THE APPLICATION OF EXPLORATORY DATA ANALYSIS IN AUDITING

By Qi Liu

A dissertation submitted to the

Graduate School- Newark

Rutgers, The State University of New Jersey

in partial fulfillment of requirements

for the degree of

Doctor of Philosophy

Graduate Program in Management

Written under the direction of

Professor Miklos A. Vasarhelyi

and approved by

_____

Professor Miklos A. Vasarhelyi

_____

Professor Alex Kogan

_____

Professor Michael Alles

_____

Professor Ingrid Fisher

Newark, New Jersey

October 2014

ABSTRACT OF THE DISSERTATION

The Application of Exploratory Data Analysis in Auditing

By Qi Liu

Dissertation Chairman: Professor Miklos A. Vasarhelyi

Exploratory data analysis (EDA), which originated centuries ago, is a data analysis approach that emphasizes pattern recognition and hypothesis generation from raw data. It is suggested as the first step of any data analysis task for exploring and understanding data, and has been applied in many disciplines such as Geography, Marketing, and Operations Management. However, even though EDA techniques, such as data visualization and data mining, have been used in some procedures in auditing, EDA has not been employed in auditing in a systematical way. This dissertation consists of three essays to investigate the application of EDA in audit research. The study contributes to the auditing literature by identifying the importance of EDA in auditing, proposing a framework to describe how auditors could apply EDA to auditing, and using two cases to demonstrate the benefits that auditors can gain from EDA by following the proposed framework.

The first essay identifies the value of EDA in auditing and proposes a conceptual framework to identify EDA's potential application areas of EDA in various audit stages in both the internal and external audit cycles, describe how auditors can apply various EDA techniques to fulfill different audit purposes, and introduce a recommended process for auditors to implement EDA. In addition, this essay also discusses how EDA can be integrated into a continuous auditing system.

The second essay examines the use of EDA in an operational audit. Traditional EDA techniques, such as descriptive statistics, data transformation, and data visualization, are applied in this credit card retention case. Descriptive statistics can reveal the distribution of the data, and data visualization techniques can display the distribution in an effective way so that auditors can easily identify patterns hidden in the data. By integrating these EDA techniques in audit tests, many critical risky issues, such as negative discount, inactive representatives, and short calls, are detected.

With the rapid development of modern data analysis, EDA techniques have been greatly enriched. Besides the traditional EDA techniques, some data mining techniques can also be used to fulfill EDA tasks. The third essay investigates the application of two data mining techniques on a Medicare dataset to assess fraud risk. Specifically, this essay utilizes clustering techniques to detect abnormal Medicare claims in terms of claim payment amount and Medicare beneficiaries' travel distances and hospital stay periods, and applies association analysis to analyze doctors' diagnoses and performed procedures to identify abnormal combinations.

In summary, this dissertation attempts to contribute to auditing literature by identifying the value that EDA can add to auditing, and illustrating some applications to demonstrate how auditors can benefit from EDA.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude and appreciation to my esteemed dissertation advisor, Professor Miklos A. Vasarhelyi. He gave me the opportunity to study at Rutgers and introduced me to the world of research. Without him, this dissertation would be impossible and I would not have gone so far in academia. He offered me great help not only on guiding my research but also on encouraging, supporting, and trusting me all along my Ph.D. studies

I am very grateful to Prof. Alex Kogan, Prof. Michael Alles, and Prof. Ingrid Fisher for serving on the committee and helping me improve my dissertation with their highly constructive comments. I am also indebted to Dr. Trevor Stewart for his valuable comments on this dissertation, and to Dr. John Peter Krahel and Dr. Ann Medinets for their support for the manuscript.

I would also like to express my special thanks to Dr. Gerard Brennan, Dr. Victoria Chiu, Prof. Glen Gray, and Barbara Jensen. I learnt so much from the collaboration with them. Thanks also go to my friends and colleagues at Rutgers: Jun Dai, Pei Li, Tiffany Chiu, He Li, Feiqi Huang, Hussein Issa, Kyunghee Yoon, Desi Arisandi, Deniz Appelbaum, Paul Byrnes, et al. It was a great pleasure to work with all of them.

For my family, no words can describe my great gratitude. Their endless love and unwavering support helped me immensely. I particularly than my husband Yuan for constantly encouraging and supporting me, my daughter Mabel for always bringing me greatest joy and sunshine, and my parents for their unconditional love leading to my success. I dedicate this work to all of you.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

1.1 Background

This dissertation incorporates three essays investigating the application of Exploratory Data Analysis (EDA) in the auditing domain. Chapter one introduces the motivation and methodology of this thesis and provides an extended literature review of the concept of EDA and its enabling techniques. The three essays are included in chapter two, three and four, respectively. The last chapter concludes the dissertation by summarizing the findings, discussing the limitations, and pointing out future research areas.

EDA, which originated centuries ago, is a data analysis approach that emphasizes pattern recognition and hypothesis generation. It can be used as the first step of any data analysis task to explore and understand the data (De Mast and Kemper, 2009a). Audit is a data-intensive process. Auditors can obtain valuable audit evidence by analyzing clients' data. Therefore, data analysis plays an important role in the audit process. However, even though EDA has been applied in many disciplines, such as Geography, Marketing, and Operations Management (Chen et al., 2011; Nayaka and Yano, 2010; Koschat and Sabavala, 1994; Wesley et al., 2006; De Mast and Trip, 2007), it has not been employed in auditing in a systematic way. Only a few EDA techniques, such as data visualization and data mining, have been used in some audit procedures.

This dissertation investigates the systematic application of EDA in the auditing domain. Therefore, the first essay proposes a conceptual framework to guide auditors' application of EDA. Particularly, the framework illustrates when EDA can be applied in an audit cycle, how various EDA techniques can benefit auditors in different audit

procedures, and specifically what activities auditors need to do in order to guarantee the best practice of EDA. Besides the application of EDA in traditional audit settings, this essay also discusses how EDA can be integrated into a continuous auditing environment.

The second essay provides a field study of EDA application in an operational audit. A real dataset from an international bank in Brazil is used in this field study that applies descriptive statistics and data visualization techniques to investigate the data related to phone calls made by bank clients to negotiate their credit card annual fees. Many critical risky issues that cannot be identified by standard audit tests, such as negative discount, inactive agents and short calls, are detected in the EDA process.

The third essay applies the proposed EDA process in the audit planning stage to assess fraud risk in 2010 inpatient Medicare claims. Descriptive statistics, cluster analysis, and association analysis are performed in this case study. By extending the analysis scope, descriptive statistics can discover abnormal claims that may be ignored by conventional audit procedures. Cluster analysis is conducted to identify abnormal claims based on claim payment amounts and beneficiaries' travel distances and hospital stay periods. Compare to conventional audit procedures and descriptive statistics analysis of a single variable, cluster analysis can not only reveal more hidden risk areas, but also narrow the scope for substantive tests to the most suspicious cases. Association analysis is applied to analyze doctors' diagnoses and performed procedures to identify abnormal combinations that can be considered as risk indicators.

1. 2 Overview of Exploratory Data Analysis

Exploratory data analysis (EDA) is a statistical data analysis approach that emphasizes pattern recognition and hypothesis generation. The concept originates from

nineteenth-century empiricism (Mulaik, 1985), but the term EDA stems from the work of John Tukey and his colleagues about four decades ago (Tukey, 1969, 1977, 1986a, 1986b, 1986c; Tukey and Wilk, 1986). Tukey (1977) characterizes EDA as (1) a philosophy or attitude, rather than a fixed set of formal procedures; (2) a focus on the comprehensive understanding of the data to extract the story behind the data; (3) the use of simple descriptive measures to summarize and re-express the data; (4) an emphasis on graphic representations of the data; (5) flexibility in both tailoring the analysis to the structure of the data and responding to the uncovered patterns; and (6) a focus on tentative model-building and hypotheses-generation. The goal of EDA is not to draw conclusions on predefined questions, but to explore the data for clues to inspire ideas and hypotheses. The role of the researchers in EDA is to analyze the data in as many ways as possible until a plausible "story" of the data appears. Therefore, EDA is speculative (pursuing potential clues), and open-ended (leaving the support of the hypotheses generated to Confirmatory Data Analysis) (De Mast and Kemper, 2009a).

Confirmatory Data Analysis (CDA) is a widely used data analysis approach that contrasts fundamentally with EDA. CDA emphasizes experimental design, significance testing, estimation, and prediction (Good, 1983). The distinction between EDA and CDA is first explicitly discussed by Tukey (1997). He likens EDA to detective work, which is the process of gathering evidence, and compares CDA to the court trial, which mainly focuses on evaluating the evidence collected by the detectives. Therefore, in practical data analysis tasks, CDA usually follows or alternates with EDA as needed (Mosteller and Tukey, 2000).

In order to get a better understanding of EDA, it is critical to recognize the differences between EDA and CDA. Those differences are summarized in Table 1 from the aspects of Reasoning Type, Goal, Applied Data and Tools, Advantages and Disadvantages. In terms of reasoning type, EDA is an abductive data analysis approach (Yu, 1994) that begins with observations and researchers' background knowledge (such as domain knowledge and widely-accepted theories and ideas) to generate hypotheses describing the most likely explanations of the patterns identified from the data. By contrast, CDA is a deductive (also named top-down) data analysis approach starting from a predefined hypothesis, then collecting data to evaluate the hypothesis. EDA is usually applied to observation data collected without well-defined hypotheses, whereas CDA is often used on the data obtained via formally designed experiments (Good, 1983) or naturally collected data with certain constraints[1]. The commonly used tools to perform EDA are descriptive statistics, such as frequency distribution, mean, standard deviation, etc., and data visualization techniques like pie chart, bar chart, scatter plot and so forth. CDA is often conducted using traditional statistical tools of inference, significance, and confidence, such as p-values, confidence intervals, and so on. One advantage of EDA is that it does not require strong predetermined assumptions. However, this does not mean that EDA is conducted without any reference (Yu, 2010). In fact, when performing EDA, researchers usually employ research questions and their domain knowledge to define the

[1] Formerly, when collecting data was expensive, researchers started their research from specific hypotheses and collected as little data as possible to verify their hypotheses. The data collected in this kind of research are experimental data. Even though currently there are still plenty of researchers who test their hypotheses in this way, with the increasing availability of data, many researchers now use available observation data to test their hypotheses. In this case, they usually add some control variables to select observation data in certain conditions in order to satisfy the assumptions included in the hypotheses.

scope of EDA, select the most appropriate EDA techniques, and choose the most likely explanations from numerous alternatives to explain the phenomena shown in the data and develop hypotheses. Another advantage of EDA is that it can promote a deeper understanding of the data by identifying prominent patterns. Its disadvantages include that it does not provide definitive answers and that it is difficult to avoid bias produced by overfitting due to the high dependence on the data. By contrast, CDA's advantages are that it can provide precise results for hypothesis testing and that it has well-established theories and methods. The disadvantages of CDA are that it may require some unrealistic assumptions causing misleading impressions in less than ideal circumstances and that it is difficult to notice unexpected results since its focus is the predefined model.

Table 1: Comparison of EDA and CDA

	Exploratory Data Analysis (EDA)	Confirmatory Data Analysis (CDA)
Reasoning Type	Abductive	Deductive
Goal	Pattern recognition and hypothesis generation	Estimation, modeling, and hypothesis testing
Applied Data	Observation data (data collected without well-defined hypothesis)	Experimental data (data collected through formally designed experiments), or observation data under certain condition (with control variable)
Tools	Descriptive statistics and data visualization	Traditional statistical tools of inference, significance, and confidence
Advantages	• No strong, pre-determined assumptions needed • Promotes deeper understanding of the data	• Precise • Well-established theory and methods
Disadvantages	• No conclusive answers • Difficult to avoid bias produced by overfitting	• May require unrealistic assumptions • Difficult to notice unexpected

		results

Compared to EDA, CDA is closer to traditional statistical inference. Therefore, it has attracted much more attention from researchers. The literature on CDA is more elaborate than the literature on EDA, both in volume of research articles and in depth of theory development. (De Mast and Trip, 2007; De Mast and Kemper, 2009a). Obviously, EDA is underappreciated in statistics and social science research (Horowitz, 1980; De Mast and Kemper, 2009b). However, this does not mean that EDA is unimportant. Actually, EDA may now be considered the prerequisite of CDA because without EDA, the results of CDA can be deceptive. For example, lack of EDA may lead to the generation of inappropriate hypotheses. Even though these hypotheses are tested to be significant in CDA processes, the conclusion is still improper. In summary, the relationships between EDA and CDA are: (1) both techniques are important; (2) EDA comes first; (3) any given study should combine both (Tukey, 1980, 1986d).

Another concept that is usually confused with EDA is Descriptive Data Analysis (DDA), which summarizes the data in a number of descriptive statistics. The main concern of DDA is the presentation of data to reveal salient features. It uses summary statistics, such as mean and standard deviation, to suppress uninformative features of the data in order to reveal the significant features. Due to the limitation of human cognitive ability, raw datasets are too complex for human understanding. DDA is designed to match the salient features of the dataset to human cognitive abilities because these techniques are usually simple and easy to comprehend (Good, 1983). DDA can be seen as part of EDA. EDA goes further than DDA because, in addition to presenting the salient

features of the data, EDA also aims to formulate hypotheses that can explain these salient features (De Mast and Trip, 2007).

Four main themes that appear through the EDA process include: Resistance, Residuals, Re-expression, and Revelation (Mosteller and Tukey, 2000). EDA usually uses resistance measures, such as median, to present the data. These measures are insensitive to outliers or skewed distribution, so they can better reflect the main body of the data. In EDA, researchers analyze residuals to distinguish dominant and unusual behavior in the data. Re-expression, such as standardization or normalization, is used in EDA to rescale the data in order to improve interpretability. Revelation, indicating data visualization, is the major contribution of EDA.

In practice, EDA is not a purely statistical task; it needs to combine statistical methods with human interpretation (also called mental models). Specifically, the steps to apply EDA in practical problem-solving issues include: (1) displaying the data; (2) identifying salient features; and (3) interpreting salient features (De Mast and Kemper, 2009). In the first step, various techniques are employed to reveal the data distribution in order to facilitate researchers' ability to recognize the hidden patterns. The selection of data to perform EDA is based on research questions, and the choice of techniques to display data depends on researchers' background knowledge. Once the distribution of the data is displayed, the researchers start to look for salient features. Typically, researchers expect the data to be neutral, uniformly or normally distributed. Deviations from expectations, such as outliers, are considered salient. After identifying the salient features, the next step is to theorize and speculate about the reasons for these patterns. Usually, researchers discuss the identified patterns with the domain experts, and come up

with the hypotheses grounded in their domain knowledge. Until now, the EDA process ended there. Then subsequent CDA studies will be conducted to validate these hypotheses, thus delivering solutions to the issues. The process of EDA in practical problem-solving issues and the role of mental models in this process are illustrated in Figure 1.

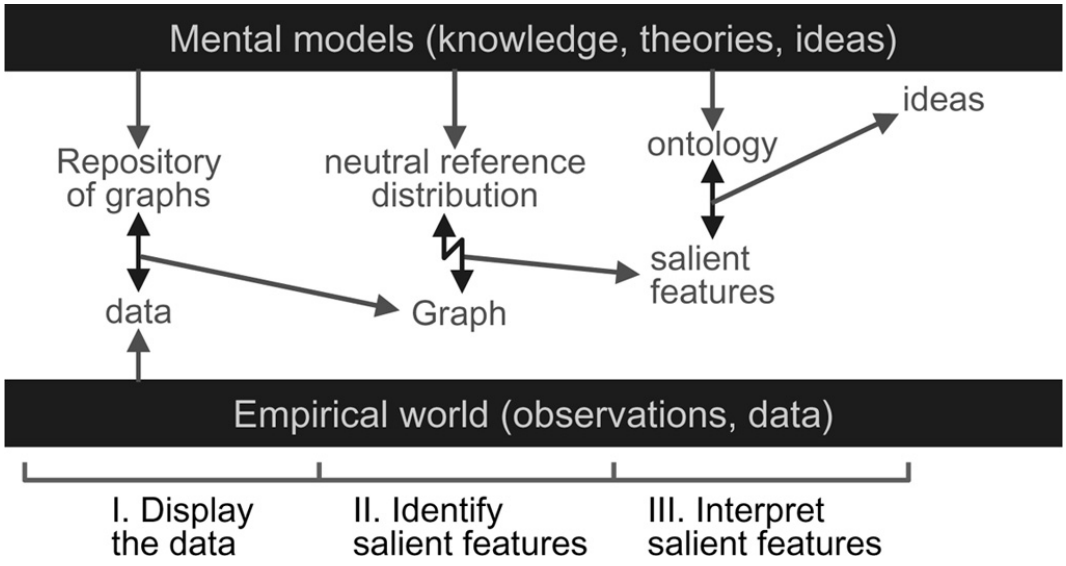


Figure 1: The Steps to Applying EDA in Problem-Solving and the Role of Mental Models in this Process (Source: De Mast and Kemper, 2009)

The previous paragraphs focused on traditional concept of EDA originated from Tukey (1977). Traditional EDA uses simple arithmetic and easy-to-draw pictures to present data (Tukey, 1977). However, this conventional definition encounters some challenges in the current "big data" era. First, with dramatically increasing data volume, these simple techniques cannot present enormous datasets effectively. For example, stem and leaf plot, one of the most commonly used traditional EDA techniques, is only useful for a dataset with less than 150 data points. With very large datasets, it will become cluttered and hard to understand. Data is ubiquitous nowadays, and users want to understand and extract useful knowledge from the data in a timely manner. Therefore,

more efficient EDA techniques are required to satisfy real-time requests. In order to cope with these challenges, many new data analysis methods were developed in the last decades. In practice, researchers are using these advanced techniques, such as data mining, to explore and visualize the data, which are EDA tasks. With these newly developed data analysis techniques, some elements of traditional EDA are no longer as necessary. Therefore, it is the time to redefine EDA in the contemporary environment.

With the emergence of new data analysis methods, the nature of EDA changes. EDA is converging with other methodologies, such as data mining. Data mining is a group of techniques to extract useful information and relationships from immense quantities of data automatically (Larose, 2005). Similar to EDA, data mining is completely data-driven because it starts without any predefined hypotheses, but aims to detect patterns that already exist in the data. Therefore, most data mining techniques can be used for EDA. In addition, data mining techniques can fulfill the missions of EDA, such as outlier detection, variable selection, and pattern recognition. Hence, data mining is considered as an extension of traditional EDA (Luan, 2002).

Because of the new features of modern data analysis techniques, the definition of EDA needs to be refined. Generally speaking, EDA should be changed from a means-oriented approach to a goal-oriented approach. The goals of modern EDA should include outlier detection, pattern recognition, and variables selection (Yu, 2010). Specifically, descriptiveness and visibility are two necessary conditions for traditional EDA. Techniques with poor visibility usually cannot be considered traditional EDA tools. However, some data mining techniques, such as neural network, which is very poor in terms of transparency and is usually considered to be a black box, should still be

considered one of the EDA tools, as long as its goal is to identify patterns in the data. Therefore, the scope of modern EDA has expanded from visualized exploratory analysis to general exploratory analysis. In addition, traditional EDA usually cannot provide conclusive answers, so it transfers the validation process to CDA. Some new data mining-based methods can validate findings and provide results comparable to CDA. So EDA is not necessarily an open-ended process; it may provide solid results as well. Table 2 summarizes and compares the characteristics of traditional EDA and modern EDA.

Table 2: Comparison Between Traditional EDA and Modern EDA

	Traditional EDA	Modern EDA
Type	Means-oriented	Goal-oriented
Scope	Visualized exploratory analysis	General exploratory analysis
Tools	• Simple arithmetic • Easy-to-draw pictures	• Descriptive statistics • Advanced data visualization techniques • Data mining techniques
Key Features	• No conclusive answers	• May provide conclusive answers

1.3 Exploratory Data Analysis Techniques

1.3.1 Traditional Exploratory Data Analysis Techniques

Since traditional EDA features simple arithmetic and easy-to-draw pictures, conventional EDA techniques mainly include descriptive statistics, basic data visualization techniques, and data transformation (Tukey, 1986a). Currently some of these traditional EDA techniques are introduced in some professional training course (for example, "Data Analysis for Internal Auditors" provided by IIA[2] and "Data Analytics for

[2] https://na.theiia.org/training/courses/Pages/Data-Analysis-for-Internal-Auditors.aspx

Auditors" provided by E&Y[3]) as recommended data analysis methods for auditors to obtain audit evidence.

Descriptive statistics provide quantitative descriptions of the observations to reveal their main features. They are most often used to examine: (1) Frequency distribution of data: how many data points fall into different ranges; (2) Central tendency: where data tends to fall; (3) dispersion (variability) of data: how spread out the data points are; (4) Skew (symmetry) of data: how concentrated data points are at the low or high end of the scale; and (5) Kurtosis (peakedness) of data: how concentrated data points are around a single value (Mann, 1995).

Frequency distribution summarizes and compresses data by grouping it into classes and recording how many data points fall into each class. For qualitative variables, each value is a class, whereas for quantitative variables, a class is usually an interval. Frequency distribution is usually measured by absolute frequency, relative frequency, cumulative absolute frequency, and/or cumulative relative frequency (Anderson et al., 2003). Absolute frequency shows the number of data points in each class. The sum of absolute frequency equals the total number of observations in the dataset. Relative frequency distribution displays the proportion of data points within each class. It is the ratio of absolute frequency to the total number of observations. The sum of relative frequencies always equals one. Cumulative absolute frequency is the total of an absolute frequency and all absolute frequencies below it. Similarly, cumulative relative frequency

[3]http://www.eytrainingcenter.com/index.php?option=com_content&task=view&id=689&Itemid=657&top_parent_id=746&parentId=657

is the total of a relative frequency and all relative frequencies below it. Table 3 demonstrates a simple data distribution using these four frequency distribution measures.

Table 3: An Example of Frequency Distribution

Score	Absolute Frequency	Relative Frequency	Cumulative Absolute Frequency	Cumulative Relative Frequency
1	2	0.14	2	0.14
2	5	0.36	7	0.50
3	4	0.29	11	0.79
4	2	0.14	13	0.93
5	1	0.7	14	100%
Total	14	1		

Central tendency indicates the middle and commonly occurring data points in a dataset. The common measures of central tendency are mean, median and mode (Dean and Illowsky, 2012). Mean is the sum of all values divided by the number of observations. Even though every data point is included in the computation of mean, mean may not always be the best measure of central tendency, especially when data is skewed. Median is a number such that half of the values in the dataset are below it and half of the values are above it. Median is not sensitive to the extreme values, so it can represent the exact middle of the data better than mean. Mode is the most frequent value in the dataset. It can indicate bimodality or multimodality in the data if more than one value occurs frequently in the dataset.

Dispersion measures indicate how spread out the data is around the mean. The most common measures of dispersion include range, variance, standard deviation and coefficient of variation (Anderson et al., 2003). Range is the difference between the lowest and highest values in the dataset. It generally describes how spread out the data is. Variance equals the sum of the squares of the difference between each data point and the

mean, divided by the total number of observations. It is often used when the variability of two or more datasets needs to be compared quickly. In general, the higher the variance, the more spread out the data is. Standard deviation is the positive square root of the variance. It refers to the average difference between the actual values and the mean. Standard deviation is used more commonly than variance in expressing the degree to which data is spread out. The coefficient of variation is simply the standard deviation divided by the mean. Since it includes the mean in its calculation, it illustrates the relative dispersion and describes the variance of two data sets better than the standard deviation does.

The asymmetry of data is measured by its skewness index (Dean and Illowsky, 2012). The skewness index can be calculated by the equation:

$$Sk = \frac{\sum F(X_i - \mu)^3}{\sigma^3}$$

Where Sk means the skewness of the data, F is the frequency of each class, σ is the standard deviation of the data, and $X_i - \mu$ stands for the difference between each item and the population mean. The ideal value of skewness index is zero, which means that the data is symmetrical. A positive skewness value indicates a distribution that is skewed to the right, whereas a negative skewness value indicates a distribution that is skewed to the left. Figure 2 shows the distributions with zero, positive and negative skewness values.

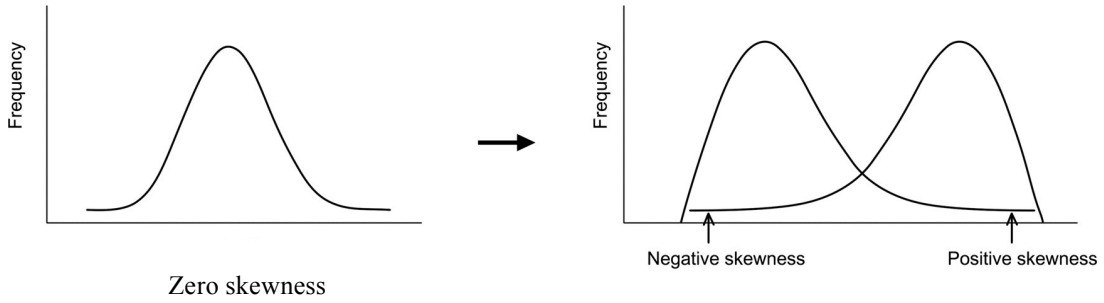


Figure 2: Distributions with Zero, Positive and Negative Skewness Values[4]

The degree of peakedness of a distribution is measured by the Kurtosis index (Balanda and MacGillivray, 1988). The equation to calculate Kurtosis is:

$$K = \frac{\sum F(X_i - \mu)^4}{\sigma^4}$$

The ideal value of the Kurtosis index is 3, which means that the data perfectively follows a normal distribution. The higher the value above 3, the more peaked is the distribution. The lower the value below 3, the more flat is the distribution. Figure 3 shows distributions with different Kurtosis values.

[4] Source: http://pic.dhe.ibm.com/infocenter/cx/v10r1m0/index.jsp?topic=%2Fcom.ibm.swg.ba.cognos.ug_cr_rptstd.10.1.0.doc%2Fc_id_obj_desc_tables.html

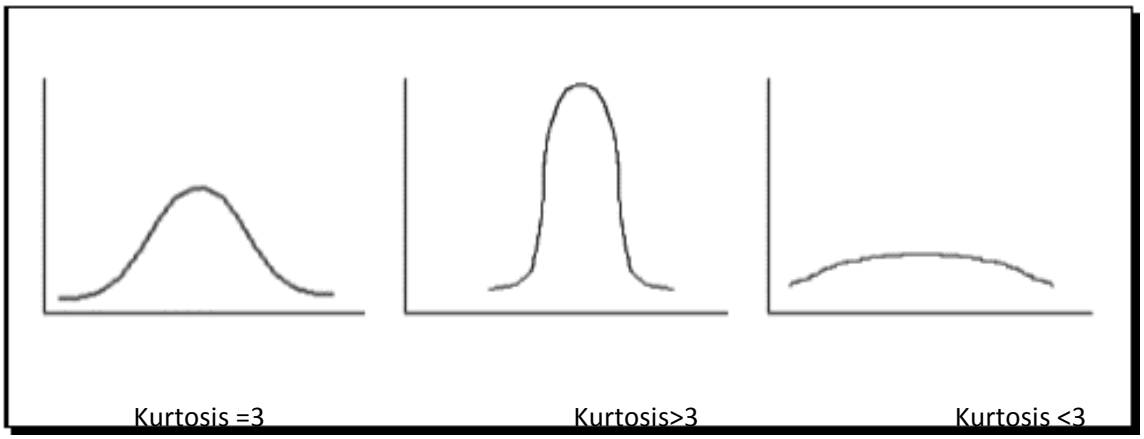


Figure 3: Distribution with Different Kurtosis Values[5]

Data visualization is the graphic presentation of data. With manipulation of graphic entities (e.g.: points, lines, shapes, images and text) and attributes (e.g.: color, size, position, and shape), patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized more easily (Cleveland, 1985). Basic data visualization tools include the pie chart, column chart, bar chart, linear chart, ogive, histogram, frequency polygon, scatter plot, box plot and stem-and-leaf plot.

A pie chart is a circle divided into sectors with areas proportional to relative size of each value (Anderson et al., 2003). It is usually used to present the absolute frequency or relative frequency distributions of qualitative variables. Figure 4 demonstrates the relative frequency distribution of employees in different work department in an organization[6].

[5] Source: http://allpsych.com/researchmethods/distributions.html
[6] Data used to create figures 3 to 7 is available at: https://www.dropbox.com/s/jmzk4gwf7ab51id/empmast.csv

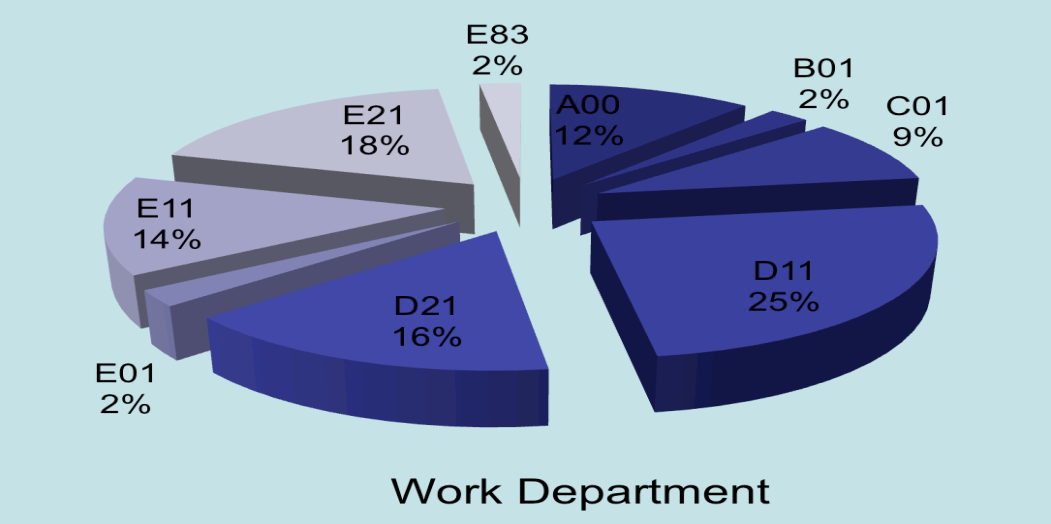


Figure 4: Pie Chart

A column chart (shown in Figure 5 (a)) displays vertical bars going across the chart horizontally, whereas a bar chart (shown in Figure 5 (b)) is similar to a column chart, but with horizontal bars (Anderson et al., 2003). Both column charts and bar charts are constructed to show the absolute frequencies or relative frequencies of qualitative variables. The height/length of each bar is proportional to the frequency. The two charts in Figure 5 display the absolute frequency distribution of the same data shown in Figure 4.

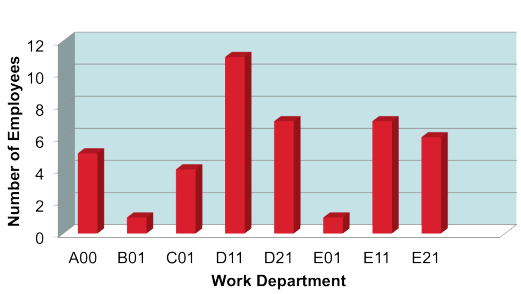


Figure 5 (a): Column Chart

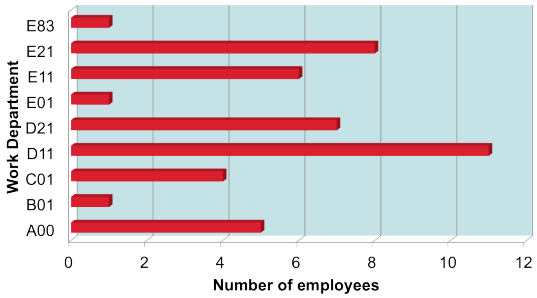


Figure 5 (b): Bar Chart

A linear chart displays a series of data points connected by straight lines (Andreas, 1965). It is suitable to present the frequency distribution of an ordinal variable.

An ogive is a kind of linear chart designed specifically to display a cumulative relative frequency distribution (Anderson et al., 2003). Figure 6 shows the relative frequency distribution and cumulative relative frequency distribution of employees' education levels in an organization using a linear chart and an ogive.

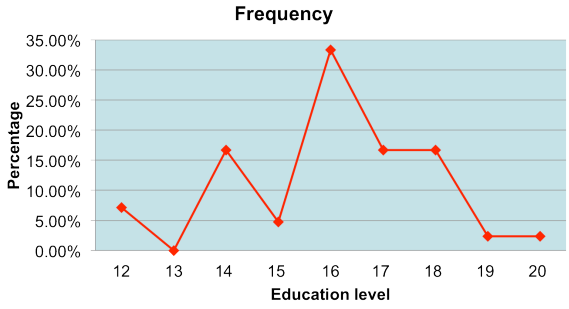


Figure 6(a): Linear Chart

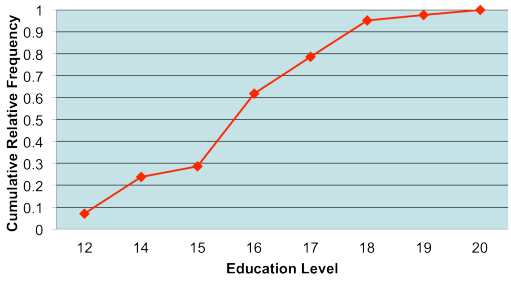


Figure 6 (b): Ogive

A histogram is a chart designed specifically to demonstrate the frequency distribution of quantitative variables (Anderson et al., 2003). It looks similar to a column chart, but in a histogram, the bars are not separated from each other. Instead of indicating a specific value for the variable, the bars in histogram denote intervals that should cover the full range of the variable. Connecting the middle point on the top of each bar with straight lines forms a frequency polygon, a linear chart showing frequency distribution of quantitative variables. Figure 7 shows a histogram and a frequency polygon displaying the frequency distribution of the pay per period for employees in an organization.

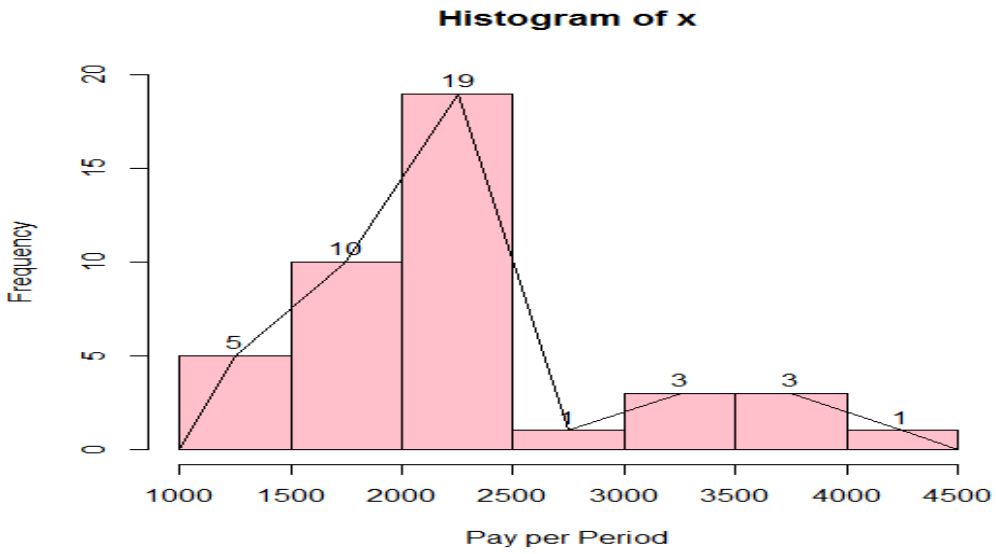


Figure 7: Histogram and Frequency Polygon

A scatter plot is a graph of plotted points, each representing a data point in the dataset. It is usually used to compare two quantitative variables (Jarrell, 1994). A trend line can be added to the scatter plot to describe the trend of the points and reveal the relationship between two variables. Figure 8 is a scatter plot with trend line presenting the relationship between pay per periods and salaries of employees in an organization.

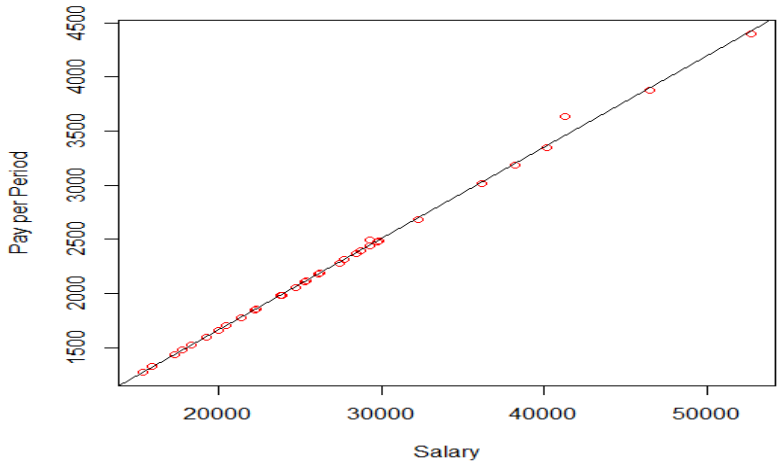


Figure 8: Scatter Plot

Instead of comparing two variables, quantile-quantile (Q-Q) plot (Wilk and Gnanadesikan, 1968) compares two data sets to check whether they can be fit into the same distribution. Quantile in the name indicates the fraction (or percent) of points below the given value. For example, the 0.4 (or 40%) quantile is the point at which 40% percent of the data fall below and 60% fall above that value. A Q-Q plot displays the quantiles of one data set against the quantiles of the other data set. Usually, a 45-degree reference line is also shown in a Q-Q plot. If the two sets come from populations with a common distribution, the points should fall approximately along with this reference line. The greater the variance from this reference line, the more possible that the two data sets come from populations with different distributions. Figure 9 demonstrates a Q-Q plot showing the distribution of two batches. According to this Q-Q plot, it is very likely that these two batches are not from populations with the same distribution.

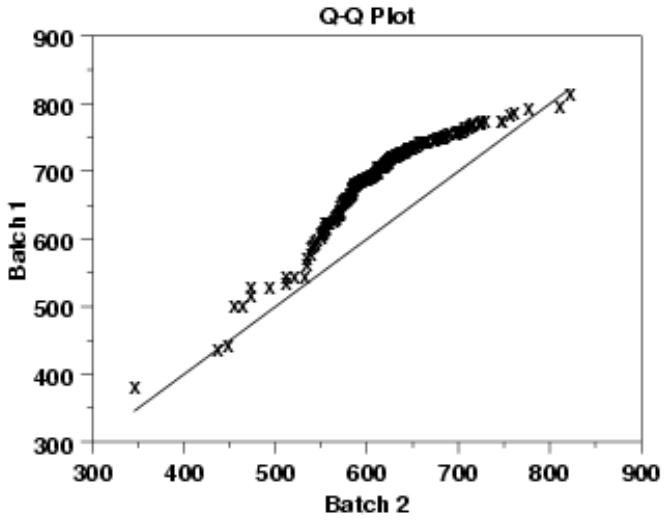


Figure 9: Q-Q plot[7]

[7] Source: http://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm

Another graphic representation that can be used to compare more than one distribution is Trellis Chart (Cleveland, 1993). A Trellis Chart displays data in smaller charts in a grid with consistent scales. It is usually used to compare the distribution of data points belonging to different categories, where the data demonstrated on each smaller chart belongs to one of these categories. The data displayed on different smaller charts are compared based on the same variables expressed in the form of X and Y axes in the charts. Trellis Charts are useful for finding patterns in complex data. Figure 10 is a Trellis Chart comparing the 2000 to 2004 sales information of cars and trucks in different regions.

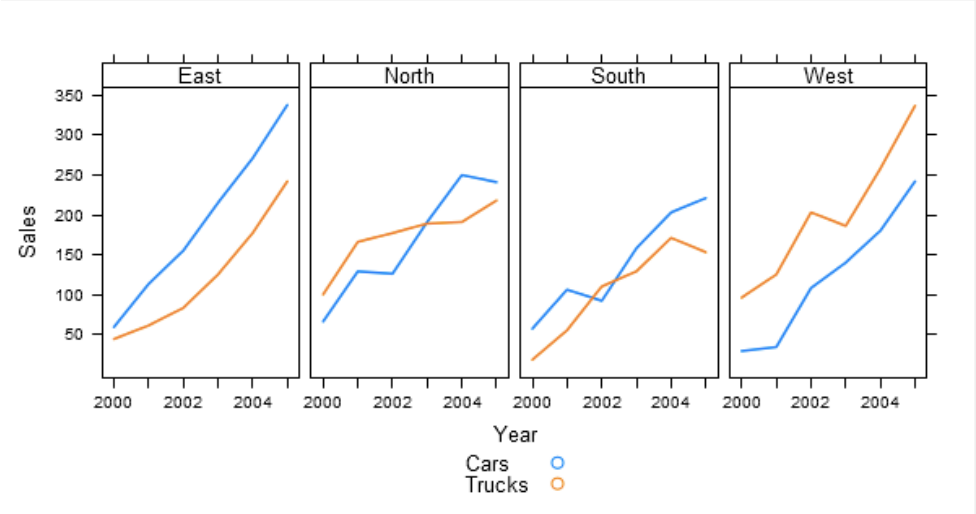


Figure 10: Trellis Chart[8]

A box plot is a graphic representation of quantitative variables based on their quartiles, as well as their smallest and largest values (Tukey, 1977). The simplest box plot contains a box and two whiskers. The bottom and top of the box are the first and third quartiles, and the line inside the box is the median. The box area represents the

[8] Source: http://trellischarts.com/what-is-a-trellis-chart

range between the first and third quartiles, which is called the interquartile range (IQR). The ends of the two whiskers indicate the maximum and minimum values of the variable. Box plots can have many variations. For example, complex box plots mark outliers (three or more times the IQR above the third quartile or below the first quartile) and suspected outliers (one and a half or more times the IQR above the third quartile or below the first quartile). In a complex box plot, if either type of outlier appears, the end of the whisker on the appropriate side changes to one and a half *IQR* from the corresponding quartile. The end of this whisker is defined as an inner fence, and the third IQR from this quartile is considered as the outer fence. Outliers in this plot are displayed as filled circles and suspected outliers are displayed as unfilled circles. Figure 11 shows an example of simple and complex box plots.

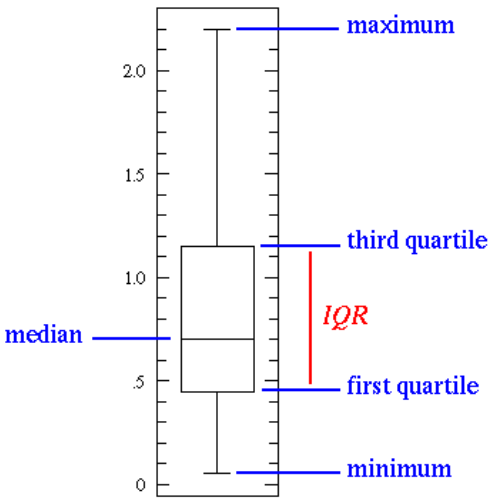


Figure 11(a): Simple Box Plot

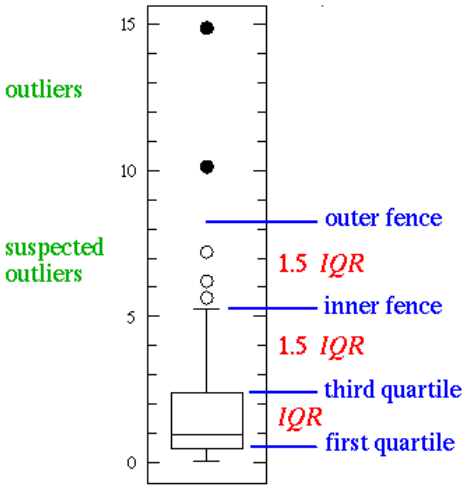


Figure 11 (b): Complex Box Plot[9]

A stem-and-leaf plot is a textual graph to present the distribution of quantitative variables (Tukey, 1977). A basic stem-and-leaf plot contains two columns separated by a

[9] Source: http://www.physics.csbsju.edu/stats/box2.html

vertical line. The left column is called the stem and the right column is called the leaf. Typically, the leaf contains the first digit of each data point and the stem contains all of the other digits. Compared with other tools, such as a histogram or box plot, the stem-and-leaf plot displays each data point. Therefore, it is not suitable for very large datasets. Thus, as the volume of data increases, the stem-and-leaf plot is used less. Figure 12 demonstrates a stem-and-leaf plot displaying the following list of values: 12, 13, 21, 27, 33, 34, 35, 37, 40, 40, 41.

Figure 12: Stem-and-Leaf Plot

To display the distribution of several datasets at the same time,

Data transformation is a technique that focuses on improving the interpretability or appearance of graphs (Tukey, 1977). For example, if displayed in original scale, the majority of the population in certain distributions may intensively locate in a small area in the graph (e.g.: Figure 13 (a)), which makes it very difficult for users to identify any patterns or trends in the data from this graph. In this case, if we can present the data in another scale where the majority of the records can distribute evenly in the display (e.g.: in Figure 13 (b)), it will be much easier for users to identify the relationships from the revised graph. The technique of choosing a new scale and redisplaying the data is called

data transformation. Specifically, when doing data transformation, a deterministic mathematical function is applied to each point in a dataset. Each original point in the dataset is replaced by a transformed value and shown in the revised graph. Various mathematical functions can be used in data transformation. Users can select the most appropriate one based on the characteristics of the original data. The most commonly used transformation functions in practice are the logarithm function (y=log(x)), which is employed in Figure 13 (b), the square root function ($y=\sqrt{x}$), and the reciprocal function (y=1/x).

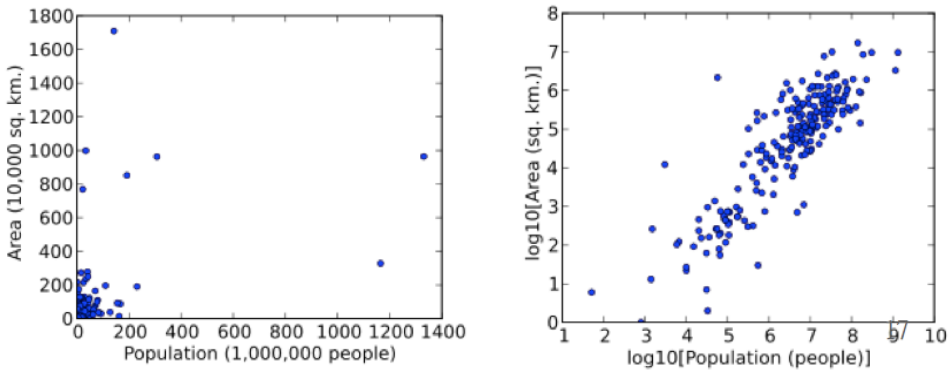


Figure 13 (a): Graph in Original Scale *Figure 13 (b): Graph After Data Transformation*[10]

1.3.2 Advanced Exploratory Data Analysis Techniques

Instead of using basic statistics and easy-to-draw graphs to show the general patterns in data, advanced EDA techniques employ more complicated models to unearth deeper relationships hidden in the data. Numerous advanced EDA techniques exist. The ones that are widely used in business comprise advanced data visualization techniques, feature selection techniques, data mining techniques, such as clustering and association analysis, text mining techniques, social network analysis, and process mining.

[10]Source: http://en.wikipedia.org/wiki/Data_transformation_(statistics)#mediaviewer/File:Population_vs_area.svg

Accounting researchers apply some of these techniques to auditing in their studies. A detailed discussion of the application of these techniques is in Chapter 2, section 2.2.2 Related Research in Auditing Discipline. However, most of these techniques have not yet been widely implemented in audit practice.

Advanced data visualization techniques are extensions of basic data visualization techniques, but they display data in more sophisticated ways so that more variables can be shown in one graph (Heer et al., 2010). Three commonly used advanced data visualization techniques are the heat map, geographic map and dashboard.

A heat map is a two-dimensional representation of data in which values are represented by colors (Wilkinson and Friendly, 2008). Different colors as well as the shades of those colors can be used to represent data values. For example, Figure 14 shows a heat map demonstrating the aggregated average response time of a website in different time slots during a six-week period. In this graph, each cell in the table represents the website's response speed in a certain time slot. The darkest green color indicates the fastest response speed. As the shade of green turns lighter, the displayed response speed becomes slower. The medium response speed is represented by yellow. Red denotes a slow response speed. The redder the color, the slower the response speed that is represented. In a heat map, the value of each data point can be displayed only by color (Figure 14 (a)), or by both color and value (Figure 14 (b)) to show more detailed information.

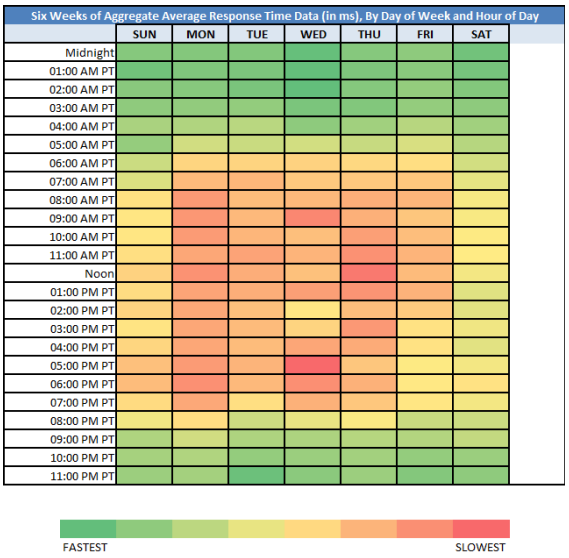


Figure 14(a): Heat Map[11]

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
0	6	8	8	8	9	12	10
1	5	5	6	7	8	10	10
2	3	4	4	5	7	7	5
3	2	3	2	3	2	4	3
4	0	1	0	1	1	1	0
5	0	0	0	0	0	0	0
6	1	1	2	1	0	0	0
7	10	10	10	9	7	6	4
8	19	18	18	17	12	14	9
9	20	18	17	15	14	18	15
10	16	14	16	16	12	17	15
11	15	14	15	15	13	14	16
12	15	14	13	12	11	11	15
13	13	12	13	13	11	11	14
14	15	12	14	14	10	12	13
15	13	14	13	13	10	11	12
16	11	10	10	10	9	10	11
17	6	5	5	5	4	6	6
18	3	2	1	2	3	3	2
19	4	3	4	5	5	6	4
20	7	7	7	8	9	10	8
21	10	9	9	10	11	11	9
22	9	10	9	12	11	13	9
23	8	9	11	10	12	12	8

Figure 14 (b) Heat Map with Values[12]

A widely used variation of the heat map is a tree map (Shneiderman, 1991), which uses rectangles to represent records. Unlike the basic heat map, where the size of each cell is the same, the rectangles in a tree map can have different size. The size and the color of the rectangles can correspond to two different values, allowing the user to perceive two variables at once. Figure 15 shows a tree map representing the total usage of renewable energy in several countries in 2010. Each rectangle in this graph indicates one country. The size of each rectangle demonstrates the total amount of renewable energy used by that country, and the color denotes the annual percentage change in this amount.

[11] Source: http://webtortoise.com/tag/heatmap/
[12] Source: http://policeanalyst.com/creating-heat-maps-in-saps-businessobjects-webis/

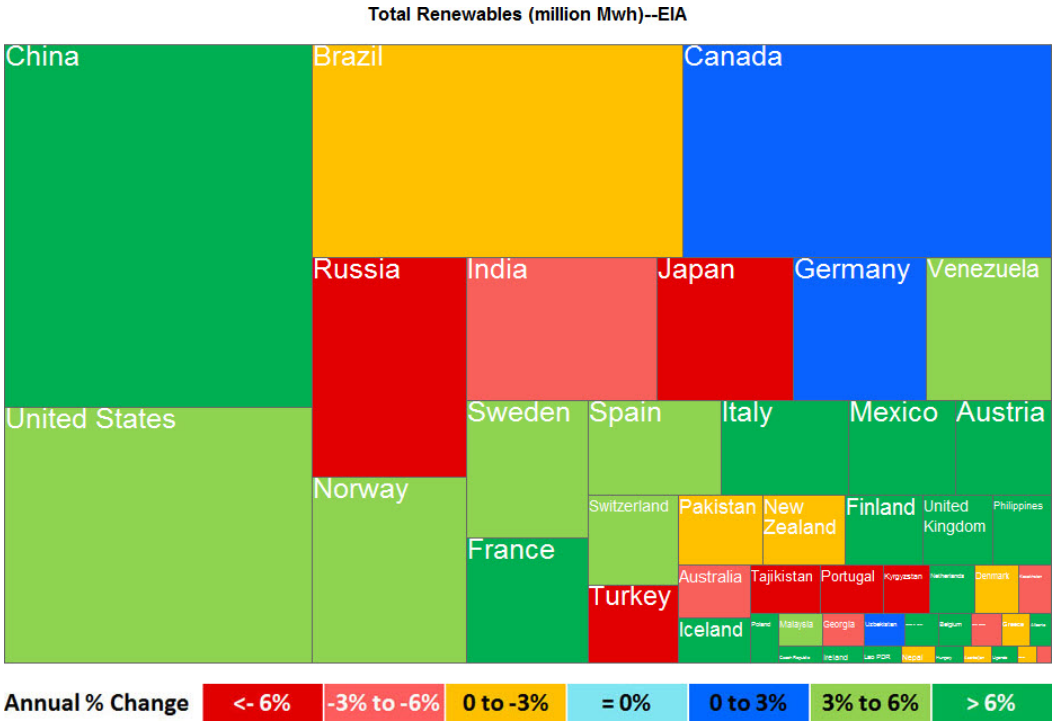


Figure 15: Tree Map[13]

A geographic map plots the geographic location information in a dataset on a geographic map (MacEachren and Kraak, 1997). Like a pie chart or column chart, it usually uses basic attributes like color and size to demonstrate desired information in each location. With a geographic map, users can easily compare the displayed features in different locations. For example, Figure 16 is a geographic map showing the distribution of the proportion of residents with no health insurance in the U.S. (Pickle and Su, 2002).

[13] Source: http://co2scorecard.org/home/researchitem/10

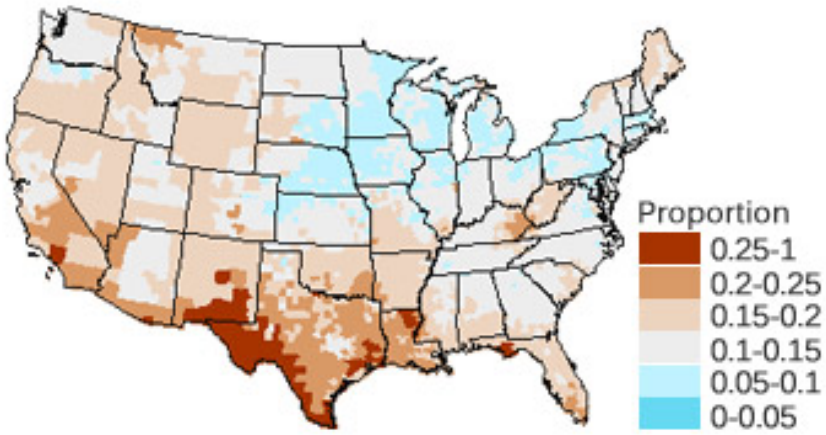


Figure 16: Geographic Map[14]

Dashboard data visualization is a comprehensive display of data. It is widely used in business to depict the performance of an enterprise, a specific department, or a key business operation (Alexander and Walkenbach, 2010). A dashboard (e.g.:Figure 17) usually provides an at-a-glance view of an organization's key indicators, such as key performance indicator (KPI) and key risk indicator (KRI), using gauge charts, heat maps, geographic maps, and other basic data visualization techniques to inform users of the current status of the organization's performance. Like a vehicular dashboard, the most commonly used component in a dashboard is the gauge chart. A gauge chart consists of a scale with tick marks and numbers, and a needle pointing to a value on that scale. Usually, a gauge chart also includes markings on the scale to indicate critical values or ranges. Unlike most of the complex data visualization techniques discussed above, which attempt to use various attributes to demonstrate many variables in the data, a gauge chart can only reflect one variable. Because of this simplicity, it is usually used to highlight the most critical indicators in an organization.

[14] Source: http://gis.cancer.gov/overview/geovisualization_tools.html

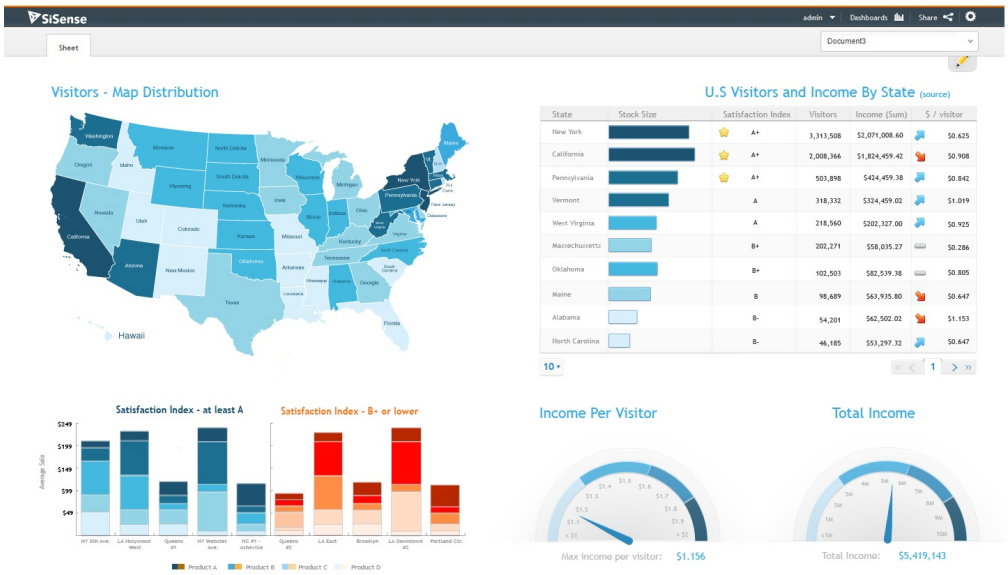


Figure 17: Dashboard Data Visualization[15]

Feature selection is another advanced EDA technique that focuses on selecting a subset of relevant variables for use in constructing a predictive model (Guyon and Elisseeff, 2003). This technique is very useful in today's "big data" environment where a dataset usually has many attributes, but only few of them are related to the analysis target. A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure that scores the different feature subsets. The basic idea of feature selection is to test each possible subset of features and select the one that can minimize the error rate.

Three main categories of feature selection algorithms are wrappers, filters, and embedded methods (Guyon and Elisseeff, 2003). Wrapper methods use a predictive model to score feature subsets based on the number of mistakes made by each subset. Filter methods select subsets of variables as a preprocessing step that is independent of

[15] Source: http://plotting-success.softwareadvice.com/case-study-alpharooms-com-improves-booking-performance-0813/

the chosen predictor. Embedded methods perform variable selection as part of the model construction process. The most popular feature selection algorithm in statistics is stepwise regression, a type of wrapper model. Stepwise regression can select variables in three approaches: (1) Forward selection: Start with no variables in the model, then add the variable that improves the model the most, and repeat this process until no more improvement can be done to the model; (2) Backward elimination: Start with all candidate variables, then delete the variable that improves the model the most after its removal, and repeat this process until no further improvement is possible; (3) Bidirectional elimination: A combination of forward selection and backward elimination with testing at each step for variables to be included or excluded (Draper and Smith, 1981; Stringer and Stewart, 1996).

Since one of the goals of data mining is to derive patterns that summarize the underlying relationship in data, which is the same as the goal of EDA, data mining techniques (e.g.: clustering and association analysis) and some variations of conventional data mining (e.g.: text mining and social network analysis) can be used as advanced techniques to conduct EDA tasks (Yu, 2010).

Cluster analysis is a data mining task that focuses on dividing data points into different groups so that the data points within a group are similar to each other and different from those in other groups (Tan et al., 2006). The groups in cluster analysis are only derived from the data. An important step in cluster analysis is the definition of similarity among the data points. Various similarity measurements can be developed to accommodate different features in the data. For example, similarities can be measured by distances, densities, connections, etc. According to these similarity measures, diverse

methods can be used in cluster analysis to separate the data. The distance-based method, density-based method, hierarchical method, graph-based method, and probabilistic method all generate many clustering algorithms (Estivill-Castro, 2002). In order to get the best results, users should choose the most suitable method for the dataset to conduct the cluster process. However, there is no existing rule to define which method should be used on certain types of data. The choice of the most suitable method depends significantly on the features of the dataset. Therefore, cluster analysis is usually considered more of an art than a science.

There are three common ways to evaluate clustering results: external indices, internal indices, and relative evaluation. External indices compare the results of a cluster analysis to externally known results, such as externally provided class labels. By contrast, internal indices evaluate the quality of a cluster analysis without referring to external information. Internal indices usually use some statistical measurements, such as the sum of squared errors (SSE), to measure how close the objects within a cluster are and how distinct they are from the objects in different clusters. We can also determine whether a cluster method fits the data by comparing the results of different cluster methods, which is called relative evaluation.

Unlike clustering, which emphasizes grouping data points, association analysis (Agrawal et al., 1993) focuses on discovering hidden, interesting relationships in datasets. Specifically, it seeks frequent patterns, associations, correlations, or rules according to the occurrences of one data point based on the occurrence of other data points in the dataset. The uncovered relationships are represented in the form of association rules like A ➔ B, where A and B can each be an attribute or a set of attributes in the dataset. This

rule suggests a strong relationship between A and B, indicating that they may appear together in the dataset.

The strength of association rules is often measured by support and confidence, which are defined by the following equations:

$$support = \frac{\sigma(A \cup B)}{N}; \qquad confidence = \frac{\sigma(A \cup B)}{\sigma(A)}$$

where σ(A∪B) is the support count of the rule (the number of times that A and B appear together in the dataset), N is the total number of records in the dataset, and σ(A) is the number of times that A appear in the dataset. Support measures how often a rule applies to a data set. An association rule with low support is not representative. Therefore, in an association analysis process, we usually set a support threshold to eliminate unrepresentative rules. Confidence describes how frequently B appears in records that contain A, and it measures the reliability of an association rule (Lai and Cerpa, 2001).

During the analysis of large, real datasets, if we consider each attribute in the rule generation process, a large number of unrepresentative rules will be generated. To make the association analysis process more efficient, we usually set a support count threshold to screen attributes that occur frequently first, and then generate rules based on these frequent attributes. Different association analysis algorithms use different methods to determine the most frequent attributes.

Most of the data mining techniques discussed above are designed to analyze structured data, such as data stored in relational databases and spreadsheets. However, in the current "big data" era, a large volume of unstructured textual data is emerging, which is not amenable to traditional data mining techniques. Text mining has developed to

obtain knowledge from such data. Generally speaking, text mining can exploit information contained in textual documents in various ways, including discovering patterns and trends in the data and identifying association among entities described in the text data (Grobelnik et al., 2001). Therefore, text mining can be seen as an EDA technique that leads to previously unknown information, or to answers for questions that were previously unanswerable (Hearst, 1999).

In most text mining processes, unstructured text data is first represented in a structured way and then various data mining techniques can be applied to the transformed data to accomplish specific objectives. A widely used method to represent text data is called a vector space model (VSM) (Salton and Yang, 1975). Based on VSM, each document is represented by a vector of terms (words) after removing stop words and applying word stemming (changing words back to their root form). The frequency that a word appears in a specific text is used to indicate the weight of the corresponding feature. Typically, a frequency measure can be binary to indicate absence or presence of a word, or it can be a number given by a mathematical function. Traditional statistical text mining methods treat text as a "bag-of-words" (Salton and McGill, 1983), where single words or word stems are used as features in the text's vector presentations. Using this method, text mining algorithms are restricted to detecting patterns only in the words used, although the semantics in the texts might be misrepresented (Bloehdorn and Hotho, 2004).

How to deal with the semantic meaning in text information is a big challenge for researchers. To cope with the complex and subtle relationships among concepts, word ambiguity, and context sensitivity, a series of semantic text mining methods (Lewis, 1992; Kozima, 1993; Antonellis and Gallopoulos, 2006; Hotho, 2002) have been

developed. Some of these techniques use phrases instead of individual words as features in the texts' vector presentations; some of them considered the locations of the words in the file, such as in the beginning, middle, or end of a paragraph or article; and others integrate linguistic features, such as synonym, antonym, and hierarchical relationships between words and phrases, into the analysis by combining some background knowledge.

Besides text mining, another variation of data mining is process mining, a technique that focuses on processes rather data. Process mining enables users to discover, control, and improve processes by analyzing event logs (Van der Aalst, 2011). An event log is a record of events. It usually includes four components: (1) the activity taking place during the event, (2) the process instance of the event, (3) the originator or party responsible for the event, and (4) the timestamp of the event (Jans et al., 2014). An example of the event log of an invoice creation event originated by John on August 5, 2013 is shown in Figure 18.

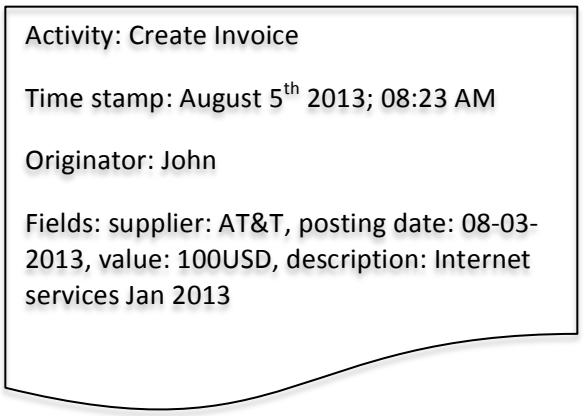


Figure 18: An Example of an Event Log (Data from Jans et al., 2014)

By applying process mining techniques to extract and analyze information stored in an event log, five types of analyses can be performed: (1) process discovery, (2) conformance check, (3) performance analysis, (4) decision mining and verification, and

(5) social networks analysis (Jans et al., 2013). Process discovery focuses on constructing a model based on low-level events to represent the actual process when an *a priori* model does not exist. Commonly used technologies for process discovery include Petri nets and probabilistic algorithms. A conformance check can be conducted when there is an *a priori* model. This model is then compared with the actual event logs to identify discrepancies. Performance analysis utilizes various measurements to represent the KPI of a process (e.g.: the minimum, maximum, and average throughput times). Decision mining and verification focus on decision points in a discovered process model. Identified decisions can be compared with the standard practice rules or policies to detect irregular behaviors. The information related to the parties involved in the event contained in the event log can be used to perform social network analysis. This analysis can ascertain relationships between individuals in the workplace to help users to gain an understanding of the actual role of each person in the process and to identify unexpected or anomalous relationships. Some of the potential applications of process mining are exploratory in nature, such as process discovery, performance analysis, and social network analysis, whereas others, like conformance checks and decision mining and verification, are typical CDA techniques aiming at evaluating the performance of prior models in practice.

Besides being used together with process mining, social network analysis can be used as a stand-alone EDA technique to explore the connections between individuals. The basic idea of social network analysis is to view interpersonal relationships in terms of a network consisting of points that represent individuals and lines that indicate relationships between individuals (Wasserman et al., 1994). Relationships in a social

network can be binary or valued with numbers indicating the weight of relationships, and can be undirected or directed with an arrow indicating the information flow direction of a relationship. For example, Figure 19 demonstrates a binary directed social network in an organization.

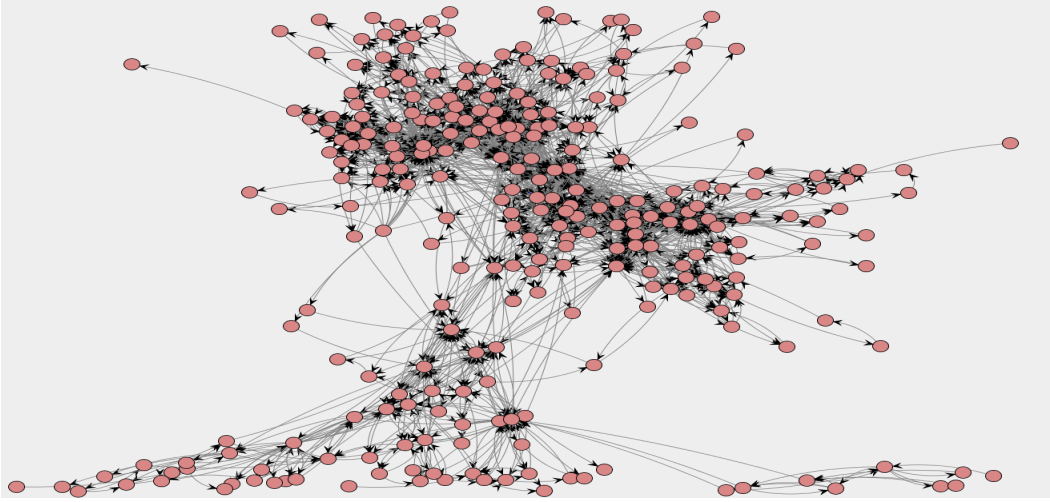


Figure 19: A Social Network in an organization (Source: Jans et al., 2014)

There are basically three types of measurements used in social network analysis: local centrality or degree, betweenness, and global centrality or closeness (Moody, 2004). Local centrality or degree calculates the number of connections an individual has with others. It is a potential sign of power. Specifically, high in-degree (many arrow pointing in) can be a sign of prominence or prestige and high out-degree (many arrows pointing out) can be a sign of influence. Betweenness denotes the extent to which an individual is situated between two groups and is a necessary route between those groups. Individuals with high betweenness have the potential to have major influence. Therefore, they can be mediators/brokers, gatekeepers, bottlenecks, or obstacles to communication. Global centrality or closeness measures the average distance between an individual and all other

individuals in a network. Individual with high global centrality are likely to know what is happening throughout the whole social network.

From these basic measurements, various mathematical and statistical tools can be used to analyze a social network to fulfill specific objectives. Therefore, social network analysis has been widely used in anthropology, biology, communication studies, economics, geography, history, information science, organizational studies, political science, social psychology, development studies and sociolinguistics (Wasserman et al., 1994). For example, it can be used in career planning to investigate how people find jobs, in organizational design to explore how an office should be laid out, and in knowledge management to study how innovations spread and who the resident subject matter experts are. Also, as mentioned above, it can be used together with process mining to reengineer business processes by identifying where the organizational disconnects and bottlenecks are.

1.4 Methodology and Research Questions

1.4.1 Design Science Approach

Traditionally, research in the area of accounting information systems is considered to be natural science research, which focuses on understanding phenomena and finding new truths (Geerts, 2011). This research follows the paradigm of design science, first defined by Simon in 1969 (Simon, 1996). Design science research attempts to create artifacts that describe how things ought to be in order to change existing situations into preferred ones to improve practice. The purpose of this dissertation is not to figure out why things work the way they do. Instead, it attempts to embed an analytical

approach into the audit process to improve audit practice. Therefore, this dissertation can be considered as design science research.

March and Smith (1995) classify artifacts into four different types: concepts, models, methods, and instantiations. Concepts indicate novel constructs within the domain, which can be used to improve the current solutions. EDA is an example of a concept that is developed to complement the traditional CDA approach. A model is a description representing the relationships among constructs. EDA techniques can be seen as models that depict the associations within data. A method is a set of steps or guidelines used to perform a task. The framework developed in this dissertation is a method for auditors to integrate EDA into the audit process. An instantiation is the realization of an artifact in its environment. Two case studies demonstrated in this dissertation follow the guidelines and steps proposed in the framework. They can be seen as two instantiations of the proposed framework.

A normative framework to conduct design science research in the information systems discipline, named design science research methodology (DSRM) was proposed by Peffers et al. in 2008. This methodology consists of six sequential activities: (1) Problem identification and motivation: Defining the specific research problem and justifying the value of a solution; (2) Definition of the objectives for a solution: Inferring the objectives of a solution from the problem definition and knowledge of what is possible and feasible; (3) Design and development: Creating the artifact; (4) Demonstration: Showing the use of the artifact to solve one or more instances of the problem; (5) Evaluation: Observing and measuring how well the artifact supports a solution to the problem; and (6) Communication: Expressing the problem and its

importance, the artifact, its utility and novelty, the rigor of its design, and its effectiveness to researchers and other relevant audiences (Peffers et al., 2008).

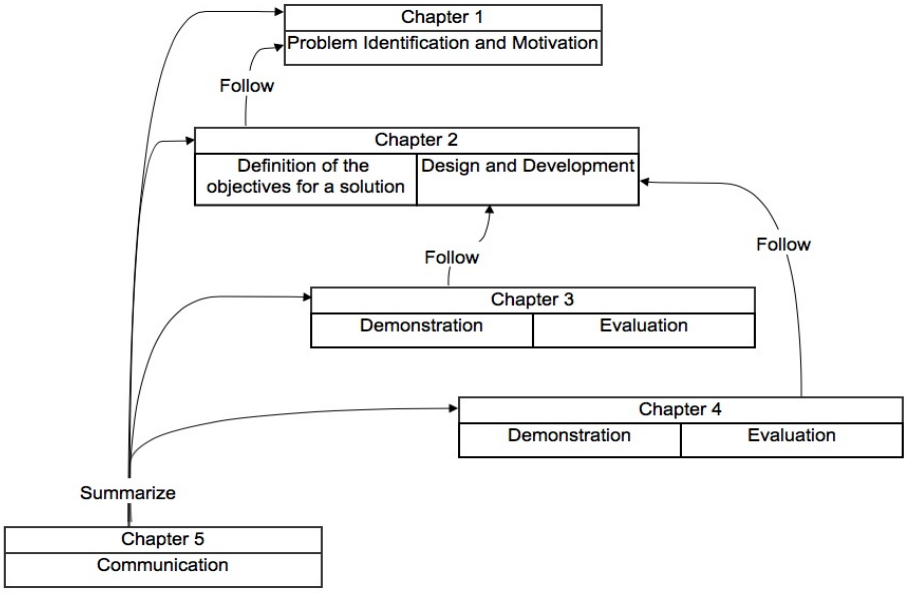


Figure 20: The Research Design in this Dissertation

Following this methodology, these activities are covered in different chapters in this dissertation. Figure 20 demonstrates the research design in this dissertation. Specifically, the first activity, problem identification and motivation, is covered in 0 by elaborating the desired and current solutions in audit practice, and specifying the research questions. In the second chapter, two activities, the definition of the objectives for a solution and the design and development, are discussed. In particular, the objectives of a new solution are explored by reviewing the literature and evaluating the applications of EDA-enabling technologies. Then the conceptual design and development of the new solution is described to address the research questions raised in the first chapter. 0 and 0 provide real case demonstrations to support the solution proposed in **Error! Reference source not found.**, and evaluate the results of each case to show the benefits of this new

solution. Finally, communication history, including related presentations and publications of this solution, is summarized in 0.

1.4.2 Motivation and Research Questions

Risk assessment is an essential step for both internal and external auditors to provide high-level assurance. In the current complex and rapid-changing business environment, organizations are more likely to face unfamiliar and unexpected risks. These emerging risks may be entirely new or they may evolve or escalate from existing risks, and they may cause serious consequences to the organization (Thomson Reuters, 2013). The failure of the banking system in 2008 demonstrates the significant impact caused by emerging risks. A definition from Reinsurance Company Swiss Re describes emerging risk as "newly developing or changing risks which are difficult to quantify and which may have a major impact on the organization". Therefore, to provide sufficient assurance, emerging risks should not be ignored in any audit process.

In fact, both internal and external audit regulators have recognized the significance of identifying emerging risks and have included it as one of the auditor's obligations in their newly issued standards or strongly recommended guidance. For example, as described in a recommended guidance for financial services issued by the IIA[16], internal auditors should identify and assess all of the key risks in the organization including ***emerging*** and systemic risk.

Similar considerations are reflected in regulations and guidance for external

[16] http://www.iia.org.uk/media/354788/0758_effective_internal_audit_financial_webfinal.pdf

auditors. For example, in the PCAOB's guidance alert No. 9[17], "Assessing and responding to risk in the current economic environment", external auditors are required to consider ***new*** risks that may exist due to current events when developing an audit plan. Also, if new risks are identified during the audit process, auditors need to reassess audit risks, update their understanding of how emerging risks may affect the company's financial reporting, and modify planned audit procedures.

Even though the importance of considering emerging risks in audit has been recognized, the way to identify emerging risks is not explicitly described in these standards and guidance. Clarified statements on audit standards newly issued by the AICPA mention that external auditors can use analytical procedure to identify new risks. For example, in AU-C sec. 315[18], "Understanding the entity and its environment and assessing the risks of material misstatement", external auditors are required to perform analytical procedures as risk assessment procedures to identify aspects of the entity of which the auditor was unaware. Also, AU-C sec. 520[19], "Analytical procedures" request that analytical procedures should also be designed and performed near the end of the audit to identify previously unrecognized risk of material misstatement.

However, traditional analytical procedures used in auditing are mainly based on analysis of history, so they have limited power to identify emerging risks because these newly developed risks lack precedent or history. EDA, as discussed in the previous sections, is a data analysis approach based on current data, which does not require any

[17] http://pcaobus.org/Standards/QandA/12-06-2011_SAPA_9.pdf
[18] http://www.aicpa.org/Research/Standards/AuditAttest/DownloadableDocuments/AU-C-00315.pdf
[19] http://www.aicpa.org/Research/Standards/AuditAttest/DownloadableDocuments/AU-C-00520.pdf

prior assumptions or benchmarks. Hence, it has the potential to ameliorate the drawbacks of the traditional audit analysis approach and facilitate auditors' discovery of emerging risks. However, arbitrarily applying EDA in certain audit procedures cannot unearth all of the hidden risks in the organization. In addition, human cognition has a number of biases that hinder a rational approach to unusual events. For example, we tend not to give credibility to that which we have not experienced (Thomson Reuters, 2013). Therefore, auditors need a standardized approach to assist them in systematically identifying, confirming, and assessing emerging and unknown risks.

This dissertation intends to explore the use of EDA in both the internal and external audit processes, develop recommended guidelines for auditors to implement EDA in both traditional and continuous audit environments, and demonstrate and evaluate the proposed guidelines in real cases. Specifically, this dissertation focuses on addressing the following research questions:

1. What value can EDA add to the audit process?
2. When should auditors use EDA in the audit process?
3. Specifically, how should auditors perform EDA in auditing?
4. What EDA techniques can be applied in different audit stages?
5. How can EDA be implemented in a continuous auditing system?
6. How can the application of EDA be demonstrated and evaluated in audit practice?

References

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data P.207

Aiken, L.S., West, S.G., Sechrest, L., and Reno, R.R. (1990). Graduate training in statistics, methodology, and measurement in psychology: A review of Ph.D. programs in North America. American Psychologist, 45, 721-734.

Alexander, M. and Walkenbach, J. (2010) Excel Dashboards and Reports. Wiley. ISBN: 978-1118490426

Anderson, D., Sweeney, D., and Williams, H. (2003). Essentials of Statistics for Business and Economics, Revised. Cengage Learning.

Antonellis, I. and Gallopoulos, E. (2006). Exploring term-document matrices from matrix models in text mining. In Proc. of the SIAM Text Mining Workshop 2006, 6th SIAM SDM Conference, Maryland.

Balanda, K. P. and MacGillivray, H.L. (1988). Kurtosis: A Critical Review. *The American Statistician*, 42(2): 111–119.

Bloehdorn, S. and Hotho, A. (2004) Boosting for text classification with semantic features. Proceedings of the Workshop on Mining for and from the Semantic Web at the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining: 70-87

Burton G. A. (1965). *Experimental psychology*. John Wiley & Sons Inc. 2nd edition.

Cleveland, W.S. (1985). The Elements of Graphing Data. Pacific Grove, CA: Wadsworth & Advanced Book Program.

Cleveland, W.S. (1993). Visualizing Data. Hobart Press.

CO_2 Scorecard Research. (2010). Site Feature: Using a Treemap to Analyze Relative Values of CO_2 Indicators. Available at: http://co2scorecard.org/home/researchitem/10

De Mast, J. and Trip, A. (2007). Exploratory data analysis in quality improvement projects. Journal of Quality Technology, 39: 301–311.

De Mast, J., and Kemper, B. P. H. (2009). Principles of Exploratory Data Analysis in Problem Solving: What Can We Learn from a Well-Known Case? Quality

Engineering, 21: 366-375.

De Mast, J. and Kemper, B. P. H. (2009 (b)). Discussion of "Principles of Exploratory Data Analysis in Problem Solving: What Can We Learn From a Well-Known Case? " – Rejoinder. Quality Engineering, 21: 382-383.

Dean, S. and Illowsky, B., (2012). Descriptive Statistics: Skewness and the Mean, Median, and Mode. Available at: http://cnx.org/content/m17104/latest/

Draper, N. and Smith, H. (1981). Applied Regression Analysis, 2d Edition, New York: John Wiley & Sons, Inc.

Estivill-Castro, V. (2002). Why so many clustering algorithms – A position paper. ACM SIGKDD Explorations Newsletter 4(1): 65-75.

Geerts, G. L. (2011). A design science research methodology and its application to accounting information systems research. International Journal of Accounting Information Systems 12(2011): 142-151.

Guyon, I., and Elisseff, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research. 3: 1157-1182.

Heer, J., Bostock, M., and Ogievetsky, V. (2010). A tour through the visualization Zoo. ACMqueue. Available at http://queue.acm.org/detail.cfm?id=1805128

Horowitz, I. L. (1980). Exploratory Data Analysis. By John W. Tukey. Administrative Science Quarterly. 25(4): 700-703.

Hotho, A., Maedche A., and Staab S. (2002). Text Clustering Based on Good Aggregations. Künstliche Intelligenz (KI), 16(4): 48-54.

Jans, M., Alles M., and Vasarhelyi M. (2013). The case for process mining in auditing: Sources of value added and areas of application. International Journal of Accounting Information Systems 14: 1-20.

Jans, M., Alles M., and 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.

Jarrell, S. B. (1994). Basic Statistics. William C Brown Communications.

Kohonen, T. (1982). Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics 43(1): 59-69.

Kozima, H. (1993). Text segmentation based on similarity between words. In Proc. of the 31st ACL, Columbus, Ohio, USA, pp. 286-289.

Grobelnik, M., Mladenic, D., and Milic-Frayling, N. (2000) Text Mining as Integration of Several Related Research Areas. Report on KDD'2000 Workshop on Text Mining

Hearst, M., (1999). Untangling Text Data Mining. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.

Lai, K., and Cerpa, N., (2001) Support vs Confidence in Association Rule Algorithms. Cerpa Proceedings of the OPTIMA Conference, October 10-12, 2001, Cuico, Chile.

Larose, D. (2005). Discovering knowledge in data: An introduction to data mining. NJ: Wiley-Interscience.

Lewis, D. D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In Proc. of SIGIR, Copenhagen, Denmark, pp. 37-50.

Luan, J. (2002). Data mining and its applications in higher education. In A. Serban & J. Luan (Eds.), Knowledge management: Building a competitive advantage in higher education. PA: Josey-Bass, pp. 17-36.

MacEachren, A.M. and Kraak, M. J. (1997). Exploratory Cartographic Visualization: advancing the agenda. Computers & Geosciences, 23(4): 335-343.

Mann, P.S. (1995). Introductory Statistics (2nd ed.). Wiley.

March ST, Smith G. (1995). Design and natural science research on information technology. Decision Support Systems.15(4): 251-266.

McCandless, D. (2009). The Visual Miscellaneum: A Colorful Guide to the World's Most Consequential Trivia. Harper Design.

Moody, J. (2004). Social Network Analysis. American Sociological Association, San Francisco, August 2004. Available at http://www.soc.duke.edu/~jmoody77/presentations/asa_snaintro.ppt

Nigrini, M. J. and Mittermaier, L. J. (1997). The use of Benford's law as an aid in analytical procedures. Auditing: A Journal of Practice & Theory. 16(2): 52-67.

Peffers K, Tuunanen T, Rothenberger MA, and Chatterjee S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems. 24 (3): 45-77

Salton, G., and McGill, M.J. (1986). Introduction to modern information retrieval. McGraw-Hill, New York, NY, US

Salton, G., Wong, A., and Yang, C.S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18 (11): 613-620

Shein, M., and Lanza, R. B. (2004) Top Audit Tests Using ActiveData for Excel. Available at: http://www.auditsoftware.net/documents/auditebooksample.pdf

Sheeiderman, B. (1991). Tree Visualization with treemaps: a 2-d space-filling approach. ACM Transactions on Graphics. 11(1): 92-99.

Simon H.A. (1996). The Sciences of the artificial. 3rd ed. Cambridge, MA: MIT Press

Stringer, K. W. and Stewart, T. R. (1996). Statistical Techniques: For Analytical Review in Auditing. Wiley.

Tan, P., Steinbach, M., and Kumar, V. (2006). Introduction to data mining. Pearson Education.

Thomson Reuters Accelus. (2013) Eye on the horizon: Internal Audit's role in identifying emerging risks.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? American Psychologist, 24: 83-91.

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

Tukey, J. W. (1980). We need both exploratory and confirmatory. The American Statistician, 34: 23-25.

Tukey, J. W. (1986a). Data analysis, computation and mathematics. In L. V. Jones (Ed.), The collected works of John W. Tukey: Vol. IV. Philosophy and principles of data analysis: 1965-1986 (pp.753-775). Pacific Grove, CA: Wadsworth. (Original work published 1972)

Tukey, J. W. (1986b). Exploratory Data Analysis as part of a larger whole. In L. V. Jones (Ed.), The collected works of John W. Tukey: Vol. IV. Philosophy and principles of data analysis: 1965-1986 (pp. 793-803). Pacific Grove, CA: Wadsworth. (Original work published 1962)

Tukey, J. W. (1986c). The future of data analysis. In L. V. Jones (Ed.), The collected works of John W. Tukey: Vol. III. Philosophy and principles of data analysis: 1949-1964 (pp. 391-484). Pacific Grove, CA: Wadsworth. (Original work

published 1962).

Tukey, J, W. (1986d). Methodological comments focused on opportunities. In L.V. Jones (Ed.), The collected works of John W. Tukey. Volume IV: Philosophy and principles of data analysis: 1965-1986 (pp. 819-867). Belmont, CA: Wadsworth. (Original work published 1980)

Tukey, J. W., and Wilk, M. B. (1986). Data analysis and statistics: An expository overview. In L. V. Jones (Ed.), The collected works of John W. Tukey: Vol. IV. Philosophy and principles of data analysis: 1965-1986 (pp. 549-578). Pacific Grove, CA: Wadsworth. (Original work published 1966)

Van der Aalst, W. (2011). Process mining: Discovery, Conformance and Enhancement of Business Processes. Springer Verlag, Berlin (ISBN 978-3-642-19344-6)

Wasserman, S., and Katherine F. (1994). Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.

Weisberg H.F (1992) *Central Tendency and Variability*, Sage University Paper Series on Quantitative Applications in the Social Sciences.

Wilk, M. B., and Gnanadesikan, R. (1968) Probability plotting methods for the analysis of data. Biometrika 55 (1): 1-17

Wilkinson, L., and Friendly, M. (2009). The History of the Cluster Heat Map. The American Statistician. 63(2): 179-184.

Yu C.H. (1994). Induction? Deduction? Abduction? Is there a logic of EDA? Paper presented at the Annual Meeting of American Educational Research Association, New Orleans, LA (ERIC Document Reproduction Service No. ED 376173)

Yu C.H. (2010). Exploratory Data Analysis in the Context of Data Mining and Resampling. International Journal of Psychological Research, 3(1): 9-22.

Chapter 2 A Conceptual Framework to Apply Exploratory Data Analysis in Audit Practice

2.1 Introduction

Auditing is based on data. The auditing process entails substantive data analysis (Liu and Vasarhelyi, 2014). Various data analysis techniques, such as regression analysis (Stringer, 1975; Kinney, 1978), ratio analysis (Beneish, 1999; Grove and Cook, 2004; Kaminski et al., 2004), and artificial neural networks (Calderon and Cheh, 2002; Lin et al., 2003; Koskivaara, 2004), have been proposed and applied in the audit process. However, most of the data analysis techniques currently used in auditing focus on validating predefined audit objectives derived from general management assertions. This kind of data analysis approach, which is known as confirmatory data analysis (CDA), can provide precise evaluation of whether an audit objective has been achieved, but it cannot discover the risks not included in the existing audit objectives.

To identify unrecognized risk areas, auditors can investigate the auditees' data to detect abnormal items hidden in the data. As discussed in the previous chapter, EDA is able to serve this purpose because it can explore the data to unearth concealed patterns and identify unusual trends. In addition, this analysis is not constrained by any specific, predefined assumptions. This feature enables auditors to extend their audit scope to go beyond the existing audit objectives and look for new risk indicators. Moreover, the goal of EDA is to discover what happened in the data and to identify outliers and anomalies, which is the same as some of the auditors' responsibilities. Therefore, EDA, by its nature, can be a useful data analysis approach for auditors to achieve some of their objectives, such as fraud detection and risk assessment.

EDA has already been widely used in many disciplines, such as geography, marketing, and operations management (Chen et al., 2011; Nayaka and Yano, 2010; Koschat and Sabavala, 1994; Wesley et al., 2006; De Mast and Trip, 2007). In contrast to this intensive activity in other academic disciplines, there have been only a handful of papers in accounting that have discussed some EDA techniques, such as cluster analysis (Thiprungsri, 2011), and their application in auditing. Moreover, these papers focus either on one EDA technique or a specific audit obligation. Ways that auditors could benefit from the EDA approach and employ various EDA techniques in different audit stages have not been properly studied in the current literature.

Compared with arbitrarily using one or more EDA techniques as audit tools in certain audit procedures, systematically integrating EDA throughout the audit cycle can maximize the value that EDA can bring to auditors. Therefore, this chapter explores the potential application of EDA in various audit stages in both the internal and external audit cycles. In addition, it also proposes detailed guidelines for auditors to implement EDA in both traditional and continuous auditing settings.

The objective of this chapter is to demonstrate how auditors can benefit from integrating EDA into the audit process by identifying the value that EDA can add when applied to auditing. Specifically:

1. EDA allows auditors to identify emerging risks and existing risks that have not previously been recognized.
2. EDA techniques can analyze the entire population of an auditee's data, not just a sample.

3. Abundant EDA techniques enable auditors to assess risks based on both financial and non-financial information that can be collected from inside or outside of the organizations (such as media press and social media websites).

4. EDA allows auditors to conduct analyses not possible with existing audit tools, such as revealing associations between entities and relationships between individuals, or uncovering business processes.

5. By combining EDA with CDA, auditors can create new audit objectives while they are testing the existing ones, thereby continuously improving audit quality and efficiency.

6. Newly developed EDA and CDA techniques provide more analytical power for auditors not only to give a high level of assurance, but also to identify new business opportunities or directions for the organization.

The reminder of this chapter is organized as follows. The next section conducts a literature review of the prior research related to EDA's current applications in accounting and other disciplines. Section 3 discusses the framework for auditors to apply EDA in both internal and external auditing contexts, including when EDA can be applied in an audit cycle, how various EDA techniques may benefit auditors in different audit procedures, and specifically what auditors need to do in order to guarantee the best practice of EDA. Section 4 then discusses how the EDA application process can be integrated into a continuous auditing environment. Finally, Section 5 offers concluding comments.

2.2. Prior Research in EDA Application

2.2.1 Related Research in other Disciplines

Since the 1980s, EDA has been applied to diverse disciplines, such as interior design, marketing, industry engineering, and geography. In 1987, Kouka and Saucier first used EDA to analyze the information of floorplan building blocks. They generated a floorplan topology from the data to facilitate the design of floorplans for customers.

In industrial engineering, EDA has been used to analyze a comprehensive multilevel contemporary cycle for stocks and flows of copper (Graedel et al., 2004). Various EDA techniques are used in this study, such as bar charts, box plots, scatter plots, and Q-Q plots. The results reveal unexpected characteristics of the copper cycles, and demonstrate the usefulness of EDA in material flow analysis. De Mast and Trip (2007) also present the value of EDA in quality improvement projects. In this article, they first propose a framework to employ EDA in quality-improvement projects, and then provide several examples to illustrate applications.

Because of its visualization focus, EDA is widely used in geography to display spatial data. For example, Gluck (2001) attempts to use EDA in spatial data analysis. He develops a tool incorporating EDA with geographical map, sound, and graphic interactions to exploit spatial datasets to gain a better understanding of the relationships among spatial, temporal and human variables. In recent geographical research, EDA is likely to be included in geography information systems (GIS) to identify spatiotemporal patterns. For instance, Shaw and Xin (2003) design a GIS that offers EDA capabilities to examine the interaction of land use and transportation at user-specified spatial and temporal scales. Chen et al. (2011) develop a space-time GIS approach that can represent

and analyze spatiotemporal activity data at the individual level. By incorporating EDA techniques, this approach uses a multi-level clustering method to investigate individual-level spatiotemporal patterns covering a set of functions, such as space-time path generation, space-time path segmentation, space-time path filter, and activity distribution/density pattern exploration. Nakay and Yano (2010) investigate the spatiotemporal patterns of crime clusters by mapping crime events in a space-time cube. They show that the space-time cube display, which is an EDA tool, can differentiate stable and transient space-time crime clusters and reveal temporal inter-cluster associations.

In business, EDA is mainly applied in marketing, and is usually used as supplementary tools to facilitate or improve the building of a confirmatory model. For example, Koschat and Sabavala (1994) combine EDA with formal modeling to investigate the effects of television advertising on local telephone usage. Their EDA techniques include data visualization, data transformation and robust non-linear smoothers[1]. Using a case study, they demonstrate the value of EDA in model specification and refinement. Wesley et al. (2006) use EDA together with comparative methods to assess how consumers' decision-making styles relate to their shopping mall behavior and their global evaluations of shopping malls. In their study, applied EDA techniques include data visualization, exploratory factor analysis and tipping point analysis[2]. EDA results support a complex view of the antecedents and consequences of

[1] This is a method to identify patterns from noisy data. It is proposed by Velleman in a paper entitled Robust nonlinear data smoothers: Definitions and recommendations and published in Proceedings of the National Academy of Sciences of the United States of America in 1977.
[2] Tipping Point Analysis is a tool to assess the impact of missing data on the conclusion of a

consumer decision-making styles. Lombardo and Valle (2011) use EDA and data mining tools to display survey and analysis results in their study evaluating employees' satisfaction at non-profit enterprises.

2.2.2 Related Research in Auditing Discipline

In the auditing domain, EDA theory has seldom been explicitly mentioned in the literature, but some EDA techniques, such as data visualization, cluster analysis, text mining, process mining and social network analysis, have been studied as audit tools to fulfill certain audit objectives, such as fraud detection.

For example Cox et al. (1997) build a suite of visual interfaces to display telephone calls and help auditors to detect telephone calling fraud. Sokol et al. (2001) apply descriptive statistics and visualization tools in a healthcare claim audit setting. They claim that descriptive statistics (e.g.: categorical statistics) and visualization graphs (e.g.: pie charts, column charts, and scatter plots) can help healthcare investigators to recognize new and unusual patterns of activity, thus allowing a better understanding of the potential problems and better use of limited health care fraud detection and investigation resources.

Feature selection techniques, such as principal component analysis, have also been used in fraud detection. Brockett et al. (2002) apply this technique to classify automobile insurance claims as fraud or non-fraud when labeled fraudulent cases are not available. According to this model, a set of variables related to automobile insurance

---

study, first introduce by Yan et al. in a paper entitled Missing data handling methods in medical device clinical trials and published in Journal of Biopharmaceutical Statistics in 2009.

claims is analyzed by principal component analysis to generate scores indicating the likelihood that the claims are fraudulent.

Cluster analysis can be used in the auditing domain to improve auditors' understanding of clients' data. For example, Thiprungsri (2012) applies cluster analysis to the transitory accounts information of an international bank to assist internal auditors in understanding the regularity and behavior of these transitory accounts. Another application area for cluster analysis is anomaly detection. Zaslavsky and Strizhak (2006) employ self-organizing maps, a cluster analysis algorithm, on credit card transactions to create a model of normal cardholder behavior and to analyze deviations in transactions, thus identifying suspicious ones. Quah and Sriganesh (2008) utilize the same algorithm to cluster online credit card transactions in order to provide a better understanding of spending patterns and achieve real-time detection of potential fraudulent cases. Thiprungsri (2011) also employs cluster analysis on an insurance company's wire transaction data to identify abnormal transactions.

Text mining, as an EDA technique, has not been studied much by auditing researchers. Holton (2009) applies a text-clustering technique to messages posted on Vault.com and Yahoo! discussion groups to identify disgruntled employees in an organization who may perpetrate occupational fraud. However, few other auditing studies consider text mining as a fraud detection tool.

Process mining was first employed in auditing by Yang and Hwang (2006), who use process mining to formalize clinical pathways in order to detect healthcare service providers' fraudulent and abusive behavior. Jans et al. (2009) propose a framework for internal fraud risk reduction in an advanced IT integration environment. In this

framework, process mining is used on system event logs to discover real business processes. Potential internal fraud risks can be identified by comparing the real business processes and the expected business processes. Jans et al. (2013) identify the general value that process mining can add to auditing, and point out that, in addition to fraud risk analysis, process mining can also be used to facilitate auditors' ability to understand a client's business and assess business and internal control risks. Furthermore, Jans et al. (2014) propose a protocol for auditors to use process mining in analytical procedures. This protocol incorporates social network analysis with process mining to identify collusive frauds by revealing unexpected relationships between individuals participating in the business process.

Other than being used together with process mining, social network analysis can also be combined with text mining to assess fraud risk. Debreceny and Gray (2011) first use text mining techniques to analyze the body of Enron emails to filter suspicious emails, and then conduct social network analysis of the senders and recipients of these suspicious emails to identify the participants in the potentially fraudulent activities.

Besides the traditional auditing context, some EDA techniques are proposed to be used in continuous auditing systems as well. Rezaee et al. (2002) develop a continuous auditing system with an integrated audit data mart to facilitate auditors' analytical requirements. According to their design, a set of EDA techniques, such as descriptive statistics, data visualization graphs, feature selection and cluster analysis, should be included in the integrated audit data mart.

Table 4 summarizes the application areas of EDA techniques in the current auditing literature. Fraud detection is the most emphasized application area of EDA; all

of the EDA techniques listed in Table 4 have been employed by auditing researchers to assess fraud risk. This may attribute to the fact that fraudulent behaviors are usually concealed in the data and may change over time (Palshikar, 2002). Therefore, deep insight into the data and extensive exploration of the data from different aspects are necessary for effective fraud investigations. EDA techniques can assist auditors to identify various unreasonable or unusual cases in the data that may be attributable to fraud.[3] In addition, it is very difficult for researchers to obtain identified fraud data. In this circumstance, EDA techniques rather than CDA techniques can be applied to identify potentially fraudulent cases. In addition to fraud detection, the other application areas (e.g.: understanding a client's business and assessing business and internal control risks) have been studied as well. But the research effort devoted to these areas is very limited. For example, cluster analysis and process mining are the two EDA techniques employed to understand clients' businesses, and only process mining has been applied to assess business risks and internal control risks. Can the other EDA techniques be used to serve these purposes? If so, how should they be applied to these areas? These two research questions are worth studying, and will be discussed in the next section. Even though some EDA techniques have been proposed for use in the continuous auditing environment, the ways that auditors should implement these techniques has not been discussed yet. Furthermore, some existing EDA techniques, such as association analysis, have not been applied in auditing yet. How these techniques can benefit auditors is also addressed in the next section.

[3] Usually, the cases identified as suspicious may be ascribed either to fraud or errors. Since fraud usually indicates intentional misbehavior, whereas an error is an unintentional mistake, the issue of whether identified instances are fraud or errors must be investigated further with additional supporting evidence.

Table 4: Summary of the Application of EDA Techniques in Current Auditing Literature

EDA technique	Application areas
Descriptive Statistics	Fraud Risk Analysis (Sokol et al., 2001) Continuous Auditing (Rezaee et al., 2002)
Data Visualization	Fraud Risk Analysis (Cox et al., 1997; Sokol et al., 2001) Continuous Auditing (Rezaee et al., 2002)
Feature Selection	Fraud Risk Analysis (Brockett et al., 2002)
Cluster Analysis	Understand client's business (Thiprungsri, 2012) Fraud Risk Analysis (Zaslavsky and Strizhak, 2006; Quah and Sriganesh, 2008; Thiprungsri, 2011) Continuous Auditing (Rezaee et al., 2002)
Text Mining	Fraud Risk Analysis (Holton, 2009)
Process Mining	Understand client's business (Jans et al., 2013) Business Risk Assessment (Jans et al., 2013) Internal Control Assessment (Jans et al., 2013) Fraud Risk Analysis (Yang and Hwang, 2006; Jans et al., 2009)
Social Network Analysis	Fraud Risk Analysis (Debreceny and Gray, 2011; Jans et al., 2014)

2.3 EDA Application Framework in Auditing

This section introduces a framework (Figure 21) to guide auditors' application of EDA. In this framework, three aspects of EDA applications are covered. (1) Audit Flow: When can auditors apply EDA in the audit process? Specifically, this discussion goes through the stages in a typical audit cycle and lists EDA's potential application in each stage by referring to both internal and external audit standards. (2) Means: How can various EDA techniques be applied to different audit requirements? All application areas of each EDA technique are illustrated. (3) Process: What are the recommended steps for

auditors to follow when implementing EDA? These steps provide formalized guidelines for auditors to integrate EDA procedures into traditional CDA-based audit tests.

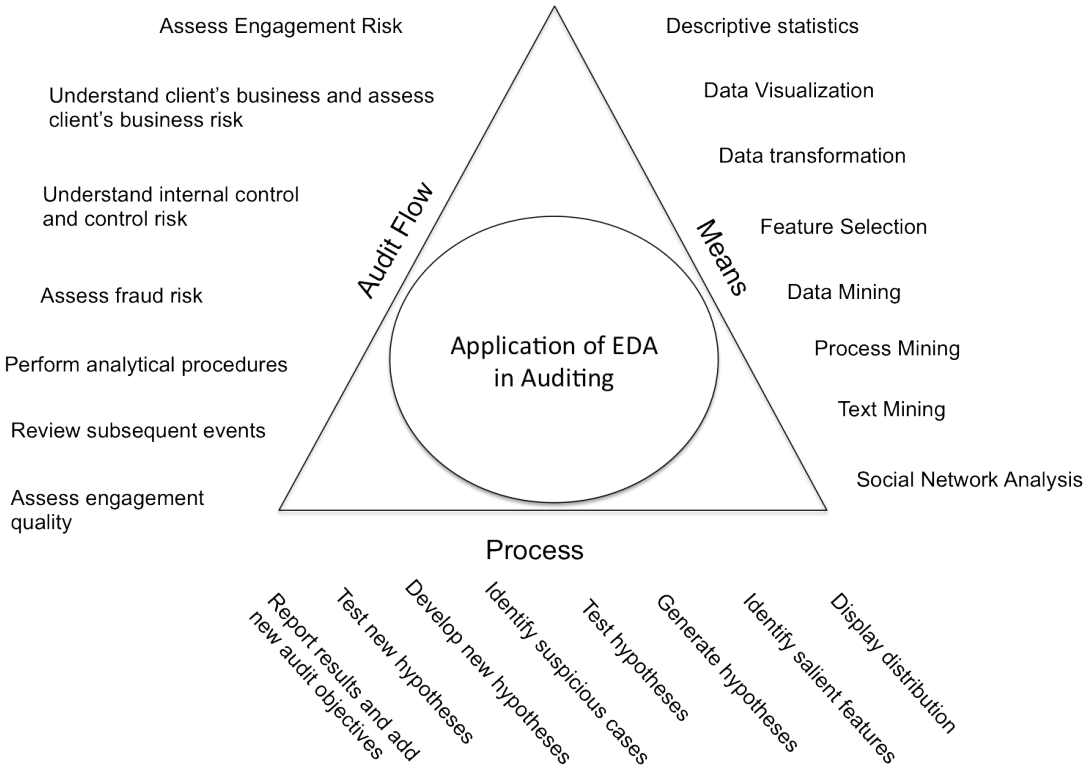


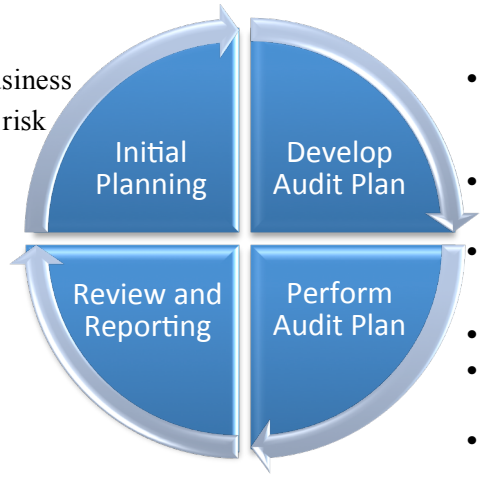
Figure 21: EDA Application Framework

2.3.1 Audit Flow

As discussed in the previous sections, EDA focuses on understanding data, recognizing patterns, and detecting outliers. Therefore, it can be used as a tool for auditors to understand clients' businesses, identify unusual situations, and discover hidden risks. These three audit purposes extend over all of the steps in a typical external audit cycle,[4] including the four basic stages (initial planning, developing an audit plan,

[4] There are different models of the audit cycle. They all contain similar phases including all of the major steps in the audit process. Here, we adopt one that used at Deloitte.

performing the audit plan, and reviewing and reporting), shown in Figure 22. Therefore, generally speaking, EDA can be applied throughout the external audit cycle.



- Assess and respond to engagement risk
- Understand client's business
- Assess client business risk
- Perform preliminary analytical procedures

- Set materiality and assess acceptable audit risk and inherent risk
- Understand internal control and assess control risk
- Gather information to assess fraud risks
- Develop overall audit

- Perform tests of controls
- Substantive tests of transactions
- Perform analytical procedures
- Test of details of balances

- Perform Subsequent events review
- Issue audit report
- Assess engagement quality

Figure 22: External Audit Cycle

Specifically, in the "initial planning" stage, EDA can be utilized to fulfill most of the audit tasks. For example, when deciding whether to accept a new client, auditors can use EDA to investigate the risks in the new client's industry, to explore its standing in the business community, and to gain an understanding of its financial stability. Also, auditors can employ EDA in preliminary analytical procedures to gain a general understanding of the client's business and identify data-driven indicators of risks or emerging risks in a client's organization. This can enable auditors to assess the client's business risks. Some of these potential EDA application areas in the initial planning stage have been pointed out in the audit standards issued by the AICPA. For example, in audit standard AU-C sec. 315[5] (Understanding the Entity and Its Environment and Assessing the Risks of Material Misstatement), it states that "Examples of matters that the auditor may consider when

[5] http://www.aicpa.org/Research/Standards/AuditAttest/DownloadableDocuments/AU-C-00315.pdf

obtaining an ***understanding*** of the nature of the entity include…accounting for ***unusual*** or complex transactions, including those in controversial or ***emerging*** areas (for example, accounting for stock-based compensation)." In addition, AU-C sec. 550 points out that auditors need to "obtain an ***understanding*** of the related party relationships and transactions" when they are doing risk assessment. Particularly, they need to consider whether a transaction (1) "has ***unusual terms*** of trade, such as unusual prices, interest rates, guarantees, and repayment terms", or (2) involves ***previously unidentified*** related parties", or (3) "is processed in an ***unusual manner***". By employing EDA in auditing, auditors are able to reveal these irregular transactions and suspicious relationships, thereby increasing the accuracy of their risk assessment. A complete list of potential application areas for EDA illustrated in AICPA clarified statements on auditing standards is included in Appendix A: Potential application areas of EDA in Clarified Statements on Audit Standards issued by AICPA.

In the "develop audit plan" stage, auditors can utilize EDA to understand a client's internal controls and assess various risks by identifying salient features in the client's data, which enables the creation of audit plans that focuses on the areas of the highest concerns. Audit standards related to the audit plan development stage are AU-C sec. 240[6] (Consideration of Fraud in a Financial Statement Audit), and AU-C sec. 500[7] (Audit Evidence). A number of potential application areas for EDA are mentioned in AU-C sec. 240. Specifically, auditors should "obtain an ***understanding*** of the entity's financial reporting process and controls over journal entries and other adjustments",

[6] http://www.aicpa.org/Research/Standards/AuditAttest/DownloadableDocuments/AU-C-00240.pdf
[7] http://www.aicpa.org/Research/Standards/AuditAttest/DownloadableDocuments/AU-C-00500.pdf

identify "individuals involved in the financial reporting process about inappropriate or ***unusual*** activity relating to the processing of journal entries and other adjustments", and analyze "significant transactions that are ***outside the normal*** course of business for the entity or that otherwise appear to be ***unusual*** given the auditor's ***understanding*** of the entity and its environment and other information obtained during the audit". These potential application areas explain why fraud detection is the dominant EDA application area in current accounting literature. AU-C sec. 500 (Audit Evidence) also indicates that, when selecting audit evidence, auditors need to identify "significant or ***unusual*** items to test". EDA can be used in analytical procedures to discover these significant or unusual items, including nonstandard journal entries and exceptional transactions, to specify the audit evidence that will need to be examined in substantive tests.

Data analysis procedures in the "perform audit plan" stage are mainly in the form of substantive analytical procedures. Substantive analytical procedures that are normally performed actually belong to CDA based on expectations developed from recorded data. However, some EDA procedures, such as variable selection, can be utilized in the expectation development process to increase the precision of developed expectation, thus improving the effectiveness of the analytical procedure. This application requirement is pointed out in paragraph .A8 in audit standard AU-C sec. 520[8] (Analytical Procedures). In addition, by performing EDA on the more disaggregated data and new audit evidence identified in this stage, auditors can discover additional risky areas that had not been recognized in the planning stage. After confirmation with CDA, newly identified risks

[8] http://www.aicpa.org/Research/Standards/AuditAttest/DownloadableDocuments/AU-C-00520.pdf

can be converted into new audit objectives and added to the audit checklist. In this way, auditors can continuously improve the audit plan during the course of audit.

Finally, in the "review and reporting" stage, EDA can be used in the subsequent events review to identify significant events that have occurred between the date of the financial statements and the date of the auditor's report. EDA can also be utilized in engagement quality assessment procedures to inspect whether there are any risks that were missed or ignored in the previous stages. Audit standards, such as AU-C sec. 330[9] (Performing Audit Procedures in Response to Assessed Risks and Evaluating the Audit Evidence Obtained), AU-C sec. 520 (Analytical Procedures), and AU-C sec. 560 (Subsequent Events and Subsequently Discovered Facts), imply these applications. Eventually, all of the EDA processes and results should be included in the audit working papers. This helps to ensure that the conclusions drawn from EDA can be relied on and are error-free.

The internal audit cycle is very similar to the external audit cycle. Therefore, EDA can also be applied to similar parts throughout the internal audit cycle. Since internal audit standards (IIA standards)[10] are not as detail as external audit standards, detailed EDA application areas are not listed in the IIA standards. However, some general EDA application areas can still be identified from the IIA standards. For example, EDA can be applied to "a ***preliminary assessment of the risks*** relevant to the activity under review" required in IIA standard 2200 (Engagement Planning), and to evaluate "the ***potential for the occurrence of fraud***" and "organization's ***risk***

[9] http://www.aicpa.org/Research/Standards/AuditAttest/DownloadableDocuments/AU-C-00330.pdf
[10] https://na.theiia.org/standards-guidance/Public%20Documents/IPPF%202013%20English.pdf

management processes" regulated in IIA standard 2120 (Risk Management). Appendix B: Potential application areas of EDA in International Standards for the professional Practice of Internal Auditing issued by IIA presents a complete list of potential application areas for EDA indicated in the IIA standards.

2.3.2 **Means**

As discussed in the last section, EDA can provide support to auditors in audit activities including assessing engagement risk, understanding a client's business and assessing a client's business risk, understanding internal control and assessing control risk, assessing fraud risk, performing analytical procedures, reviewing subsequent events, and assessing engagement quality. This section focuses on how auditors can apply different EDA techniques (introduced in Chapter 1, section 1.3 Exploratory Data Analysis Techniques) to various audit evidence to accomplish these purposes.

The first application area in an audit cycle is engagement risk assessment. When assessing a client's engagement risk, auditors are mainly concerned about the client's financial situation, integrity, and reputation. In this stage, auditors can apply data visualization techniques to industry-wide and organization-wide financial data to explore a client's standing in the business community and the client's financial stability. To investigate a client's integrity and reputation, auditors usually need to ask the predecessor auditors or gather information from third parties, such as local attorneys and banks. If information collected in this stage is in text format, text mining techniques can be applied to search for clues related to predecessor auditors' and third parties' attitudes about the client's integrity and reputation.

After accepting the client, auditors need to understand the client's business and assess its business risk, which requires auditors to understand the client's industry and external environment, business operations and processes, management and governance, objectives and strategies, and measurement and performance. Among these five aspects, an understanding of the client's industry and external environment can be supported with data visualization techniques by graphing economic information related to the client's industry and its major competitors. For example, auditors can understand the impact of economic volatility on the industry by displaying the last ten years of revenue for the client's major competitors. In addition, text mining allows auditors to examine industry-wide text information, such as industry-specific regulations and news reports, to gain an understanding of the client's external environment. In the client's organization, if information systems are used to manage its business operation and processes, auditors can apply process mining to the system event logs to understand the client's business processes. They can perform social network analysis in combination with the process mining to gain insight into employees' roles in each business process. Another concern related to the client's business operation is related parties. Descriptive statistics can be used to analyze the client's business transactions to identify any significant or unusual transactions associated with related parties. To understand the client's organizational structure, management style, and business objectives and strategies, auditors usually need to examine various organizational documents, such as minutes of meetings and all kinds of contracts. Text mining can also be applied here if these documents can be transferred into digital format. Text mining can also be used on external information, such as client-related news reports and posts on social media websites, to facilitate the auditors'

understanding of the client's performance. To achieve the same goal, auditors can use data visualization techniques to display key performance indicators in the organization.

If the auditee utilizes an information system to perform and manage its internal controls, process mining can be applied to the internal control system event logs to discover the internal control processes within the organization, which will allow the auditors to have a comprehensive understanding of how the internal controls are actually performed in the client's organization. By comparing the identified internal control process with the ideally designed process, auditors can reveal potential risk areas in current internal control practices.

As shown in Table 4: Summary of the Application of EDA Techniques in Current Auditing fraud risk assessment is a major application area for EDA in auditing. Since fraud may happen at any step of business processes, and the behavior of fraud perpetrators may evolve according to existing fraud detection methods, auditors need to understand and analyze organizational data thoroughly when assessing fraud risk. Therefore, both traditional and advanced EDA techniques can be applied to achieve this purpose. For example, to identify instances of potential fraud in the accounting data, descriptive statistics and data visualization techniques can be used on accounts-related information and business transactions to reveal unreasonable trends and unexpected distributions in the data. Sometimes, data transformation techniques are required in combination with data visualization techniques to improve the interpretability or appearance of graphs. Data mining techniques, such as clustering and association analysis, can unearth unusual accounts and transactions hidden in the data. For example, cluster analysis can be utilized to identify suspicious vendors by their payment profile. In

this case, vendors that provide similar products or services are clustered into the same groups. If a vendor is paid for very different products or services compared to the other vendors in the same group, it is considered as a suspicious vendor. Furthermore, association analysis can be performed to generate common combinations of raw materials that are usually purchased from vendors. Unusual situations can be identified when odd combinations are discovered. Internal auditors may expand the scope of their examination to operational data in order to identify deficiencies in business operations. In addition to potential fraud in business operations, internal audit analysis results can discover other valuable information as well, such as potential new business opportunities.

Besides accounting and transaction data, other types of information can also be analyzed by EDA techniques to assess fraud risk. For example, social network analysis can be applied to communication records, such as phone calls, text messages, and emails, to discover suspicious relationships between individuals inside and outside of the organization. This analysis is usually used together with text mining to investigate the text information contained in the communications. Social network analysis can also assist auditors to identify undisclosed related parties. Moreover, process mining can be employed on more detailed business processes to reveal weak segments that are vulnerable to fraud.

In analytical procedures conducted at the “perform audit plan” stage, feature selection techniques are traditionally used to choose the most suitable variables for confirmatory models, such as regression models, used to build expectations. In addition, as mentioned in last section, a new application of EDA in analytical procedures is to analyze more disaggregated data used in the analytical procedures to explore risk areas

that had not been identified during the planning stage. Since the data used in analytical procedures are usually in traditional formats, EDA technologies can be employed in this stage, include descriptive statistics, data visualization, data transformation, and data mining techniques. Salient features identified by EDA techniques can then be used to develop new expectations or revise current expectations that need to be tested by confirmatory methods. Confirmed expectations can be considered as new audit objectives and added to the existing objective list to create a close-loop learning environment for the audit process. For new audit evidence identified in this stage, similar processes should be applied to analyze the new audit evidence thoroughly. Detailed steps to apply EDA in analytical procedures are discussed in the next section.

Since the subsequent events review and engagement quality assessment are performed near the audit report date, it is infeasible to conduct complicated analyses at that time. Therefore, descriptive statistics and data visualization techniques are suitable to be applied to these two application areas. Specifically, they can be used to explore subsequent transactions to identify extreme transactions that may have an effect on the current year's financial statements. They can also be included in the analytical procedures in the completion phase to analyze the audit evidence used in the previous audit procedures to reveal unrecognized or missed risk areas. Table 5 provides a summary of how different EDA techniques can be applied to various audit evidence in each audit stage.

Table 5: Summary of the Application of EDA Techniques in Auditing

Audit tasks	**EDA technologies**	**Audit Evidence**
Assess engagement	Data visualization	Industry-wide and organization-wide

risk		financial data
	Text mining	Text documents collected from predecessor auditors and third parties
Understand client's business and assess client's business risk	Descriptive statistics	Business transactions
	Data visualization	Economic information related to the client's industry and its major competitors Key performance indicators
	Process mining	Business process event logs
	Social network analysis	
	Text mining	Industry-wide text information Client-related text information from both inside and outside of the organization.
Understand internal control and assess control risk	Process mining	Internal control process event logs
Assess fraud risk	Descriptive statistics	Accounting data Operational data (Internal audit)
	Data visualization	
	Data transformation	
	Data mining	
	Social network analysis	Communication data
	Text mining	
	Process mining	Business process event logs
Perform analytical procedures	Descriptive statistics	Audit evidence specified in the planning stage
	Data visualization	

	Data transformation	New audit evidence identified in this stage
	Feature selection	
	Data mining	
Review subsequent events	Descriptive statistics	Subsequent transactions
	Data visualization	
Assess engagement quality	Descriptive statistics	Audit evidence used in previous audit processes
	Data visualization	

2.3.3 **Process**

From the previous discussion, EDA is a powerful tool for auditors, and it can be used in almost every stage in the audit cycle. However, inappropriate or incomplete use of EDA can counteract its benefits and may lead to inaccurate results. Inspired by the framework proposed by De Mast and Kemper (2009),[11] a suggested procedure for auditors to perform EDA is proposed in this section to enhance the appropriate use of EDA. This process considers the specific features of auditing, prevents the misuse of EDA, and ensures the correctness of EDA and subsequent CDA results.

Figure 23 demonstrates this process. The first step in this process is to display the distribution of the fields related to the audit subject. Different EDA techniques can be employed in this step to analyze different types of data. For example, descriptive statistics are usually used on numerical or categorical data, text mining is for text data, social network analysis is applied to individuals, association analysis is used for items, and process mining is applicable to event logs. Then various data visualization techniques

[11] This framework is discussed on page 7

can be used to display these distributions. Theoretically, distribution analysis can be performed on every field in the dataset. However, considering the timing and cost constraints in auditing, it is infeasible for auditors to do this. Therefore, the most efficient way for auditors to conduct EDA is to start from the audit subject-related fields. For example, in an accounts payable audit, auditors should apply descriptive statistics to fields like payment amount, payment nature, and account number, social network analysis to vendors and purchase clerks, and association analysis to purchased items.

Then auditors can identify salient features from the displayed distributions. These salient features do not necessarily represent risk indicators; they may be caused by some special circumstances. Therefore, after the salient features are identified, auditors can generate potential explanations according to their own knowledge and experience, or they can communicate their findings to management in order to get explanations for these salient features. These explanations can then be used to create hypotheses to be tested in the following CDA steps.

In the next step, auditors use CDA techniques or substantive tests to verify the hypotheses derived from the explanations identified in the previous step. If all of the exceptional cases match with the hypotheses, the analysis terminates. Otherwise, auditors can consider the cases that cannot be justified by the explanations as suspicious cases.

After the suspicious cases are identified, auditors can start another round of EDA by using the techniques that can examine relationships between variables, such as scatter plot and feature selection, to explore possible causes for the suspicious cases. Identified possible causes can then be utilized to develop new hypotheses, which should be verified by CDA techniques or substantive tests in the following steps.

Confirmed hypotheses may reveal previously unknown risk areas. Therefore, auditors should report these suspicious cases along with the verified causes. The audit report should also include the risks inferred from the findings and recommendations for management to address these risks. In addition, new audit objectives, derived from the confirmed hypotheses, need to be added to the standard audit objective list so that the same risk will be tested in the next audit cycle.

Even though the proposed process includes eight normative steps, auditors can be flexible in its implementation, depending on the actual audit situation. For example, if there is no reliable information to test the hypotheses generated in the third step, auditors can skip the fourth step and perform more sophisticated EDA techniques, such as clustering, in the fifth step to identify suspicious cases. Patterns identified in this step can also be used as support evidence to verify the hypotheses. In addition, all eight steps may not need to be covered in every audit stage. For instance, the analytical procedures conducted in the "Initial Planning" stage usually stop at step three. By contrast, for the analytical procedures in the "Perform Audit Plan" stage, auditors should execute all the eight steps. Thus, when the auditors conduct CDA to evaluate certain audit objectives, they can simultaneously perform EDA on the same fields. As mentioned in the previous discussion, this is the most efficient way for auditors to implement EDA.

Another noteworthy feature in this process is that both the EDA and CDA approaches are included in this process. Therefore, by following this process, auditors use both EDA and CDA to perform audit tests, which is consistent with the best data analysis practice proposed by Tukey (1980, 1986d, 2000).

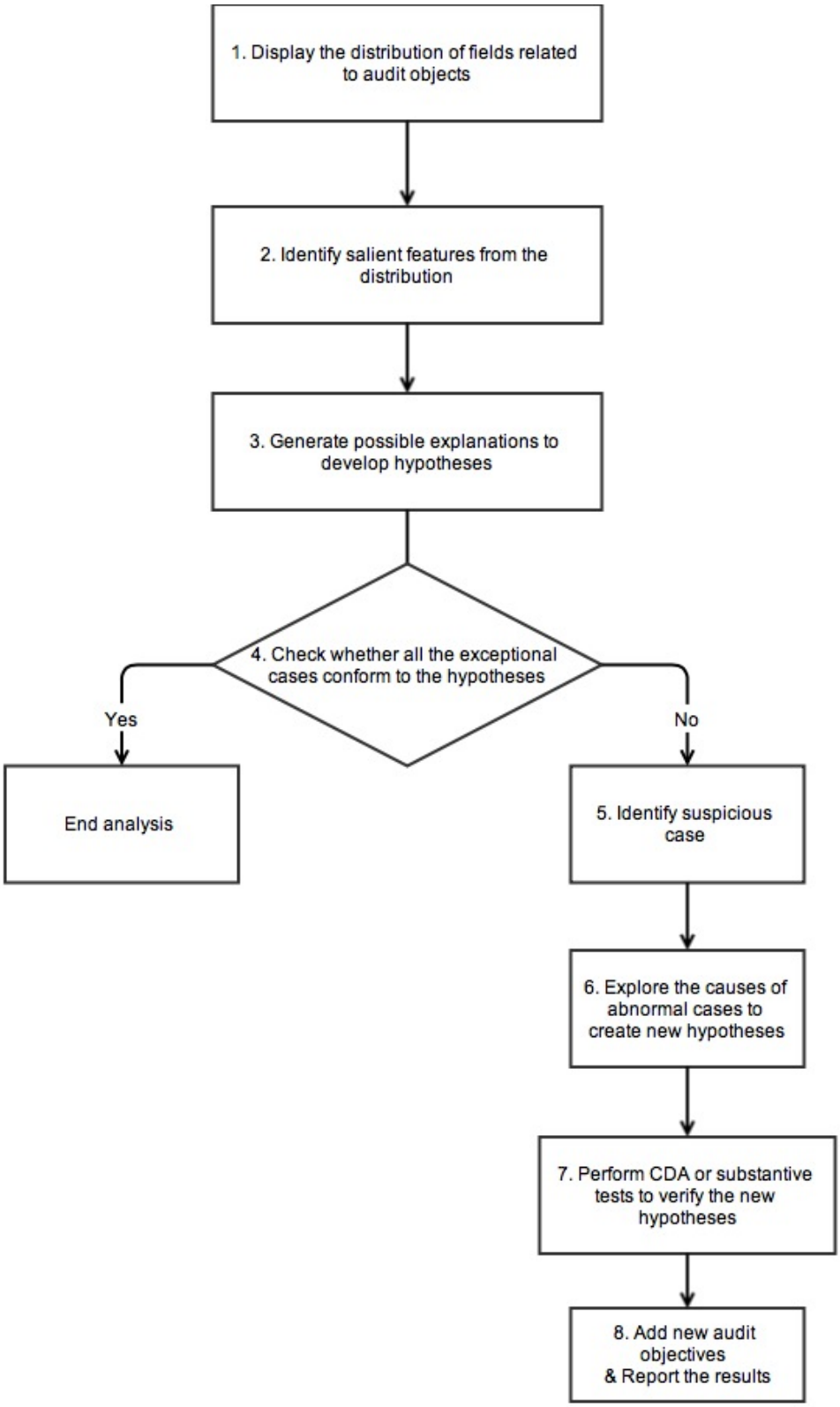


Figure 23: Steps to Perform EDA in Auditing

2.4 The Application of EDA in Continuous Auditing Environment

2.4.1 Overview of the Continuous Auditing Environment

Continuous auditing is a process of constantly testing transactions based on criteria prescribed by the auditor, and identifying anomalies (exceptions) on which the auditor can then perform additional procedures (Alles, 2007). The main features of continuous auditing include: (1) Continuous auditing can provide real-time or near real-time assurance to the organization and can support more frequent reporting; (2) Continuous auditing focuses on both financial and nonfinancial information, thus expanding the scope of traditional audit from providing assurance on financial statement alone to offering strategic-level guidance for the organization; (3) Continuous auditing changes the audit process from periodic testing to audit by exceptions identified by the continuous auditing system; (4) Information technology plays a key role in continuous auditing by automating the audit tests, analyzing transactions, and testing controls.

Continuous auditing is composed of three main parts: continuous data assurance, continuous controls monitoring, and continuous risk monitoring and assessment (Vasarhelyi, 2011). Continuous data assurance focuses on continuously and automatically analyzing transaction-level data in order to provide more detailed assurance (Kogan, 2014). In a continuous auditing system, audit tests are automated, and they constantly test transactions so that errors, anomalies, and exceptions can be identified in a timely manner for auditors to review. Continuous control monitoring is a complementary process of continuous data assurance, which relies on automatic procedures to monitor business process controls continuously (Alles, 2006). Continuous risk monitoring and assessment is a real-time integrated risk-assessment approach. It can assess risk exposures by

aggregating data across different functions in the organization to provide reasonable assurance on a firm's risk assessments (Vasarhelyi, 2011).

2.4.2 Integrating EDA into a Continuous Auditing System

As discussed in the previous section, EDA can be used in every stage in the audit cycle to identify hidden risks and emerging risks in an organization. However, the audit cycle in a traditional auditing setting is not applicable to a continuous auditing environment. In continuous auditing, most of the activities in the "Initial Planning" and "Develop Audit Plan" stages of the traditional audit cycle are performed during the continuous audit system design and development phase. Therefore, EDA should be applied thoroughly when designing the continuous auditing system following the process proposed in Chapter 2, section 2.3.3 **Process**in order to identify as many existing risks as possible. Audit tests that are used to examine the audit objectives generated from the EDA process need to be automated and built into the continuous auditing system.

After a continuous auditing system has been developed, EDA can be integrated as part of the continuous risk monitoring and assessment module. Continuous risk assessment requires continuous performance of EDA in order to identify the emerging risks in a timely manner. However, EDA requires the data population to contain a certain number of observations. For example, identifying any patterns or outliers from one or two records is not possible. To solve this problem, a certain amount of historical data can be stored in the continuous risk assessment module as the basis of the analysis. Patterns identified in historical data should be saved as benchmarks for subsequent analysis. When new data is input into the system, it should be analyzed by EDA techniques together with the base dataset. With new data continuously being added to the base

dataset, some old records that are no longer suitable to indicate current risks should be eliminated periodically from the dataset. Figure 24 demonstrates the change in the data used in EDA in a continuous auditing system, where rectangles with bold line represent the datasets for EDA at particular times.

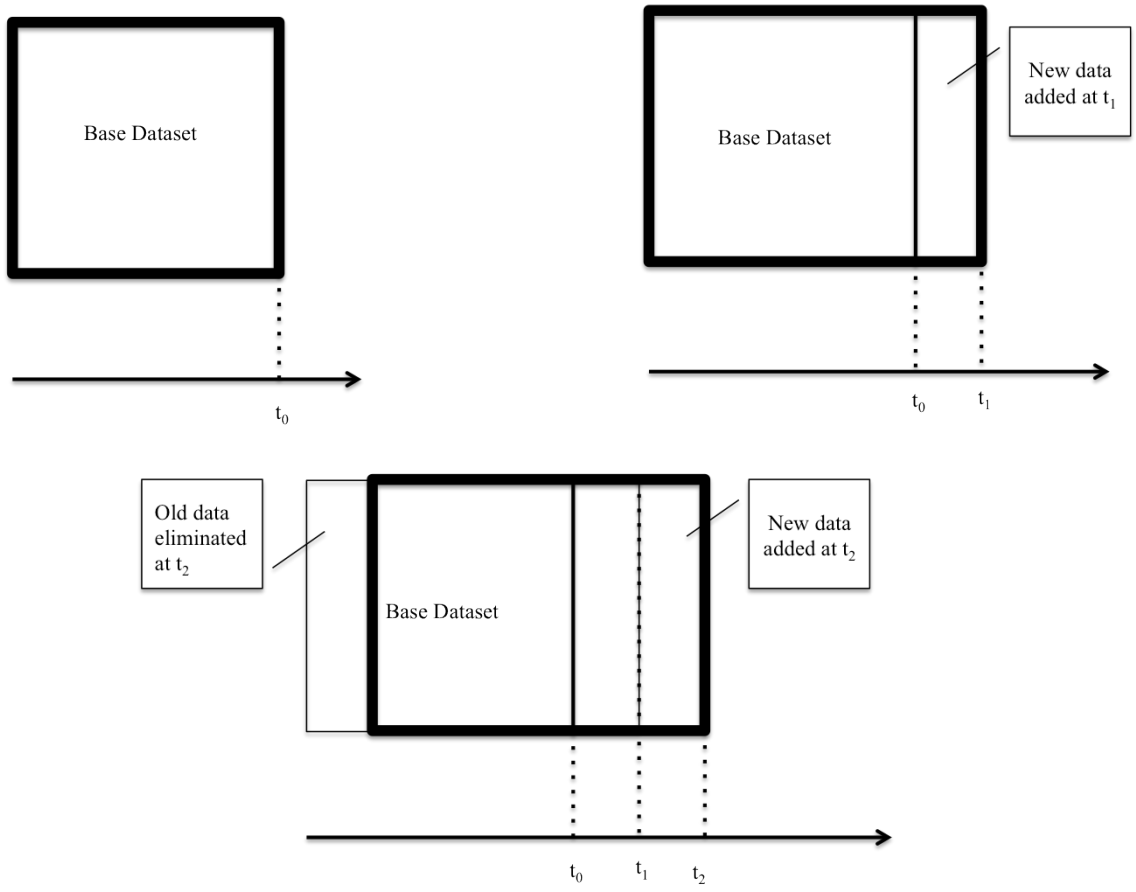


Figure 24: EDA Dataset in a Continuous Auditing System

In order to be integrated into the continuous auditing system, the EDA process needs to be automated. However, EDA is also an interactive process that requires auditors' definitions of salient features and choices of techniques and parameters, so not all of the activities in the EDA process can be automated. Nevertheless, some steps in the EDA process can still be automated to accommodate the continuous auditing environment. Figure 25 demonstrates how the EDA process can be automated in the

continuous auditing environment. In the first step, data distributions can be automatically displayed by assigning suitable EDA techniques to different types of data. For example, descriptive statistics of numeric fields can be routinely calculated and displayed. Social network analysis can be automatically performed on the fields containing individuals' identification information, such as vendors and clients. Association analysis can be automatically applied to items fields, such as purchase items or sold/returned items. The next three steps, identifying salient features, generating hypotheses, and choosing suitable tests to verify the hypotheses, need human involvement in a traditional audit setting. These activities can be automated by using expert systems or other artificial intelligent technologies to formalize auditors' and managements' judgments. For example, since salient features are usually patterns that do not comply with normal situations, all of the normal situations need to be defined by rules in the expert systems as benchmark. By comparing the benchmark distribution with the incoming distribution, exceptions and new patterns can be automatically identified. However, some exceptions may actually have legitimate reasons. Therefore, all of the knowledge related to these special cases needs to be formalized in the expert system to provide possible explanations for identified salient features. Based on these explanations, hypotheses can be generated. Each explanation should be associated with a specific test to examine the corresponding hypothesis. The cases that cannot be justified by those explanations should be reported to the auditors or managements because they cannot be interpreted by the system's knowledge. Then auditors and/or managements can manually select additional analyses and tests to explore and confirm the causes of these unexplained cases. Based on the tests results, they can judge whether a particular case is an opportunity or a risk to the

organization. Confirmed risks should be fixed, and derived new audit objectives need to be routinely tested to continuously enrich and improve the continuous auditing system.

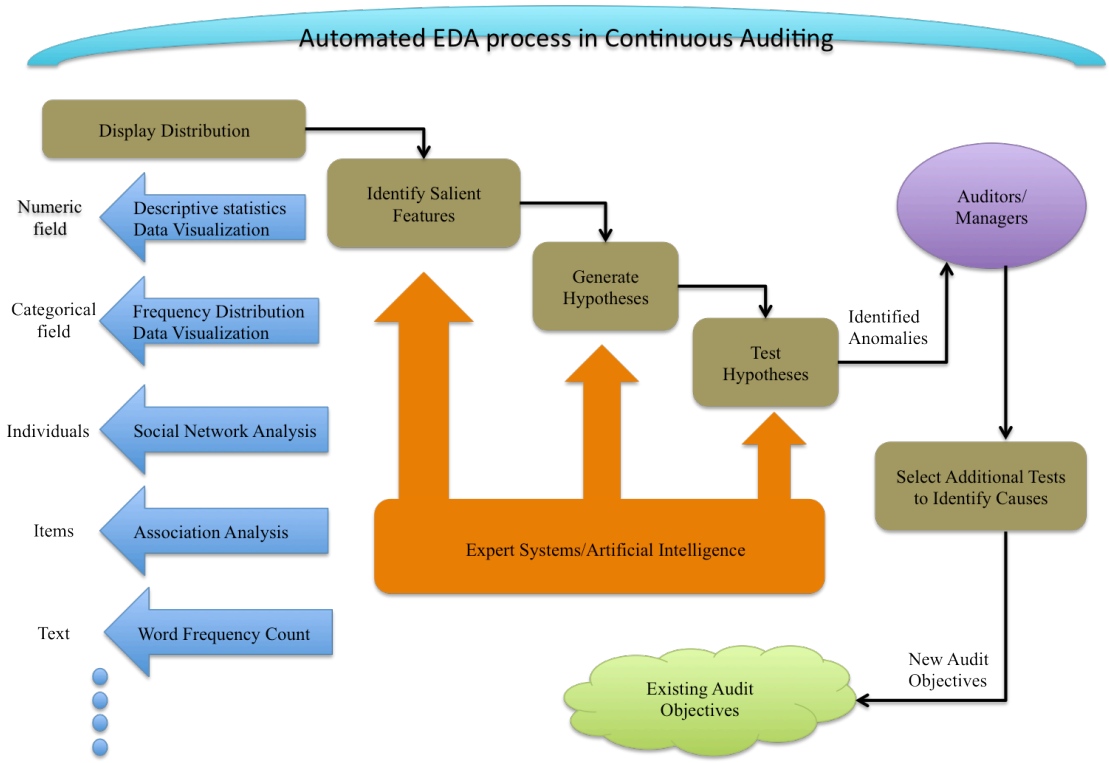


Figure 25: Automated EDA Process in Continuous Audit

2.5 Conclusions

EDA is a data analysis approach that focuses on discovering patterns hidden in the data, which has the potential to assist auditors in identifying emerging risks and opportunities in the organization. Therefore, this chapter explores the application of EDA in the auditing domain. In current accounting literature, some EDA techniques have been studied for specific audit targets, but EDA has not yet been considered as a data analysis approach that should be systematically applied in the audit process. Therefore, this chapter proposes a recommended framework to assist auditors in systematically applying EDA. First, the framework examines internal and external audit standards to determine when auditors can apply EDA in the audit process, then it investigates how various EDA

techniques can be applied to different audit requirements, and finally it suggests specific steps for auditors to follow when implementing EDA.

The application areas identified in this chapter comprise assessing engagement risk, understanding the client's business and assessing the client's business risk, understanding internal control and assessing control risk, assessing fraud risk, performing analytical procedures, reviewing subsequent events, and assessing engagement quality. Eight EDA techniques (descriptive statistics, data visualization, data transformation, feature selection, data mining, text mining, social network analysis, and process mining) have been discussed in this framework. To guide the auditor's use of these techniques, a recommended EDA application process is proposed. This process has eight normative steps: (1) Displaying the distribution of fields related to the audit object; (2) Identifying salient features; (3) Generating hypotheses to explain the salient features; (4) Testing the hypotheses; (5) Identifying suspicious cases; (6) Developing new hypotheses to interpret abnormal cases; (7) Testing new hypotheses; and (8) Adding new audit objectives and reporting the results.

In addition to considering the application of EDA in a traditional, periodic audit setting, this chapter also covers the application of EDA in the continuous auditing environment. The complete application of EDA in continuous auditing includes performing EDA in the continuous auditing system design and development phase and integrating the EDA function into the continuous auditing system as part of the continuous risk monitoring and assessment module. To adapt EDA to the continuous auditing environment, some steps in the EDA application process need to be automated with the help of artificial intelligent techniques, such as expert systems.

References

Alles, M., Kogan, A., Vasarhelyi, M. A., Warren, J. D. (2007). BNA Accounting Policy & Practice Portfolios Portfolio 5405 Continuous Auditing. Accounting Policy & Practice Series. ISSN 1933-0243

Alles, M., Brennan, G., Kogan, A., and 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: 137-161.

Beneish, M. (1999). The Detection of Earnings Manipulation. Financial Analysts Journal, Sep./Oct. 1999: 1-11.

Brockett, P. L., Derrig, R. A., Golden, L. L., Levine, A., Alpert, M. (2002). Fraud Classification Using Principal Component Analysis of RIDITs. The Journal of Risk and Insurance. 69(3): 341-371.

Calderon, T. G., and Cheh, J. J. (2002). A Roadmap for Future Neural Networks Research in Auditing and Risk Assessment. International Journal of Accounting Information Systems, 3(4): 203-236.

Chen, J., Shaw, S.L., Yu, H., Lu F., Chai Y., and Jia Q. (2011). Exploratory data analysis of activity diary data: a space-time GIS approach. Journal of Transport Geography. 19: 394-404.

Cox, K., Eick, S., Wills, G., and Brachman, R.J. (1997). Visual data mining: Recognizing telephone calling fraud. Data Mining and Knowledge Discovery, 1: 225-231.

Debreceny, R.S., and Gray, G. L. (2011). Data Mining of Electronic Mail and Auditing: A Research Agenda. Journal of Information Systems. 25(2): 195-226.

De Mast, J. and Kemper, B. P. H. (2009). Principles of Exploratory Data Analysis in Problem Solving: What Can We Learn from a Well-Known Case? Quality Engineering, 21: 366-375.

De Mast, J. and Trip, A. (2007). Exploratory data analysis in quality improvement projects. Journal of Quality Technology, 39:301–311.

Dilla, W., Janvrin, D. J., and Raschke, R. (2010). Interactive Data Visualization: New Directions for Accounting Information Systems Research. Journal of Information Systems, 24(2): 1-37.

Gluck, M. (2001). Multimedia Exploratory Data Analysis for Geospatial Data Mining: The Case for Augmented Seriation. Journal of the American Society for Information Science and Technology. 52(8): 686-696.

Graedel, T. E., Bertram, M., Kapur, A., Reck, B., and Spatari, S. (2004). Exploratory Data Analysis of the Multilevel Anthropogenic Copper Cycle. Environmental Science & Technology, 38(4): 1253-1261.

Grove, H., and Cook, T. (2004). Lessons for Auditors: Quantitative and Qualitative Red Flags. Journal of Forensic Accounting, 4: 131-146.

Holton, C. (2009). Identifying disgruntled employee systems fraud risk through text mining: a simple solution for a multi-billion dollar problem. Decision Support Systems, 46(4): 853–864.

Jans M. (2009). Internal fraud risk reduction by data mining and process mining: framework and case study (PhD Thesis). Diepenbeek: Hasselt University.

Jans, M., Alles, M., and Vasarhelyi, M. (2013). The case for process mining in auditing: Sources of value added and areas of application. International Journal of Accounting Information Systmes, 14: 1-20.

Jans, M., Alles, M., and Vasarhelyi, M. (2014). A field study on the use of process mining of event logs as an analytical procedure in auditing. The Accounting Review. Forthcoming.

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

Kinney, W. R. (1978). ARIMA and Regression in Analytical Review: an Empirical Test. The Accounting Review, 53(1): 48-60.

Kirkos, E., Spathis, C., and Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications 32(4): 995-1003.

Kogan, A., Alles, M. G., Vasarhelyi, M. A., and Wu, J. (2014). Design and Evaluation of a Continuous Data Level Auditing System. Auditing: A Journal of Practice & Theory. Forthcoming.

Kouka, E. F. M. and Saucier, G. (1987). An application of exploratory data analysis techniques to floorplan design. 24th ACM/IEEE Design Automation Conference, pp 654658.

Koschat, A. M., and Sabavala, D. J. (1994). The effects of television advertising on local telephone usage: exploratory data analysis and response modeling. Marketing Science, 13(4): 374-391.

Koskivaara, E. (2004). Artificial neural networks in analytical review procedures, Managerial Auditing Journal, 19(2): 191– 223.

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.

Lombardo, R., and Valle, E. D. (2011). Data Mining and Exploratory Data Analysis for the Evaluation of Job Satisfaction. iBusiness, 3: 372-382.

Nakaya, T. and Yano, K. (2010). Visualising Crime Clusters in a Space-time Cube: An Exploratory Data-analysis Approach Using Space-time Kernel Density Estimation and Scan Statistics. Transactions in GIS, 14(3): 223-239.

Palshikar, G. K. (2002). The Hidden Truth. Feature. May 28, 2002.

Phua, C., Lee, V., Smith, K., and Gayler, R. A comprehensive survey of data mining-based fraud detection research, Working Paper, http://arxiv.org/abs/1009.6119, 2010.

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

Rezaee, Z., Sharbatoghlie, A., Elam, R., and McMickle, P. L. (2002). Continuous Auditing: Building Automated Auditing Capability. Auditing: A Journal of Practice & Theory. 21(1): 147-163.

Shaw, S. L., and Xin, X. (2003). Integrated land use and transportation interaction: a temporal GIS exploratory data analysis approach. Journal of Transport Geography 11: 103-115.

Sokol, L., Garcia, B., Rodriguez, J., West, M., and Johnson, K. (2001). Using data mining to find fraud in HCFA health care claims, Topics in Health Information Management 22(1): 1–13.

Stringer, K. W. (1975). A Statistical Technique for Analytical Review. Journal of Accounting Research, 13(3): 1-9.

Thiprungsri, S. and Vasarhelyi, M. (2011). Cluster Analysis for anomaly detection in

accounting data: an audit approach. The International Journal of Digital Accounting Research, 11: 69-84.

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

Vasarhelyi, M. A. (2011). The Coming Age of Continuous Assurance. Insights. Melbourne Business and Economics, April 2011. 23-30.

Wesley, S., Lehew, M., and Woodside, A. G. (2006). Consumer decision-making styles and mall shopping behavior: Building theory using exploratory data analysis and the comparative method. Journal of Business Research, 59: 535-548.

Yang, W., and Hwang, S. (2006). A process-mining framework for the detection of healthcare fraud and abuse. Expert Systems with Applications, 31(1): 56–68.

Zaslavsky, V., and Strizhak, A. (2006). Credit card fraud detection using self-organizing maps. Information & Security, 18: 48–63.

Chapter 3 An Application of Exploratory Data Analysis in Auditing -- Credit Card Retention Case

3.1 Introduction

Over the last decades, operational risks in the banking system attracted both regulatory and academic attention due to the devastating losses experienced by 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 over fraudulent sales practices over 13 years[2] (Muermann and Oktem, 2002). The 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 realize their strategies and objectives. This chapter demonstrates an application of EDA in a real operational audit setting to support the conceptual framework proposed in the previous chapters and illustrate how internal auditors can benefit from this approach.

This case study follows the EDA application steps proposed in Chapter 2, section 2.3.3 **Process** to analyze a credit card annual fee discount dataset from an international bank in Brazil. In this case study, the EDA process is mainly applied in the "Perform Audit Plan" stage (in Figure 22) where three specific audit objectives have already been developed. The results of the EDA process are compared with the results of conventional audit procedures. The comparison outcomes demonstrate that, following the EDA,

[1] http://online.wsj.com/news/articles/SB10129910421902036 40
[2] http://caselaw.findlaw.com/us-3rd-circuit/1362355.html

comprehensive findings can easily be obtained even with simple statistics and visualization techniques.

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

3.2 The Audit Problem

3.2.1 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 don't want to pay these fees may call the bank asking for cancelation or a fee reduction. In these circumstances, bank representatives negotiate with the clients about the fees. Finally, based on clients' backgrounds, representatives can offer appropriate discounts. During the discount negotiation process, bank representatives should follow the bank policy; they cannot offer discounts higher than their authority. And within their jurisdiction, they should also give top priority to the benefit of the bank. In other words, they should offer the lowest discounts acceptable to the clients.

3.2.2 Audit Objectives

The initial audit scope suggested by the bank is to identify the bank representatives whose behavior in course of the annual fee negotiation 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 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.

In order to achieve this audit objective, all related fields need to be thoroughly explored for irregularities, making it a suitable scenario for EDA. Auditors will gain understanding about the process, identifying risks and problems within this process and their internal control system.

3.3 Methodology

3.3.1 Data

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

The account master data is a large dataset 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 dataset. The fields in 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 limit and late pay amount. 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 account, and number of cards. Most of them are necessary to test original audit objectives, such as call length, annual fee, and output annual fee; while some of them are new attributes added during EDA process, for example, supervisor number and number of cards. The names, source database, and descriptions of these attributes are listed in Table 6.

Table 6: 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 & 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 clients
Number of Cards (Account Master)	Number of cards associated with each account

Among these fields, call length, original fees, actual fees, and number of cards are continuous variables. Representatives ID, supervisors ID, clients ID, account sequential number, customer service center location are nominal variables. In order to protect clients' privacy, account sequential number and clients ID are encrypted in the dataset. The encryption method preserves the integrity of the original data; each original value corresponds to a unique cipher text.

3.3.2 Data Preprocess

Discounts offered by bank representatives play an important role in the process of loss of revenue analysis. However, in the raw retention data, there is not a field directly reflecting discount. Two existing fields related to discount are original fees and actual fees (fees after negotiation). The difference represents the discount, and this figure must be calculated before conducting EDA. Specifically, discount is the difference between original fees and actual fees divided by the original fees. The formula used to calculate discounts is as follows.

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

Certain analyses (e.g. credit card quantity check on page 91), 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 dataset. These many-to-one datasets can be joined based on this relationship. The joining process uses the account sequential number field as it exists in both datasets and is the unique identifier in the VBA data.

3.3.3 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.

3.4 Results and Discussion

3.4.1 Policy violating bank representatives and negative discounts

3.4.1.1 Conventional Audit Procedures

In order to determine whether bank representatives are violating bank policy, the maximum discount each bank representative 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 there were any bank representatives offering more than 100% discounts. Internal auditors can perform this test by simply applying a filter to select all the records with discounts greater than 100%. This filter returned no records, indicating that there was no bank representative violating bank policy. Hereto, this audit objective is confirmed by conventional audit procedure. Auditors can check this box on their checklist and move to the next one.

3.4.1.2 EDA process

Display Distributions

Following the EDA process shown in Figure 23, the first step is to display the distribution of related fields. Since 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 7.

Table 7: Descriptive Statistics of Discounts

Field Name	Mean	Median	Minimum	Maximum	Standard deviation
Discount	-2.326.04%	60%	-27,944,522.22%	100.00%	219933.88%

Identify Salient Features

According to the 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 representatives violate bank policy.

Besides this, a salient feature can be observe from this table is that the minimum discount is a large negative value (-27,944,522.22). The mean is also negative (-2326.04), which means 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 Figure 26) also reveals that only 0.15% (286) discounts are negative.

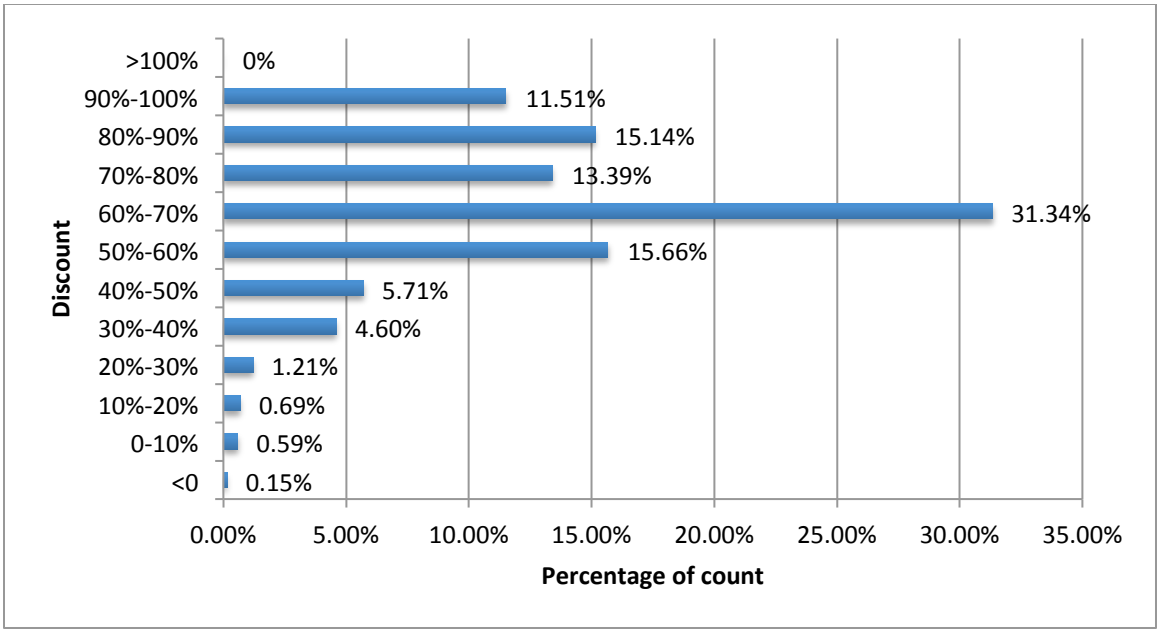


Figure 26: Frequency Distribution of Discounts

According to the formula for discount, 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.

Generate Hypothesis

After discussion, the bank's internal auditors yielded a potential explanation: in some cases, a group of people (e.g. a family) have 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. Since the actual fee for all cards may surpass the original fee for one card, we hypothesized that negative discounts are due to group discounts offered to clients with more than one credit cards.

Test hypothesis and Identify Suspicious Cases

To gain insight into negative discounts, the frequency distribution of negative discounts is calculated and displayed in a line chart (Figure 27).

Figure 27: Distribution of Negative Discounts

Figure 27 demonstrates a multimodal and discontinuous distribution of negative discounts featuring three separate clusters. The first contains 27 (10%) of the records with extreme discounts (lower than -5,000,000%). The second includes 69 (24%) data points associated with relatively significant discounts (between -10,000% and -100%). The third and largest cluster involves 190 records having small discounts (less than -50%). The negative discounts in this cluster may due to group discounts. However, this explanation cannot be applicable to the negative discounts in the other two clusters because of their exceptional 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. One easy verification for these data points is to determine

whether a given client has multiple cards. The results of this verification are shown in Figure 28.

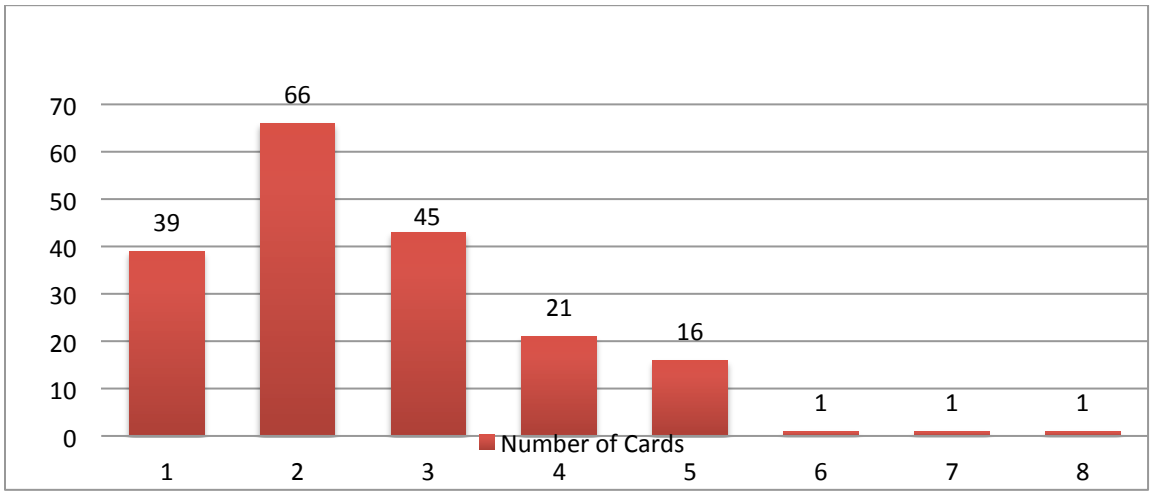


Figure 28: Frequency Distribution of Number of Cards of the 190 Cases with Reasonable Negative Discounts

According to Figure 28, there are 39 (20.5%) out of 190 clients having only one credit card, who should not be given group discounts. Therefore, these 39 accounts are also considered suspicious.

Investigate Causes of Suspicious Cases to Create New Hypothesis

Since original fees and actual fees are the two determinant factors in calculating discounts, the relationship between negative discounts and these two figures are examined in order to investigate the cause of this distribution. As the ranges of the variables are very wide, to display the data satisfactorily, the values need to be transformed to another scale. Specifically, the values of original and actual fees are transformed to their logarithmic value. Due to the negative values, the logarithmic value of the absolute value of negative discounts are calculated. Then scatter plots are used to

display the relationship between discount and actual and original fees (shown in Figure 29).

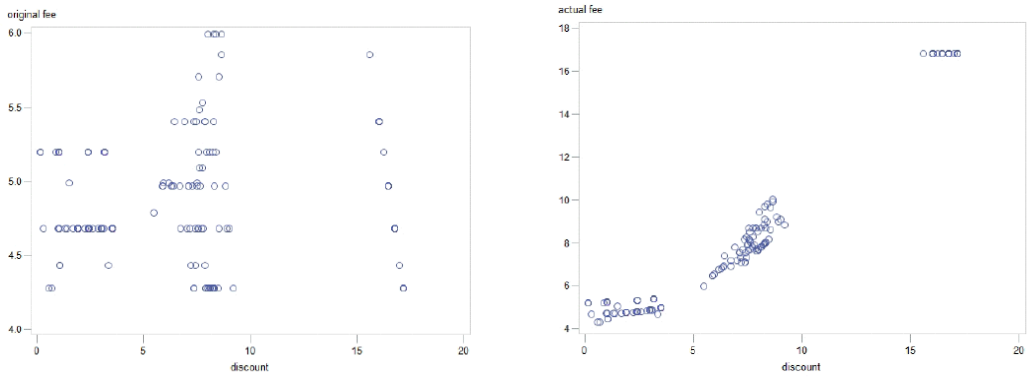


Figure 29: Relationships Between Negative Discounts and Original and Actual Fees

Figure 29 reveals that three clusters negative discounts evenly distribute among original fees. While, the same three clusters can also be observed in the scatter plot of discount and actual fees. Hence, the new hypothesis is that these large negative discounts are caused by irregular actual fees.

Test New Hypothesis

Since the number of extreme negative discounts is manageable, a substantive test is performed to investigate the specific reason of these extreme negative discounts. Among these 96 cases, 27 negative discounts are due to obvious input error (dates are mistakenly inputted as the actual fees). The other 69 negative discounts are caused by round unreasonable large actual fees. These records may also include input error such as incorrect placing of a decimal point.

Develop New Audit Objectives and Report Results

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 our analysis were reported to the bank, and a new audit objective was developed: actual fees were recorded correctly.

The 39 suspicious cases with reasonable negative discounts were reported to internal auditors for their further investigation, limiting our ability to learn more. We did, however, create a new audit objective: negative discounts have been offered to clients with multiple cards.

In summary, after performing EDA process in testing this audit objective, 135 abnormal cases are identified, while no anomaly can be identified using conventional audit procedure. In addition to these exceptional cases, two new audit objectives are generated and two new internal control functions are suggested.

3.4.2 Lazy bank representatives and inactive representatives

3.4.2.1 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% (hereafter "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 Figure 30. 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 lazy, they can identify 59 such representatives (with ratio greater than 0.5).

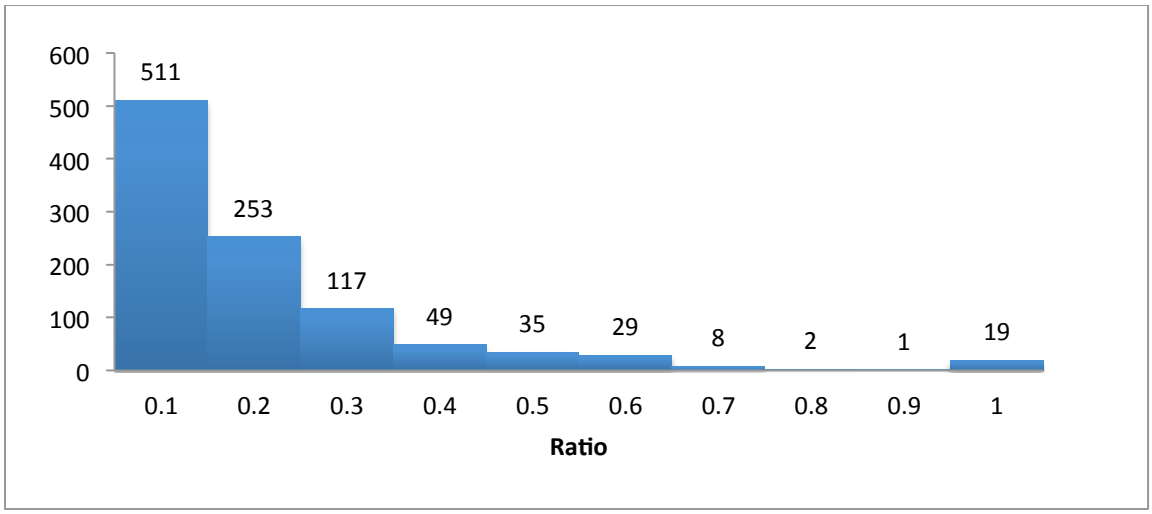


Figure 30: Frequency Distribution of the Ratio of 100% Discounts to All Discounts Offered by Each Bank Representative

3.4.2.2 EDA process

Display Distribution

In the EDA process, the representatives who offered 100% discounts are first identified since they are the main concern of this audit objective. Among all 1151 representatives, 1024 representatives offered 100% discount at least once. A comparison of these representatives' appearances in the full retention dataset and the 100% discount subset is shown in Figure 31.

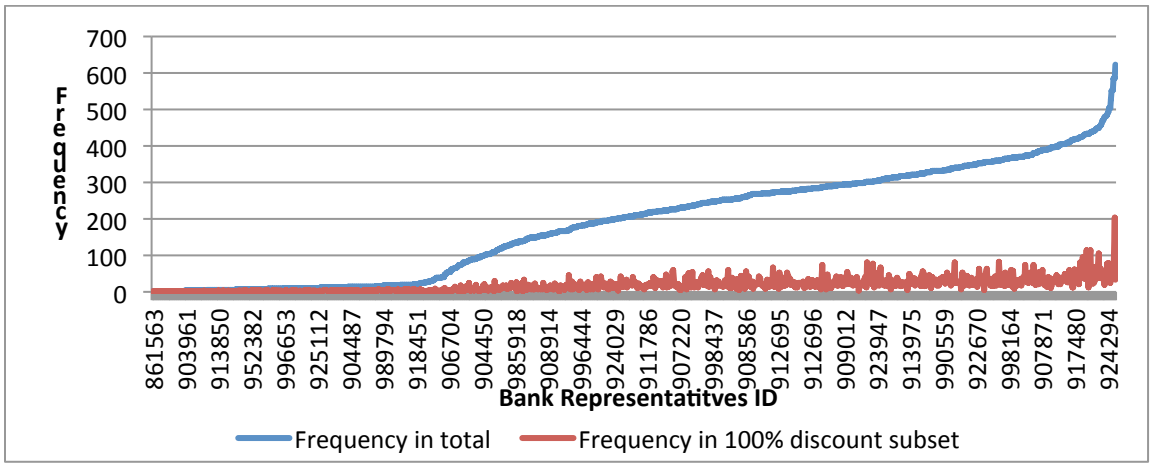


Figure 31: Distribution of Bank Representatives Offered 100% Discounts in the Whole Retention Data and the 100% Discount Subset

Identify Salient Features

Figure 31 demonstrates that, generally speaking, 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' frequencies very close to zero. It is illogical 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 8.

Table 8: Descriptive Statistics of Frequency Distribution of Bank Representatives

Mean	Standard Deviation	Minimum Value	Maximum Value	Count
170	148	1	623	1151

Table 8 reveals that the average number of calls answered by the 1151 bank representatives is 170. 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 and Cook, 1983). Therefore, the 403 representatives answering 22 or fewer calls (170-148) are considered suspicious.

Generate Hypothesis

According to the bank, a potential explanation of these representatives is that they are supervisors. It is reasonable for the supervisors to 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.

Test Hypothesis and Identify Suspicious Cases

After comparing with these 403 representatives' IDs with supervisors' IDs, 33 of them are confirmed as supervisors. This leaves 370 still-suspicious representatives. Therefore, in all the 1151 bank representatives, 748 (65%) of them are active and 403 (35%) are inactive, among which 33 (3%) are supervisors and 370 (32%) are suspicious inactive representatives. The distribution of bank representatives is shown in Figure 32: *Distribution of Bank Representatives*Figure 32.

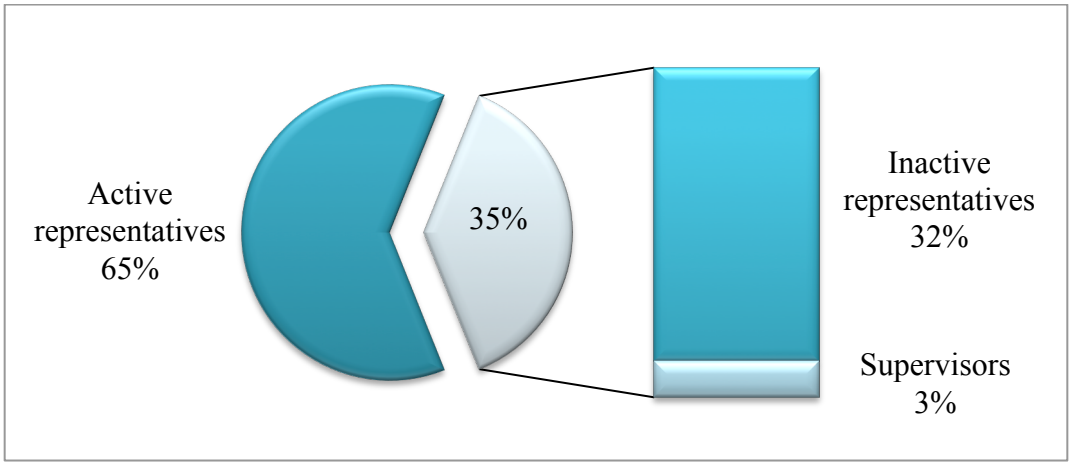


Figure 32: Distribution of Bank Representatives

Investigate Causes of Suspicious Cases to Create New Hypothesis

To locate the cause of this issue, the distributions of inactive and active representatives in different customer service centers are compared in Figure 33. It reveals that 85.95% of 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 internal auditors in the bank, they suggested another potential explanation of 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. Thereby, the hypothesis created in this step is that non-supervisory inactive bank representatives are interns.

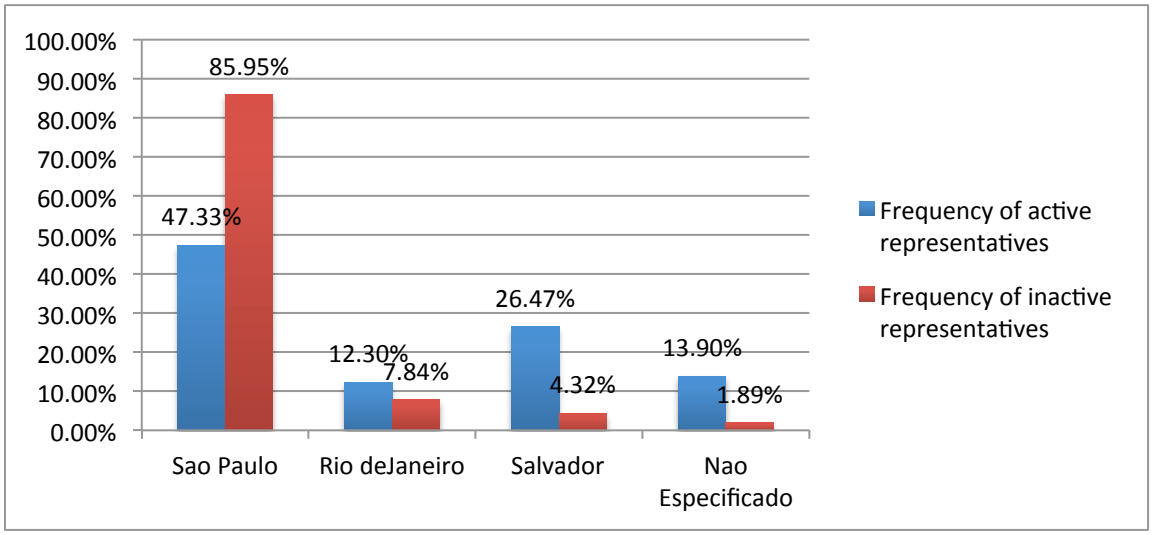


Figure 33: Distributions of Inactive and Active Representatives in Different Customer Service Centers

Test New Hypothesis

Because of limited access to other supporting audit evidence, test of this hypothesis cannot be performed, but it can enhance internal auditors' analysis.

Develop New Audit Objectives and Report Results

Confirmation of the new hypothesis yields a new audit objective: all non-supervisory permanent bank representatives were active bank representatives.

370 inactive bank representatives are identified through the EDA process. These representatives are all identified as suspicious in conventional audit procedures. All bank representatives with a 100% discount ratio greater than 0.35 are actually inactive bank representatives. Therefore, compared with conventional audit procedures, EDA allows internal auditors to obtain a more comprehensive set of anomalies. In addition, EDA discovers that a potential cause of these inactive representatives is related to customer service center location.

3.4.3 Non-negotiation bank representatives and short calls

3.4.3.1 Conventional audit procedures

The third representative of interest doesn't negotiate with clients, instead immediately offering a discount. Since these calls should have relatively short duration, internal auditors can sort the call duration field to find unreasonably short calls (e.g. calls shorter than 60 seconds). This conventional audit procedure has identified 28,027 unreasonably short calls fielded by 933 bank representatives.

3.4.3.2 EDA Process

Display Distribution

In the EDA process, some descriptive statistics of call duration field are calculated to display its distribution. The results are shown in Table 9. 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 while 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 less than 600 seconds (shown in Figure 34).

Table 9: Descriptive Statistics of Call Duration

Minimum	Maximum	Mean	Median	90th Percent	Count
10	6561	255	206	514	195694

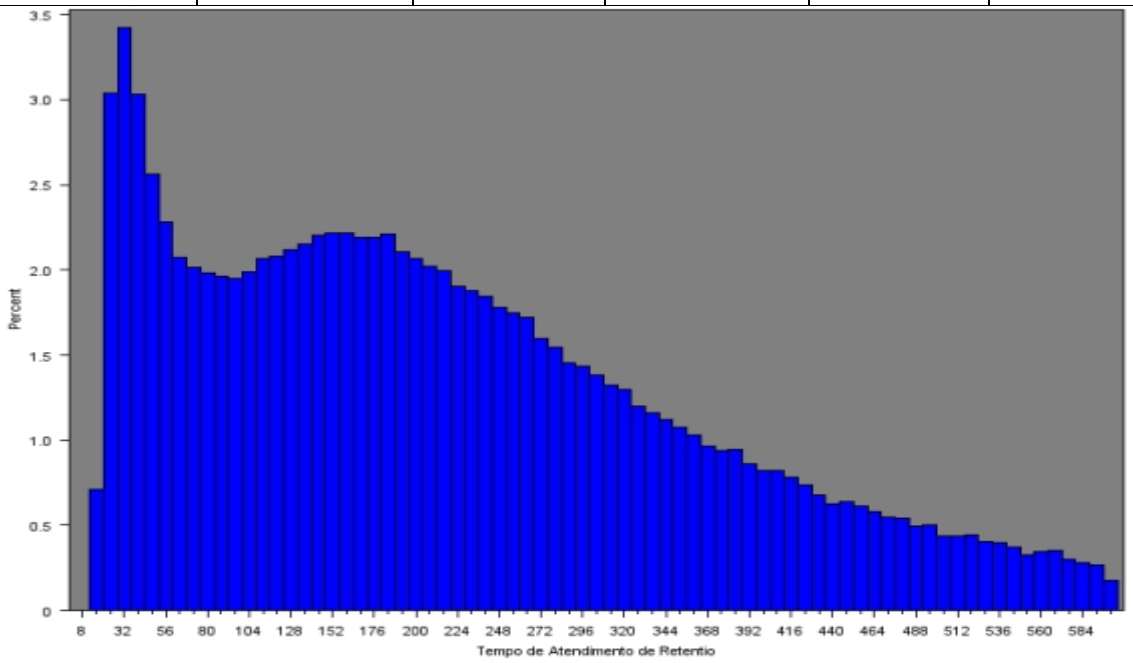


Figure 34: Frequency Distribution of Call Duration Less Than 600 Seconds

Identify Salient Features

From this distribution two peaks are observed: one is between 2 to 3 minutes, and the other one is between 20 to 60 seconds. It is reasonable for a customer to negotiate 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 to 60 seconds.

Generate Hypothesis

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.

Test Hypothesis and Identify Suspicious Cases

Since in the retention database each call is related to a discount, no discount should be associated with these calls. In the retention dataset, there are 28,027 calls under 60 seconds each. Only 121 short calls resulted in no discount; the other 27,906 short phone calls have a nonzero discount and are therefore considered suspicious.

Investigate Causes of Suspicious Cases to Create New Hypotheses & Test New Hypotheses & Develop New Audit Objectives and Report Results

As was true with the previous audit objective, it is not possible for us to directly identify the causes of these suspicious cases due to limited data access. These findings are therefore 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 the effective phone calls lasted more than 1 minute.

The findings from EDA are generally consistent with those from conventional audit procedures. Therefore, besides 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 standalone examination.

3.5 Conclusion

Using real datasets from an international bank in Brazil, this chapter provides an example on how EDA can be applied by internal auditors 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 EDA process. By comparing the two sets of results it demonstrates the incremental contribution of EDA process in uncovering risk areas that cannot be detected by conventional audit procedures.

Specifically, the datasets consists of information related to phone calls made by clients intending to negotiate their credit card annual fees. The original audit objectives are to identify bank representatives who may cause loss of 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 more 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 effective phone calls lasted less than 1 minute

Based on these findings four new audit objectives can be developed and added to the existing audit objectives to continuous 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 1 minute

One limitation of this field study is that 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. In addition, this field study only applied three traditional EDA techniques. More advanced EDA techniques are demonstrated in Chapter 4. Another limitation of this case study is its potential lack of generalizability, but this is an issue only for results specific to the bank and not for the general conclusion that EDA can identify abnormal cases and risk areas that conventional analytical procedures cannot.

References

Beckman, R. J., and 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

Muermann, A., and Oktem, U. (2002) The Near-Miss Management of Operational Risk. The Journal of Risk Finance. 4 (1):25-36

Chapter 4 An application in Healthcare Fraud Detection

4.1 Introduction

Healthcare has become a major expenditure in the US since 1980. Both the size of the healthcare sector and the enormous volume of money involved make it an attractive fraud target. According to the Office of Management and Budget, about 9.5%, or around $47.8 billion of the US's Medicare expenditure was lost to fraud in 2013[1]. Therefore, effective fraud detection is important for reducing the cost of U.S. healthcare system.

Detecting healthcare fraud and abuse, however, requires intensive medical knowledge. Many health insurance systems rely on human experts to manually review insurance claims and identify suspicious ones. This results in both time-consuming processes of system development and claim review, especially for large insurance programs.

Automated claims processing are being increasingly implemented to address these costs. These systems identify areas requiring special attention such as erroneous or incomplete data input, duplicate claims, and non-covered services. Although these systems may be used to detect certain types of fraud, their fraud detection capabilities are usually limited since the detection mainly relies on pre-defined simple rules specified by domain experts (Li et al., 2007).

---

[1] http://www.healthcarepayernews.com/content/medicare-medicaid-error-estimates-grew-2013#.U6hF2xbhzBE

Numerous researchers have attempted to increase these systems' effectiveness by developing more sophisticated antifraud approaches incorporating data mining, machine learning, or other methods (Chan and Lan, 2001; He et al., 1997, 2009; Liou et al., 2008; Major and Riedinger, 2002; Musal, 2010; Otega et al., 2006, Sokol et al., 2001; Viveros et al., 1996; Williams and Huang, 1997; Yamanishi et al, 2004). Advanced EDA techniques such as clustering (Musal, 2010) and process mining (Yang, 2006) were applied in some of the research. Compared with existing fraud detection systems, these approaches focus on more complex tasks such as automatic learning of fraud patterns, prioritizing suspicious cases by assigning "fraud likelihoods," and identifying new types of fraud.

However, these studies are essentially technical in character, focusing on the development of various methods to identify different types of fraudulent behaviors in healthcare practice from a computer science or information technology perspective. Few discuss the problem from an accounting or auditing point of view. Since both internal and external auditors's responsibilities include fraud detection, integration of these healthcare fraud detection techniques into the audit process, and the associated benefits, are worth studying. This chapter attempts to fill this gap and demonstrate how auditors can take advantage of advanced EDA techniques to assess healthcare fraud risk by following the conceptual EDA application process proposed in **Error! Reference source not found.**. Specifically, a real Medicare inpatient claim dataset purchased from Center for Medicare and Medicaid Services (CMS) is used in this case study. Two advanced EDA techniques – cluster analysis and association analysis – are applied in addition to traditional EDA methods. Whereas Chapter 3's application focused on the "Perform Audit Plan" stage,

this study uses EDA in the "Develop Audit Plan" stage (shown in **Error! Reference source not found.**) to identify high risk cases. As in the previous chapter, EDA results are compared with conventional audit results to determine additional benefit.

This chapter begins with an introduction of US healthcare system and its fraud behavior; followed by a description of the methodology including the dataset used in the analysis, the analysis process, and the algorithms for cluster analysis and association analysis. Results are then presented and discussed, followed by implications and limitations of the study.

4.2 Background of US Healthcare System and its Fraud Behavior

The public healthcare system in US consists of two major programs: Medicare and Medicaid services. Medicare is a social insurance program administered by the United States government that provides health insurance coverage to (1) people age 65 or older, (2) people under 65 with certain disabilities, and (3) people of all ages with end-stage renal disease, i.e., permanent kidney failure requiring dialysis or a kidney transplant. Medicare provides three types of services: hospital insurance (part A), medical insurance (part B) and prescription drug coverage (parts C and D). Medicaid is a state-administered program and each state sets its own guidelines regarding eligibility and services, but generally speaking, it is available only to low-income individuals and families as determined by federal and state law.

Both Medicare and Medicaid involve three parties: (1) service providers, including doctors, hospitals, ambulance companies, and laboratories; (2) insurance subscribers, who are beneficiaries of Medicare and Medicaid services; (3) insurance carriers, who pay healthcare costs on behalf of their subscribers, including federal and

state governmental health departments. According to which party commits the fraud, healthcare fraud behaviors can be classified as follows (NHCAA, 2005):

- Service provider fraud:
  (a) Billing services that are not actually performed;
  (b) Unbundling, i.e., billing each stage of a procedure as if it were a separate treatment;
  (c) Upcoding, i.e., billing more costly services than those actually performed;
  (d) Performing medically unnecessary services solely for the purpose of generating insurance payments;
  (e) Misrepresenting non-covered treatments as medically necessary for the purpose of obtaining insurance payments;
  (f) Falsifying patients' diagnoses and/or treatment histories to justify tests, surgeries, or other procedures that are not medically necessary.
- Insurance subscriber fraud:
  (a) Falsifying records of employment/eligibility for qualifying for Medicare and Medicaid services;
  (b) Filing claims for medical services which are not actually received;
  (c) Using other persons' coverage or insurance card to illegally claim insurance benefits.
- Insurance carrier fraud:
  (a) Falsifying reimbursements;

(b) Falsifying benefit/service statements.

- Conspiracy fraud: fraud involving more than one party, e.g., a patient colluding with his physician, fabricating medical service and transition records to deceive governmental health departments for illegitimate Medicare/Medicaid reimbursements

Among the four healthcare service participants, the service provider has the most opportunities to perpetrate healthcare fraud. Since service provider fraud can cause great damage to the healthcare system (NHCAA, 2005), it attracts large amount of research efforts. In current literature, about 69% of studies have been devoted to detecting service provider fraud, while the research efforts on the other three types of fraud are limited (31% for insurance subscribers' fraud and 0% for insurance carriers' and conspiracy fraud; Li et al., 2007).

4.3 Methodology

4.3.1 Healthcare Data

In current literature, data for healthcare fraud detection come mostly from insurance carriers, including governmental health departments and private insurance companies. Major governmental health departments that have been reported in the literature include the Bureau of National Health Insurance (NHI) in Taiwan (Chan and Lan, 2001; Yang and Hwang, 2006; Liou et al., 2008), and the Health Insurance Commission (HIC) in Australia (Viveros et al., 1996, He et al., 1997, 2000; Williams and Huang, 1999; Yamanishi et al., 2004). Healthcare data from private insurance companies have also been used by several researchers (Major and Riedinger, 2002; Ortega et al., 2006).

Whatever the source, the most used data in healthcare fraud detection are insurance claims. An insurance claim involves the participation of an insurance subscriber and a service provider. The claim data have two characteristics. First, they contain a rich amount of attributes to describe the behaviors of the involved service providers and insurance subscribers, allowing for detection of the types of fraud committed by these two parties. Second, each claim usually contains unique identifiers for the involved service provider and insurance subscriber, respectively. Using these identifiers enables a global view of a service provider's behaviors over time and across different subscribers, and a similar view of a subscriber's behavior over time and across different service providers. These perspectives are important when identifying provider and subscriber fraud.

Practitioner data can also be used in fraud detection (Viveros et al., 1996) by providing a general description of service providers in a certain time period. This data include personal information on providers as well as measures of services such as the cost, usage, and quality. Practitioners data is usually used with insurance claim data in supervised fraud detection methods to provide a description of the practice and to identify the selection and frequency of tests.

Most claim and practitioner data used in current literature relates to outpatient services from individual service providers. Inpatient claims are rarely investigated in extant research. In addition, even though the U.S. healthcare system is suffering great losses from fraud, U.S. Medicare and Medicaid data are seldom used in academic research. To fill in these gaps, this study utilizes Medicare inpatient claim data to investigate fraud in hospital services.

The data used in this study is purchased from the Center for Medicare and Medicaid Services[2]. It includes all the Medicare inpatient claims in 2010, consisting of 12,453,186 records and 1627 fields. These fields include insurance subscriber information (age, sex, medical status code, etc.), insurance provider information (Provider number, providers' state, etc.), physicians' information (claim operating physician number, claim attending physician number), diagnosis information (diagnosis code count, etc.), payment/payer information (claim payment amount, payer code, etc.), and claim information (claim total charge amount, claim diagnosis code count, claim admission date, etc.). The data is anonymized for privacy purposes; a random identification number replaces all identifying information, such as name, address, and ZIP code.

Among these 1627 fields, 1181 fields are left blank for all the records, either because the fields have not been filled in or because the attributes are not used. For the other 446 attributes, 134 have more than 50% missing values and 55 attributes have a single value. These attributes are excluded from analysis. A summary of this basic information is presented in Figure 35. This leaves 257 attributes, which are listed in Appendix C: Usable fields in 2010 Inpatient Medicare Claim Data. The fields included in the

All Attributes	1627
Less:	
Blank Attributes	-1181
Filled Attributes with more than 50% missing values	-134
Attributes with Single Values	-55
Remaining Attributes	**257**

[2] http://www.cms.gov/

following analysis are selected from this list.

Figure 35: Pre-Analysis Attribute Filtering

4.3.2 Analysis Process

4.3.2.1 Conventional audit procedures

To gain a basic understanding of the conventional audit procedures auditors usually perform to analyze healthcare claims, domain experts (internal auditors of New Jersey Medicaid Program[3]) are consulted. The auditors' major concern is the payment: They want to assure that the payments paid to diagnose and treat beneficiaries' diseases are not excessive. The field directly related to claim payment is claim payment amount. Auditors consider payments that are high relative to other payments for a similar procedure to be high-risk. Specifically, the threshold to distinguish extreme large payment amount and normal payment amount is defined as the mean plus three standard deviations. Therefore, in the conventional audit procedure, the mean, standard deviation, and threshold of each diagnosis' payment amounts are calculated. Then each claim's payment amount is compared with the threshold corresponding to its diagnosis. The claims with payment amounts exceeding threshold are considered as high-risk claims.

[3] Because of geographical restriction, it is impossible for us to work with internal auditors in the Medicare program. We therefore coordinated with internal auditors in the New Jersey Medicaid program. As Medicaid and Medicare programs are similar in terms of fraudulent behavior and data types, the conventional audit procedures performed by internal auditors in New Jersey Medicaid program are also applicable for similar Medicare data.

Payment can also be used with other risk indicators (such as the ones recommended by ValueOptions[4]) to prioritize the suspicious cases that need further investigation. For example, as service providers are the most suspect party in healthcare service participants, service providers obtaining extreme large payment amounts from Medicare or filing large number of claims are considered potential fraud perpetrators. High-risk claims associated with these service providers should be given higher priority for further investigation. Similarly, the definition of these exceptional service providers can based on the same statistical measure used to determine high-risk claims.

4.3.2.2 EDA Process

4.3.2.2.1 Attributes Selection

This case study follows the EDA process proposed in Chapter 2, section 2.3.3 **Process**. In this case, there are 257 useable fields in the dataset. Budget and time constraints make it impossible for auditors to examine all of them. Therefore, fields that can be used to assess Medicare fraud risk must be identified and included in the EDA process before proceeding. According to the fraudulent behaviors presented in 4.2 Background of US Healthcare System and its Fraud Behavior, factors directly relating to Medicare fraud detection include service provider identification, claim payment amount, diagnosis and treatment of each claim, beneficiary identification, number of days in hospital, and distance between beneficiary residence and hospital location.

Among these factors, service provider identification (PRVDR_NUM) and beneficiary identification (DESY_SORT_KEY) are intuitively related to Medicare fraud

[4] Downloadable at: http://www.valueoptions.com/providers/Compliance/FraudandAbuse.pdf

detection because they can be used to create individual provider and beneficiary profiles. In addition to these two IDs, claim number (CLAIM_NO), the unique identifier of Medicare claims, is also added to this analysis to support some data manipulation activities, such as analysis results comparison. As is true for conventional audit procedures, claim payment amount (CLM_PMT_AMT) can serve as a measurement to evaluate the appropriateness of claim payment and is therefore included in the EDA process as well.

Several other fields were used to supplement the more obviously relevant ones listed above. The number of covered days a beneficiary stayed in hospital (CLM_UTLZTN_DAY_CNT) can be used to assess the necessity of Medicare services (Blanchard, 2007). A long hospital stay should result only from serious disease; otherwise a red flag should be created. Apart from this, distance between beneficiary's residence and hospital location can also be considered as a Medicare fraud indicator. Medicare beneficiaries, either senior, disabled or seriously ill, would likely prefer a provider within a relatively short distance. Longer distances between a beneficiary's residence and a service provider's location may be indicate fraudulent behaviors such as beneficiary identification theft, billing of nonexistent medical services, or collusive fraud involving both service providers and beneficiaries. A benign alternative explanation for the long travel distance is that no local provider can treat the beneficiary's condition, necessitating a longer trip. This situation is usually associated with serious or rare conditions, usually requiring longer hospital stays and larger payment amounts. Therefore, the number of days beneficiaries stayed in hospital and the distance between

beneficiaries' residences and hospital locations can be used together with payment amounts to identify potential fraudulent cases.

Because the distances between residences and hospital locations do not exist in the original Medicare claim dataset, they need to be calculated before performing the analysis. As discussed in the last section, personally identifying information, including addresses of both subscribers and providers, are eliminated from the database. Only beneficiaries' living states (BENE_RSDNC_SSA_STD_STATE_CD) and counties (BENE_RSDNC_SSA_STD_CNTY_CD) and service providers' locating states (NCH_PRVDR_STATE_CD) are available. The latitude and longitude of the center of each county and state are collected from the US census website and mapped to the Medicare dataset according to the SSA code[5] of each county and state, yielding estimated locations. The distances between beneficiaries and service providers are estimated by the following expression, which based on the formula to calculate the distance between two points on a sphere.

Distance=3958.758349716768 * ARCOS(SIN(Beneficiary's latitude / 57.2958) * SIN(Provider's latitude/57.2958) + COS(Beneficiary's latitude/57.2958) * COS(Provider's latitude/57.2958) * COS(Beneficiary's longitude/57.2958 – Provider's longitude/57.2958))

Besides payment, hospital stay, and distance related information, diagnoses and procedures information in the original dataset can also be utilized to test the

[5] SSA code is a coding system used by U.S. social security administration. In this coding system each state is represented by a two-digit code, and each county is represented by a five-digit code with the first two digits indicating the state of the county.

reasonableness of Medicare claims. For example, they can be used to test whether irrelevant procedures are billed for certain diagnoses. In the raw Medicare claim dataset, one claim can be associated with up to 10 diagnoses and 6 procedures. Each of these diagnoses and procedures are stored in a separate field, for 16 total (CLM_DGNS_CD{1-10} and CLM_PRCDR_CD{1-6}). Thus, 21 original fields and 1 derived attribute, listed in Table 10, are finally selected for EDA.

Table 10: Attributes Selected in EDA Process

Attribute Name	Description
CLAIM_NO	Claim number
DESY_SORT_KEY	Beneficiary identifier
CLM_PMT_AMT	Claim payment amount
PRVDR_NUM	Provider number
CLM_UTLZTN_DAY_CNT	Claim utilization day count
Distance (derived)	Distance between beneficiary's residence county and provider's state
CLM_DGNS_CD{1-10}	Claim diagnosis code
CLM_PRCDR_CD{1-6}	Claim procedure code

4.3.2.2.2 Cluster Analysis

Descriptive statistics and data visualization are first performed to display the distribution of the selected fields and help identify salient features. After that, potential explanations are generated and tested based on the available data. However, since payment amounts, number of days in hospital, and distance between beneficiaries and service providers are three interrelated fields, their relationships might be ignored in separate analyses. Therefore, a multivariate cluster analysis (introduced in **Error!**

Reference source not found.) is performed to provide a more comprehensive distribution of these three attributes and their relationships, ideally identifying additional cases that would have gone undetected during individual analysis.

Because payment amount, hospital stay period, and distance between beneficiaries and service providers are numeric attributes, simple K-means clustering algorithm is used for cluster analysis. K-means is a simple, well-known algorithm for clustering. It is less computationally intensive than many other algorithms, making it a preferable choice for large datasets (Tan et al, 2006). The steps in the K-means algorithm can be explained as follows (Roiger et al, 2003):

1. Choose a value for K, the total number of clusters to be determined.
2. Choose K instances (data points) within the dataset at random. These are the initial cluster centers.
3. Use simple Euclidean distance to assign to remaining instances to their closest cluster center.
4. Use the instances in each cluster to calculate a new mean for each cluster.
5. If the new mean values are identical to the mean values of the previous iteration the process terminates. Otherwise, use the new means as cluster center and repeat steps 3-5.

One way to measure the quality of clustering results is to use a silhouette (calculated as follows), which reflects how closely a data point matches to data within its cluster and how loosely it is matched to data of the neighboring cluster (the cluster with the lowest average distance from the data point; Rousseeuw, 1987).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the average dissimilarity of data point i with all other data within the same cluster, and $b(i)$ is the lowest average dissimilarity of data point i to any other cluster which i is not a member. Silhouette value can range from -1 to 1. A silhouette score close to 1 implies the data point is in an appropriate cluster, while a silhouette score close to -1 implies the data point is in the wrong cluster. The average silhouette of the entire dataset can also be used to assess the number of clusters in K-mean clustering algorithm. According to the algorithm, the number of clusters has to be decided at the beginning of the analysis. Therefore, in this analysis, several different numbers of clusters are tested, and average silhouette scores of each setting are calculated and compared. The results from the settings with the high average silhouette scores are analyzed.

Since the values of these three attributes are not in the same scale, they are standardized by the following formula before cluster analysis so that their values can be compared to minimize the effect of scale differences.

$$x_{new} = \frac{x - \mu}{\sigma}$$

where μ is the mean of the distribution and σ is the standard deviation.

After cluster analysis is conducted, its results are interpreted and compared with the results of single attribute analysis. The potential causes of identified anomalies and new audit objectives are discussed. The process of this cluster analysis is shown in Figure 36.

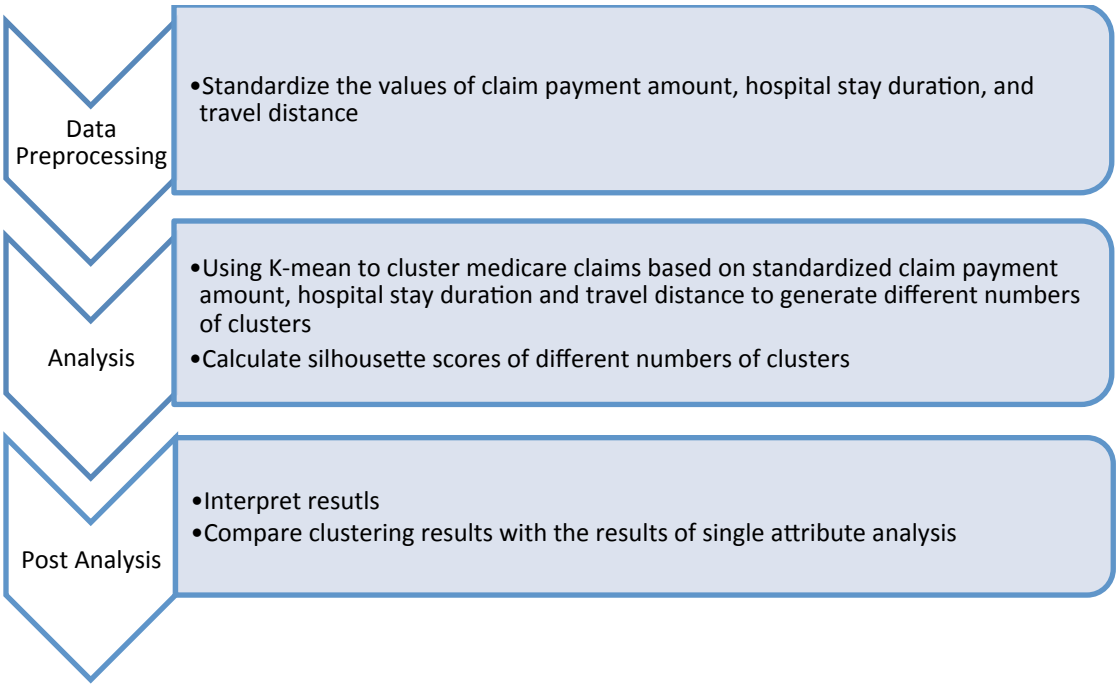


Figure 36: Cluster Analysis Process

4.3.2.2.3 Association Analysis

For the 16 diagnosis and procedure fields, simple frequency distribution or category statistics cannot generate very useful insight to assess fraud risk. Valuable information hidden in these fields is the co-occurrence of diagnoses and procedures. Common coincidences of diagnoses and procedures indicate normal treatments or complications, while rare combinations can be fraud risk indicators. Association analysis (introduced in Chapter 1, section 1.3.2 Advanced Exploratory Data Analysis Techniques) can generate rules from data to reflect relationships among items. It is applied in this study to reveal hidden relationships in diagnoses and procedures.

This study uses Apriori, one of the most commonly used association analysis algorithms. Apriori uses predefined minimum support (a measurement indicating how often a rule is applicable to a given dataset) to find all frequent itemsets in a database, then analyzes

these itemsets and utilizes the minimum confidence (a measurement indicating how strong a rule is) constraint to form rules. In this analysis, three sets of minimum confidence and minimum support values are tested and their results are compared. The general association analysis process applied in this study is shown in

Figure 37, which includes two major phases: application of association analysis algorithms and post-process analysis of results.

In the post-process results analysis stage, we first compare the results from different settings of minimum confidence and support levels to determine the combination that can generate the most valuable rules. These rules should not only reveal the commonly occurred diagnoses or procedures, respectively, but also discover frequent combinations of diagnoses and procedures. Experts can review these rules for validation and to learn new patterns from strong rules to support medical research. Rules with high confidence can also be used as benchmarks to filter out abnormal combinations of diagnoses, tests, or diagnoses and tests that don't follow these rules. Claims with these abnormal combinations can be considered high-risk, requiring further investigation. Validated association rules can be added to the existing audit checklist as new audit objectives.

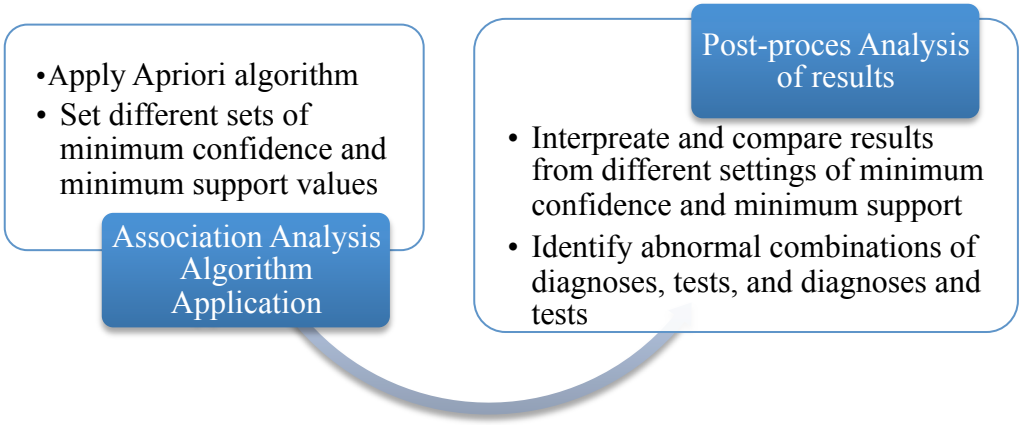


Figure 37: Association Analysis Process

4.3.3.3 Tools

The software packages used in this study are IBM SPSS Modeler[6] and SAS[7]. Data manipulation and descriptive statistics are conducted in SAS. Preprocessed SAS files are then imported into SPSS Modeler for cluster and association analysis. The results are exported back into SAS for interpretation, comparison and detailed investigation.

4.4 Results and Discussion

4.4.1 Conventional audit procedures results

Following the conventional audit procedures described in 4.3.2.1 Conventional audit procedures, 180,644 high-risk payments are identified. Since this number exceeds auditors' manageable amount for substantive testing, additional indicators must be used to prioritize claims for investigation. As discussed in 4.3.2.1 Conventional audit procedures, two possible fraud risk indicators are service providers who obtained extremely large

[6] http://www-01.ibm.com/software/analytics/spss/products/modeler/index.html
[7] http://www.sas.com/en_us/home.html

payments and who filed exceptionally large numbers of claims. To identify these service providers, a frequency distribution is calculated and claim payment amounts are summarized by service providers.

Descriptive statistics of service providers' frequency distribution are displayed in Table 11. The average number of claims filed by the 8302 service providers is 1500, and the standard deviation is 2534.86. According to the criteria used to define high-risk payment (mentioned in 4.3.2.1 Conventional audit procedures), service providers who filed more than 9104 (1500.02+3*2534.86) claims are considered high-risk. There are 192 such service providers.

Table 11: Descriptive Statistics of Service Providers' Frequency Distribution

Mean	Standard deviation	Minimum	Maximum	Count
1500.02	2534.86	1	39200.00	8302

Descriptive statistics of providers' payment summary are displayed in Table 12, which reveals that the average amount paid to these service providers is $15,111,119.19 and the standard deviation is $29,306,417.04. Based on the same criteria, the threshold payment amount for suspicion is $103,030,370.31 (15,111,119.19+29,306,417.04*3), which creates 178 exceptional service providers. 130 of these service providers have already been identified as service providers filed large number of Medicare claims. A total of 240 service providers are identified in these two analyses, which relate to 56,267 high-risk payments.

Table 12: Descriptive Statistics of Service Providers' Payment Summary

Mean	Standard deviation	Minimum	Maximum	Count

15111119.19	29306417.04	-375	229205708.79	8302

In summary, conventional audit procedures can identify 180,644 high-risk claims with large payments. 56,267 of these payments are associated with service providers that either filed a large number of claims or obtained exceptionally large payments. These claims should have higher priority for further investigation.

4.4.2 EDA results

4.4.2.1 Display Distributions

In the first step, distributions of selected fields are calculated and displayed in various forms. Numeric variables, such as payment amount, hospital stay period, and beneficiaries' travel distance are displayed in a boxplot (shown in Figure 38, Figure 39, and Figure 40, respectively). Frequency distributions and summary statistics based on the three numeric variables are displayed in Table 14 and Table 14 for beneficiary number and service provider number, respectively. Since each claim has more than one diagnosis and procedure, summary statistics are not applicable for diagnoses and procedures. Hence, only frequency distributions are computed for these two attributes, and these are displayed in Table 15.

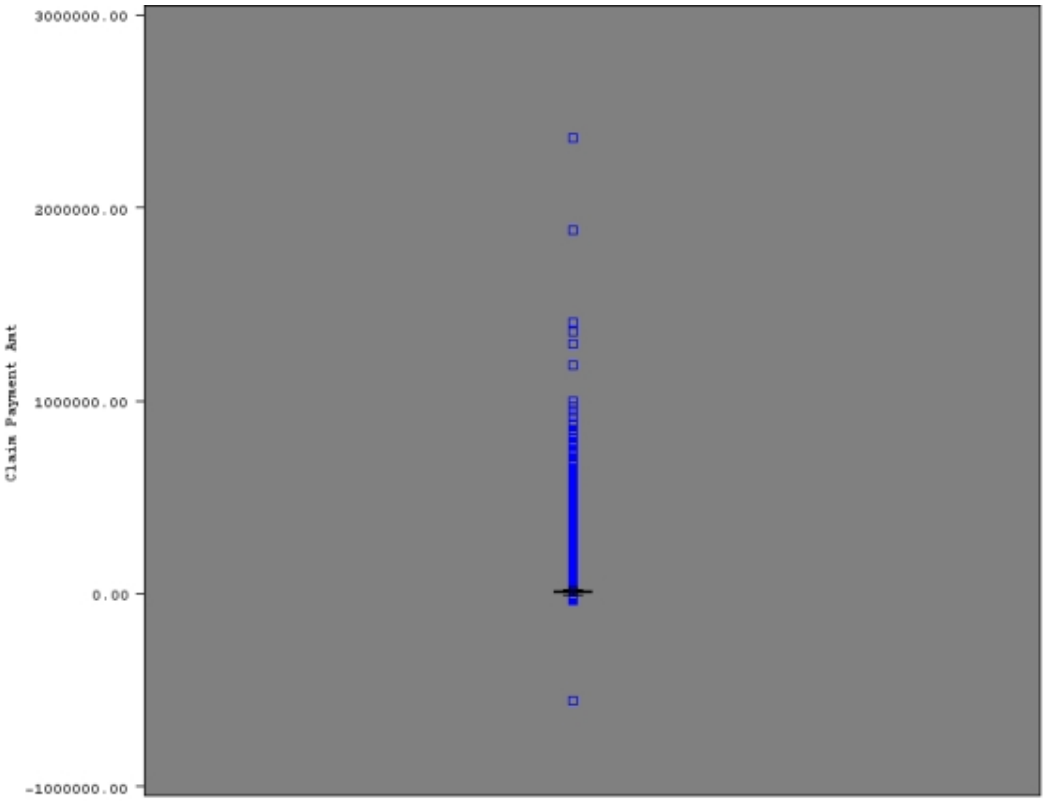


Figure 38: Distribution of Claim Payment Amount

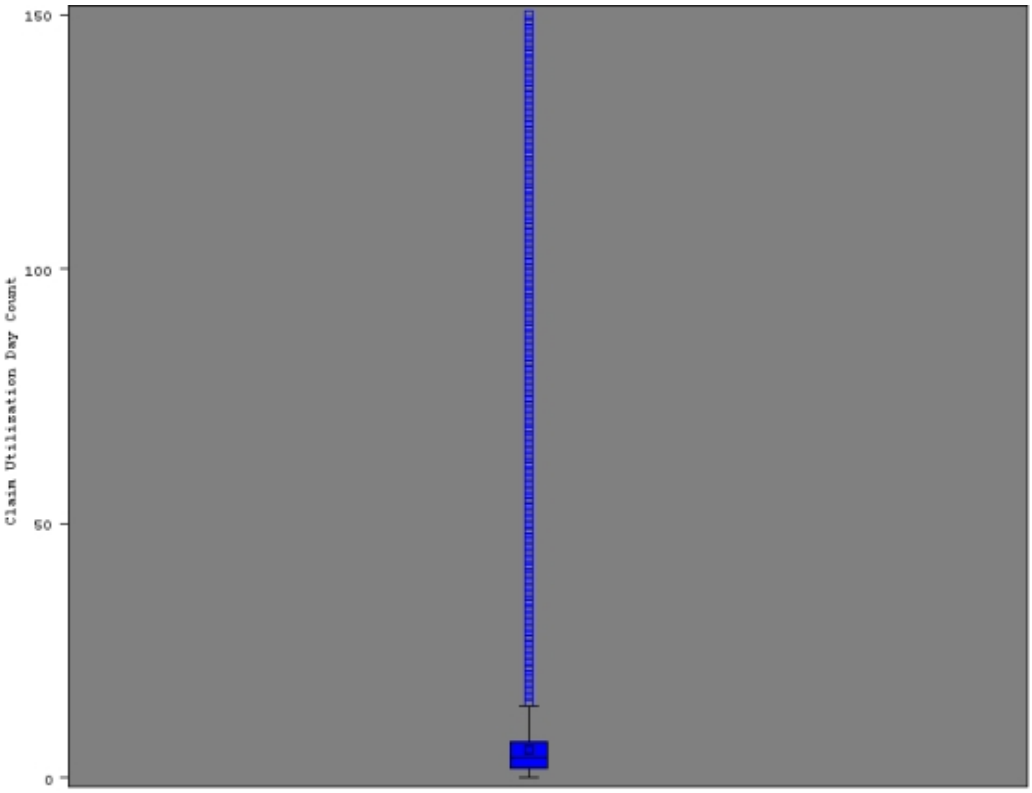


Figure 39: Distribution of Hospital Stay Period

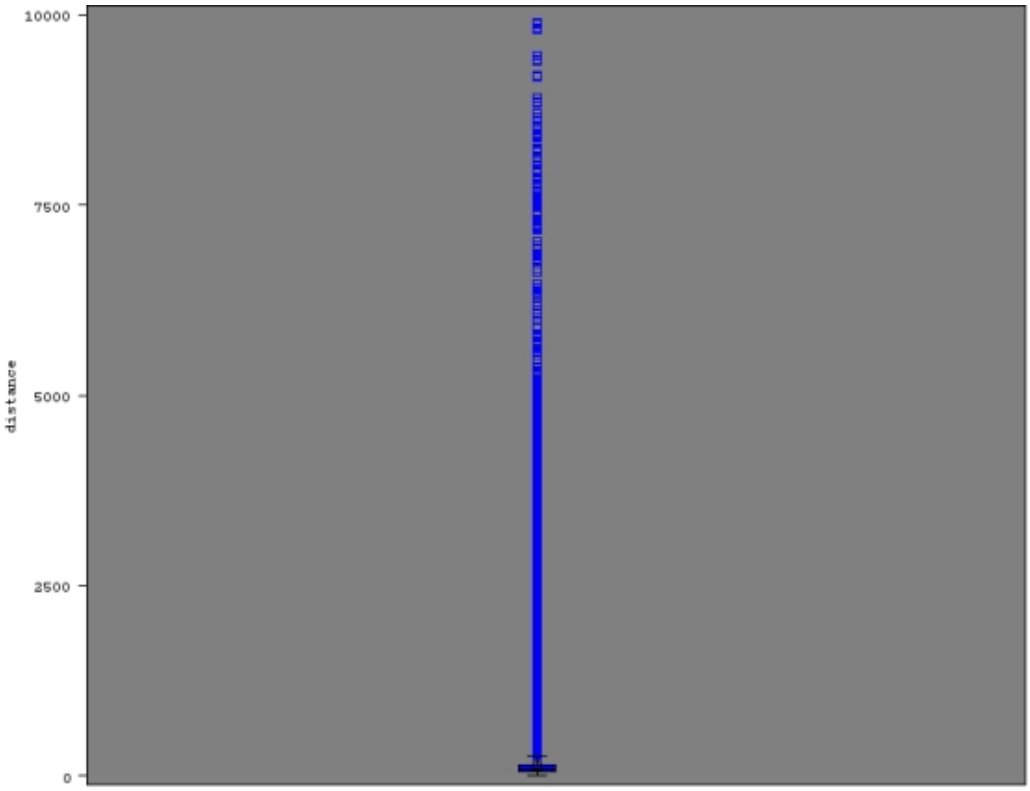


Figure 40: Distribution of Travel Distance

Table 13: Descriptive Statistics of Beneficiary Related Distributions

Distribution Name	Mean	Standard Deviation	Minimum	Maximum
Frequency distribution	1.7448999	1.3800865	1	103
Payment summary	17480.15	24400.12	-534561.55	3384174.56
Hospital Stay summary	9.7433405	13.9969875	0	668
Travel Distance summary	218.6821081	398.0258267	0	110925.55

Table 14: Descriptive Statistics of Service Provider Related Distributions

Distribution Name	Mean	Standard Deviation	Minimum	Maximum
Frequency Distribution	1500.02	2534.86	1	39200.00
Payment summary	15111119.19	29306417.04	-375	229205708.79
Hospital Stay summary	8354.34	13108.51	0	197988
Travel Distance	187506.53	356079.6	6.8939004	4546519.78

summary				

Table 15: Descriptive Statistics of Frequency Distribution of Diagnosis and Procedure

Variable Name	Mean	Standard Deviation	Minimum	Maximum	Count
Diagnosis	10177.5	215258.83	1	22093016	12236
Procedure	4637.43	32895.20	1	1195806	3969

In addition to descriptive statistics, an association analysis is performed on diagnoses and procedures to show their relational distribution. Specifically, association rules are obtained using minimum confidence (C_{min}) of 85%, and three different values for minimum support (S_{min}): l%, 0.5%, and 0.25% (Lai and Cerpa, 2001). The numbers of association rules with different confidence levels obtained in each of the experiments are shown in Table 16**Error! Reference source not found.** (the complete list of association rules generated from these three experiments is in Appendix D: Association Rules Generated from Medicare database). Generally speaking, more rules can be generated when S_{min} is reduced. In addition, more high confidence rules (C>90%) are generated in the experiments with lower $S_{min,}$ which implies natural correlations between certain diagnoses and procedures.

Table 16: Distribution of Generated Association Rules with Different S_{min} and C_{min}

	S_{min}=1%	S_{min}=0.5%	S_{min}=0.25%
C>95%	2	11	75
90%<C<95%	1	11	39
85%<C<90%	2	4	24
Total	5	26	138

The first experiment is set to identify those diagnoses and procedures which appeared in various combinations with S_{min} > 1% (or 124,531 Medicare claims). A minimum confidence of 85% is set to select the data into 5 production rules. The major results are:

1. In the identified combinations, the most commonly co-occurring diagnosis and procedure are a diagnosis of *Osteoarthrosis, localized, not specified whether primary or secondary, lower leg* (71536[8]) and procedure *Total knee replacement* (8154), appearing together in 1.665% of the Medicare claims. The rule obtained shows that there is a 97.29% chance that if a diagnosis of *Osteoarthrosis, localized, not specified whether primary or secondary, lower leg* (71536) is given, a *Total knee replacement* (8154) would be performed.

2. Among the obtained rules, the most reliable rule (having confidence=99.868%) is the co-occurrence of *insertion of drug-eluting coronary artery stent(s)* (3607) and *percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy* (0066).

The second experiment is set to identify those diagnoses and procedures which appeared in various combinations with $S_{min} > 0.5\%$ (or 62,265 transactions). $C_{min} = 85\%$ is set to reduce the data into 26 production rules. A greater amount of knowledge through the behavior patterns is gained by setting $S_{min} > 0.5\%$ rather than $S_{min} > 1\%$. The strongest rule in the 26 production rules has confidence equals to 99.993%.

Lowering S_{min} to 0.25% (or 31,133 transactions), with retaining $C_{min}= 85\%$ produces 138 rules and more detailed information. Many rules generated in this experiment have three or four antecedents, while most of the rules obtained in the first

[8] This is the code of *Osteoarthrosis, localized, not specified whether primary or secondary, lower leg*. In this database diagnoses and procedures follow the ninth revision of International Classification of Diseases (ICD9) maintained by the world health organization. Downloadable at http://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/codes.html

experiment (with S_{min}=1%) have only one antecedent, which implies that the rules produced in this experiment reflect more specific conditions.

For example, the strongest rule obtained in this experiment is: Given a diagnosis of *Coronary atherosclerosis of native coronary artery* (41401) and the performance of an *Insertion of drug-eluting coronary artery stent(s)* (3607), a *Coronary arteriography using two catheters* (8856), and a *Left heart cardiac catheterization* (3722), there is a 100% chance that a *Percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy* (0066) would also be claimed, accounting for 0.271% of the cases (33748 Medicare claims).

4.4.2.2 Identify Salient Features

According to Figure 38, Figure 39, and Figure 40, 188,662 large payments, 224,191 long hospital stays, and 206,359 long travel distances can be observed. Another noteworthy feature in the distribution is the existence of negative payments. What these negative payments mean and in which situation they occur need to be identified and verified.

The fourth row in Table 14 summarizes the number of days beneficiaries stayed in hospital in the year 2010. A salient feature can be identified from there, where the maximum value of beneficiary's hospital stay period is 668, whereas, there are only 365 days in a year. Therefore, a beneficiary should not stay in the hospital for more than 365 days within a year. In the raw data, 28 beneficiaries, who have spent more than 365 days in hospital, are found.

For diagnoses and procedures, descriptive statistics and distribution analysis cannot reveal very useful information. For example, frequency distribution can only show that there are 12,236 different diagnoses and 3,970 unique procedures in the dataset. The most common diagnosis is unspecified essential hypertension (4019) (5,107,503) and the most frequently claimed procedure is transfusion of packed cells (9904) (1,195,806).

After performing association analysis, the combinations of diagnoses and procedures described in the obtained association rules should be reviewed by experts to assess their appropriateness. Inappropriate combinations can imply misuse of procedures. Confirmed high confidence rules can be used to identify abnormal cases (salient features) from the dataset. In the first experiment (S_{min}=1%), applying obtained high confidence rules (C>95%) yields 212 noncomplying claims. Only 9 of these could have been detected by conventional audit procedures.

Eleven high confidence rules generated in the second experiment (S_{min}=0.5%) reveal 12,298 abnormal cases, among which 177 claims could have been identified in conventional audit procedures. All 212 exceptional claims discovered in the previous test are also detected in this analysis.

Using the rules developed in the third experiment (S_{min}=0.25%) yields 29,530 exceptional claims including all anomalies from the previous experiments. Therefore, generally speaking, at the same confidence level, association rules generated in the lower support level can discover more exceptional cases. Auditors can prioritize these identified anomalies by testing rules with higher confidence or generated from higher support levels. Association rules having 100% confidence (such as the rule illustrated in the last

paragraph) cannot directly identify any anomaly, but they are good candidates for audit objectives that can be tested on a new dataset.

4.4.2.3 Generate and Testing Hypotheses

One innocuous explanation for large payment amount, long hospital stay period, and long travel distance is a serious health issue. Claims having all the three features may be reasonable cases. Those with either long travel distances and short hospital stay periods or large payment amounts and short hospital stay periods are intuitively more suspicious. Hence, one hypothesis relating to this issue can be that long hospital stay periods should be accompanied by either large payment amounts or long travel distances. However, it is not easy to test this hypothesis and identify further suspicious cases using a single variable, and in any case, this measure generates many anomalies, hindering prioritization. Therefore, a cluster analysis considering the relationships among all these three variables is performed. The analysis results may be able to provide clues to validate the hypothesis[9] and highlight more suspicious cases.

According to CMS, negative payment amount can be presented in two situations: (1) When a beneficiary is charged the full deductible during a short stay and the deductible exceeded the amount Medicare pays, or (2) when a beneficiary is charged a coinsurance amount during a long stay and the coinsurance exceeds the amount Medicare pays[10]. These two situations can be combined in one hypothesis: a negative payment appears when the deductible or coinsurance amount exceeds the Medicare payment.

---

[9] As indicated in **Error! Reference source not found.**, advanced EDA techniques such as cluster analysis may provide conclusive results. Therefore, their results may be able to use to test hypotheses as well.

[10] http://www.resdac.org/resconnect/articles/120

Claims with negative payment amount should have either a deductible or coinsurance. Two new variables – beneficiaries' deductible amount (NCH_BENE_IP_DDCTBL_AMT) and beneficiaries' coinsurance amount (NCH_BENE_PTA_COINSRNC_AMT) – are used.

For the beneficiaries who claimed more than 365 days hospital stay in 2010, one possible reason is that the claims start from 2009. The equivalent hypothesis of this explanation is that the beneficiaries who stayed more than 365 days in the hospital are admitted before 2010. Therefore, some days in 2009 are added together with the days in 2010 as hospital stay period. To test this hypothesis, these beneficiaries' admission date (CLM_ADMSN_DT) and discharge date (NCH_BENE_DSCHRG_DT) are used to calculate the real number of days they stayed in hospital. Then calculated values are compared with the hospital stay periods recorded in the claims to inspect whether the claims truly reflect the actual number of days the beneficiaries stayed in hospital. If claimed hospital stay periods are longer than the actual days beneficiaries stayed in hospitals, Medicare may overpaid the service providers for these claims.

Because every patient has different health issues, it is possible that special procedures are performed for a particular patient due to his/her complicated health problem. Therefore, a general hypothesis for the salient features identified in the association analysis can be that anomalies identified in association analysis are caused by patients' special health conditions. To test this hypothesis, domain experts need to review related records of these anomalies to determine whether they are caused by special medical situations.

4.4.2.3 Identify Suspicious Cases

The aforementioned tests revealed the following:

1. 7 out of 12,417 claims with negative payment amount are not associated with a positive deductible or coinsurance amount.

2. 25 out of 28 beneficiaries who were paid for more than 365 days were not actually in the hospital for so long.

3. Among the 138 claims relating to these 25 beneficiaries, 6 potential duplicate claims are identified. Each of the claims has exactly the same payment amount, hospital stay period, and diagnoses and procedures as another claim in the dataset.

Cluster analysis is then performed to discover suspicious cases from beneficiaries' travel distance, payment amount, and hospital stay period. Several numbers of clusters, from 2 to 9, have been tested. The relationship between the number of clusters and resulting silhouette coefficient (SC) is plotted in Figure 41, which shows that SC scores decrease as the number of clusters increases. Generally speaking, a SC score greater than 0.51 suggests a reasonable structure, while a score greater than 0.71 suggests a strong structure (Lewis, 2010). 6 cluster counts (2 to 7) result in strong clusters. The highest SC score is 0.806 for 2 clusters, decreasing slightly to 0.803 with 3 clusters. Then it dramatically dropped to 0.734 with 4 clusters. 5, 6, and 7 clusters each have a score of 0.714. 8 and 9 clusters yield SC scores of 0.7 and 0.675, respectively.

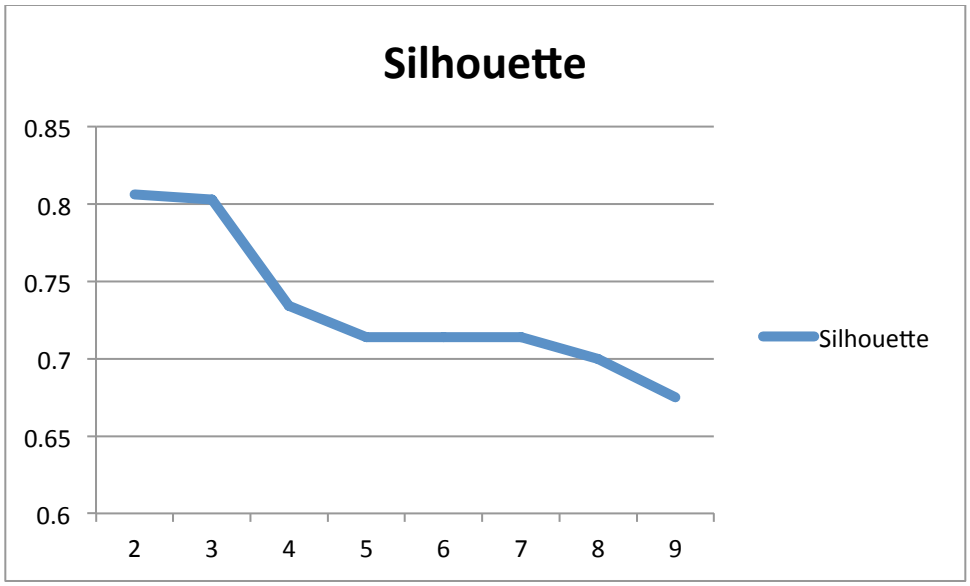


Figure 41: Number of Clusters and Resulting Silhouette Coefficient

When the numbers of clusters are selected as 2 the algorithm can generate highest quality clusters. In this experiment, about 92.2% of claims fall into cluster 1 and 7.8% are in cluster 2. The distribution of these two clusters and the absolute distributions of the input variables in each cluster are shown in Figure 42. Claims in the larger cluster have smaller payment amounts and shorter hospital stay durations compared to those in the smaller cluster. This is reasonable: the larger cluster relates to ordinary diseases and smaller cluster may be associated with more serious diseases, therefore, no anomalies are directly shown in this clustering result. In addition, beneficiaries' travel distance doesn't have enough discriminating power to distinguish the claims grouped in these two clusters (the background color of this variable shows lighter). Therefore, this information is not fully used in this analysis.

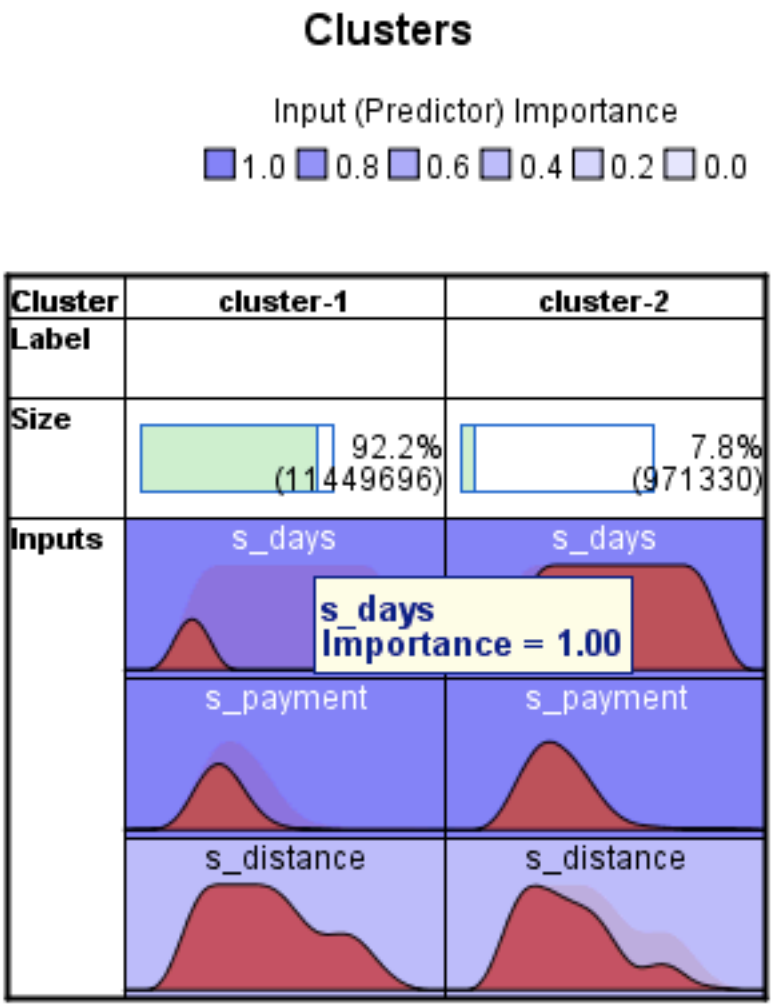


Figure 42: Cluster Analysis Results of 2 Clusters

A three-cluster model yields a smaller cluster (Figure 43). The claims in this small group, 1.6% of the entire population, are associated with long travel distance, small payment amount, and short hospital stay. Recall that reasonable long distance travels should be due to serious diseases, which may lead to longer hospital stays and larger payment amounts. The claims in the third cluster are more likely to be suspicious. Since there are 195,343 claims in the third cluster, and since not all these claims feature long distance, more clusters must be used to narrow the investigation scope.

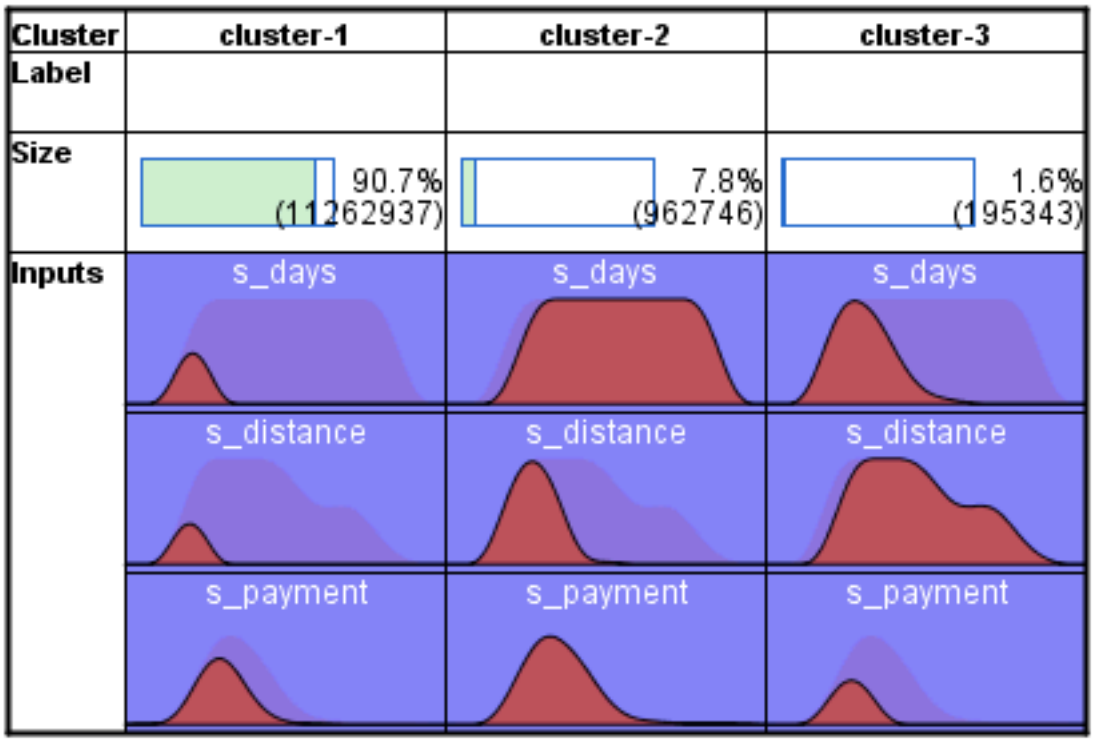


Figure 43: Cluster Analysis Results of 3 Clusters

Results for 7 clusters, the largest cluster count with a strong SC score, are shown in Figure 44. According to this result, claims in clusters 1, 5, 6, and 7 have relatively short travel distance, short hospital stay period, and small amount of payment. Cluster 2 relates to long hospital stay period, short travel distance, and relatively large amount of payment. As discussed before, these two patterns are considered normal. More than 99.9% of claims are assigned to these clusters. Clusters 3 and 4 contain 3,671 and 47 claims, respectively. Claims in cluster 3 have long travel distance, short hospital stay period, and small payment amount. This distribution is similar to the smallest cluster in the 3-cluster experiment, but the travel distances in this cluster are longer, implying more suspicious claims. Cluster 4 contains claims with large payment amount and short hospital stay period. This is a new abnormal pattern revealed in this analysis.

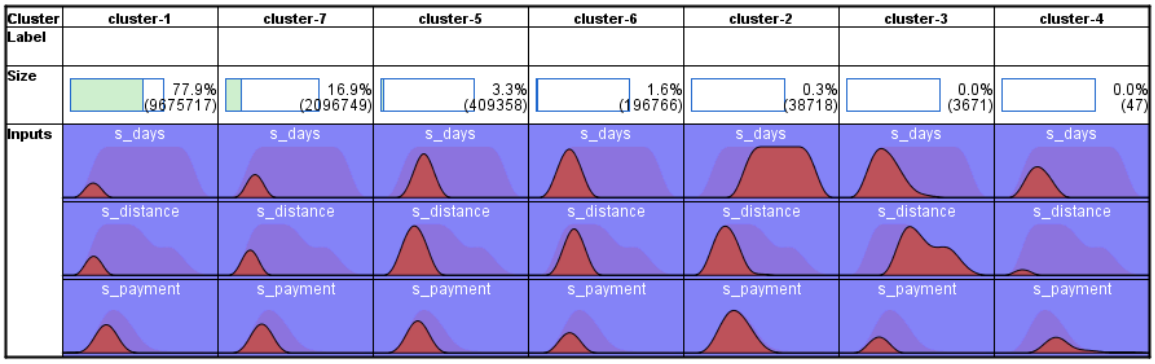


Figure 44: Analysis Results of 7 Clusters

The seven-cluster analysis reduces suspicious claims from 195,343 to 3718 (3671+47), which are more feasible for auditors to examine. These 3718 claims are associated with 2733 beneficiaries and 1222 providers. Compare this result with the results of single variable analysis, in the 188,662 large payments, 224, 191 long hospital stays and 206, 359 long travel distances identified in 4.4.2.2 Identify Salient Features, 3668 claims are also discovered in cluster analysis. In addition, 50 new suspicious claims are detected in cluster analysis. Conventional audit procedures only identified 119 out of these 3718 claims and 36 out of these 1222 providers. Cluster analysis provides more insight and reveals more suspicious cases to narrow down the scope of further investigation. In addition, unlike conventional audit procedures, this analysis does not discriminate between fraud types.

4.4.2.4 Explore the Causes of Exceptional Cases to Develop New Hypotheses, Test New Hypotheses, and Generate New Audit Objectives and Report Finding

The causes of all the previously identified suspicious claims need to be investigated with additional supporting evidence or experts' domain knowledge. For example, claims associated with long travel distance and a short hospital stay may occur when a beneficiary is traveling. To check this hypothesis, auditors need to refer to more

detailed doctor visit records and beneficiaries' medical history for more evidence. Once the identified anomalies/suspicious cases are confirmed as irregularities, the following new audit objectives can be created:

1. Negative payment amount was presented when deductible amount or coinsurance amount exceeded the amount Medicare pays.
2. One beneficiary was paid for at most 365 days' hospital stay per year.
3. The number of days paid by Medicare did not exceed the actual number of days the beneficiary stayed in hospital. .
4. Large payment amounts and long travel distances were associated with long hospital stays.
5. If *Coronary atherosclerosis of native coronary artery* (41401) was diagnosed and the an *Insertion of drug-eluting coronary artery stent(s)* (3607), a *Coronary arteriography using two catheters* (8856), and a *Left heart cardiac catheterization* (3722) were performed, a *Percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy* (0066) was performed as well. [11]

4.5 Conclusion

In conclusion, this chapter demonstrates how EDA can be applied to healthcare data to assess fraud risk. Specifically, real Medicare inpatient claims are analyzed in this case study. Both conventional audit procedures and EDA process are conducted. EDA generates descriptive statistics and performs cluster and association analysis. Compared

[11] This is an example of audit objective that derived from obtained association rules. Each verified association rule could be considered as a new audit objective.

with the conventional audit procedures, EDA process can not only reveal more hidden risk areas but also narrow down the scope for substantive testing to the most suspicious cases.

Unlike the case study presented in Chapter 3, where EDA is employed at " Perform Audit Plan" stage, this field study applies EDA at the "Develop Audit Plan" stage. Therefore, the general audit objective of this case study is fraud risk assessment no specific audit objective exists in this case study. Conventional audit procedures focus on analyzing payment amount and provider profile, using statistical measures to define abnormal healthcare claims. Medicare claims with extreme large payment amounts are their main concern. However, sophisticated fraudulent cases may not include large payments; therefore, this kind of fraud can not necessarily be identified by conventional audit procedures.

EDA considers more attributes, including beneficiaries' hospital stay period, travel distance, and diagnoses and performed procedures. These fields and their relationships are explored. The major findings are:

1. Descriptive statistics discover 25 beneficiaries who were inappropriately paid for more than 365 days in 2010
2. Cluster analysis identifies 3,671 Medicare claims with long travel distances, short hospital stay periods, and small payment amounts; and 47 claims with large payment amounts and short hospital stay periods.
3. Association analysis creates up to 75 strong rules to describe relationships among diagnoses and procedures, revealing at least 212 exceptional Medicare claims from the data.

As in the first case study, a lack of supporting evidence means that suspicious claims cannot be verified. Therefore, complete EDA processes are not performed. In addition, only 21 out of 257 fields that directly relate to healthcare fraudulent behaviors are included in this case study for demonstration purpose. It is possible that some unselected fields can also be used as fraud indicators, but due to limited domain knowledge and supporting information, analysis of these fields was not performed. Results show that EDA allows auditors to gain detailed insight of the data and to identify abnormal cases that standard analytical procedures cannot.

References

Blanchard, T.P. (2007). Medicare Medical Necessity: Avoiding Overpayments, Penalties and Fraud Allegations. Health, Civil No. 05-00521 JMS/LEK. Available at: http://www.healthlawyers.org/Events/Programs/Materials/Documents/PHYHHS11/blanchard.pdf

Chan C.L., Lan C.H. (2001) A data mining technique combining fuzzy sets theory and Bayesian classifier—an application of auditing the health insurance fee. In Proceedings of the International Conference on Artificial Intelligence, 402–408

He H., Wang J., Graco W., Hawkins S. (1997) Application of neural networks to detection of medical fraud. Expert Syst Appl 13:329–336

He H., Hawkins S., Graco W., Yao X. (2000) Application of Genetic Algorithms and k-Nearest Neighbour method in real world medical fraud detection problem. Journal of Advanced Computational Intelligence and Intelligent Informatics 4(2):130–137

Lai, K., and Cerpa, N., Support vs Confidence in Association Rule Algorithms. Cerpa Proceedings of the OPTIMA Conference, October 10-12, 2001, Cuico, Chile.

Lewis, P. D. (2010) R for Medicine and Biology. Jones & Bartlett Learning

Li J., Huang K.Y., Jin J., Shi, J. (2007) A Survey on statistical methods for healthcare fraud detection, Healthcare Management Science 11 (3): 275-287

Liou F.M., Tang Y.C., Chen J.Y. (2008) Detecting hospital fraud and claim abuse through diabetic outpatient services. Helth Care Manage Sci 11: 353-358

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

Musal R.M. (2010) Two models to investigate Medicare fraud within unsupervised databases. Expert Systems with Applications 37: 8628-8633

NHCAA (2005) The Problem of Healthcare Fraud: A serious and costly reality for all Americans, report of National Healthcare Anti-Fraud Association (NHCAA)

Ormerod T., Morley N., Ball L., Langley C., Spenser C. (2003) Using ethnography to design a Mass Detection Tool (MDT) for the early discovery of insurance fraud. In Proceedings of the ACM CHI Conference

Ortega P. A., Figueroa C. J., Ruz G. A. (2006) A medical claim fraud/ abuse detection system based on data mining: a case study in Chile. In Proceedings of International Conference on Data Mining, Las Vegas, Nevada, USA

Roiger R. J. and Geatz. M. W. (2003). Data Mining: A Tutorial-Based Primer (International Edition). Pearson Education, USA.

Rousseeuw P. J. (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65

Sokol L., Garcia B., West M., Rodriguez J., Johnson K. (2001) Precursory steps to mining HCFA healthcare claims. In Proceedings of the 34th Hawaii International Conference on System Sciences

Tan P.N., Steinbach M., Kumar V. (2006) Introduction to Data Mining, Pearson Education.

Thiprungsri, S. (2011) Cluster analysis for anomaly detection in accounting Data: An Audit Approach. The international Journal of Digital Accounting Research. 11: 69-84

Viveros M.S., Nearhos J.P., Rothman M.J. (1996) Applying data mining techniques to a health insurance information system. In Proceedings of the 22nd VLDB Conference, Mumbai, India, 286– 294

Williams G., Huang Z. (1997) Mining the knowledge mine: The Hot Spots methodology for mining large real world databases. Lect Notes Comput Sci 1342:340–348

Yang W.S., Hwang S.Y. (2006) A process-mining framework for the detection of healthcare fraud and abuse. Expert Syst Appl 31:56–68

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 8:275–300

Chapter 5 Conclusion and Future Research

5.1 Summary

Exploratory data analysis (EDA) is a statistical data analysis approach originating centuries ago. It analyzes data in as many ways as possible to discover hidden patterns and identify clues to inspire ideas and hypotheses. Traditional EDA is a mean-oriented approach that presents data using simple arithmetic and images, and it results in descriptive statistics, data visualization and data transformation. Modern EDA uses a goal-oriented approach aiming at outlier detection, pattern recognition, and variable. Advanced EDA techniques include feature selection, data mining techniques such as cluster analysis and association analysis, text mining, process mining, and social network analysis.

Since the goals of EDA are similar to the objectives of some essential audit steps, EDA can be a useful approach for auditors to fulfill some of their tasks such as fraud detection and risk assessment. In addition, it can supplement currently used confirmatory data analysis (CDA) to discover previously unknown risks, improving audit quality. However, even though EDA is extensively studied in many disciplines such as geography, marketing, and operations management, there have been only a handful of papers in accounting that have discussed EDA. One possible reason for this phenomenon is the high effort formerly involved in data collection, leading auditors to focus on doing more with less. Thus, they prefer to use traditional sampling and CDA approaches in the audit process. The increasing availability of data enables EDA for more auditors. Current EDA related papers in auditing focus on either the application of one EDA technique in an auditing context or the use of certain EDA techniques to fulfill one audit task. Compared

with arbitrarily choosing EDA techniques for certain audit procedures, systematically integrating EDA throughout the audit cycle can maximize its value. This dissertation investigates the value EDA can bring to auditors and how it can be systematically applied in auditing.

The first essay proposes a conceptual framework for an auditing application of EDA. The framework illustrates when EDA can be applied in the audit cycle, how various EDA techniques benefit auditors in different audit procedures, and EDA-related auditing best practices. EDA can be applied to seven audit activities: assessing engagement risk, understanding client's business and assessing client's business risk, understanding internal control and assessing control risk, assessing fraud risk, performing analytical procedures, reviewing subsequent events, and assessing engagement quality.

Eight types of techniques enable EDA in the audit cycle: descriptive statistics, data visualization, data transformation, feature selection, data mining, text mining, social network analysis, and process mining. Traditional EDA techniques, such as descriptive statistics and data visualization, allow auditors to understand trends and distributions of both financial and operational data, which are the basis of most audit steps. Advanced data mining techniques can help auditors to gain a deeper understanding of the patterns hidden in the data to reveal evidence that cannot be easily seen. Text mining can be used to screen electronic documents, which may be needed in engagement risk, business risk and fraud risk assessment. Process mining can be used to analyze business and internal control processes, enabling auditors to understand clients' business and internal control processes and assess related risks. Social network analysis can be combined with process mining or text mining to analyze relationships between participants involving in business

activities or communications. Revealed relationships can assist auditors in understanding clients' businesses and assessing business and fraud risks.

In addition, an eight-step application process is proposed to guide auditors' implementation of EDA: display the distribution of fields related to an audit object, identify salient features, generate hypotheses, test hypotheses, identify suspicious cases, explore the causes of abnormal cases to generate new hypotheses, test new hypotheses, and add new audit objectives and report the results. This process includes both EDA and CDA steps to guarantee the most reliable analysis results. In addition, even though the proposed process contains eight normative steps, it should be flexibly implemented by auditors on a case-by-case basis.

Besides the application of EDA in traditional audit settings, this essay briefly discusses how EDA can be used in a continuous auditing environment. Generally speaking, EDA should be performed in the continuous auditing system design and development phases and integrated into the continuous auditing system as part of continuous risk monitoring and assessment. Even though EDA can bring numerous benefits to auditors, the biggest drawback of this approach is that it is a very interactive (and therefore effort-intensive) process requiring continuous human intervention. Some researchers have automated EDA processes with the help of artificial intelligence techniques (St. Amant and Cohen, 1998; Becher et al., 2000). Based on these studies, a conceptual schema is proposed to automate some steps of the suggested EDA process where the integration of EDA into continuous auditing system is supported by artificial intelligence techniques such as expert systems.

The second essay provides a field study of the application of EDA in the performance stage of an operational audit. In this study, the eight-step EDA application process is performed on a real dataset from an international bank in Brazil. Descriptive statistics, data transformation, and data visualization techniques are used to assess internal control risk and detect fraud. Many critical risk issues not identified by standard audit tests, such as negative discount, inactive agents, and short calls, are detected by EDA techniques, which demonstrate that comprehensive findings can easily be obtained even with simple statistics and visualization techniques.

The third essay applies the proposed EDA process in the audit planning stage to assess fraud risk in 2010 inpatient Medicare claims. Descriptive statistics as well as cluster analysis and association analysis are performed in this case study. By extending the scope of analysis, descriptive statistics can discover abnormal claims that are ignored by conventional audit procedures. Cluster analysis is conducted to identify abnormal claims based on payment amount, beneficiaries' travel distance, and hospital stay period. Compared with conventional audit procedures and single variable analysis, cluster analysis can not only reveal more hidden risk areas but also narrow down the scope of substantive testing to the most suspicious cases. Association analysis focuses on discovering the relationships and common combinations among diagnoses and procedures. Auditors can assess the appropriateness of common combinations of diagnoses and procedures to identify extensive misuse of certain procedures and set the rules with high confidence as audit objectives to reveal abnormal behaviors. The EDA approach doesn't emphasize any specific type of fraud, such as the provider fraud

emphasized by conventional audit procedures; therefore, beneficiary and collusive fraud, overlooked in current healthcare fraud detection literature, can also be discovered.

The applications of the first essay's proposed conceptual model in the second and third essays are demonstrated in Table 17.

Table 17: Mapping of EDA Applications in the Chapter 3 and 4 to the Suggested Conceptual Framework Proposed in Chapter 2

EDA Framework\Case Study		Credit Card Retention	Healthcare Fraud Detection
Audit Flow		Perform Audit Plan	Audit Plan Development
Mean		• Descriptive Statistics • Data Visualization	• Descriptive Statistics • Data Visualization • Cluster Analysis • Association Analysis
Process	1. Display Distribution	Objective 1: Display the frequency distribution of Discount Objective 2: Display the frequency distribution of bank representatives Objective 3: Display the frequency distribution of call duration	1. Display the frequency distribution of payment amount, hospital stay period, and travel distance 2. Calculate the descriptive statistics of beneficiaries, service providers, and diagnoses and procedures 3. Perform association analysis of diagnoses and procedures
	2. Identify Salient Features	Objective 1: Negative discounts Objective 2: Inactive bank representatives Objective 3: Unreasonably short calls	1. Extremely large payment amount, long hospital stay period, and long travel distance 2. Negative claim payment amount 3. Some beneficiaries stayed in hospital for more than 365 days in 2010 4. Rare combinations of diagnoses and procedures in some claims
	3. Generate Hypotheses	Objective 1: Negative discounts are due to	1. Long hospital stay periods are accompanied by

		group discounts offered to clients with more than one credit cards Objective 2: Bank representatives who answered few phone calls are supervisors Objective 3: Unreasonably short calls are accidentally disconnected because of network problems	either large payment amounts or long travel distances 2. Negative payment amount is presented when the deductive or coinsurance amount exceeds the amount Medicare pay 3. Beneficiaries who stayed more than 365 days in the hospital were admitted before 2010 4. Anomalies identified in association analysis are caused by patients' special health conditions
	4. Test Hypotheses & 5. Identify Suspicious Cases	Objective 1: 96 negative discounts are too large for group discounts and 39 smaller negative discounts cannot be explained by group discounts Objective 2: 91.8% inactive bank representatives are not supervisors Objective 3: 99.57% of short calls were not accidentally disconnected	1. 7 negative payments have zero deductible amount and coinsurance amount 2. 25 beneficiaries who were paid for more than 365 days are not actually stayed in hospital for such long 3. 6 potential duplicate claims 3671 claims have long travel distances, short hospital stay periods, and small payment amounts 4. 47 claims have large payment amounts and short hospital stay periods
	6. Investigate Causes of Suspicious Cases to Develop New Hypotheses	Objective 1: Large negative discounts are caused by irregular actual fees Objective 2: Non-supervisory inactive bank representatives are interns	Suspicious claims associated with long travel distance and short hospital stay happen when beneficiaries are on vacation in the places far from their residences
	7. Test New	Objective 1: Large	Need additional information

	Hypotheses	negative discounts are caused by input errors	to complete
	8. Create New Audit Objectives and Report Results	Objective 1: (1) Actual fees were recorded correctly; (2) negative discounts have been offered to clients with multiple cards. Objective 2: all non-supervisory permanent bank representatives were active bank representatives Objective 3: all the effective phone calls lasted more than 1 minute.	1.Negative payment amount was presented when deductible amount or coinsurance amount exceeded the amount Medicare pays. 2. Beneficiary was paid for at most 365 days' hospital stay per year. 3. The number of days paid by Medicare did not exceed the actual number of days the beneficiary stayed in hospital. 4. Large payment amounts and long distance travels were associated with long hospital stays. 5. All the verified association rules can be considered as new audit objectives.

The main contribution of this dissertation lies in the intersection of auditing and data analysis. It introduces a new approach, EDA, to auditors. EDA may fundamentally change existing audit data analysis based on testing predefined audit objectives derived from fixed management assertions. In the contemporary business environment, traditional audit data analysis method is not sufficient for auditors to provide high-level assurance; new forms of risks can be missed in the audit process. EDA is more in the big data era. By integrating EDA into audit processes, auditors can gain deeper insight into clients' data by performing new forms of analysis (such as cluster analysis and association analysis) and take advantage of more diverse information to obtain new types of audit

evidence. In addition, EDA enables auditors to discover not only emerging risks but also new opportunities for the organization, expanding audit scope from providing assurance service to offering strategic level guidance to the organization. Moreover, EDA approach allows auditors to generate risk-oriented audit objectives, which ensures the quality of their service. Even though auditors can gain numerous benefits from EDA, they may also encounter some challenges when applying EDA in practice. For example, a practical issue of EDA is over analyzing. Theoretically, users should perform as many analyses as possible to explore the data. However, in really, while constrained by various factors, when should auditors stop exploring the data is a question worth considering. In addition, as discussed in the previous paragraphs, EDA process is difficult to automate. How to build an artificial intelligent model that can generate explanations for emerging features identified from the data is also a challenge for both auditors and IT practitioners. Table 18 summarizes the benefits and challenges EDA can bring to auditors.

Table 18: Summary of the Benefits and Challenges of Applying EDA in Auditing

Benefits	Challenges
• Broader audit scope • Identify emerging risks and opportunities • Enable new forms of analysis • Take advantage of diverse information • Establish risk-oriented audit objectives	• When to stop? • How to automate?

The basic idea of this proposed framework was presented at the Strategic and Emerging Technologies (SET) workshop during the 2013 American Accounting Association (AAA) annual meeting. Preliminary results of the case studies were reported in 12th and 13th Bryant University XBRL and Healthcare Standardization Conference and

the 28th and 29th World Continuous Auditing and Reporting Symposia. Demand for integrating EDA into the audit process was discussed in Journal of Information Systems (Liu and Vasarhelyi, 2014).

5.2 Limitations

The main limitation of this dissertation lies in the unavailability of supporting evidence to verify suspicious cases. In both case studies, the verification of identified exceptions requires additional information. Without this information the causes of the exceptional cases cannot be confirmed. Therefore, in both case studies, some EDA processes are not completed. In reality, auditors do not have this problem.

In addition, due to the lack of appropriate data, the proposed EDA process is only applied to the planning and performance stages of the internal audit. Not all techniques included in the proposed framework are demonstrated in the case studies for this reason.

5.3 Future Research

This study, as an initial examination of the systematic application of EDA in auditing, is a starting point. There are many possibilities for future research.

1. Demonstrate the application of EDA in other types of audit. Due to limited data access, the application of EDA following the proposed process is only applied in operational audit and forensic audit contexts in this dissertation. The same process should also be tested in other audit settings (e.g. the financial audit) to demonstrate its universal value.

2. Apply the proposed EDA process to the other stages of audit. As shown in the

proposed framework, EDA can be employed in all the four stages throughout the audit cycle. The case studies in this dissertation illustrate how EDA can be used for planning and performance. The application of this approach in the initial planning stage, to assess the engagement risk and understand client business, and in the review and reporting stage, should also be studied.

3. Explore the application of other EDA techniques in auditing. The proposed framework includes eight EDA. Due to lack of appropriate data, only four EDA techniques are applied in the case studies in this dissertation. The application of the other four EDA techniques (feature selection, process mining, text mining and social network analysis) following the proposed EDA process needs to be demonstrated.

4. Recommend the best EDA techniques for auditors to use in specific contexts. Auditors may feel confused when they have abundant EDA techniques to choose to accomplish a specific task. Therefore, after exploring the application of all the EDA techniques in various audit contexts, a detailed framework can be developed to list the most appropriate EDA techniques for specific audit purposes. This framework can provide guidance for auditors' choice of EDA techniques in practice.

5. Develop information systems for EDA application in a continuous auditing environment. EDA automation is only discussed at the theoretical level in this dissertation. In the future researchers can attempt to implement this concept.

6. Investigate behavioral issues related to EDA. Since it is an interactive process,

human judgments, such as field selection, explanation generation, parameter setting, and outlier definition, are needed when performing EDA. Behavioral issues may crop up. Future research can examine the behavioral issues involved with the human judgment required by EDA.

7. Even though the importance of data analysis has been recognized by auditors, it has not been reflected in audit standards. Current audit standards must be modified to enable audit data analysis (Titera, 2013). Therefore, one goal of this research is to include EDA as an audit analysis approach in the revised audit standards. Hence, the integration of EDA into audit standards should be researched.

In summary, EDA is a very promising area for auditing research. This dissertation is only the first attempt to apply EDA in the field of auditing. With sufficient follow-up, the value of EDA to auditors can be comprehensively discovered.

Reference

Becher, J. D., Berkhin, P., Freeman, E. (2000) Automating Exploratory Data Analysis for Efficient Data Mining. KDD '00 Procedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 424-429

Liu, Q., and Vasarhelyi, M. (2014) Big questions in AIS research: measurement, information processing, data analysis, and reporting. Journal of Information Systems, 28 (1): 1-17

St. Amant, R., and Cohen, P. R. (1995) Issues in Automating Exploratory Data Analysis. AAAI Technical Report SS-95-03.

St. Amant, R., and Cohen, P. R. (1998) Intelligent Support for Exploratory Data Analysis. Journal of Computational and Graphical Statistics. Vol 7 (4): 545-558

Titera, W. T. (2013) Updating Audit Standard – Enabling Audit Data Analysis. Journal of Information Systems, 27 (1): 325-331

Appendix A: Potential application areas of EDA in Clarified Statements on Audit Standards issued by AICPA

Audit Standard	Potential application areas of EDA
AU-C sec. 240 Consideration of Fraud in a Financial Statement Audit	**.22** Based on analytical procedures performed as part of risk assessment procedures,[8] the auditor should evaluate whether ***unusual or unexpected*** relationships that have been identified indicate risks of material misstatement due to fraud. To the extent not already included, the analytical procedures, and evaluation thereof, should include procedures relating to revenue accounts. (Ref: par. .A24–.A26 and .A46) **.27** The auditor should treat those assessed risks of material misstatement due to fraud as significant risks and, accordingly, to the extent not already done so, the auditor should ***obtain an understanding*** of the entity's related controls, including control activities, relevant to such risks, including the evaluation of whether such controls have been suitably designed and implemented to mitigate such fraud risks. (Ref: par. .A36–.A37) **.32** Even if specific risks of material misstatement due to fraud are not identified by the auditor, a possibility exists that management override of controls could occur. Accordingly, the auditor should address the risk of management override of controls apart from any conclusions regarding the existence of more specifically identifiable risks by designing and performing audit procedures to, etc. **a.** test the appropriateness of journal entries recorded in the general ledger and other adjustments made in the preparation of the financial statements, including entries posted directly to financial statement drafts. In designing and performing audit procedures for such tests, the auditor should (Ref: par. .A47–.A50 and .A55) i. obtain an ***understanding*** of the entity's financial reporting process and controls over journal entries and other adjustments, [12] and the suitability of design and implementation of such controls; ii. make inquiries of individuals involved in the financial reporting process about inappropriate or ***unusual*** activity relating to the processing of journal entries and other adjustments; , etc..

	c. evaluate, for significant transactions that are ***outside the normal*** course of business for the entity or that otherwise appear to be ***unusual*** given the auditor's ***understanding*** of the entity and its environment and other information obtained during the audit, whether the business rationale (or the lack thereof) of the transactions suggests that they may have been entered into to engage in fraudulent financial reporting or to conceal misappropriation of assets. (Ref: par. .A54) **.A21** Those charged with governance of an entity oversee the entity's systems for monitoring risk, financial control, and compliance with the law. In some circumstances, governance practices are well developed, and those charged with governance play an active role in oversight of the entity's assessment of the risks of fraud and of the relevant internal control. Because the responsibilities of those charged with governance and management may vary by entity, it is important that the auditor ***understands*** the respective responsibilities of those charged with governance and management to enable the auditor to obtain an understanding of the oversight exercised by the appropriate individuals. **.A37** It is, therefore, important for the auditor to ***obtain an understanding*** of the controls that management has designed, implemented, and maintained to prevent and detect fraud. **.A49** When identifying and selecting journal entries and other adjustments for testing and determining the appropriate method of examining the underlying support for the items selected, the following matters may be relevant: • *The characteristics of fraudulent journal entries or other adjustments*. Inappropriate journal entries or other adjustments often have unique identifying characteristics. Such characteristics may include entries (*a*) made to ***unrelated, unusual, or seldom-used*** accounts***;*** (*b*) made by individuals who typically do not make journal entries; (*c*) recorded at the end of the period or as post closing entries that have little or no explanation or description; (*d*) made either before or during the preparation of the financial statements that do not have account numbers; or (*e*) containing round numbers or consistent ending numbers.

	• *The nature and complexity of the accounts*. Inappropriate journal entries or adjustments may be applied to accounts that (*a*) contain transactions that are ***complex or unusual*** in nature, (*b*) contain significant estimates and period-end adjustments, (*c*) have been prone to misstatements in the past, (*d*) have not been reconciled on a timely basis or contain unreconciled differences, (*e*) contain intercompany transactions, or (*f*) are otherwise associated with an identified risk of material misstatement due to fraud. In audits of entities that have several locations or components, consideration is given to the need to select journal entries from multiple locations.
AU-C sec. 315 Understanding the Entity and Its environment and Assessing the Risks of Material Misstatement	**.A1** Obtaining an understanding of the entity and its environment, including the entity's internal control (referred to hereafter as an *understanding of the entity*), is a continuous, dynamic process of gathering, updating, and analyzing information throughout the audit. The understanding of the entity establishes a frame of reference within which the auditor plans the audit and exercises professional judgment throughout the audit when, for example • identifying areas for which special audit consideration may be necessary (for example, related party transactions, the appropriateness of management's use of the going concern assumption, considering the business purpose of transactions, or the existence of complex and ***unusual*** transactions); **.A7** Analytical procedures performed as risk assessment procedures may identify aspects of the entity of which the auditor was ***unaware*** and may assist in assessing the risks of material misstatement in order to provide a basis for designing and implementing responses to the assessed risks. **.A8** Analytical procedures may enhance the auditor's understanding of the client's business and the significant transactions and events that have occurred since the prior audit and also may help to identify the existence of ***unusual*** transactions or events and amounts, ratios, and ***trends*** that might indicate matters that have audit implications. ***Unusual or unexpected*** relationships that are identified may assist the auditor in identifying risks of material misstatement, especially risks of material misstatement due to fraud.

AU-C sec. 330 Performing Audit Procedures in Response to Assessed Risks and Evaluating the Audit Evidence Obtained	**.A73** An audit of financial statements is accumulative and iterative process. As the auditor performs planned audit procedures, the audit evidence obtained may cause the auditor to modify the nature, timing, or extent of other planned audit procedures. Information may come to the auditor's attention that differs significantly from the information on which the risk assessments were based. For example • analytical procedures performed at the overall review stage of the audit may indicate a ***previously unrecognized risk*** of material misstatement. In such circumstances, the auditor may need to reevaluate the planned audit procedures, based on the revised consideration of assessed risks for all or some of the classes of transactions, account balances, or disclosures and related assertions.
AU-C sec. 500 Audit Evidence	**. A22** *Scanning* is a type of analytical procedure involving the auditor's exercise of professional judgment to review accounting data to identify significant or **unusual** items to test. This may include the identification of **unusual** individual items within account balances or other data through the reading or analysis of, for example, entries in transaction listings, subsidiary ledgers, general ledger control accounts, adjusting entries, suspense accounts, reconciliations, and other detailed reports. Scanning may include searching for large or **unusual** items in the accounting records (for example, **nonstandard** journal entries), as well as in transaction data (for example, suspense accounts and adjusting journal entries) for indications of misstatements that have occurred.
AU-C sec. 520 Analytical Procedures	**.A8** The expected effectiveness and efficiency of a substantive analytical procedure in addressing risks of material misstatement depends on, among other things, (*a*) the nature of the assertion, (*b*) the plausibility and predictability of the relationship, (*c*) the availability and ***reliability*** of the data used to develop the expectation, and (*d*) the ***precision*** of the expectation. **.A25** A wide variety of analytical procedures may be used when forming an overall conclusion. These procedures may include reading the financial statements and considering (*a*) the adequacy of the evidence gathered in response to ***unusual or unexpected*** balances identified during the course of the audit and (*b*) ***unusual***

	or unexpected balances or relationships that ***were not previously identified***. Results of these analytical procedures may indicate that additional evidence is needed. **.A26** The results of analytical procedures designed and performed in accordance with paragraph .06 may identify a previously ***unrecognized risk*** of material misstatement. In such circumstances, section 315 requires the auditor to revise the auditor's assessment of the risks of material misstatement and modify the further planned audit procedures accordingly.
AU-C sec. 550 Related Parties	**.A20** *Considerations specific to smaller entities*. Control activities in smaller entities are likely to be less formal, and smaller entities may have no documented processes for dealing with related party relationships and transactions. An owner-manager may mitigate some of the risks arising from related party transactions or potentially increase those risks through active involvement in all the main aspects of the transactions. For such entities, the auditor may obtain an ***understanding*** of the related party relationships and transactions, and any controls that may exist over these, through inquiry of management combined with other procedures, such as observation of management's oversight and review activities and inspection of available relevant documentation. **.A33** In the presence of other risk factors, the existence of a related party with dominant influence may indicate significant risks of material misstatement due to fraud. For example • an ***unusually high turnover*** of senior management or professional advisors may suggest unethical or fraudulent business practices that serve the related party's purposes. • the use of business intermediaries for significant transactions for which there appears to be no clear business justification may suggest that the related party could have an interest in such trans- actions through control of such intermediaries for fraudulent purposes. **.A41** In evaluating the business rationale of a significant related party transaction outside the entity's normal course of business, the auditor may consider the following: • Whether the transaction

	— ***has unusual terms of trade***, such as unusual prices, interest rates, guarantees, and repayment terms — involves previously ***unidentified*** related parties — is processed in an ***unusual*** manner
AU-C sec. 560 Subsequent Events and Subsequently Discovered Facts	**.09** The auditor should perform audit procedures designed to obtain sufficient appropriate audit evidence that all subsequent events that require adjustment of, or disclosure in, the financial statements have been ***identified***. The auditor is not, however, expected to perform additional audit procedures on matters to which previously applied audit procedures have provided satisfactory conclusions. (Ref: par. .A2–.A3)

Appendix B: Potential application areas of EDA in International Standards for the professional Practice of Internal Auditing issued by IIA

2010 **Planning**	**2010.A1 –** The internal audit activity's plan of engagements must be based on a documented ***risk assessment***, undertaken at least annually. The input of senior management and the board must be considered in this process.
2120 **Risk Management**	**2120.A2** – The internal audit activity must evaluate the ***potential for the occurrence of fraud*** and how the organization manages fraud risk. **2120.C1** – During consulting engagements, internal auditors must ***address risk*** consistent with the engagement's objectives and be alert to the existence of other significant risks. **2120.C2** – Internal auditors must incorporate knowledge of risks gained from consulting engagements into their ***evaluation of the organization's risk management processes***. **2120.C3** – When assisting management in ***establishing or improving risk management processes***, internal auditors must ***refrain from assuming any management responsibility*** by actually managing risks.
2200 **Engagement Planning**	**2210.A1** – Internal auditors must conduct a ***preliminary assessment of the risks*** relevant to the activity under review. Engagement objectives must reflect the results of this assessment.
2400 **Disseminating Results**	**2440.A2 –** If not otherwise mandated by legal, statutory, or regulatory requirements, prior to releasing results to parties outside the organization the chief audit executive must: • ***Assess the potential risk*** to the organization;

Appendix C: Usable fields in 2010 Inpatient Medicare Claim Data

Field Name	Description
CLAIM_NO	LDS Claim Number
DESY_SORT_KEY	LDS Beneficiary Identifier
BENE_RSDNC_SSA_STD_STATE_CD	State Code from Claim (SSA)
CLM_THRU_DT	Claim Through Date (Determines Year of Claim)
CLM_QUERY_CD	Claim Query Code
PRVDR_NUM	Provider Number
CLM_TOT_SGMT_CNT	Claim Total Segment Count
CLM_SGMT_NUM	Claim Segment Number
CLM_FREQ_CD	Claim Frequency Code
BENE_RSDNC_SSA_STD_CNTY_CD	County Code from Claim (SSA)
FI_NUM	FI Number
BENE_SEX_IDENT_CD	Gender Code from Claim
BENE_RACE_CD	Race Code from Claim
BENE_BIRTH_DT	Beneficiary Birth Date
CWF_BENE_MDCR_STUS_CD	Beneficiary Medicare Status Code
CLM_PRNCPAL_DGNS_CD	Claim Principal Diagnosis Code
CLM_PMT_AMT	Claim Payment Amt
NCH_PRMRY_PYR_CLM_PD_AMT	NCH Primary Payer Claim Paid Amt
NCH_PRMRY_PYR_CD	NCH Primary Payer Code
FI_CLM_ACTN_CD	FI Claim Action Code
NCH_PRVDR_STATE_CD	NCH Provider State Code
PTNT_DSCHRG_STUS_CD	Patient Discharge Status Code
CLM_PPS_IND_CD	Claim PPS Indicator Code
CLM_TOT_CHRG_AMT	Claim Total Charge Amt
IP_CLM_DGNS_CD_CNT	Inpatient/SNF Claim Diagnosis Code Count
IP_CLM_PRCDR_CD_CNT	Inpatient/SNF Claim Procedure Code Count
IP_CLM_RLT_COND_CD_CNT	Inpatient/SNF Claim Related Condition Code Count
IP_CLM_RLT_OCRNC_CD_CNT	Inpatient/SNF Claim Related Occurrence Code Count
IP_CLM_VAL_CD_CNT	Inpatient/SNF Claim Value Code Count
IP_REV_CNTR_CD_I_CNT	Inpatient/SNF Revenue Center Code Indicator Count
CLM_ADMSN_DT	Claim Admission Date
CLM_SRC_IP_ADMSN_CD	Claim Source Inpatient Admission Code
CLM_ADMTG_DGNS_CD	Claim Admitting Diagnosis Code
NCH_PTNT_STUS_IND_CD	NCH Patient Status Indicator Code
CLM_PASS_THRU_PER_DIEM_AMT	Claim Pass Thru Per Diem Amt
NCH_BENE_IP_DDCTBL_AMT	NCH Beneficiary Inpatient Deductible Amt
CLM_TOT_PPS_CPTL_AMT	Claim Total PPS Capital Amt

CLM_PPS_CPTL_FSP_AMT	Claim PPS Capital FSP Amt
CLM_PPS_CPTL_OUTLIER_AMT	Claim PPS Capital Outlier Amt
CLM_PPS_CPTL_DSPRPRTNT_SHR_AM	Claim PPS Capital Disproportionate Share Amt
CLM_PPS_CPTL_IME_AMT	Claim PPS Capital IME Amt
CLM_PPS_CPTL_DRG_WT_NUM	Claim PPS Capital DRG Weight Number
CLM_UTLZTN_DAY_CNT	Claim Utilization Day Count
BENE_TOT_COINSRNC_DAY_CNT	Beneficiary Total Coinsurance Days Count
BENE_LRD_USE_CNT	Beneficiary LRD Used Count
CLM_NUTLZTN_DAY_CNT	Claim Non Utilization Days Count
NCH_BLOOD_PT_FRNSH_QTY	NCH Blood Pints Furnished Quantity
NCH_BENE_DSCHRG_DT	NCH Beneficiary Discharge Date
CLM_DRG_CD	Claim Diagnosis Related Group Code
CLM_DRG_OUTLIER_STAY_CD	Claim Diagnosis Related Group Outlier Stay Code
NCH_DRG_OUTLIER_APRV_PMT_AMT	NCH DRG Outlier Approved Payment Amt
CLM_ATNDG_PHYSN_NPI_NUM	Claim Attending Physician NPI Number
CLM_OPRTG_PHYSN_NPI_NUM	Claim Operating Physician NPI Number
CLM_OTHR_PHYSN_NPI_NUM	Claim Other Physician NPI Number
ORG_NPI_NUM	Organization NPI Number
CLM_DGNS_CD{1-10}	Claim Diagnosis Code 1-10
CLM_POA_IND_SW{1-10}	Claim Present on Admission Code 1-10
CLM_PRCDR_CD{1-6}	Claim Procedure Code 1-6
CLM_PRCDR_PRFRM_DT{1-6}	Claim Procedure Performed Date 1-6
CLM_RLT_COND_CD{1-4}	Claim Related Condition Code 1-4
CLM_RLT_OCRNC_CD{1-5}	Claim Related Occurrence Code 1-5
CLM_RLT_OCRNC_DT{1-5}	Claim Related Occurrence Date 1-5
CLM_VAL_CD{1-6}	Claim Value Code 1-6
CLM_VAL_AMT{1-6}	Claim Value Amount 1-6
REV_CNTR_CD{1-30}	Revenue Center Code 1-30
REV_CNTR_UNIT_CNT{1-29}	Revenue Center Unit Count 1-29
REV_CNTR_RATE_AMT{1-29}	Revenue Center Rate Amount 1-29
REV_CNTR_TOT_CHRG_AMT{1-29}	Revenue Center Total Charge Amount 1-29
REV_CNTR_NCVR_CHRG_AMT{1-29}	Revenue Center Non-Covered Charge Amount 1-29

Appendix D: Association Rules Generated from Medicare database[1]

Generated Association Rules when S_{min}=1% and C_{min}=80%

Consequent	Antecedent	Support %	Confidence %
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD2 = 3607	1.288	99.868
CLM_PRCDR_CD1 = 8154	CLM_DGNS_CD1 = 71536	1.665	97.29
CLM_PRCDR_CD2 = 8856	CLM_PRCDR_CD3 = 8853	1.034	91.579
CLM_DGNS_CD2 = 5856	CLM_DGNS_CD3 = 40391	1.513	88.904
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD3 = 3722	1.027	87.943

[1] For better presentation, the diagnoses and procedures are shown in ICD9 code. The coding tables can be downloaded at http://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/codes.html

Generated Association Rules when S_{min}=0.5% and C_{min}=80%

Consequent	Antecedent	Support %	Confidence %
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.538	99.993
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD2 = 3607	0.657	99.988
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.573	99.959
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD2 = 3607 and CLM_DGNS_CD1 = 41401	0.672	99.935
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD2 = 3607	1.288	99.868
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD1 = 0066	0.584	98.87
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722	0.647	98.487
CLM_PRCDR_CD1 = 8154	CLM_DGNS_CD1 = 71536	1.665	97.29
CLM_PRCDR_CD2 = 8856	CLM_PRCDR_CD3 = 8853 and CLM_PRCDR_CD1 = 3722	0.869	96.29
CLM_PRCDR_CD1 = 8151	CLM_DGNS_CD1 = 71535	0.643	95.645
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD1 = 0066	0.605	95.494
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD1 = 0066	0.638	94.693
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.573	93.919
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.573	93.888
CLM_DGNS_CD2 = 5856	CLM_DGNS_CD3 = 40391 and CLM_PRCDR_CD1 = 3995	0.658	93.414
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD1 = 0066	0.8	92.259
CLM_PRCDR_CD2 = 8856	CLM_PRCDR_CD3 = 8853	1.034	91.579
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD1 = 0066	0.638	91.461
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856	0.637	90.576
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856	0.817	90.347

CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853	0.844	90.23
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722	0.647	90.226
CLM_DGNS_CD2 = 5856	CLM_DGNS_CD3 = 40391	1.513	88.904
CLM_PRCDR_CD1 = 3722	CLM_PRCDR_CD3 = 8853 and CLM_PRCDR_CD2 = 8856	0.947	88.4
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD3 = 3722	1.027	87.943
CLM_PRCDR_CD1 = 9339	CLM_PRCDR_CD2 = 9383	0.563	87.167

Generated Association Rules when S_{min}=0.25% and C_{min}=80%

Consequent	Antecedent	Support %	Confidence %
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607 and CLM_DGNS_CD1 = 41401	0.271	100
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD2 = 3607 and CLM_DGNS_CD1 = 41401	0.333	99.998
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD2 = 3607	0.25	99.997
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD2 = 3607	0.431	99.994
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.426	99.994
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD2 = 3607	0.281	99.994
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD2 = 3607	0.265	99.994
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.259	99.994
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.259	99.994
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.538	99.993
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.271	99.991

CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.445	99.991
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD2 = 3607	0.657	99.988
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD2 = 3607	0.466	99.979
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607 and CLM_DGNS_CD1 = 41401	0.29	99.975
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.573	99.959
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD2 = 3607	0.308	99.951
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD2 = 3607 and CLM_DGNS_CD1 = 41401	0.672	99.935
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722	0.342	99.915
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856	0.34	99.915
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD2 = 3607 and CLM_DGNS_CD1 = 41071	0.271	99.87
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD2 = 3607	1.288	99.868
CLM_PRCDR_CD2 = 3783	CLM_PRCDR_CD1 = 3772	0.451	99.785
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856	0.355	99.767
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722	0.385	99.714
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD4 = 8856	0.354	99.665
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD2 = 3606	0.426	99.644
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853	0.364	99.563
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722	0.342	99.413

CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD1 = 0066	0.342	99.412
CLM_PRCDR_CD1 = 8154	CLM_DGNS_CD1 = 71536 and CLM_DGNS_CD2 = 2851	0.294	99.399
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD2 = 3607	0.25	99.394
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.25	99.394
CLM_PRCDR_CD1 = 3812	CLM_PRCDR_CD2 = 0040 and CLM_DGNS_CD1 = 43310	0.325	99.387
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD4 = 8856	0.375	99.289
CLM_PRCDR_CD1 = 8154	CLM_DGNS_CD1 = 71536 and CLM_DGNS_CD2 = 4019	0.27	99.066
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD2 = 3607	0.431	98.897
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.431	98.897
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD1 = 0066	0.584	98.87
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD1 = 0066 and CLM_DGNS_CD1 = 41401	0.273	98.739
CLM_PRCDR_CD1 = 8154	CLM_DGNS_CD1 = 71536 and CLM_DGNS_CD3 = 4019	0.301	98.615
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722	0.647	98.487
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_DGNS_CD1 = 41401	0.281	98.447

CLM_PRCDR_CD1 = 3772	CLM_PRCDR_CD2 = 3783	0.458	98.143
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD2 = 3607	0.265	97.757
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.265	97.757
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD1 = 0066	0.362	97.576
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853	0.364	97.476
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856 and CLM_DGNS_CD1 = 41401	0.277	97.374
CLM_PRCDR_CD1 = 8154	CLM_DGNS_CD1 = 71536	1.665	97.29
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_DGNS_CD1 = 41401	0.281	97.086
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856 and CLM_DGNS_CD1 = 41401	0.349	97.056
CLM_PRCDR_CD1 = 8154	CLM_DGNS_CD1 = 71596	0.459	96.455
CLM_PRCDR_CD2 = 8856	CLM_PRCDR_CD3 = 8853 and CLM_PRCDR_CD1 = 3722	0.869	96.29
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.259	96.27
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.259	96.267
CLM_PRCDR_CD5 = 8853	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.259	96.145

CLM_PRCDR_CD5 = 8853	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.259	96.142
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD1 = 0066	0.353	96.102
CLM_PRCDR_CD5 = 8853	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD1 = 0066	0.354	95.936
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.445	95.921
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.445	95.918
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD4 = 8856	0.354	95.861
CLM_PRCDR_CD5 = 8853	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856	0.355	95.794
CLM_PRCDR_CD1 = 8151	CLM_DGNS_CD1 = 71535	0.643	95.645
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD1 = 0066	0.605	95.494
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.271	95.439
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.271	95.436
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.466	95.33
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD2 = 3607	0.466	95.319

CLM_PRCDR_CD5 = 8853	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.271	95.314
CLM_PRCDR_CD5 = 8853	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.271	95.312
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD3 = 3722 and CLM_DGNS_CD1 = 41401	0.438	95.255
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD1 = 0066	0.372	95.056
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD1 = 0066 and CLM_DGNS_CD1 = 41401	0.299	95.029
CLM_PRCDR_CD5 = 8853	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD1 = 0066	0.372	94.892
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD1 = 0066 and CLM_DGNS_CD1 = 41401	0.284	94.884
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD1 = 0066	0.638	94.693
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.265	94.684
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD2 = 3607	0.265	94.682
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD4 = 8856	0.375	94.601
CLM_PRCDR_CD5 = 8853	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD4 = 8856	0.375	94.535
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD1 = 0066	0.362	94.326
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD5 = 8853	0.364	93.993

CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.573	93.919
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607	0.573	93.888
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066 and CLM_DGNS_CD1 = 41401	0.29	93.536
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD2 = 3607 and CLM_DGNS_CD1 = 41401	0.29	93.513
CLM_DGNS_CD2 = 5856	CLM_DGNS_CD3 = 40391 and CLM_PRCDR_CD1 = 3995	0.658	93.414
CLM_PRCDR_CD2 = 8856	CLM_PRCDR_CD3 = 8853 and CLM_DGNS_CD1 = 41401	0.264	93.339
CLM_PRCDR_CD2 = 9383	CLM_PRCDR_CD1 = 9339 and CLM_DGNS_CD1 = V5789	0.454	92.982
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.466	92.461
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD2 = 3607	0.466	92.447
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD1 = 0066	0.8	92.259
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD1 = 0066	0.384	92.179
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD4 = 8856 and CLM_PRCDR_CD1 = 0066 and CLM_DGNS_CD1 = 41401	0.368	92.136
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722	0.385	92.131
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD2 = 3607	0.281	92.123
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.281	92.123
CLM_DGNS_CD1 = 43310	CLM_PRCDR_CD1 = 3812 and CLM_PRCDR_CD2 = 0040	0.351	92.091

CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD6 = 0045	0.467	91.678
CLM_PRCDR_CD2 = 8856	CLM_PRCDR_CD3 = 8853	1.034	91.579
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD1 = 0066	0.638	91.461
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD1 = 0066 and CLM_DGNS_CD1 = 41401	0.299	91.319
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.308	91.174
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD2 = 3607	0.308	91.134
CLM_DGNS_CD1 = 43310	CLM_PRCDR_CD1 = 3812	0.46	91.05
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853 and CLM_DGNS_CD1 = 41401	0.368	91.029
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856	0.637	90.576
CLM_PRCDR_CD1 = 3722	CLM_PRCDR_CD3 = 8853 and CLM_DGNS_CD1 = 41401	0.264	90.425
CLM_PRCDR_CD1 = 9339	CLM_PRCDR_CD2 = 9383 and CLM_DGNS_CD1 = V5789	0.467	90.399
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD4 = 8856	0.817	90.347
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD5 = 8853	0.844	90.23
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD5 = 8853 and CLM_PRCDR_CD3 = 3722	0.647	90.226
CLM_DGNS_CD2 = 5856	CLM_PRCDR_CD2 = 3995 and CLM_DGNS_CD3 = 40391	0.26	89.744
CLM_PRCDR_CD3 = 3722	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD1 = 0066	0.428	89.609
CLM_PRCDR_CD1 = 3722	CLM_PRCDR_CD2 = 8856 and CLM_DGNS_CD1 = 41401	0.311	89.173
CLM_PRCDR_CD5 = 8853	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.281	89.111
CLM_PRCDR_CD5 = 8853	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD2 = 3607	0.281	89.109

CLM_PRCDR_CD5 = 8853	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722 and CLM_PRCDR_CD1 = 0066	0.384	88.956
CLM_DGNS_CD2 = 5856	CLM_DGNS_CD3 = 40391	1.513	88.904
CLM_PRCDR_CD5 = 8853	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD3 = 3722	0.385	88.777
CLM_PRCDR_CD1 = 3722	CLM_PRCDR_CD2 = 8853 and CLM_PRCDR_CD3 = 8856	0.282	88.611
CLM_PRCDR_CD1 = 3722	CLM_PRCDR_CD3 = 8853 and CLM_PRCDR_CD2 = 8856	0.947	88.4
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.308	88.007
CLM_PRCDR_CD1 = 5123	CLM_PRCDR_CD2 = 8753	0.267	87.973
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD2 = 3607	0.308	87.971
CLM_PRCDR_CD1 = 0066	CLM_PRCDR_CD3 = 3722	1.027	87.943
CLM_PRCDR_CD4 = 0040	CLM_PRCDR_CD3 = 0045	0.338	87.728
CLM_PRCDR_CD3 = 8856	CLM_PRCDR_CD2 = 8853 and CLM_PRCDR_CD1 = 3722	0.287	87.176
CLM_PRCDR_CD1 = 9339	CLM_PRCDR_CD2 = 9383	0.563	87.167
CLM_PRCDR_CD4 = 8856	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD1 = 0066	0.428	86.896
CLM_PRCDR_CD1 = 3722	CLM_PRCDR_CD2 = 8853	0.331	86.653
CLM_DGNS_CD1 = V5789	CLM_PRCDR_CD2 = 9383 and CLM_PRCDR_CD1 = 9339	0.491	86.062
CLM_PRCDR_CD5 = 8853	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD2 = 3607 and CLM_PRCDR_CD1 = 0066	0.308	85.808
CLM_PRCDR_CD5 = 8853	CLM_PRCDR_CD6 = 0045 and CLM_PRCDR_CD2 = 3607	0.308	85.771
CLM_PRCDR_CD2 = 8853	CLM_PRCDR_CD3 = 8856 and CLM_PRCDR_CD1 = 3722	0.293	85.259
CLM_PRCDR_CD3 = 8856	CLM_PRCDR_CD2 = 8853	0.331	85.249

VITA

Qi Liu

1985.4	Born in Wuhan, Hubei Province, China.
2003	Graduated from Shuiguohu High School, Wuhan, Hubei Province
2003-2007	B.S., Electronic Commerce. Wuhan University, China
2004-2007	B.S., Law. Wuhan University, China
2007-2009	M.S., Management, Conservatoire National des Arts et Metiers, France
2009-2014	Ph.D. in Management (Accounting Major), Rutgers University

Publications

2012	R. P. Srivastava, and Q. Liu. 2012. Editor's Note on the Special Issue of JIS on XBRL, Journal of Information Systems, Vol. 26 (1), pp. 97-101
2014	Q. Liu, and M. A. Vasarhelyi. 2014. Big Questions in AIS Research: Measurement, Information Processing, Data Analysis, and Reporting. Journal of Information Systems, Vol. 28 (1), pp. 1-17
	G. Gray, V. Chiu, Q. Liu, and P. Li. 2014. The Expert Systems Life Cycle in AIS Research: What does it mean for future AIS research? International Journal of Accounting Information Systems. Available online 4 July 2014
	V. Chiu, Q. Liu, and M. A. Vasarhelyi. 2014. The Development and Intellectual Structure of Continuous Auditing Research. Journal of Accounting Literature. Available online 12 August 2014