe et al., 2014). This risk is usually caused by excessive competition among credit card companies for new customers (Ausubel, 1997; Chen & Huang, 2011; Chung & Suh, 2009; Gross & Souleles, 2002; Holmes & Ghahremani, 2015). Credit card delinquency adversely affects the health of the credit card industry because failing to evaluate credit card delinquency risk effectively leads to high non-performing ratio, increased debt collection costs, and/or growing bad debt accounts (Chen & Huang, 2011). As a result, it is crucial for creditors or other financial institutions that are threatened by the resulting financial distress to classify cardholders' behaviors and characteristics and to predict the risk of credit card delinquency (Lin, 2009; Shi, Peng, Kou, & Chen, 2005).

Decision Tree is a mature symbolic data mining technique that "organizes information extracted from a training dataset in a hierarchical structure composed of nodes and ramifications" (Nie, Rowe, Zhang, Tian, & Shi, 2011). Because of its simplicity and flexibility, Decision Tree has had many successful applications for credit card-related problems, such as credit card churn (Nie et al., 2011), credit card fraud (Sahin, Bulkan, & Duman, 2013), and credit card scoring (Lee, Chiu, Chou, & Lu, 2006). However, limited literature applies Decision Tree to predicting credit card delinquency. One study (Yeh & Lien, 2009) compares the predictive performance of several data mining techniques, including K-nearest neighbor classifiers (KNN), logistic regression, discriminant analysis, Naive Bayesian classifier, artificial neural networks (ANNs), and classification trees (CTs) in evaluating the probability of credit card clients' default. However,

[1]Credit files refer to the raw data in the credit file databases of the credit bureaus.

Audit Analytics in the Financial Industry, 71–83

doi:10.1108/978-1-78743-085-320191006

the data that they examine contain 25,000 records with 5,529 default records (22.12%). This is a relatively small data set, and the issue of data imbalance is not as severe as that of most real-life big data sets.

A noteworthy issue is the effect of the current Big Data era (Appelbaum, Showalter, Sun, & Vasarhelyi, 2019). Thanks to the development of information technology, Big Data, which is characterized by high levels of volume, velocity, variety, and veracity, is generated and collectable (Appelbaum et al., 2015; Laney, 2001). As Big Data provides more complete information than traditional data can, analyzing Big Data helps researchers and decision-makers to obtain more accurate predictions. One issue in Big Data mining is data imbalance, which means that the amount of normal data is much greater than the incidents of abnormal data. As a result, it is necessary to search for solutions to deal with data imbalance when examining real-life Big Data.

As little attempt has been made to examine the potential of the Decision Tree approach systematically for credit card delinquency forecasting on imbalanced Big Data, this chapter seeks to closes this gap. Unlike other research that proposes novel classification approaches combining different methods for specific prediction tasks (e.g., Lin, 2009; Shi et al., 2005; Yeh & Lien, 2009), this chapter adopts the C5.0 Decision Tree algorithm to focus on the execution and the implications of the Decision Tree model. This chapter also employs an under-sampling method to handle imbalanced Big Data, which contributes to the existing literature by providing an example of applying the Decision Tree approach to investigate large scale, real-life data.

Two types of factors are generally associated with credit card delinquency: personal characteristics of cardholders and their past spending behaviors (Chen & Huang, 2011). Banks and other financial institutions have increasingly accumulated large amounts of such data. The data used in this chapter are collected from a major bank in South America. The data cover a broad range of information about cardholders and their accounts, such as gender, address, marital status, purchase and payment history, and credit limits. Examining such data helps creditors to issue credit card responsibly, make wise management decisions, and monitor the activities of targeted customers. Therefore, this study explores how credit card holders' personal characteristics and spending patterns are related to payment delinquency.

This study identifies the 10 most important predictors of credit card delinquency. Among these predictors, results show that the value of all unpaid cash withdrawals is the most powerful factor for predicting credit card delinquency, whereas the total amount of installments, the average percentage of the amount that exceeds the installment limit, the quantity of accumulated payments, and the cardholder's residential region are the least important factors. The results also show that the predictive performance of the C5.0 model is satisfactory for big real-life data.

To evaluate the accuracy of the C5.0 decision tree model further, a classification and regression tree (C&RT) model is used to conduct the same research procedure, and the results of both classifiers are compared. The analytical results show that the C5.0 model outperforms the C&RT model in terms of Type II error rate,

sensitivity, *G*-mean, and area under the curve (AUC), whereas the overall accuracy, Type I error rate, and specificity are slightly lower than for the C&RT model.

2. Related Research

Existing literature often applies data mining techniques to analyze problems related to credit cards. Research in this field mainly centers on credit scoring, credit risk, and credit fraud. This research stream usually concerns two dimensions: (1) novel classification methods combining various data mining techniques to develop the prediction model; and (2) statistical hypothesis tests to compare the performance of different data mining methods.

Lee et al. (2006) employ C&RT and multivariate adaptive regression spines to explore their effectiveness in performing credit scoring tasks, compared to the traditional discriminant analysis, logistic regression, neural networks, and support vector machine approaches. Chen, Ma, and Ma (2009) also focus on credit scoring. They propose a hybrid support vector machine technique and verify the feasibility and effectiveness of the proposed model in terms of classification rate and Type II errors. Lin (2009) proposes a two-stage hybrid approach to assess credit risk and points out that the proposed approach outperforms the conventional logistic regression, logarithm logistic regression, and ANN approaches. Twala (2010) investigates five classifiers (ANN, decision trees, Naïve Bayes classifier, KNN, and logistic discrimination) for credit risk forecasting and provide evidence that the predictive accuracy of classifiers can be improved by using classifier ensembles. Chen and Huang (2011) demonstrate that ANN and Decision Tree techniques can successfully identify influential factors for predicting credit risk.

Nie et al. (2011) apply logistic regression and Decision Trees to develop credit card churn prediction models. They use information about the customers, their credit cards, risks, and transaction activity as predictors, and they concluded that logistic regression performs a little better than Decision Trees.

Bhattacharyya, Jha, Tharakunnel, and Westland (2011) attempted to detect credit card fraud using support vector machines, random forest, and logistic regression. They demonstrate that the overall performance of random forests is better across performance measures.

However, there is limited research that applies Decision Trees to predict credit card delinquency. Using a relatively small sample (25,000 records) of credit card data, Yeh and Lien (2009) examine the ability of six data mining techniques to predict the default risk of credit card clients. By comparing the area ratio[2] in the lift chart for classification accuracy estimation and employing a novel technique called the sorting smoothing method (SSM) for predictive accuracy among the six techniques, they find that the ANN performs better than the KNN classifiers, logistic regression, discriminant analysis, Naive Bayesian classifier, and CTs in terms of both classification and predictive accuracy.

[2] Area ratio $= \dfrac{area\ between\ model\ curve\ and\ baseline\ curve}{area\ between\ theoretically\ best\ curve\ and\ base\ line\ curve}$ (Yeh & Lien, 2009).

3. Methodology

In this chapter, Decision Tree is used to predict credit card delinquency. Decision Tree is a popular machine learning technique that is widely adopted to solve various real-world problems (Tsai & Chiou, 2009). It has a tree-like structure composed of nodes and branches that split a data set into branch-like segments. The nodes represent decisions and the branches are their possible consequences. As the paths from root to leaves represent classification rules that describe the relationship between inputs and targets, it is easy to understand and interpret the results of Decision Tree models (Mitchell, 1997). A Decision Tree can be used on data sets including both numerical and categorical data (Lorena & de Carvalho, 2007). Therefore, the Decision Tree method is chosen to predict credit card delinquency in this study because credit card data usually contains both continuous and categorical data fields.

Furthermore, real-world credit card data are usually big and complex, has many variables that contains detailed information, and may have missing data or outliers. These characteristics support the choice of Decision Tree to conduct this research because it overcomes scale difference among variables. In addition, it does not require normalization or scaling, as traditional techniques do, because the tree structure will remain the same with or without these transformations. Another feature of Decision Tree is that it is non-parametric, so missing values or non-linear relationships between variables do not affect the model's development or performance. Decision Trees are also not sensitive to outliers as the splitting takes place based on proportions of samples within the split ranges, not on absolute values.

This study uses C5.0, a major algorithm of Decision Tree, which can be applied to big datasets and is suitable for analyzing continuous and categorical independent and dependent variables. C5.0 is primarily used in data mining situations when the objective is to classify an object into two or more populations and a large number of potential variables must be considered (Lee et al., 2006). For performance evaluation purposes, the results are compared to the results from a model using a C&RT, another well-known Decision Tree algorithm that can also be conducted on both continuous and categorical variables.

The research methodology is shown in Fig. 1. In the data pre-processing and partitioning phase, the original data sets are pre-processed by merging datasets to collect both personal and behavior information for credit card holders, and then random under-sampling is used to eliminate data imbalance. To avoid influencing the over-fitting, training data and data are designated. The product of data pre-processing is separated into two parts: 50% becomes the training set and the remaining 50% becomes the testing set. As the focus of the study is on credit card accounts, some transactional data fields are aggregated to create various derived account-level attributes. Delinquent accounts are those that are permanently blocked by the bank due to delinquency issue. In the modeling phase, a C5.0 model is developed to forecast delinquency. In the performance evaluation phase, a series of traditional evaluation metrics are examined, including overall accuracy, Type I and Type II error rates, sensitivity, specificity, *G*-mean, AUC, and Gini (dispersion). The predictive ability of C5.0 is also compared with that

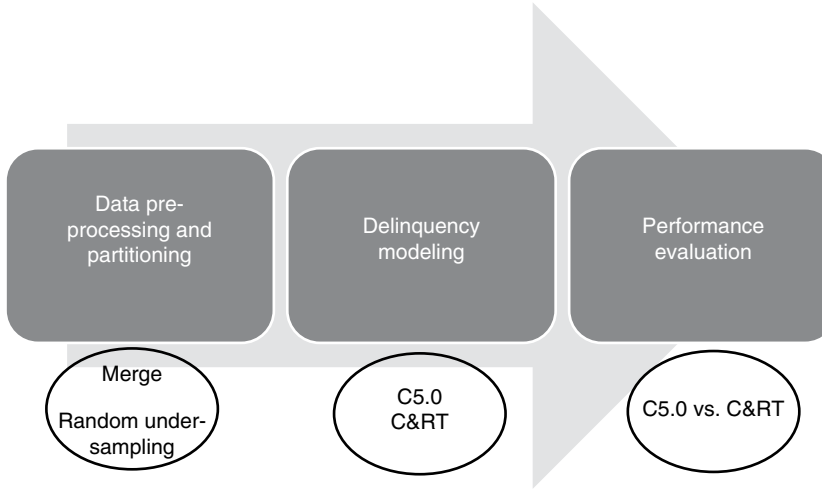


Fig. 1. Research Methodology.

of C&RT, another popular and mature Decision Tree algorithm. This study uses SPSS enterprise modeler to develop and test the model.

4. Experiment and Results

4.1. Data Pre-Processing and Partitioning Phase

Data: The data are obtained from credit card data sets of a major bank in South America. The data are obtained by merging two subsets: detailed transaction records of credit card accounts in July 2013 and personal information of credit cardholders, as recorded in September 2013. The merged dataset has 95 fields with 6,516,045 records, of which 45,017 are delinquent and 6,471,028 are non-delinquent. The data is divided into training and testing data sets, each with 50% of the total observations.

Random under-sampling: The data have unbalanced class sizes of delinquent and non-delinquent records; the non-delinquent records far outnumber the delinquent ones. In the training data, there are 22,481 delinquent records and 3,237,576 normal records. As random under-sampling of the majority class is generally found to be better than other sampling approaches (Van Hulse, Khoshgoftaar, & Napolitano, 2007), random under-sampling of the non-delinquent class is used to obtain training data with reasonable class distributions. Therefore, 0.92% of the normal class of training data is randomly selected so that the new training data have 22,481 delinquent records and 29,903 normal records.

Variables: The dependent variable of the proposed model is "Ifpblk", which equals 1 if the credit card account was permanently blocked by the bank in September 2013 due to delinquency, and 0 otherwise. As this research seeks to identify causes for credit card delinquency, it is not possible to rely on existing theory or hold any assumptions before data mining. In other words, this study does

not suppose that some factors would affect the dependent variable in advance. Instead, it considers as many variables as possible; all variables are included from the transactional records and personal information discussed above.[3]

4.2. Delinquency Modeling Phase

In this phase, a model is developed by using the C5.0 algorithm on the training set. The testing data are imbalanced with 3,233,452 normal records and 22,536 delinquent ones. To enhance the predictive power of the constructed model and to ensure that the result of the performance evaluation is robust, a 10-fold cross-validation is used before applying the constructed model to the testing data. The training data are randomly divided into 10 sub-groups. When testing Group 1 data to check the predictive power of the constructed model, the other nine groups (Groups 2–10) are used as sub-training data for Group 1 model construction. Similarly, when testing Group 2 data, the data in Groups 1 and 3– 10 are used as sub-training data for Group 2 model constructions, and so on. After 10 tests using the same method, the average of the 10 test results is taken as the average rate of prediction accuracy. Then the testing data are used to test the constructed model.

A Decision Tree model has the implicit ability to perform feature selection, and the top few nodes where the tree is split are essentially the most important variables within the dataset. Table 1 displays the 31 input fields of the model.

Based on the C5.0 Decision Tree analysis, the top 10 predictors of credit card default are displayed in Table 2 in the order of their importance. The most important variable is the value of all unpaid cash withdrawals (SLD_TOT_CSH) with a predictor importance of 0.3. Other strong indicators are the portion of an authorized transaction that exceeds the credit limit (PCT_POL) and the average of credit card revolving payments that exceeds the revolving credit limit (UTILIZACAO_ROT_Mean) with predictor importance of 0.14 and 0.12, respectively. Moderately important variables include the cash withdrawal limit (CASH_LIM_VLR), the frequency of the transactions for the account (FREQ), and the total value of authorized transactions for the account (MOED_TRANS_VLR_Sum). The least important among the top ten variables include the total amount of installments (PARC_QTD), the average percentage of the amount that exceeds the installment limit (PARC_EXC_PCT_Mean), the quantity of accumulated payments (QTDEPG), and the customer's residential region code (CODREGRE), each with a predictor importance of 0.03.[4]

4.3. Performance Evaluation Phase

As shown in Table 3, the training data have an estimate accuracy rate of 87.75% for predicting credit card delinquency, with an error rate of 12.25%.

---

[3]Due to the length limit, details of all candidate variables in this study are not provided.
[4]Due to the length limit, the classification rules generated by the model are not reported.

Table 1. Input Variables of the C5.0 Model.

Input Variable	Description	Data Source
SLD_TOT_CSH	The value of unpaid cash withdrawal	Transaction records
QTCASHAF	The purchase value of the invoice	Transaction records
VAL_PGTO_TK	The value of payment	Transaction records
PCT_POL	The value of the portion of the authorized transaction that has exceeded the credit limit	Transaction records
PARC_QTD	The amount of installment	Transaction records
PARC_EXC_PCT_Mean	The average percentage of the portion of the transaction that has exceeded the installment limit	Derived variable
IDADECTA	The number of the times that the account is billed	Personal information
FREQ	The frequency of transactions of the account	Derived variable
QTDEPG	The quantity of accumulated payments	Personal information
UTILIZACAO_ROT_Mean	The average percentage over the limit of credit card payment	Derived variable
RENDADCL	Annual revenue claimed by the customer	Personal information
QTADIC	Quantity of additional assets	Personal information
VLIMPARC	Actual installment limit	Transaction records
IDACTAME	Account age in month	Personal information
CODREGRE	Region code	Personal information
CODESTRES	State code	Personal information
IDADASSO	Customer age	Personal information
SLD_TOT_ROT	Balance of credit card	Transaction records
QTDECRDS	Total number of cards under the account	Personal information
RENDCONF	Annual revenue confirmed	Personal information
MOED_TRANS_VLR_Sum	The sum of the value of authorized transactions	Derived variable
SALDO_CONTA	Credit card balance when the transaction is processing	Transaction records

Table 1. (*Continued*)

Input Variable	Description	Data Source
CASH_LIM_VLR	Cash withdrawal limit	Transaction records
QTCMPRPE	Credit card limit of the card	Personal information
Excesso_calc_Mean	The value of the authorized transaction that has exceeded the limit of credit card payment	Derived variable
UNT_VBX_POL	The amount of authorized transaction per day	Transaction records
VLIMRT_POR	The limit of credit card payment for the account	Transaction records
VLLICRAN	Previous total credit limit (including credit card, cash withdrawal, and installment)	Personal information
SEXO	Gender of the customer	
VLLICRAT	Current total credit limit	Personal information
CASH_EXC_PCT	Percentage of the portion exceed the cash withdrawal limit	Transaction records

Table 4 shows the pattern of matches between a predicted field and its target field for the training data, where rows are defined by actual values and columns are defined by predicted values.

The model is tested on the testing dataset that comprises 3,255,988 samples (3,233,452 normal records and 22,536 delinquency records). Table 5 shows that the testing data have an estimate accuracy rate of 86.57% for predicting credit card default with an error rate of 13.43%.

However, overall accuracy is inadequate as a performance measure because non-delinquent cases dominate in the data, so a prediction of all cases into the majority class will show a high performance value (Bhattacharyya et al., 2011). As a result, multiple measures are needed to provide a comprehensive perspective on the classifier's performance.

Table 6 shows the following results, where positive corresponds to delinquency cases and negative corresponds to non-delinquency cases:

True Positive (TP) = 19,458 True Negative (TN) = 2,799,141
False Positive (FP) = 434,311 False Negative (FN) = 3,078

Type I and Type II error: Type I and Type II error rates are among the most popular measures of the predictive performance of a classifier. A Type I error (false positive) occurs when an account that was not delinquent is misclassified as delinquent, and a Type II error (false negative) happens when a delinquent

Table 2. Predictor Importance.

Variable	Description	Predictor Importance
SLD_TOT_CSH	The value of unpaid cash withdrawal	0.30
PCT_POL	The value of the portion of an authorized transaction that has exceeded the credit limit	0.14
UTILIZACAO_ ROT_Mean	The average of credit card revolving payments that exceeded the revolving credit limit	0.12
CASH_LIM_VLR	The cash withdrawal limit	0.07
FREQ	The frequency of transactions for the account	0.06
MOED_TRANS_ VLR_Sum	The total value of authorized transactions for the account	0.05
PARC_QTD	The value of installments	0.03
PARC_EXC_ PCT_Mean	The average percentage of the value of the transactions that exceeded the installment limit to total transaction value	0.03
QTDEPG	The quantity of accumulated payments	0.03
CODREGRE	The residential region code of the customer	0.03

Table 3. Overall Accuracy of the Training Data.

	Number of Records	Rate (%)
Correct	45,579	87.75
Wrong	6,362	12.25
Total	51,941	

account is misclassified as non-delinquent. Table 7 shows that the Type I error rate is 13.43% and Type II error rate is 13.66%.[5]

[5]Type I error rate = false positive/(true negative + false positive).
Type II error rate = false negative/(false negative + true positive).

Table 4. Coincidence Matrix of training data for Predicted and Actual Delinquent Accounts.

	Predicted Normal Accounts	**Predicted Delinquent Accounts**
Actual normal accounts	25,533	3,927
Actual delinquent accounts	2,435	20,046

Table 5. Overall Accuracy of Testing Data.

	Number of Records	**Rate (%)**
Correct	2,818,599	86.57
Wrong	437,389	13.43
Total	3,255,988	

Table 6. Coincidence of Matrix of testing data for Predictions and Actual Results.

	Predicted Normal Accounts	**Predicted Delinquent Accounts**	**Type I/II Error Rate (%)**
Actual normal accounts	2,799,141	434,311	Type I: 13.43
Actual delinquent accounts	3,078	19,458	Type II: 13.66

Table 7. Performance Comparison for C5.0 and C&RT on Testing Data.

	C5.0	**C&RT**
Accuracy	0.866	0.929
Type I error rate	0.134	0.068
Type II error rate	0.137	0.452
Sensitivity	0.863	0.548
Specificity	0.866	0.932
G-mean	0.864	0.715
AUC	0.927	0.770

Sensitivity, specificity, and G-*mean*: The accuracy of the classification of accounts as delinquent or non-delinquent cases can be assessed through sensitivity and specificity. Along with the *G*-mean, they can indicate desired performance characteristics.

Sensitivity measures the proportion of true positives that are correctly identified out of all true positives and false negatives. By contrast, specificity measures

the proportion of true negatives that are correctly identified out of all true negatives and false positives. The *G*-mean gives the geometric mean of the positive and negative accuracies. These performance measures are defined with respect to the confusion matrix above.

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN}) = 0.863$$
$$\text{Specificity} = \text{TN}/(\text{FP}+\text{TN}) = 0.866$$
$$G-\text{mean} = \sqrt{(\text{Sensitivity} \times \text{Specificity})} = 0.864$$

AUC: AUC is the area under the Receiver Operating Characteristics (ROC) curve,[6] providing a single number (between 0 and 1) summary for the performance of model (Lin, Huang, & Zhang, 2003). The higher the value of AUC, the better the model is. The AUC for this model equals 0.927.

Follow the same model development and evaluation processes, another model is built using a C&RT algorithm so that the prediction performance of the two models can be compared. Table 7 exhibits the explanatory power of the C5.0 and C&RT models in predicting credit card delinquency, as measured by a series of metrics. The detailed predictive results (including overall accuracy and the coincidence of matrix for training and testing data) of the C&RT model are omitted. The predictive performance of the C5.0 model is better than that of C&RT in terms of Type II error rate, sensitivity, *G*-mean, and AUC. The overall accuracy and the specificity of the C&RT model are higher than that of C5.0 because C&RT has fewer Type I errors. However, as the cost to control Type I errors is much lower than to control Type II errors, this study considers the C5.0 model to be preferable.

5. Conclusion

This chapter demonstrates the effectiveness of Decision Tree techniques in predicting credit card delinquency for a real-world credit card big data set. The C5.0 algorithm is used to develop the model. By analyzing a real-life credit card data set, the model identifies the 10 most important predictors, among which the value of all unpaid cash withdrawal (SLD_TOT_CSH) is the most powerful one with the predictor importance of 0.30.

As the results show, the predictive performance of the C5.0 model is satisfactory for this big real-life data set. To evaluate the accuracy of the C5.0 Decision Tree model further, a C&RT model is used to conduct the same research procedure, and the results of both classifiers are compared. The analytical results show that the C5.0 model outperforms the C&RT model in terms of Type II error rate, sensitivity, *G*-mean, and AUC, whereas the overall accuracy, Type I error rate, and specificity are slightly lower for the C5.0 model than for the C&RT model.

[6]The ROC is a graphic plot of the sensitivity (i.e., the number of true positives vs the total number of events), and 1-specificity (i.e., the number of true negatives vs the total number of non-events).

This study demonstrates that data mining techniques can help a financial institution to identify the predictors of delinquent credit card accounts efficiently and effectively. In addition, these results imply that Decision Tree models can be applied to solve other real-world big data problems.

References

Appelbaum, D., Showalter, D. S., Sun, T., & Vasarhelyi, M. A. (2019). Analytics knowledge required of a modern CPA in this real-time economy: A normative position. Working Paper.

Ausubel, L. M. (1997). Credit card defaults, credit card profits, and bankruptcy. *American Bankruptcy Law Journal*, *71*(Spring), 249–270.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, *50*(3), 602–613.

Chen, S. C., & Huang, M. Y. (2011). Constructing credit auditing and control and management model with data mining technique. *Expert Systems with Applications*, *38*(5), 5359–5365.

Chen, W., Ma, C., & Ma, L. (2009). Mining the customer credit using hybrid support vector machine technique. *Expert Systems with Applications*, *36*(4), 7611–7616.

Chung, S.-H., & Suh, Y. M. (2009). Estimating the utility value of individual credit card delinquents. *Expert Systems with Applications*, *36*(2), 3975–3981.

Gross, D. B., & Souleles, N. S. (2002). An empirical analysis of personal bankruptcy and delinquency. *Review of Financial Studies*, *15*(1), 319–347.

Holmes, T. E., & Ghahremani, Y. (May 27, 2015). Credit card delinquency statistics. Retrieved from https://www.nasdaq.com/article/credit-card-delinquency-statistics-cm480951

Laney, D. (2001) 3D data management: Controlling data volume, velocity and variety. Gartner Report. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lin, S. L. (2009). A new two-stage hybrid approach of credit risk in banking industry. *Expert Systems with Applications*, *36*(4), 8333–8341.

Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics and Data Analysis*, *50*(4), 1113–1130.

Ling, C. X., Huang, J., & Zhang, H. (2003). AUC: A statistically consistent and more discriminating measure than accuracy. In *Proceedings of the international joint conferences on artificial intelligence* Acapulco, Mexico (pp. 519–526).

Lorena, A. C., & de Carvalho, A. C. P. L. F. (2007). Protein cellular localization prediction with support vector machines and decision trees. *Computers in Biology and Medicine*, *37*(2), 115–125.

Mitchell, T. (1997). *Machine learning*. New York, NY: McGraw-Hill.

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, *38*(12), 15273–15285.

Ratcliffe, C., McKernan, S.-M., Theodos, B., Kalish, E. C., Chalekian, J., Guo, P., & Trepel, C. (2014). *Delinquent debt in America*. The Urban Institute Report. Retrieved from http://www.urban.org/research/publication/delinquent-debt-america/view/full_report

Sahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, *40*(15), 5916–5923.

Shi, Y., Peng, Y., Kou, G., & Chen, Z. (2005). Classifying credit card accounts for business intelligence and decision making: A multiple-criteria quadratic programming approach. *International Journal of Information Technology and Decision Making*, *4*(04), 581–599.

Tsai, C. F., & Chiou, Y. J. (2009). Earnings management prediction: A pilot study of combining neural networks and decision trees. *Expert Systems with Applications*, *36*(3), 7183–7191.

Twala, B. (2010). Multiple classifier application to credit risk assessment. *Expert Systems with Applications*, *37*(4), 3326–3336.

Van Hulse, J., Khoshgoftaar, T., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on machine learning*, Corvallis, Oregon, June. pp. 935-942.

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473–2480.

This page intentionally left blank

Part III

Analytics in Insurance Audits

This page intentionally left blank

Chapter 7

Cluster Analysis for Anomaly Detection in Accounting*

Sutapat Thiprungsri

1. Introduction

Understanding the data is an important step in any analysis. When planning an audit, auditors need to understand the client and the client's data in order to plan effective and efficient audit procedures. Auditors can use several methods to obtain information and knowledge about their clients. For example, auditors can calculate statistical values, such as maximum, minimum, median, and standard deviation. In addition to these classic statistical methods, powerful data mining methods are gaining popularity as computer systems become less expensive. For example, cluster analysis can be used as an important exploratory tool for knowledge discovery.

This research study applies a conceptual clustering tool to a data set from a real company. The objective is to demonstrate the use of this method for exploratory data analysis (EDA) in a real-world setting. The chapter begins by providing an explanation of EDA and discussing how cluster analysis can be used for data exploratory purposes by outlining the audit problems that the sample company is facing. The structure of the data set, the research methodology, and the results are then presented.

2. Exploratory Data Analysis

Practical data analysis can be divided into two phases: exploratory and confirmatory (Tukey, 1977). EDA emphasizes flexible searching for clues and evidence, whereas confirmatory data analysis stresses evaluating the available evidence (Hoaglin, Mosteller, & Tukey, 1983). The method of choice for evaluating

*This chapter is based on the study that is published in the *International Journal of Digital Accounting Research* (Thiprungsri & Vasarhelyi, 2011).

Audit Analytics in the Financial Industry, 87–110

doi:10.1108/978-1-78743-085-320191007

data is a matter of philosophy. EDA can be characterized as a search for regularity or structure among objects in an environment, and the subsequent interpretation of discovered regularity (Fisher & Langley, 1985; Tukey, 1977). The most common forms of exploratory data examination are classical statistics and specialty graphical displays. Many tools can be used to represent structures in data using classical method, including numeric summaries (e.g., means, median, minimum, maximum, and summation) and statistical distributions (e.g., stem-and-leaf displays, bar graphs, and other types of graphs).

With the advent of advanced technologies and the decreasing cost of computer systems, a technique using a series of artificial intelligence (AI) methods for EDA is becoming more popular. This technique uses AI for machine learning. The major distinction between methods of machine learning and statistical data analysis comes from differences in the ways that the techniques represent data and structure within the data. For example, machine learning methods are strongly biased toward symbolic data representation (Fisher & Langley, 1985). Whether using statistical or machine learning methods, the objective of the data exploratory task is to construct classification schemes for an initially unclassified set of data.

3. Cluster Analysis for Data Exploratory Purposes

One alternative to using machine learning methods for data exploratory purposes is to use cluster analysis. Fisher and Langley (1985) define the abstract clustering task as follows:

> Given: A set of object, O.
>
> Goal: Distinguish clusters (subsets of O) $S_1,\ldots, S_n$, such that intra-cluster object similarity of each S_i tends to be maximized, and the inter-cluster object similarity over all S_j's tends to be minimized. A collection of clusters is termed a "clustering."

Michalski (1980) considers conceptual clustering as an extension to the method of numerical taxonomy. Sokal and Seneath (1963) define numerical taxonomy as a classification system in biology that uses numerical methods to group taxonomic units based on their character states, rather than using subjective evaluation of properties. The similarity between two objects is the value of a numeric function applied to the descriptions of the two objects (Everitt, 1980).

The numerical taxonomy technique implicitly assumes that objects or observations can be represented using continuous values. However, not all objects or observations can be measured as continuous values. There are four general levels of measurement (nominal, ordinal, interval, and ratio), which lead to four kinds of scales (Kerlinger & Lee, 2000). The lowest level of measurement is nominal measurement, which is used for nominal categories. For example, gender can be categorized into male and female; religion can be categorized into Catholic, Protestants, Buddhist, Muslim, and others. The second level of measurement is

ordinal measurement. This level requires that the objects in a set can be rank ordered on an operationally defined characteristic or property (Kerlinger & Lee, 2000). For example, customer satisfaction can be ranked as high, medium, or low. In this example, it is possible to say high > medium > low. However, it does not give information on how different each step on the scale is from others. The third measurement level is interval measurement. In addition to having the characteristics of nominal and ordinal scales, interval has numerically equal distances between each level of the interval scale for the property being measured (Kerlinger & Lee, 2000), and can be added and subtracted. For example, height, weight, and age can be considered as interval measures. The highest level of measurement is ratio measurement. In addition to containing characteristics of nominal, ordinal, and interval measures, ratio has an absolute or natural zero, which means that if an object is measured as zero on the ratio measurement, it has none of the property being measured.

Some objects can be measured using multiple measurements. For example, if the object is a person, the values (or variables or attributes) of interest could be age, height, weight, marital status, educational background, race, etc. Some of these variables are ordinal measures (i.e., age, height, and weight) and others are nominal values (i.e., marital status, education background, and race). To use a numerical taxonomy to represent similarity among people would be a challenge because these objects are represented with several types of measures.

Moreover, although numerical taxonomy techniques are useful, the resulting clusters or groups may not easily be characterized in a generalized conceptual language that can be used to hypothesize about future observations (Fisher & Langley, 1985). Michalski (1980) addresses the problem of determining conceptual representations of object clusters and defines conceptual clustering. Given a set of concepts C that may be used to describe structures within object set O, Michalski (1980) defines the similarity between two objects, A and B, as:

$$Similarity(A,B) = f(A',B',O',C)$$

In simple terms, the similarity between two objects depends on the quality of the concepts used to describe them. Conceptual clustering defines the similarity between objects as a set of values common to all objects in an object group.

Cluster analysis has been extensively used in marketing and other fields to understand the patterns and behavior of the data set. For example, Erdogan, Deshpande, and Tagg (2007) examine consumer media habits by utilizing cluster analysis to understand readership patterns in a medical journal. The results of this type of study help interested parties, such as the pharmaceutical industry, to allocate appropriate resources to their marketing campaigns. Chang, Lai, and Chang (2006) investigate consumers' typical modes of expression by using hierarchical clustering analysis to re-categorize the total set of expression categories. The categories are clustered into meaningful and distinct expression modes. Lim, Acito, and Rusetski (2006) use the case survey methodology to group the characteristics of international marketing campaigns of multi-national firms using

hierarchical clustering via Ward's method to minimize within-cluster variance. Srivastava, Leone, and Shocker (1981) use a hierarchical clustering approach to cluster products based on a substitution-in-use in market structure analysis. Two important factors for the analysis are the degree of substitution of the product and the method. Morwitz and Schmittlein (1992) examine sales forecasts based on purchase intentions utilizing various methods of partitioning to determine whether segmentation methods can improve the aggregate forecasts. Shih and Liu (2003) study customers' relationships to management using a *K*-means clustering methodology to group customers with similar lifetime value or loyalty in the retail hardware business.

Although there is much potential for using cluster analysis on accounting data to understand the nature of transactions, few studies have used this approach. Therefore, more research is needed in this area.

4. The Audit Problem

This research study deals with a large international bank that identifies transitory accounts as a major area of risk. Transitory accounts act as temporary resting places for bank transactions that cannot be completed immediately after being entered into the bank's system. The transaction is kept in a transitory account until the issue is resolved. Problems involving transitory accounts may include a non-existent destination account for money transfers, a higher transaction amount than the balance of the account, or an inactive account. Once the problem is resolved, the transaction is cleared. The remaining balance should then be "0."

The audit/management questions involve:

- Why did a transaction end up in a transitory account?
- What was done to resolve the issue with the transaction?
- Was the action/change made for the transaction proper?

Questions about what happens to transactions posted to transitory accounts are of great concern to the bank. Large, incorrect entries could eventually create materially incorrect financial statements and substantial losses due to fraud or error. Consequently, transitory accounts are included in a continuous auditing program aimed at monitoring key accounts and filtering potentially fraudulent transactions.

Little information is known about the regularity and/or behavior of transitory accounts. Therefore, they represent a suitable scenario for EDA. The auditor will gain an understanding about these transitory accounts and can identify possible risks and problems with the system.

5. Data

5.1. General Information

The sample transaction data set covers the period from January 2008 through December 2008. It encompasses information from 16 transitory accounts that

have been pre-selected by the bank as the accounts of interest. A detailed description of each account given by the bank is listed in Table 1. The purpose and origin of each transitory account is different. For example, account 302 is for operations in process, which is a form of inter-departmental transfer; Account 61930 is for wire transfers for financial applications; Accounts 70050 and 70068 are for returns of inter-departmental transfers. No matter where these transactions originated, they are posted to the transitory account because the system cannot complete the transactions at the time they are presented. The problems can occur for various reasons, including incomplete account numbers, incorrect account numbers, and/or insufficient funds.

The frequency distribution of transactions in each account is presented in Table 2.

Table 1. Detailed Information on the Transitory Accounts.

Account Number	Description
302	Operations in process: inter-departmental transfer
1155	Banker's check still outstanding
5738	(-) DAV (clearance) debit reclassification: checking acct
21776	Return of the check paid by clearing house
21830	Adjustment DAD (clearance of compensation) TITULOS/check/DOC
32360	Incorrect cash in treasury (maybe bulky multi-purpose acct)
45136	FAI: pending operations in process
58122	SPB (cleaning house for TED): credits between banks resubmitted
60836	Rejected received TED
61042	Clearance credit reclassification
61930	Wire transfer for financial application
66613	(-) Reclassification of debit CY (name of system) to correct wrong classification
68128	USB (department of process) occurrences debtors
70050	Returns of inter-departmental transfer: one dept charges the other and turned out to be wrong so returned
70068	Returns of inter-departmental transfer: one dept charges the other and turned out to be wrong so returned
94870	Value received Bandeirantes TED STR

Table 2. Frequency Distribution of Transactions.

Account Number (LANVFCDPCB)	Frequency	Percent (%)
302	5,652	1.41
1155	688	0.17
5738	49,539	12.37
21776	28,741	7.18
21830	23,926	5.98
32360	67,187	16.78
45136	73,375	18.32
58122	20,395	5.09
60836	91,660	22.89
61042	12,114	3.03
61930	1,042	0.26
66613	2,426	0.61
68128	19,565	4.89
70050	2,715	0.68
70068	889	0.22
94870	503	0.13
Total	400,417	100.00

5.2. Attributes

The 16 attributes extracted from each transaction are shown in Table 3.

The first three attributes, LANVFCDUNID, LANVFCDPCB, and LANVFNUNSUE, are identification types. They provide the branch number, account number, and transaction number. Four of the attributes provide information related to date and monetary value when the transaction is originally posted and when the transaction is last posted on the transitory account. These attributes are LANVFDTLANC (original date), LANVFVLLANC (original monetary value posted), LANVFDTULBX (last date posted) and LANVFVLSALD (remaining balance). Six of the attributes (LANVFCDORLC, LANVFINRESP, LANVFCDSTPR, LANVFINEXCE, LANCDMULIUNID, and LANCDEVENNEGO) are no longer used by the bank. Two of the attributes give information on the nature of the transaction. LANVFINDBCR indicates whether the transaction is debit (8) or credit (9) type, and ANVFCDFUNC specifies whether the transaction is entered automatically or manually.

The attribute that appears to contain the most detailed information related to the transaction is the comment field or LANVFDCCOMP. This attribute is rich with information content. It can contain many details, such as account number, branch number, dates, and reasons why the transaction is posted to the transitory account.

Table 3. Attribute Information.

Attribute Name	Description
LANVFCDUNID	Bank unit no. branch or administrator
LANVFCDPCB	Internal Acct #, Old branch #, and the table name
LANVFNUNSUE	(unique) Transaction #
LANVFDTLANC	Date of entry in the table
LANVFDCCOMP	Comment
LANVFINDBCR	8 for debit, 9 for credit
LANVFVLLANC	Amount
LANVFVLSALD	(residual) amt (= balance)
LANVFDTULBX	Date of last entry on the balance
LANVFCDFUNC	Both 0 and *999999999* for automatic entry
LANVFCDORLC	System of origin that feeds the table. Validating from what system it is coming from (not used any more)
LANVFINRESP	Manual entry indicator (not used any more)
LANVFCDSTPR	Status of processing (not used any more): 0 – sent but not processed, 1 – processed
LANVFINEXCE	Account overage (not used any more)
LANCDMULIUNID	Multi-branch data (not used any more)
LANCDEVENNEGO	Manual entry indicator (not used any more)

This information can be entered automatically by the system, entered manually by an employee of the bank, or a combination of automatic and manual entry. Generally, there is no predetermined format for the way that comments should/could be entered. Although this attribute seems to contain valuable information, it is extremely difficult, if not impossible, to process the information for further analysis due to its free-field format. Therefore, a parsing procedure is developed to make the attribute readable and understandable to the computer system. The parsing procedure aims at breaking the comments into smaller, more manageable forms.

The structures of the comment vary widely from account to account (e.g., comments belonging to account #302 are totally different from those belonging to account #5738), so the parsing procedure must be tailored to a specific account. Once the procedure is fully developed, it can become a prototype for developing similar procedures for other accounts.

5.3. Transitory Account for Debit Reclassification of Checking Account

Because of the differences in the structure of each comment, the parsing procedure developed is specific to the account. The account selected for this study is

#5738. It is a transitory account for debit reclassification of a checking account. The reasons for selecting this account include the large number of observations, which are sufficient for the analysis, and the organized structures of the comment, which make it possible to develop the parsing procedures.

Each attribute is closely examined to determine if it can be used as input for clustering procedures. Four attributes (LANVFINDBCR, LANVFCDSTPR, LANVFINEXCE, and LANCDMULIUNID) have only one possible value. Two date attributes (LANVFDTLANC and LANVFDTULBX) have errors: They have "1900" as the year value. Therefore, they are excluded for the analysis. One attribute (LANVFDCCOMP) contains a free format field with comments from employees concerning the particular transactions. Two attributes (LANVFVL-LANC and LANVFVLSALD) are continuous values. From the original sixteen attributes, LANVFCDFUNC, LANVFCDORLC, LANVFINRESP, and LANCDEVENNEGO are the remaining candidates for conceptual clustering, so the cluster analysis is performed on these four attributes. The frequency distribution for this data is shown in Table 4.

The parsing procedure for Account #5738 is presented in the next section, followed by the clustering procedure and the analysis of the results.

6. Parsing Procedure

The structure of the data file and an example of the values for Account #5738 are presented in the Fig. 1. The open comment attribute, LANVFDCCOMP, is rich with additional information. It is either automatically entered by the system,

Table 4. Distribution of the Four Remaining Attributes.

LANVFCDFUNC		**LANCD£V£NN£GO**	
0	49,316	603	49,307
1120423	208	30,397	6
1173444	12	(blank)	226
1180885	3	Total	49,539
Total	49,539		
LANVFINRESP		**LANVFCDORLC**	
A	3	11	222
U	223	15	109
(blank)	49,313	29	3
Total	49,539	41	49,199
		51	6
		Total	49,539

LANVFCDUNID	LANVFCDPCB	LANVFNUNSUE	LANVFDTLANC	LANVFCDTPTR	LANVFDCCOMP	LANVFINDBCR	LANVFVLLANC	LANVFVLSALD	LANVFDTULBX	LANVFCDFUNC	LANVFCDORLC	LANVFINRESP	LANVFCDSTPR	LANVFINEXCE	LANCDMULIUNID	LANCDEVENNEGO
1	5738	10200001941	15-Jan-08	0	AGENCIA = 1 CONTA = 2121828 72 - LCTO BLOQUEADO P/ CONTA PARALISADA 02030 DEVOLUCAO CHEQUE DEPOSITADO CN00012121828200801159940 62BCN21DA	8	900	0	16-Jan-08	0	41		1	0	0	603
1	5738	20200001941	15-Jan-08	0	AGENCIA = 1 CONTA = 2121828 72 - LCTO BLOQUEADO P/ CONTA PARALISADA 02030 DEVOLUCAO CHEQUE DEPOSITADO CN00012121828200801159940 63BCN21DA	8	805	0	17-Jan-08	0	41		1	0	0	603
46	5738	20800046941	15-Jan-08	0	AGENCIA = 46 CONTA = 2023243 03449 *ASSINATURA EDITORA ABRIL IB 88497554244039810310361601081501	6	51.4	0	18-Jan-08	0	41		1	0	0	603

Fig. 1. Data Structure.

manually entered by a bank employee, or a combination of both. In addition to account and branch number, LANVFDCCOMP generally contains information related to reasons why the specific transaction could not be completed and was consequently transferred to the transitory account. Possible reasons include incomplete account number, insufficient funds, and inactive account. When the information is passed on through computer systems, the data in this field are combined, making it unusable for further analysis. Therefore, the comment field must be parsed so that the individual comment items can be used. Perl 5.10 is used for the parsing.

The automatic or partially automatic comments are comprised of 2–5 preset strings. These preset strings are separated from each other by blank spaces. Therefore, the blank spaces are used as the criteria for parsing these automatic and partially automatic comments.

The open comment attribute, LANVFDCCOMP, is separated from the other attributes for the parsing procedure. After the comment attribute is parsed into a set of smaller and understandable attributes, they are integrated back into the original list of attributes. The maximum of five attributes would be created for each comment field. They are named "part1", "part2", "part3", "part4", and "part5". The parsing procedure is shown in Fig. 2.

7. Banks Processing System

In addition to the preset strings, the comment field also includes a string that appears to be coded information transferred between various computer systems.

The bank's main system is created from hundreds of smaller systems including legacy and newer systems. Therefore, the system has created a coding system for transferring between computer systems so that the system analyst can identify the system that is the source of the information.

7.1. System Identification

The characteristic of a system identification string normally begins with characters that are followed by a string of numbers. An example of the system identification string is shown in Fig. 3. The characters are a code name for a specific system, whereas the numeric string appears to have a structure or format. However, at present, it is not possible to parse the numeric string in order to get to the real information embedded. Therefore, the string is coded as "SYS" to represent the information transferred with the computerized code. This string is also assigned to one of the parsed comment fields ("part1", "part2", "part3", "part4"and "part5") depending on its respective position.

The parsing procedure is presented in Fig. 3. In this illustration, the system identification string is coded as "SYS" in the last step.

The parsed comment fields are used as the attributes for clustering. However, they are coded prior to feeding them to the data mining software in order to reduce the memory and disk requirements needed for the operation. Most preset strings are reduced from 30–50 characters to less than 10 characters. Some preset

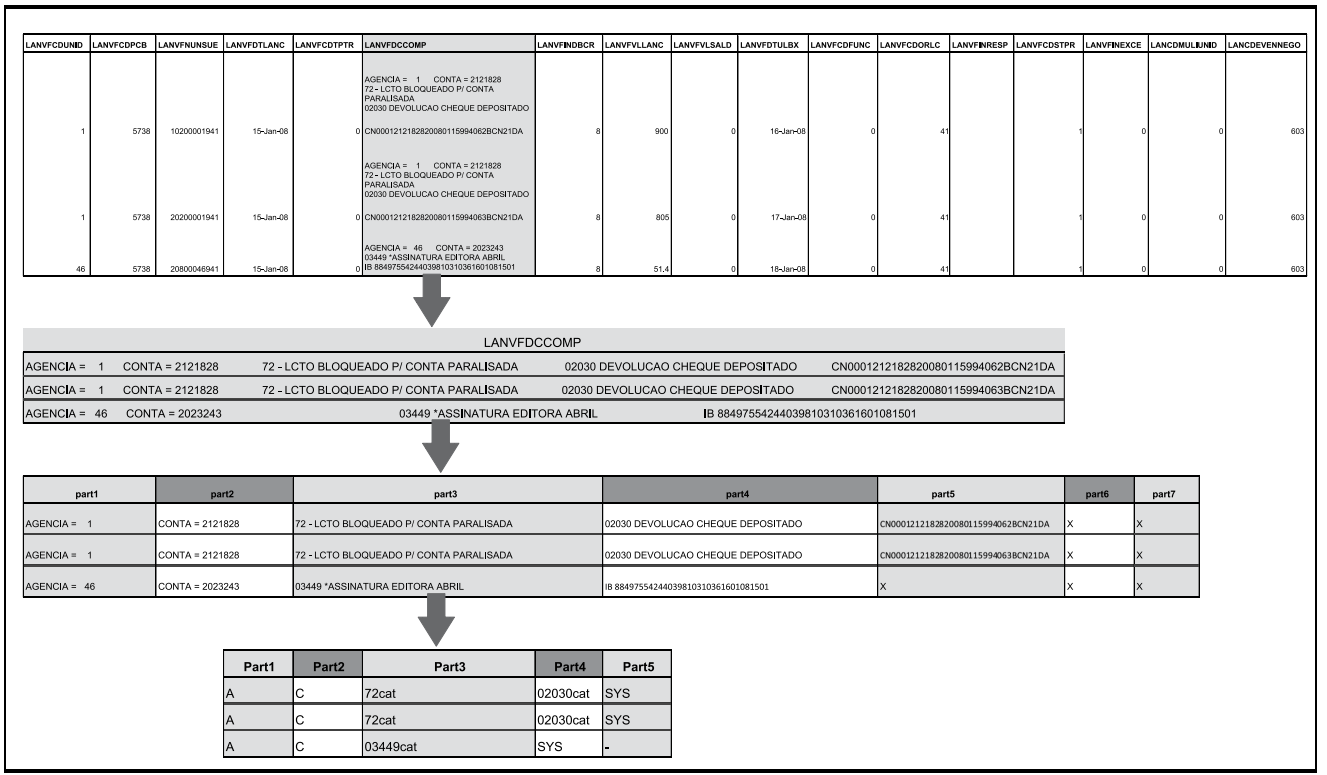


Fig. 2. Parsing Procedure.

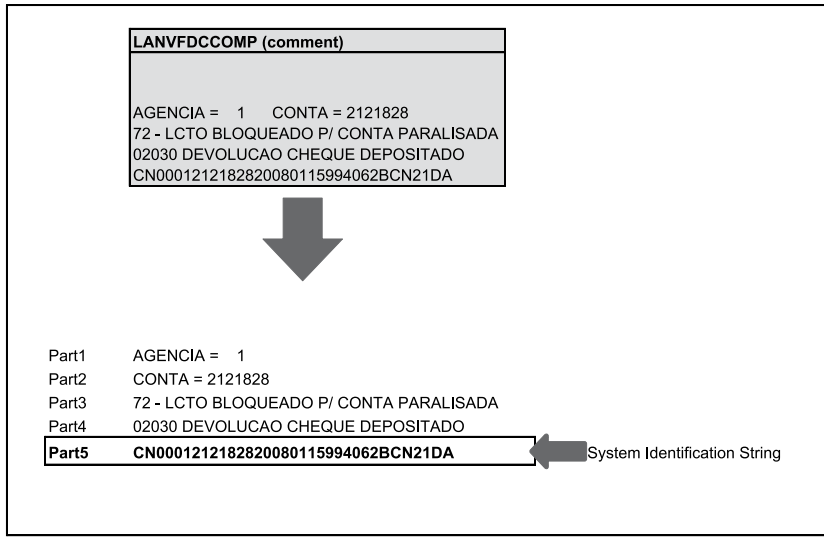


Fig. 3. Example of System Identification String in the Comment.

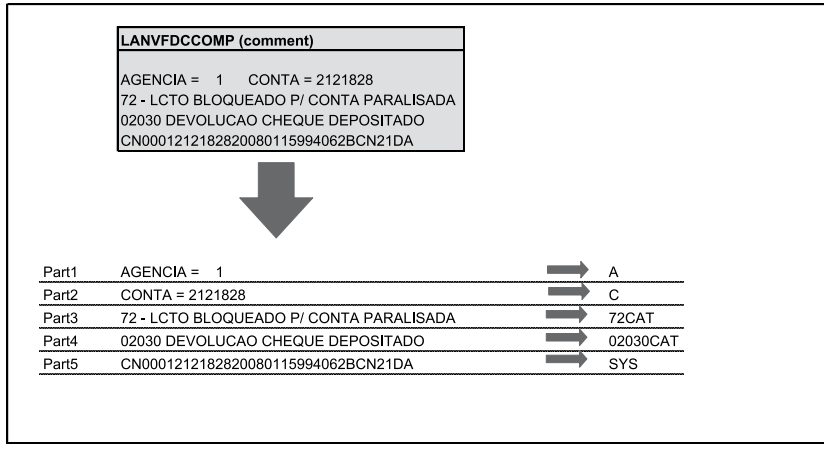


Fig. 4. Comment Coding Illustration.

strings are reduced to only one character. For example, “AGENCIA = 1100” is shortened to “a” to represent information about “agency” or branch. The objective at this coding stage is not to pick up information on specific identity, such as “1100” from the string to represent “AGENCIA = 1100”, but to recognize the pattern that this string refers to or gives information on the agency number or branch number. In addition, as mentioned in the previous section, the alphanumeric string representing the computer-coded information transferred between computer systems are coded as “SYS”. The steps taken to code the comment are shown in Fig. 4.

7.2. Clustering Procedure

The parsed comment field is used as the input variable for clustering. Because neither the distribution of the input variables nor other parameters necessary for cluster analysis are known, the expectation–maximization (EM) algorithm will be adopted for the clustering (Witten & Frank, 2005). The EM algorithm involves two steps: (1) "expectation" is the calculation of the cluster probabilities and (2) "maximization" of the likelihood of the distributions given the data. For this method, only the cluster probabilities, not the clusters themselves, are known for each observation. They could be considered as weights. Witten and Frank (2005) state that if w_i is the probability that observation i belongs to cluster A, the mean and standard deviation for cluster A are

$$\mu_A = \frac{w_1 x_1 + w_2 x_2 + \ldots\ldots + w_n x_n}{w_1 + w_2 + \ldots\ldots + w_n}$$

$$\sigma_A^2 = \frac{w_1 (x_1 - \mu)^2 + w_2 (x_2 - \mu)^2 + \ldots + w_n (x_n - \mu)^2}{w_1 + w_2 + \ldots + w_n}$$

The EM algorithm converges toward a fixed point that can be identified by calculating the overall likelihood that the data come from this data set, given the values for all parameters. However, it will never actually reach that point. This likelihood can be calculated by multiplying the probabilities of the individual observations i:

$$\prod_i (p_A \Pr[x_i \mid A] + p_B \Pr[x_i \mid B])$$

where the probabilities, given clusters A and B, are determined from the normal distribution function $f(x;\mu,\sigma)$ (Witten & Frank, 2005). The likelihood is a measurement of the goodness of the clustering process. This number should increase after each iteration.

7.3. Results

Although the most commonly used techniques for EDA are statistical methods, especially graphical displays, cluster analysis can also be used for EDA. As discussed earlier, cluster analysis can be used to find the common characteristics in a data set. Understanding the characteristics of a data set can help the user to make sense of the data. There are many possible uses for this technique including developing specific strategies or policies, and anomaly detection techniques.

Using the four original attributes, LANVFCDFUNC, LANVFCDORLC, LANVFINRESP, and LANCDEVENNEGO with the EM algorithm for conceptual clustering creates three clusters. The result is shown in Fig. 5. Based on the four original attributes, 99% (49,199) of the transactions are grouped into cluster 0. Only 223 transactions are grouped into cluster1, and 117 transactions are grouped into cluster2. When observing the original data set, it becomes clear that cluster1 and cluster2 are transactions for which comments are entered

```
=== Run information ===
Scheme:       weka.clusterers.EM -I 100 -N -1 -M 1.0E-6 -S 100
Relation:     5,738Main
Instances:    49,539
Attributes:   7
              LANVFCDFUNC
              LANVFCDORLC
              LANVFINRESP
              LANVFCDSTPR
              LANCDEVENNEGO
Test mode:    evaluate on training data
=== Model and evaluation on training set ===
EM
==
Number of clusters selected by cross validation: 3
Cluster
Attribute                     0          1          2
                         (0.99)          0          0
LANVFCDFUNC
  mean                        0 1,124,089.5          0
  std.dev.           75,255.23  13,675.25     3.4191

LANVFCDORLC
  mean                       41    11.0179    17.2051
  std.dev.               2.3463     0.2673     8.1624

LANVFINRESP
  U                      49,200        224        115
  A                           1          1          4
  [total]                49,201        225        119
LANVFCDSTPR
  mean                        1          1     0.9487
  std.dev.                0.011      0.011     0.2206

LANCDEVENNEGO
  mean                      603   606.6251 2,130.9904
  std.dev.           327.8754          0 6,571.7138

Clustered Instances
0     49,199 ( 99%)
1        223 (  0%)
2        117 (  0%)
Log likelihood: -16.99899
```

Fig. 5. Clustering Result from Using the Four Remaining Attributes.[1]

[1]The screenshot was captured from the software WEKA (Frank, Hall, & Witten, 2016).

manually. The reasons given for those transactions vary, but the nature of the comments is the same: they are manual comments. They could have been entered either partially or completely manually.

Fig. 6 shows that when the four original attributes are used as the clustering attributes, this data set can be grouped to a set of meaningful clusters. Dots with the same shade are the transactions in the same cluster. The more homogeneous the shades in the group are, the clearer or better the clustering results appear to be.

The resulting cluster is clear (i.e., unseen observations can easily be classified into groups) and meaningful, but it may not necessarily be useful. For example, in this clustering result, it is possible to tell which transactions are entered manually. Whether the knowledge about the nature of the transaction is useful is an important question to consider.

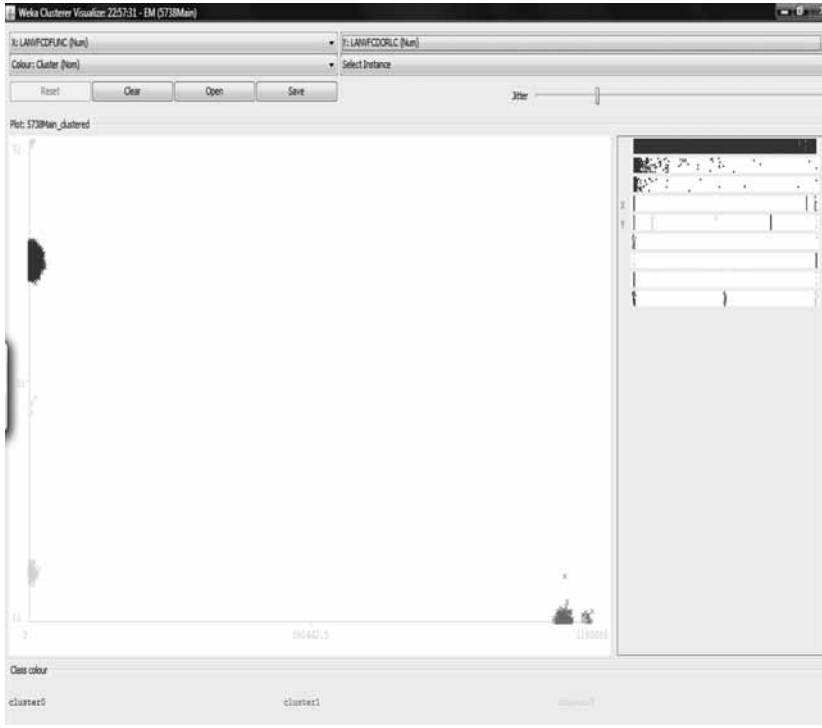


Fig. 6. Visualization of Clustering Results (LANVFCDFUNC and LANVFCDORLC).[2]

[2]The screenshot was captured from the software WEKA (Frank et al., 2016).

When using the open comment attribute with the EM algorithm, seven clusters are created. Dominant characteristics or the content of the comment in each cluster are summarized in Table 5.

Transactions with similar comments are grouped together. For example, the transaction with non-existent accounts (01 – CONTA INEXISTENTE) would be in cluster2. There is additional information relating to the non-existent accounts presented in part4. These clusters represent different reasons that the transactions are transferred to this transitory account. The visualization of the assigned clusters is shown in Figs. 7 and 8.

With the limitations of the visualization in this chapter, the attributes and their relationships can only be shown three dimensions at most: *x*-axis, *y*-axis and color.

Table 5. Content of the Comment Fields in Each Cluster.

Cluster No.	**Content of the Comment Fields**
cluster0	70 – RESOLUCAO 2025-CLIENTE NAO RECADASTRADO with other message
cluster1	02 – VALOR LANCTO MAIOR QUE O SALDO with 01902 *TAXA ADMINISTRATIVA OR 02 – VALOR LANCTO MAIOR QUE O SALDOb with 06246 *DEB AUT REVISTA SELECOES
cluster2	01 – CONTA INEXISTENTE with other message
cluster3	Transaction with the comment that have only one message (in addition to the agencia, conta, and the alphanumeric phase generated from the system) AND the alphanumeric phase generated from the system
cluster4	50 – LANC.INVALIDO P/PRODUTO HOT-OVER OU COMPROR with other message
cluster5	71 – LCTO BLOQUEADO P/CONTA NAO HABILITADA with other messages 72 – LCTO BLOQUEADO P/CONTA PARALISADA with other messages
cluster6	02 – VALOR LANCTO MAIOR QUE O SALDO with other message but NOT 01902 *TAXA ADMINISTRATIVA OR 06246 *DEB AUT REVISTA SELECOES – Transaction with only one message (in addition to AGENCIA and CONTA) WITHOUT alphanumeric phrase generated from the system – Transactions that have the messages too varied to parsed

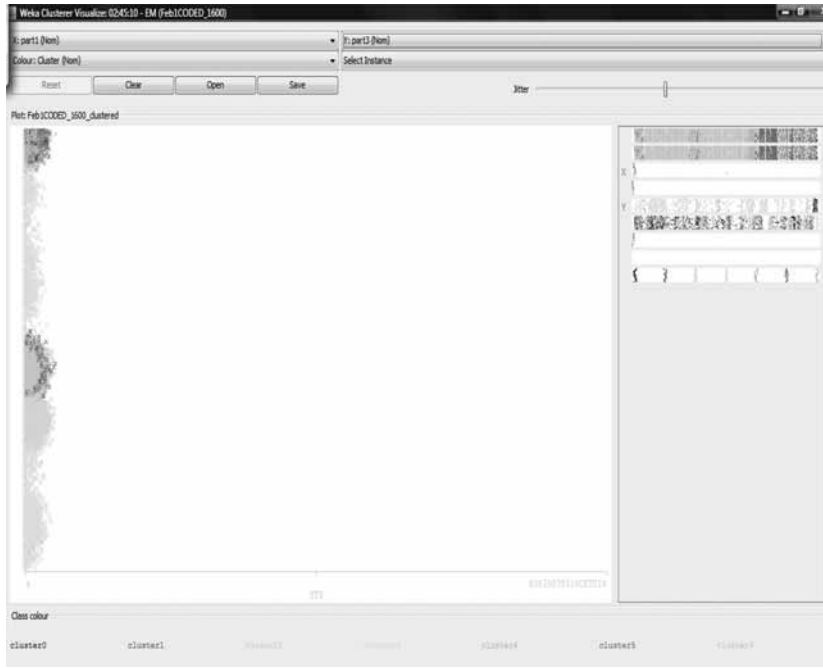


Fig. 7. Assigned Clusters and the Value in part1 and part3 of the Comment Field.[3]

Therefore, even though cluster analysis is a multidimensional analysis (i.e., several attributes can be used as the input at the same time), only three dimensions of the input attributes and the results can be shown. The assigned clusters are visualized by using color; the other two attributes would be visualized using the *x*-axis and *y*-axis. The use of color makes the distinction and/or the interpretation of the results easier. By observing dots that represent observations in different colors, the reader can easily perceive that they form different groups or clusters.

Although the free form comment field has no standardized format, it could be parsed or separated into four parts. Most transactions have the branch (AGENCIA) and account number (CONTA) as the first two parts of the free form comment. The remaining two parts are mainly coded messages (a coded number and a written message). Because the first two parts of the comment are mostly (but not always) the same, it is justifiable to select part3 and part4 for visualization purposes.

Transactions having similar comments are grouped into the same clusters. The graphical visualization shows that there are clusters. Dots represent individual transactions, and the same colors represent the same clusters. Fig. 8 highlights that the most grouped transactions create large areas of a single color. The *x*-axis

[3]The screenshot was captured from the software WEKA (Frank et al., 2016).

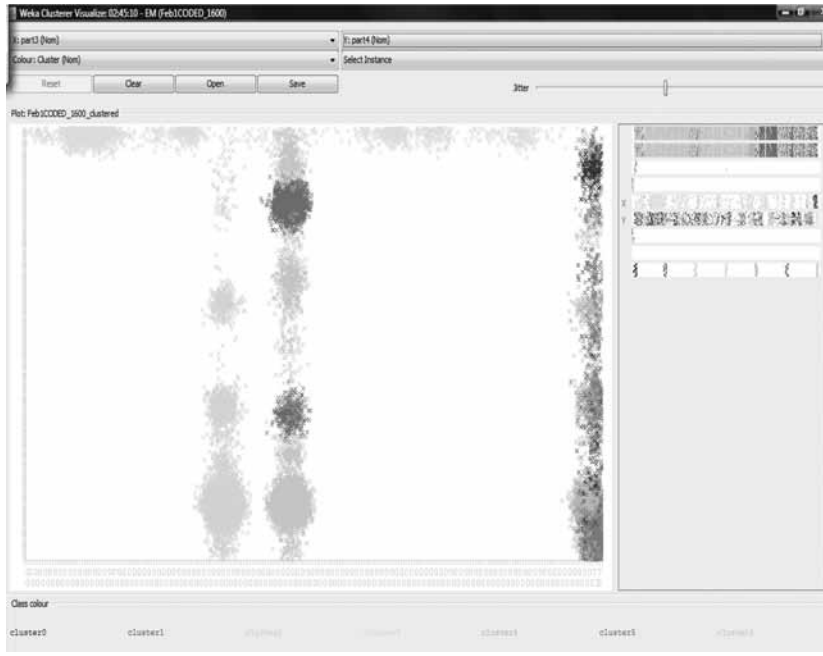


Fig. 8. Assigned Clusters and the Value of part3 and part4 of the Comment Field (2).[4]

represents part3 of the comment, and the *y*-axis represents part4. The large areas of single color demonstrate that transactions from the same clusters have the same values in part3 and part4.

Table 6 shows the number of transactions in each cluster. The majority of transactions (70.20%) are in cluster2 (38.36%) and cluster6 (31.84%). The cluster with the smallest number of transactions is cluster0 (0.99%). The remaining clusters (cluster1, cluster3, cluster4, and cluster5) contain 8.09%, 5.20%, 8.63%, and 6.88% of the transactions, respectively.

This data set consists of transactions entered into transitory account #5738 by over 1,000 branches (AGENCIA) of a major foreign bank. Some branches are bigger and originate more transactions than others do. The number of transactions by branches ranges from 1 to over 1,500 transactions. In terms of the number of transactions originated, the top 20 branches comprise 20% of all transactions in transitory account #5738. In order to gain a better understanding of the data set, the clustering results by branch are examined. The distribution of the clustered transactions in the top 20 branches is examined by branch in Table 7.

[4]The screenshot was captured from the software WEKA (Frank et al., 2016).

Table 6. Number of Transactions by Clusters.

Cluster Name	Number of Transaction	Percentage
cluster0	491	0.99
cluster1	4,010	8.09
cluster2	19,005	38.36
cluster3	2,575	5.20
cluster4	4,277	8.63
cluster5	3,408	6.88
cluster6	15,773	31.84
Total	49,539	100.00

Table 7 shows that different branches have different dominating types of comments. For example, Branches 7227, 1715 and 986 originate many transactions that are in cluster4, whereas Branches 90 and 444 originate many transactions that are either Message 02[5] plus an additional comment, comments with one message, or comments that are impossible to parse. By contrast, Branches 329, 130, 7335, and 811 originate two dominating type of comments or reasons: Message 01[6] plus an additional comment AND transactions with incomplete comments.

When examining transactions from a particular transitory account in a particular branch, there should be only one cluster because a transitory account is supposed to be used for the same purpose at all branches. This suggests a possible explanation for why the observed concentrations exist: different branches use the same transitory account for the different types of pending transactions. For example, Account #5738 might be designated for transactions without account numbers, whereas Account #60836 might be for transactions with insufficient funds). Thus, transactions from clusters other than the one with the highest population may be considered anomalies.

From the data set, 1,016 transactions are generated from Branch 1715. These transactions are grouped into cluster2 (5), cluster4 (1,003), cluster5 (1) and cluster6 (7). More than 98% of these transactions are in cluster4. The remaining five transactions from cluster2, one transaction from cluster5, and seven transactions from cluster6 are in the minority. Using this rationale to evaluate the clustering results for Account #5738, these 13 minority transactions from Branch 1715 can be flagged as possible anomalies. The total number and the percentage of transactions in these groups are so small that they may be illegitimate. The reasons related to these transactions may seem valid, but the small volume suggests doubt about their presence. Further investigation and/or tests are needed to check the legitimacy of these transactions.

The distributions of transactions grouped into clusters are shown in Table 8, which lists each branch and how its transactions are distributed into clusters, and

[5]VALOR LANCTO MAIOR QUE O SALDO.
[6]CONTA INEXISTENTE.

Table 7. Distribution of Transactions into Clusters by the Top 20 Branches.

Branch	cluster0	cluster1	cluster2	cluster3	cluster4	cluster5	cluster6	Total
7227		6	78		1,560	5	38	1,687
1715			5		1,003	1	7	1,016
90	5	5	26	2	1	12	954	1,005
7246		9	865	1		6	23	904
329		4	324	1			539	868
444		6	48	1		1	470	526
927		457	22			1		480
201		377	51	2	16	9	7	462
379		9	393		7	2		411
130			234	3	3	3	142	385
529	4	6	286	1	1	5	58	361
986		5	48		290	7	8	358
7335		12	183	2			111	308
870		3	277	2		6		288
206			254	6		3	12	275
562		3	239				31	273
811		9	74			3	184	270
351		12	219	9		5	4	249
126			4	233		10	1	248
1693		21	185	2		3	35	246

Fig. 9, which presents the same information graphically. Looking at the top 20 branches by the number of transactions indicates that most branches have 1 or 2 dominant types of transactions. For example, 80% of transactions in Branches 927 and 201 are in cluster1 (Message 02); 75% of transactions in Branches 7246, 379, 529, 870, 206, 562, 351, and 1693 are in in cluster2 (Message 01); 93% of transactions in Branch 126 are in cluster3; 80% of transaction in Branches 7227, 1715, and 986 are in cluster4 (Message 50[7]); and 89% of transactions in Branches 90 and 444 are in cluster6 (incomplete comment). No branch in the top 20 group originates the majority of the transactions in cluster0 (Message 70[8]) or cluster5 (Message 71[9] or Message 72[10]). From this knowledge, it is possible to write a

[7]LANC.INVALIDO P/PRODUTO HOT-OVER OU COMPROR.
[8]RESOLUCAO 2025-CLIENTE NAO RECADASTRADO.
[9]LCTO BLOQUEADO P/CONTA NAO HABILITADA.
[10]LCTO BLOQUEADO P/CONTA PARALISADA.

Table 8. Percentage of Transactions Grouped into Clusters for the Top 20 Branches.

Branch	cluster0 (%)	cluster1 (%)	cluster2 (%)	cluster3 (%)	cluster4 (%)	cluster5 (%)	cluster6 (%)	Total (%)
7227		0.3557	4.6236		92.4718	0.2964	2.2525	100
1715			0.4921		98.7205	0.0984	0.6890	100
90	0.4975	0.4975	2.5871	0.1990	0.0995	1.1940	94.9254	100
7246		0.9956	95.6858	0.1106	0.0000	0.6637	2.5442	100
329		0.4608	37.3272	0.1152			62.0968	100
444		1.1407	9.1255	0.1901		0.1901	89.3536	100
927		95.2083	4.5833			0.2083		100
201		81.6017	11.0390	0.4329	3.4632	1.9481	1.5152	100
379		2.1898	95.6204		1.7032	0.4866		100
130			60.7792	0.7792	0.7792	0.7792	36.8831	100
529	1.1080	1.6620	79.2244	0.2770	0.2770	1.3850	16.0665	100
986		1.3966	13.4078		81.0056	1.9553	2.2346	100
7335		3.8961	59.4156	0.6494			36.0390	100
870		1.0417	96.1806	0.6944		2.0833		100
206			92.3636	2.1818		1.0909	4.3636	100
562		1.0989	87.5458				11.3553	100
811		3.3333	27.4074			1.1111	68.1481	100
351		4.8193	87.9518	3.6145		2.0080	1.6064	100
126			1.6129	93.9516		4.0323	0.4032	100
1693		8.5366	75.2033	0.8130		1.2195	14.7776	100

program to detect whether an incoming transaction does not have characteristics similar to the others in its group.

Using simple expressions in Perl programing can change the complex field, which is initially impossible to analyze, into a computable form. After the field is parsed, the patterns and other information can be revealed. Data mining on this data can then be performed.

Cluster analysis can be used effectively for this EDA. At first glance, it might appear as if simple counting can be used for this purpose. For example, transactions with Message 70 can be grouped into cluster0, whereas transactions with Message 50 can be grouped into cluster4. However, cluster analysis is able to group the transactions with similar mistakes into the same groups in a way that a simple counting of frequency distribution cannot accomplish because clustering can incorporate more than one attribute in the consideration of grouping.

For example, cluster analysis actually places transactions with Message 02 into two separate groups. Some transactions with Message 02 are in cluster1, whereas

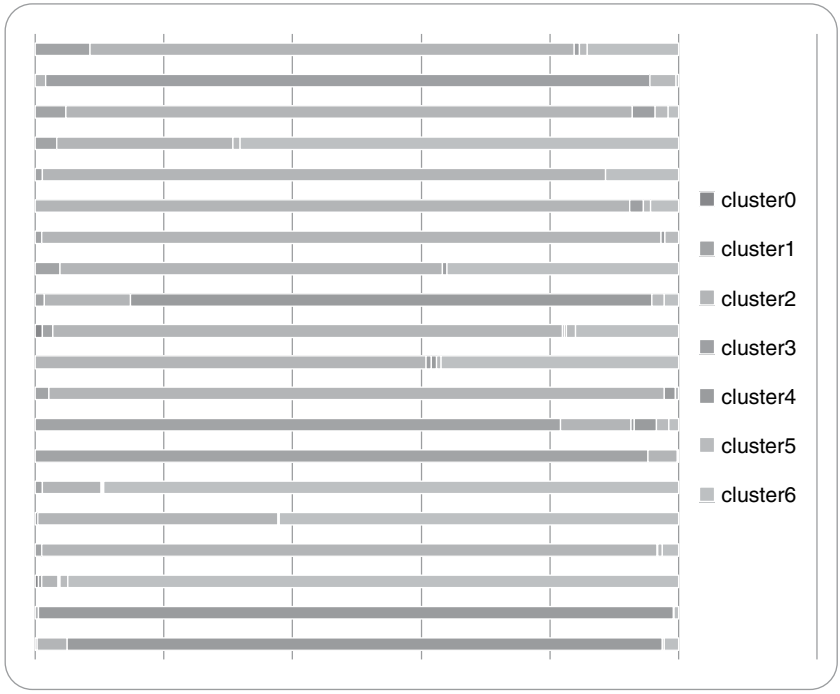


Fig. 9. Distribution of Transactions into Clusters by Top 20 Branches.

others are in cluster6. The difference between the two clusters is that cluster1 represents Message 02 with either Comment 01902[11] or Comment 06246[12] as the second part of the comment. By contrast, Message 02 transactions in cluster6 do not have either of those comments. Moreover, some Message 02 transactions that are grouped into cluster6 represent transactions in which the additional, second phrase information is missing. If the Message 02 transactions have either Comment 01902 or Comment 06246 as the second phrase, they will be grouped into cluster1.

Although the results from clustering using the four original attributes and using the parsed comments are not equivalent, there are some points that should be noted. First, clustering using the four original attributes generates three clusters, two of which are a subset of cluster results from the parsed comment. Transactions that are grouped in cluster1 and cluster2 based on the four original attributes are grouped into cluster6 using the parsed comment field. Second, most transactions (99%) are grouped into cluster0 when the four original attributes are used, despite the fact that these transactions actually differ in their nature. Therefore, the open comment attribute, once parsed, can be a useful variable for conceptual clustering.

[11]*TAXA ADMINISTRATIVA.
[12]*DEB AUT REVISTA SELECOES.

8. Conclusions

Cluster analysis is used extensively in marketing as a technique to discover hidden patterns and structures, such as market segments. The ultimate benefits depend on the goal of the clustering. For example, after identifying market segments, a marketer may use the information to develop marketing strategies to serve some specific market segments or to develop strategies targeted at individual segments. Thus, cluster analysis may be used as a tool to help discover hidden information without the need for prior knowledge about the data set.

Using a real data set from an international bank, this study provides an illustration of how auditors may apply cluster analysis to gain knowledge about a data set. This data set consists of transactions that are posted into transitory accounts because they cannot be completed at the time they are entered into the system. The transactions are posted to transitory accounts temporarily before employees can find the solution to the problem. Once the issue is resolved, the transactions are cleared, leaving an ending balance of zero in the transitory accounts. Little information exists relating to the nature of these transitory accounts. Therefore, cluster analysis is an excellent method for exploring this type of data set. In this study, the EM clustering technique is used, resulting in seven clusters. The resulting clusters may be considered as major sub-groups. A better understanding of the data set may be developed from this analysis, and this knowledge may be useful in the audit planning process.

Cluster analysis is a useful technique for EDA. However, the ultimate benefits of the cluster analysis depend on the objective of the clustering and the way that the resulting clusters are used.

References

Chang, H., Lai, H. H., & Chang, Y. M. (2006). Expression modes used by consumers in conveying desire for product form: A case study of a car. *International Journal of Industrial Ergonomics*, *36*(1), 3–10.

Erdogan, B. Z., Deshpande, S., & Tagg, S. (2007). Clustering medical journal readership among GPs: Implications for media planning. *Journal of Medical Marketing*, *7*(2), 162–168.

Everitt, B. (1980). *Cluster analysis*. New York, NY: Halsted Press.

Frank, E., Hall, M. A., & Witten, I. H. (2016). *The WEKA workbench. Online appendix for "Data mining: Practical machine learning tools and techniques"* (4th ed.). Morgan Kaufmann.

Fisher, D., & Langley, P. (1985). Conceptual clustering and its relation to numerical taxonomy. In *Workshop on artificial intelligence and statistics*, AT&T Bell Laboratories, Princeton, NJ.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1983). *Understanding robust and exploratory data analysis*. New York, NY: Wiley.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research* (4th ed.). Fort Worth, TX: Harcourt College Publishers.

Lim, L., Acito, F., & Rusetski, A. (2006). Development of archetypes of international marketing strategy. *Journal of International Business Studies*, *37*(4), 499–524.

Michalski, R. (1980). Knowledge acquisition through conceptual clustering: A theoretical framework and algorithm for participating data into conjunctive concepts. *International Journal of Policy Analysis and Information Systems*, *4*(3), 219–243.

Morwitz, V. G., & Schmittlein, D. (1992). Using segmentation to improve sales forecasts based on purchase intent: Which "intenders" actually buy? *Journal of Marketing Research*, *29*(4), 391–405.

Shih, Y. Y., & Liu, C.-Y. (2003). A method for customer lifetime value ranking – Combining the analytic hierarchy process and clustering analysis. *Database Marketing and Customer Strategy Management*, *11*(2), 159–172.

Sokal, R. R., & Seneath, P. H. A. (1963). *Principles of numerical taxonomy*. San Francisco, CA: W.H. Freeman.

Srivastava, R. K., Leone, R. P., & Shocker, A. D. (1981). Market structure analysis: Hierarchical clustering of products based on substitution-in-use. *Journal of Marketing 45*(3), 38–48.

Thiprungsri, S., & Vasarhelyi, M. A. (2011). Cluster analysis for anomaly detection in accounting data: An audit approach. *International Journal of Digital Accounting Research*, *11*(17), 69–84.

Tukey, J. W. (1977). *Exploratory data analysis*. Boston, MA: Addison-Wesley Publishing.

Witten, I. H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques* (2nd ed.). Burlington, MA: Morgan Kaufmann Publishers.

Chapter 8

Multi-dimensional Approaches to Anomaly Detection: A Study of Insurance Claims*

Basma Moharram

1. Introduction

The advent of Big Data and the corresponding increases in the affordability and effectiveness of data processing and storage have made automated anomaly detection both more feasible (Chandola, Banerjee, & Kumar, 2009; Patcha & Park, 2007) and more necessary (Chen, Chiang, & Storey, 2012). As more types of data become available in greater volumes and at greater velocity, the use of manual detection methods will become increasingly costly, imprecise, and non-representative, making automated anomaly detection an increasingly attractive prospect. The development of automated anomaly detection methods has been addressed in a variety of literature streams, notably health care (Campbell & Bennett, 2001; Solberg & Lahiti, 2005; Wong, More, Cooper, & Wagner, 2003), intrusion detection (Eskin, 2000; Hu, Liao, & Vemuri, 2003; Lee, Stolfo, & Mok, 2000), and credit card fraud detection (Bolton & Hand, 2002; Ghosh & Reilly, 1994). Recent increases in the prevalence of insurance fraud (Major & Riedinger, 2002) make its analysis particularly fertile ground for quantitative anomaly detection.

The National Healthcare Anti-Fraud Association estimates that 10% of all insurance claims contain some elements of fraud (Major & Riedinger, 2002). Traditional fraud detection involves checking documents manually for abnormalities. Examples of abnormalities would be cross-outs, similar handwriting of provider and patient, or photocopied bills instead of the originals (Major & Riedinger, 2002). This process is time-consuming and difficult to automate. As fraud detection problems involve huge data sets, researchers have been motivated to develop an electronic fraud detection (EFD) system for insurance claims (Hodge, 2001; Major & Riedinger, 2002; Viaene, Dedene, & Derrig, 2005; Viaene, Mercedes, Guillen, Gheel, & Dedene, 2007).

*This chapter is based on the second chapter of the author's dissertation (Moharram, 2016).

Audit Analytics in the Financial Industry, 111–144

doi:10.1108/978-1-78743-085-320191008

This chapter looks into identifying anomalies in life insurance data and focuses on two business cycles; the claim payment cycle and revenue cycle. Before these anomalies can be considered fraudulent, the claims that seem suspicious enough to require further investigation must be identified. Many research studies are concerned with detecting outliers or anomalies (Bakar, Mohemad, Ahmad, & Deris, 2006; Breunig, Kriegel, Ng, & Sander, 2000; Hawkins, Williams, & Baxter, 2002; Knorr, Ng, & Tucakov, 2000; Williams & Baxter, 2002; Yu, Sheikholeslami, & Zhang, 2002). This research stream focuses primarily on how to "calculate" the outliers. Some studies use distance-based outlier calculations (e.g., Knorr et al. 2000). One study (Yu et al., 2002) introduces a new way to calculate outliers called FindOut, which is based on removing clusters from the original data and then identifying the remaining outliers. Another study (Breunig et al., 2000) uses density-based calculations of outliers.

This research focuses on something different. Rather than determining how to calculate the outliers, this chapter explores which attributes should be used to calculate the outliers. The choice of attributes should make sense to the problem in hand, and the problems in this case, is detecting the anomalies in life/disability insurance claims in a way that helps to test two main audit assertions: (1) Is the claim settlement reasonable? and (2) Is the claim itself legitimate?

First, this chapter proposes a continuous auditing framework for the life insurance claim payment cycle. One step in this framework is to detect claim anomalies, so analytical procedures are developed to help the auditor detect claims anomalies in testing the audit assertions for the reasonableness of the claim amount and the legitimacy of the claim itself. To accomplish this, claims data provided by a leading international insurance company are analyzed using a multi-dimensional approach in which the data attributes are divided into different groups (dimensions). These dimensions are assessed logically to find insurance claim anomalies, and then the intersections of the anomalies detected in each dimension are used to find the high priority outliers. As an additional way to prioritize the outliers, the belief function is used to assign a "risk score" to the different branches within the insurance company. This risk score can then be used to prioritize the claims outliers. Finally, a model for detecting premium outliers is presented.

2. Related Literature

2.1. Insurance Outlier Detection

One of the earliest studies on detecting insurance fraud is a paper by Artıś, Ayuso, and Guillen (1999) in which they build a discrete logit model for fraud behavior based on classifying automobile insurance claims data to determine whether the claims are fraudulent. Their model is based on the idea that individuals' behavior seeks to maximize their utility function, which considers the benefit of committing the fraud times the probability of not being detected compared to the cost of punishment times the probability of being detected. In their discrete choice models, Artıś, Ayuso, and Guillen (2002) classify the claims as either legitimate, fraud for the benefit of oneself, or fraud for the benefit of others.

Another study by Major and Riedinger (2002) focuses on comparing the performance of a certain insurance provider with its peers to identify fraud among health providers. They develop an EFD system to be used by the investigative consultant reviewing claims issued by health providers. The EFD provides a "multiple frontier summary" report for the consultant. This report lists any unusual providers identified by the system. For each provider, the investigative consultant can call up a frontier summary report, with the details of the provider and the behavioral patterns for which this provider is identified as unusual.

This EFD architecture consists of five layers. The first layer is the behavioral heuristic measurement with twenty-seven heuristics in five categories. The second layer is the information, frontier, and rules layer, which compares measurements among the providers and flags those that are out of line relative to their peer group. The third layer is the data exploration layer in which the EFD can supply the extracted claim records to the consultant. The fourth layer is the decision and action layer in which, if the consultant decides to investigate a case further, the system would issue a memo translating the frontier summary report into business-oriented terms along with the consultant's recommendations and send it to the appropriate regional investigator. The fifth and final layer is the enhancement layer. This chapter, will use the Major and Riedinger's (2002) work to develop one of the claims anomalies detection dimensions in this study.

Another research study that focuses on detecting fraud in insurance claims is by Pathak, Vidyarthi, and Summers (2005), who create a fuzzy expert system to detect anomalies in insurance claims that have already been settled. Their model is based on three measures: an ambiguity index, the degree of incomplete information in the claim, and the level of discretion used by the claim settlers. For each measure, they set a low, mid, or high level based on the data itself and on the auditor's judgment. Then they set rules for "If … THEN" statements to categorize each claim as either genuine or not. The authors do not define any variables driving these measures. Nevertheless, their three measures are used in this chapter to evaluate different dimensions for detecting anomalies in claims.

Yamanishi, Takeuchi, Williams, and Milne (2004) have an interesting way to detect outliers. They introduced a theoretical framework for an online, unsupervised outlier detector called "SmartSifter". Every time a new instance is input, the online process requires the system to evaluate its deviation from the normal pattern. In addition, every time an instance is input, the system employs an online learning algorithm to update the model. The authors contrast their concept of online outlier detection to the traditional batch detection process in which outliers can only be detected after seeing the entire database. They also develop two different algorithms, Sequentially Discounting Laplace Estimation, which learns the histogram density for the categorical domain, and Sequentially Discounting Expectation and Maximizing, which learns the finite mixture model for the continuous domain. One important aspect of those two algorithms is that they gradually discount the effect of past examples in the online process. SmartSifter assigns a score to each piece of input data based on the learned model, and measures the change to the model after learning. A high score indicates a high possibility that a particular piece of data is an outlier. The authors test their new

system using simulated data and conclude that it works better than other systems in terms of accuracy and computation time.

2.2. Belief Function

Shafer (1996) states that "the theory of belief functions provides a non-Bayesian way of using mathematical probability to quantify subjective judgments." The basic difference between probability theory and a belief function is in the assignment of uncertainties to a set of mutually exclusive and exhaustive states, referred to as a "frame" (Srivastava, 1993). The belief function bases degrees of belief or trust for one specific question on the probabilities for related questions. The belief function measures the evidence that supports a specific event, the ambiguity that represents the part for which there is no evidence to support any outcome, and the plausibility of the event that combines the evidence and the ambiguity of the event.

Three basic functions are important to understand the use of belief functions in decision-making (Srivastava & Mock, 2000): basic belief mass functions, belief functions, and plausibility functions.

*Basic belief mass function (*M-*values):* For a decision problem with n possible elements forming a mutually exclusive and exhaustive set represented by $\{a_1, a_2, \ldots, a_n\}$, this set is called a frame, and it is represent by the symbol Θ. The *m*-values are defined as the level of support directly obtained from the evidence to support a specific element. These m-values can be assigned to all single elements, to all subsets that can be derived from the frame, and also to the entire frame itself. All *m*-values must add to one.

Belief function: Belief on a set of elements (A) of frame Θ is defined as the total belief on (A). This represents the sum of all the *m*-values assigned to the elements contained in (A) plus *m*-values assigned to (A).

$$\mathrm{Bel(A)}=\sum_{\mathrm{B}\subseteq\mathrm{A}}\mathrm{m(B)}$$

where B is any subset that belongs to A.

Plausibility function: The plausibility of an element or a set of elements (A) of frame Θ is defined to be the maximum possible belief that could be assigned to (A) if all future evidence is in support of (A). In other words, it could be defined as 1−*m*-values assigned to all other elements or set of elements not A ($\sim A$).

Ambiguity function: The ambiguity in an element (A) is defined as the difference between the plausibility of this element and the belief in it.

Belief functions have a number of features that argue for their more extensive use in auditing (Srivastava & Mock, 2005). One of these features is the "rigorous definition of risk" in comparison to probability functions. In case of complete ignorance when an auditor does not have any evidence on whether there is management fraud in the financial statements, the probability functions will assign both events (fraud and no fraud) a probability of 0.5 to represent uncertainty. By contrast, the belief functions will more clearly assign a belief of zero to each

event, and the plausibility of each event will be assigned a value of one. The difference between the plausibility of each event and its belief (1−0 = 1) is a rigorous measure of the "ambiguity" of the event, which the probability functions cannot represent. In a different situation, where the auditor has partial knowledge, the authors give an example of an auditor who has 0.3 evidence that there is fraud in the financial statements. The probability functions represent this as P(fraud) = 0.3 and P(no-Fraud) = 1−0.3 = 0.7 even though there is no evidence about the no-fraud event. The belief function, on the other hand, represents the risk more realistic way by saying that Belief(fraud) = 0.3, Belief(no-Fraud) = 0, and the remaining 0.7 will be undecided. That makes the plausibility of Fraud = 1 in the belief function, whereas probability of fraud is 0.3 in the probability function (Srivastava & Mock, 2005). Researchers have argued in favor of the belief function because it gives a more plausible and conservative valuation of risk that is more suitable for auditing (Shafer & Srivastava, 1990; Srivastava & Mock, 2005). Research studies have explored the applicability of belief function in auditing and assurance services (Mock, Sun, Srivastava, & Vasarhelyi, 2009; Shafer & Srivastava, 1990; Srivastava, 1993; Srivastava & Mock, 2000, 2005, 2011; Srivastava & Shafer, 1992; Srivastava, Mock, & Turner, 2007; Srivastava, Rao, & Mock 2013; Sun, Srivastava, & Mock, 2006). The belief function has also been used in other fields, such as financial portfolio management, data mining, image processing, agriculture, water treatment, and forecasting demand (Srivastava & Mock, 2005).

3. Data

This chapter uses data from a leading international insurance company that deals with many forms of insurance. For this study, only life/disability insurance claims are used.

The database contains 2,763,591 unique policies, with effective dates ranging from November 1998 through January 2014. For this study, there are 28,490 records of life/disability insurance claims paid in the period between January 2013 and July 2014. Of these records, 82% (23,392 records) relate to individual life/disability insurance, and 18% (5,098 records) relate to group insurance. In the individual insurance setting, the client is an individual person who bought an insurance policy against his/her life or, in some cases, against someone else's life. On the other hand, a group insurance policy is bought by an organization against the life or disability of its members or employees. When dealing with individual policies, the insurance company has direct contact with the person who is being issued the policy. In this case, the insurance company can collect all of the personal information needed to make the decision whether to issue the policy. For group policies, the insurance company has no direct contact with the organization's employees. In fact, the insurance company in this study does not record any information regarding the client organization's employees until a claim is actually filed.

The claims data include information about the insured, the coverage, the beneficiaries, and the payment. Only 33% (7,746) of the individual claims were successfully joined with their corresponding policies. Fig. 1 summarizes the data.

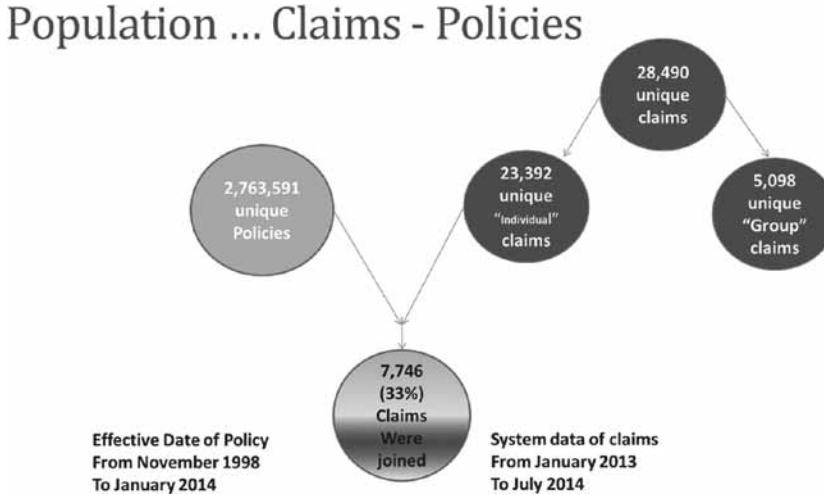


Fig. 1. Population.

4. Continuous Auditing Framework for Life Insurance

The continued use of traditional auditing practices in modern accounting information systems prevents the economy from exploiting the full potential of these systems. Chan and Vasarhelyi (2011) describe seven dimensions that distinguish continuous auditing from traditional auditing: (1) more frequent or continuous audits; (2) a proactive audit model; (3) automated audit procedures; (4) evaluation of the work and role of auditors; (5) change in the nature, timing, and extent of auditing; (6) use of data modeling and analytics for monitoring and testing; and (7) change in nature and timing of audit reports. In their paper, Chan and Vasarhelyi (2011) also describe the stages of applying continuous auditing. The first stage is to identify business process for which the audit can be continuous in terms of the availability and accessibility of the data, and then to identify the audit procedures and the type of monitoring and testing that can be automated. The second stage is to develop benchmarks for evaluating future data based on historical data already available through modeling, estimation, classification, association, and/or clustering.

In this chapter, the continuous auditing view is applied to the life/disability insurance industry. As recommended by Chan and Vasarhelyi (2011), the procedure starts with identifying a specific business process that can be automated, in this case, the Claim Payment process. Next, the set of controls and tests that can be automated within this business process are identified. Fig. 2 shows the proposed continuous auditing model for the claim initiation step.

As Fig. 2 shows, a claim is first submitted to the company in the claim initiation step. Before the company starts to process the claim, a few tests should be performed through a simple check with the policies that the company is maintaining. The first test is to determine whether this claim is filed against a valid policy that is maintained by the company. The second test checks whether the

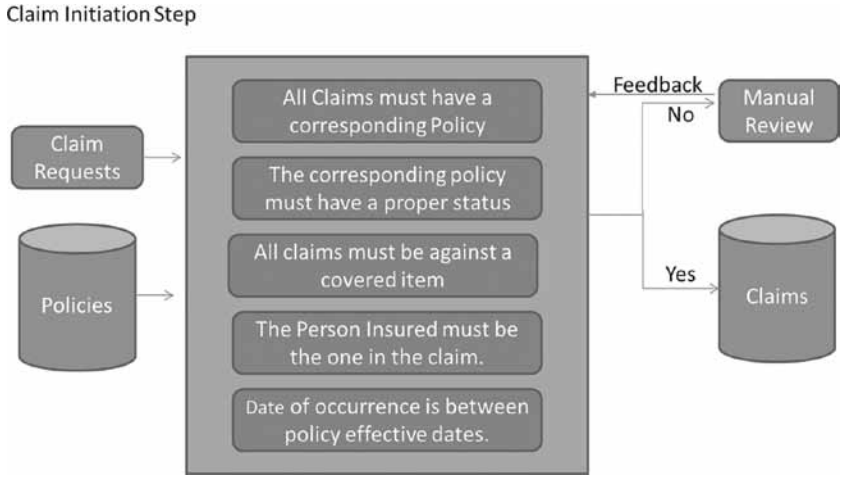


Fig. 2. Claim Initiation.

corresponding policy has a proper status. For example, a policy might be canceled or the company might have deactivated it because the client is not paying the premium. In these cases, the company might not be liable to pay the claim. The third test is to check whether the item for which the claim is filed is covered under the corresponding policy. For example, if the claim is filed to cover funeral expenses, the third test will check whether funeral expenses are covered under the corresponding policy. The fourth test checks whether the person in the claim is the same person insured under the corresponding policy. The last test in the claim initiation process checks whether the date of occurrence of the event causing the claim lies between the beginning and ending effective dates of the policy. If all the five tests come back positive, then the claim can be initiated and entered into the claims database to start the approval process. If any of the tests comes back negative, then the claim should be flagged for manual review. The results of the manual process are then fed back into the system as feedback. The claim initiation step in the claim payment process is a perfect example of continuous auditing as a part of a business process that can be automated in a simple way.

Fig. 3 shows the proposed continuous auditing model for the claims payment cycle for the next step, claim validation.

Now that it has been determined that the claim is filed against a valid policy, the next step is to assess whether the claim itself is legitimate. The first test is simple. No client can die twice. So previous claims filed against the same policy or another policy for the same insured person are checked for inconsistencies that might include two deaths and/or two funerals for the same client, or excessively frequent disabilities. A manual check will be needed in these cases as some of them might still be legitimate, such as the case of frequent disabilities. The next test checks for the relationship between the reason for the claim and the item against which the claim is filed. A claim might have one of many reasons like death of the insured person, hospitalization of the insured person, an accident

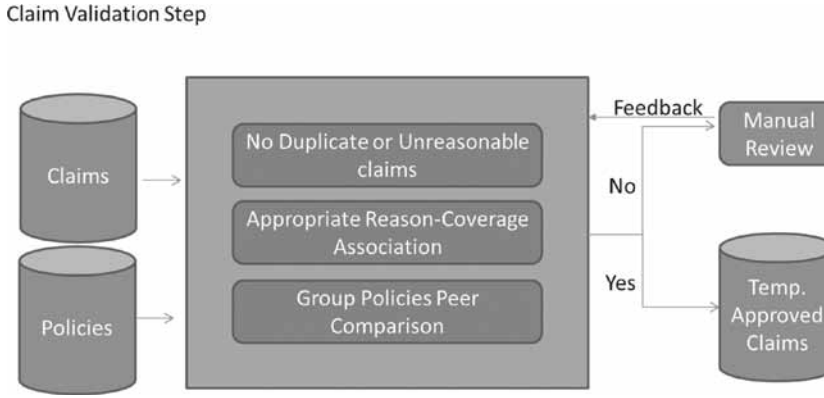


Fig. 3. Claim Validation.

causing disability, or something else. When a claim is made, it must be filed against a specific item insured under the policy. This item could be funeral and/or grave expenses, hospital expenses, medical assistance, or something else. This test uses the historical data to learn the possible association between the reason for the claim and the item against which the claim is filed. This can be used to anticipate whether a certain claim needs further manual investigation. For example, if the reason for the claim is illness, it would be reasonable to be filed against hospital expenses, but unreasonable to be filed against funeral expenses. This test is discussed in detail later in this chapter. The next test focuses only on group policies that are bought by a company or an institution to insure its employees. It is called a group policy because it is one policy that covers a group of insured personnel. In some cases, the insurance company does not collect any information on the insured personnel unless something happens to one of them, which makes the control and auditing of these claims more difficult than for individual claims. The group peer comparison test clusters the insured companies into identical groups in terms of the types of policies they have and the items insured under these policies. The test compares the filing pattern of an insured company with the filing pattern of its peer group. As with individual claims, if any of the tests comes back negative, the claim will be reviewed manually, and the feedback will be used to update the continuous auditing system. Otherwise, the claim will be approved and stored in a temporary approved claim data set while waiting for valuation and payment.

The final step in this continuous auditing vision for life/disability insurance is the claim valuation and payment shown in Fig. 4.

This process starts with the temporary approved claims data set, which contains the claims that passed the previous two steps. Now that the claim has been validated, it is time to determine how much the insurance company should pay. With a life insurance policy, the client or the insured person is entitled to part of the face value of the policy in the case of disability. This portion of the face value is determined based on the type and severity of the disability. The insurance company has a list of possible disabilities and the percentage of face value associated

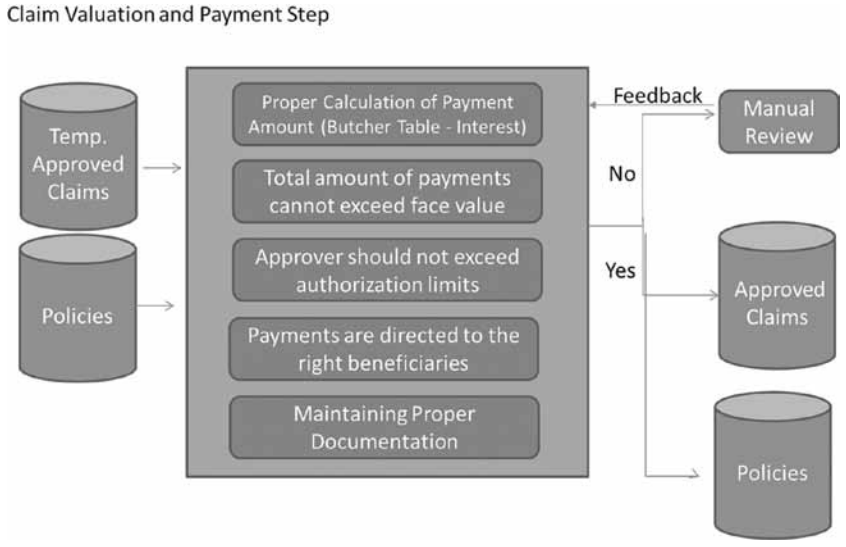


Fig. 4. Claim Valuation and Payment.

with each of them. Insurance companies call this list "the butcher table". Another factor that affects the amount that the insurance company must pay is the interest. While the calculations of interest may differ from one company to another, the rule is that the insurance company has to pay interest if the claim payment is delayed beyond a specific grace period.

The first test in this step is to make sure that the amount of the claim approved is the correct amount that the insurance company must pay. The test is a simple recalculation of the amount of the claim based on the "butcher table" and the company's interest calculation policy. The second test searches for all previous claims paid by the insurance company against the policy of the current claim. This test adds all previous claim amounts and compares the total to the face value of the corresponding policy. If the outstanding balance is sufficient to pay the current claim, then the claim moves on to the next test. If not, the claim is flagged for manual review. The next test is to verify that the approver of this claim has not exceeded his or her authorization limit for the current period. This test searches for all amounts approved by the current claim's approver during the current period, which could vary depending on the company's policy. Then, the test adds these amounts and compares the total to the approver's authorization limit based on his/her position within the insurance company. The next test concerns the fact that the claim could be payable to one or more beneficiaries. These beneficiaries must be identified in advance in the policy contract. The test compares the original beneficiaries and their bank accounts with the person(s) to whom the insurance company is directing the claim payment. The last test checks for proper documentation. The insurance company should maintain an updated list of its policies. That means updating the outstanding balance of the face values after paying claims, changing the status of the policy after cancelation, expiration, or

a death claim. As always, if any of the tests come back with abnormalities, the claim will be reviewed manually, and then the feedback will update the continuous auditing system. Otherwise, the claim will be approved and stored in the approved claims data set while waiting to be paid. The corresponding policies will also be updated in the policies table. As part of the feedback, this study looks at all true positives and checks which assertions are the most important. Then true positives or ideal positives for each assertion are constructed. False positives are also identified to set a benchmark for comparison with exceptions.

5. Claims Anomalies Detection

This section expands on the analytical procedures discussed in the claims validation step of the framework. The analytical procedures are designed to help the auditor to detect claims anomalies when testing two main audit assertions: (1) Is the claim settlement reasonable? and (2) Is the claim itself legitimate? To do this, claims data provided by a leading international insurance company are used in a multi-dimensional approach that divides the attributes into different groups or dimensions. Each dimension is used logically to find insurance claim anomalies, and then the intersections of the anomalies detected in each dimension are used to highlight the high priority outliers.

5.1. Is the Claim Settlement Reasonable?

The total payment to the client or beneficiaries is used to assess the reasonableness of the settlement amount. According to the policy of this particular insurance company, if the claim is not paid within 30 days after the claim is filed, an interest amount is calculated and added to the settlement. For individual life insurance claims, the total payments to the beneficiaries should be equal to the face value of the policy plus any interest earned beyond the allowed 30-day period. Detecting anomalies in this case involves testing the relationship between the interest payments and the period between filing the claim and paying the settlement. From the available data, the attributes needed include the lump sum of total payments to the beneficiaries, the date of filing the claim, and the date of payment. These attributes can be used to compute Days before Payment, which represents the number of days between filing the claim and paying the settlement beyond the allowed 30-day period, and Estimated Daily Interest, which is the positive difference between the amount actually paid and the face value of the policy divided by the Days before Payment. Fig. 5 shows the distribution of the estimated daily interest.

Fig. 5 shows that there are seven obvious anomalies above the line with more than 1% daily interest after the 30-day period.

Two more dimensions are also added to the analysis: (1) the relationship between Estimated Daily Interest and the reason for the claim (e.g., death of the insured person); and (2) the relationship between Estimated Daily Interest and the type of coverage (e.g., paying for funeral expenses).

One major limitation of this analysis is that it can only be applied to individual life insurance claims because the insurance company pays the full face value of

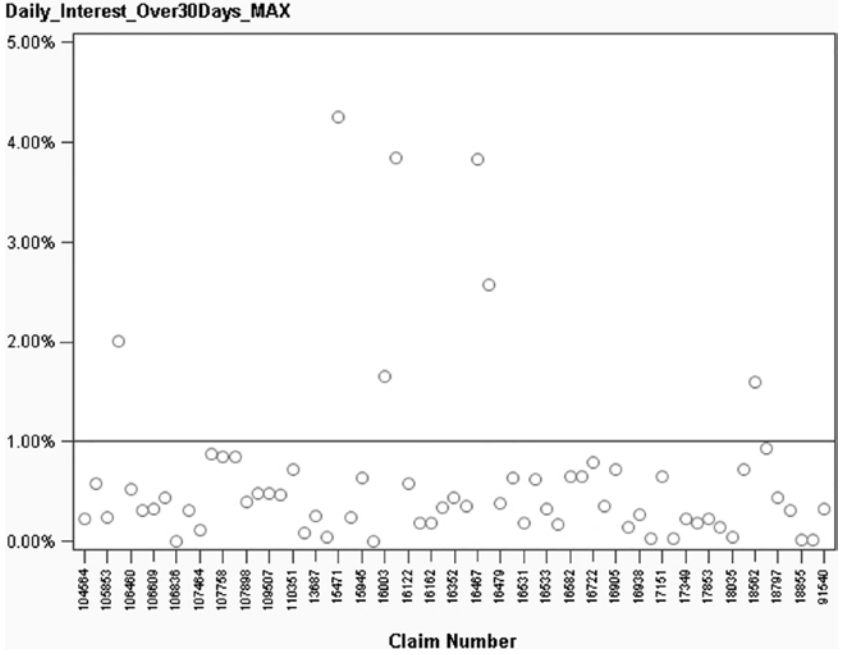


Fig. 5. Estimated Daily Interest (After the 30-Day Period).

the policy upon the death of the insured person in an individual claim. For group life insurance policies, the insurance company pays a percentage of the face value of the entire group policy as a lump sum based on some predetermined criteria (e.g., the deceased employee's salary), so the only data available are the total payment for each claim. Due to the lump sum payments and the nature of the group life insurance settlement calculations, there is no precise way to calculate the estimated interest amount for group life insurance policies.

This analysis also cannot be used for disability insurance claims because only a portion of the face value of the policy is paid based on the type of the disability. The insurance company provides percentages for various disabilities, which they call the "butcher table". However, a link between these percentages and the actual claims data cannot be established using the available data. So again, there is no way to calculate the estimated interest amount.

5.2. Is the Claim Itself Legitimate?

For the second main audit assertion, the legitimacy of the claim, anomalies are checked using three dimensions: (1) reason-coverage association; (2) group similarities; and (3) timeline.

5.2.1. Reason-Coverage Association. When people buy insurance policies from this insurance company, they get to choose among different types of policies. Each product provides insurance against a specified set of events, and a unique coverage code is used to identify each insured event, such as death, funeral

expenses, headstone cost, or disability. A policyholder or beneficiary who files a claim must specify the reason for filing the claim and the coverage code for that claim. For example, if a beneficiary is filing for funeral expenses, the claim reason would be the death of insured person and the claim coverage code would be the code for funeral expenses. If the beneficiary wants to file for headstone cost, there would be a separate claim in which the claim reason would be the death of the insured person and the claim coverage code would be the code for funeral expenses.

The data set for this study has 9 different claim reasons for filing and 35 different coverage codes. For the purpose of this analysis, the coverage code for each claim is matched with the associated claim reason. For each claim reason, a matrix is constructed. Each matrix lists the claim numbers by rows and the coverage codes by columns. Fig. 6 shows the matrix for Reason 1. For each claim, a "1" indicates that this claim was filed against that specific type of coverage (COV). For example, Claim 1234 was filed for Reason 1 and was filed against COV 1, COV 2, and COV 4, but not COV 3. When all claims filed for Reason 1 are added to the matrix, the results are examined to determine whether there are different sets of coverage codes for that claim reason. Fig. 7 shows that three claims were filed for Reason 1, but they were not all filed against the same coverage items. Specifically, Claim 1234 and Claim 1236 were filed against COV 1, COV 2, and COV 4, whereas Claim 2234 was filed against all four coverage items. Thus, there are two different sets of coverage codes for Reason 1: Coverage Set "A" {COV 1, COV 2, COV 4} and Coverage Set "B' {COV 1, COV 2, COV 3, COV 4}. Next, the filing rate against each coverage set is calculated. In the example shown in Fig. 7, claims against Coverage Set "A" were filed twice out of a total of three claims, so its filing rate is 66.7%, whereas only one claim was filed against Coverage Set "B" out of a total of three claims, so its filing rate is 33.3%.

In this reason–coverage association, the lower the filing rate against a specific coverage set compared to other coverage sets, the more suspicious the claim is. Fig. 7 shows an example from the data set and indicates the distribution of the 17 different coverage sets found for the reason "Death by Disease". In this case, the anomalies would be the least common sets (less than 5% filing rate). These policies would require additional investigation.

5.2.2. Group Similarities. The second dimension in testing the legitimacy of the claim is group similarities. Here, "group" refers to group insurance policies in which the client is a corporation insuring against the death or disability of its employees. In this dimension, two group policies are identified as being similar only if both policies insure against the same coverage set. Once the similar policies are identified, they are put into groups. Then, the claim filing pattern of each

Policy Number	Cov 1	Cov 2	Cov 3	Cov 4
1234	1	1	0	1
1236	1	1	0	1
2234	1	1	1	1

Fig. 6. Reason–Coverage Association – Example.

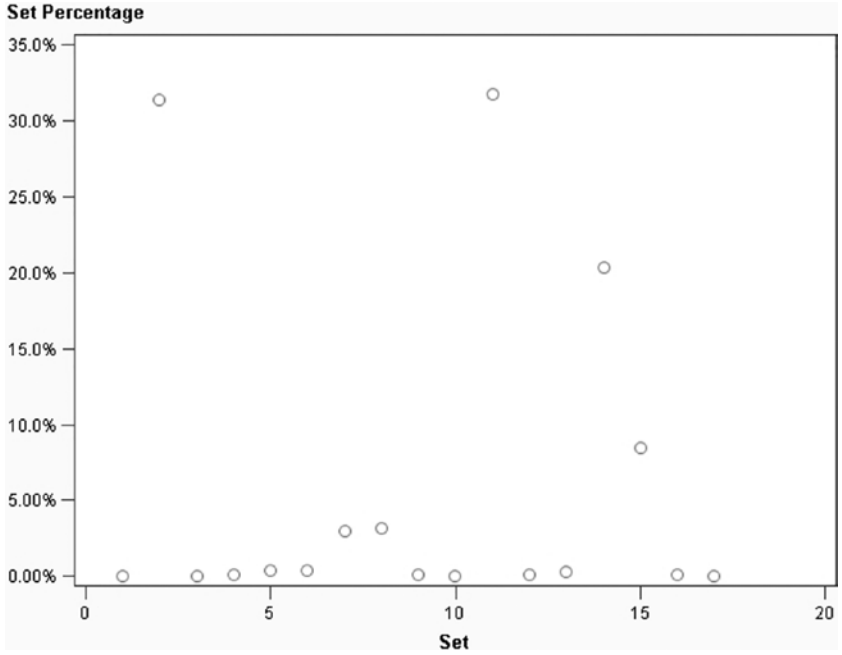


Fig. 7. Reason–Coverage Association – Filing Rate.

policy in a group is tested against its group's average. For example, Fig. 8 shows a sample list of group policies and the coverage sets for which they are insured. The first three policies all insure against a particular coverage set {COV 1, COV 3}, so they are considered to be members of one group. The last two policies, shown in green, insure against another coverage set {COV 2, COV 3}, so they are considered to be members of a separate group.

After defining each group based on the similarity of their coverage choices, the claim filing pattern of each policy within a given group is studied in terms of the most frequent coverage codes for its filings. Then, the filing pattern for each policy is compared to the average filing pattern for its group. Fig. 9 shows the filing patters for the first three policies in Fig. 8.

Fig. 9 shows that Policy 1234 files 60% of its claims against COV 1 and 40% against COV 3, and Policy 1235 has similar behavior with 65% of its claims filed against COV 1 and 35% against COV 3. By contrast, Policy 1236 filed only 10% of its claims against COV 1 and 90% against COV 3. Compared with the rest of this group, Policy 1236 would be considered an anomaly that could require further investigation.

5.2.3. Claim Timeline. The third dimension used to detect anomalies in terms of the legitimacy of the claim itself is the claim timeline, shown in Fig. 10.

The claim timeline shows the five key dates in the data set: (1) the date when the policy was issued; (2) the date when the event occurred; (3) the date when the insurance company first received "notice" of the event (i.e., when the insured

Policy Number	Cov 1	Cov 2	Cov 3
1234	1	0	1
1235	1	0	1
1236	1	0	1
2134	0	1	1
2135	0	1	1

Fig. 8. Group Similarities – Identifying Groups.

Policy Number	Cov 1	Cov 2	Cov 3
1234	60%	0	40%
1235	65%	0	35%
1236	10%	0	90%

Fig. 9. Group Similarity – Filing Pattern.

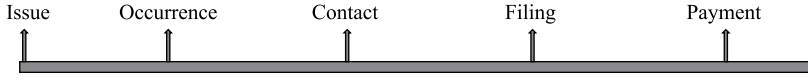


Fig. 10. Claim Timeline.

person or beneficiaries first inform the insurance company about the claim); (4) the date when the claim was filed; and (5) the date when the claim was paid. The significant issue here is anomalies in this timeline, such as an excessive time between the occurrence of the event and the insurance company being informed of the event.

In this timeline, key considerations include: (1) number of days from the effective date of the policy to occurrence; (2) number of days from occurrence to notice; (3) number of days from notice to filing; and (4) number of days from filing to payment.

The anomalies of interest would be an exceptionally short time between issuance and occurrence or an exceptionally long time between occurrence and notice, notice and filing, or filing and payment.

Fig. 11 shows the box-and-whisker plot for Days to Occurrence, which is the number of days between the effective date of the policy and the occurrence of the reason for the claim.

Fig. 11 shows the distribution of Days to Occurrence, which ranges from zero to 9,410 days with a median of 59 days. The lower whisker is on 0 days and the upper whisker is on 300 days, which indicates that most of the claims are filed against incidents that happen within a year after the policy was issued. The upper outliers in Fig. 12 range from 300 days up to 4,995 days. These outliers are actually profitable for the insurance company because it is able to collect premiums for a longer period before it has to pay a claim, so they do not represent potential

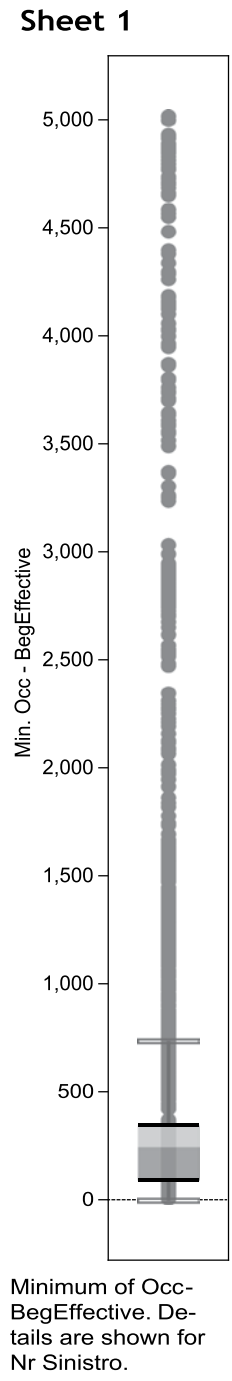


Fig. 11. Box-and-Whisker Plot for Days to Occurrence.

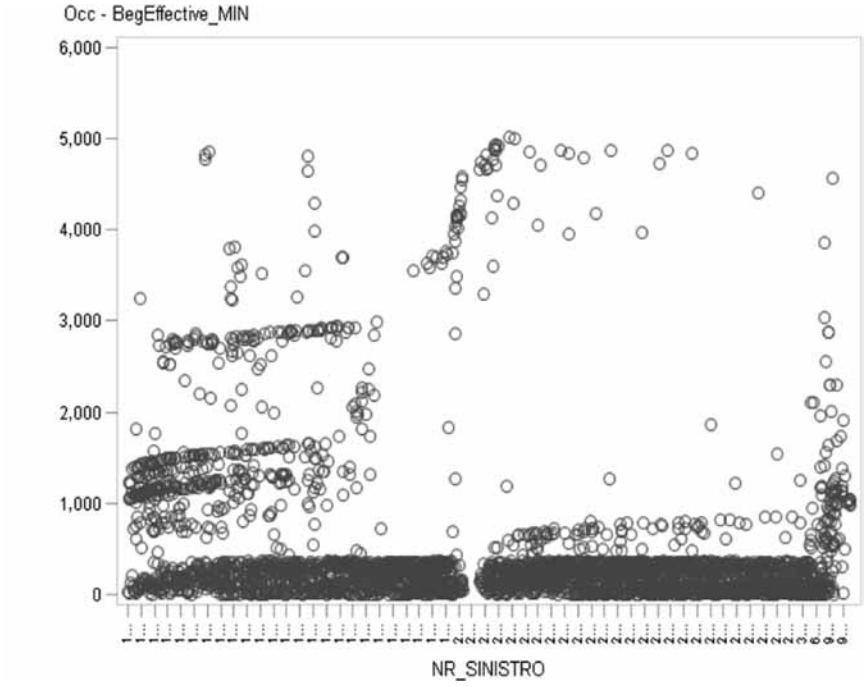


Fig. 12. Distribution of Days to Occurrence.

problems. The main concern is with claims that are filed within a short time after the policy's effective date.

Fig. 12 is a different representation of the same data. The horizontal axis shows the claims, and the vertical access shows the Days to Occurrence. The graph shows that most of the claims are filed on incidents that happen within the first year of purchasing the insurance policies.

To prioritize the claims anomalies based on Days to Occurrence, the percentile distribution of Days to Occurrence are shown in Table 1.

As seen in Table 1, the percentiles suggest that the claims with Days to Occurrence of two days or less should be assigned the highest priority for examination as possible anomalies, followed by claims between two and eleven Days to Occurrence, and so on.

Fig. 13 shows the distribution of Days from Occurrence to Notice.

Fig. 13 shows that most of the data lies below the 500-day line, although some of these claims have Days from Occurrence to Notice that exceed 5,000 days.

Table 1. Days to Occurrence Percentile.

Obs	P_1	P_5	P_10	P_25	P_50	P_75	P_90	P 95	P_99	P_100
1	2	11	26	85	252	348.5	1,208	1,636	4,160	5,017

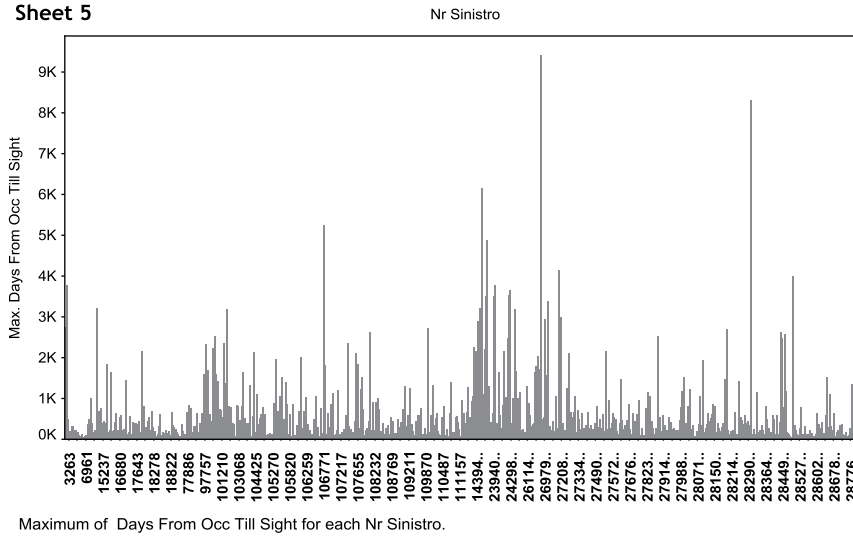


Fig. 13. Days from Occurrence to Notice – Distribution.

Fig. 14 shows the box-and-whisker plot for Days from Occurrence to Notice, which ranges from zero to 9,410 days with a median of 59 days. The lower whisker is 0 and the upper whisker is 300 days. This indicates that the normal situation is for the policyholder or beneficiaries to contact the insurance company within the first year after the occurrence of the event. For claims with more than a year between the occurrence of the event and contact with the insurance company, the auditor should investigate whether the claim is legitimate.

The concern with the Days to Occurrence in the previous section is the lower outliers, whereas the concern in this section for Days from Occurrence to Notice is the upper outliers. Why would a client wait for years before notifying the insurance company about an event that is covered under an insurance policy? Table 2 shows the percentile distribution of Days from Occurrence to Notice.

Table 2 suggests that any claim above the 90th percentile (338 days) should be considered an anomaly. The higher priority should be given to claims with more than 2,010 Days from Occurrence to Notice, followed by claims with 630 to 2,010 days, followed by claims with 338 to 630 days.

6. Risk Scoring

This section discusses a methodology that can be used to assign risk scores to specific unit (e.g., a particular authorizer of claims, branch of the insurance company, or geographic region) based on the evidence collected during the audit. The scoring methodology is based on the belief function and evidential reasoning (Shafer & Srivastava, 1990; Srivastava & Mock, 2000, 2005; Srivastava & Shafer, 1992). The resulting risk scores can then be used to prioritize the claims outliers.

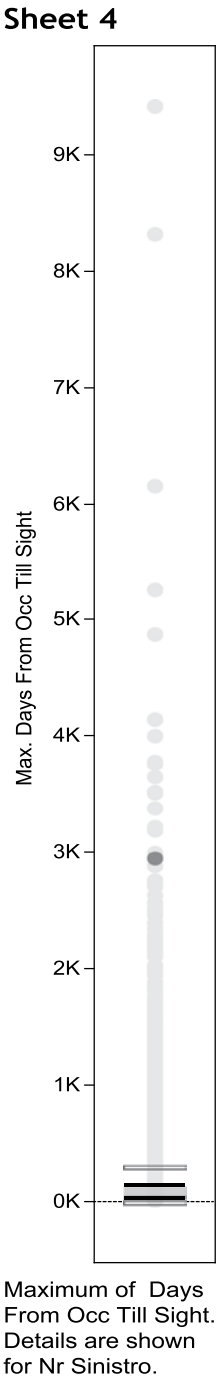


Fig. 14. Box–Whisker Days from Occurrence to Notice.

Table 2. Days from Occurrence to Notice – Percentiles.

Obs	P_1	P_5	PJO	P_25	P_50	P_75	P_90	P_95	P_99	P_100
1	2	10	16	28	56	128	338	630	2,010	9,410

The audit assertions in this study concerning whether the claim settlement is reasonable and whether the claim itself is legitimate have binary results: either pass (a) or fail ($\sim a$). Furthermore, there is no ambiguity in the evidence collected as it is designed to test these assertions for each claim transaction. However, there can be a problem with missing values. Some claim transactions cannot be tested due to missing values that are essential to run the test. For example, to check whether the total amount of a claim does not exceed the face value of the policy requires a link between the claim and the policy. If this link is broken because of a missing value (e.g., a missing the policy number in either the claims or the policy data sets), the test cannot run for that specific transaction. Similarly, when checking whether the amount of the claim was paid to the right beneficiaries, the test cannot be run if the beneficiary information is missing from either data set.

In the case of missing data, the question is how a unit of interest should be treated since it did not passed the test, but it did not fail either. This study proposes that belief functions, with their inherent ability to incorporate ambiguity, may be able to describe these cases of missing evidence. Based on Dempster's rule of combination and evidential reasoning, which combines available evidence from different sources to create a belief function (Dempster, 1967), evidential diagrams can be built that will summarize the results of the tests of the assertions and give an aggregate risk score for the unit in question.

The following example uses data for two different branches of the insurance company to illustrate how to assign a risk score to each of these branches. Table 3 shows the distribution of the claims between the two branches:

To determine whether claims are legitimate, one test is whether all claims have corresponding policies. The results for the two sample branches are shown in Table 4:

Table 3. Each Branch's Share of Total Claims for Both Branches.

Branch	Percentage of Total Records
A	64.78
B	35.22

Table 4. Test Results – All Claims Have Corresponding Policies.

Branch	Pass (%)	Fail (%)	Unknown (%)
Branch A	86	14	0
Branch B	96	4	0

Table 5. Test Results – All Insured Events Occurred while the Policies Were in Effect.

Branch	Pass (%)	Fail (%)	Unknown (%)
A	75	12	13
B	73	23	4

The results show that 86% of Branch A's claims had a corresponding policy on file and 14% did not, whereas 96% of Branch B's claims had a corresponding policy on file and 4% did not. There is no ambiguity in the results of this test because the answer is known for all claims.

Another test of claim legitimacy is whether the insured events occurred while the policies were in effect. Table 5 shows the results of this test for the two branches:

The results show that 75% of Branch A's claims passed the test and 12% failed, but 13% were unknown because of missing values. For Branch B, 73% of the claims passed the test, 23% did not, and 4% were unknown because of missing values. Applying the formula for the belief function to Branch A yields the following results:

$$\begin{aligned}
&\text{Branch A_Test1: } m1(\{\text{a}\}) = 0.86,\ m1(\{\sim\text{a}\}) = 0.14,\ m1(\{a, \sim a\}) = 0.0\\
&\text{Branch A_Test2: } m2(\{a\}) = 0.75,\ m2(\{\sim a\}) = 0.12,\ m2(\{a, \sim a\}) = 0.13\\
&\text{Conflict} = [m1(a).m2(\sim a)] + [m1(\sim a).m2(a)] = [0.86 \times 0.12] +\\
&\qquad\qquad [0.14 \times 0.75] = 0.01\\
&K = 1 - \text{Conflict} = 1 - 0.01 = 0.99.
\end{aligned}$$

7. Premium Outliers Detection

An important task in auditing the life/disability insurance revenue cycle is to check the valuation of the premiums. The obvious way to do this check is to recalculate the premium based on the data available for the corresponding policy. However, to do this, the factors incorporated into the premium calculations must be identified.

7.1. Factors Affecting Premium Calculations

The premium calculations for a particular policy depends on four factors: (1) the type of policy; (2) the risk factors of the insured person; (3) the importance of the insured person to the insurance company (VIP status); and (4) the desired profit margin of the insurance company.

7.1.1. Type of the Policy. The type of the policy is the way that an insurance company defines the policy. The first aspect in defining the policy is to determine the type of insurance that this policy offers. An insurance company may provide different types of insurance, including life insurance, auto insurance, rental insurance, and/or health insurance. The second aspect is the type of product offered under this policy. For example, an insurance company might offer different products under life insurance, such a product specifically designed for women, a product specifically designed for families, and so on. The third aspect would

be the list of items that are covered under each policy. A particular life insurance policy might cover hospital expenses, temporary income, death, and funeral expenses, or it might only cover hospital expenses and death. The premiums for these two policies should be different because they cover different items, even though both policies are life insurance and both may be same product (e.g., life insurance for families). The last aspect concerning the type of policy that affects the premium is the face value of the policy. Two policies with the same type of insurance (life), the same product (life insurance for families), and the same list of covered items, but different face values should have different premiums.

7.1.2. Risk Factors of the Insured Person. Each type of insurance has specific risk factors. For an automobile insurance policy, the risk factors might include the driving history, age, and location of the policyholder. By contrast, the risk factors for a fire insurance policy would include the size, age, and construction of the building.

This study focuses on life insurance. It is intuitive that the life expectancy of the insured will have a negative relationship with the premium he/she will pay, other things being equal. So, what affects the life expectancy of a person? The first aspect that comes to mind is age. Generally, a younger person is expected to live longer than an older person will. Most insurance companies employ actuaries to estimate how long a person is expected to live. A second aspect is the health of the insured person. A third aspect is the insured person's profession because some professions have higher mortality rate than others. For example, a coal miner is expected to have more health issues than a university administrator at the same age. The fourth aspect might be any risky sports or habits. For example, a biker might have a higher mortality risk than a non-biker of the same age and profession, and a smoker has a higher mortality risk than a non-smoker does.

7.1.3. Importance of the insured person to the insurance company. Just as any other businesses might offer special discounts for its VIP clients or customers, an insurance company might offer special pricing for its special clients. For example, someone who plans to get several insurance policies (e.g., life, disability, home, car, and/or boat) from the same company may be offered a special "bundle" price that is lower than the sum of the individual premiums for those policies.

7.1.4. The Insurance Company's Profit Margin. The premium on any insurance policy serves two purposes. The sum of the present values of the annual premiums over the policy period must cover the expected value of contractual payout under the policy, adjusted for the estimated or expected risk that the payout will occur. Any portion of the premium in excess of that amount goes to cover the insurance company's operating costs and provide a profit margin.

7.2. Robust Regression

Robust regression is an important tool for analyzing data that are contaminated with outliers (Chen & Meer, 2003). It can be used to detect outliers and to provide stable results in the presence of outliers. To achieve this stability, robust regression uses various statistical methods to limit the influence of outliers and stabilize the results. These methods measure the proportion of outliers in the data, which helps the regression model to limit the influence of these outliers on the results.

The choice of the method to be used in a robust regression depends on the characteristics of the data (Chen & Meer, 2003). One of the simplest method is Huber's (1973) M estimation (maximum likelihood). It is used extensively to analyze data when it can be assumed that the outlier contamination is mainly in the response direction. However, the M estimation method cannot be used if the contamination is in the explanatory direction of the data. Another method is least trimmed squares (LTS) estimation (Rousseeuw & Van Driessen, 2006; Rousseeuw & Yohai, 1984). Unlike ordinary least square, which minimizes the sum of squared residuals over n points, LTS attempts to minimize the sum of squared residuals over a subset k of those n points. The remaining $(n–k)$ points are not used because they are presumed to be outliers. The breakdown value is a measure of the proportion of contamination that a procedure can withstand and still maintain its robustness. LTS is considered to be a high breakdown value method. A third method is the S estimation (Rousseeuw & Yohai, 1984), which finds a line (or plane or hyperplane) that minimizes a robust estimate of the scale of the residuals. S estimation and LTS have the same breakdown value, but S estimation has higher statistical efficiency than LTS estimation. A fourth method is the MM estimation introduced by Yohai (1987), which combines the high breakdown property of S estimation with higher statistical efficiency.

As the main objective in this section is to find outliers in the premium based on the known data about the policy and the insured person, it can be assume that any contamination in the data will be in the response part (the premiums) rather than the exploratory part (details about the policy and the insured person). With this assumption in mind, the M estimation method is chosen to deal with the outlier problem by creating a robust regression.

7.3. The Model

Based on the factors affecting the premium calculation discussed above, the model for this part of the study is developed using the ROBUSTREG procedure in SAS with the M estimation method to stabilize the results.

The data for the type of policy have three variables: (1) the type of insurance policy (life or disability); (2) the type of the product (e.g., Accidents, Life, Life Uniclass, and Life Women); and (3) the type of coverage (e.g., death by accident, death by disease, funeral, grave, and hospital expenses).

7.3.1. Type of Policy – Insurance. Table 6 shows the frequency of different types of insurance policies in this data.

Code 93 stands for group life policies, Code 81 stands for individual disability policies, and Code 82 stands for group disability policies. At this insurance company, Code 81 (individual disability) includes only individual policies, but Code 93 (group life) and Code 82 (group disability) might actually include both group policies and individual policies. The accurate measure of whether a particular policy belongs to an individual or a group is the existence of either an individual unique identifier (Social Security Number), or a company unique identifier (tax ID). Of course, the existence of a date of birth in the data would obviously indicate an individual policy.

Table 6. Types of Insurance Policies.

COD_RAMO	Frequency	Percent	Cumulative Frequency	Cumulative Percent
93	6,059,531	53.69	6,059,531	53.69
81	3,398,601	30.12	9,458,132	83.81
82	1,827,207	16.19	11,285,339	100.00

The factors that affect the calculations of the premium should be different for group policies and individual policies. The model in this portion of the study focuses on individual policies, so the relevant factors are those that affect premium calculations for individual policies and the relevant characteristics relate only to individual policies. To restrict the data to individual policies, two conditions are added: (1) The individual unique identifier (SSN) is not missing; and (2) The birth data are not missing. Based on these two conditions, all of the policies under Code 93 (group life) actually should be classified as individual policies rather than group policies. For the model in this analysis, the data are restricted to Code 93 (group life) with the added condition that the policies actually belong to individuals.

7.3.2. Type of Policy – Product. There are 81 different types of products in the entire data set, but as the model will only use Code 93 (group life) policies, only the products that belong to this type of insurance are considered, resulting in 50 different products. Table 7 shows the distribution of the 15 products that constitute more than 90% of the policies under Code 93.

To demonstrate the model and the premium calculation for this study, only data for Product 157 will be used.

7.3.3. Type of Policy – Coverage. There are 69 different types of coverage in the data set, but as the model will be only for Code 93 and Product 157, there are 12 different coverage options available for this combination. Table 8 shows the distribution for these coverage options.

To demonstrate the model and the premium calculation for this study, only the data for Coverage 203, which constitute 16% of Code 93 and Product 157 data, will be used.

This combination of Code 93, Product 157, and Coverage 203 results in a total of 25,497 records. Before using the data, three additional tests are run: (1) to make sure that the premium has a positive value; (2) to make sure that the face value of the policy has a positive value; and (3) to make sure the policy is active. The data set has a variable that describes the status of the policy. A given policy can be active (Code 1), canceled by the client (Code 2), or terminated by the company (Code 10). The distribution of the policy status for the 25,497 records is shown in Table 9.

Table 9 indicates that 92.29% of the records are active policies. These 23,530 records will be used in the model. Table 10 shows the record selection process.

Fig. 15 shows a visualization of the premiums and face values of the final 23,530 records selected for the model.

Table 7. Distribution of Products.

COD_ PRODUTO	Frequency	Percent	Cumulative Frequency	Cumulative Percent
167	1,395,500	23.03	1,395,500	23.03
242	966,625	15.95	2,362,125	38.98
168	948,033	15.65	3,310,158	54.63
431	336,616	5.56	3,646,774	60.18
429	292,844	4.83	3,939,618	65.02
420	257,218	4.24	4,196,836	69.26
428	192,376	3.17	4,389,212	72.43
405	173,160	2.86	4,562,372	75.29
430	168,908	2.79	4,731,280	78.08
427	166,010	2.74	4,897,290	80.82
157	159,002	2.62	5,056,292	83.44
493	119,947	1.98	5,176,239	85.42
403	114,971	1.90	5,291,210	87.32
425	112,853	1.86	5,404,063	89.18
150	96,420	1.59	5,500,483	90.77

Table 8. Coverage Distribution.

COD_ GARANTIA	Frequency	Percent	Cumulative Frequency	Cumulative Percent
203	25,497	16.04	25,497	16.04
221	25,497	16.04	50,994	32.07
242	25,497	16.04	76,491	48.11
482	25,490	16.03	101,981	64.14
624	25,457	16.01	127,438	80.15
480	18,559	11.67	145,997	91.82
435	6,933	4.36	152,930	96.18
210	1,854	1.17	154,784	97.35
223	1,854	1.17	156,638	98.51
224	1,854	1.17	158,492	99.68
205	465	0.29	158,957	99.97
211	45	0.03	159,002	100.00

Table 9. Policy Status Distribution.

COD_SITU AC AO_ APOLIC E	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	23,530	92.29	23,530	92.29
2	1,967	7.71	25,497	100.00

Table 10. Data Records Selected for the Model.

Conditions	Number of Records
	11,285,339
Insurance type = 93 (required for insurance type factor)	6,059,531
Individual unique identifier (SSN) (required to prove individual policies)	6,059,531
Birthdate (required to calculate age)	6,059,531
Product type = 157 (required for product type factor)	159,002
Coverage type = 203 (required for coverage type factor)	25,497
Premium has a positive value (to eliminate data entry errors)	25,497
Face value has a positive value (to eliminate data entry errors)	25,497
Policy status is active	23,530

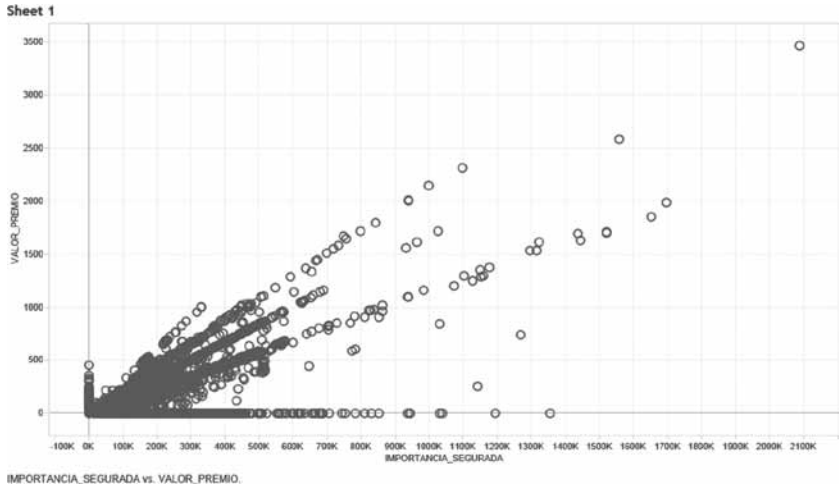


Fig. 15. Premiums and Face Values of Records for Robust Regression.

7.3.3. Life Expectancy of Insured Person. Various factors may affect the life expectancy of the insured person, including age, health, profession, and risky sports or habits. However, the only personal information in the data set is the birthdate of the insured person, which can be combined with the date of purchasing the policy for the first time (initial effective date of the policy) to calculate the age of the insured person at the time when the policy was purchased.

7.3.4. Other Factors. As discussed previously, the importance of the insured person to the insurance company and the insurance company's profit margin may be factors that affect the premium calculations for the policy. Unfortunately, there are no indicators for these factors in the data, so the model is limited to the type of policy factors and only the age of the insured person as an indicator of life expectancy due to a lack of data. The performance of the model is likely to be negatively affected by this data limitation.

7.4. SAS ROBUSTREG

The original model was

$$\text{Premium} = \beta_0 + \beta_1 \text{ Insurance Type} + \beta_2 \text{ Product Type} + \beta_3 \text{ Coverage Type} + \beta_4 \text{ Face Value} + \beta_5 \text{ Age} + \beta_6 \text{ Health} + \beta_7 \text{ Profession} + \beta_8 \text{ Risky Choices} + \beta_9 \text{ Importance} + \beta_{10} \text{ Profit}$$

Due to a lack of the necessary data for health, profession, risky choices, importance, and profit, the model becomes:

$$\text{Premium} = \beta_0 + \beta_1 \text{ Insurance Type} + \beta_2 \text{ Product Type} + \beta_3 \text{ Coverage Type} + \beta_4 \text{ Face Value} + \beta_5 \text{ Age}$$

As discussed above, the selection of a specific Insurance Type (Code 93), Product Type (Product 157), and Coverage Type (Coverage 203) results in the following model:

$$\text{Premium} = \beta_0 + \beta_1 \text{ Face Value} + \beta_2 \text{ Age}$$

Table 11 shows some statistical descriptions of the face value, age, and premiums in the selected data.

7.5. Results

Table 12 shows the results of running the SAS ROBUSTREG with M estimation on the selected data, and Table 13 shows goodness-of-fit measures to evaluate the model's performance.

As expected, these results indicate that the coefficients for the face value of the policy and the age of the policyholder are significant predictors of the insurance policy premium ($p < 0.0001$), and the model explains 42% of the variability in policy premiums. Presumably, the predictive ability of the model would improve

Table 11. Summary Statistics.

Summary Statistics						
Variable	**Q1**	**Median**	**Q3**	**Mean**	**Standard Deviation**	**MAD**
IMPORTANCIA_SEGURADA	61,655.4	111,607	205,341	147,578	123,407	93,328.9
Age	49.8301	55.4808	61.5283	55.8372	8.5482	8.6499
VALOR_PREMIO	99.8505	130.4	242.6	184.5	173.6	99.1340

Variable	**Minimum**	**Maximum**	**Mode**
Age	26.6465753	95.0082192	60.6876712
IMPORTANCIA_SEGURADA	2.7001000	2,088,930.62	47,579.44
VALOR_PREMIO	0.0044223	3,463.54	103.2124727

Table 12. ROBUSTREG Results.

Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Intercept	1	56.7042	2.1241	52.5410	60.3674	712.64	<0.0001
IMPORTANCIA_SEGURADA	1	0.0011	0.0000	0 0011	0.0011	167,139	<0.0001
Age	1	−0.3760	0.0377	−0.4499	−0.3021	99.43	<0.0001

Table 13. ROBUSTREG Goodness of Fit.

Goodness-of-Fit	
Statistic	**Value**
R-Square	0.4215
AICR	44,505.57
BICR	44,530.75
Deviance	76,055,543

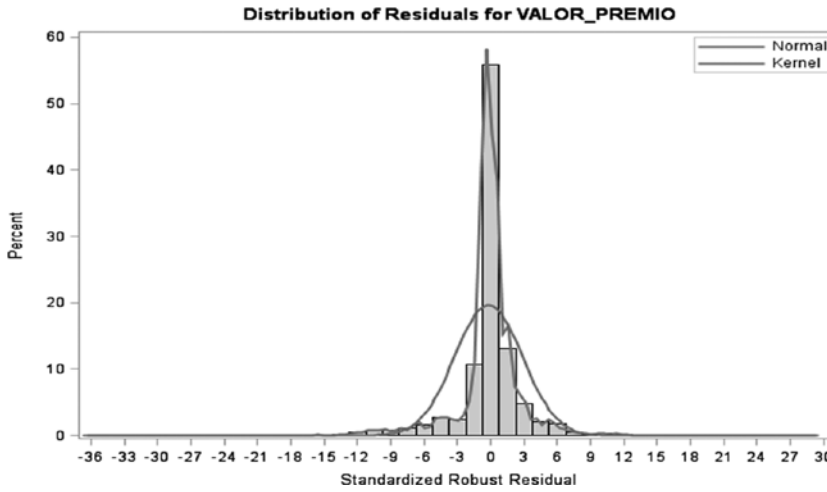


Fig. 16. Distribution of Residuals.

significantly if the factors that were omitted due to lack of available data (health, profession, risky choices, importance, and profit) could be included.

To define the outliers, the ROBUSTREG procedure calculates standardized residuals. Each residual is the positive or negative difference between the actual premium stated in the data set and the calculated premium based on the model. To calculate the standardized residual, the procedure estimates a scale. In this case, the scale is estimated to be 41.34. For example, if the calculated premium based on the model is BR[1] 1,254 and the actual premium in the sample data set for that policy is BR 250, the absolute residual premium would be BR 1,004. As the scale is estimated to be 41.34, the standardized absolute residual would be 24.28 (1,004/41.34). The outliers are defined based on a standardized residual cutoff point. If the absolute standardized residual lies below that cutoff point, the record is considered to be normal, but if it is above that cutoff point, the record is considered to be an outlier. The ROBUSTREG procedure allows users to set their own standardized residual cutoff point (k) or choose the default cut-off point for the procedure, which is ±3.

The graph in Fig. 16 shows the distribution of the standardized residuals for the premiums in the data set used to test the model.

As seen in Fig. 16, the standardized residual ranges from −36 to 30. About 55% of the records have a residual of zero, and about 85% of the records have a standardized residual between −3.0 and 3.0. Any data point with standardized residual beyond this range (−3.0 to +3.0) is considered an outlier. Table 14 shows the outlier cutoff point used and the proportion of the outliers based on this cutoff.

[1]All monetary values are in Brazilian Real (BR).

Table 14. Outliers at a Cutoff Point of ± 3.

Diagnostics Summary		
Observation Type	**Proportion**	**Cutoff**
Outlier	0.1582	3.0000

Using the default cutoff point of ±3, the graph in Fig. 17 visualizes the model's results for the relationship between premiums and face values of the 23,530 data records tested.

The points shown in dark grey represent the records that are not flagged as outliers; this means that the absolute standardized premium residuals in these cases are less than three. By contrast, the points shown in orange represent the records that are flagged as outliers because their absolute standardized premium residuals are greater than or equal to three. The blue points shown in the graph divide the outliers into two groups. The first group, located above the blue line, includes records for which the insurance company is collecting more premiums than predicted by the model. The second group, located below the blue line, includes records for which the insurance company is actually collecting less premiums than the model predicts it should. While the first group is profitable for the insurance company, the second group is suspicious. An auditor might be interested in studying the second group more thoroughly. The most suspicious group of outliers is highlighted by the black oval. These records have almost an equal value of premiums that are very close to zero, whereas the face values are increasing to almost BR 1,400,000.

As mentioned before, the lack of available data for some variables that should be included in the premium calculations is likely to impair the model's

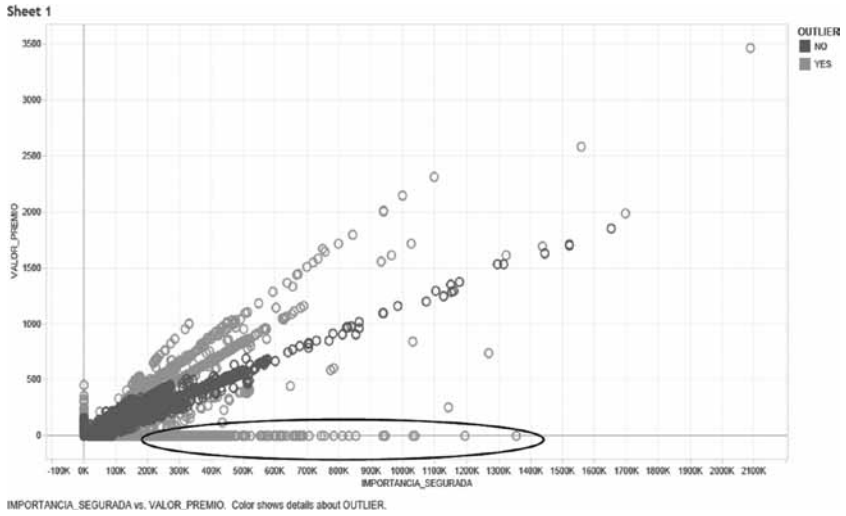


Fig. 17. Premium-face Value Outliers for a Cutoff Point of ±3.

performance. However, other factors that are not included in the model may also affect the premium calculations. To demonstrate an example of this fact, the ages of the insured individuals are grouped as follows in Table 15.

Since insurance type, product type, and coverage type are all controlled through the selection of the data to test, the only two factors remaining in the model as predictors of insurance premiums are the age of the policyholder when the policy was first implemented and the face value of the policy. If these are the only factors affecting the premium calculations, then the range of residuals in each group should be minimal. However, when the distribution of the standardized residuals in each age group is graphed, it becomes apparent that this is not the case, as shown in Fig. 18.

The graph in Fig. 19 shows that the distribution of residuals within each group varies dramatically in some groups. For people in their 20s and 30s, the distribution span of the residuals is minimal, whereas for older policyholders, the span is significant and ranges from −35 to +30. The most likely explanation for this result is increased variability in the health status of older policyholders, which could not be captured in the model because these data are not available in the data set.

The ROBUSTREG procedure allows the cutoff point to be changed from the default value of ±3, depending on the user's understanding of the data. To see

Table 15. Age Groups.

Age	Group
Under 18	Child
18–Under 30	Twenties
30–Under 40	Thirties
40–Under 50	Forties
50–Under 60	Fifties
60–Under 80	Senior
80 and above	Elder

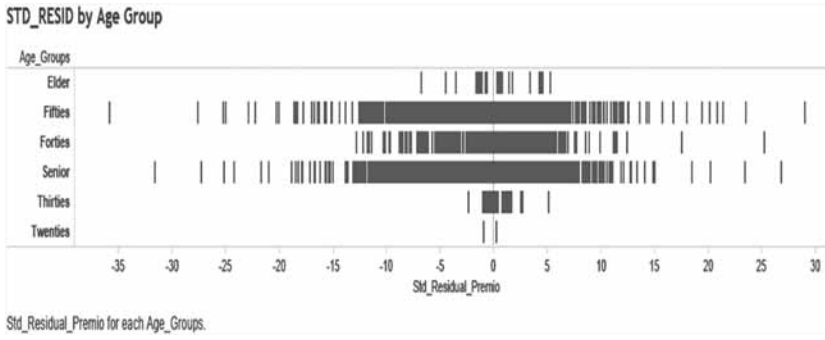


Fig. 18. Distribution of Standardized Residual by Age Group.

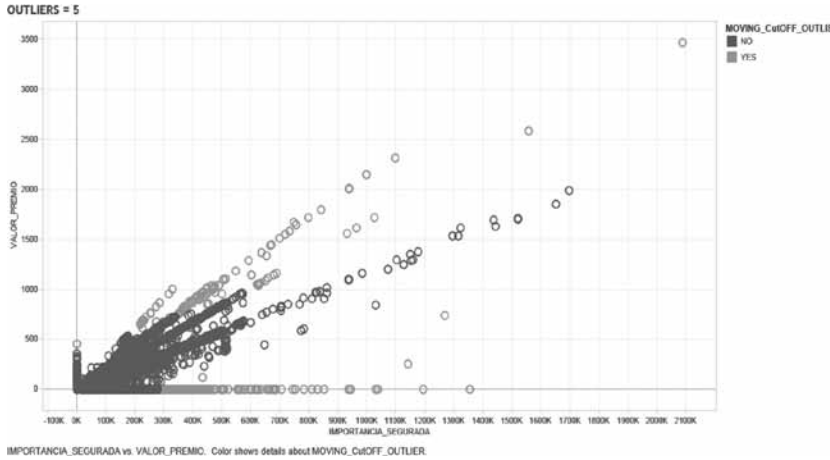


Fig. 19. Premium-face Value Outliers with a Cutoff Point of ±8.

the effect of changing this parameter, the cutoff is expanded to a very loose ±8 instead of ±3, which should increase the number of records that fall within the acceptable predicted range and reduce the number of records identified as outliers. The results of this change are shown in Table 16.

Compared to the results in Table 14, in which 15.82% of the records that are identified as outlier when the cutoff point is set at ±3, Table 16 shows that the model identifies 4% of the records as outliers with the looser cutoff point of ±8. The graph in Fig. 19 shows the visualization of the premiums and face values of the 23,530 data records based on the model, with the 8.0-cutoff point.

8. Conclusion and Limitations

This chapter focuses on detecting anomalies in life/disability insurance. In the claim payment business cycle, a methodology is defined to test two audit assertions: Is the claim settlement reasonable? Is the claim itself legitimate? A multi-dimensional approach is used in which the available attributes are divided into different groups (dimensions). The dimensions are interest payments, reason-coverage association, timeline, and group similarities. As an additional dimension, a belief function is used to create a risk score for each branch of the company. Each dimension is used to identify insurance claim anomalies logically. Then, the weighted average of the dimensions is used to priorities the anomalies that are

Table 16. Outliers at a Cutoff Point of ± 8.0.

Observation type	# of Records	Proportion of Outliers	Cutoff Point
Outliers	869	0.04	8.00

found. Finally, in the premium collection business cycle, an example of a model to detect premium outliers is constructed.

This chapter illustrates how design science can be used to showcase ideas. A key challenge in this study for both the claim payment and premium collection business cycles is data availability, yet these difficulties should motivate further research. In field work, auditors face many challenges in accessing the data they need to fulfill their duties and form their opinion even if their clients are fully digitalized and technologically capable of providing the information required. Future research needs to find ways to deal with this problem and investigate alternative possible solutions, such as the use Audit Data Standards and macroeconomic data.

References

Artıś, M., Ayuso, M., & Guillen, M. (1999). Modelling different types of automobile insurance fraud behaviour in the Spanish market. *Insurance: Mathematics and Economics*, *24*(1–2), 67–81.

Artís, M., Ayuso, M., & Guillén, M. (2002). Detection of automobile insurance fraud with discrete choice models and misclassified claims. *Journal of Risk and Insurance*, *69*(3), 325–340.

Bakar, Z. A., Mohemad, R., Ahmad, A., & Deris, M. M. (2006). A comparative study for outlier detection techniques in data mining. In *IEEE conference on cybernetics and intelligent systems*. Bangkok, Thailand. Retrieved from http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4017846

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science 17*(3), 235–249.

Breunig, M., Kriegel, H., Ng, R., & Sander, J. (2000). LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data* (pp. 93–104). Texas, USA: Dallas.

Campbell, C., & Bennett, K. P. (2001). A linear programming approach to novelty detection. In *Advances in neural information processing systems* (pp. 395–401).

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 1–6.

Chan, D. Y., & Vasarhelyi, M. A. (2011). Innovation and practice of continuous auditing. *International Journal of Accounting Information Systems*, *12*(2), 152–160.

Chen, H., Chiang, R., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*(4), 1165–1188.

Chen, H., & Meer, P. (2003). Robust regression with projection based M-estimators. In *Proceedings ninth IEEE international conference on computer vision* (*ICCV*). (pp. 878–885), Nice, France.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, *38*(2), 325–339.

Eskin, E. (2000). Anomaly detection over noisy data using learned probability distributions. In *Proceedings of the international conference on machine learning* (pp. 255–262). Stanford, CA, USA.

Ghosh, S., & Reilly, D. L. (1994). Credit card fraud detection with a neural-network. In *Proceedings of the twenty-seventh Hawaii international conference on system sciences* (pp. 621–630). Wailea, HI, USA.

Hawkins, S., He, H., Williams, G., & Baxter, R. (2002). Outlier detection using replicator neural networks. In *Proceedings of the 4th international conference on data warehousing and knowledge discovery (DaWaK)* (pp. 170–180). Aix-en-Provence, France.

Hodge, F. D. (2001). Hyperlinking unaudited information to audited financial statements: Effects on investor judgments. *The Accounting Review*, *76*(4), 675–691.

Hu, W., Liao, Y., & Vemuri, V. R. (2003). Robust support vector machines for anomaly detection in computer security. In *IEEE international conference on machine learning and applications (ICMLA)* (pp. 168–174). Los Angeles, California.

Huber, P. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, *1*(5), 799–821.

Knorr, E. M., Ng, R. T., & Tucakov, V. (2000). Distance-based outliers: Algorithms and applications. *The International Journal on Very Large Data Bases*, *8*(3–4), 237–253.

Lee, W., Stolfo, S. J., & Mok, K. U. I. W. (2000). Adaptive intrusion detection: A data mining approach. *Artificial Intelligence Review*, *14*(6), 533–567.

Major, J. A., & Riedinger, D. R. (2002). EFD: A hybrid knowledge/statistical-based system for the detection of fraud. *The Journal of Risk and Insurance*, *69*(3), 309–324.

Mock, T. J., Sun, L., Srivastava, R. P., & Vasarhelyi, M. (2009). An evidential reasoning approach to Sarbanes–Oxley mandated internal control risk assessment. *International Journal of Accounting Information Systems*, *10*(2), 65–78.

Moharram, B. (2016). *Auditing in environments of diverse data*. Doctoral dissertation, Rutgers University Graduate School, Newark, NJ.

Patcha, A., & Park, J. M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, *51*(12), 3448–3470.

Pathak, J., Vidyarthi, N., & Summers, S. L. (2005). A fuzzy-based algorithm for auditors to detect elements of fraud in settled insurance claims. *Managerial Auditing Journal*, *20*(6), 632–644.

Rousseeuw, P. J., & Van Driessen, K. (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, *12*(1), 29–45.

Rousseeuw, P., & Yohai, V. (1984). Robust regression by means of S-estimators. In *Lecture notes on statistics* (Vol. *26*, pp. 256–272). New York, NY: Springer.

Shafer, G. R. (1996). *The art of causal conjecture*. Cambridge, MA: MIT Press.

Shafer, G. R., & Srivastava, R. P. (1990). The Bayesian and belief-function formalisms: A general perspective for auditing. *Auditing: A Journal of Practice & Theory*, *9*, 110–137.

Srivastava, R. P. (1993). Belief functions and audit decisions. *Auditors Report*, *17*(1), 8–12.

Srivastava, R. P., & Mock, T. J. (2000). Belief functions in accounting behavioral research. In *Advances in accounting behavioral research* (Vol. *3*, pp. 225–242). Bingley: Emerald Group Publishing Limited.

Srivastava, R. P., & Mock, T. J. (2005). Why we should consider belief functions in auditing research and practice. *Auditors Report*, *28*(2), 1–8.

Srivastava, R. P., & Mock, T. J. (2011). The Dempster–Shafer theory of belief functions for managing uncertainties: An introduction and fraud risk assessment illustration. *Australian Accounting Review*, *21*(3), 282–291.

Srivastava, R. P., Mock, T. J., & Turner, J. L. (2007). Analytical formulas for risk assessment for a class of problems where risk depends on three interrelated variables. *International Journal of Approximate Reasoning*, *45*(1), 123–151.

Srivastava, R. P., Rao, S. S., & Mock, T. J. (2013). Planning and evaluation of assurance services for sustainability reporting: An Evidential Reasoning approach. *Journal of Information Systems*, *27*(2), 107–126.

Srivastava, R. P., & Shafer, G. R. (1992). Belief-function formulas for audit risk. *The Accounting Review*, *67*(2), 249–283.

Solberg, H. E., & Lahti, A. (2005). Detection of outliers in reference distributions: Performance of Horn's algorithm. *Clinical Chemistry*, *51*(12), 2326–2332.

Sun, L., Srivastava, R. P., & Mock, T. J. (2006). An information systems security risk assessment model under the Dempster–Shafer theory of belief functions. *Journal of Management Information Systems*, *22*(4), 109–142.

Viaene, S., Dedene, G., & Derrig, R. (2005). Auto claim fraud detection using Bayesian learning neural networks. *Expert Systems with Applications*, *29*(3), 653–666.

Viaene, S., Mercedes, A., Guillen, M., Gheel, D. V., & Dedene, G. (2007). Strategies for detecting fraudulent claims in the automobile insurance industry. *European Journal of Operational Research*, *176*(1), 565–583.

Williams, G., & Baxter, R. (2002). A comparative study of RNN for outlier detection in data mining. In *IEEE international conference on data mining* (pp. 1–16). Maebashi City, Japan.

Wong, W., Moore, A., Cooper, G., & Wagner, M. (2003). Bayesian network anomaly pattern detection for disease outbreaks. In *IEEE international conference on machine learning and applications (ICMLA)* (pp. 808–815). Washington, DC.

Yamanishi, K., Takeuchi, J., Williams, G., & Milne, P. (2004). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery*, *8*(3), 275–300.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, *15*(2), 642–656.

Yu, D., Sheikholeslami, G., & Zhang, A. (2002). FindOut: Finding outliers in very large datasets. *Knowledge and Information Systems*, *4*(4), 387–412.

Part IV

Audit Analytics in Transitory Systems

This page intentionally left blank

Chapter 9

Development of an Anomaly Detection Model for a Bank's Transitory Account System[1]

Yongbum Kim

1. Introduction

The goal of anomaly detection models is to screen or filter out true anomalies from a population. However, while focusing on the power of these models, both researchers and practitioners tend to neglect their practicability, which refers to whether the models can be implemented efficiently and effectively in real-world situations. If monitoring is performed infrequently (e.g., annually or semi-annually), the practical aspects of implementation may not be a great concern since all of the data can be downloaded or transferred to a designated place and examined by the models. However, if continuous monitoring is necessary, the issue of practicability becomes significant, especially to practitioners.

One of a bank's major functions is to transfer funds through a "wire transfer" from one customer to another. While the sender can be either an account holder or a non-account holder at the bank, the transfer's recipient is generally an account holder. However, it is not uncommon in the process of a fund transfer that the recipient of the wire transfer is not identified immediately when a bank receives funds from a sender, for example, because of an incorrect account number. When a bank cannot identify the recipient immediately, the funds need a place to stay for the time being. A transitory bank account is a special account to serve this purpose by holding the money until the correct recipient is clearly identified. Although the time to identify the recipient ranges from a few seconds to several months, most wire transfers do not stay in a transitory account for long. A bank may have many transitory accounts, depending on its needs and purposes. The bank in this study has about 10,000 transitory accounts.

[1]This chapter is based on the study that is published in *Journal of Information Systems* (Kim & Kogan, 2014).

Audit Analytics in the Financial Industry, 147–164

doi:10.1108/978-1-78743-085-320191009

Detecting abnormal transactions out of millions of transactions in about 10,000 transitory bank accounts is a challenging, a time-consuming, and an expensive task in terms of data processing. An online or real-time monitoring system must be developed carefully since it will have an impact on a company's daily data processing resources. Otherwise, it may consume too much of the firm's data processing resources, interrupt regular business activities, and possibly distort the data processing system. These factors may make a company hesitant to integrate an internal control (IC) system into its existing data processing systems. Furthermore, the complexity of IC screening models described in the literature can serve as another barrier to implementing a screening model in practice. Most of the fraud detection models in the literature require mathematical and/or statistical expertise that most internal auditors do not possess or understand. One practical solution for internal auditors can be to develop a screening model consisting of a series of generic rules that do not contain complicated mathematical or statistical algorithms. Given these concerns, it would be useful to develop a screening model that can be applied to transitory bank accounts without significantly consuming and possibly interrupting the bank's data processing systems.

However, this strategy has a potential weakness because enhanced practicability implies that an anomaly detection model should be sufficiently simple and generic to implement on the current data processing system, which may reduce the power of the model. Thus, this approach requires a trade-off between the power of a screening model and its practicability, so the success of this approach depends on sacrificing a little of the model's power in order to keep it simple and easy to implement.

2. Objectives

A transitory account is a temporary buffer for funds in transit until the destination is identified by human intervention and updated by a manual process that transfers the funds from the transitory account to the designated account. Therefore, a transitory account can be vulnerable to anomalies, including internal fraud, when its activities and the employees in charge are not rigorously monitored and the transactions are not verified. The monitoring and verification process can be done with traditional manual process or with a help of modern technology.

The purpose of this chapter is to develop and test IC monitoring models that will detect anomalies out of the millions of transactions in approximately 10,000 transitory bank accounts. Based on considerations of efficiency and practicability, the model should be capable of being implemented on all of the bank's transitory accounts without consuming significant data processing resources. The transactions flagged by the model are then cross-checked by the actual internal auditors to estimate the model's power, and the results can be used to improve the model further.

In Section 3, the data set in this study and the screening rules used to detect possible anomalies are discussed, and the test results are described and evaluated. Section 4 offers conclusion and suggestions for future research.

3. Methodology

3.1. Phase I

3.1.1. Data The data for this study are transitory account transactions from one of the largest banks in Brazil. The data set includes 16 selected transitory accounts out of the 10,000 possible accounts. The dates in each account have different ranges. The narrowest range is about a year (from late May 2007 to early August 2008), while the longest is about three years (from early October 2005 to early August 2008). Table 1 shows details of the date ranges by account. As Table 1 shows, most of the data fall in the narrowest range.

Before data cleaning, there were 580,020 records, including 2 records with missing values. Among the 580,018 records without missing values, 121,899 pairs of records are found to be identical. After the duplicate records are removed, the resulting 458,119 transactions contain 221 pairs that have the same values for all attributes except the balance field that indicates the remaining amount to be cleared. Since the records with lower balance amounts are generally the more recent ones, unless the balance is found to increase, the records with the larger balance amounts are considered to be prior versions of subsequently updated records, so they are also eliminated from the data set. The final data set has 457,898 observations. The descriptive statistics for the final data set are shown in Tables 2 and 3 by variable and account.

Table 1. Data – Date Ranges.

Account	Distinct Days	Oldest	Latest	Range
5738	518	10/04/2005	08/11/2008	1,043
45136	244	02/02/2006	08/11/2008	922
60836	202	04/02/2007	08/11/2008	498
32360	233	01/26/2006	08/11/2008	929
61042	227	04/30/2007	08/11/2008	470
21830	232	04/03/2006	08/11/2008	862
21776	226	05/25/2006	08/11/2008	810
68128	226	04/30/2007	08/11/2008	470
58122	186	05/29/2007	08/11/2008	441
302	221	06/28/2006	08/11/2008	776
70050	210	05/07/2007	08/11/2008	463
70068	190	05/10/2007	08/11/2008	460
1155	173	05/25/2007	08/07/2008	441
94870	155	05/29/2007	08/11/2008	441
61930	177	05/24/2007	08/11/2008	446
66613	167	02/28/2007	08/11/2008	531

Table 2. Descriptive Statistics – Amount.

Account	Variable	*n*	nmiss	Average	Median	Stdev	Min	Max
1155	Amount	694	0	7,434.65	1,777.02	19,147.03	1.00	170,073.99
21776	Amount	25,719	0	2,084.08	607.54	7,235.72	4.95	440,000.00
21830	Amount	21,983	0	1,036.28	68.04	8,130.35	0.01	843,000.00
302	Amount	5,116	0	690,248.20	286.35	14,636,008.00	0.01	418,030,303.00
32360	Amount	62,916	0	666.39	50.00	7,829.16	0.01	899,348.33
45136	Amount	65,289	0	216,625.50	236.70	4,347,807.50	0.01	311,084,647.00
5738	Amount	133,564	0	706.29	5.40	26,615.34	0.01	4,252,752.50
58122	Amount	18,021	0	3,532,071.00	109,614.30	11,406,518.00	0.01	309,072,377.00
60836	Amount	79,652	0	38,148.42	7,518.42	601,805.66	0.01	70,000,276.00
61042	Amount	19,283	0	5,855.88	368.88	225,951.27	0.01	30,040,000.00
61930	Amount	729	0	5,037,741.00	900,000.00	13,936,447.00	15.63	230,000,642.00
66613	Amount	773	0	9,765,888.00	10,000.00	216,507,723.00	0.03	5,899,996,308.00
68128	Amount	19,755	2	43,530.23	177.79	615,870.99	0.01	31,867,577.10
70050	Amount	3,010	0	7,905.24	284.19	109,117.21	0.01	4,261,950.73
70068	Amount	915	0	429,573.40	1,600.00	8,036,053.80	0.01	241,203,449.00
94870	Amount	479	0	37,519.10	10,000.00	88,582.10	0.40	900,037.44

Table 3. Descriptive Statistics – Balance.

Account	Variable	*n*	nmiss	Average	Median	Stdev	Min	Max
1155	Balance	694	0	424.64	0	7,057.19	0	163,907.49
21776	Balance	25,719	0	85.46	0	1,052.05	0	67,824.68
21830	Balance	21,983	0	89.99	0	1,004.78	0	58,905.00
302	Balance	5,116	0	140,012.80	0	6,183,851.20	0	386,445,649.00
32360	Balance	62,916	0	75.59	0	2,388.80	0	465,570.00
45136	Balance	65,289	0	4,436.78	0	753,667.69	0	189,000,000.00
5738	Balance	133,564	0	48.41	0	4,088.69	0	820,000.00
58122	Balance	18,021	0	7,705.93	0	589,623.84	0	74,220,000.00
60836	Balance	79,652	0	0.00	0	0.00	0	0.00
61042	Balance	19,283	0	1,754.47	0	216,470.28	0	30,040,000.00
61930	Balance	729	0	473,788.90	0	4,445,412.30	0	105,000,000.00
66613	Balance	773	0	171,286.40	0	3,404,279.70	0	92,999,468.60
68128	Balance	19,755	2	48.96	0	1,140.55	0	101,701.00
70050	Balance	3,010	0	444.06	0	3,133.40	0	100,000.00
70068	Balance	915	0	1,810.18	0	21,662.35	0	502,144.85
94870	Balance	479	0	304.32	0	5,380.13	0	116,895.33

3.1.2. Screening Rules The IC screening model in this study is a collection of rules that will be applied to the transactions to examine whether they may have anomalies. As the first step, the scenarios or cases that are of the most concern to internal auditors are identified. The materiality of transaction amounts is the primary concern for internal auditors when they investigate the potential for frauds, so this analysis focuses more on transactions with large amounts than those with smaller amounts. The second concern is the effect of the model's implementation on the bank's systems. Consequently, the model should require as little computational resources as possible. The final issue is that manual entries are much riskier than automatic/systematic entries. Considering that the minimum requirements for the model are to choose accounts with large values, use less computational resources, and identify manual entries, several monitoring rules are developed and tested. The overall blueprint for the bank's IC system implementation in this case is shown in Fig. 1.

For Level 1 of the procedure shown in Fig. 1, practitioners selected 150 specific accounts based on the decision efforts required to audit them. Based on the selected transitory accounts, the Level 2 general screening rules should support mainframe-level implementation, and the Level 3 detailed screening rules aim at terminal or personal computer-level monitoring. Depending on the power of the systems, both the Level 2 and Level 3 monitoring rules can be implemented at the mainframe-level.

In order to develop general screening rules that do not consume too many system resources, a collection of procedures is applied. This process named P-rule is to find the transactions with material amount, based on the fact that the majority of transactions have relatively small amounts. Thus, the distribution of the amounts is positively skewed and has a high peak. The overall logic of the procedure is shown in Fig. 2.

If there is a specific material amount above which all records should be examined, its location, represented by percentile, will be changed according a shape of the distribution that are measured by the skewness (= degree of asymmetry) and

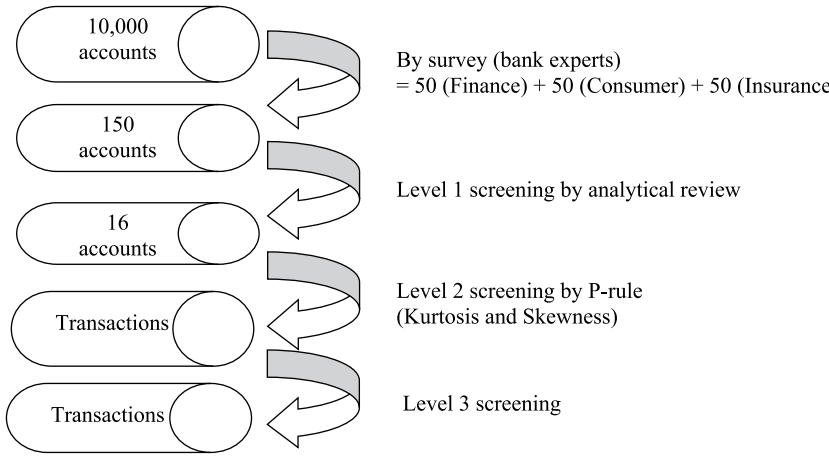


Fig. 1. Internal Control Procedure.

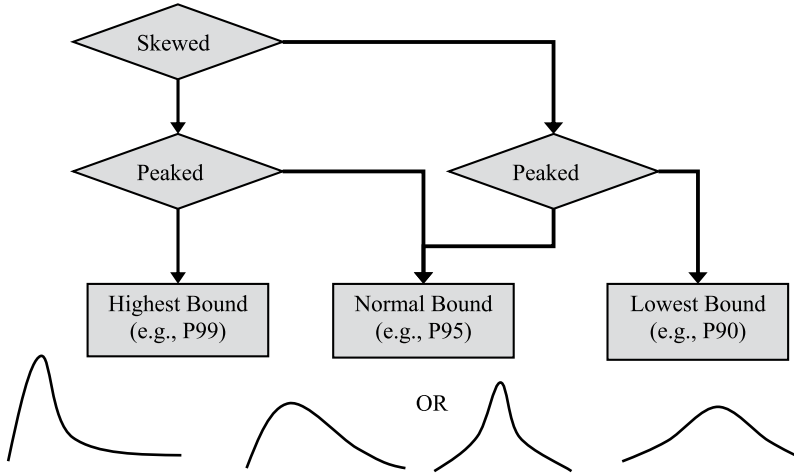


Fig. 2. P-Rule Flowchart.

kurtosis (= degree of peakedness, or more precisely, degree of tail sizes). For example, assume that two data sets that have the same number of the transactions and the same means. If all of the amounts are extremely positively skewed (the decision degree of skewness = 1) and peaked (the decision degree of kurtosis = 1) in one data set, then only a few observations may exceed the material level. By contrast, if the distribution is less positively skewed and less peaked in the other data set, then more observations are likely to exceed the cutoff point. The main reason to use this distribution information is to reduce the computational cost. If the computer system is more powerful, a method using prediction intervals could be substitute for the simpler percentile rules. Although the decision criteria chosen in this study for the degrees of skewness and kurtosis equal 1, these are arbitrary cutoffs. In addition, the parameters for cutoff distributions can be changed as well.

Based on the P-rule, two general screening methods can be applied to each account. The first method is to examine the daily sums to identify suspicious days and then investigate all transactions on the flagged days. If fraud occurs on particular days, the daily sums will be abnormally high, so this method can detect anomalies easily. However, some days may have large sums simply because there was an unusually large number of transactions on those days, so all transactions, even those with the highest values, may actually not be materially significant. Alternatively, a day might have a small daily sum that results from only a few transactions with very large amounts. In that case, some transactions that exceed the material amount cannot be detected by applying the P-rule to daily sums. The second method is to apply the P-rule to all individual transactions and flag the abnormally high transactions directly. This method requires more discrete tests, but it is more suitable to detect anomalies if the frauds occur over a period of time or randomly. In this study, both methods are used to complement each other. After applying these two screening rules, the union of both sets of potential anomalies will be the final alarms for further investigation. This process is shown in Fig. 3.

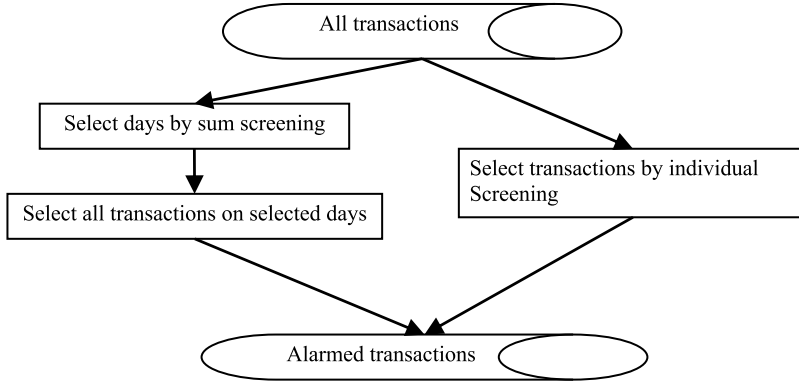


Fig. 3. Identifying and Combining Anomalies.

To identify manual entries, this study utilizes the manual indicator. If this value is either 0 or 999999999, the record was entered automatically and manual otherwise. It was expected that more manual entries than automatic entries would trigger the alarm based on the Level 2 screening rules. However, the results shown in Table 4 indicate that more automatic entries are flagged, although a significantly larger proportion of the manual entries are identified as potential anomalies.

In addition to the manual entries, duplicate amounts may act as a fraud indicator. As fraudulent transactions tend to have large amounts, they can easily be detected. Hence, a better idea for the fraudster, but a worse idea for the company, would be to split the fraudulent transaction into two or more smaller amounts. The splitting might be done in equal amounts (tested here) or different amounts (tested in a later section).

The logic for testing evenly split fraudulent transactions is relatively simple. Since one transaction belongs only to one branch on one day and branches are in separate locations, the transactions in a given branch on a given day are examined for the existence of duplicate amounts. Interestingly, the results show that there are many duplicate amounts. In an extreme case, there are 39 duplicate amounts at a particular branch in one day. Since all transactions are already screened by the Level 2 rules, most of them have relatively large amounts. If both factors are considered, this may indicate anomalies that are more likely fraudulent. Table 5 summarizes the results.

Table 4. Flagged Records among Manual and Automatic Entries.

Type	Population	Alarmed	Percentage (%)
Manual	53,591	465	0.87
Automatic	404,307	4,420	1.09
Total	457,898	4,885	1.07

Table 5. Transfers with Duplicate Amounts.

Duplicates	**2**	**3**	**4**	**5**	**8**	**11**	**12**	**23**	**24**	**27**	**39**
Branches/day	93	9	4	3	2	1	1	1	1	1	1

Another potential fraud indicator can be the number of alarms in a branch in a day. If a certain branch has more alarms than other branches, this may signal potential problems. In this study, two alarms are considered as the criterion. A detailed summary is shown in Table 6.

One possible fraud indicator might be the existence of negative values for transaction amounts based on the fact that each account has either a normal debit or credit balance and all accounts have a debit/credit indicator. Any account with an abnormal balance must contain an abnormal transaction. However, a pilot test finds that literally no case violates this rule. Another possible fraud indicator might be the aging of transactions. Since transitory accounts are temporary places to keep money, they should be cleared in a short time. However, a pilot test showed that many transactions remain in temporary accounts for six months or more.

3.1.3. Results and Discussion. Since each Level 3 screening rule can have different weight for its importance, an alternative strategy to identify anomalies is the use of a Venn diagram, as shown in Fig. 4.

This diagram is based on the assumption that while it is uncommon to have two transactions with the same large amount, it would be exceptionally rare to have those transactions occur at a branch that has more than one transaction flagged on a given day. Similarly, a higher proportion of manual transactions are likely to be flagged, but finding manual transactions in conjunction with duplicate large transaction amounts or occurring at a branch with more than one flagged transaction in a day would be highly unusual. The results of these analyses are shown in Fig. 5.

Based on the Venn diagram, the final alarms for further investigation (tests of details) could include the 3 transactions that are duplicates and were manually entered, the 92 transactions that occurred in branches with more than one flagged transaction on a given day and were manually entered, and the 8 transactions that shared all three characteristics for a total of 103 transactions. This represents a very small portion of the whole population of 457,898 transactions (0.02%). This may be due to the use of overly strict rules chosen to reduce false positives. If the screening rule parameters are changed, the number of observations selected for further testing may increase significantly.

Rather than assuming that the data set in the study had been audited and did not have any errors, this study utilizes the actual confirmation by the internal auditors. Since they believe that some transactions are truly fraudulent, it is not practical to assume that the data set is anomaly-free. In addition, if the alarmed transactions are truly fraudulent, the power of the IC model can easily be confirmed. Moreover, even if the model does not detect any actual fraudulent cases, the model can be used to deter potential future fraud. For this to work, employees

Table 6. Summary of Accounts with Two or More Alarms in a Day.

Acct	Population				All Alarmed Transactions				Alarms ≥2			
	Day *	Obs	Day	Branch	Day *	Obs	Day	Branch	Day *	Obs	Day	Branch
302	2,577	5,116	221	729	30	52	28	5	7	29	7	1
1155	635	694	173	297	6	6	6	6	–	–	–	–
5738	22,350	133,564	518	1,235	658	1,335	210	312	90	767	68	58
21776	16,781	25,719	226	1,130	471	557	126	253	46	132	29	28
21830	15,142	21,983	232	1,157	208	219	154	57	8	19	8	3
32360	28,455	62,916	233	1,021	569	629	180	241	40	100	31	26
45136	32,359	65,289	244	1,189	435	652	164	34	122	339	83	10
58122	372	18,021	186	3	117	192	99	2	59	134	55	2
60836	276	79,652	202	2	181	797	181	1	165	781	165	1
61042	5,825	19,283	227	992	157	191	100	63	26	60	23	13
61930	177	729	177	1	7	7	7	1	–	–	–	–
66613	569	773	167	325	7	7	7	4	–	–	–	–
68128	12,321	19,757	226	1,045	156	197	108	21	34	75	33	3
70050	1,931	3,010	210	642	29	30	25	21	1	2	1	1
70068	636	915	190	132	9	10	9	5	1	2	1	1
94870	155	479	155	1	3	4	3	1	1	2	1	1
All	140,561	457,900	535	1,358	3,043	4,885	233	627	600	2,442	192	122

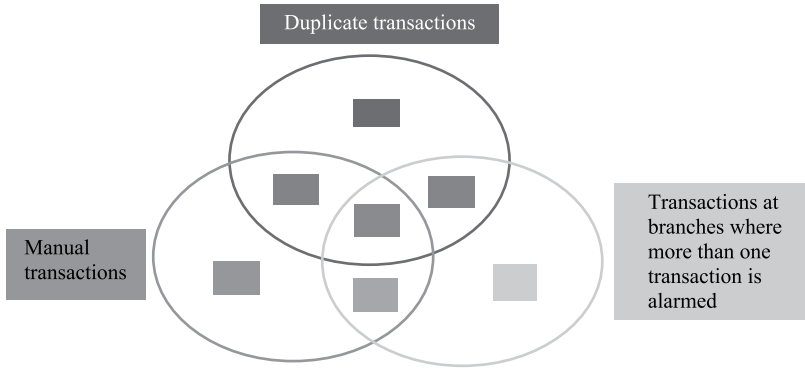


Fig. 4. Example of Venn Diagram for Potentially Anomalous Transactions.

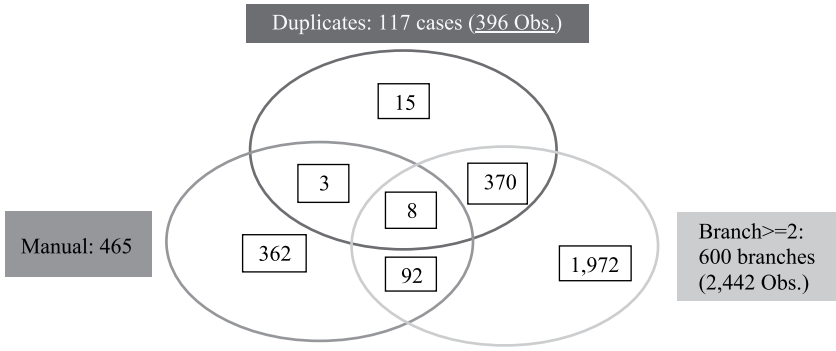


Fig. 5. Venn Diagram of Actual Potentially Anomalous Transactions.

must be aware of the model's existence and use, as well as the possible consequences if they are caught. In any cases, the model can be useful to the bank and can be a touchstone for future study.

3.2. Phase II

3.2.1. Data. While Phase I shows a wide range of information regarding the transitory account system, its result clearly suggests the need for further investigation to improve the model's effectiveness. To that end, the data set in Phase II is separated into two parts: one for model training and the other for model testing. If the transitory accounts in this study have been used for a long time and their transactions have similar patterns over the time, then it can be assumed that a model developed with a training set of past data can predict the behavior of transactions in a future data set. The Phase II anomaly detection model is developed based on this assumption.

The training data set was extracted over 11 months from January 2008 through November 2008, whereas the test data set covers the 3 months from December 2009 through February 2010. Any potential effects of this time gap are unknown, so

interpretation of the results should be done with caution. However, considering that each transitory account was created for a specific purpose, its transactions should have similar behaviors unless the business environment changes significantly. As in Phase I, the Phase II data are cleaned by discarding transactions beyond common date ranges. After data cleaning, the transitory accounts for training have 400,466 transactions, and the test set has 75,236 observations. However, the test set has unexpected outcomes for less frequently used accounts. Not all accounts are used intensively, and those with small numbers of transactions (e.g., Accounts 61930 and 94870) have become almost dormant. Since the anomaly detection model is developed and tested for each account, this natural selection will not affect the overall performance of the testing procedure in this study, but those nearly extinct accounts should probably be omitted in the future. Table 7 shows the details for each account.

3.2.2. Screening Rules. Although an indicator like duplicate records is strong evidence of an anomaly, it represents only one anomaly type. Increasing the overall effectiveness of the model calls for indicators that can testify other types of anomalies. To meet this end, five new anomaly indicators, based on various anomaly scenarios and data analyses, are added to those in Phase I. The newly introduced indicators are (1) age of a transaction; (2) transaction initialized on weekends; (3) transaction cleared on weekends; (4) transaction cleared before it is initialized; and (5) duplicate transaction numbers.

Table 7. Transitory Account Details.

Account	Table	cnt	min_amtDate	max_amt Date	range_amt Date
1155	Train	688	01/15/2008	11/20/2008	310
21776	Train	28,744	01/15/2008	11/20/2008	310
21830	Train	23,950	01/15/2008	11/20/2008	310
302	Train	5,654	01/15/2008	11/20/2008	310
32360	Train	67,188	01/15/2008	11/20/2008	310
45136	Train	73,389	01/15/2008	11/20/2008	310
5738	Train	49,539	01/15/2008	11/20/2008	310
58122	Train	20,395	01/15/2008	11/20/2008	310
60836	Train	91,660	01/15/2008	11/20/2008	310
61042	Train	12,114	01/15/2008	11/20/2008	310
61930	Train	1,042	01/15/2008	11/19/2008	309
66613	Train	2,426	01/15/2008	11/20/2008	310
68128	Train	19,568	01/15/2008	11/20/2008	310
70050	Train	2,715	01/15/2008	11/19/2008	309
70068	Train	891	01/15/2008	11/19/2008	309
94870	Train	503	01/15/2008	11/20/2008	310

Account	Table	Cnt	Min_amt Date	max_amt Date	range_amt Date
1155	Test	63	12/22/2009	02/18/2010	58
21776	Test	2,476	12/02/2009	02/23/2010	83
21830	Test	4,676	12/01/2009	02/23/2010	84
302	Test	1,056	12/02/2009	02/23/2010	83
32360	Test	13,652	12/01/2009	02/22/2010	83
45136	Test	14,745	12/01/2009	02/23/2010	84
5738	Test	2,620	12/01/2009	02/23/2010	84
58122	Test	1,620	12/23/2009	02/23/2010	62
60836	Test	24,668	12/23/2009	02/23/2010	62
61042	Test	5,872	12/01/2009	02/23/2010	84
61930	Test	3	01/05/2010	02/18/2010	44
66613	Test	24	12/14/2009	02/22/2010	70
68128	Test	3,586	12/01/2009	02/23/2010	84
70050	Test	78	12/09/2009	02/23/2010	76
70068	Test	97	12/03/2009	02/23/2010	82
94870	Test	–	–	–	–

Age of a transaction: Since a transitory account holds a transaction in transit, an excessively long stay means that there is difficulty finding its destination, which is unusual. Hence, it is important to know how long a transaction stays in a transitory account until it is cleared.

Transactions initialized or cleared on weekends: A lack of sufficient monitoring of an employee's activity is known to facilitate internal fraud and/or produce more careless errors. Compared to weekdays, weekends have fewer observers to check transactions that are either initialized or cleared.

Transaction cleared before it is initialized: A transaction must be initialized before it can be cleared. If events do not occur in that order, it signals an apparent anomaly.

Duplicate transaction numbers: Transaction numbers must be unique in order to differentiate one record from others. Duplicate numbers may cause malfunctions and/or unexpected outcomes in a relational database. It is also possible that one transaction might mask another transaction's activity if they have the same transaction number.

3.2.3. Results and Discussion. The parameters for the Level 2 screening developed by the training data set are applied to the test set. After applying the Level 2 screening, 2,066 transactions remain for the Level 3 screening in which the 9 anomaly indicators are tested with these remaining observations. As a result, 993 transactions are flagged by 1 or more anomaly indicators. The larger number of indicators makes the Venn diagram less useful to show results. Instead, Table 8 presents the Level 3 screening result.

Table 8. Level 3 Screening Details.

Wrong_ Dr Cr	manual	Dup	Aging	amt_ weekends	bal_ weekends	wrong-Date	Dup-Trans-actio-nID	multi-Flags	cnt
0	0	0	0	0	0	0	0	0	1,073
0	0	0	0	0	0	0	0	1	521
0	0	0	0	0	0	0	1	0	2
0	0	0	1	0	0	0	0	0	4
0	0	0	1	0	0	0	0	1	3
0	0	1	0	0	0	0	0	1	365
0	0	1	1	0	0	0	0	1	4
0	1	0	0	0	0	0	0	0	60
0	1	0	0	0	0	0	0	1	30
0	1	0	1	0	0	0	0	0	4

Here, it is difficult to decide the number of flagged transaction for further investigation by internal auditors. Table 8 shows that most of the flagged transactions are identified by one or two indicators; only four transactions are flagged by three anomaly indicators. Since the relative importance of each indicator cannot be measured in an objective way, it is difficult to say which transactions are more likely to be anomalous. Although the criteria in Phase I can be used as a touchstone, they reveal little about transactions flagged by the new indicators. Thus, the only feasible option to decide which transactions need auditor verification could be to consult the internal auditors. However, this remedy is just a temporary fix, so it is necessary to find a more systematic way to tackle this problem. One solution would be a larger set of anomaly indicators for the Level 3 screening. The maximum number of flags found for a transaction is only three, which does not mean that an anomalous transaction has only those characteristics. Instead, it may mean that the anomaly detection model in Phase II does not completely capture anomaly characteristics. However, the expansion of anomaly indicators is not a feasible solution due to the limited number of variables available. For example, some transactions have non-zero balances that are less than the original amounts, so partial account clearance is possible. If this is the case, information about transaction clearance would be useful to create additional anomaly indicators.

3.3. Phase III

3.3.1. Data. In response to the discussion in Phase II, Phase III focuses on expanding the list of anomaly indicators with other attributes that were not used in the previous models. After discussing this issue with the bank's internal auditors, the bank offered to provide sensitive additional data related to the

regularization of transactions, which is the process of making a transaction balance zero by finding its intended destination and transferring the funds from the transitory account to the intended recipient.

As in Phase II, the data sets in the Phase III consist of two parts: one set to train the model and the other set to test the model. After data cleaning, 75,236 transactions remain for the training data and 54,768 transactions for the test data. Details for these two data sets are shown in the Table 9.

3.3.2 Screening Rules. When the newly provided data set is analyzed, three additional anomaly indicators are found to be useful: (1) the number of regularizations; (2) the age of the regularization; and (3) manual regularization.

Number of regularizations: Typically, a transaction is fully cleared with a single process. However, a certain transaction may need more than one regularization process to make its balance zero, which increases the likelihood of a potential anomaly since it can avoid a possible transaction amount check. This scheme to avoid an authorization limit test is similar to splitting the transfer.

The age of the regularization: A regularization process generally takes a day or less. Thus, the time until the first regularization process can be a good indicator to capture anomalous behavior. In fact, it may be more important than the age of a transaction since it indicates the first action taken concerning the transaction.

Table 9. Data Sets for Training and Testing in Phase III.

Account ID	**cnt_train**	**cnt_test**
1155	63	27
21776	2,476	1,322
21830	4,676	3,356
302	1,056	702
32360	13,652	7,615
45136	14,745	9,864
5738	2,620	2,706
58122	1,620	1,650
60836	24,668	23,708
61042	5,872	2,437
61930	3	6
66613	24	13
68128	3,586	1,160
70050	78	75
70068	97	127
Total	75,236	54,768

Manual regularization: Most regularization processes are handled automatically. Since manual processes are more likely to be erroneous or fraudulent than automatic processes, it may be possible to capture anomalous transactions by utilizing the fact that a transaction is processed manually.

With addition of these 3 indicators, the model in the Phase III consists of 12 anomaly testing rules.

3.3.3. Results and Discussion. After applying the Level 2 screening rules with parameters decided by the training data set, 529 transactions in the test data set are selected. Among those, 248 transactions are flagged by 1 or more anomaly indicators, as shown in Table 10.

Based on these results, the 46 transactions with 2 or more flags are selected for further investigation. Considering the number of newly added anomaly indicators, the difference from the Phase II model seems significant. The flags by indicators are summarized in Table 11.

The next step would be for the bank's internal auditors to examine these 46 transactions to determine whether they actually do represent fraud, and that feedback would provide insight into the effectiveness of the model. Although the bank chose not to investigate these transactions, the results of this study highlight

Table 10. Number of Flags for Transactions in Phase III.

Score	**0**	**1**	**2**	**3**	**4**
cnt_transactions	281	202	37	8	1

Table 11. Number of Flags by Indicator in Phase III.

DrCr	0	0	0	0	0	0	0	0	0	0
WrongDate	0	0	0	0	0	0	0	0	0	0
Manual	0	0	0	0	0	0	1	1	1	1
Duplicates	0	0	0	0	1	1	0	0	0	0
Age	0	0	0	1	0	0	0	0	0	0
Amt_weekend	0	0	0	0	0	0	0	0	0	0
Bal_weekend	0	0	0	0	0	0	0	0	0	0
Duplicate transaction ID	0	0	0	0	0	0	0	0	0	0
Multi regul	0	1	1	0	0	0	0	0	0	1
Manual regul	1	1	1	0	0	1	0	1	1	1
Regul_weekend	1	0	1	1	1	0	1	0	1	1
Age to first regul	0	0	0	1	0	0	0	0	0	0
Suspicion Score	2	2	3	3	2	2	2	2	3	4
cnt_transactions	16	4	5	1	7	2	6	2	2	1

a potential problem that may be encountered while developing an anomaly detection model and its feasible remedy. More than 80% of the flagged transactions that violated only one test were excluded although those tests suggest the need for immediate attention. This problem will persist as long as an equal weighting system is used in a rule-based model and the number of anomaly indicators is insufficient. Alternatively, the relative weights of individual indicators can be used to discriminate flagged transactions. However, as discussed previously, it is difficult to measure the relative importance of anomaly indicators in an objective way until more characteristics of anomalous transactions are uncovered.

4. Conclusion, Limitations, and Future Research

In this study, three anomaly detection models are developed to detect anomalous transactions in a bank's transitory accounts. Phase I serves as a pilot study, while Phase II and Phase III are suggested to improve it. Although various attempts have been made to enhance the accuracy and effectiveness of the model, the number of available variables limits the creation of enough anomaly indicators to capture the true characteristics of anomalous transactions. This may be due to the fact that the detection model is a general model that is applied to all transitory accounts.

Future study might consider account-specific models. Some transitory accounts in this study have the similar numbers of transactions and similar distributions, whereas others accounts vary significantly from one another. If an IC screening model includes information on account variability, it may filter the anomalies more accurately. However, the number of models will inevitably increase as the degree of heterogeneity among the transitory accounts becomes large. In the worst case, if all accounts to be examined are distinctively different, a separate screening model would be needed for each one. Thus, the IC screening models might become too complex to be implemented and maintained. Although the bank correspondents mention that there are three types of transitory accounts (finance, consumer, and insurance), these distinctions may not be sufficient to convey the unique features of each transitory account. Direct evidence about account differences can be analyzed from their descriptive statistics. For comparison, assume that 2 numbers are considered to be similar to each other if the difference is less than 10% of the smaller one. For example, the account differences become clear when Accounts 1155 and 61930 from this study are compared. Both accounts have similar date ranges (441 and 446 days, respectively) and a similar number of transactions (694 and 729, respectively). However, their medians are significantly different (1,777.02 and 900,000.00, respectively). Using the general screening rules, many transactions in Account 61930 will not be flagged even if those amounts are beyond the predetermined materiality level since most transactions have large amounts. By contrast, the flagged transactions of Account 1155 may have relatively small amounts that are far below the materiality level. Consequently, a general screening model that does not consider these account-specific characteristics may not work properly unless the characteristics of all the accounts are sufficiently homogeneous.

Considering this potential drawback, the first step to develop account-specific screening models should be to identify and group accounts that have similar characteristics. Grouping criteria could be the date range, the number of distinct days, the number of transactions, their descriptive statistics (mean, median, and standard deviation), and/or the bank's categories for the transitory accounts (finance, insurance, and customer).

Another suggestion is to use relationships among variables. The screening rules in this study are mainly for individual attributes, such as transaction amounts and manual/automatic indicators. Interestingly, there are attributes that are closely related to each other. One example is the association between the transaction amounts and the balance amounts. By definition, the transitory accounts are designated to keep unidentified or insufficiently identified money temporarily, so the account balances should become zero fairly soon. One way to utilize this relationship could be with continuity equations (Kogan, Alles, Vasarhelyi, & Wu, et al., 2011), which assume that the inflow and outflow of a system are the same in equilibrium. Multiple (time series) regression method that includes both numerical and categorical variables could also be developed. Alternatively, clustering methods can be used to divide the transactions into several groups, and then the group(s) with the fewest members can be investigated as potential anomalies.

In the collection of screening rules, the inclusion and exclusion of a rule can easily be processed since each rule is a discrete piece that does not affect the other rules. Thus, if a screening rule is removed, only the transactions identified by that rule will be removed. In addition, if a new screening rule is added, the only change will be additional transactions flagged by the rule. However, in the multiple variable models, the addition and deletion of rules may require a complex process and understanding. For example, if a rule is removed in a network model that is popular in fraud literature, the resultant transactions flagged by the model can be very different from those before the change, and interpreting the differences is unlikely to be intuitive because multiple variable models utilize not only the variables themselves, but also the relationships among them. Consequently, if a variable is deleted, the impact will not only be on the variable itself, but also on its relationships to the other variables. The addition of a variable will have a similar impact.

References

Kim, Y., & Kogan, A. (2014). Development of an anomaly detection model for a bank's transitory account system. *Journal of Information Systems*, *28*(1), 145–165.

Kogan, A., Alles, M. G., Vasarhelyi, M. A., & Wu, J. (2011). Analytical procedures for continuous data level auditing: Continuity equations. *Auditing: A Journal of Practice and Theory*, *33*(4), 221–245.

Chapter 10

Development of an Anomaly Detection Model for an Insurance Company's Wire Transfer System*

Yongbum Kim

1. Introduction

Various data processing systems are implemented in industries to optimize their resource usage and operations. Although new products are frequently introduced to the market, some companies still maintain their old systems, referred to as legacy systems.

A legacy system is a computerized data processing system that is highly customized and system-specific. Although it is believed to be less efficient than cutting-edge database management systems (DBMSs), a well-structured and maintained legacy system may function as well as those currently in the market. However, companies do try to adopt newer computer systems and migrate the existing systems for the long-term proven benefits that the more sophisticated systems can provide. Despite these benefits, migration into a new system is not always an easy task, especially when a company cannot justify the benefits compared to the huge initial investment and maintenance costs. That is why some companies convert their legacy systems into the new systems gradually rather than abruptly, if they convert them at all. This can be the only option for a company that frequently merges with other companies to expand its business.

The insurance company in this study has grown through a succession of mergers and acquisitions, and has adopted a gradual migration to its data processing system. When it acquires another company in the same industry, the acquired company usually has its own data processing system that is not 100% compatible with the parent insurance company even if they have similar business products.

*This chapter is based on the authors' study that is published in the *Journal of Emerging Technologies in Accounting* (Kim & Vasarhelyi, 2012).

Audit Analytics in the Financial Industry, 165–200

doi:10.1108/978-1-78743-085-320191010

In order to resolve this problem, the parent company keeps the system of the acquired company and migrates transactions from the subsidiary's system into its own system. This complicated data processing structure makes it difficult to monitor and control transactions. The parent company has to monitor only the migrated transactions from the feeding systems during the period when it is not feasible to access them directly because of system incompatibility.

Mergers and acquisitions are not uncommon in business. However, gradual system and database integration exposes distinctive barriers due to its inherent drawbacks. As in the bank case in the previous chapter, lack of information about anomalous transactions serves as another obstacle in this study.

Most data fields in the insurance company's system are entered manually because of the complex migration process and lack of internal controls (ICs). In an enterprise resource planning (ERP) system or other highly sophisticated data processing system, most data fields are entered either automatically or semi-automatically as a form of input controls. For example, timestamps and dates may be entered automatically, and product codes may be entered semi-automatically by selecting a value from a menu. This could also be true in a well-designed legacy system. However, that is not the case for this acquired company. Manual entries are more common, and this causes various data integrity problems, such as violations of references or domain integrity. For example, a customer can have more than one name depending on who inputs the customer's name into the system. As a result, many exceptions should actually be considered normal transactions. Although the information about these exceptions will be helpful to improve the firm's IC in the future, data integrity problems remain a big obstacle when building screening rules.

Another obstacle is that the company lacks historical information about its own anomalies including fraudulent scandals. Lack of past experience about anomalies does not necessarily mean that there were no anomalies. Instead, it might indicate that the company's IC system could not detect them or did not have any modules for anomaly detection. This lack of knowledge requires that the anomaly detection model be built from scratch. In this case, two possible assumptions could be made for transactions. Since little is known about the extent of anomalies in the company, it might be assumed that the data are audited and free of irregularities and material errors. As a matter fact, this assumption would be too strong and inappropriate considering that external auditors are not responsible for fraud detection itself, although they are responsible for evaluating the IC system in terms of material misstatements. Consequently, it may be more appropriate to assume that the data contain irregularities and/or material errors.

This study involves a major US insurance company that is proceeding toward developing a continuous audit/fraud detection process. To the end, the company decided that a research team would cooperate with its internal audit organization to develop basic modeling and analysis methodologies in parallel with its internal audit process. The project plan entailed a set of progressive steps in the development of an automated discrepancy detection process. Once the process and model are developed, the data extraction process is made frequently and systematically, moving toward more frequent data screening to monitor for potential fraud. The

model proposed in the project is similar to an external stand-alone system that is used to extract and analyze data for exceptions in continuous auditing (Murthy & Groomer, 2004; Rezaee, Sharbatoghlie, Elam, & McMickle, 2002; Vasarhelyi & Halper, 1991; Woodroof & Searcy, 2001). The wire payment data are extracted from the production legacy information systems and is analyzed externally. This is beneficial as running an automatic fraud detection system can be stressful on the production system, which might cause the system to operate sub-optimally. Pathak, Vidyarthi, and Summers (2005) find that auditing transactions in batches is more cost effective than initiating periodic audits. The model in this study proposes that fraud detection should occur in batches. Before an audit, the internal auditors can extract the desired data and run the fraud detection mechanism. Any resulting exceptions can be investigated during their regular audit. The wire transfer process is chosen as a desirable first target because of: (1) data availability; (2) the volume and importance of the process; (3) the availability of knowledgeable and competent internal audit staff for knowledge engineering; and (4) the timing of the audit.

The wire transfer payment process did not seem as well controlled as other processes in place. Furthermore, the company did not have documented historical information about past fraud occurrences. However, this lack of past experience about the existence of fraud does not necessarily mean that there is no fraud in the wire transfer process. Consequently, it is appropriate that the fraud detection and prevention model should be based on the unsupervised method. The objective is to create a statistical model to detect potential anomalies within wire payment transactions. Internal auditors can then investigate selected transactions for anomalies.

From the perspective of analysis method, the indicators are divided into conditional indicators, which are pass/fail type tests, and statistical indicators, which utilize statistical methods like correlation. Each indicator is equally weighted, although it is likely that certain indicators are more important than the others are. Equal weighting is chosen because the absolute degree of each indicator's effect on the final decision cannot be measured in a systematic, globally agreeable way. Wires with suspicion scores that are higher than a threshold are flagged and forwarded to the internal audit team for a test of details. Investigation results and feedback from the audit team become a direct input to modify the model for fine-tuning.

2. Objectives

A check is a useful payment method in practice. Despite its frequent usage, electronic fund transfers (or wire transfers) are taking over the role of checks in the current business world. Wire transfers have become popular because of their distinctive benefits, such as practical convenience and economics. In a wire transfer, a customer does not need to appear in a bank to fill out a necessary document and the cost to process the transaction is much smaller than for its paper-based counterpart. Less human efforts and less resource consumption result in less cost for a fund transfer, so wire transfers are becoming more attractive than other payment methods.

However, wire transfers have disadvantages too. Unlike traditional payments methods (e.g., checks), a wire transfer leaves little or no physical trace that can be verified. Although audit trails and logs exist in a processing system, access to them requires expertise in DBMSs because they are in electronic forms. As a result, most of the information is highly vulnerable to unauthorized modification if it is not appropriately managed. Thus, having few physical traces is both the main reason for cost savings and a potential source of anomalous transactions, such as fraud and errors.

This apparent drawback is more significant for a cash outflow (wire-out) than a cash inflow (wire-in) because the former decreases a company's cash, whereas the latter increases it. However, wire transfer payments to insurance customers are important. In response to this concern, this study examines how to develop an anomaly detection model for the insurance company's wire transfer payment system.

This anomaly detection model for wire transfer payments is developed in order to identify fraudulent and erroneous transactions. The model consists of a series of anomaly indicators designed to detect wire transfers with abnormal behaviors. Each wire transfer goes through the model and its suspicion score is calculated. If the score is beyond a certain threshold, then it is labeled as potentially anomalous and forwarded to the internal audit team for further investigation. After investigation by the audit team, the model is fine-tuned based on the result and their feedback.

3. Methodology and Results

3.1. Overview

The wire transfer process at the insurance company consists of three stages: initiation, approval, and settlement (or payment). Each wire transfer requires a minimum of one initiator and one approver. Depending on the nature of the transaction, certain types of wire transfers require two approvals when they do not have prior information. Once a wire payment is approved, the wire is imported into the wire transfer payment system as shown in Fig. 1.

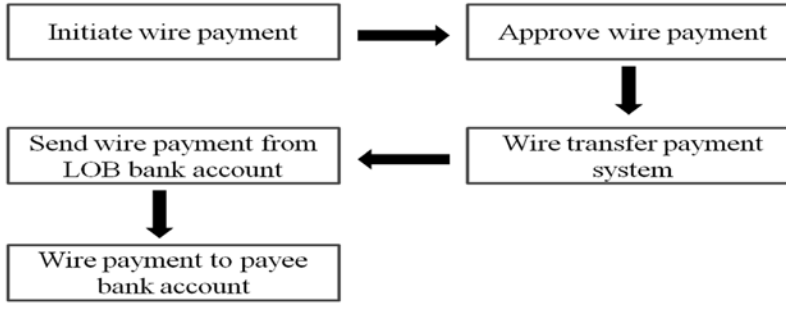


Fig. 1. Steps in Wire Payment.

Although computerized systems exist, the company also maintains all of the relevant physical documents. When a wire transfer needs to be processed, one or more physical documents are filled out to record the customer's information. After the documentation is complete, only part of the customer information is entered into the computerized systems. As a result, certain customer information exists only in a physical form. Although the internal auditors argued that the firm maintained both physical and electronic systems concurrently due to frequent mergers and acquisitions, they did not provide a clear reason for those entries in the physical documents that were required, but not entered into the systems. However, it is not surprising that the DBMS of the acquired company has a different data structure from that of the acquirer. Those differences can be small, but are not always 100% compatible. When this discrepancy is not easily resolved, a solution can be to maintain both systems and merge the two systems gradually. Eventually, there only one system will be left after the process is complete. During this gradual merging process, there may be no systems that can record certain customer's information unless the merged company goes through the expensive process of adopting new forms for existing customers' information. Hence, the use of physical documentation may be necessary to save costs.

Another problem due to frequent mergers and acquisitions is that the same or similar file with the same purpose can exist in multiple systems. Although the parent company's main system governs all sub-systems that the merged companies are running, it is impractical for the main system to have a data structure that can encompass all sub-systems, especially when another merger or acquisition is likely to occur in the near future. Consequently, it is an economical solution to maintain a separate file in each sub-system, even though it is likely to cause problems with data discrepancy.

The most imminent data discrepancy issue is the difficulty in enforcing common input and output controls. This can make the whole system vulnerable to control deficiencies that are represented by data integrity problems related to entity, referential, and domain integrity.

Entity integrity: An entity integrity violation occurs when a specific transaction cannot be identified because of duplicate primary key values. The wire payment data do not have an entity integrity violation because its primary key has unique values for transactions, although the other two data integrity violations are found in the data.

Referential integrity: A referential integrity problem occurs when a key in a referencing table does not appear on its referenced master file. For example, when a customer master file is recorded by the customer service department and the sales department manually enters a customer's name, that customer's name may not exist in the master file. This discrepancy would continue to exist until the two tables are reconciled. If referential integrity is strictly enforced, the sales department is unable to input a customer record for a customer whose name is not in the master file. This can also happen when a referencing system uses multiple names for the same customer. For example, the sales department may enter "Traveler" or "Travelers" in reference to "Travelers Co." in the master file. Thus, a referential integrity violation can take place whenever an information system allows manual inputs and the referential integrity is not completely enforced.

Domain integrity: Domain integrity means that all data in a field must have valid values. Typical examples of domain integrity problems would be skipping required values or entering the wrong type of values. For example, domain integrity is violated if a string of characters is entered into a numeric field, such as "one hundred" for "100". Data may have this problem when historical records are missing or unavailable due to an update. For example, an employee's authorization limit in an employee table will be increased if he/she is promoted and will be set to zero if the person retires or leaves the company. Unless the employee table is designed to keep all historical changes, it is likely to contain only the latest information. If data backups are performed infrequently, some of the past information may not be available.

Another obstacle in this study is that the internal auditors do not have past information about their company's own frauds or material mistakes. Lack of past information leads to a problem in determining vulnerable IC areas. Consequently, an anomaly detection model must be developed from scratch, although some of risky areas can easily be identified, such as amounts of wire payments. Overall, some indicators in the previous case study are used as a starting point in this study, while new screening rules are developed to meet the project's specific goals.

The model development and testing in this study are performed quarterly. Each quarterly cycle has test results and feedbacks from the internal audit team that become input for the next quarter's test. This evolving process entails revision and expansion of the previous quarter's model. As a result, the latest model is the most extensive and sophisticated. However, to understand the developing process, all of the past models and their results will be discussed as well. This process will be particularly beneficial to a company that tries to introduce a new anomaly detection system or revise its current model.

3.2. Phase I (September 2008)

3.2.1. Data. The data set in this study is wire transfer payments made by the insurance company from October 2007 to September 2008, consisting of about 230,000 wire payments to over 10,000 payees. Approximately 90% of the wire transfer payments went to 10.25% of the payees. Over half of the payees (62.82%) engaged in only 1 transaction and almost all of the payees (93.84%) had less than 30 transactions. The data sets provided by the insurance company have seven tables (files). The table primarily used in this study is All_Wires, which has 27 attributes. After removing irrelevant records including monthly amount totals, there are 229,531 transactions. The other six tables are master files that are referenced by the All_Wires table. The master files keep employee information, such as start date, employment status, rank status, and authorization limits. These attributes are used mainly to check employees' authorization limits. Descriptive statistics for four numeric attributes (wire amounts, initialization limits, approval limits, and settlement limits) are shown in Table 1.

There are four types of wire transfers: random, repetitive, concentration, and batch. A random wire needs only one payment (e.g., payment due to a car accident), whereas a repetitive wire requires multiple payments (e.g., pension payments).

Table 1. Descriptive Statistics.

var_name	Wire Amount	Approver Authorization limit	Initiator Authorization Limit	Settler Authorization Limit
N	229,531	8,239	8,239	8,239
Average	4,793,957	167,685,975	80,232,688	606,870
Median	70,242	10,000,000	0	0
Std	79,213,746	452,077,500	325,524,629	24,628,745
Min	0	0	0	0
Max	13,260,787,693	9,814,999,869	7,806,759,586	1,000,000,000

A concentration wire is initiated in the process of fund optimization. For example, if a line of business (LOB) is short of money, funds in other LOBs are transferred to the LOB in need. Batch wires are used for a practical convenience. Each batch wire is a collection of transactions that consist of the three other wire types.

In order to verify domain integrity, frequency checks are performed for each attribute in the All_Wires table. No attributes violate domain integrity. However, missing values were found for twelve attributes. Table 2 shows the attributes and their number of null values.

Some attributes can have null values. For example, the Datasource field in the All_Wires table must have null values if a wire was either random or repetitive because of the nature of random and repetitive wires. However, a missing value for the RountingNum that records the receiver's bank account is clearly anomalous. Since it is practically impossible to determine the reason for the null values, a MissingRNo field is added to the All_Wires table. The field has "N" if the record has the routing number and "Y" otherwise.

3.2.2. Model Development Process. The overall model development consists of five stages based on data mining methods. First, data files and analyses related to anomaly detection are collected from internal auditors. An initial brainstorming session is held with the internal auditors based on their analyses and pilot tests on the collected data to identify potentially risky areas. Next, anomaly indicators are created based on the brainstorming result. This process is laborious because known anomaly instances are rare. Thus, a newly generated anomaly indicator may have a different behavioral pattern than expected. For example, an anomaly indicator might test whether an employee initiates a wire transfer with an unusually small amount compared to his/her other initiated wires. This indicator assumes that wires initiated by an employee have a narrow range of large amounts, whereas they actually have such a wide range that it is impossible to capture a wire transfer with an unusually small amount. After anomaly indicators are generated, the anomaly detection model is tested with the wire transfer data. Each indicator has a different weight with respect to its likelihood to detect an anomaly. Once a particular wire transfer payment has passed through all

Table 2. Attributes.

var_name	Total_cnt	Valid_cnt	Null_cnt
APPROVER1ID	229,531	229,444	87
APPROVER1LOB	229,531	188,709	40,822
APPROVER1NAME	229,531	229,444	87
APPROVER2ID	229,531	59,828	169,703
APPROVER2LOB	229,531	50,784	178,747
APPROVER2NAME	229,531	59,828	169,703
COUNTBIZUNIT	229,531	83,313	146,218
DATASOURCE	229,531	61,173	168,358
INITIATORLOB	229,531	185,061	44,470
REPREF	229,531	169,697	59,834
ROUTINGNUM	229,531	228,055	1,476
TRANREF	229,531	227,141	2,390

indicators and received a score for each violation, an aggregate total is calculated. If the total score for that wire transfer payment is above a given threshold, it is flagged for investigation. Finally, flagged observations are verified by the internal auditors and the verification results are used to update and fine-tune the model. This process is reiterated until a satisfactory model is derived. The notable feature of this model development process is that it is iterative and interactive. The overall process is shown in Fig. 2.

The initial phase of the study involves obtaining a general understanding of the company's wire transfer payment system and the corresponding data. Understanding the system and its ICs is important to facilitate the creation of indicators and algorithms to supplement and support the controls in place. To understand

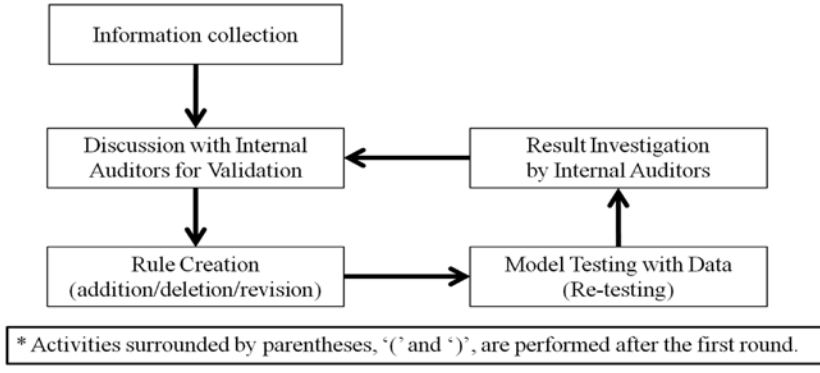


Fig. 2. Model Development Process.

the data, their characteristics and layouts are obtained, along with descriptive statistics, such as average, range (minimum and maximum), and distributions. This gives a general quantitative understanding of the data and the types of wire transfer payments being made. Next the research team and the internal IT audit team brainstorm ideas for meaningful indicators that may identify anomalous transactions. These indicators are transformed into statistical algorithms that utilize statistical data mining techniques.

The indicators consist of three types of statistics: prediction, correlation, and frequency tests. Applying these tests to the data allows anomalies or patterns to be identified. Each indicator is scored based on anomaly risk: 1=low risk; 3=moderate risk; and 5=high risk. Scores are based on the professional judgment of the internal auditors. After running the wire transfers through the various indicators, the suspicion scores are aggregated and used to determine the cutoff/threshold for further investigation by the internal audit team. After completion of their investigation, the internal auditors verify whether the flagged transactions are fraudulent.

In addition, the internal auditors suggest how to improve the model and the indicators. The model should constantly evolve to adapt to new findings. Since fraud is persistent in nature, the fraud detection/prevention process should be run and updated in a continuous manner. The target tests performed by the company are not discussed in this study in order to prevent harming the insurance company's fraud detection efforts.

3.2.3. Screening Rules. Potentially risky areas suggested by the audit team are investigated first: (1) whether the payee transactions payment amount is out of the range of payment amounts; (2) whether the payee transaction payment trend line over time has a positive slope; (3) whether the payee is unusual from a certain sender; (4) whether the initiator/approver transaction payment amount is out of the range of baseline payment amounts; (5) whether the transaction amount is out of the range of normal activity from this bank account; (6) whether the transaction initiator is not a normal sender from this bank account; (7) whether the transaction payee is not a normal receiver from this bank account; and (8) whether a bank account is associated with many other types of transactions. The initial tests to tackle these risky areas are summarized in Table 3.

The first fraud indicator tests for an amount anomaly for each payee. Because of domain integrity problem, the payee names are not a reliable way to identify each payee uniquely. To make the problem worse, because of the lack of a master file that contains payee IDs and their bank account information, it must be assumed that each payee uses only one bank account for wire receipts.

Based on statistical and interpretational constraints, payees are categorized into four types: P1 if the total number of wires=1; P2 if the total number of wires=2; P3 if the total number of wires is between 3 and 29; and P4 if the total number of wires=30 or more. These categories reveal that about 90% wire transfers belong to 10% of payees. In other words, most payees were involved in very few transactions. Specifically, 62.82% of the payees were involved in only one transaction and 93.84% of the payees were involved in less than 30 transactions. Considering that there are 13,145 payees, only 800 of them had more than 30 wire transfers.

Table 3. Initial Tests of Risky Areas.

Potential Fraud Indicators	Possible Screening Rules to Test
The payment amount to a payee is abnormally large or small	Amount range for each payee (or all payees) and check outliers
The payee transaction payment trend line over time has a positive slope	Correlation between date (or sequence numbers) and payee amounts for each payee
The payee is an outlier to payee baseline activity. (Payment sent to a payee that normally does not receive payments)	Payee frequency by each initiator and check the payees that have the lowest frequencies
The initiator/approver transaction payment amount is out of the range of baseline payment amounts	First, check the transaction amounts with their authorization amounts. Second, calculate 90, 95, or 99PI. And then find the transactions that are beyond these bounds
The transaction amount is out of range of normal activity from this bank account	The 90, 95, and 99 PI amounts for each sending/receiving bank account and check the exceptions
The transaction initiator is not a normal sender from this bank account	First, check the list of sender bank account, then create exception lists of initiators by originating bank account
The transaction payee is not a normal receiver from this bank account	A list of payees by sending banks who have least frequency
Access to the bank account is commingled with many other types of transactions	A list of bank accounts with wire types that have the least frequency

3.2.4. Indicators. The anomaly indicators are categorized into two groups based on their analytical approach: target tests and trend tests. A target test is a pass/fail type of anomaly indicator, such as whether an employee approves a wire transfer beyond his/her authorization limit, whether a payee exists on a payee master file, and whether a wire is sent to a country known as a financial safe harbor. Seven target tests, listed in Table 4, are included in the model. The insurance company performs these tests.

By contrast, trend tests utilize statistical methods like prediction intervals, correlation tests, and frequency tests. Measures for these tests are generally continuous, so a threshold must be set to determine whether a wire payment is risky. Twelve anomaly indicators, listed in Table 5, are included in the model.

3.2.5. Prediction Interval Test. The prediction interval test involves stratifying payees into four categories: (1) payees with 1 wire payment; (2) payees with 2 wire payments; (3) payees with 3–29 wire payments; and (4) payees with 30 or more wire payments. Wires are stratified by their group for statistical

Table 4. Target Tests.

Description: Target tests
T1: Payee does not receive payments from more than one initiator
T2: Payee does not receive payments from more than one approver
T3: Initiated transaction date is after the initiator's termination date or before hire date
T4: Approved transaction date is after the approver 1's termination date or before hire date
T5: Approved transaction date is after the approver 2's termination date or before hire date
T6: Receiver is located in a country known as financial safe harbors
T7: Multiple payments on the same day in the aggregate exceeds approvers limit for payment to a single payee

Table 5. Trend Tests.

Description: Trend tests
A: The payee transactions payment amount is out of the range of payment amounts
B: The payee transaction payment trend line over time has a positive slope
C: The payee is an outlier to payee baseline activity. (Send to a payee that normally do not send to)
C1: New initiator
C2: New approver 1
C3: Potential collusion
D: Transaction payment amount is out of the range of baseline payment amounts
D1: Initiator's
D2: Approver 1's
E: The transaction amount is out of range of normal activity from the sending bank account
F: The transaction initiator/approver is not a normal sender from the sender's bank account
F1: Initiator
F2: Approver 1
G: The transaction payee is not a normal receiver from this bank account
H: Access to the bank account is commingled with all types of transactions.

interpretations. The stratification process is made for payees, initiators, and approvers. In addition, alternative prediction intervals of 90%, 95%, 99% are also considered. A higher confidence level will have fewer outliers; conversely, a lower confidence level will result in more outliers. For payees with only one wire payment, a prediction interval is estimated by grouping a given payee's wire payment together with other payees who have one wire payment in order to determine which payments are abnormal compared to the group. The prediction interval is also applied to payees with thirty or more wires. For payees who have only 2 wire payments and for payees with 3–29 payments, clustering may be useful to detect outliers.

3.2.6. Correlation Test. The correlation test examines whether a payee's payments increased in a consistent manner. Activity monitoring (Fawcett & Provost, 1999) is adopted for this type of test. It requires maintaining a usage profile for each payee or employee in order to identify any deviation in activity. Unlike the prediction interval test, the correlation calculation requires at least three observations, so the wire transfers are stratified into two groups: those with more than two wire payments, and those with one or two wires. The statistical significance of the degree of overall increase in wire amounts is determined by the correlation value and its p-value. In the literature, various correlation values are suggested to decide whether observations are positively correlated. Although global standards do not exist for strength of correlation, coefficients between 0.3 and 0.7 are generally considered to indicate moderate correlation, so a threshold 0.5 is used in this study as a cutoff that is conservative, but not too strict.

3.2.7. Frequency Test. Based on the definition of typical or normal activity patterns, wire payments that are anomalies or outliers can be determined. Frequency tests can help define what should be considered typical or normal activity patterns since infrequent activities may indicate potential errors or fraud. A frequency test indicator entails examining each payee and the employee initiating the wire payment to determine which pairs have unusual activity.

For example, an unusual activity could be a payee interacting with an uncommon employee or group of employees for the first time. A pilot test in this study shows that a payee typically encounters many different initiators and approvers in the company. Consequently, encounters with only one initiator or approver is considered abnormal.

3.2.8. Scoring System. The scoring of each indicator is developed with the assistance of the internal audit team. The knowledge engineering of experienced professionals (Vasarhelyi & Halper, 1991) helps to determine whether indicators are considered abnormal or potentially fraudulent in nature. Furthermore, an effective internal audit team may have the ability to identify indicators that suggest fraud (AICPA, 2002). In response to these arguments, each indicator is assigned a score based on perceived risk. However, it is not an easy task to measure the relative importance of anomaly indicators, especially when their effects are close, but clearly different. Weight assignment becomes even more challenging as the number of anomaly indicators increases in response to the dynamic nature of model development.

After each indicator is processed through the statistical algorithms, a violation total for each wire transfer is computed. A wire transfer that violates more than a

designated threshold is subject to investigation by the internal auditors. However, it is difficult for internal auditors to allocate large amounts of time to investigate exceptions in practice because of the cost barrier. In running the initial model, enormous numbers of exceptions are found. The cognitive effect of information overload is not a trivial issue (Kogan, Sudit, & Vasarhelyi, 1999). An overload of alarms will have a negative effect on the internal audit department's willingness to adopt an anomaly detection system.

As a result, the threshold scores must be increased in order to reduce the number of flagged wire transfers. The summary statistics for the aggregated scores is shown in Table 6.

The thresholds assigned to determine the flagged wire transfers are 10 for the trends tests, 25 for the target tests, and 20 for total tests. These thresholds produce 106 flagged wires, which the internal audit team deems information overload. Therefore, the number of exceptions is reduced to 47 by increasing the levels of thresholds: 11 for the trends tests, 25 for the target tests, and 22 for total tests. This is a more reasonable number of flagged transactions for internal auditors to investigate. Table 7 shows some of the flagged wires with fictitious numbers.

3.2.9. Results and Discussion. *Results:* The internal audit team investigates the 47 wire transfer payments during their regular audit, and no evidence is found that these wires are either fraudulent or erroneous. The investigation results show that most of the wires are flagged when they are the only payment to a payee. This occurs because a wire transfer violates two target tests and three trend tests when a payee has only one payment. In addition, the wires beyond the target test thresholds are intercompany transactions. For reasons that the internal audit team cannot identify, those wires violate five target tests. This finding implies that the weighting system needs changes to reduce the effect of a single violation on the overall weighting score. Although the anomaly detection model does not find anomalous wire transfers, this does not mean that there are no anomalous payments. Instead, it may indicate the need to revise and fine-tune the model. The company intends to include the fraud detection process as a part of its regular audit, retaining these indicators for future detection or preventive measures. In addition, the company is interested in refining the indicators and adding new ones to screen for anomalies. In fact, the company should consistently reevaluate and revise the model, considering the highly adaptive nature of fraud perpetuators.

IC issues: During this study, the effectiveness of the company's ICs comes into question. Three main issues emerges: (1) Certain controls meant to segregate the duties of employees are violated; (2) Terminated employees remain able to process payments; and (3) Wire payment limits are circumvented because employees with even $0 authorization limits are able to process wire payments. These major IC issues are brought to the attention of the internal audit department and are investigated. The internal auditors find that there are inconsistencies between the wire transfer payment process records and human resource records. This discrepancy is caused by the company keeping only the most recent information. For example, a terminated employee may have been an active employee when he/she initiated or approved a wire transfer. Although these IC violations appear to be potential fraud indicators, investigation of their nature and frequency suggests

Table 6. Summary Statistics for Aggregate Scores.

score_trend	cnt_wires
0	183,534
1	25,933
2	3,141
3	5,179
4	1,271
5	266
6	1,005
7	8,217
8	707
9	209
10	58
11	9
12	2

Score_target	cnt_wires
0	195,948
5	18,841
10	14,401
15	334
25	7

Score_total	cnt_wires	Score_total	cnt_wires
0	163,304	13	67
1	23,204	14	18
2	2,437	15	139
3	3,562	16	187
4	964	17	7,106
5	14,962	18	538
6	2,652	19	136
7	1,490	20	41
8	1,717	21	7
9	361	22	27
10	5,406	23	2
11	897	24	1
12	299	25	6
		32	1

Table 7. Flagged Wires.

ID	Score_trend	Score_target	Score_total	Amount
3259892	11	10	21	989,343,618.35
3278185	0	25	25	150,000.00
3296478	7	15	22	4,473.33
3314771	11	10	21	135,350,000.00
3333064	7	15	22	25,657.09
3351357	7	15	22	53,077,500.00
3369650	11	10	21	1,235,418.75

that these violations are due to poor database management. However, this does not exclude a chance of fraudulent activity.

Scoring system issues: During the investigation, the internal audit department notes that some wires were flagged because of systematic causes that were mainly attributed to the target tests. Each of the target test indicators was scored at a high risk level. As a result, a flag raised by a few of these indicators will most likely hit the threshold for investigation. This may suggest that an equally weighted scoring system may be more useful as a starting point than an unequally weighted one. However, some indicators are clearly more important than others. As long as the indicators are subjectively assigned weights, this issue may return. Further deliberation will be necessary to find less subjective weighing methods for the indicators. This finding further illustrates that any fraud system must be developed and updated continuously as new flaws with the current system surface.

Discussion: This study provides a pilot investigation for anomaly detection at a major US insurance company. Although the literature discusses numerous methods to detect fraud, few studies use real company data. In this study, a data mining approach is used to detect anomalous wire transfers, based on statistical algorithms created to detect data abnormalities. Since much of the prior research uses complex methods, such as neural networks and clustering, to detect anomalous transactions, the use of simple statistics, such as prediction interval, frequency test, and correlation test, may seem trivial. However, it is not uncommon for simple methods to be as robust and powerful as more sophisticated methods, and sometimes more accurate. This study leaves out data analysis for payees with between two and twenty-nine wire payments due to lack of statistical significance. Future research can look into implementing other types of statistical methods, such as clustering, to detect abnormal or patterned activity. The company plans to pursue data mining further in a continuous effort to detect fraud.

This study provides a learning experience for academics by showing how an anomaly detection and prevention model is implemented. In addition, this study shows that internal auditors can run anomaly detection and prevention activities on a frequent basis instead of during an annual audit. Two issues require further consideration: (1) Highly subjective weighting of most indicators yields

extreme numbers of violations; and (2) As the pilot study progresses, some indicators may need to be adjusted because of an increasing understanding of data characteristics.

3.3. Phase II

3.3.1. Data. After the Phase I study, the insurance company decided to include the anomaly detection model in its regular audit. However, the Phase II data sets provided by the audit team have several changes compared to the previous ones. The most distinctive difference is a major DBMS update made since the previous audit. The changes include data structure changes, data reformat and conversion, and data migration. As a result, the new data set has fewer transactions over a longer time period. In addition, there are five newly included variables: (1) status of the wire; (2) bank confirmation code; (3) Office of Foreign Assets Control (OFAC) status; (4) payee address; and (5) a comment that is similar to an existing reference variable. With the addition of these 5 new variables, the total number of variables becomes 32. Since a majority of the data variables in Phase II are not affected by this change, the model from Phase I is used as a starting point to develop the new model.

As in Phase I, the data set in this study is wire transfer payments made by the insurance company. The data span 18 months, from January 2008 through June 2009. After 28 irrelevant records are excluded, the All_Wires table consists of 201,476 wire payments with 32 variables. The other six tables have the same data structure as in Phase I. Approximately 90% of these wire transfer payments belong to about 26.5% of the payees compared to 10.25% in Phase I, implying that each major payee has fewer transactions during this period. The proportion of payees engaged in only one transaction (57.82% in Phase II compared to 62.82% in Phase I) and the proportion of payees with less than thirty transactions (92.59% in Phase II compared to 93.84% in Phase I) are very similar.

3.3.2. Model Development Process. The anomaly detection model in Phase II is based on the five-stage model used in Phase I. The main goals of Phase II are (1) revision of indicator weights for better scoring; (2) addition of new indicators; (3) materiality considerations; (4) inputs into the current quarterly audit; and (5) reclassification of tests.

Revision of indicator weights: In Phase I, a few target tests have too much weight, and the effects of the other indicators are relatively ignored. Although the use of a relative weighting system is desirable because of the differentiated importance of the indicators, it is nearly impossible to create a consensus weight system due to the subjective nature of "importance". It is well known in research that quality cannot easily be measured by quantity. As long as the degree of risk is qualitative, the resulting weights will be relative, not absolute. However, relative weights represented by numbers are subject to change whenever new indicators are added. In order to minimize the number of false alarms due to the weighting system, the weights of several indicators are changed. As a result, three target tests that had 5-point weights in Phase I are merged into trend tests. For example, an indicator examining whether a payee receives wires from only one initiator is

merged into a broader trend indicator examining whether a payee has a new initiator because the only time that payees can have only one initiator is when they are new customers. This new indicator is considered less risky and is assigned three points if violated.

Addition of new indicator: After discussion with the internal audit team, 18 anomaly indicators are added for validation. The validation process is indispensable to identify IC exceptions. New indicators examine aging, potential collusion, segregation of duty, split wire, referential integrity, process day, invalid variable value, proper approval, OFAC process, and similarity by clustering. These new indicators are summarized in Table 8.

Materiality considerations: Wire transfers with small dollar amounts are excluded in the verification process. A good detection model should be powerful (few false negatives) and efficient (few false positives). A model's power is its ability to detect anomalies accurately, but this quality often results in a large number of false positives. Although an accurate model is more desirable than an efficient one, the verification process is costly, laborious, and/or time-consuming, so a company cannot afford to expend additional resource on excessive false positives to check their abnormality. Although it would be idealistic to examine all flagged wire transfers regardless of the number of false positives, practicality must be considered while developing the detection model. To reduce the number of flagged wires, the most common discriminating criterion is their amounts. In a regular audit, a small error can ignored, based on the assumption that it does not have a material effect. Following this approach, anomalies with amounts less than $2,000 are excluded since this is the materiality threshold used in the company's regular audits.

Input for the quarterly audit: Outcomes of anomaly detection become part of the company's quarterly audit. As a result, feedback from the verification process becomes more timely and model development becomes more practical. Since the wire transfers in Phases I were audited and no anomalous wire transfers were found, all wire transfers before the current audit period (i.e., April, May, and June) will be considered free of anomaly. Therefore, wire screening will be applied only to the latest quarter.

Reclassification of tests: The target tests in Phase I were conducted by the internal audit team. However, this reduces the consistency of model development in two ways. First, the data and programs related to target tests were not made available, so their verification was not possible. Second, the overlapping of two pairs of trend tests and target tests made model development less efficient. To make it worse, the outcomes of those tests were somewhat different. In order to eliminate these inconsistencies found in Phase I, the indicators in Phase II are classified into trend tests and control tests, while excluding three target tests because of data unavailability. In Phase II, a test is considered a control test if it is a pass/fail or yes/no type and a trend test otherwise. This reclassification helps to build a framework of detection model development. Fifteen indicators relate to trend tests, and five indicators are classified as control tests. Furthermore, the 15 trend tests are sub-divided into 24 indicators, and the 5 control tests have 6 sub-indicators.

Table 8. New Trend Tests and Control Tests.

Trend tests	
I	Date between the initiated date and the fund disbursal date are abnormally far (>20 days)
L	Payee does not receive payments from more than one initiator/approver 1 combination. (This does not include any "switches.")
	L1: (A, B) --> Unique count (A,B) = 1
	L2: (A, B) & (B, A) --> 1) the same as collusion but more than 2 wires, 2) Switch among <L.1>
M	An initiator/approver1 has only one payee (opposite direction to the two target tests: a payee has only one initiator/approver 1)
	M1: Initiator
	M2: Approver 1
N	Switches of initiator and approvers
	N1: Across payees
	N2: Within each payee
O	SplitWireTest for initiators: This is not possible because of multiple authorization limits. Instead, existence test is performed
P	Below low bounds but having large amounts (LPL01 and ≥$2,000)
	P1: Initiator
	P2: Approver1
	P3: Sender
T	Clustering
Control tests	
J	Wires initiated/disbursed on weekends or holidays (Saturday/Sunday/ Holiday)
	J1: Initiated date
	J2: Disbursed date
K	Violation of Referential Integrity: Repetative wires = = > This belongs to a trend test because of data incompleteness. (some repetative wireIDs do not appear on the master file.)
Q	Missing routing number
R	OFAC status = 'F' (Failed) --> Approval Status = 'A' (Approved)
S	No approvers

3.3.3. Screening Rules. The Phase I results and feedback are used as the starting point in Phase II. Wire verification results from the audit team and discussion of new variables due to the recent database update suggest new areas of potential risk. The newly added anomaly indicators examine those risky areas: (1) whether the wire transfer process takes an unusually long time (i.e., over 20 days) (2) whether a payee has an unusual connection with an initiator or an approver (i.e., a payee receives wires from only one initiator or approver); (3) whether a payee, an initiator, and an approver have a close connection (i.e., a payee receives wires from only initiator-approver pair); (4) whether an initiator and an approver have a close connection (i.e., possible collusion) by switching their roles; (5) whether a wire transfer beyond the authorization limits is initiated or approved through splitting; (6) whether a wire with an unusually small amount is processed; (7) whether a wire is processed on non-work days (i.e., weekends and holidays); (8) whether a repeating wire transfer does not have a record in its master file; (9) whether a wire does not have a payee's routing number that identifies his/her destination account; (10) whether a wire against an OFAC status check is processed; and (11) whether a wire is processed without proper approval.

3.3.4. Indicators. As mentioned in the previous section, the anomaly indicators are separated into two groups based on their analysis approach: control tests and trend tests. Like the target tests in Phase I, control tests are pass/fail or yes/no types of anomaly indicators that examine the existence of any violations against the company's operational control policies. For example, a wire transfer violating the OFAC policy should not be approved even if it is initiated. Accordingly, a wire transfer is anomalous if it fails the OFAC test, but is approved. Another example is a wire transfer processed on a non-work day, such as a weekend or holiday. The company is open on weekdays and closed on weekends and holidays, so a wire transfer on a non-work day is clearly anomalous. In addition, non-work days are vulnerable to internal fraud because the wire transfer is performed without being monitored by other employees. Six anomaly indicators are used for control tests in Phase II.

As in Phase I, trend tests utilize statistical methods and include tests that are not control tests. Twelve trend tests are newly included in Phase II. One exception is a split-wire test. Although it should be considered a control test, it is classified as a trend test because the answer is not dichotomous. Multiple authorization limits are assigned to some employees due either to incorrect data extraction or improper database management. Since the cause of multiple authorization limits is undetected by both the internal audit team and the IT team, conservatively, the smallest authorization limits are applied to this test. Hence, some wire transfers that are considered to violate the split-wire test may actually represent multiple authorization violations. Twenty-four indicators are used for trend tests in Phase II.

3.3.5. Prediction Interval Test. All of the anomaly indicators from Phase I that use prediction intervals are included in Phase II and one new indicator is added. The newly added indicator examines whether a wire transfer has an unusually low amount. This indicator is designed to detect an internal fraud in which small amounts of money are embezzled.

3.3.6. Correlation Test. No changes are made to the anomaly indicators that apply correlation tests.

3.3.7. Frequency Test. This type of anomaly indicator is extensively used in Phase II. An anomalous activity is a rare event, so a frequency test can be the best test for this type of anomaly detection. In addition to the seven indicators from Phase I, eight new indicators are implemented in Phase II. For example, one test examines whether an initiator and an approver switch their roles for a payee. Assume that John and Jack are employees of the company, and Jane is a payee. John initiates a wire transfer to Jane and Jack approves it. Later, Jack initiates another wire transfer to Jane and John approves. This practice clearly violates segregation of duty and shows a possible collusion risk related to all three participants. However, this sequence of events might happen by mistake or in error if the information system allows it, although it is not a good practice. Taken together, the indicator implies potential collusion, rather than offering direct evidence of collusion.

3.3.8. Scoring System. During Phase I, it was found that the weighting system produced unexpected outcomes. For example, five points were assigned to each target test in Phase I. The relatively high weights on the target tests resulted in false positives that were caused by an unknown systematic problem. In addition, the over-emphasis on the target tests made the trend tests relatively less important, so their effects on anomaly detection were minimized.

This unexpected outcome resulted from the effects of subjectively assigned weights on the anomaly indicators, which over-weighted the target tests. It is difficult, if not impossible, to quantify anomaly risk levels in an objective measure because risk is qualitative in nature. One example from Phase I is a target test examining whether a payee has only one approver. Although the test assumes that it would be risky for a payee's transactions to be processed by only one approver, it produced an unexpected outcome. If a payee is new to wire transfers, the indicator flags the transaction immediately. However, the risk related to the wire transfer for a new customer is lower than that of a wire transfer whose payee has many other wire transfers. This example illustrates that the weight on this indicator must be revised.

To mitigate this problem, four of the seven target tests are reclassified as trend tests with weight changes, and the remaining three target tests are discarded because of data deficiency. For example, the indicator for identifying wire transfers with only one approver is divided into two cases. If a payee is a new customer with only one transaction, one point is assigned (low risk), whereas if the payee has more than one transaction, three points are assigned (medium risk). In addition, newly added trends tests are given an equal weight of one point except for the trend test that examines whether employees switch their roles as an initiator and an approver, which is assigned three points. The reason to assign equal weights to new indicators is that their risk levels are unknown and difficult to measure. If more information related to their inherent risks becomes known, their weights should be changed accordingly.

The problem of subjectively weighting indicators may not be resolved completely even with those changes. However, it suggests that the impact and

undesirable consequences related to the weighting system should be considered with more caution.

Limited human resources represent a great obstacle to the verification process in Phase II as it was in Phase I. The number of wire transfers that the audit team can reasonably investigate is about thirty. With this practical restriction, the thresholds for the two types of tests are carefully chosen to produce the maximum number of flagged wires. As a result, the chosen thresholds are nine or higher for the trend tests, and three or higher for the control tests if they have routing numbers and approvers, but one or higher if they do not have routing numbers and approvers. Since problems related to missing approvers (215 wires) and missing routing numbers (214 wires) are more than expected, they are reported to the audit team separately for further investigation. In this phase, a threshold for total score is not used because the number of control tests is relatively small (six tests) compared to the trend tests, and only a few wires violate these tests. Hence, computing the total score is not of great use. In addition, the control tests are defined as events that must not happen if related controls are in place. Therefore, a wire that violates even a single control test is worthy of investigation. After applying the screening thresholds, 26 wire transfers are selected and sent to the audit team for verification.

3.3.9. Results and Discussion. *Results:* Although 26 wire transfers are recommended for verification, three are discarded by the audit team due to the $2,000 materiality threshold. Two of these wires are for $0.01 and one is for $1,572.78. The audit team considers that those amounts are negligible in audit practice.

The internal audit team then investigates the remaining 23 wire transfer payments as part of their regular audit and finds no evidence to supports that these wires are either fraudulent or erroneous. However, the investigation result shows that these wire transfers have three features. First, some of these wires are sent to tax authorities, so their payees have no personal interest in them. Second, the wires processed on non-work days occurred because the related employees worked on those days because of their personal holiday schedules. Finally, some of the flagged wires were sent to the company's subsidiaries. The audit team argues that the money is traceable since it is still inside the company, so it does not bear any risks related to internal fraud. Despite their strong claim, it is not clear that all internal accounts are properly controlled. Although management fraud might still be related to these transactions, it is beyond the scope in this study. If it can be assumed that internal transfers are free of internal fraud, as the internal auditors state, it may be possible to exclude the concentration wires (i.e., internal wire transfers to optimize the company's fund usage). This exclusion will be applied in the next phase.

Overall, the audit team does not find any evidence to support the existence of any anomalies. However, this does not mean that all of the wire transfers in the quarter from April through June are free of anomaly. Instead, it may mean that the current detection model is not powerful enough to catch anomalous wire transfers or that the indicator weights are not properly measured. In either case, the detection model still has a room for improvement. Newly emerging issues in this phase are summarized below.

No approvers: A wire transfer process consists of three steps: initiation, approval, and disbursement. Consequently, a wire transfer must have an initiator and at least one approver. Among these processes, the approval requires an approver ID if the transfer is manually processed or a preset value for the approver ID if it is automatically processed. According to this policy, the approver ID field must have a value and cannot be empty. However, it is found that some wire transfers have null values for the field. After investigating the cause of missing entries, the IT team finds that those wires are from the pension system, so their approvals exist in the administrative system, but are not properly carried over to the wire transfer system. However, the IT team fails to resolve why this problem occurs. The DBMS may not be managed seamlessly, or mistakes may occur during the data extraction process. Regardless, this indicator may need to exist until its cause is clearly identified and resolved.

Multiple authorization limits: An employee can be assigned to more than one LOB in this company and the assigned LOBs generally have multiple wire types. Since authorization limits should relate only to the employee's rank, they also relate to the employee's LOB. Consequently, it is possible for an employee to have multiple authorization limits for the same type of wire transfers in the company. In fact, 203 out of 418 initiators in this data set have this problem. In the audit team's opinion, it is more appropriate for an employee to have only one authorization limit for a given wire type. Therefore, this situation may imply that employee authorizations are loosely controlled, and stronger controls should be implemented. Until the problem is resolved, conservatively, the authorization limit check will use the lowest authorization number.

Referential integrity: Referential integrity requires that an attribute value referenced by a table must exist in the table that it references. This property is one of the most fundamental concepts in a relational database system. If a table in a relational database system violates referential integrity, it can result in a catastrophe that destroys the whole database. In this study, the All_Wires table references other tables for further details. Among those tables, the Templates table is a master file that is referenced by repetitive wires in the All_Wires table. Consequently, if a repetitive wire record exists on the All_Wires table, it must exist on the Templates table as well. However, repetitive wires do not appear on the Templates table in some cases, although they belong to the All_Wires table. For example, a repetitive wire (ID=5549) does not exist on the Templates table even though it is in the All_Wires table. Since this violation can seriously damage the whole database system, it calls for immediate action to investigate its cause. Although some possible causes are incorrect data extraction and improper table matching, an investigation by the IT team shows that the true reason is the least expected one. According to the IT team, the All_Wires table violates referential integrity because records in the Templates table can be deleted even though they are referenced by the All_Wires table. In a relational database system, a record on a master file should be deleted only after all referencing records are deleted in order to safeguard the entire database. On the contrary, the company does not enforce referential integrity strictly, which may deteriorate its database in the future. The IT team must consider this problem seriously to prevent possible disasters.

Missing routing numbers: A payee can be identified by payee name, payee address, or a bank account. In order to distinguish an individual payee, a unique identifier, called a key, is necessary. Since a given payee name can mean more than one person, a payee name cannot be used as a key. Although a payee address could be a candidate key, it is prone to too many errors to be used in this study. The address field is manually entered, so that it is entirely up to an employee what is entered into the system. For example, a payee address "161 Washington Street, Newark, New Jersey" might be recorded as "161 Washington St.," or "161 Washington Street, Newark". Although they seem to indicate the same payee, it is difficult to determine whether they are actually the same. Hence, the most appropriate candidate key is a payee's bank account, which consists of a routing number and an account number. Since the two components are recorded separately, it is important that both variables have valid values. The audit team confirms that these two variables must have values to identify the destination of a wire transfer. However, there are cases in which the routing number has a null value, meaning that further information is necessary to verify that the funds are truly sent to the intended recipient. After investigating a cause of missing routing numbers, the IT team finds that a mapping table that is not currently available might be used to relate a certain wire transfer to its intended bank account. However, they fail to explain why only some wire transfers have this problem or to provide the mapping table. This phenomenon should be considered as an anomaly until the mapping table is provided and proven to be error-free.

Discussion: Based on the pilot study in Phase I, the study in Phase II presents the development process and results of the second-generation model for anomaly detection. Major features of the Phase II model include: (1) eighteen new indicators among which, four indicators are conversions of the target tests in Phase I; (2) revision of indicator weights due to unexpected outcomes in Phase I to reflect their potential effects on anomaly risk; and (3) identification of newly found problems with missing routing numbers, lack of approvals, referential integrity violations, and multiple authorization limits.

Although some flagged wire transfers seem highly suspicious, investigation by the audit team does not find any evidence of anomaly. However, some of the newly raised issues need further investigation and evidence. For example, the missing routing number problem will be resolved only when the mapping table becomes available.

This study provides a learning experience about the second-generation anomaly development process. The feedback from Phase I facilitates the model development in Phase II by providing direction and details. Knowledge about anomaly detection will continue to be accumulated as the development process continues. Phase III will consider the newly found problems, revise the factors that need fine-tuning, and modify the raw data set to narrow the scope of wire transfers.

3.4. Phase III

3.4.1. Data. Although the models in Phase I and Phase II identified many problems during model development and testing, they need to be improved for

better detection power. After components of wire transfers are examined in detail, rebuilding the categorizations of anomaly indicators is suggested as a way to improve the anomaly detection model.

The categorization of anomaly indicators plays an important role in developing and evaluating anomaly indicators and screening wire transfers for anomalies. Indicator categorization facilitates finding missed risky areas and discovering potentially anomalous wire profiles. Although categorization is useful and important, it was not systematically developed in Phases I and II. Hence, it is reasonable to arrange anomaly indicators in a logical way to improve the quality of the anomaly detection model. This requires that the screening process be changed.

The data set in Phase III is an undated expansion of the Phase II data set. Wire transfers from July through September 2009 are added to Phase II, and the master files, such as employee records, are updated to reflect any changes during the period. With these changes, the entire data set spans over 21 months from January 2008 through September 2009, and consists of 260,762 wire payments. After excluding 40 summary observations and 1,239 rejected wires, 259,483 wire transfers remain in the data set.

3.4.2. Model Development Process. Phase III starts with feedback from Phase II and discussion with the audit team. The main goals of Phase III are: (1) revising indicator categorization for more reasonable classification; (2) adding new indicators; and (3) comparing flagged wires based on the old and new categories.

First, the anomaly indicator categorization is reconstructed in Phase III. Although minor changes were made to the indicator classification in Phase II, a more systematic approach is needed for its development in Phase III. New categories divide the anomaly indicators based on how rigid their discrimination standards are. That is, yes/no or true/false types of anomaly indicators are more rigid thresholds than statistical types since no cutoff values are needed for the parameters. For example, an answer to a question that examines whether a wire transfer is initiated on a non-work days is either "yes" or "no," so no parameter values need to be decided in advance. In addition, the new categorization assigns equal weights to the anomaly indicators. Thus, the weights of anomaly indicators are controlled at the aggregate level, rather than at the individual level. This change facilitates anomaly indicator development and weight assignment. As discussed in Phase II, weights assigned to individual anomaly indicators are subject to changes as their effects on anomaly detection become better understood.

Second, new anomaly indicators are developed based on feedback from Phase II and periodic discussion with the audit team on various hypotheses that are derived from data analyses. One example is expansion of the split-wire test. In Phase II, the split-wire test was performed in terms of initiators. Since wire splitting is one of the most common internal fraud methods, the test is expanded to include approvers.

Finally, new indicator categorization is examined by comparing the flagged wire transfers according to the existing categories with those flagged by the new categories. One concern about the newly suggested categories is how to weight each type. Although the yes/no type of anomaly indicators seem to have more

significant effects than statistical tests, it is difficult to determine degree of the gap between them. This issue is discussed in the section on the new scoring system.

3.4.3. Screening Rules. The anomaly detection model in Phase III starts with anomaly indicators and feedback from Phase II. In addition to the anomaly indicators used in Phase II, two types of anomaly indicators are expanded, one type is newly added, and one indicator is dropped. First, the split-wire test is expanded to consider a wider variety of cases. In addition to initiators examined in Phase II, Phase III also considers approvers and examines them both by wire type. Second, clustering indicators identify anomalous wire transfers using three hierarchical clustering methods: flexible beta clustering, two-stage density linkage, and Ward's clustering. For these clustering methods, anomalous wire transfers are defined as the observations in the smallest clusters that represent unusual or infrequent events. Since the smallest cluster is subject to change, depending on the method used, the clustering results may not give concrete evidence on anomalies. Third, an indicator examining segregation of duty is newly added in Phase III. Segregation of duty is an essential component of IC. Its targets are relationships between the initiator and approver 1 and between approver 1 and approver 2. Finally, an indicator to identify wire transfers without approvers is dropped because insufficient information about acceptable values prevents its implementation. Until relevant information becomes available, this test will be discontinued.

3.4.4. Indicators. Indicator categorization is supposed to facilitate development and management of anomaly indicators. Although some changes are made to indicator categorization in Phase II, it is still not completely systematic. Since a systematic framework can help to find risky areas and to manage existing anomaly indicators, it is crucial to have a well-developed anomaly indicator classification. Another benefit of indicator categorization is that a scoring system can be more easily controlled. If indicators are categorized by relative risk, it will be easier to interpret the meaning of a wire transfer's suspicion score.

Anomaly indicators are divided into trend tests and control tests in Phase II. However, those category names are somewhat misleading because not all trend tests actually test trends. To avoid this confusion, new categories separate anomaly indicators into statistical tests and conditional tests. The statistical tests utilize statistical methods, such as prediction intervals, correlations, and clustering, whereas the conditional tests use frequencies or yes/no questions to screen wire transfers. While statistical tests require the user to determine various parameter values (e.g., significance level) in order to decide whether a wire transfer is anomalous, conditional tests do not need this procedure. For this study, the most commonly used parameter values are chosen, but the effect of that choice on anomaly detection is unknown. Since statistical tests need more human intervention, these tests are more affected by the user's choices. Based on these considerations, conditional tests are assumed to be more powerful than statistical tests.

Since little is known about the new categorization, anomaly indicators in Phase III are categorized using both the old and new categorizations. Then, the outcomes of the two classifications are compared to identify the similarities and differences between them. Categorization based on the old framework is the same

as in Phase II except that the split-wire test and the segregation of duty test now belong to control tests. However, categorization in the new scheme is much easier than the old method. If an indicator uses a frequency test or yes/no question, it belongs to conditional tests. Otherwise, it belongs to statistical tests.

New split-wire tests are developed with yes/no type of questions. One point is assigned to a wire split that initiates or approves a wire transfer by avoiding the employee's authorization limit. Similarly, the test of segregation of duty assigns one point if a wire transfer violates the rule.

3.4.5. Scoring System. The anomaly scoring system in this study is highly dependent on indicator categorization. When wire transfers are screened, thresholds are applied to the sub-total of each category and the total suspicion score. In addition, unequal weights are assigned in the old categorization of trend tests and control tests, whereas equal weights are assigned in the new framework of statistical tests and conditional tests. A critical issue about the scoring system is how to assign a weight to each indicator. Since the relative risk of an individual indicator is difficult to measure, grouping indicators based on risk levels may help to manage them. Therefore, the indicators are divided into a lower risk group and a higher risk groups. After grouping, the difference between the two groups is measured and weights commensurate with relative risk are assigned. This process is as challenging as assigning weights to individual variables. An alternative way is to handle them separately by applying different thresholds to each category as in Phase I and Phase II.

The maximum number of flagged wire transfers that can be verified by the internal audit team is limited to 30 because of the limited human resources. Hence, the threshold for each category is determined to flag up to 30 wire transfers. In the old categorization, 19 wire transfers are flagged after applying thresholds, "[(Trend≥10) OR (Control≥2 AND Total≥5)] AND Amount≥$2,000". In the new categorization, "(Statistical≥6) OR (Conditional≥5)" is used for thresholds, and nineteen wire transfers are flagged for further investigation. Comparing flagged wire transfers by both methods shows that eleven wire transfers are selected in common. Wire transfers flagged by both categorizations have higher suspicion scores, whereas those selected by only one method have relatively low scores, so that they are more affected by the nature of the categorization. This difference can be reconciled by loosening the resource limit. Accordingly, this problem will continue as long as the number of wire transfers is capped to thirty. Although the new categorization seems to provide a more intuitive framework and results than old one, the union of both sets of results is suggested to the audit team to compare the effectiveness of each method. As a result, 38 wire transfers are delivered to the audit team in this quarter.

3.4.6. Results and Discussion. *Results:* The 38 wire transfers are investigated by the internal audit team during their quarterly audit. After verification, the audit team concludes that the suggested wires are neither fraudulent nor erroneous. Although the investigation does not find any anomalous wire transfers, the flagged transactions have similar interesting characteristics as those found in Phase II. First, most of the flagged wires are sent to other internal departments

or the company's subsidiaries. The audit team says that the money is traceable since it is still inside the company, so it does not bear any risks related to internal fraud. Despite their claim, it is unclear whether all internal accounts have sufficient IC. If they are properly controlled, as the internal auditors assert, internal money transfers may be free of internal fraud. Although these wire transfers may involve management fraud that possibility is beyond the scope of this study. If it can be assumed that internal transfers are free of internal fraud, the concentration wires (i.e., internal wire transfers to optimize the company's fund usage) could be excluded from testing. Second, some of the flagged wires are sent to tax authorities that do not have a personal interest in the company. Therefore, it is reasonable to assume that those wires do not bear anomaly risks. Third, the flagged wires processed on non-work days are time-sensitive and sent to other internal departments, so employees must process them on those days. In addition to these characteristics noted in Phase II, some of the flagged transactions are batch wires newly introduced during the latest quarter. Since they do not have a sufficiently long history, their violations are mainly related to the control tests that examine abnormal frequencies. This problem will vanish once these transactions have stayed in the payment system for sufficient time. These batch transfers will be included in the next quarter.

In order to screen wire transfers more efficiently, the less risky transactions are excluded and the whole detection process is performed again. After excluding summary observations (40), rejected wires (1,239), internal transfers and transfers to subsidiaries or the IRS (93,030), and newly added batch types (23,760), 142,693 wire transfers remain in the data set. Approximately 90% of these wire transfers belong to 17% of the payees after data cleaning. Their categories, suspicion scores, and thresholds are shown in Table 9, and a summary of results by categorization is shown in Table 10.

As before, excluding the less risky wire transfers, the wire transfers with high suspicion scores are not affected by categorization. This result shows that a transition to new categorization has benefits including easier screening control and better management of anomaly indicators.

In this phase, the audit team does not find any evidence to support the existence of anomalies in the flagged wire transfers. However, this does not mean that all the wire transfers in the quarter are free of internal fraud. Instead, it may imply that the current detection model is not powerful enough to catch anomalous wire transfers or that the indicators are not properly weighted. In either case, it is evident that the detection model needs revision for improvement.

HR records: Employees' authorization limits have an important role in the wire transfer process. As found in Phase II, the company's HR file shows that multiple authorization limits can be assigned to an employee. In addition, it is not possible to track employees' authorization limits once they leave the company because their records are erased. This is supplementary evidence that the company needs more caution about its database management because this problem can be catastrophic, especially in auditing. This illustrates that the inspection of past data should not be overlooked while investigating the recent transactions.

Table 9. Flagged Wire Transfers Using the Old and New Models.

Previous	
Trend	**Wires**
0	9,163
1	13,770
2	2,543
3	851
4	535
5	109
6	90
7	133
8	34
9	22
10	6
11	1
Control	**Wires**
0	26,977
1	258
2	14
4	8
Total	**Wires**
0	9,006
1	13,849
2	2,555
3	907
4	539
5	114
6	91
7	133
8	34
9	22
10	6
11	1

Table 9. (*Continued*).

New	
Statistical	**Wires**
0	25,625
1	1,079
2	420
3	93
4	30
5	8
6	2
Conditional	**Wires**
0	9,624
1	14,694
2	2,382
3	504
4	39
5	14
Total	**Wires**
0	9,006
1	14,222
2	3,072
3	735
4	138
5	66
6	14
7	3
8	1

Discussion: The Phase III study presents the development and results of the third-generation model for anomaly detection. Major features in Phase III include: (1) Anomaly indicators are added, modified, and dropped based on the feedback from Phase II, and the old model that characterizes indicators into trend tests and control tests is compared to the new model that categorizes indicators into statistical tests and conditional tests; (2) Weights are revised for indicators when the new categorization is used by treating each category a group, which

Table 10. Summary of Results by Categorization.

	Previous	New
Criteria	(Trend≥9) OR (Control≥2 and Total ≥4)	(Statistical≥5) OR (Conditional) ≥5 OR (Total≥6)
Amount	≥2,000	≥2,000
The number of flagged wires	24	28
Comment	To flag as many wire transfers as possible up to 30. (Trend≥9) and (Total≥9) produce the same result	To flag as many wire transfers as possible up to 30

facilitates use of the scoring system; and (3) An additional problem concerning employee record files is found due to inadequate database management, which prevents tracking of authorization limits for employees who left the company.

Despite high suspicion scores, no flagged wire transfers are found to be anomalous. The newly raised issues and feedback from the regular audit require further investigation and implementation in the next phase.

This study provides a learning experience about the third-generation anomaly development process. New findings in this phase will serve as a basis for model development in the next phase.

3.5. Phase IV

3.5.1. Data. The model in Phase IV starts with the Phase III model and feedback from the last quarter's audit. After the components of wire transfers are examined in detail, several ideas are developed to improve the anomaly detection model.

First, the model may perform better if less risky wire transfers are excluded. Among the various types of wire transfers in the data, some have higher risk than the others. Since detection of internal fraud is a more important goal than detection of simple errors, focusing only on the more risky wire transfers can help to improve the discriminating power of the model. One way to focus on risky wires is to exclude wire transfers whose destination is in the company or its subsidiaries. Wire transfers can be divided into internal and external payees based on the destination of the recipients. An internal payee is either a department in the company or a subsidiary of the company. Since funds related to internal wire transfers do not leave the company, it will be reasonable to assume that the internal transactions bear low internal fraud risk, so they can be excluded. Another group of wire transfers that can be excluded is rejected wires. An initiated wire transfer is not always approved. Since rejected wires do not affect cash outflows, it is reasonable to exclude them. Finally, wire transfers to the IRS are free of anomaly risk since they are for a taxation purpose. Hence, they can reasonably be excluded from this study.

Second, an anomaly detection model will perform better if wire transfers prone to false positives are excluded in a data cleaning process. As found in the

previous quarter, batch transactions violate many anomaly tests because of their short history. These false alarms reduce the model's detection power so they are excluded in Phase IV.

The data set in Phase IV consists of the data set in Phase III and the fourth quarter (i.e., October, November, and December) of 2009. Thus, it spans two years from January 2008 through December 2009, consisting of 323,917 wire payments. After excluding 52 summary observations, 1,510 rejected wires, 113,816 internal transfers to subsidiaries or the IRS, and 298 batch type transfers, 208,241 wire transfers remains in the data set. Overall, approximately 90% of the wire transfers belong to 26% of the payees after data cleaning.

3.5.2. Model Development Process. Phase IV starts with the model from Phase III and takes the feedback and discussion with the audit team into consideration. All of the anomaly indicators in Phase III are transferred to Phase IV and one indicator that was dropped in Phase II because of data unavailability is added again. Due to the non-disclosure agreement with the insurance company, its detail cannot be discussed. In addition, the anomaly indicators using prediction intervals are revised in more meaningful manner.

Another significant change in Phase IV is that past transactions are separated from the latest quarter to develop better anomaly indicators. Until the last quarter, anomaly indicators with prediction intervals are built with all available transactions, and then those behaving differently from the population are labeled as potentially anomalous. A problem with this approach is that the universal behavior is affected by the target transactions in the latest quarter, although the abnormalities in this group are still unknown. Instead of this approach, the model in Phase IV excludes the target transactions in the current period to determine the prediction intervals, which are then applied to the target transactions to signal potentially anomalous ones.

3.5.3. Screening Rules. The model in Phase IV includes all of the anomaly indicators in Phase III and one from Phase II. The prediction intervals from past transactions are used to predict the abnormality of current transactions, which makes the indicators more statistically robust.

After these changes, the Phase IV detection model has six types of statistical tests (11 indicators) and 15 types of conditional tests (27 indicators).

3.5.4. Scoring System. The suspicion scoring system in Phase IV inherits the Phase III scoring, except for the weighting system. While Phase III shows both unequal and equal weighting systems, this phase presents only the equal weighting system. Although unequal weighting is more realistic, the subjectivity of assessing relative risk often lead to over- or under-estimation of indicator weights as observed in the previous phases. In order to avoid this issue, anomaly indicators in Phase IV are equally weighted.

The limitation of human resources that can be assigned to wire transfer verification affects the thresholds used to determine the number of flagged wire transfers. As in the previous quarters, the default number of wire transfers for verification is thirty, which is not a large number. The thresholds for categories are determined to satisfy this constraint, while allocating the number evenly to each category in order to avoid heavy dependence on only one test category. As a result, 20 wire transfers that meet the threshold condition, "(Statistical≥5) OR

Table 11. Flagged Wire Transfers in Phase IV.

	Previous	New
Criteria	(Trend≥9) OR (Control≥2 AND Total ≥4)	(Statistical≥5) OR (Conditional) ≥5 OR (Total≥6)
Amount	≥2,000	≥2,000
The number of flagged wires	24	28
Comment	To flag as many wire transfers as possible up to 30. (Trend≥9) and (Total≥9) produce the same result	To flag as many wire transfers as possible up to 30

Statistical	**cnt_wires**
0	52,574
1	2,092
2	485
3	281
4	92
5	12
6	2

Conditional	**cnt_wires**
0	3,4472
1	17,787
2	2,183
3	987
4	80
5	22
6	5
7	2

Total	**cnt_wires**
0	32,666
1	18,227
2	2,824
3	1,425
4	266
5	84
6	32
7	9
8	3
10	2

(Conditional≥6) OR (Total≥7)" are flagged and sent to the audit team for further investigation. Table 11 summarizes suspicion scores, thresholds, and the number of flagged wires transfers by test category.

3.5.5. Results and Discussion. *Results:* One of the obstacles to the development of an anomaly detection model is that relevant information is highly decentralized. That is, there are no master files that explain what values can or must exist for each variable, such as wire type or payee. Although the audit team ascertains that the information is kept in physical forms, it is challenging to trace the documents' traits. This might happen because the wire payment system collects transactions from a number of subsidiaries (about 2,300 in and out of the USA) that become part of the company as a result of mergers and acquisitions, and the data transition system is not completely established. This decentralized system makes it difficult to identify exceptional cases. The verification results in this phase shows that this issue can reduce the quality of an anomaly detection model.

Another obstacle is that the verification process is conducted only when the audit team can assign its human resources during regular quarterly audits. The response to requests to verify certain wire transfers were often delayed until the next regular audit cycle, so the results has to be reflected in the subsequent phase.

The 20 flagged wire transfers are inspected by the audit team. Although no wire transfers are found to be anomalous, the investigation result shows unexpected facts that raise three issues. First, additional less risky areas are found concerning loans and subsidiaries. Although loan-related wire transfers are sent to outside payees, they will be collected in the future. In other words, the risk related to this type of wire transfers is a potential default that is connected to a collection process rather than the payment system. It is also found that certain subsidiaries are not in the subsidiaries master file. Accordingly, wire transfers of the newly found subsidiaries will be excluded for the same reason as those of subsidiaries in the master file.

Second, some flagged wire transfers were approved by supervisors whose approvals are needed only for suspicious or unusual wire transfers. Clearly, the supervisors recognized that those wire transfers were anomalous and needed a special care, which implies that the anomaly detection model in this phase is working as intended.

Third, further investigation of wire transfers approved by the supervisors discloses an undesirable truth that one employee has two IDs. An employee ID is like a social security number in that a person can have only one. This violates the entity integrity that a primary key must be uniquely assigned to an entity. Since this situation is contrary to the conditions needed for a relational database, an immediate request is sent to the audit team. After examination, the audit team gives an explanation that the supervisor worked both in the company and at home. For a virtual private network that enables an employee to access the company's data processing system remotely, the employee used the different ID. Despite this justification, it is highly doubtful that a separate ID was necessary for a remote access. Although a request for much closer investigation of this case was made, further inspection was not conducted. This instance highlights the need for better control over the company's DBMS.

Overall, no supportive evidence is found that the flagged wire transfers are anomalous. However, the existence of the flagged wire transfers that were approved by supervisors implies that the detection model is appropriately identifying anomalous wire transfers. A future phase will consider these findings for further improvements.

Discussion: The anomaly detection model in Phase IV is adjusted to reflect the findings and issues in the previous phases. Major findings and changes include: (1) A control test that was dropped in Phase III has been added back. The test becomes feasible after relevant information is provided by the audit team. (2) Upper and lower bounds of anomaly indicators that use prediction intervals are computed based on past data and applied to target transactions. (3) Wire transfers related to loans or unlisted subsidiaries are newly found and estimated to be less risky areas. These will be excluded in the next phase. (4) Evidence implying that the detection model is working is found. This is supported by the fact that some flagged wire transfers are approved by supervisors instead of ordinary approvers. (5) A violation of entity integrity is found in approvers. Although supervisors are supposed to have superior power, it does not mean that they have a right to have two IDs. This must not be taken for granted. Rather, it may require an urgent action from the audit team or the IT department. The next phase will start with resolving newly found problems, such as modification of data cleaning process.

4. Conclusion, Limitations, and Future Research

This project on an insurance company's wire payment system provides an ongoing anomaly detection model. The study starts with a pilot study to test the possibility of implementing an anomaly detection model in the company's regular audit. Since it is the company's first attempt to apply an anomaly detection model to their regular audit, the project faces challenges that are caused by misunderstanding and miscommunication. However, the most difficult challenge is due to the lack of direct accessibility to necessary information. Although the company uses a highly computerized database system, the internal audit team cannot directly access the information in need. This difficulty is mainly caused by the way that the company has become a large enterprise. The insurance company grows by acquiring other businesses that have their own DBMSs. After each acquisition, the new subsidiary's database systems need to be merged into the parent company, but the conversion of systems and data is costly and time-consuming. As a solution, the company converts and merges only information that is indispensable for running business and leaves the subsidiary in charge of the remaining data. This becomes a problem when employees of subsidiaries leave the company. Since they are the only people who know where the information is kept and what it means, their termination causes breaks in information linkages. After decades of acquisitions, the absence of these links interrupts the communication of information that is not centralized. In the age of high information technology, it is generally assumed that a company's data are electronically stored, but that is not the case at this firm, which makes it difficult to find detailed information in a timely manner.

This study provides a variety of useful findings about the company's wire payment system. First, it provides evidence that unsupervised methods can be a useful tool for anomaly detection. When a company first starts an anomaly detection activity, it is highly likely that no prior information is available. Once an anomaly detection model is in place, modifying and improving it are less difficult. This study helps the company by presenting a detection model at its beginning stage and its subsequent modification processes. Second, this study presents findings about the company's DBMS. In the age of information technology, ERP systems can be indispensable to businesses because of their time and cost savings. However, if a company uses a computerized data processing system that is customized for its business environment, differences between parent and subsidiary systems can be so large that complete integration of the two systems may be overwhelming, if not impossible. As a result, this insurance company has a mix of integrated and disintegrated database. Despite the firm's efforts to unify all of its related systems, total integration may be an unattainable goal due to limited resources and costs. During development of the anomaly detection models, many flaws have been discovered. Some of them can threaten the integrity of the current information system, while others can be fixed with gradual changes. Finally, the project helps to identify potentially risky areas that the company did not consider before. At the same time, it is also determined that certain types of wire transfers may not need as much care as others. With these findings, it becomes possible to narrow the scope of transactions that should be investigated. This is valuable information for the firm and for future studies of anomaly detection.

This study helps academics by showing how an anomaly detection model is implemented and improved in practice. In addition, this study shows that anomaly detection activities can be useful during regular audits to help internal auditors identify possible weak or risky areas and transactions.

Although this study does not identify actual anomalous wire transfers, many issues are raised, resolved, and dropped over the development process. The findings provide indirect evidence of the model's operation, but that is insufficient to claim that the model is working as it is supposed to. Future research needs to consider all of these findings and take more care with data cleaning in order to focus only on the most relevant pool of transactions. Exclusion of irrelevant or less risky transactions can help reduce possible noise that can adversely affect the quality of an anomaly detection model.

References

American Institute of Certified Public Accountants (AICPA). (2002). *Consideration of fraud in a financial statement audit*, SAS 99. New York, NY: AICPA.

Fawcett, T., & Provost, F. (1999). Activity monitoring: Noticing interesting changes in behavior. In *ACM SIGKDD 5th international conference on knowledge discovery and data mining* (pp. 53–62).

Kim, Y., & Vasarhelyi, M. A. (2012). A model to detect potentially fraudulent/abnormal wires of an insurance company: An unsupervised rule-based approach. *Journal of Emerging Technologies in Accounting*, *9*(1), 95–110.

Kogan, A., Sudit, E. F., & Vasarhelyi, M. A. (1999). Continuous online auditing: A program of research. *Journal of Information Systems*, *13*(2), 87–103.

Murthy, U. S., & Groomer, M. S. (2004). A continuous auditing web services model for XML-based accounting systems. *International Journal of Accounting Information Systems*, *5*(2), 139–163.

Pathak, J., Vidyarthi, N., & Summers, S. L. (2005). A fuzzy-based algorithm for auditors to detect elements of fraud in settled insurance claims. *Managerial Auditing Journal*, *20*(6), 632–644.

Rezaee, Z., Sharbatoghlie, A., Elam, R., & McMickle, P. L. (2002). Continuous auditing: Building automated auditing capability. *Auditing: A Journal of Practice and Theory*, *21*(1), 147–163.

Vasarhelyi, M. A., & Halper, F. B. (1991). The continuous audit of online systems. *Auditing: A Journal of Practice and Theory*, *10*(1), 110–125.

Woodroof, J., & Searcy, D. (2001). Continuous audit: Model development and implementation within a debt covenant compliance domain. *International Journal of Accounting Information Systems*, *2*(3), 169–191.

Part V

Audit Analytics for Lawsuit Risk Detection

This page intentionally left blank

Chapter 11

A Legal Risk Prediction Model for Credit Cards

Feiqi Huang, Qi Liu and Miklos Vasarhelyi

1. Introduction

Operational risk is a serious issue in the banking and finance industry (Hellwig, 1995). The definition of operational risk is "the risk of a change in value caused by the fact that actual losses, incurred for inadequate or failed internal processes, people and systems, or from external events (including legal risk), differ from the expected losses" (Basel Committee on Banking Supervision, 2001). Among all operational risks, legal risk is special and important because, unlike most other operational risks, legal risk cannot be traded away in any market (Molot, 2009). In recent years, legal risk and legal expense have become the most common area of risk management encountered by corporate counsel (Glidden, Lea, & Victor, 2016, chapter 12). Companies can incur extremely large legal expense, even though they finally win the lawsuit. Reports show that large global banks' legal expenses are more than $100 billion (Kapner, 2013).

Prior literature claims that legal risk could be a good indicator of the weakness of internal control and an omen of bad operational performance in the future (McNulty & Akhigbe, 2014). In addition, SAS No.109 requires auditors to have a sufficient understanding of the entity, environment, and internal control (AICPA, 2006). Compared with the backwards focus of traditional audit, the new audit focus is forward looking or predictive (Kuenkaikaew & Vasarhelyi, 2013). By conducting predictive analysis, auditors can to help firms to predict and prevent potential lawsuits and remediate any weaknesses in internal control.

Using a South American financial group's credit card data sets, this paper studies the application of predictive analysis on legal risk. The objective of this research is to predict and prevent lawsuits against the financial group by building various prediction models. Using nine different classification algorithms, this study achieves a high area receiver operating characteristic (ROC) curve and high recall rate, suggesting that the prediction model can filter potential lawsuit

Audit Analytics in the Financial Industry, 203–213

doi:10.1108/978-1-78743-085-320191011

successfully. To the best of our knowledge, there is no existing literature that uses a predictive model to analysis legal risk based on credit card holders' information. There are two reasons that lawsuit prediction has not been studied. First, public companies are not required to disclose legal expense in their financial statements, so legal risk and legal expense may be overlooked by the literature. Second, legal expense and legal risk information relate to key business information for any organizations, so it is difficult for researchers to get access to a firm's legal risk data sets.

Section 2 in this chapter reviews the prior literature about operational risk, legal risk, and predictive analysis. Section 3 describes the data sets, methods, and performance measurement. Several prediction models are built, and the results are discussed and compared in Section 4. Section 5 presents the conclusion and avenues for future research.

2. Literature Review

2.1. Operational and Legal Risk

In The New Basel Capital Accord (Basel II), published by the Basel Committee, operational risk is defined as "the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events," including legal risk (Cornalba & Giudici, 2004). In Basel II, three methods can be used to calculate the operational annual capital charge: the basic indicator approach, the standardized approach, and the advanced measurement approach. Prior literature has conducted many quantitative analyses for operational risk after Basel II. For instance, Cornalba and Giudici (2004) use actuarial models, causal models, and Bayesian networks to analyze operational risk. Shevchenko (2010) points out that a Bayesian framework is suited for operational risk modeling. Böcker and Klüppelberg (2010) use spectrally positive Lévy processes to analyze the risk.

However, legal risk is not like other operational risks that have been fully explored by quantitative analysis (Chavez-Demoulin, Embrechts, & Neslehova, 2006). Until now, legal risk has no standard definition. Legal risk can be regarded as the likelihood of being taken into court. McCormick (2010) believes that legal risk is an institution's risk of loss that is primarily caused by: (a) a defective transaction; (b) a claim (including a defense to a claim or a counterclaim) being made, or some other event that results in a liability for the institution, or other loss (e.g., the termination of a contract); (c) failing to take appropriate measures to protect assets (e.g., intellectual property) owned by the institution; or (d) change in law. Large companies with deep pockets can be especially prone to legal risk as the rewards for any plaintiffs can be considerable. Customer lawsuits against banks are a serious problem. For instance, Bank of America paid $410 million to settle a class action lawsuit in 2012. However, to the best of our knowledge, there is no existing literature that focuses on legal risk prediction, so this study is the first research to predict legal risk, especially concerning lawsuit cases.

2.2. Predictive Analysis

Predictive analytics is conducted using statistical techniques (e.g., modeling, machine learning, or data mining) to analyze historical and current data and make predictions about future events (Nyce, 2007). Popular methodologies in predictive analysis include linear regression, logistic regression, probit regression, time series models, neural networks, multilayer perceptrons, and Decision Trees. Predictive analysis is widely used in various industries, such as health care, pharmaceuticals, business, engineering, and manufacturing.

In business areas, predictive models exploit patterns found in historical data to identify risks and opportunities. For instance, plenty of research assesses fraud risk in accounting by using neural networks, Bayesian belief network, Decision Trees, logistic model, and other predictive model (Bell & Carcello, 2000; Kirkos, Spathis, & Manolopoulos 2007; Lin, Hwang, & Becker, 2003; Sharma, 2012; Zhou & Kapoor, 2011). Research on credit card scoring problems and bankruptcy predictions are common applications of predictive analysis (Abdou & Pointon, 2011; Atiya, 2001; Min & Lee, 2008; Nanni & Lumini, 2009; Tsai & Wu, 2008). Predictive models are also widely applied in financial distress prediction. Traditional statistical methodologies, such as discriminant analysis, probit regression, and logistic regression, have been widely used to predict corporate financial distress. Recent literature also uses various neural networks models to predict financial distress.

This chapter applies predictive analysis to legal risk to predict or prevent lawsuits against a bank in a real-world business setting.

3. Methodology

3.1. Data Description

This chapter uses data sets related to the credit card business of a major South American financial group. The data sets contain information on cardholders, lawsuit cases, complaint records, customer default data, and credit card restriction data. Cardholder data describe each account holders' personal information, and it contains 289 variables and 67,049,047 observations. Lawsuit data record information for each lawsuit case from January 2008 to August 2013, and they contain 256 variables and 1,495,673 instances. Complaint data show clients' complaint records from July 2013 to September 2013. This data set has 26 variables and 1,116,386 records. Default data contain 50 variables and 53,224,215 observations concerning credit card holders' default information. The last data set concerns credit card restriction. It has 27 fields and 197,950,335 records. After discussion with domain experts, 30 variables are chosen for the prediction model. These variables are presented in Table 1.

This table contains 42,235,966 distinct clients. Only 1.4% of them have sued the bank, so this data set contains highly imbalanced data. Following the data mining literature to deal with unbalanced data, rare class data are over-weighted by randomly extracting 450,000 non-lawsuit clients and 412,928 lawsuit clients to form the balanced training data. For the testing data set, 45,880 lawsuit records

Table 1. Variable Definitions.

No.	Variables	Source	Description
1	bj_indicator	Lawsuit information	Whether the client sues the bank
2	CDCPFCGC_CRIPTO	Clients information	Identification number
3	CDPROFIS	Clients information	Code that uniquely identifies an individual's occupation
4	CODCEPRE_CRIPTO	Clients information	Residential zip code
5	CODESTRES	Clients information	State residential code
6	CODREGRE	Clients information	Region residential code
7	FISOUJUR	Clients information	Identifier of the type of person and company
8	SEXO	Clients information	Gender
9	IDADASSO	Clients information	Age
10	active_cards	Clients information	Number of active cards
11	average_limit_active	Clients information	Average credit limit of active cards
12	average_limit_inactive	Clients information	Average credit limit of inactive cards
13	block_indicator	Clients information	Whether block happened before lawsuit
14	inactive_cards	Clients information	Number of inactive cards
15	income_dif	Clients information	The difference between confirmed annual income and declared annual income
16	fq_count	Complaint information	Number of complaint records
17	fq_indicator	Complaint information	Whether the client has complained to the bank
18	ca_account	Default information	Number of credit cards default records

Table 1. (*Continued*)

No.	Variables	Source	Description
19	ca_indicator	Default information	Whether the client has a credit cards default record
20	avg_risk_10	Default information	Average value of total risk overdue up to 10 days late
21	avg_risk_360	Default information	Average value of total risk arrears over 360 days late
22	avg_risk_11_30	Default information	Average value of total risk won between 11 and 30 days late
23	avg_risk_121_180	Default information	Average value of total risk won between 121 and 180 days late
24	avg_risk_181_240	Default information	Average value of total risk won between 181 and 240 days late
25	avg_risk_241_360	Default information	Average value of total risk won between 241 and 360 days late
26	avg_risk_31_60	Default information	Average value of total risk won between 31 and 60 days late
27	avg_risk_61_90	Default information	Average value of total risk won between 61 and 90 days late
28	avg_risk_91_120	Default information	Average value of total risk won between 91 and 120 days late
29	bx_count	Restriction information	Number of credit cards restriction records
30	bx_indicator	Restriction information	Whether the client has a credit cards restriction record

and 3,192,230 non-suing records are chosen, which keeps the same lawsuit ratio as the original data set.

3.2. Methods

In the process of building prediction models, SAS is used to preprocess data and SPSS Modeler is used to build prediction models. Nine different supervised learning algorithms are run to build prediction models for the lawsuit data set:

C5.0, CHAID, decision list, C&R tree, QUEST (quick, unbiased, and efficient statistical tree), Bayesian network, discriminant, neural network, and logistic regression.

The Decision Tree is a flowchart-like structure that is composed of internal decision nodes and terminal leaves. An internal decision node represents a test function $f_m(x)$ with discrete outcomes labeling the branches, a branch represents the outcome of the test, and each leaf node defines a class label. The path from root to leaf represents classification IF-THEN rules that are easily understandable (Alpaydin, 2010). Decision Tree C5.0 is used in this prediction model, which is improved by computer science researcher Ross Quinlan based on his own algorithm Decision Tree C4.5. Compared to C4.5, C5.0 offers a number of improvements, such as high speed, memory efficiency, a smaller size result tree, different weights, and automatic winnowing.

Logistic regression is a type of probabilistic statistical classification model (Bishop, 2006). It is based on the cumulative logistic probability function and has been widely applied to classify binary dependent variables. This method categorizes a binary dependent variable with one or more independent variables and uses probability scores as the predicted value of the dependent variable. Logistic regression is a widely used statistical technique when the probability of an outcome is related to a series of potential predictor variables (Hosmer & Lemeshow, 2004).

3.3. Performance Measure

The data sets for this study is imbalanced as less than 2% of clients have ever sued the bank. Imbalanced data mean that the number of instance in one class is much fewer than the instances in another class. This feature implies that predictive accuracy, which is a common measure of the performance of prediction model, might not be appropriate (Chawla, 2005). To illustrate this issue, assume that a data set contains 10,000 clients and only 10 clients sue the bank. If the model predict that all the clients do not sue the bank, the predictive accuracy of the model will be 99.9%. However, that model simply avoids the rare class, which is the focus of this prediction model. To avoiding this problem, this study uses a ROC curve, recall, and precision as the measures of model performance. ROC is a graphic approach for presenting classifier performance over a range of tradeoffs between true positive and false positive error rates (Swets, 1988). The more area covered by the ROC curve, the better the model is. Recall presents the model's capability to predict positive instances, and precision reflects the accuracy of predicted positive instances. The main goal for learning from imbalanced data sets is to improve the recall without hurting the precision (Chawla, 2005).

4. Result and Discussion

4.1. Preliminary Prediction Model

The preliminary prediction model contains the dependent variable bj_indicator and ten independent variables: IDADASSO, SEXO, FISOUJUR, CDPROFIS,

income_diff, active_cards, inactive_cards, complain_indicator, average_limit_active, and average_limit_inactive. All independent variables come from the cardholder information data set and complaint data set. Table 2 shows the models built based on the training data set. In terms of ROC curve area, models built by algorithms C5.0, CHAID, and C&R tree are the best prediction models. C5.0 has the best performance by achieving 97.4% area under the ROC curve for the model that includes all variables. Untabulated results show that the number of active cards and the number of inactive cards have strong impacts on lawsuit action.

4.2. Final Prediction Model

In discussion with the company's internal auditors, they suggested that credit card block time and the clients' addresses may be important. In addition, credit card default information and restriction information are taken into consideration. Therefore, several more variables are added to prediction model. After using the new training data, Table 3 indicates that the best four algorithms are C5.0 (99.1% ROC), neural network (97.5% ROC), CHAID (97.4% ROC), and logistic regression (94.9% ROC). When the best performance model, the C5.0, is used with the testing data, the model achieves 95.63% recall rate and 18.91% precision rate. In addition, an alternative model is built by using the ensemble method of applying the best three models to the testing data and conducting majority voting to decide class. The recall (precision) rate of the ensemble model is 93.63% (18.59%). Given that the C5.0 model achieved higher recall and precision, the C5.0 Decision Tree is the better model.

According to the C5.0 model, 26 variables are used in the Decision Tree. The depth of this tree (the longest path from root to leaf) is 24 and it contains hundreds of rules, so it is not easy to present the whole tree structure and explain all the thresholds. Fig. 1 shows the five most important variables in the C5.0 model: (1) the number of inactive cards; (2) an indicator for whether the client's cards are blocked before the lawsuit; (3) the number of active cards; (4) age; and (5) an indicator for whether the credit card is restricted. More inactive cards and block_indicator will lead a higher likelihood of initiating a lawsuit case, possibly because more inactive cards and block cards may bring more unexpected cost for cardholders, which may lead to lawsuits.

It might be supposed that the complaint indicator would have a significant impact on lawsuits, but the results do not support that supposition, possibly because the complaint data set covers only three recent months, making this variable unimpotant in the model.

4.3. Discussion

Usually, when efforts are made to improve recall, it will hurt precision. How to find the trade-off between recall and precision is related to factors like the business environment and management's goals. During the learning process, a cost matrix can be imposed on false positive or false negtive records, which

Table 2. Model Performance (Tranining Data Set).

Model	Build Time	Max Profit	Max Profit Occurs in (%)	Lift (Top 30%)	Overall Accuracy (%)	No. Fields Used	Area Under Curve
C5.0	1	1,945,938….	54	2.023	92.792	10	0.974
CHAID	<1	1,769,579….	57	1.972	89.071	5	0.954
C&R Tree	<1	1,644,886….	61	1.829	86.552	7	0.917

Table 3. Model Performance (Testing Data Set).

Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift (Top 30%)	Overall Accuracy (%)	No. Fields Used	Area Under Curve
C5.0	4	1,861,412,718	49	2.075	95.53	26	0.991
Neural network	72	1,671,285	49	2.048	91.04	27	0.975
CHAID	1	1,685,802.401	50	2.036	91.404	8	0.974
Logistic regression	4	1,509,635	49	2.033	87.026	28	0.949

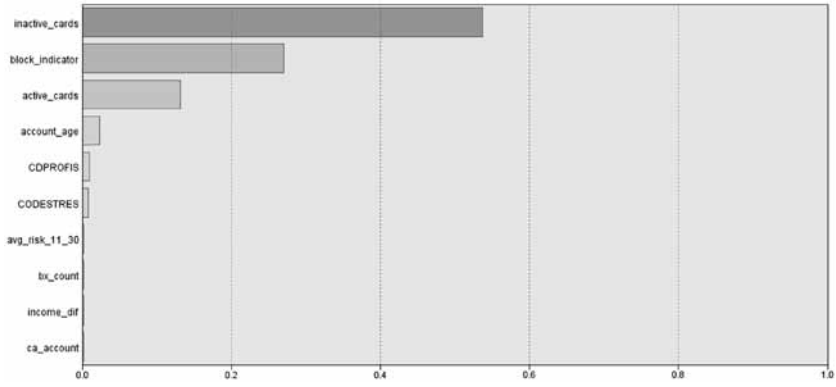


Fig. 1. The Five Most Important Variables in the C5.0 Model.

reflects the preference for recall versus precision. For instance, internal auditors can adjust the cost matrix parameters to minimize the cost, based on the future costs of failing to recognize a "lawsuit client" and misunderstanding a good customer.

5. Conclusions

Legal risk has become the most common area of risk for which companies may incur extremely large legal expenses. Therefore, the ability to predict potential future lawsuits would be a valuable part of financial institutions' strategies. However, unlike other operational risks, legal risk has not been explored using quantitative analysis in prior literature at least partially due to the lack of lawsuit data.

Using a South American financial group's credit card data sets, this research builds predictive models to predict and prevent lawsuit against the financial group. After comparing the results of nine different classification algorithms, the results in this study suggest that the Decision Tree C5.0 model achieves the best performance in predicting potential lawsuit cases based on clients' information. Specifically, this model achieves a high area ROC curve and high recall rate, suggesting that the prediction model can filter potential lawsuit successfully.

One way to extend this research might include performing factor analysis[1] to classify multiple variables into several factors, which would reduce the number of dimensions in the prediction model and improve its performance. In addition, analyzing the cause of the lawsuits and building separate prediction models based on different lawsuit causes could be useful to improve the model's accuracy. Finally, using geography-related data to predict potential lawsuit conspiracies would be an interesting follow-up for future research.

[1]A statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables.

References

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance, and Management*, *88*(June), 59–88.

AICPA. (2006). Understanding the Entity and Its Environment and Assessing the Risks of Material Misstatement. Statement on Auditing Standards No. 109. New York, NY: AICPA.

Alpaydin, E. (2010). *Introduction to machine learning* (2nd ed.). Cambridge, MA: MIT Press.

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks*, *12*(4), 929–935.

Basel Committee on Banking Supervision. (2001). The New Basel Capital Accord. Bank for International Settlements.

Bell, T., & Carcello, J. (2000). A decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing: A Journal of Practice and Theory*, *9*(1), 169–178.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Böcker, K., & Klüppelberg, C. (2010). Multivariate models for operational risk. *Quantitative Finance*, *10*(8), 855–869.

Chavez-Demoulin, V., Embrechts, P., & Neslehova, J. (2006). Quantitative models for operational risk: Extremes, dependence and aggregation. *Journal of Banking and Finance*, *30*(10), 2635–2658.

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 875–886). Boston, MA: Springer.

Cornalba, C., & Giudici, P. (2004). Statistical models for operational risk management. *Physica A: Statistical Mechanics and Its Applications*, *338*(1–2), 166–172.

Glidden, C. B., Lea, C. W., & Victor, M. B. (2016). Evaluating legal risks and costs with decision tree analysis. In R. L. Haig (Ed.), *Successful partnering between inside and outside counsel* (pp. 12-1–12-25). West Group, American Corporate Counsel Association. Retrieved from www.westlaw.com

Hellwig, M. (1995). Systemic aspects of risk management in banking and finance. *Revue Suisse d'Economie Politique et de Statistique [Swiss journal of economics and statistics]*, *131*, 723–737.

Hosmer, D. W., Jr, & Lemeshow, S. (2004). Applied logistic regression. Hoboken, NJ: John Wiley and Sons.

Kapner, S. (2013). Banks looking at $100 billion legal tab. *Wall Street Journal*. Retrieved from https://www.wsj.com/articles/SB10001424127887323466204578382261903001702

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, *32*(4), 995–1003.

Kuenkaikaew, S., & Vasarhelyi, M. A. (2013). The predictive audit framework. *The International Journal of Digital Accounting Research*, *13*(April), 37–71.

Lin, J. W., Hwang, M. I., & Becker, J. D. (2003). A fuzzy neural network for assessing the risk of fraudulent financial reporting. *Managerial Auditing Journal 18*(8), 657–665.

McCormick, R. (2010). *Legal risk in the financial market*. Oxford: Oxford University Press.

McNulty, J. E., & Akhigbe, A. (2014). Bank litigation, bank performance and operational risk: Evidence from the financial crisis. Retrieved from https://ssrn.com/abstract=2463373 or http://dx.doi.org/10.2139/ssrn.2463373

Min, J. H., & Lee, Y. C. (2008). A practical approach to credit scoring. *Expert Systems with Applications*, *35*(4), 1762–1770.

Molot, J. T. (2009). A market in litigation risk. *The University of Chicago Law Review*, *76*(1), 367–439.

Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications 36*(2), 3028–3033.

Nyce, C. (2007). *Predictive analytics*. White paper. American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America.

Sharma, A. (2012). A review of financial accounting fraud detection based on data mining techniques. *International Journal of Computer Applications*, *39*(1), 169–178.

Shevchenko, P. V. (2010). Implementing loss distribution approach for operational risk. *Applied Stochastic Models in Business and Industry*, *26*(3), 277–307.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293.

Tsai, C. F., & Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, *34*(4), 2639–2649.

Zhou, W., & Kapoor, G. (2011). Detecting evolutionary financial statement fraud. *Decision Support Systems*, *50*(3), 570–575.

This page intentionally left blank

Part VI

Audit Analytics in the Payment Process

This page intentionally left blank

Chapter 12

Analyzing Payment Data and Its Process: A Bank Case

Karina Chandia and Miklos Vasarhely

1. Introduction

In the business world, the use of sophisticated machines and new technologies is increasing. The introduction of these innovations has contributed positively to the auditors' work, specifically in managing huge amounts of data and using that data to make decisions.

In that context, the auditor's role is to determine whether the client's transactions are recorded in accordance with accounting standards. For this purpose, auditors have improved their methodologies to understand and to analyze accounting information, often by using statistical models that are available in the market. Some of those models are based on strict assumptions, such as normal distribution of the errors, constant variance, and non-correlation between the independent variables. However, other tools are seldom used as their use requires specific knowledge that goes beyond traditional audit work requirements. For instance, procedures like neural networks or spectral analysis have shown powerful results in the detection of anomalies, but they are scarcely ever used in accounting, auditing, or management.

Audit decisions, such as evaluating the effectiveness of internal controls, must often be made under conditions of subjectivism and/or uncertainty. Subjectivism occurs because the process of evaluation requires the person making decisions to exercise judgment, but past studies conclude that auditors who perform the same task may, in fact, make different decisions (Biggs & Mock, 1983). Thus, the decision-making process is subjective. By contrast, uncertainties relate to events that have not yet occurred. For example, managers must evaluate internal controls based on threats that could affect the company in the future, so quantifying the possibility of the threat's occurrence could create imprecise measures (de Korvin, Shipley, & Omer, 2004). As a way to assess internal controls over various accounting processes, it is useful to

Audit Analytics in the Financial Industry, 217–235

doi:10.1108/978-1-78743-085-320191012

consider a technique that would allow decisions to be made under conditions of subjectivism and uncertainty.

One methodology that could reduce subjectivism and uncertainties is the fuzzy logic approach, based on fuzzy set theory (Zadeh, 1965). According to Levy and Yoon (1995), fuzzy logic "prescribes a mathematical formalism to handle the vagueness of human knowledge conveyed through natural language." Thus, using fuzzy logic could make it possible to infer an even base on uncertain premises (Levy & Yoon, 1995). Hence, uncertainties are reduced because the use of discrete values (or different membership) allows the decision-maker to know the risk in a more certain way.

Fuzzy logic not only reduces uncertainties and subjectivism, but also provides other advantages when it is applied to assessing internal control. First, the output of the fuzzy logic approach permits the decision-maker to assess the firm's degree of risk in a more realistic way. The inputs of the model are beliefs or estimation of the risk made by experts, and the output of the model is a number between 0 and 1, which could be interpreted as the level of risk of the company, department, or assertion that is the subject of analysis. Second, fuzzy logic requires that companies assess the risks related to their various business processes and cycles, so the creation of the risk model will contribute to the auditor evaluation of the internal control over a specific area. Third, auditors must judge each area detected in the fuzzy model approach, so there will be a formalization of the company's risks and controls at that point. Finally, the fuzzy logic approach can be modified if it is found not to be useful or realistic.

In this chapter, the methodology of fuzzy logic helps to create a generic risk model for assessing internal controls in a bank's payment cycle. The literature has little empirical research focusing on specific processes within a company, such as the payment process. Regulators also do not give specific guidance on internal processes, which creates an opportunity to make different inferences or interpretations that affect the decision-making process. Therefore, it is important to consider specific guidelines or more standardized processes that could help the information users to make their decisions. One way to do so is by guiding companies in their internal process. The fuzzy logic approach may also help to reduce the uncertainties in the environment, which would allow managers to react in ways to reduce the risks that are present in the firm's internal processes. It is also possible that this model could be applied to the company to assess risk, and could be extended to other critical areas of the company.

In addition to the application of the fuzzy logic, this study assesses irregularities and/or anomalies through various statistical models that help the user to understand the data. For that purpose, this study uses frequencies, charts, and metrics to diagnose conflicting data observations related to outliers and leverage. Then, robust regression based on MM estimators is used to address the outliers and leveraged observations.

Section 2 reviews the literature on fuzzy logic and detection of anomalies. The methodology used in this study is described in Section 3. Conclusions and comments are discussed in Section 4.

2. Literature Review

2.1. Fuzzy Logic

Fuzzy logic is concerned with formal principles for approximated reasoning (Zadeh, 1988). Unlike classical logical systems, fuzzy logic creates models under uncertainty. Zadeh, known as the father of fuzzy logic, first defined fuzzy logic in 1965 as "a class of objects with a continuum of grades of membership" (Zadeh, 1965). Those objects or classes of objects do not have a specific membership or characteristic (Tam, Leung, & Chiu, 2002).

As fuzzy logic reduces uncertainty, it has been applied in many areas of business and information technology, such as building an intelligent data warehouse (Krishna & Kumar De, 2001), selecting suppliers (Bayrak, Çelebi, & Taşkin, 2007; Bevilacqua & Petroni, 2002; Famuyiwa, Monplaisir, & Nepal, 2008), assessing electronic commerce transactions (Akhter, Hobbs, & Maamar, 2005), valuing firms (Magni, Malagoli, & Mastroleo, 2006), analyzing acquisition (Glackin, Maguire, McIvor, Humphreys, & Herman, 2007), making foreign currency risk decisions (Lee & Wong, 2007), supply chain management (Tang, Lau, & Ho, 2008), forecasting models (Li & Cheng, 2009) loan decisions (Che, Wang, & Chuang, 2010), and automobile warranty systems (Lee, Lee, & Moon, 2010). Lee (1990a, 1990b) describes the application of fuzzy logic in controls, such as water quality control, automatic train operation systems, automatic container crane operation systems, elevator control, fuzzy logic controller hardware systems, fuzzy memory devices, and fuzzy computers.

Fuzzy logic has also been used in going concern assessment. For example, Lenard, Alam, and Booth (2000) use a fuzzy clustering based on fuzzy logic to identify characteristics that may indicate whether a firm requires a going concern modification. In addition, it has been used in the detection of frauds. Pathak, Vidyarthi, and Summers (2005) use a fuzzy algorithm to detect elements of fraud in insurance claims. Insurance companies face a problem related to assessing the fair value of the claims (Pathak et al., 2005). In that context, human adjusters might fraudulently collude with claimants, affecting the monetary interest of the insurers (Pathak et al., 2005). Dubinsky and Warner (2008) propose fuzzy logic for detecting duplicate payments as an efficient way to reduce internal control problems. Chang, Tsai, Shih, and Hwang (2008) design an audit detection risk assessment that could improve the efficiency of detection risk and reduce the possibility of audit failure.

In internal control evaluation and materiality assessment, fuzzy logic has been applied in cash receipts and shipping control system by Cooley and Hicks (1983). They present a method for the evaluation of internal control systems that combines linguistic information with rigorous mathematical aggregation (Cooley & Hicks, 1983). Rangone (1997) suggests a fuzzy linguistic framework that links organizational effectiveness, key success factors, and performance measures. According to Baldwin, Brown, and Trinkle (2006), Lin finds that a fuzzy neural network for financial detection performs better in this area than considering only an artificial neural network. In related work, de Korvin et al. (2004) formulate a

security risk assessment model based on fuzzy logic. Comunale and Sexton (2005) apply fuzzy logic to the assessment of materiality. They demonstrate that a fuzzy expert system can enable the auditor to incorporate qualitative factors, which can be helpful in drawing conclusions.

2.2. Detecting Anomalies in Accounting Data

Analytical procedures have been used for many years to detect certain financial statements errors. According to Chen and Leitch (1999), trend analysis, ratio analysis, and accounting changes are among the non-statistical procedures used in this area. However, they also point out that many researchers, including Kinney (1978), Loebbecke and Steinbart (1987), Knechel (2007), and Wilson and Colbert (1989) have found that statistical rules are generally more effective than non-statistical rules.

Statistical methodologies have been used to look for better models to detect errors. Chen and Leitch (1999) compare four time series models (Census X-11, autoregressive integrated moving average [ARIMA], a stepwise regression model, and the Martingala model), using quarterly data. The results suggest that all four models provide the same degree of effectiveness, but the stepwise model performs the best among these models.[1]

Statistical methodologies are also effective in looking for key elements in a data set, such as outliers, which are unusual transactions that follow a different distribution than the rest of the data. If outliers correspond to patterns or errors in the recording of economic event, detecting them is important because they affect the decision-making process. Other researchers have developed effective mechanisms to detect anomalous observations. Barnett (1978) proposes a method to identify significant outliers that have a strong influence on the estimation results. Those points are known as leverage points. Similarly, Cook (1979) proposes a mechanism to detect highly leveraged outliers that can influence conclusions.

Robust regression models have been used to control for outliers and leverage points. The most powerful model, proposed by Yohai (1987), is based on MM estimators. This model shows a higher power response in the presence of outliers or unusual transactions than other models can. In addition, this model can be applied without assuming a normal distribution, or when heteroscedasticity may be a problem (Yaffee, 2002).

3. Methodology and Results

3.1. Fuzzy Logic

This study uses fuzzy logic to assess internal control for a bank's payment cycle, based on the specification of de Korvin et al. (2004). This methodology is based on a belief matrix Q, and it is explained in seven steps:

[1]See Chen and Leitch (1999) and Omura and Willet (2006) for additional information.

Step 1: Define the risk matrix. The rows $t_i (i = 1,2,\ldots,n)$ represent the possible threats, and the columns $d_j (j = 1,2,\ldots,n)$ represent the internal control concerns associated with data.

Step 2: Define the risk associated with the exposure to loss due to the failure or absence of internal controls. Each value, r_{ij}, is defined as the risk associated with threat t_i and internal control concern d_j. Fuzzy sets can be used to define scenarios to represent underlying decision uncertainties.

Step 3: Now that the different scenarios associated with each risk are defined, the possible values for r_{nn} range from 0 to 1 in 0.1increments. Thus, it is possible to evaluate the risk for each scenario.

Step 4: Q_{ij} is the belief matrix for r_{ij}. It is possible to define A_k, where $k = 1, N, \left[(A_k \circ Q_{ij})/y\right] = \text{Sup}\ A_k = (x)\hat{}\ Q_{ij}(x,y)$, where A is a fuzzy subset of the attribute space and Q is a relationship from the space of attributes to $[0,1]$.

Step 5: To determine the belief matrix Q, the operator $\varnothing$ is defined on $[0,1]*[0,1]$ by : $a \varnothing b = \text{Sup}\ \{c\ 0 \le c \le 1, a \wedge c \le b\}$, such that if $a \le b$ then $a \varnothing b = 1$ and if $a <$ b then a $\varnothing\ b = b$.

Letting R_{ij}^w be the fuzzy value risk corresponding to fuzzy attribute A_w and T_w is the set of all relations Q_w satisfying $A_w \circ Q_w = r_{ij}^w$ for w fixed, it is assumed that $\bigcap_{w=1}^{n} \mathrm{T}_w \neq \varnothing$. Then the maximal solution of the system $A_w \circ Q_w = r_{ij}^w$ is given by $Q = \bigcap_{w=1}^{n} Q_w$, where Q_w is the maximal solution of $A_w \circ Q_w = r_{ij}^w$ for w fixed.

Step 6: If the current state is a fuzzy subset of r_{ij}, say $\sum_p P_{ap}/p$, then the resulting fuzzy risk is given by $R_{ij}(v) = \text{Sup}\ A(P) \wedge Q_{ij}\ (P,v)$.

Step 7: The overall risk is determined by taking all $R_{ij}(v)$ such that $G_{ij} = \sum_v v\ R_{ij}(v) / \sum_v R_{ij}(v)$. A high G_{ij} indicates a serious problem, whereas a low G_{ij} indicates minimal concern for the decision-maker.

3.2. Application of Fuzzy Logic

In this section, fuzzy logic is used to assess internal controls over the payment cycle. This model can be applied for any company by varying the inputs, which are the degrees of beliefs or possibilities for receiving threats, and by increasing or decreasing the number of controls used in a firm's specific case. The model includes the elements that are in all payment cycles. Then, the controls are defined based on three subsystems: process, system, and staff. Finally, the internal control risks for those three subsystems are calculated.

3.2.1. Elements in the Payment Cycle. The elements in the payment cycle include agents, events, and resources. The resources are the products or services acquired and the cash expended. The events occur when resources are acquired and when payments for those resources are made. The agents are the vendors who supply the goods or services, the employees (bookkeepers, managers, assistants, and cashiers) who process the transactions and the clients who receive the goods or services. These elements are shown in Table 1.

The acquisition process is an essential part of the payment cycle because the department that receives goods or services originates the movements of the payments. The first step in assessing risk in the payment cycle is to evaluate whether

Table 1. Elements in the Payment Cycle.

Resources	Events	Agents
Service/Product	Acquisition	Vendors
Cash	Payment	Employees
		Clients

the acquisition process has been completed (Arens, Elder, & Beasley, 2010). Normally, this means that the company received the goods or service, although companies sometimes pay in advance. When a company verifies the input from the acquiring department, the information includes the approval of the expense, the unique number of the voucher, and information about the vendor, which must coincide with the vendor file. For the resources acquired and the cash expenses, the company must prove that they exist or will exist in the future by analyzing the supporting documentation (Arens et al., 2010).

3.2.2. Types of Controls. To simplify the model, the controls are defined depending on whether they relate to process, staff, or systems. Table 2 presents the template of some controls that are included in the payment cycle.

In Table 2, the first column shows the three elements: process, staff, and system. The second column lists concerns for each element. For instance, concerns about the acquisition process include incomplete information about the purchase event or a fraudulent purchase, calculation errors and recording errors that directly affect the reports, and lack of monitoring of the information processed. The third column gives an example of the control concern, and the last column shows the different states that affect the perception of the degree of the risk.

A control concern is defined as a critical part of the process (in this case, the payment cycle) that could be altered or infringed either by accidental errors or by deliberated acts. Both of these situations will be included in the model as threats that need to be assessed (de Korvin et al., 2004). An accidental error is a non-reflective act that can affect the accuracy of a process, whereas a deliberated act is considered to be a reflective act that has a specific purpose.

As a result, it is possible to identify a matrix of risks for payment cycle that are present in a particular process, such as the payment cycle.

The internal control concerns, shown in the first column of Table 3, are classified according to threats, so each internal control concern can be activated either by accidental errors or by deliberate acts.

Each internal control concern or risk is assessed according to the possible states and threats. Thus, the matrix of risks, which must be assessed by experts who know the company well, can have two possible outcomes that assess the risk level for accidental errors and for deliberate acts. Table 4 shows an example of this type of risk analysis for the System Section.

Each of these situations must be evaluated based on the belief given by the auditors or others who have experience assessing the controls in this specific company because the assessment of internal control risks depends on the actual

Table 2. Internal Control Concerns, Definitions, and States.

	Internal Control Concern	Definition	States
Process	Incomplete acquisition process	Incomplete acquisition information or fraudulent purchase	Included in payment report or not
	Calculation errors	Tax was not applied correctly	Included in report or not
	Recording errors	Payment date was before issuing date	Affects movement of money or not
		Payment date was before recording date	
		Issuing date was before recording date	
		Duplicate transactions	
	Revision of information	Manager does not check periodically	Sporadic or never revised
Staff	Inadequate segregation of duties	Authorization was made by bookkeeper	High hierarchy or low hierarchy
		Authorization was made by cashier	
		Recording was made by cashier	
System	Lack of system monitoring	There is no verification flowchart	Sporadic or always
		There is no supervision of operations	
		Lack of revision of the systems	
	Data accessibility	There is no security policies in the system	Protected or not

Table 3. Matrix of Risks.

Internal Control Concern	States
Incomplete acquisition process	Included in payment requirement report or not
Calculation errors (taxes, etc.)	Included in report or not
Recording errors (date, amounts, etc.)	Affect the movement of money or not
Revision of the information	Sporadic or never revised
Segregation of duties	High hierarchy or low hierarchy
Lack of supervision	Sporadic or always
Data accessibility	Protected or not

Table 4. Matrix of Risk for the System Section.

	Lack of Supervision	Data Accessibility
Accidental Errors	Sporadic or always	Protected or not
Deliberate	Sporadic or always	Protected or not

current and past characteristics of the firm in those areas. The output of the fuzzy logic assessment is interpreted as the magnitude of risk that the company has at that point in time, expressed as a percent.

Two simulated cases are used to illustrate the fuzzy logic process. The first case considers a random assignment of the degrees of risk, and the second case considers simulated auditor responses to a questionnaire, shown in the Appendix at the end of this chapter. The simulation includes 150 companies classified as low, medium, or high risk. This case is used to measure the effectiveness of fuzzy logic, so Type I and Type II errors are calculated after assessing the risk levels for those companies.

3.3. Case I: Random Assignment

To show how fuzzy logic works, an example is created for the System area of the payment cycle. In Table 5, each node (A–D) corresponds to the matrices associated with the level of risk for two possible internal control weaknesses (lack of supervision and data accessibility) and two types of problems (accidental errors or deliberate acts). Matrix A and Matrix B are used here as an example.

Matrix A indicates the level of risk when the lack of supervision is accidental. In this case, the lack of supervision could be sporadic or it could occur always. In the belief matrix in Table 6, auditors assign lower risk when there is a sporadic lack of security, but higher risk when there is always a lack of security.

Matrix B indicates the level of risk when the lack of security is a deliberate act. As in Matrix A, the problem could occur always or it could be sporadic.

Table 5. Risk Assessed Based on Belief Risks.

	Lack of Supervision	**Data Accessibility**
Accidental Errors	A = ?	C
Deliberate Acts	B = ?	D

Table 6. Belief Matrix for Accidental Errors in Lack of Supervision.

Matrix A	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
Always									0.5		
Sporadic		0.2	0.4								

Table 7. Belief Matrix for Deliberate Act in Lack of Supervision.

Matrix B	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
Always										0.5	0.7
Sporadic		0.2	0.5								

As shown in Table 7, auditors assign higher risk to the event when the deliberate lack of supervision always occurs and lower risk when the lack of supervision is sporadic.

Clearly, if the lack of supervision is deliberate, it is more likely that the systems can decrease the credibility of the information that is processed in the payment process.

To calculate the risks for each matrix, the fuzzy sets must be defined. In the case of matrix A and B, the fuzzy sets are

Matrix A	Matrix B
– A1 = 0.5/0.8	– B1 = 0.5/0.9 + 0.7/1
– A2 = 0.2/0.1 + 0.4/0.2	– B2 = 0.2/0.1 + 0.5/0.2

Thus, the assessed risk for each level of control risk for lack of supervision is determined as follows:

For matrix A: (0.5*0.8 + 0.2 * 0.1 + 0.4 * 0.2)/(0.5+0.2+0.4) = **0.45**
For matrix B: (0.5 * 0.9 + 0.7 * 1+0.2*0.1+0.5*0.2)/(0.5+0.7+0.2+0.5) = **0.67**

Thus, according to the beliefs of the auditors, there is a **45%** risk associated with accidental lack of supervision, but a **67%** risk when the lack of supervision is produced by a deliberate act.

Applying the same process for Matrixes C and D, we obtain four risks, and the risk associated with the System area is the maximum value among Matrixes A–D. For example, if the risk for Matrix C is 0.34 and the risks for Matrix D is 0.48, then the risk for the System process is: max (0.45, 0.67, 0.34, 0.48) = **0.67**.

To assess the risk for the entire payment cycle, the risks for the Process and Staff areas must be determined as well. The overall level of risk for the payment cycle is then estimated by considering the maximum value for each category. Assuming that the Process area has a risk of 0.84, the Staff area has a risk of 0.5, and the System area has a risk of 0.67, it becomes clear that the payment cycle is very risky and has clear defects in internal controls. Therefore, the firm should increase the number of controls in the payment department. As shown in Table 8, it is also possible to compute the level of assurance of the controls (LAC) for the company based on the following definition:

LAC = 1−R, where R is the level of the risk for the specific control

Table 8 indicates that Process is the most risky subsystem for payments. Hence, managers should develop additional controls to avoid possible threats that could affect the reliability and accuracy of payment information that is processed in the company. The other two subsystems are also fairly risky, so managers should consider increasing the number of controls and verifications for those areas as well.

Table 8. Risk by Areas and LAC by Area.

	Risk	**LAC**
Process	0.84	0.16
Staff	0.50	0.50
Systems	0.67	0.33

3.4. Case II: Effectiveness of the Model

To assess the effectiveness of the fuzzy logic model in the payment cycle, information is simulated for three different types of companies, assuming that they were previously assessed as low, medium or high risk by their auditors. Those companies are then classified depending on the probability that they will suffer a violation in their process. Thus, the low-risk companies have a high probability to be affected by a low-risk event (0−0.10), the medium-risk companies are more likely to be affected by a medium-risk event (0.11−0.49), and it is more probable that the high-risk companies would be affected by a high-risk event (0.5−1). Table 9 shows the means of the simulated data for each group of companies.

Table 9 shows that the low risk companies have a about a 76% probability to be altered by a low-risk event, the medium risk companies have a 77% average probability to be affected by a medium-risk event and the high-risk companies have 75% on average to be affected by a high risk event.

To assess the power of fuzzy logic, the rate of Type I errors (false negative) and Type II errors (false positive) are calculated. The results are shown in Table 10.

There is 0% false positive rate and 10% of false negative rate for low-risk firms, which is a good result considering that 10% of the low-risk companies were classified as having a medium- or high-risk of being altered by an event. This situation does not happen in the cases of medium or high-risk companies.

Table 9. Mean Risk for Low-Risk, Medium-Risk, and High-Risk Companies.

Companies	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
Low risk	0.77	0.76	0.02	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.03
Medium risk	0.03	0.03	0.76	0.76	0.78	0.03	0.02	0.03	0.02	0.03	0.03
High risk	0.03	0.02	0.02	0.03	0.03	0.73	0.78	0.77	0.75	0.76	0.78

Table 10. Effectiveness of Fuzzy Logic Assessment.

	False Positive Rate	**False Negative Rate**	**Specificity (%)**	**Sensitivity (%)**
Low-risk firms	0	10	100	90
Medium-risk firms	0	0	100	100
High-risk firms	0	0	100	100

Sensitivity is the detection capacity of fuzzy logic. In this case, the test shows a high capacity (90%) for detecting events in the group defined as low risk companies. In that case, some of the low risk companies will have underestimated their internal control needs, which implies that more revisions and assessment of these processes should be considered. This is not negative; the extra revisions may find that the area/process was not risky as suggested by using fuzzy logic. It would be a bad result if a medium-risk or high-risk company was defined as low risk because it would imply that the perception of strong internal controls in this area is wrong, but this condition was not found.

3.5. Detecting Anomalies in Accounting Data

When dealing with huge amount of accounting data, the first step is to describe the data using summary statics, such as frequencies and plots, and then to detect duplicates, outliers and leverage points.

The real company used in this part of the study sells credit cards, savings cards, life insurance, and house insurance. The period analyzed runs from November 3, 2009 through April 30, 2010. In total, there are 123 days with an average of 1,000 transactions per day.

Three files are used in this study. The first file includes employee descriptions, branch descriptions, and the summary transactions file, which contains the canceled, inactive, or reimbursed products, linked products, and products that were bought for other employees.

Table 11 shows the summary by product. It indicates that credit cards had the greatest sales volume during this period, but the number of the cancelation was greater for life insurance. Table 12 presents the descriptive analysis for each variable.

Table 11. Description of Variables by Products.

Product	Branches	Employees	Sales	Cancellation	Reimbursed	Linked	Inactive	Claimed	Sales by the Same Employee
Credit card	3,530	11,750	1,081,857	44,321	612	183,633	32,003	23,543	10,367
Save card	3,147	10,233	607,189	40,436	2,151	112,805	12,625	12,973	2,400
House insurance	3,190	15,771	158,059	18,347	915	21,253	1,505	4,172	3,469
Life insurance	3,485	19,451	405,539	98,960	3,623	75,110	6,123	10,023	2,780

Table 12. Descriptive Analysis by Variable.

Variable	Std. Dev	Sum	Kurtosis	Skewness
Cancelation	0.2858	202,064	6.2467	2.8717
Reimbursed	0.0568	7,301	303.5430	17.4798
Linked	0.3794	392,801	0.9460	1.7164
Inactive	0.1505	52,256	38.1317	6.3350
Claimed	0.1483	50,711	39.4443	6.4377
Sale to same employee	0.0915	19,016	113.4692	10.7457

The analysis that describes the data (normality, homoskedasticity, and collinearity) and detects anomalies (outliers and leverage points) in the data is considered from two points of view: branches (n = 3,754) and employees (n = 21,900).

The outliers correspond to observations with greater studentized residuals, which means that each residual is divided by its standard deviation. First, the residuals are calculated. Then, the observations are ranked according to the absolute studentized residuals. Data points are considered to be outliers if their studentized residuals are greater than two.

Leveraged points correspond to observations for which the distance from the ith observation to the average of all x observations is greater than $(2k+2)/n$, where k is the number of predictors in the regression model, and n, the number of observations. Outliers with greater leverage are defined as problems that require subsequent analysis.

3.5.1. Data Description: Branches and Employees. Normality, heteroscedasticity and collinearity are tested. The results show a strong presence of heteroscedastic errors in the model. To address this issue, a transformation is applied to the response variable, but the data continue to have non-constant variances. Furthermore, the sample is not normally distributed. The assumption is that the high presence of outliers interferes in the characteristics of the data. Therefore, a robust regression is needed because it will be not be distorted by the presence of unusual observations.

3.5.2. Detection of Outliers and Leverage Points: Branches and Employees. Outliers and leverage points are detected using the STATA program. The results indicate that 119 branches are considered to be outliers, and of those branches, 112 have a high leverage. The summary statistics for this analysis are detailed in Table 13.

The criteria shown in Table 13 are relevant for determining those points that have both high leverage and high residuals. Fig. 1 shows the relationship between Leverage and R^2 for branches that are outliers and have high leverage. Fig. 2 shows the same relationship for employees.

3.6. Authorization Limit Problem

It is possible that employees split payments to circumvent transaction limits and record duplicate payments. To solve this problem, duplicate payments are

Table 13. Summary Statistics for Outliers and Leveraged Points.

	Cancel	Sales	Reimbursed	Inactive	Complained	Linked	Sales per Employee	Sales per Different Employee	Q Employee
Sum	38,456	256,403	1,225	4,359	6,582	55,037	6,315	1,730	37,195
Average	323.1	2,154.6	10.29	36.63	55.31	462.4	53.06	14.53	312.56
Max	1,221	6,670	61	126	263	2,427	123	69	1,980

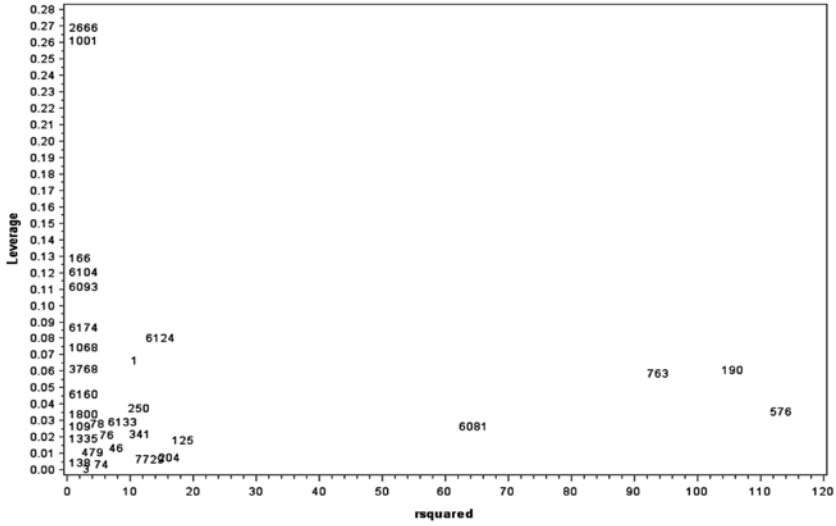


Fig. 1. Branches with High Leverage and Outliers.

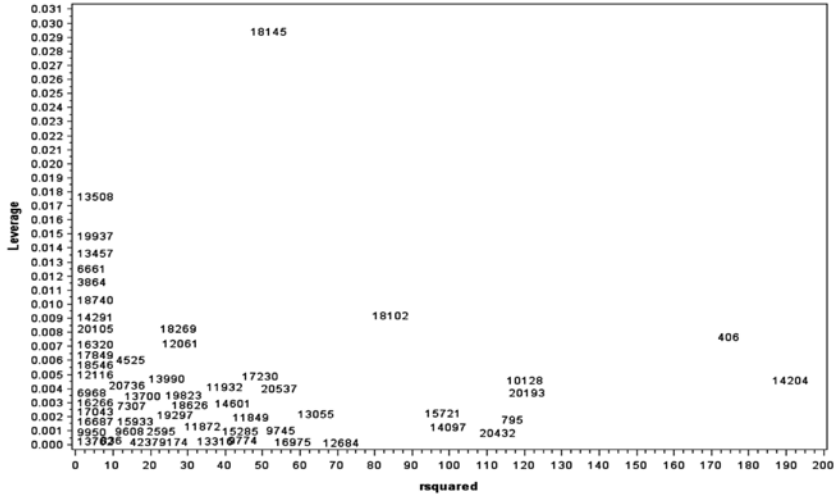


Fig. 2. Employees with High Leverage and Outliers.

identified, and then the associated authorization limits and segregation of duties are analyzed. First, the process must be understood through a flowchart and descriptive statistics. In that process, anomaly indicators are defined by vendors, transactions, and employees.

Data: Two data sets are needed for this analysis: payments and authorization limits. The data cover January through September 2010, which correspond to 4,060,803 records covering 271 days for 173,128 different vendors and 11,076 employees who approved transaction.

Methodology: First, the frequencies and means in the database are determined. Next, duplicates payments and split payments are identified. Finally, to give insights about the data, it is relevant to assess the percentage of errors (or problems) that is exhibited by the data. Using indicators (Indicator per transaction) that goes from 0 to 5, we can see which transaction, need further revision.

$$\text{Indicators per transactions} = \text{LimitPP} + \text{LimitPM} + \text{Segregation} + \text{Dates} + \text{Split}$$

where

- LimitPP is a dummy variable that takes 1 if the employee associated with this transaction exceeds the limit per payment, and 0 otherwise.
- LimitPM a dummy variable that takes 1 if the employee associated with this transaction exceeds the limit per month, and 0 otherwise.
- Segregation the same approver and recorder, is a dummy variable that takes 1 if the person who make the approval is the same as the person who record the transaction, and 0 otherwise.
- Dates the error in recorded date, and is a dummy variable that takes 1 if there are error in the recorded date, and 0 otherwise.
- Split the split payments behavior is the dummy variable that takes 1 if there are issues with the split payment behaviour. (same vendor, same approver, same date of approval, same date of paying, and same number of document).

Results: Out of 4,060,803 transactions, 2,479,555 transactions have one problem, 14,604 transactions have two problems, 139 transactions have three problems, and two transactions have four problems next analysis, should include finding a pattern, so it is easier to understand which transactions are much more susceptible to exhibit errors. The empirical constraints in this study include matching data with different timing, delayed feedback from companies, and difficulty getting ideal data to analyze.

4. Conclusion

According to COSO, internal control is designed to assist organizations in achieving their objectives. The COSO board also recognizes that the monitoring of controls is a key component of the effectiveness of the internal control assessment. In fact, the more frequently auditors monitor internal control, the more efficient their internal control assessment will be. However, the large volume of data and the limited time available to make decisions, auditors face a problem during the audit when they look for patterns.

Given the importance of evaluating a client firm's internal controls, this chapter presents a methodology based on fuzzy logic that can be used to assess internal controls for a firm's payment cycle and it was also presented an statistical tool to detect outliers in credit card data. This statistical methodology can reduce the impact of problems with outliers or heteroscedasticity that impede the use of traditional statistical methodologies.

Fuzzy logic allows risk to be assessed based on the belief of experts, people who know the actual problems that exist in a particular company's unique situation. In addition, fuzzy logic can help auditors to deal with the uncertainties involved

with the internal control assessment. The output of the fuzzy logic approach permits auditors to assess the degree of assurance offered by the firm's internal controls in a more realistic way. Based on beliefs, the output obtained from the fuzzy logic method is a number between 0 and 1 that can be interpreted as the level of risk for a particular process. Auditors must make judgments about each area, and these results can become the basis for a formal document that will contribute to managers' understanding of the effectiveness of the internal controls in each specific area. The fuzzy logic approach can also be modified as needed because it is easy to change the parameters as the company grows or develops a more robust internal control model.

This fuzzy logic model permit auditors to evaluate the fragility of a firm's internal controls, so it can help managers to be proactive in reducing internal control threats. The existence of fragile internal controls decreases their level of assurance for the company, and a low level of assurance reduces the credibility and effectiveness of the resulting information. Indeed, evaluating and reducing threats in internal controls helps to increase the usefulness of financial information not only for external users of financial information, but also for managers' decision-making processes.

On the other hand, the statistical tools presented in this chapter, to analyze credit card data, helps auditors to understand the data even thought they are dealing with a huge amount of data. First of all, the detection of outliers, using errors studentized, helps auditors to detect those transactions that are unusual, and then, to conclude about the impact of the errors presented, it is possible to generate conclusions without being interfered by outliers (using Robust Regressions).

References

Akhter, F., Hobbs, D., & Maamar, Z. (2005). A fuzzy logic-based system for assessing the level of business-to-consumer (B2C) trust in electronic commerce. *Expert Systems with Applications*, *28*(4), 623–628.

Arens, A., Elder, R. J., & Beasley, M. S. (2010). *Auditing and assurance services: An integrated approach*. London: Pearson Education.

Baldwin, A. A., Brown, C. E., & Trinkle, B. S. (2006). Opportunities for artificial intelligence development in the accounting domain: The case for auditing. *Intelligent Systems in Accounting, Finance and Management*, *14*(3), 77–86.

Barnett, V. (1978). The study of outliers: Purpose and model. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *27*(3), 242–250.

Bayrak, M. Y., Çelebi, N., & Taşkin, H. (2007). A fuzzy approach method for supplier selection. *Production Planning and Control*, *18*(1), 54–63.

Bevilacqua, M., & Petroni, A. (2002). From traditional purchasing to supplier management: A fuzzy logic-based approach to supplier selection. *International Journal of Logistics Research and Applications*, *5*(3), 235–255.

Biggs, S. F., & Mock, T. J. (1983). An investigation of auditor decision processes in the evaluation of internal controls and audit scope decisions. *Journal of Accounting Research*, *21*(1), 234–255.

Chang, S.-I., Tsai, C.-F., Shih, D.-H., & Hwang, C.-L. (2008). The development of audit detection risk assessment system: Using the fuzzy theory and audit risk model. *Expert Systems with Applications*, *35*(3), 1053–1067.

Che, Z. H., Wang, H. S., & Chuang, C.-L. (2010). A fuzzy AHP and DEA approach for making bank loan decisions for small and medium enterprises in Taiwan. *Expert Systems with Applications*, *37*(10), 7189–7199.

Chen, Y., & Leitch, R. (1999). An analysis of the relative power characteristics of analytical procedures. *Auditing: A Journal of Practice & Theory*, *18*(2), 35–69.

Comunale, C. L., & Sexton, T. R. (2005). A fuzzy logic approach to assessing materiality. *Journal of Emerging Technologies in Accounting*, *2*(1), 1–15.

Cook, D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, *74*(365), 169–174.

Cooley, J. W., & Hicks, J. O., Jr. (1983). A fuzzy set approach to aggregating internal control judgements. *Management Science*, *29*(3), 317–334.

Dubinsky, B., & Warner, C. (2008). Uncovering accounts payable fraud using "fuzzy matching logic" Part 1. *Business Credit*, *110*(3), 6–9.

Famuyiwa, O., Monplaisir, L., & Nepal, B. (2008). An integrated fuzzy-goal-programming-based framework for selecting suppliers in strategic alliance formation. *International Journal of Production Economics*, *113*(2), 862–875.

Glackin, C., Maguire, L., McIvor, R., Humphreys, P., & Herman, P. (2007). A comparison of fuzzy strategies for corporate acquisition analysis. *Fuzzy Sets Systems 158*(18), 2039–2056.

Kinney, W. R., Jr. (1978). ARIMA and regression in analytical review: An empirical test. *The Accounting Review*, *53*(1), 48–60.

Knechel, W. R. (2007). A simulation study of the relative effectiveness of alternative analytical review procedures. *Decision Sciences*, *17*(3), 376–394.

de Korvin, A., Shipley, M. F., & Omer, K. (2004). Assessing risks due to threats to internal control in a computer-based accounting information system: A pragmatic approach based on fuzzy set theory. *Intelligent Systems in Accounting, Finance and Management*, *12*(2), 139–152.

Krishna, P. R., & Kumar De, S. (2001). A fuzzy approach to build an intelligent data warehouse. *Journal of Intelligent & Fuzzy Systems*, *11*(1–2), 23–32.

Lee, C. C. (1990a). Fuzzy logic in control systems: Fuzzy logic controller − Part I. *IEEE Transactions on Systems, Man and Cybernetics*, *20*(2), 404–418.

Lee, C. C. (1990b). Fuzzy logic in control systems: Fuzzy logic controller − Part II. *IEEE Transactions on Systems, Man and Cybernetics*, *20*(2), 419–435.

Lee, S.-H., Lee, S., Moon, K. (2010). A fuzzy logic-based approach to two-dimensional automobile warranty system. *Journal of Circuits, Systems and Computers*, *19*(1), 139–154.

Lee, V. C. S., & Wong, H. T. (2007). A multivariate neuro-fuzzy system for foreign currency risk management decision making. *Neurocomputing*, *70*(4–6), 942–951.

Lenard, M. J., Alam, P., & Booth, D. (2000). An analysis of fuzzy clustering and a hybrid model for the auditor's going concern assessment. *Decision Sciences*, *31*(4), 861–884.

Levy, J., & Yoon, E. (1995). Modeling global market entry decision by fuzzy logic with an application to country risk assessment. *European Journal of Operational Research*, *82*(1), 53–78.

Li, S.-T., & Cheng, Y.-C. (2009). An enhanced deterministic fuzzy time series forecasting model. *Cybernetics and Systems: An International Journal*, *40*(3), 211–235.

Loebbecke, J. K., & Steinbart, P. J. (1987). An investigation of the use of preliminary analytical review to provide substantive audit evidence. *Auditing: A Journal of Practice & Theory*, *6*(2), 74–89.

Magni, C. A., Malagoli, S., & Mastroleo, G. (2006). An alternative approach to firms' evaluation: Expert systems and fuzzy logic. *International Journal of Information Technology and Decision Making*, *5*(1), 195–225.

Omura, T., & Willett, R. (2006). Using automated equilibrium correction modelling in analytic review. *Managerial Auditing Journal*, *21*(2), 207–223.

Pathak, J., Vidyarthi, N., & Summers, S. L. (2005). A fuzzy-based algorithm for auditors to detect element of fraud in settled insurance claims. *Managerial Auditing Journal*, *20*(6), 632–644.

Rangone, A. (1997). Linking organisational effectiveness, key success factors and performance measures: An analytical framework. *Management Accounting Research*, *8*(2), 207–219.

Tam, C. M., Tong, T. K. L., Leung, A. W. T., & Chiu, G. W. C. (2002). Site layout planning using nonstructural fuzzy decision support system. *Journal of Construction Engineering and Management*, *128*(3), 220–231.

Tang, X., Lau, H., & Ho, G. (2008). A conceptual fuzzy-genetic algorithm framework for assessing the potential risks in supply chain management. *International Journal of Risk Assessment and Management*, *10*(3), 263–271.

Wilson, A. C., & Colbert, J. (1989). An analysis of simple and rigorous decision models as analytical procedures. *Accounting Horizons*, *3*(4), 79–83.

Yaffee, R. (2002). Robust regression analysis: Some popular statistical package options. Retrieved from www.nyu.edu/its/socsci/Docs/RobustReg2.pdf | 2003-04-30

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, *15*(2), 642–656.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353.

Zadeh, L. A. (1988). Fuzzy logic. *Computer*, *21*(4), 83–93.

Appendix: Questionnaire

Please complete the following questionnaire to show your assessment of the company's level of risk for internal control concerns from accidental errors or deliberate acts related to the company's Acquisition Process. You should use numbers in the interval (0,1). You do not need to complete the entire matrix; only show where you have greater or lower concerns.

Case 1: Accidental Error

The columns from 0 to 1 represent the weights of the risk in each area. For instance, if there is high probability of an accidental problem in the Acquisition Process, and this error would be included in the payment report, fill in the first row indicating your risk perception in this situation. If the errors would not be included in the payment report, fill in the second row.

Case 2: Deliberate Act

Follow the same procedure used in Case 1, but in this case, consider the risk of an internal control concern about a deliberate act that may affect the Acquisition Process.

Example: In this example, the bold numbers show a participant's assessment of the level of risk for internal control concerns from accidental errors and deliberate acts.

	Acquisition Process											
Case 1:	Accidental error	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
	Included in payment requirement report								**0.1**		**0.2**	
	Not included in report		**0.5**									
Case 2:	Deliberate act	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
	Included in payment requirement report							**0.4**			**0.7**	
	Not included in report		**0.1**									

This page intentionally left blank