Auditing in Environments of Diverse Data

By Basma Moharram

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

Dr. Miklos A. Vasarhelyi

Approved by

Dr. Miklos A. Vasarhelyi

_____

Dr. Alexander Kogan

_____

Dr. Dan Palmon

_____

Dr. JP Krahel

Loyola University, Maryland

_____

Newark, New Jersey

October 2016

ABSTRACT OF THE DISSERTATION

Auditing in Environments of Diverse Data

By BASMA MOHARRAM

Dissertation Director:

Dr. Miklos A. Vasarhelyi

This dissertation has three essays. The first essay examines several analytical models that can be used in the insurance industry. We develop a continuous auditing framework for the insurance industry's claims payment process. We propose analytical models to detect anomalies in the already paid claims and in the premium calculations.

The second essay proposes a framework for developing the Audit Data Standards (ADSs) for specific industries. We apply the framework on the insurance industry. In specific, we apply it on claim payment business cycles of the life / disability insurance industry. We then use these ADSs to develop an interactive auditor dashboard.

The third essay examines a possible way for the auditor to overcome data limitations. We examine the added value of using the macroeconomic indicators to improve the prediction and error detection performance of the statistical models used by the auditor through two research questions. our first research question

investigated the effect of using macroeconomic indicators in the prediction models by comparing the Mean Absolute Percentage Error (MAPE) with and without the use of the macroeconomic indicators. our second research question investigated the effect of using macroeconomic indicators in the prediction models when the independent variables contain undetected errors. We tested each research question in three different situations; the macroeconomic variable is used individually, collectively, or with the peer data. our results came in favor of the macroeconomic indicator's use, specifically when used along with the peer data.

Preface

Acknowledgements

I would like to gratefully acknowledge the contributions and guidance of the members of my dissertation committee. I would also like to thank my family for their love and support.

Table of Contents

List of Tables

List of Illustrations

Chapter One: INTRODUCTION

Recently, there has been much pressure exerted over the auditing profession to improve and to keep up with the current technology and the changes in the investors' needs. Advocates for continuous auditing and continuous assurance emphasize that the archival audit where the auditor go to the audited entity at the end of the year to examine financial statements and issue an ex-post opinion on these statements should, and will, be replaced by a more timely assurance services (Alles et al. 2006; Alles, Kogan, and Vasarhelyi 2002; Vasarhelyi, Alles, and Kogan 2004). The idea of the continuous auditing and assurance has emerged, in part, from the availability of the required technology. Thanks to the current technology, companies are able to collect transactional data almost instantly making the availability of the data almost continuous (Alles et al. 2002).

In our first essay we propose a continuous auditing framework on an international insurance company. We specifically use the life and disability insurance. We suggest different methodologies to help the auditors better perform their tasks. However, dealing with such data we faced a lot of difficulties with the data needed for our work. We realized that the auditor's problem is not the data availability, it's is the data's accessibility.

Auditors face many challenges in accessing the data they need to fulfill their duties and form their opinion even when their clients are fully digitalized and technologically capable of providing the needed data. Zhang et al (2012) state that without open access to data, innovative audit tools and techniques might be

disregarded. Researchers argued that there is a disparate need to standardize the data that should be available to the auditor (Moffitt and Vasarhelyi 2013; Vasarhelyi 2013; Zhang et al. 2012). The standardized data should facilitate the auditor's work by giving him access to the data he needs and also by paving the road to the standardized audit applications. Efforts toward issuing data standards have been already launched. The first data standards to be issued were standards for general accounts that most if not all the companies have in common like the General ledger and Accounts Receivables (Committee 2013a, 2013b; Zhang et al. 2012)

In our second essay, we propose framework that provide a structure for the process of generating the audit data standards for specific industries. We follow our first essay by applying this framework on the insurance industry. We specifically work on life and disability insurance and we generate a proposed set of audit data standards. We then use the proposed audit data standard in an interactive audit dashboard to help the auditor in performing his audit.

Up till now the data standards sill face other challenges. One of the main challenges is the enforcement of such standards. So far the standards are being issued as recommendations that are not enforced by the FASB. The hope is that in time as professionals grow more accustomed to the idea and as the standards are better established there will be some kind of enforcement.

While the data standards might take some time before its auditors are able to benefit from them, researchers are trying to find other ways to enhance the performance of their analytical procedures even with the restricted access to clients' data. Analytical reviews are supposed to help the auditor reach a

reasonable expectations of what the account balances should be so he can determine the extent to which the actual balances deviate from these expectations (Lev 1980). The main stream of research focused on improving the performance of statistical models used in analytical procedures to predict the account balances and detect any discrepancies. Since firms do not operate in vacuum, it is expected that economy wide factors and industry wide factors affect the operations of companies operating in such economy and in the specific industry (Lev 1980). One of the research studies that aimed to improve the performance of the statistical models used in analytical procedures proposed the use of Gross National Product "GNP" and Total Corporate Profits after tax "TCP" to improve the prediction models of a company's sales, operating income and net income (Lev 1980). Another study proposed the use of data from peer companies to improve the prediction and error detection performance of statistical models (Hoitash, Kogan, and Vasarhelyi 2006).

In our third essay, we propose using macro-economic indicators to improve the prediction and error detection performance of the statistical models. We add to Lev's 2008 research the use of monthly data instead of the annual data and the use of multiple macro indicators. We also test the effectiveness of the macroeconomic indicators in detecting coordinated errors and in mitigating the effects of misstated accounts.

Chapter Two: Weighted Multi-Dimensional Approaches to Anomaly Detection: A Study of Insurance Claims

Introduction

The advent of big data and corresponding increases in the affordability and effectiveness of data processing and storage have made automated anomaly detection both more feasible (Chandola, Banerjee, and Kumar 2009; Patcha and Park 2007a, 2007b) and more necessary (Chen, Chiang, and Storey 2012). As more and more data types become available in greater volumes and at greater velocity, the use of manual detection methods will become increasingly costly, imprecise, and non-representative, making automated anomaly detection an increasingly attractive prospect. The development of automated anomaly detection methods has been addressed in a variety of literature streams, notably health care (Campbell and Bennett 2001; Solberg and Lahti 2005; Wong et al. 2003), intrusion detection (Eskin 2000; Hu, Liao, and Vemuri 2003; Lee, Stolfo, and Mok 2000), and credit card fraud detection (Bolton and Hand 2002; Ghosh and Reilly 1994).

Recent increases in the prevalence of insurance fraud (Major and Riedinger 2002) make its analysis particularly fertile ground for quantitative anomaly detection. The National Healthcare Anti-Fraud Association (NHCAA) estimates that 10 percent of all insurance claims contain some elements of fraud (Major and Riedinger 2002) Traditional methods of claim fraud detection involved manually checking documents for abnormalities. Examples of abnormalities would be cross-outs,

similar handwriting of provider and patient, photocopied bills instead of originals (Major and Riedinger 2002). The process was time consuming and difficult to automate. With the use of electronic systems and since the fraud detection problems involve huge data sets, researchers were motivated to search for an electronic fraud detection system for insurance claims (Hodge 2001; Major and Riedinger 2002; Viaene et al. 2007; Viaene, Dedene, and Derrig 2005).

This research paper looks into identifying anomalies in life insurance data. We focus on two business cycles; the Claim payment cycle, and the Revenue cycle. We cannot call these anomalies “fraudulent” just yet. We just try to detect these claims that looks suspicious enough to require further investigation. Many research studies were concerned with the outliers or anomalies detection (Bakar et al. 2006; Breunig et al. 2000; Hawkins et al. 2002; Knorr, Ng, and Tucakov 2000; Williams and Baxter 2002; Yu, Sheikholeslami, and Zhang 2002). Their research was mainly focused on how to “calculate” the outliers. Some of them used distance based outlier calculations (Knorr et al. 2000). One study (Yu et al. 2002) introduced a new way to calculate outliers. They called it FindOut which is mainly based on removing the clusters from the original data and then find the outliers. Another study (Breunig et al. 2000) worked on density based calculations of outliers. This research focuses on something different. We don’t focus on how to calculate the outliers; we rather focus on which attributes should we use to calculate the outliers. The choice of the attributes should make sense to the problem in hand. And the problem in our hands are detecting anomalies of

Life/Disability Insurance Claims in a way that help us test two different audit assertions; is the claim settlement reasonable? Is the claim itself legitimate?

We start by proposing a continuous auditing framework for the claim payment cycle of life insurance. One of the steps in this frame work is to detect claim anomalies. We then follow by proposing analytical procedures to help the auditor detect claims anomalies in testing the above mentioned audit assertions. To do this we use claims data provided to us by a leading international insurance company. We use a weighted multi-dimensional approach in which we divide the attributes we have into different groups (dimensions). We use each dimension to logically find insurance claim anomalies. then we prioritize the anomalies based on the weighted average of the dimensions that identified this record as an anomaly. As an additional way of prioritizing the outliers, we use the belief function to give a “Risk Score” to the different branches within the insurance company. The anomalies detected will then be weighted by the risk score of the company branch that generated it.

Our last contribution in this paper is a model for detecting premium outliers.

Related Literature

Insurance Outlier Detection

One of the earliest interesting studies on detecting fraud in insurance was a study published by (Artís, Ayuso, and Guillen 1999). In their study they built a discrete logit model for fraud behavior based on classified automobile insurance claims

data (Fraudulent, or not). Their model is based on the utility function and that the individual's behavior aims at maximizing this utility function. The utility function takes into consideration the benefit of committing the fraud (times the probability of not being detected) and the amount of punishment (times the probability of being detected). In their Discrete-Choice Models (Artís, Ayuso, and Guillén 2002) classify the claims as either Legitimate, Fraud for benefit of self, fraud for benefit of others. Another study focused on comparing the performance of certain insurance provider with his peers to identify fraud of the health providers. In their research study (Major and Riedinger 2002) they developed an "Electronic Fraud Detection – EFD" system to be used by the investigative consultant reviewing claims issued by health providers. The EFD provide a "Multiple Frontier Summary" report for the consultant. This report lists the unusual providers identified by the system. For each provider, the investigative consultant can call up a frontier summary report (shown in the next graph) with the details of the provider and the behavioral patterns for which this provider is identified as an unusual one.

Copyright 1990, The Travelers EFD February 9, 1990

---------- Provider Identification Section ----------

Name: ;NXXXXXXXXXXX CENTER

TIN: 2-636009999-001 Type: CHIROPRACTOR Org: ASSOCIATION
Address: 825 XXXXX Avenue, Anytown, USA 12345
Total Paid: $33,728.27 Span of Services: 3 JAN 88 - 6 FEB 89 Total Frontier: 914

------ Behavior Pattern Section ------

Group Pattern	Mean	Peers	#Ent	Frontier
FINANCIAL:XXXXXXXXXX	157.133	117.647	347	2NDORDER
FINANCIAL:ASSIGNMENTS/NUMBER	0.798	0.720	435	2NDORDER
FINANCIAL:XXXXXXXXXXXXXX	18.391	15.723	470	2NDORDER
LOGISTICS:XXXXXXXXX	0.107	0.090	675	1STORDER
MEDICAL:XXXXXXXXXXX	0.617	0.531	47	1STORDER
MEDICAL:ODD SERVICE VS PRVTYPE	0.257	0.098	501	1STORDER

Figure 1:Major & Riedinger 2002's' Summary report

The summary report can show that a certain provider has “Odd Service” of about 25% of the services he provided while the mean of the whole peer group is only 10%. The odd service is a service not expected from this type of providers. In this case, the consultant would have to decide whether to refer this provider to a further investigation or not. Their EFD architecture consisted of five layers. The first layer is the behavioral heuristic measurement with 27 heuristics in five categories. The second layer is the information, frontier and rules layer. This layer compares measurements among the providers and flags those who are out of line relative to their peer group. The third layer is the data exploration one. In this layer the EFD can supply the extracted claim records to the consultant. The forth layer is the decision and action layer. In this layer, if the consultant decided to further investigate the case, the system issue a memo translating the frontier summary report into business oriented terms along with the consultant’s recommendations and send it to the appropriate regional investigator. The fifth and final layer is the enhancement layer. As we will discuss later in this chapter, we follow the research

by (Major and Riedinger 2002) in one of our claims anomalies detection dimensions. Another research study that focused on detecting fraud in settled insurance claims was a study by Pathak et al (Pathak, Vidyarthi, and Summers 2005). They created a fuzzy expert system to detect anomalies in the already settled claims of any type of insurance. Their model is based on three measures; Ambiguity index, degree of incomplete information of the claim, and level of discretion used by the claim settlers. For each measure they set a Law, Mid, and high levels based on the data itself and on the auditor's judgment. Then they set rules of "If … THEN" statements to place each claim in one of two classes (genuine or not). The authors did not define any variables driving the measures. We use this study's three measures in evaluating our different dimensions for detecting anomalies in claims. Another study (Yamanishi et al. 2004) has a very interesting way for making a continuous online outlier detection. They introduced a theoretical framework for an online unsupervised outlier detector called "SmartSifter". They define the "online process" as such that every time a new instance is input, it's required to evaluate its deviation from the normal pattern. They said that the concept of "online outlier detection" is in contrast to the "Batch detection process" where outliers can only be detected after seeing the entire database. Every time an instance is input, the system employs an online learning algorithm to update the model. The authors developed two different algorithms, SDLE – Sequentially Discounting Laplace Estimation and SEDM – Sequentially Discounting Expectation and Maximizing. The SDLE algorithm was developed to learn the histogram density for the categorical domain while the SEDM was developed to

learn the finite mixture model for the continuous domain. One important aspect about those two algorithms is that they gradually discount the effect of past examples in the online process. SmartSifter assigns a score to each input datum on the basis of the learned model, measuring the change to the model after learning. A high score indicates a high possibility that the datum is an outlier. The authors tested their new system using simulated data and concluded that it works better than other systems in terms of accuracy and computation time.

Belief Function

Shafer states “the theory of belief functions provides a non-Bayesian way of using mathematical probability to quantify subjective judgments” (Shafer 1996). The basic difference between probability theory and the belief function is in the assignment of uncertainties to a set of mutually exclusive and exhaustive states (referred to as a “frame”) (Srivastava 1993). The belief function bases degrees of belief or trust for one specific question on the probabilities for (a) related question(s). The belief functions measure the following; the belief which is the evidence we have that supports a specific event, the ambiguity which represent the part we don’t have any evidence for to support any outcome, and the plausibility of the event which is combined evidence and ambiguity of the event.

There are three basic functions which are important to understand the use of belief functions in decision making (Rajendra P Srivastava and Mock 2000); basic belief mass functions, belief functions, and plausibility functions.

- Basic belief mass function (m-values): If we have a decision problem with *n* possible elements forming a mutually exclusive and exhaustive set represented by $\{a_1, a_2, \ldots a_n\}$, we call this set a *frame* and we represent it by the symbol Θ. We can define the m-values as the level of support directly obtained from the evidence to support a specific element. The m-values can be assigned to all single elements, to all subsets that can be derived from the *frame*, and also to the entire *frame*. All m-values add to one.
- Belief Function: Belief on a set of elements (A) of a frame Θ is defined as the total belief on (A). This represents the sum of all the m-values assigned to the elements contained in (A) plus m-values assigned to (A).

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

Where B is any subset that belongs to A.

- Plausibility Function: The plausibility of an element or a set of elements (A) of a frame Θ is defined to be the maximum possible belief that could be assigned to (A) if all future evidence were in support of (A). In other words, it could be defined as 1 – m-values assigned to all other elements or set of elements not A (~A).
- Ambiguity Function: The ambiguity in an element (A) is defined as the difference between the plausibility of this element and the belief in it.

The belief functions have a number of features that argue for their more extensive use in auditing (Srivastava and Mock 2005). One of these features is the “rigorous definition of risk” in comparison to the probability functions. In case of complete

ignorance, an auditor who does not have any evidence on whether or not there is management fraud in the financial statements, the probability functions will assign both events (fraud, no fraud) a probability of 0.5 to represent uncertainty. The belief functions on the other hand will accurately assign a belief of zero to each event. The plausibility of each event will be assigned “1”. The difference between the plausibility of each event and its belief (1 – 0 = 1) is a rigorous measure of “Ambiguity” of the event that the probability functions cannot represent. In another situation, where the auditor has partial knowledge as opposed to complete ignorance, the authors give an example of an auditor who has 0.3 evidence that there is fraud in the statements. The probability functions will represent this as P(fraud) = 0.3 and P(no-Fraud) = 1 – 0.3 = 0.7 even though there is no evidence about the no-fraud event. The belief function on the other had will represent the risk in a more accurate way saying that Belief(fraud) = 0.3, Belief(no-Fraud) = 0, and the remaining 0.7 will be undecided. That will make the plausibility of Fraud is 1 in the belief function and probability of fraud is 0.3 in probability functions (Srivastava and Mock 2005). Researchers have argued in favor of the belief function as it gives a more accurate and conservative valuation of risk which is more suitable for the auditing (Shafer and Srivastava 1990; Srivastava and Mock 2005). Research studies have explored the applicability of using the belief function in auditing and assurance services (Lili et al. 2006; Mock et al. 2009; Shafer and Srivastava 1990; Srivastava 1993; Rajendra P. Srivastava and Mock 2000; Srivastava and Mock 2005, 2011; Srivastava, Mock, and Turner 2007; Srivastava, Rao, and Mock 2013; Srivastava and Shafer 1992). The belief function has been

also used in other fields of study such as financial portfolio management, data mining, image processing, agriculture, water treatment, and forecasting demand (Srivastava and Mock 2005).

Data

We have data from a large international insurance company. The company deals with many forms of insurance. For the purpose of this paper, we are only working on “Life/Disability” insurance claims.

We have 2,763,591 unique policies with effective dates ranging from November 1998 till January 2014. We have 28,490 records of Life/Disability Insurance claims paid in the period between January 2013 and July 2014. Of these records, 82% (23,392) are related to “Individual” Life/Disability Insurance and 18% (5,098 records) are related to “Group” insurance. In the “Individual” insurance settings, the client is an individual person who bought an insurance policy against his/her life or, in some cases, against someone else’s life. On the other hand, a “Group” insurance policy is bought by an organization against the life (or disability) of its members or employees. When dealing with “Individual” policies, the insurance company has direct contact with the person it is issuing the policy against his/her life. In this case, the insurance company can collect all the personal information it needs to make the decision of whether or not it should issue the policy. On the other hand, when dealing with “Group” policies, the insurance company has no direct contact with the organization’s employees. In fact, the insurance company we are working with does not record any information regarding the client organization’s employees until a claim is actually filed.

The claims data we have include information about the Insured, the coverage, the beneficiaries, and the payment. Only 33% (7,746) of the "individual" claims were successfully joined with their corresponding policies. Figure 2 summarizes out data.

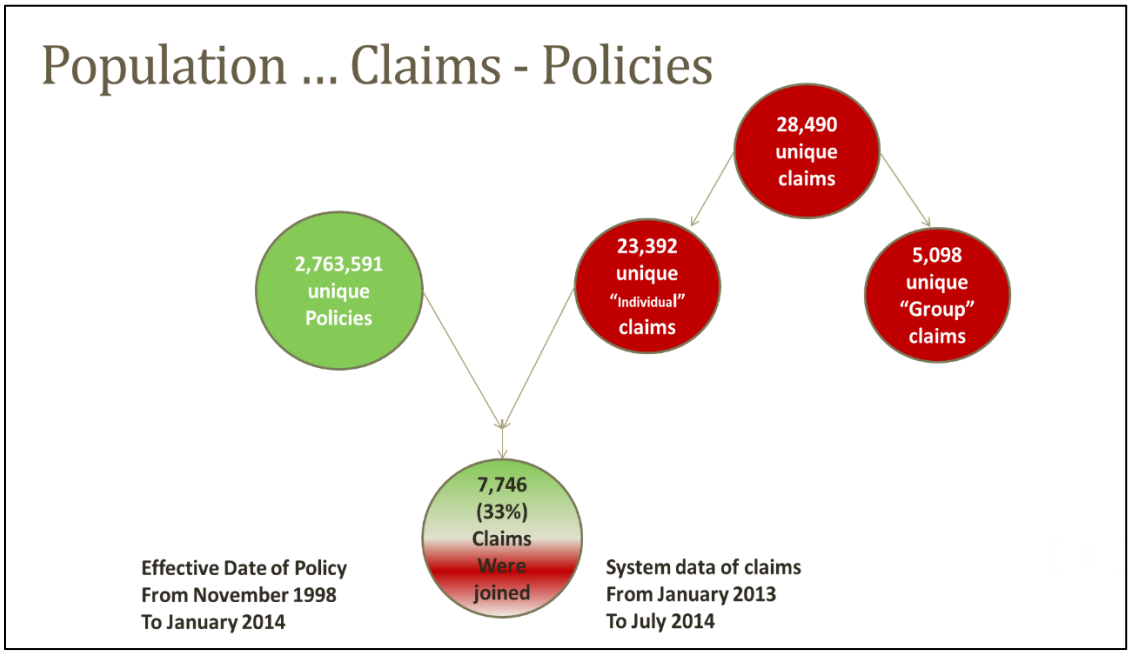

Figure 2: Population

Continuous Auditing Framework for Life Insurance

The continuing use of traditional auditing practices in the current state of the accounting information systems prevent the economy from exploiting the full potential of such systems. (Chan and Vasarhelyi 2011) described seven dimensions that distinguish the continuous auditing over the traditional one; more frequent audits (or continuous ones), proactive audit model, automated audit procedures, evaluation of work and role of auditors, change in the nature, timing, and extent of auditing, use of data modeling and analytics for monitoring and testing, and the change in nature and timing of audit report. In their paper, (Chan

and Vasarhelyi 2011) describe four stages of applying the continuous auditing. The first stage is to identify business process where the audit can be continuous in terms of the availability and accessibility of the data and then identify the audit procedures and the type of monitoring and testing that can be automated. The second stage is to develop benchmarks for evaluating future data based on the historical data we already have though modeling, estimation, classification, association, and/or clustering. In this paper, we propose continuous auditing view for the life/disability insurance industry. As recommended by (Chan and Vasarhelyi 2011) we start by identifying a specific business process that can be automated which is the Claim Payment process. We then identify the set of controls and tests that can be automated within this business process. The following graph shows the proposed continuous auditing model for the claim initiation step.

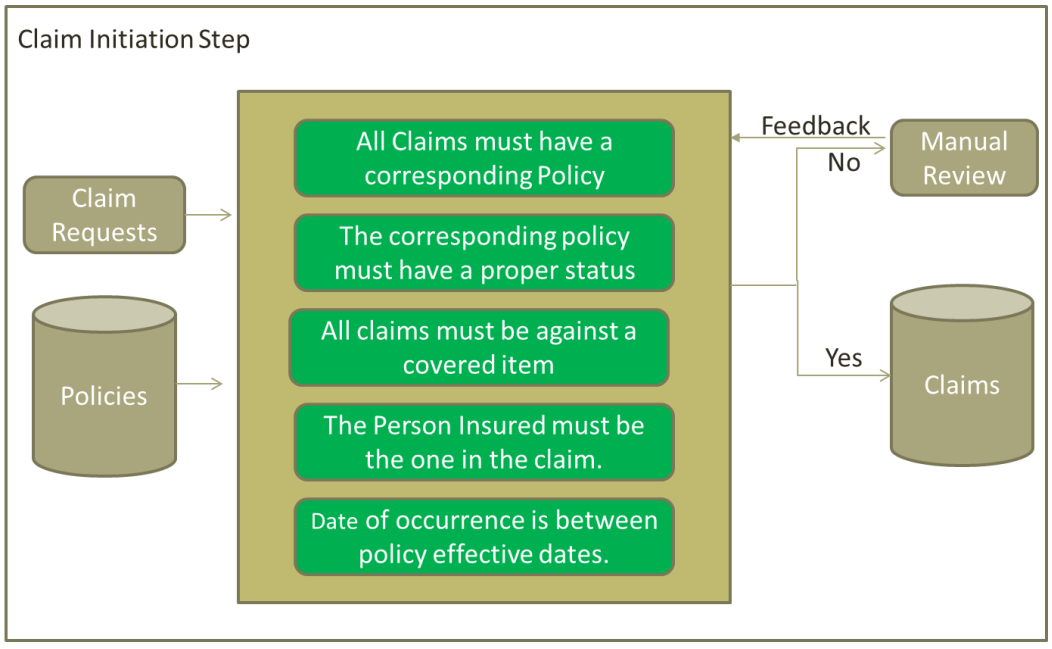

Figure 3: Claim Initiation

As Figure 3 shows, in the claim initiation step a claim gets submitted to the company. Before the company starts processing the claim, a few tests should be performed. The tests are performed through a simple check with the policies the company is maintaining. The first test is to check that this claim is filed against a valid policy that is maintained by the company. The second test checks if that corresponding policy have a proper status. For example, a policy might be canceled or the company might have deactivated it because the client was not paying his premium. In any of these cases, the company might not be liable to pay the claim. The third test is to check that item the claim is filed against is covered under the corresponding policy. For example, if the claim is filed to cover for funeral expenses, the third test will check if the “funeral expenses” are covered under the corresponding policy. The fourth test checks if the person in the claim is the same person insured under the corresponding policy. The last test in the claim initiation process checks if the date of occurrence of the event causing the claim lies between the beginning and ending effective dates of the policy. If all the five tests came back positive, then the claim can be initiated and entered into the claims database to start the approval process. If any of the tests came back negative, then the claim should be flagged for manual review. The results of the manual process are fed back into the system as feedback. The claim initiation step in the claim payment process is a perfect example of continuous auditing as a part of a business process that can simply be automated.

The Figure 4 shows the proposed continuous auditing model for the same business cycle, the claims payment cycle, for the next step, claim validation.

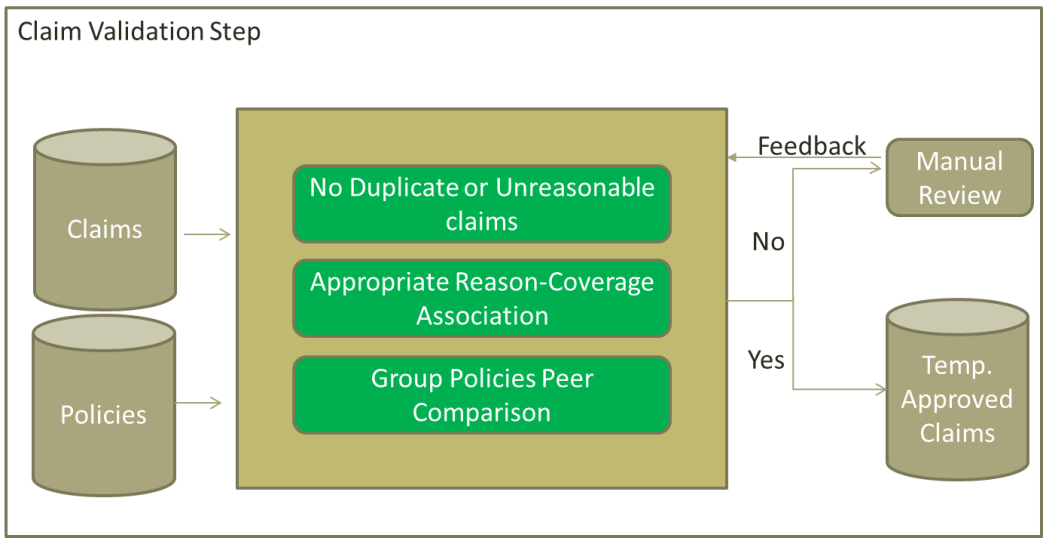

Figure 4: Claim Validation

Now that we know that the claim is filed against a valid policy, we need to know if the claim itself is legitimate. The first test is simple. No client can die twice. So we check for previous claims filed against the same policy, or another policy for the same insured person, for inconsistencies. Inconsistencies here could mean double death, two funerals for the same client, or too frequent disabilities. A manual check will be needed in these cases since some of them might still be legitimate (like the too frequent disabilities case). The next test checks for the relationship between the reason of the claim and the item the claim is filed against. A claim might have one of many reasons like death of the insured person, hospitalization of the insured person, an accident causing disability, or something else. When a claim is filed, it has to be filed against a specific item insured under the policy. This item could be funeral expenses, grave, hospital expenses, medical assistant, or something else. This test uses the historical data to learn the possible association between the reason of the claim and the item the claim is filed against. We then use this to anticipate whether a certain claim needs further manual investigation. For

example, if the reason of the claim is illness, it would be reasonable to be filed against hospital expense, but unreasonable to be filed against funeral expenses. This test will be discussed in details in a later section of this paper. The next test is only focused on Group Policies. A group policy is a policy that is bought by a company or an institution to insure its employees. It is called group policy because it is one policy with a group of insured personnel. In some cases, the insurance company does not collect any information on the insured personnel unless something happens to one of them. Which makes the control and auditing of these claims much harder than individual claims. The group peer comparison test group the insured companies into identical groups in terms of the type of policies they have and the items insured under these policies. The test compares the filing pattern of an insured company with the filing pattern of its peer group. Again, if any of the tests came back negative, the claim will be manually reviewed and the feedback will update the continuous auditing system. Otherwise, the claim will be approved and stored in a temporary approved claim dataset waiting for valuation and payment.

The final step we are discussing in this continuous auditing vision for life/disability insurance is the claim valuation and payment step shown in the graph below.

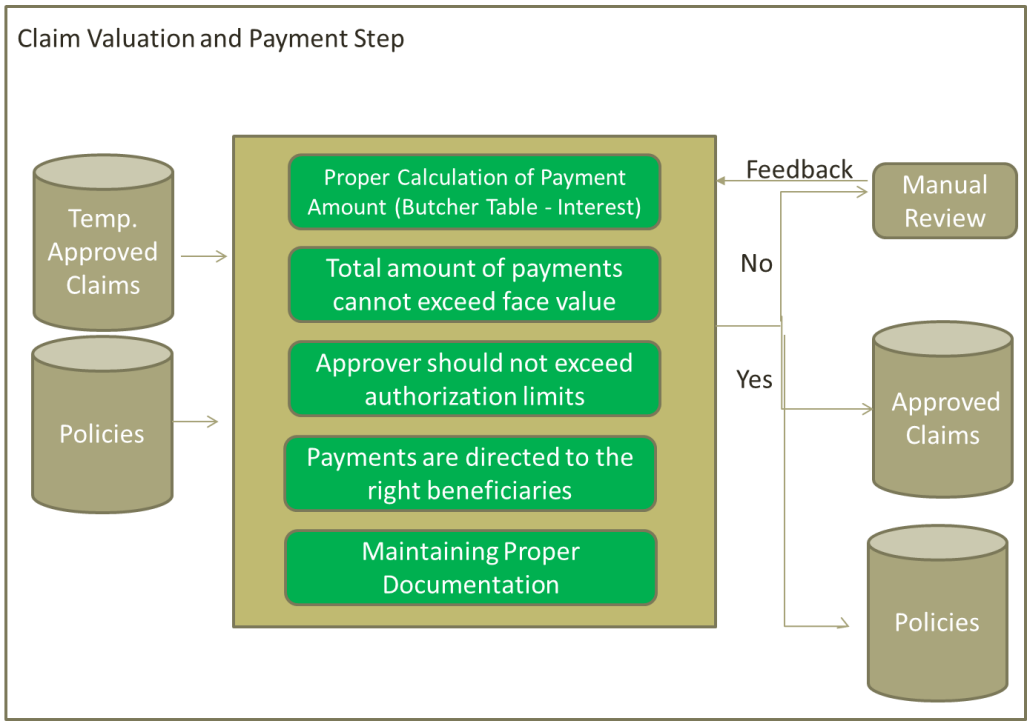

Figure 5: Claim valuation and Payment Step

So we start with the temporary approved claims dataset. These are the claims that passed the initial two steps. Now that the claim has been validated, it's time to determine how much the insurance company should pay. With life insurance the client or the insured person is entitled to part of face value of the policy in case of disability. This part of face value is determined based on the type and severity of the disability. The insurance company has a list of possible disabilities and the percentage of face value associated with each one of them. They call this list "the butcher table". Another factor that affects the amount of payment that the insurance company must pay is the interest. While the calculations of interest may be different from one company to the next one, the rule is the insurance company has to pay interest if the claim payment was delayed beyond a specific grace period. From the above, the first test in this step is to make sure that the amount

of claim approved is the correct amount the insurance company must pay. The test is a simple recalculation of the amount of the claim based on the “butcher table” and the interest calculation policy of the company. The second test searches for all previous claims paid by the insurance company against the policy of the current claim. The test adds all previous claim amounts and compares it to the face value of the corresponding policy. If the outstanding balance is sufficient to pay the current claim, then the claim is passed to the next test. If not, then the claim is flagged for manual review. The next test in this step is to verify that the approver of this claim has not exceeded his authorization limit of the current period. The test searches for all the amounts approved by the current claim’s approver during the current period (the period could vary depending on the company policy). The test adds these amounts and compare it with the approver’s authorization limit based on his/her position within the insurance company. The next test is concerned with the beneficiaries. The claim could be payable to one or more beneficiaries. These beneficiaries must be identified in advance in the policy contract. The test compares the original beneficiaries and their bank accounts with the person(s) the insurance company is directing the claim payment to. The last test checks for maintaining prober documentation. The insurance company should maintain an updated list of its policies. That means updated the outstanding balance of the face values after paying claims, changing the status of the policy after cancelation, expiration, or after death claim. As always, if any of the tests came back with any abnormalities, the claim will be manually reviewed and then the feedback will update the continuous auditing system. Otherwise, the claim will be approved and

stored in the approved claims dataset waiting. The corresponding policies will also be updated in the policies table. As part of the feedback we look at all true positives and check which assertions were most important. We then construct true positive or ideal positives for each assertion, and also false positives to set a benchmark to compare the exceptions to them.

Claims Anomalies Detection

In this section we discuss in details analytical procedures mentioned in claims validation step in the framework. The analytical procedures are to help the auditor detect claims anomalies in testing two audit assertions; is the claim settlement reasonable? And is the claim itself legitimate? To do this we use claims data provided to us by a leading international insurance company. We use a multi-dimensional approach in which we divide the attributes we have into different groups (dimensions). We use each dimension to logically find insurance claim anomalies. then we prioritize the anomalies based on the weighted average of the dimensions that identified this record as an anomaly.

Is The Claim Settlement Reasonable?

For the reasonability of the settlement amount we looked at the total payment to the client (or beneficiaries). According to the policy of the insurance company we are working with, if the claim was not paid within thirty days of the day of filing the claim, an interest amount is calculated and added to the settlement. For the

individual life insurance claims, the total payment to the beneficiaries should be equal to the face value of the policy plus any interest earned over the allowed 30-day period. Detecting anomalies in this case involves testing the relationship between the interest payments and the period between filing the claim and paying the settlement. When looking into the data we have, the attributes we used here were as follows; a lump sum of total payments to the beneficiaries, date of filing the claim, and date of payment. From these attributes (Group 1) we can get the following

- Days till Payment: the number of days between filing the claim and paying the settlement (Date of Payment – Date of Filing the claim).
- Estimated Daily Interest: the interest amount is estimated to be the positive difference between amount actually paid and the face value of the policy. The estimated daily interest is calculated by dividing the estimated interest amount by the “Days till Payment”.

The following figure (Figure 6) shows the distribution of the estimated daily interest.

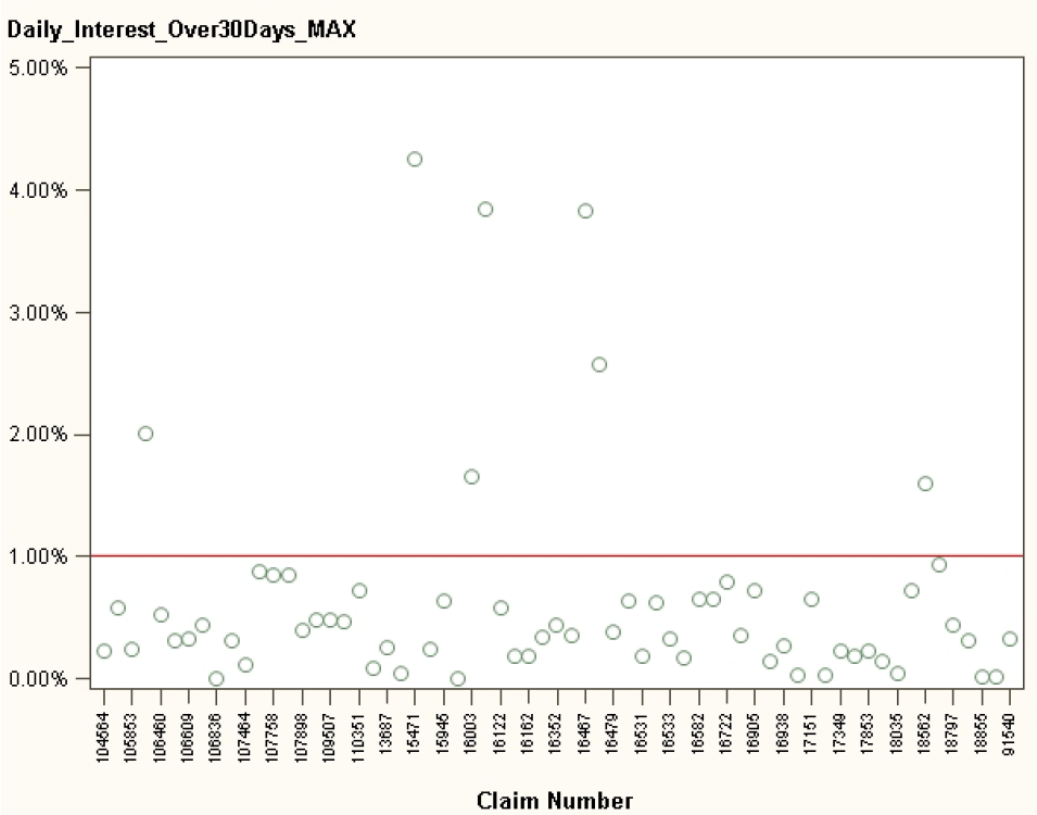

Figure 6: Estimated Daily Interest (After the first 30 days)

From Figure 6 we notice that the obvious anomalies (from this point of view) are seven claims (above the red line) with a more than 1% daily interest (after the first 30 days).

We also added to this analysis two more dimensions which are the relationship between the “Estimated Daily Interest” and the reason of the claim (the reason the client or the beneficiaries filed the claim. Example would be death of the insured person) and the relationship between the “Estimated Daily Interest” and the type of coverage (what the policy actually covers. Examples would be paying for funeral expenses, headstone, etc.).

One major limitation to this analysis is that it can only be applied to the “Individual Life Insurance” claims. We cannot apply this analysis to the “Group” Life Insurance claims. As we stated above we only have a lump sum of the total payment for each claim. The interest amount was estimated to be any excess of payment beyond the face value of the life insurance policy. For individual life insurance policies, the insurer company is to pay the full amount of the face value upon the death of the insured person. But for the group life insurance policies, the insurer company only pays a percentage of the face value of the policy based on some predetermined criteria such as the deceased employee’s salary. So given our data (lump sum of payment) and the nature of the group life insurance settlement calculations, we have no precise way of calculating the estimated interest amount.

This analysis cannot either be used with the Disability Insurance claims that we have. For disability claims, only a portion of the face value of the policy is paid based on the type of the disability. If the insured person had lost the right arm, he would be entitled to a different percentage of the face value if he had lost the left arm. The insurer company provided us with these percentages. They call it the “Butcher Table”. But a link between these percentages and the actual claims data could not be established. So again, we had no precise way of calculating the estimated interest amount.

Is the Claim Itself Legitimate?

For the second audit assertion, the legitimacy of the claim, we check for anomalies using three different dimensions; Reason-Coverage association, Group Similarities, and Timeline.

Reason-Coverage Association

When a person buys an insurance policy from the insurer company we are dealing with, he gets to choose from a set of different products they have (different types of policies). Each product provide insurance against a specified set of events. Each event is identified by a unique code; we call them the “Coverage” codes. For example, when a person buys a life insurance policy, he gets to choose the type of life insurance policy he wants. Life insurance policy type A would provide insurance against death, funeral expenses, headstone cost, and disability. Life insurance policy type B would provide insurance against death, funeral expenses, headstone cost, but not disability. When this person dies later on and his beneficiary files a claim he has to specify the reason for filing the claim (in this case the death of the insured person) and the coverage codes he is filing the claim against. So if the beneficiary is filing against funeral expenses, the claim reason would be the death of insured person, the claim coverage code would be the code for funeral expenses. If he wants to file for headstone cost, there would be another claim with a claim reason being the death of the insured person and the claim coverage code would be the code for funeral expenses.

In our dataset, we have 9 different claim reasons filed against 35 different coverage codes. For the purpose of this analysis we check for which kinds of coverage codes

was a certain “claim reason” filed against. For each claim reason we build a matrix. The matrix lists the claim number by rows and the 35 coverage codes by columns. For example, consider Figure 7. It shows the matrix for “Reason 1”. In this matrix we only consider claims that were filed for “Reason 1”. For the first claim, claim 1234, we check, was the claim filed against coverage code “COV 1”? If so, then we put “1” in the cell. And so on. So claim number 1234 was filed for “Reason 1” and was filed against COV 1, COV 2, COV 4, but not COV 3. When we are done with all the claims filed for that reason we end up with different sets of coverage codes per claim reason. Again, looking at figure 2, we have a total of 3 claims filed against “Reason 1”. We ended up with 2 different sets of coverage codes per that reason; Coverage Set “A” {Cov 1, Cov 2, Cov 4} and Coverage Set “B’ {Cov 1, Cov 2, Cov 3, Cov 4}. We then calculate the filing rate against each coverage set. For example, in Figure 2, Coverage Set “A” was filed against 2 times out of a total of 3 claims so its filing rate is 2/3 while Coverage Set “B” was filed against only 1 time out of a total of 3 claims so its filing rate is 1/3.

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

Figure 7: Reason-Coverage Association - Example

For the purpose of this Reason-Coverage association, the lower the filing rate against a specific coverage set in comparison with the other coverage sets, the more suspicious is the claim. Figure 8 below shows an example from the dataset. The figure shows the distribution of the different coverage sets we found for the reason of “Death by Disease”. We had a total of 17 coverage sets. In this case, the anomalies would be the least used sets (less than 5% filing rate.

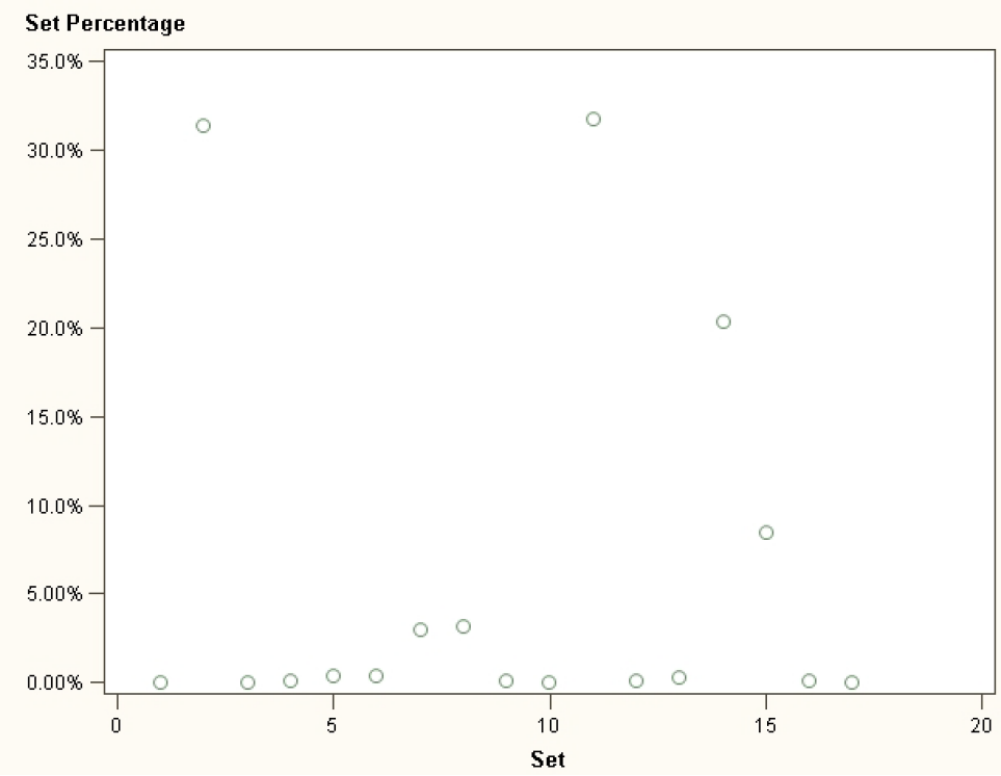

Figure 8: Reason-Coverage Association - Filing Rate

Group Similarities

The second dimension in testing the legitimacy of the claim is the Group Similarities. By “Group” here we mean “Group insurance policies” when the client is a corporation insuring against the death or disability of its employees. In this dimension we look only into these “Group policies”. We identify two “Group Policies” as being similar when and only when both of them insure against the exact same coverage set. Once we identify the similar policies we put them in groups and then test the claim filing pattern of each policy against the average of its group’s. For example, consider Figure 9. It shows an example of list of group policies and the coverage sets they are insuring against. The first three (blue policies) insure against the same coverage set {COV 1, COV 3}. So, they are considered to be a similar group. The last two policies (green policies) insure against the same coverage set {Cov 2, Cov 3}. So they are considered to be a similar group.

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

Figure 9: Group Similarities - Identifying Groups

Now after we define the group similarity, we study the claim filing pattern of each policy in terms of which coverage codes does it file against more often. And then we compare the filing pattern of each policy to the average of its group. For example, consider Figure 10 below.

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

Figure 10: Group Similarity - Filing Pattern

In Figure 10, we see the “blue” group’s filing pattern. So policy number 1234 filed 60% of its claims against COV 1 and 40% against COV 3. Policy number 1235 was a little bit similar in its behavior. It filed 65% of its claims against COV 1 and 35% against COV 3. On the other hand, policy number 1236 filed only 10% of its claims against COV 1 and 90% against COV 3. Compared with the rest of the blue group, policy 1236 would be considered as an anomaly.

Claim Timeline

The third and last dimension used to detect anomalies in terms of the legitimacy of the claim itself is the claim timeline (Figure 11).

Figure 11: Claim Timeline

The claim timeline is a representation of five dates we have on the dataset; the date of issuance of the policy, date of occurrence of event, date of noticing the event (sight), date of filing the claim, and date of paying the settlement. What we are looking into here is the anomalies in this timeline. For example, a too long period of time between the occurrence of the event and the noticing of the event (when talking with the insurance company, they explained that “Noticing” of event or “Sight date” is actually the date the insured person or his beneficiaries first contacted the insurance company about this claim).

So, in this timeline we consider the following:

- Number of days from the beginning of the effective date of the policy till Occurrence.
- Number of Days from occurrence till sight.
- Number of days from sight till filing.
- Number of days from filing till payment.

The anomalies defined here would be an exceptionally long period between occurrence and or notice and filing, or filing till payment. Also an exceptionally short period between issuance and occurrence.

The Figure 12 shows the box-and-whisker plot for the “Days till Occurrence”, which is the number of days between the beginning of the effective date of the policy and the occurrence of the reason of the claim.

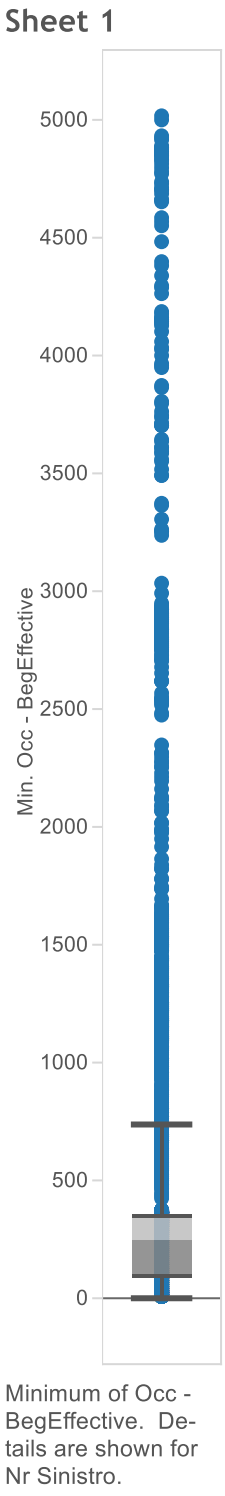

Figure 12:Box and Whisker Plot for Days till Occ

The graph gives us an idea about the distribution of the “days till occurrence”. The number of days ranges from zero to 9,410 days. The median is on a 59. The lower whisker is on a zero and the upper whisker is on 300. The graph shows that most of the claims (the four quartiles) are filed against incidents that happened within a year of the issuance of the policy. And then what was considered outliers by the box and whisker graph are what’s above the 300 days up to 4995 days. While the upper outliers are actually profitable to the insurance company (Been able to collect premiums for a longer period before it had to pay a claim), we are concerned with the claims filed within a short period of time from the beginning of the effective date.

The Figure 13 is another representation of the same data. The horizontal access shows the different claims, while the vertical access shows the “days till Occurrence”. It’s obvious from the graph that most of the claims are filed on incidents that happened within the first year of purchasing the insurance policies.

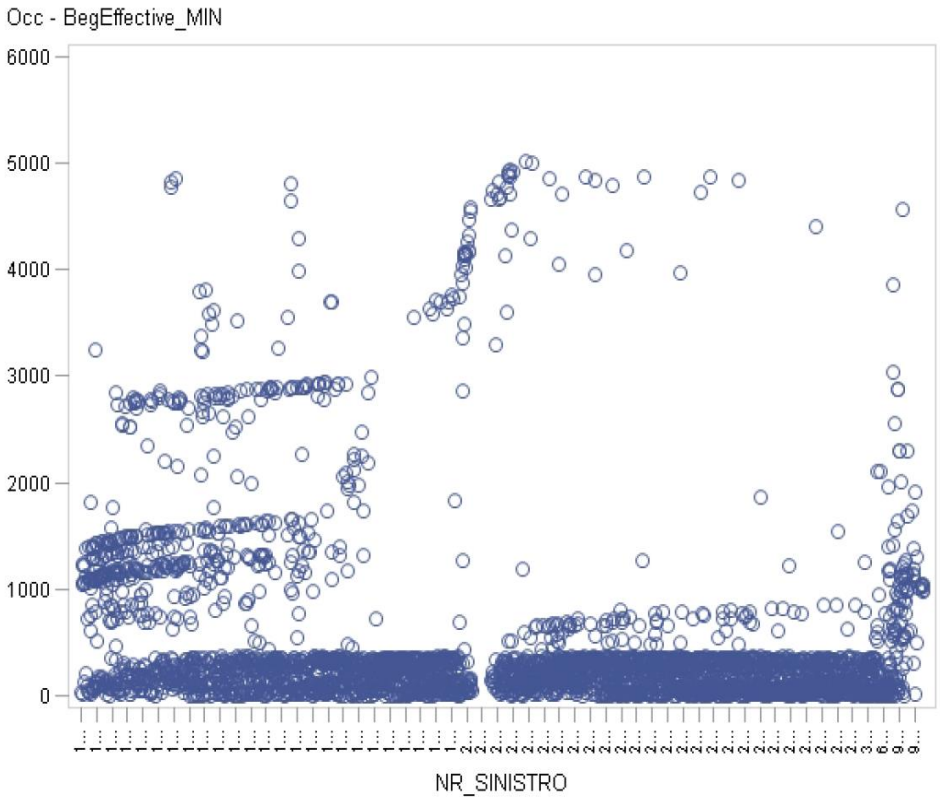

Figure 13:Distribution of Days Till Occ

To priorities the claims anomalies form the Days Till Occurrence's point of view, we need to study the percentile distribution of the "Days till occurrence" as shown in table 6.

Table 1:Days Till Occurrence Percentile

Obs	P_1	P_5	P_10	P_25	P_50	P_75	P_90	P_95	P_99	P_100
1	2	11	26	85	252	348.5	1208	1636	4160	5017

As seen in table 6, based on the percentiles, we suggest that the claims with "days till occurrence" equal to two days or less, be assigned the highest priority as an anomaly in the claims from the point of view of the days till occurrence followed by claims with more than 2 days but less than 11 days and so on.

The Figure 14 shows the distribution of the “Days from Occ till sight”.

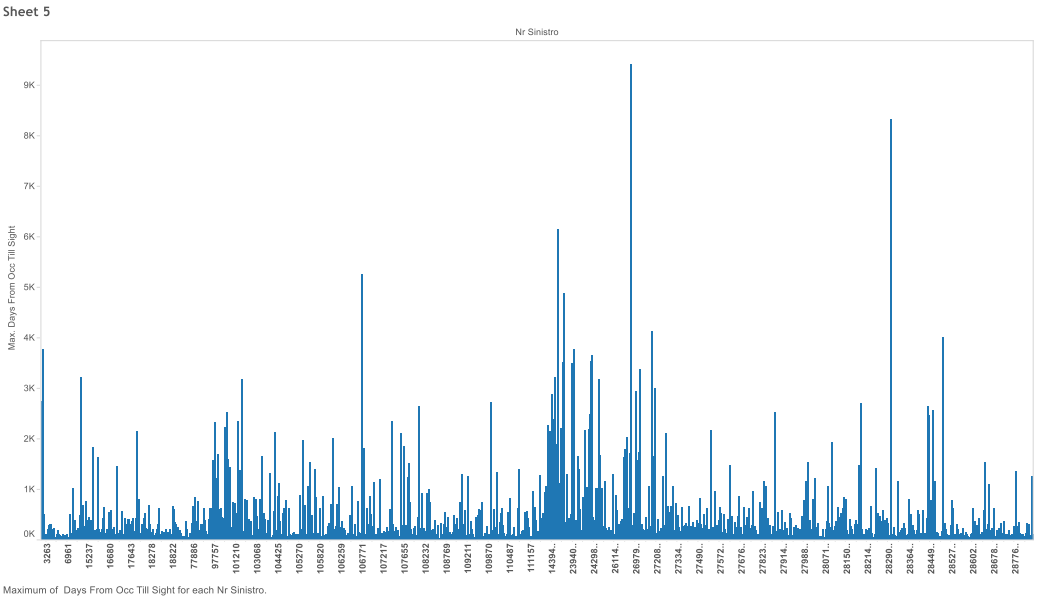

Figure 14: Days from Occ till Sight - Distribution

Figure 14 shows clear outliers as most of the data lies below the 500 days’ line, some of the data goes up to 9,410 days.

The Figure 15 shows the box-and-whisker plot for the “Days From Occ till Sight”, which is the number of days between the occurrence of the reason of the claim and the first time the client contacted the insurance company about this claim. The number of days ranges from zero to 9,410. The median is on 59. The lower whisker is on zero (no estimated outliers below the lower whisker) and the upper whisker is on 300. In other words, the normal situation would be that the client contacts the company within the first year of the occurrence of the incident. But with a period of 9,410 days between the occurrences of the incidents and contacting the company, the auditor has to investigate further whether this is a legitimate claim.

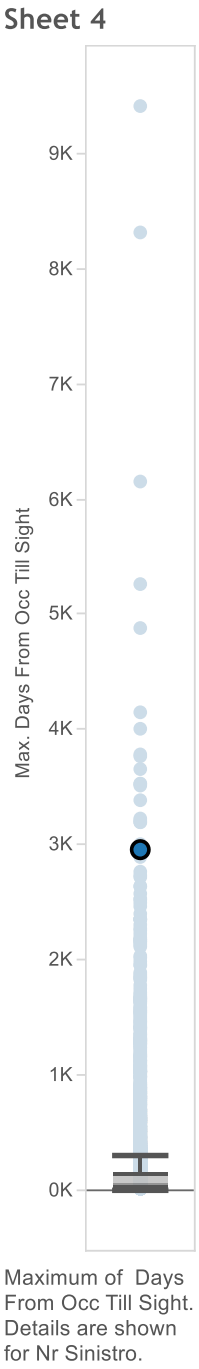

Figure 15: Box-Whisker Days Fromm Occ Till Sight

While our concern with the “Days till Occurrence” was the lower outliers, our concern here is the upper outliers. Why would a client wait for years before

contacting the insurance company about an incident that is covered under the insurance policy he bought? The following table (table 7) shows the percentiles of the “Days from Occ till Sight”

Table 2: Days from Occ till Sight Percentile

Obs	P_1	P_5	P_10	P_25	P_50	P_75	P_90	P_95	P_99	P_100
1	2	10	16	28	56	128	338	630	2010	9410

From table 7, we suggest that anything above the 90th percentile (338 days) should be considered an anomaly. The higher priority will be given to anything above 2010, followed by anything above 630 but less than 2010, followed by anything above 338 and less than 630.

Prioritizing the Anomalies based on the weighted dimensions

we can now prioritize the anomalies based on the weighted average of the dimensions that triggered them as anomalies. The suspicious score of each anomaly should be defined as follows (Issa 2013)

$$ISS(X_i) = \sum WD_i \, VD_J$$

Where $ISS(X_i)$ is the Initial Suspicion Score of claim X before considering the branch risk score $_i$.

WD_j is the weight of Dimension D_j

And VD_j is the binary variable that equals one if Claim Xi violates Dimension D_j, and 0 otherwise

Risk Scoring

In this section we are discussing a methodology that can be used to assign a risk score to a specific unit based on the evidence that we collected during the audit. The scoring methodology is based on the Belief Function and Evidential Reasoning (Lili et al. 2006; Shafer and Srivastava 1990; Rajendra P. Srivastava and Mock 2000; Srivastava and Mock 2005; Srivastava and Shafer 1992). This risk score would then be used to prioritize the claims outliers.

The assertions we discussed earlier have binary results; either pass (a) or fail (~a). In this case there will be no ambiguity in the evidence we collect specially that these assertions are meant to test each claim transaction. The ambiguity in our case comes from the missing values. Some claim transactions cannot be tested due to missing values that are essential to run the test. For example, to check if the total amount of claim does not exceed the face value on the policy we need link the claims with the policies. If this link is broken because of missing values (missing the policy number in either the claims or the policies datasets), the test cannot run for this specific transaction. Another example is when we check if the amount of the claim was paid to the right beneficiaries (deposited into the right bank account). If the beneficiary information is missing from either datasets, the test cannot run. In this case, how should we consider this transaction? We cannot say it passed the test but we cannot say it failed it either. We thought that the belief function with its designed capability of incorporating ambiguity can better describe these cases of evidence. Also, with the incorporation of the Dempster's rule of combination and evidential reasoning, we can build evidential diagram to

summarize the results of the assertions we ran and give an aggregate risk score of the unit in question. The unit in question could be either the authorizer of the claims, branch of the insurance company, or the geographical region of the claims depending on what we want to aggregate and score. The following is an example from the data we have on how to assign a risk score to the different branches of the insurance company.

The data we have represents claim transactions from two different branches of the insurance company. The distribution was as follows:

Branch	Percentage of Total Records
Branch A	64.78%
Branch B	35.22%

Table 3: Different branches' share of total data

For the first assertion, all claims must have a corresponding policy, the results we got was as follows

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

Table 4: an example of the first assertion results

The results show that 86% of Branch A's claims had a corresponding policy on files and 14% didn't while 96% of branch B's claims had a corresponding policy on files and 4% didn't. There was no ambiguity in this test's results.

For the fifth assertion, the date of occurrence must be between the effective dates of the policy, the results were as follows:

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

Table 5: Example of Fifth assertion results

The results show that 75% Branch A's claims passed the test, 12% failed, and 13% were unknown because of missing values. For Branch B, 73% of the claims passed the test, 23% didn't, and 4% were unknown because of missing values.

Branch A_Test1: m1({a}) = 0.86, m1({~a}) = 0.14, m1({a, ~a}) =0.0

Branch A_Test2: m2({a}) = 0.75, m2({~a}) = 0.12, m2({a, ~a}) =0.13

Conflict= [m1(a).m2(~a)] + [m1(~a).m2(a)] = [0.86 x 0.12] + [0.14 x 0.75] = 0.21.

K = 1 – Conflict = 1 – 0.21 = 0.79

Branch A_Total:

$mt(\{a\}) = K^{-1}[m_1(a)m_2(a)+m_1(a)m_2(\{a, \sim a\})+m_1(\{a,\sim a\})m_2(a)]$

= [0.86*0.75 + 0.86*0.13 + 0.0*0.75] / 0.79 = 0.757 / 0.79 = 0.96

$Mt(\{\sim a\}) = K^{-1}[m_1(\sim a)m_2(\sim a)+m_1(\sim a)m_2(\{a,\sim a\})+m_1(\{a,\sim a\})m_2(\sim a)]$

= [0.14*0.12 + 0.14*0.13 + 0*0.12] / 0.79 = 0.035/0.79= 0.04

$mt(\{a, \sim a\}) = K^{-1}m_1(\{a, \sim a\})m_2(\{a, \sim a\})$

= 0 * 0.13 / 0.79 = 0

Bl ({a} = 0.96 + 0 = 0.96

Pl({~a}) = 1- 0.96 = 0.04

Final Anomaly Suspicious Score

we can now use the dimension to prioritize the anomalies. The risk score of the branch that generated the anomaly. The suspicious score of each anomaly should now be defined as follows (Issa 2013)

$$ISS(X_i) = \sum WD_i \, VD_J \, B_r$$

Where $ISS(X_i)$ is the Initial Suspicion Score of claim X before considering the branch risk score.$_i$

WD_j is the weight of Dimension D_j

And VD_j is the binary variable that equals one if Claim Xi violates Dimension D_j, and 0 otherwise

The higher the suspicious score, the more the auditor is encouraged to audit this anomaly.

Premium Outliers Detection

One of the important aspects in auditing the Life/Disability insurance Revenue cycle is to check for the valuation of the premiums. The obvious way to do this check is to recalculate the premium based on the data available on the corresponding policy. The first step we have to go through to perform this check is to determine what factors are incorporated into the premium calculations.

Factors Affecting Premium Calculations

The premium calculations of a certain policy would depend on four factors; the type of the policy, the life expectancy of the insured person, the importance of the

insured person to the insurance company (VIP status), and the profit margin of the insurance company.

Type of the policy:

The type of the policy is the way the company defines the policy. The first aspect in defining the policy is to determine the type of insurance this policy offer. One single insurance company can provide different types of insurance. Examples of that could be life insurance, auto insurance, rental insurance, and/or health insurance. The second aspect is the type of product that covers this policy. For example, an insurance company might offer different products under life insurance. A product specifically designed for women which they would call “Life Insurance for Women”, a product specifically designed for families which they would call “Life Insurance for families”, and so on. The third aspect would be the list of items that are covered under each policy. So, a life insurance policy might cover hospital expenses, temporary income, death, and funeral expenses. Or, it might only cover hospital expenses and death. The premium for the both of these policies should be different because they cover different items, even though both policies are life, and both are under the same product (life insurance for families). A last aspect that affects the premium from the point of view of the “type of police” is the face value of the policy. So, two policies with the same type of insurance (life), same product (life for families), same list of covered items, but different face values should have different premiums.

Life Expectancy of the Insured Person:

It is intuitive that the life expectancy of the insured will have a negative relationship with the premium he/she has to pay, other things being equal. So, what affects the life expectancy of a person? The first aspect that comes to mind, is the age. Other things held constant, a younger person is expected to live for more years to come than an older person. Actuaries have their own calculations of how long a person is expected to live. A second aspect is health condition of the insured person. A third aspect is the profession of the insured person. Some professions have higher mortality rate than others, for example, a coal mine worker is expected to have more health issues than university administrator at the same age. The forth aspect is the affiliation of the insured person to risky sports or habits. For example, a biker might have a higher risk than a non-biker of the same age and profession. A smoker has a higher risk than a non-smoker.

Importance of the insured person to the insurance company

Just like any business might offer special discounts for its VIP list of clients or customers, an insurance company might offer special treatment for its special clients. The question here that we cannot answer at this moment is what makes a client a special client.

The Insurance Company's Profit Margin

The premium is supposed to serve two purposes for the insurance company. It is supposed to cover its estimated or expected risk and it is also supposed to provide a profit margin for the insurance company.

Robust Regression

Robust regression is an important tool for analyzing data that are contaminated with outliers (Chen and Meer 2003). It can be used to detect outliers and to provide stable results in the presence of outliers. In order to achieve this stability, robust regression limits the influence of outliers. There are different statistical methods that can be used with the robust regression to stabilize the results. These methods are used to measure the proportion of outliers in the data. This will help the regression model in limiting the influence of these outliers on the results. The choice of the method to be used depends on the characteristics of the data. Several methods can be selected with robust regression (Chen and Meer 2003). One method is the M estimation (Maximum Likelihood) which was introduced by Huber (Huber 1973). It is considered as the simplest approach both computationally and theoretically. It is used extensively in analyzing data for which it can be assumed that the contamination (outlier) is mainly in the response direction. The M estimation method cannot be used if the contamination is on the explanatory direction of the data. Another method is the Least Trimmed Squares (LTS) estimation. It is a high breakdown value method introduced by Rousseeuw (Rousseeuw 1984; Rousseeuw and Van Driessen 2006). As opposed to the Least Square which minimizes the sum of squared residuals over n points, the LTS attempts to minimize the sum of squared residuals over a subset, k, of those points. The (n-k) points are not used (assumed outliers). The breakdown value is a measure of the proportion of contamination that a procedure can withstand and still maintain its robustness. A third method is the S estimation (Rousseeuw and

Yohai 1984). This method finds a line (plane or hyperplane) that minimizes a robust estimate of the scale of the residuals. With the same breakdown value, it has a higher statistical efficiency than LTS estimation. The fourth method is the MM estimation introduced by Yohai (Yohai 1987). It combines high breakdown value estimation and M estimation. It has both the high breakdown property and a higher statistical efficiency than S estimation.

Since our main objective in this section is to find outliers in the premium based on the data we know about the policy and the insured person, we assume that in our case the contamination in the data will be in the response part (the premiums) rather than the exploratory part (policy and insured person details). With this assumption in mind, we chose to work with the M-Estimation method with the procedure ROBUSTREG provided by SAS version 9. The following section gives more details on our model.

The Model

Based on the factors affecting premium calculation discussed above, we develop our model using SAS's ROBUSTREG procedure with the M-Estimation method to stabilize the results.

For the type of the policy, our data has three variables, the type of the insurance (life – disability), the type of the product (accidents, Life, Life Uniclass, Life Women, … etc.), and type of coverage (Death by accident, Death by Disease, Funeral, Grave, Hospital Expenses, … etc.).

Type of Policy – Insurance

Table 6 shows the different types of insurance policies in our data.

Table 6: Types of Insurance Policies

COD_RAMO	Frequency	Percent	Cumulative Frequency	Cumulative Percent
93	6059531	53.69	6059531	53.69
81	3398601	30.12	9458132	83.81
82	1827207	16.19	11285339	100.00

The code 93 stands for “Group – Life” type of insurance policies. 81 stands for “Individual – Disability”. And 82 stands for “Group – Disability”. The company we are dealing with explained that while an “individual” insurance policy (code 81) includes only individual policies, “Group” insurance policies (codes 93 and 82) might actually include both group policies and individual policies. The accurate measure is the existence of either an individual unique identifier (a type of Social Security Number), or a company unique identifier (a type of a tax ID). Also, the existence of a date of birth would obviously indicate an “individual policy”.

The factors that affects the calculation of the premium should be different in case of group policies than those in case of individual policies. In our case, the factors that we discussed in the above section are those that affects premium calculations of individual policies. That means that in this model we are only interested in characteristics of individual policies. To restrict the data to those of the individual policies we added two conditions; the individual unique identifier (SSN) is not missing, and the birth data is not missing. After adding these two conditions, we

found out that all the policies under the code 93 (Group – Life) actually belonged to individual policies rather than group policies.

For this model, we will restrict the type of insurance to type 93 (group – life) with the added condition of belonging to individual policies.

Type of Policy – Product

We have 81 different types of products in our data. But since we are running the model only on the insurance type (93), we are only considering products that belong to this type of insurance. We end up with 50 different products. Table 7 shows the distribution of 15 of these products that constitute more than 90% of the policies under insurance type 93.

COD_PRODUTO	Frequency	Percent	Cumulative Frequency	Cumulative Percent
167	1395500	23.03	1395500	23.03
242	966625	15.95	2362125	38.98
168	948033	15.65	3310158	54.63
431	336616	5.56	3646774	60.18
429	292844	4.83	3939618	65.02
420	257218	4.24	4196836	69.26
428	192376	3.17	4389212	72.43
405	173160	2.86	4562372	75.29
430	168908	2.79	4731280	78.08
427	166010	2.74	4897290	80.82
157	159002	2.62	5056292	83.44
493	119947	1.98	5176239	85.42
403	114971	1.90	5291210	87.32
425	112853	1.86	5404063	89.18
150	96420	1.59	5500483	90.77

Table 7: Distribution of Products

For the purpose of this model and to take into consideration the effect of the type of product on the premium calculation, we are only using the data for the product 157 (Insurance – UBB).

Type of Policy – Coverage

We have 69 different types of coverage in our data. But since we are running the model only on the insurance type (93) and product (157), we end up with the following 12 different coverages.

Table 8: Coverage Distribution

COD_GARANTIA	Frequency	Percent	Cumulative Frequency	Cumulative Percent
203	25497	16.04	25497	16.04
221	25497	16.04	50994	32.07
242	25497	16.04	76491	48.11
482	25490	16.03	101981	64.14
624	25457	16.01	127438	80.15
480	18559	11.67	145997	91.82
435	6933	4.36	152930	96.18
210	1854	1.17	154784	97.35
223	1854	1.17	156638	98.51
224	1854	1.17	158492	99.68
205	465	0.29	158957	99.97
211	45	0.03	159002	100.00

For the purpose of this model and to take into consideration the effect of the type of coverage on the premium calculation, we are only using the data for the coverage (203) which constitute 16% of type of insurance 93 and product 157 data.

This concludes the description of the type of policy that we have in the data. We end up with a total number of records equal to 25,497. We ran three additional tests; the first one to make sure that the premium has a positive value, the second to make sure that the face value has a positive value, and the third to make sure the policy is active. In our data, we have a variable that describes the status of the policy. A given policy can either be active (code 1), canceled by the client (code 2), or terminated by the company (code 10). The distribution of the policy status in the 25,497 records that we ended up with was as follows:

Table 9: Policy Status Distribution

COD_SITUACAO_APOLICE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	23530	92.29	23530	92.29
2	1967	7.71	25497	100.00

As we see in the above table, 92.29% of the records belonged to active policies. In our model, we are only using these records with active policies.

The following table summarizes the data records that we selected for our model.

Conditions	Number of Records
	11,285,339
Insurance Type = 93 (Required for Insurance type factor)	6,059,531
Individual Unique Identifier (SSN) (required to prove individual policies)	6,059,531
Birthdate	6,059,531

(Required to calculate age)	
Product Type = 157 (Required for product type factor)	159,002
Coverage type = 203 (Required for Coverage type factor)	25,497
Premium has a positive value (to eliminate data entry errors)	25,497
Face value has a positive value (To eliminate data entry errors)	25,497
Policy status is active	23, 530

The following graph shows a visualization of the premiums and face values of the final 23,530 records of data we have selected for the model.

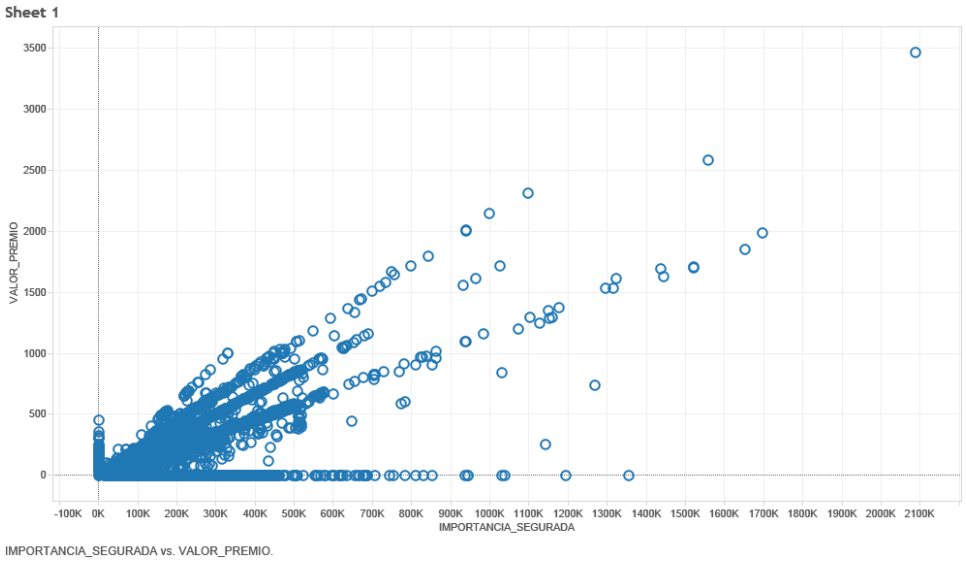

Figure 16: Premium and face values robust regression aggregate

Life Expectancy of Insured Person

For the life expectancy of the insured person we mentioned different factors in the above section. We talked about the age, the health condition, the profession, the affiliation with risky sports or habits. While all of these factors contribute to the life expectancy of the insured person, our data only contains the birthdate of the insured person. We used that date and the date of purchasing the policy for the first time (the beginning effective date of the policy) to calculate the age of the insured person at the time of purchasing the policy.

Age of Insured Person at time of Purchase in Years

= (Date of purchasing the policy – Birthdate) / 360

Other Factors

In the above section, we also talked about the importance of the insured person to the insurance company and the insurance company's profit margin as factors that

affect the premium calculations of the policy. Unfortunately, we do not have any indicators of these factors in our data. We limit our model to the type of policy factors and only the age of the insured person as an indicator of life expectancy due to the lack of data. We expect the performance of the model to be negatively affected by this limitation.

SAS ROBUSTREG

Our original model was as follows

$$\text{Premium} = \text{intercept} + \beta_1\ \text{InsuranceType} + \beta_2\ \text{ProductType} + \beta_3\ \text{CoverageType} + \beta_4\ \text{Facevalue}$$

$$+ \beta_5\ \text{Age} + \beta_6\ \text{Health} + \beta_7\ \text{Profession} + \beta_8\ \text{RiskyAff} + \beta_9\ \text{Importance} + \beta_{10}\ \text{Profit}$$

For the lack of the data we adjust the model to the following

$$\text{Premium} = \text{intercept} + \beta_1\ \text{InsuranceType} + \beta_2\ \text{ProductType} + \beta_3\ \text{CoverageType} + \beta_4\ \text{Facevalue}$$

$$+ \beta_5\ \text{Age}$$

As discussed above, we control for the Insurance Type, Product Type, and Coverage Type. So, we end up with the following model

$$\text{Premium} = \text{intercept} + \beta_1\ \text{Facevalue} + \beta_2\ \text{Age}$$

The followings are some statistical descriptions of the face value, age, and premiums that we have in the data.

Table 10:Summary Statistics

Summary Statistics						
Variable	Q1	Median	Q3	Mean	Standard Deviation	MAD
IMPORTANCIA_SEGURADA	61655.4	111607	205341	147578	123407	93328.9
Age	49.8301	55.4808	61.5288	55.8372	8.5482	8.6499
VALOR_PREMIO	99.8505	130.4	242.6	184.5	173.6	99.1340

Variable	Minimum	Maximum	Mode
Age	26.6465753	95.0082192	60.6876712
IMPORTANCIA_SEGURADA	2.7001000	2088930.62	47579.44
VALOR_PREMIO	0.0044223	3463.54	103.2124727

The Results

The following table shows the result of running the SAS ROBUSTREG with M-Estimation on the data we selected

Table 11: ROBUSTREG Results

Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	56.7042	2.1241	52.5410	60.8674	712.64	<.0001
IMPORTANCIA_SEGURADA	1	0.0011	0.0000	0.0011	0.0011	167139	<.0001
Age	1	-0.3760	0.0377	-0.4499	-0.3021	99.43	<.0001

The following table shows some goodness-of-fit measures to evaluate the model's performance

Table 12:ROBUSTREG Goodness of fit

Goodness-of-Fit	
Statistic	Value
R-Square	0.4215
AICR	44505.57
BICR	44530.75
Deviance	76055543

As we see in Table 12 our model explains 42% of the variance. We expect better results when the other independent variables are included.

To define the outliers, the RobustReg procedure calculates a standardized residual. The residual is the difference (either negative difference or positive difference) between the actual premium stated in the dataset and the calculated premium based on the model. To calculate the standardized residual, the procedures estimates a scale. In this case the scale was estimated to be 41.34. For example, if the calculated premium based on the model was BR 1,254 while the actual premium on the dataset for the sample policy was BR 250, the absolute residual premium would be BR 1,004. Since the scale is estimated to be 41.34, the standardized absolute residual would be 24.28 (1,004 / 41.34). The outliers are defined based on a standardized residual point (Cutoff). If the absolute standardized residual lies below that cutoff point, the record is said to be normal. If it was above that cutoff point, the record is said to be an outlier. The RobustReg procedures allows you to set your own standardized residual cutoff point (*k*). The default cutoff point for the procedure is 3.

The following graph shows the distribution of the standardized residuals for the premiums in the dataset we selected.

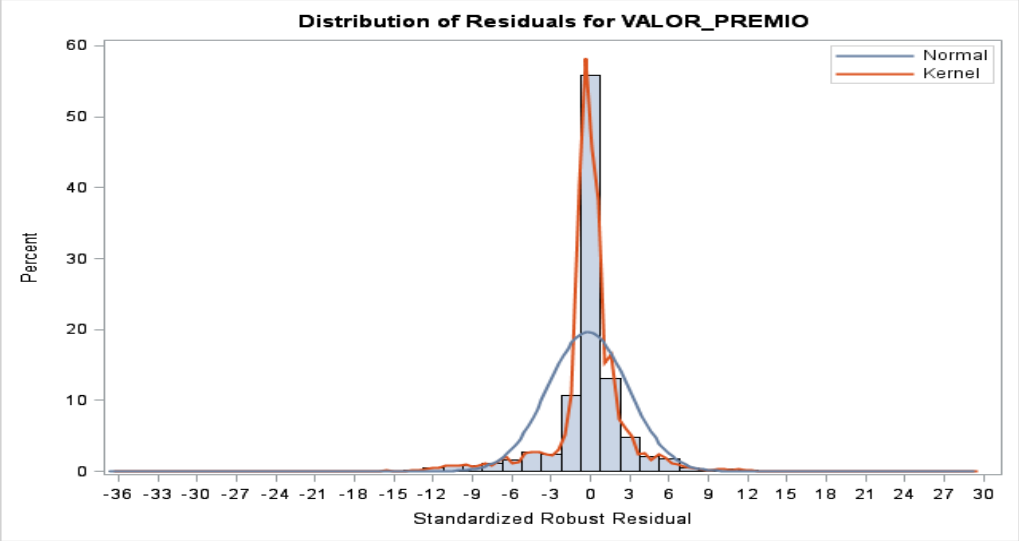

Figure 17: Distribution of Residuals

As seen in Figure 17, the standardized residual ranged from -36 to 30. Around 55% of the records we selected had a residual of zero. Around 85% of the records had a standardized residual between -3.0 and 3.0. Any data point with standardized residual beyond this range (-3.0 to + 3.0) will be considered an outlier.

The following table shows the outlier cutoff point used and the proportion of the outliers based on this cutoff.

Table 13: Outlier Cutoff Point

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

Using the default cutoff point of three, the following graph shows a visualization of the premiums and face values of the final 23,530 records of data based on the model.

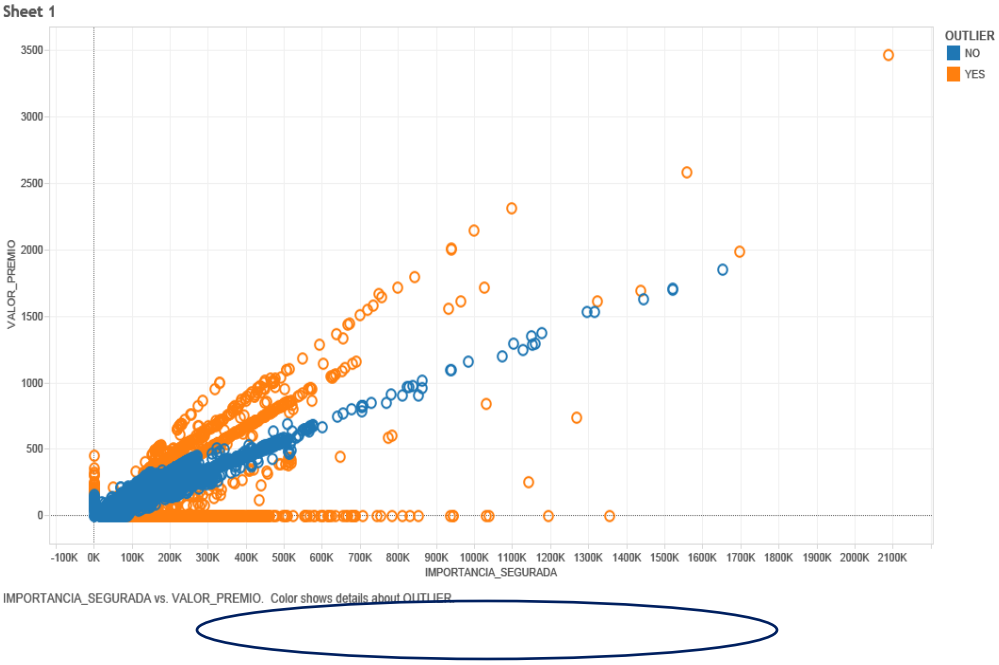

Figure 18: Premium-Face value Outliers (3.0 cutoff point)

The points shown in blue represents the records that were not flagged as outliers. That means that the absolute standardized premium residual in these cases was less than three. On the other hands, the points shown in orange represents the records that were flagged as outliers. Their absolute standardized premium residual was equal to or greater than three. The blue points shown in the graph divides the outliers into two groups. The first group which, is located on top of the blue line, shows records where the insurance company is collecting more premium than it should. The second group, which is located under the blue line, shows records where the insurance company is actually collecting less premium than it

should. While the first group is profitable to the insurance company, the second group is suspicious. An auditor might be interested in studying the second group more thoroughly. The most suspicious group of outliers is the one highlighted by the black oval. These records have almost an equal value of premium which is very close to zero while the face value is increasing close to BR 1,400,000.

As we mentioned before, due to the lack of some variables in our dataset that should be included in the premium calculations, we expect the model's performance to be negatively affected. There are other factors affecting the premium calculations that are not included in our model. To show an example of this fact, we grouped the age of the insured persons as follows.

Table 14: Age Groups

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

Since we controlled for insurance type, product type, and coverage type, the only two other factors remaining are the age and the face value.

We then studied the distribution of the standardized residual per each age group. If these were the only factors affecting the premium calculations, then the residual span in each group should be minimal. However, by studying the following graph, we realize that is not the case.

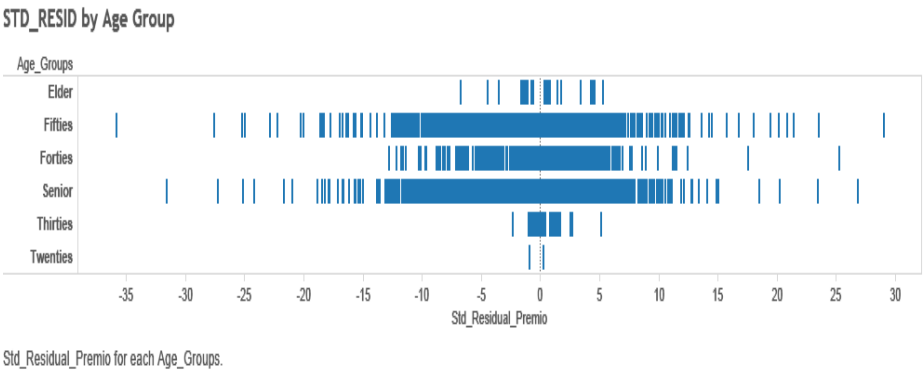

Figure 19: Standardized Residual Per Age Group

Figure 19 shows the standardized residual distribution per age group. As we see in the graph, the distribution of residual within each group varies dramatically in some groups. While in the twenties and thirties the distribution span is minimal, in other groups like the fifties and senior the span is huge from – 35 to +30. We suspect that this span will be probably explained by the health status data.

As we mentioned before, the procedures allow us to manipulate the cutoff points depending on our understanding of the data. We decided to stretch the cutoff to a very loose point of 8 instead of the default point of 3.

Table 15: Outliers at cutoff point 8.0

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

Even with a loose cutoff point of 8, we still get a large number of records defined as outliers. The following graph shows the visualization of the premiums and face values of the 23,530 records of data based on the model at the cutoff point of 8.0.

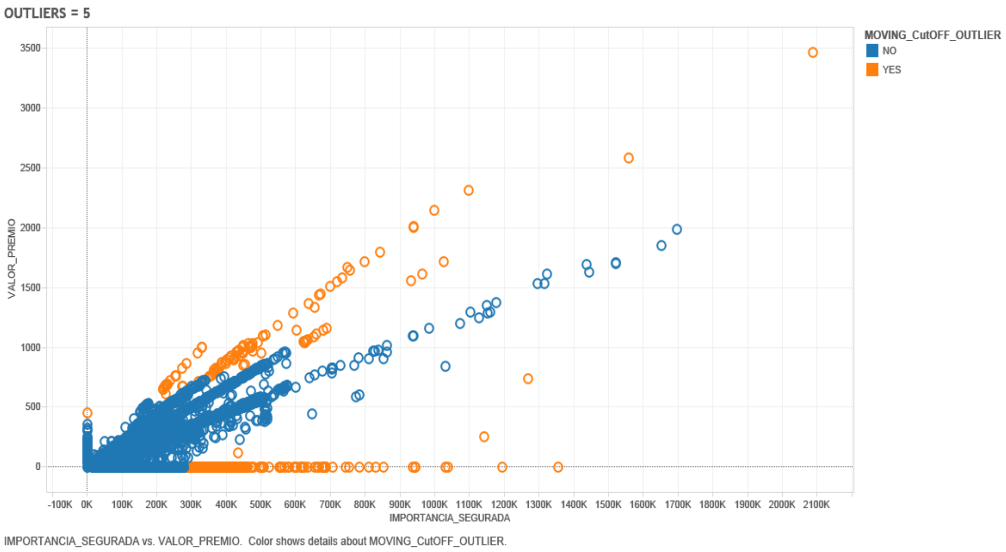

Figure 20: Premium-Face value Outliers at 8.0 cutoff point

Conclusion and Limitations

This research focused on detecting anomalies of Life/Disability Insurance. we first worked on the claim payment business cycle where we define a methodology to test two different audit assertions; is the claim settlement reasonable? Is the claim itself legitimate? We used a multi-dimensional approach in which we divided the attributes we have into different groups (dimensions). Our dimensions were Interest payments, reason-coverage association, timeline, and group similarities. We also used the belief function to give a risk score for each branch of the company we are using its data as an extra dimension. We used each dimension to logically find insurance claim anomalies. And then we use the weighted average

of the dimensions to priorities the anomalies we find. Our original plan was for the research to be empirical, but due to difficulties we experienced in obtaining the data we ended up with a design science chapter in which we used examples to show case our ideas. We then worked on the premium collection business cycle and built a model to detect premium outliers.

During our work on this chapter for both business cycles (Claims and premiums), we had a lot of difficulties in collecting the data. These difficulties have been the motivation for the next two chapters. Auditors do face many challenges in accessing the data they need to fulfill their duties and form their opinion even when their clients are fully digitalized and technologically capable of providing the needed data. The next two chapters deal with this problem and investigate different possible solutions.

Chapter Three: ADS Generation Framework and Interactive Auditor Dashboard: A Life / Disability Insurance example.

Introduction

Thanks to the current technology, companies are able to collect transactional data almost instantly making the availability of the data almost continuous (Alles et al. 2002). But, the auditor's problem in this case is not the data availability, it is the data's accessibility. Auditors face many challenges in accessing the data they need to fulfill their duties and form their opinion even when their clients are fully digitalized and technologically capable of providing the needed data. Zhang et al (2012) state that without open access to data, innovative audit tools and techniques might be disregarded. Researchers argued that there is a disparate need to standardize the data that should be available to the auditor (Moffitt and Vasarhelyi 2013; Vasarhelyi 2013; Zhang et al. 2012). The standardized data should facilitate the auditor's work by giving him/her access to the data he needs and also by paving the road to the standardized audit applications. Efforts toward issuing data standards have been already launched by the AICPA. The first data standards to be issued were standards for general accounts that most if not all the companies have in common like the General ledger, Accounts Receivables along with a Base Standard (Committee 2013a, 2013b). Later in 2015, AICPA updated the General Ledger and the Base Standards and issued two new standards; Order

to cash, and Procure to Pay (Committee 2015a, 2015b, 2015c, 2015d; Dai, Li, and Vasarhelyi 2016). The standards still face challenges. One of the main challenges is the enforcement of such standards. So far the standards are being issued as recommendations that are not enforced by the GASB. The hope is that in time as professionals grow more accustomed to the idea and as the standards are better established there will be some kind of mandate and enforcement.

So far, there is not any specific framework the ADS generators can follow. The current ADS generation process is based mainly on meetings and discussions between the AICPA committee members and a group of established auditors.

In this research we propose a framework to develop the Audit Data Standards (ADSs) for specific industries. As an example, we apply this framework on the insurance industry. In specific, we apply it on one business cycle of the life / disability insurance; the claim payment cycle. After generating the ADS for this cycle, we use them in creating an interactive audit dashboard that helps the auditor in performing his audit.

The following section will discuss the related literature of both the ADSs and the data visualization, then we will discuss the models used to generate the data standards and the interactive dashboard.

Related Literature

ADS

The AICPA Assurance Services Executive Committee believes that audit data standards (ADS) will contribute to the efficiency and effectiveness of the audit

process (Committee 2015a). Researchers have been advocating the need for the ADS in the auditing literature (Moffitt and Vasarhelyi 2013; Vasarhelyi 2013; Vasarhelyi, Alles, and Williams 2010; Zhang et al. 2012). Some of them argued that the ADS will facilitate the auditor's work and some argued that without them innovative audit tools and techniques might be disregarded. In 2013, the AICPA Assurance Services Executive Committee issued the first ADSs; Base Standard, General Ledger Standard, and Account Receivables Standard (Committee 2013a, 2013b). In 2015, the committee issued an update for the Base Standard and the General Ledger Standard and issued two new standards; Order to cash and procure to pay (Committee 2015a, 2015b, 2015c, 2015d). The standards are voluntarily applied as they are not enforceable. They are just recommendations and firms can defiantly provide more data to the auditor if they both agree on that. The standards provide a description of the data fields needed, their data type, format, and layout. They also provide guidance as to how to name the data files. The data are described in both flat file format and XBRL GL format. Each data standard describes the data files used, how they are related to each other, and how they are related to the data files in other standards if applicable.

The ADS generation model we are discussing in this chapter is based on the notion that auditing is industry specific. There is an ability to use the same audit approaches and procedures among industry clients which leads to audit firm specialization (Bills, Jeter, and Stein 2015; Cahan, Jeter, and Naiker 2011; Cairney and Young 2006; Fung, Gul, and Krishnan 2012). This ability has been established in the literature as early as the 1980s (Danos and Eichenseher 1982, 1986;

Eichenseher and Danos 1981). One of the key industry characteristics that allow for the transfer of audit processes across clients is operational homogeneity (Bills et al. 2015; Cairney and Young 2006). Operational homogeneity refers to the similarity of the cost structure of the firms in the same industry which allows the auditors to reapply their audit approaches to other clients in the same industry (Bills et al. 2015; Cairney and Young 2006). Firms of the same industry operate with the same type of technology under the same economic conditions and same regulations. This supports the notion that auditing is industry specific. Another industry-specific characteristic that supports the notion that auditing is industry specific is the accounting complexity (Bills et al. 2015). Complex accounting issues and procedures increase the risk of material misstatement, which requires specific audit plans and procedures to respond to these additional risks. Since firms of the same industry share the same accounting complexity, the auditing plans and procedures can be referred to as industry specific.

Visualization

The issue of comparing the usefulness of different presentation formats has been researched for a long time (Anderson and Mueller 2005; Kaplan 1988; Schulz and Booth 1995; Taylor and Anderson 1986). Most of the research compared between the effectiveness and/or efficiency of using graphical versus tabular presentation in decision making. The earlier research's results were not conclusive. Some of the research suggested that graphical presentation is not better than tabular one if not worse (Henry C. Lucas 1981; Lucas and Nielsen 1980; Lusk and Kersnick 1979; Watson and Driver 1983), while other research suggested that graphical

presentation leads to better decision making (Benbasat and Schroeder 1977; Feliciano, Powers, and Kearl 1963; Zmud 1978). The conflicting results at that time were mainly attributed to the variety of the tasks used in the experimental studies (Benbasat and Dexter 1986; DeSanctis 1984; Dickson, DeSanctis, and McBride 1986; Jarvenpaa and Dickson 1988; Jarvenpaa, Dickson, and DeSanctis 1985). These studies urged for a theoretical basis to base the selection of the presentation format on. (Vessey 1991) introduced the cognitive fit theory. The theory of cognitive fit proposes that a problem may be represented either spatially (graphically) or symbolically (tabular). The task required of the decision-maker may also be described as spatial or symbolic. A spatial task is where the meaningfulness of all the pieces of data taken together is greater than the sum of meaningfulness of the pieces taken separately. A symbolic task requires that one extract an individual data value (Vessey 1991). According to the cognitive fit theory of Vessey, if the type of presentation “matches” the type of task, the task will become less complex and decision makers will be both more efficient and more effective in decision making. Research findings generally support the theory of cognitive fit (Speier 2006; Tuttle and Kershaw 1998; Vessey and Galletta 1991). Although cognitive fit theory specifically addresses spatial and symbolic tasks, studies have examined other task types. For example, a study examined the effects of external problem representation on simple, range, and integrated tasks with results showing that accuracy improves when tables are used on range questions and graphs are used on integrated questions (Kelton, Pennington, and Tuttle 2010). Some studies (Frownfelter-Lohrke 1998; Kelton et al. 2010)

considered the learning curve with spatial graphical presentation. For example, Anderson and Mueller (Anderson and Mueller 2005) examined the interaction of experience with presentation format in auditing judgments. They specifically studied the effect on the analytical review judgments. Their results suggest that the data format or presentation (graphs) has more effect on students as opposed to actual practitioners (auditors with substantial experience with tables). They suggest further research on other factors and the inclusion of interactive graphs.

Some other studies examined the use of both presentation format at the same time. The results of these studies found that the use of a combination of tabular and graphical presentation provides the best information for decision making (Frownfelter-Lohrke 1998; Kelton et al. 2010). Since the financial statements are used for different types of decision task, the use of both formats proved to provide to be the most useful for the users of the financial statements.

3In summary, most of the studies use financial statements information for their research. We have not seen any use of big data in this matter. Mostly, they ask the participant to predict current balance based on previous years' balances. Also, we have not seen any interactive visualization studies on auditors or the use multi-variable graphs.

The ADS Generation Model

The model incorporates the views and experiences of the different parties involved in the process of auditing. In this paper, those parties are identified as the AICPA as representative of the professional society and the main supporter of the ADS,

the auditors as the ones actually performing the audit, the clients as the ones providing the data to be audited, the researchers as providers and reviewers of different types of audit analytics, and finally audit apps creators such as ACL and CASEWARE. The following graph summarizes the different parties and their input to the audit process.

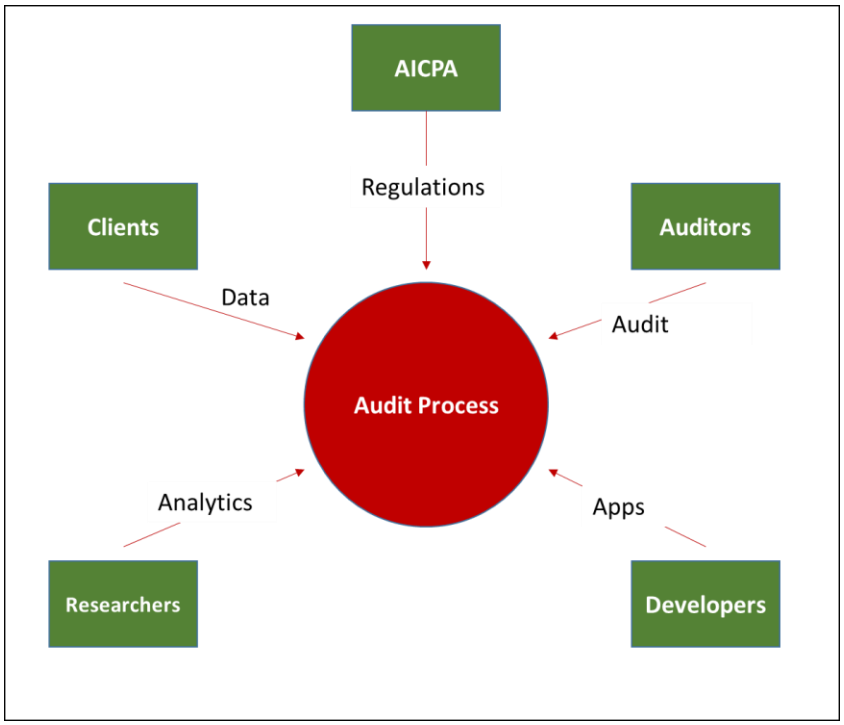

Figure 21: Different Parties involved in Auditing

As was explained in the "Related Literature - ADS" section of this chapter, the audit process is usually industry specific(Bills et al. 2015; Cahan et al. 2011; Cairney and Young 2006; Danos and Eichenseher 1982; Eichenseher and Danos 1981), the suggested ADS generation model starts with determining the specific industry the ADS will be generated for. This industry is then broken down into its different business cycles. The second step would be to focus on a specific business cycle and break it down into the basic functions or procedures that are

carried out within this cycle. We then go to the published AICPA audit guide. In their audit guide they provide industry-specific examples of audit procedures the auditor should perform in the audit. But since they are “examples”, these audit procedures do not constitute a full comprehensive audit plan. The ADS setters will now be able to generate a first version of the required ADS. For a more comprehensive view, the model uses actual industry-specific audit plans generated by experienced auditors. The audit plans serve two proposes. The first purpose is to get the auditor’s prospective on which data is needed to perform the audit and generate an opinion. The second purpose is to get the client’s prospective on which data the clients actually have and are able to provide. The more diverse the collected audit plans are in terms of the related clients, the more these audit plans will give an indications of the type of data the clients in this industry tend to keep. To generate an up to date ADS requirement, the model also incorporate the audit researchers’ developments and reviews of audit analytics by reviewing the literature and incorporating the needed data to perform the most significant audit analytics in this industry. As other developers of audit analytics, the work of major providers of audit apps (such as ACL and CASEWARE) (Dai et al. 2016) can also be incorporated into the ADS model. It has been argued though that the audit apps respond to the auditors’ needs but do not create them. In other words, they should not be part of the inputs of generating ADS, but rather they should be designed to use the output ADS. For that reason, we are not including them into the ADS generation model. Having modified the first version of the ADS

with the inputs from the last three parties (Auditors, researchers, and audit apps developers) we now have the second version of the ADS.

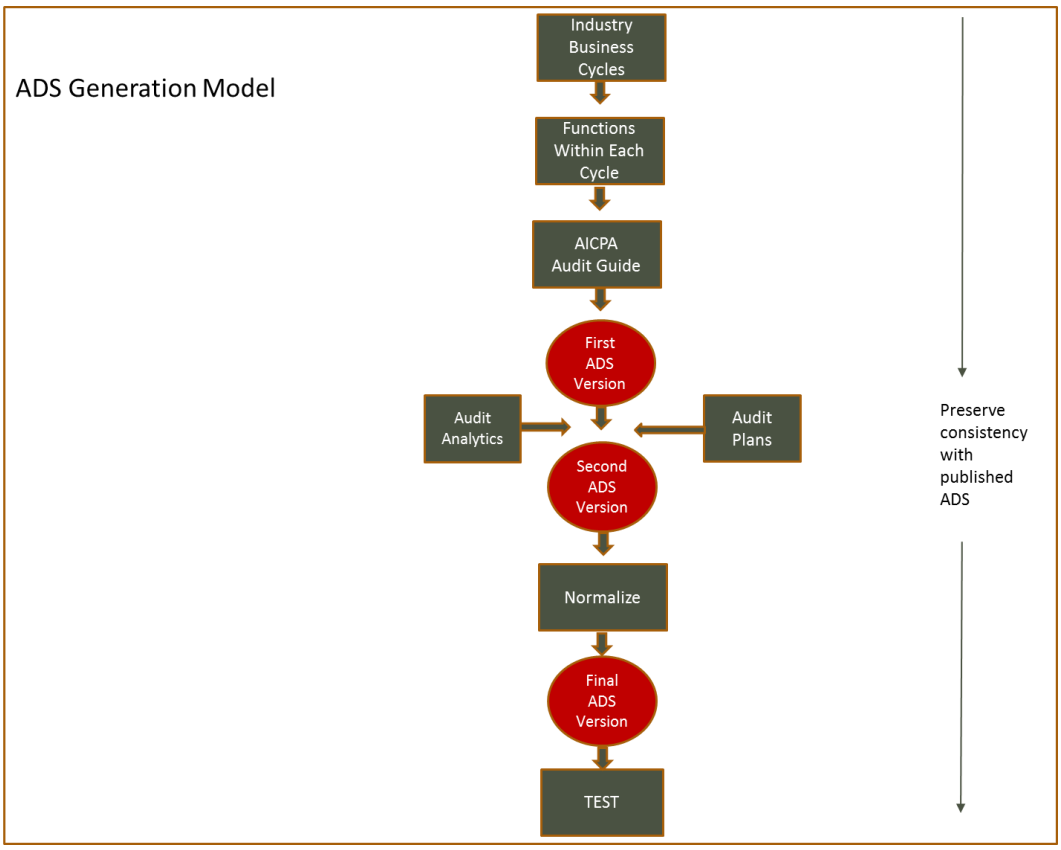

Figure 22: ADS Generation Model

Database normalization is the process of organizing the fields and tables of a relational database to minimize redundancy. Data redundancy occurs in database systems which have a field that is repeated in two or more tables. Normalization usually involves dividing large tables into smaller (and less redundant) tables and defining relationships between them. The objective is to isolate data so that additions, deletions, and modifications of a field can be made in just one table and then propagated through the rest of the database using the defined relationships.

Benefit and Claim Payments Cycle

Determining the Main functions within the business cycle

Figure 23 shows an example of the main functions within a claim payment cycle in an insurance company

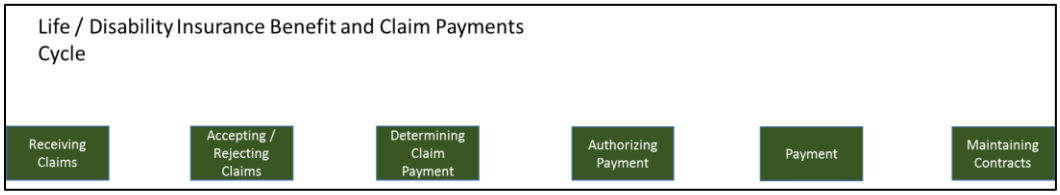

Figure 23: Claim Payment Cycle - Main functions

The insurance company receives a claim. A claim file is then established, assigned a sequential number, and recorded in the claim register. Then, the company has to decide whether to accept or reject this claim. The decision should be based on well-established policies. For example, the claim processing personnel determine whether the contract was in force at the time of the event and whether the claim is covered under the contract. If the company decided to accept the claim, the next step would be to determine the proper amount of money that it needs to pay in response to this claim. Again, determining the amount of payment must be based on well-established policies of the insurance company as well as the terms of the contract between the insurance company and the insured person. All relevant contract data is considered in calculating the benefit amount, including surrender charges, deductibles, and copayments for accident and health contracts, policy loans, advance premiums, dividends on participating contracts, and any other agreements in the contract. In the case of coinsurance, payment must be adjusted accordingly. Once the amount of payment is determined, the transaction must be

properly authorized. Then comes the actual payment step to the proper account. The appropriate benefit recipient is identified. The recipient could be the contract holder in the case of investment contracts, surrenders, or annuities; or he can be a designated beneficiary, estate, or trustee in the case of life insurance contracts. The last step related to claim payment is to maintain proper documentation of the transaction performed. When the benefit claim is either paid or denied, the claim file is annotated to indicate that payment was made. The claim register is updated to show that the claim has been closed and other applicable systems, such as statistical data relating to claims databases, outstanding amount of the insurance policy, status of the insurance policy, the profile of the insured person master files, benefit liabilities and related systems file are updated (AICPA audit guide).

First ADS Version

This section shows an example of how to use the AICPA's recommendations in generating the first ADS version. In this example we are using the AICPA guide for auditing a life and health insurance company. In specific, this example is using the AICPA recommendations concerning the audit objective "All benefits or claims paid or incurred represent valid obligations of the life insurance entity under the contracts in force".

Audit Recommendations – Understanding Internal Control

The AICPA audit guide for Life and Health insurance companies, paragraph .41 of AU section 314, states that “understanding the entity and its environment and assessing the risks of material misstatement”, the auditor should consider the following factors when attempting to understand the internal control of the company; the control environment, risk assessment, control activities, information and communication, and monitoring systems.

The following are some of the AICPA’s guidance to the auditor for internal control assessments for these types of insurance companies and the recommendations of the needed dataset fields.

- The control Environment represents the collective effect of the various factors on establishing, enhancing, or maintaining the effectiveness of specific control policies or procedures of the entity. The guidance recommends the following:
    - The Benefit operations are highly decentralized of benefit operations (approving claims, determining and processing amount due).
    - Staff is experienced or sufficient in relation to the complexity and volume of transactions involved in the benefit processing.

    For these audit guidance the focus is on the personnel of the insurance companies and their responsibilities in handling the insurance claims. We recommend fields describing the employees responsible for each data point in the expense cycle.

Emp_ID Emp_Name Level_Of_Education

Experience

- Control Activities are those policies and procedures that help ensure that management directives are carried out (AICPA audit guide). Whether the control activities are automated or manual, they have various objectives and are applied at various organizational and functional levels. The auditor should obtain an understanding of those control activities relevant to the audit. The guide provided the following examples of typical internal control activities related to the benefit and claim payments and also to other contract liability transactions. Each activity is followed by the recommended dataset fields.

    - Proper authorization of transactions and activities: Availability of written guidelines for claim processing, assigning appropriate individuals the responsibility for approval of benefit payments and determinations of amounts.

      For this audit activity, the auditor needs to gain an understanding of the employees' responsibilities within this business cycle as well as their authorization limits. The following fields are recommended:

      Emp_ID Emp_Name Position Responsibility

      Authorization_Limit Authorization_limit_Type

The responsibility field will describe the different responsibilities of the corresponding position. The authorization limits can be determined based on a specific period (monthly, weekly, daily) or as per transaction. That is why the authorization_Limit_Type is needed. The limit should be expected to vary based on the employee's position.

- Segregation of duties: Claims processing, benefit payments, premium billing and collection, key information systems functions, master file maintenance, and general accounting activities are appropriately segregated.

  For this activity, the auditor needs to collect information about the responsibilities of each employee involved in this business cycle to test whether the segregation of duties is appropriate. The recommended fields would be

  Emp_ID Emp_Name Position Responsibility

- Some of the AICPA's recommended activities for understanding internal control are not about the dataset fields in themselves as much as they are about the controls over the dataset itself. Therefore, we are not providing field recommendations for this kind of activities.

For example, one of the activities recommended by the AICPA to understand internal control is to check for adequate safeguard of access to and use of assets and accounting records, an activity that does not rely on the dataset fields.

Audit Recommendations – Auditing Procedures

The AICPA audit recommendations for life and health insurance companies also included some recommended auditing procedures to audit if all benefits or claims paid or incurred represent valid obligations of the life insurance entity under the contracts in force. They serve as examples only and are not all-inclusive. The following is a summary of these procedures and our recommended dataset fields needed to perform those procedures.

- Confirmation of benefits or claims paid to contract holders or beneficiaries in the period under review.

  To perform this procedure, the auditor needs information about the amount paid and the payee. The recommended fields are as follows

  Claim_ID Payee_Name Payee_address Approved_Amount Approved_Date Paid_Amount Pay_Date

- Individual benefit or claim payments are approved by appropriate personnel.

To perform this procedure, the auditor needs information about the approver of the claim payment. The recommended fields are as follows

Claim_ID Amount Approver_By Position Responsibility

- Verify the contract was in force at transaction date.

  To perform this procedure, the auditor needs information about the policy (contract) effective dates, the policy's status (active or canceled), and the event's date (that event that caused the claim). The recommended fields are as follows

  Contract_ID Beg_effective_date End_Effective_Date Status Claim_ID Event_Date

- Recalculate the benefit.

  To perform this procedure, the auditor needs information about the contract (policy) coverage, the event causing the claim. Since the amount payable for a specific claim is affected by the amount already paid on previous claims filed against the same contract, the information should be gathered on the claim payment currently being investigated and all previous claims filed against the contract. The recommended dataset fields are as follows.

Contract_ID Coverage Claim_ID Reason

Payable_Percentage Face_Value

Approved_amount

Based on the discussion above, the first ADS version in our example would be as follows:

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
1	**Emp_ID**	1	TEXT	100	Identifier that is unique for each employee. May require concatenation of multiple fields.
2	**Emp_First_Name**	1	TEXT	100	First name of employee
3	**Emp_Last_Name**	1	TEXT	100	Last name of employee
4	**Level_of_Education**	1	TEXT	256	Description of the level of education the employee has.
5	**Experience**	2	TEXT	256	Description of the kind of expertise the employee has.
6	**Position**	1	TEXT	25	The current position (title) of the employee.
7	**Responsibility**	2	TEXT	100	A description of the responsibilities of this employee in his current position.
8	**Authorization_Limit**	1	Num	-	Monetary value in local currency describing the maximum amount of money the employee is authorized to approve. If none, then it can be given a zero value.
9	**Authorization_Limit_Type**	2	TEXT	25	The type of the authorization limit (yearly, monthly, weekly, daily, transaction). A specific employee can have more than one limit type. For example, $100,000 monthly limit but not to exceed $50,000 per transaction.
10	**Claim_ID**	1	TEXT	100	Unique identifier for the claim. May require concatenation of different fields.
11	**Payee_First_Name**	1	TEXT	100	First Name of the person receiving the payment from the insurance company.

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
12	**Payee_Last_Name**	1	TEXT	100	First Name of the person receiving the payment from the insurance company.
13	**Payee_address**	1	TEXT	256	The address of the person receiving the payment from the insurance company.
14	**Approved_Amount**	1	Num	-	Total amount approved for this specific claim.
15	**Paid_Amount**	1	Num	-	The amount paid for this claim at the specified date (in case of installments).
16	**Pay_Date**	1	DATE		The date the paid amount was paid in.
17	**Approved_By**	1	TEXT	25	User ID (from User_Listing file) for person who approved the entry.
18	**Approved_Date**	1	DATE		The date the entry was approved.
19	**Contract_ID**	1	TEXT	100	Unique identifier for the contract. May require concatenation of different fields.
20	**Beg_Effective_Date**	1	DATE		The starting date of the contract.
21	**End_Effective_Date**	1	DATE		The ending date of the contract.
22	**Status**	1	TEXT	25	The current status of the contract (active, suspended, canceled, etc.)
23	**Event_Date**	1	DATE		The date of the event causing the claim (Date of death, date of accident, etc.)
24	**Coverage**	1	TEXT	100	The events that the insurance contract cover against (death, accidents, hospitalizations, etc.)
25	**Payable_Percentage**	1	Num	-	The percentage of the face value of the contract payable upon the occurrence of the specified event based on the terms of the contract.
26	**Face_Value**	1	Num	-	The total face value of the contract in local currency.
27	**Claim_Reason**	1	Text	100	The event causing the claim (Date of death, date of accident, etc.)
28	**Segment01**	2	TEXT	25	Reserved segment field that can be used for profit center, division, fund, program, branch, project, and so on.
29	**Segment02**	2	TEXT	25	Same as above.

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
30	**Segment03**	2	TEXT	25	Same as above.
31	**Segment04**	2	TEXT	25	Same as above.
32	**Segment05**	2	TEXT	25	Same as above.

Second ADS Version

To continue our example, we use the data analytics used in chapter two (Multi-Dimensional Approaches to Anomaly Detection) related to the same audit objective we used the first ADS version ("All benefits or claims paid or incurred represent valid obligations of the life insurance entity under the contracts in force") as an example of how to use the existing audit analytics to build on the first ADS version and create the second ADS version.

- Claim Initiation Step:
    - All Claims must have a corresponding Policy

      For this test the auditor only needs information about the claim and which policy (contract) it was filed against.

      Claim_ID Contract_ID

    - The corresponding policy must have a proper status.

      After finding the policy (contract) the claim was filed against, the auditor needs to check if the policy is in an active status, not suspended or canceled for any reason. The recommended fields are:

      Claim_ID Contract_ID Status

- All Claims must be against a covered item.

  The claim reason must be insured against in the policy contract. To test for this, the auditor needs the following:

  Claim_ID Claim_Reason Contract_ID Coverage

- The person Insured must be the one in the claim.

  The insured person in the contract must be the same person involved in the claim reason (Event).

  Claim_ID Event_Person Contract_ID Insured

- Date of occurrence must be between policy effective dates.

  The auditor needs the following fields

  Contract_ID Beg_effective_date End_Effective_Date Claim_ID Event_Date

- Claim Validation step:
  - No duplicate or unreasonable claims.

    The auditor checks for cases like two death events of the same insured person, a hospitalization after death, same kind of disability filed for more than once, etc.

    Claim_ID Reason Event_Person Event_Date Reason_Details Documents_Collected

- Appropriate Reason-Coverage Association.

  The claim should be filed against an appropriate coverage as per the insurance contract. For example, in case of a death event, it will be appropriate to file for a coverage such as funeral expenses and headstone cost but in case of a disability event, filing against these coverages should be a red flag for the auditor. The recommended fields are as follows:

  Claim_ID Claim_Reason Contract_ID Coverage

- Group Policies Peer Comparison.

  Only works in case of group insurance policies. The auditor puts the group policy holders into similar groups. We identify two “Group Policies” as being similar when both of them insure against the exact same coverage set, and works in same type of industry. Once we identify the similar policies we put them in groups and then test the claim filing pattern of each policy against the average of its group’s. For this we need the following fields

  Contract_ID Insured_Type Insurance_Type Industry Type Coverage Claim_ID Claim_Reason

- Timeline

Looking for anomalies using the timeline of the policy-claim. For example, the period between buying the contract and filing the first claim against it, and the period between filing all the needed document and approving the claim.

Contract_ID Issuance_Date Beg_Effective_Date
Claim_ID Event_Date Contact_Date
Filing_Date Approved_Date

- Claim Valuation and Payment Step:
  - Proper calculation of payment amount.

    Other than the previously mentioned recalculation of payment amount, the auditor need to check if proper interest calculation was made based on the number of days between the filing of the claim and the payment date.

    Contract_ID Coverage Claim_ID Claim_Reason
    Reason_Details Payable_Precentage
    Face_Value Approved_Amount Filing_Date
    Pay_Date Daily_Interest_Rate
  - Total amount of payments cannot exceed face value.

    Contract_ID Coverage Claim_ID Claim_Reason
    Reason_Details Payable_Precentage
    Face_Value Approved_Amount Pay_Date
  - Approver should not exceed authorization limits.

Approved_by Position Responsibility

Authorization_Limit Authorization_limit_Type

Claim_ID Approved_Date

- Payments are directed to the right beneficiaries.

  Claim_ID Payee_Name Payee_Address

  Contract_ID Beneficiary_Name

  Beneficiary_Address

Based on the discussion above, the second ADS version in our example would be as follows:

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
1	**Emp_ID**	1	TEXT	100	Identifier that is unique for each employee. May require concatenation of multiple fields.
2	**Emp_First_Name**	1	TEXT	100	First name of employee
3	**Emp_Last_Name**	1	TEXT	100	Last name of employee
4	**Level_of_Education**	1	TEXT	256	Description of the level of education the employee has.
5	**Experience**	2	TEXT	256	Description of the kind of expertise the employee has.
6	**Position**	1	TEXT	25	The current position (title) of the employee.
7	**Responsibility**	2	TEXT	100	A description of the responsibilities of this employee in his current position.
8	**Authorization_Limit**	1	Num	-	Monetary value in local currency describing the maximum amount of money the employee is authorized to approve. If none, then it can be given a zero value.

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
9	**Authorization_Limit_Type**	2	TEXT	25	The type of the authorization limit (yearly, monthly, weekly, daily, transaction). A specific employee can have more than one limit type. For example, $100,000 monthly limit but not to exceed $50,000 per transaction.
10	**Claim_ID**	1	TEXT	100	Unique identifier for the claim. May require concatenation of different fields.
11	**Payee_First_Name**	1	TEXT	100	First Name of the person receiving the payment from the insurance company.
12	**Payee_Last_Name**	1	TEXT	100	First Name of the person receiving the payment from the insurance company.
13	**Payee_address**	1	TEXT	256	The address of the person receiving the payment from the insurance company.
14	**Approved_Amount**	1	Num	-	Total amount approved for this specific claim.
15	**Paid_Amount**	1	Num	-	The amount paid for this claim at the specified date (in case of installments).
16	**Pay_Date**	1	DATE		The date the paid amount was paid in.
17	**Approved_By**	1	TEXT	25	User ID (from User_Listing file) for person who approved the entry.
18	**Approved_Date**	1	DATE		The date the entry was approved.
19	**Contract_ID**	1	TEXT	100	Unique identifier for the contract. May require concatenation of different fields.
20	**Beg_Effective_Date**	1	DATE		The starting date of the contract.
21	**End_Effective_Date**	1	DATE		The ending date of the contract.
22	**Status**	1	TEXT	25	The current status of the contract (active, suspended, canceled, etc.)
23	**Event_Date**	1	DATE		The date of the event causing the claim (Date of death, date of accident, etc.)
24	**Coverage**	1	TEXT	100	The events that the insurance contract cover against (death, accidents, hospitalizations, etc.)

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
25	**Payable_Percentage**	1	Num	-	The percentage of the face value of the contract payable upon the occurrence of the specified event based on the terms of the contract.
26	**Face_Value**	1	Num	-	The total face value of the contract in local currency.
27	**Claim_Reason**	1	Text	100	The event causing the claim (Date of death, date of accident, etc.)
28	**Event_Person**	1	TEXT	100	The person subjected to the claim event and caused the claim.
29	**Insured**	1	TEXT	100	Unique identifier of the insured person in the contract.
30	**Reason_Details**	1	TEXT	256	The details of the reason of the claim (if the reason is death, the details would be the cause of death, if the reason is disability, the details will mention the exact type of disability, etc.)
31	**Documents_Collected**	1	TEXT	256	The type of physical documents the insurance company collected to support the claim.
32	**Filing_Date**	2	DATE		The date the client filed all the required documents with the company, in case of a claim.
33	**Issuance_Date**	1	DATE		The date the policy (Contract) was issued
34	**Contact_Date**	1	DATE		The date the client first contacted the company regarding a claim.
35	**Insured_Type**	1	Boolean		Either person (Individual) or business entity (Group)
36	**Insurance_Type**	1	TEXT	100	Type of insurance (Life, Disability, Health, Auto, Home, etc.)
37	**Industry_Type**	2	TEXT	100	In case of group policies, this is the type of industry the policy holder operates in.
38	**Daily_Interest_Rate**	1	NUM		The daily interest rate applicable in case the insurance company doesn't pay the claim amount in time.
39	**Beneficiary_First_Name**	1	TEXT	100	First name of beneficiary
40	**Beneficiary_Last_Name**	1	TEXT	100	Last name of beneficiary
41	**Beneficiary_Address**	1	TEXT	256	Address of beneficiary.

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
42	**Segment01**	2	TEXT	25	Reserved segment field that can be used for profit center, division, fund, program, branch, project, and so on.
43	**Segment02**	2	TEXT	25	Same as above.
44	**Segment03**	2	TEXT	25	Same as above.
45	**Segment04**	2	TEXT	25	Same as above.
46	**Segment05**	2	TEXT	25	Same as above.

Dataset Normalization

To continue the example, we apply the dataset normalization rules on the second ADS version above to generate the final ADS Version.

Emp _Listing_YYYYMMDD

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
1	**Emp_ID**	1	TEXT	100	Identifier that is unique for each employee. May require concatenation of multiple fields.
2	**Emp_First_Name**	1	TEXT	100	First name of employee
3	**Emp_Last_Name**	1	TEXT	100	Last name of employee
4	**Level_of_Education**	1	TEXT	256	Description of the level of education the employee has.
5	**Experience**	2	TEXT	256	Description of the kind of expertise the employee has.
6	**Position**	1	TEXT	25	The current position (title) of the employee.
7	**Responsibility**	2	TEXT	100	A description of the responsibilities of this employee in his current position.

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
8	**Authorization_Limit**	1	Num	-	Monetary value in local currency describing the maximum amount of money the employee is authorized to approve. If none, then it can be given a zero value.
9	**Authorization_Limit_Type**	2	TEXT	25	The type of the authorization limit (yearly, monthly, weekly, daily, transaction). A specific employee can have more than one limit type. For example, $100,000 monthly limit but not to exceed $50,000 per transaction.
10	**Segment01**	2	TEXT	25	Reserved segment field that can be used for profit center, division, fund, program, branch, project, and so on.
11	**Segment02**	2	TEXT	25	Same as above.
12	**Segment03**	2	TEXT	25	Same as above.
13	**Segment04**	2	TEXT	25	Same as above.
14	**Segment05**	2	TEXT	25	Same as above.

Client_Master_YYYYMMDD (This table was taken from the AICPA Customer_Master – O2C to serve as the insurance company's client master with some modifications)

Field #	Field Name	Level	Flat File Data		Description
			DataType	Length	
1	Client_ID	1	TEXT	100	Identifier of the Client
2	Client _Name	1	TEXT	100	Name of client
3	Client_Type	1	BOOL	1	Individual (0) or business entity (1)
4	Industry_Type	2	TEXT	100	Type of industry the client works in (in case of business entity client).
5	Client_Physical_Street_Address2	1	TEXT	100	The physical street address line 2 of the Client.

Field #	Field Name	Level	Flat File Data		Description
			DataType	Length	
6	Customer_Physical_City	1	TEXT	100	The physical city where the Client is located.
7	Customer_Physical_State_Province	1	TEXT	6	The physical state or province where the Client is located. Recommend ISO 3166-2.
8	Customer_Physical_ZipPostalCode	1	TEXT	20	The zip code where the Client is physically located.
9	Customer_Physical_Country	1	TEXT	3	The country code where the client is physically located. Recommend ISO 3166-1 Alpha 2 or ISO 3166-1 Alpha 3 format (XX or XXX).
10	Customer_TIN	1	TEXT	100	The client's tax identification number.
11	Client_Billing_Address1	1	TEXT	100	The billing address line 1 of the Client.
12	Client_Billing_Address2	1	TEXT	100	The billing address line 2 of the Client.
13	Client_Billing_City	1	TEXT	100	The billing city of the Client.
14	Client_Billing_State_Province	1	TEXT	6	The billing state or province of the Client. Recommend ISO 3166-2.
15	Client_Billing_ZipPostalCode	1	TEXT	20	The billing zip code of the Client.
16	Client _Billing_Country	1	TEXT	3	The billing country code of the Client. Recommend ISO 3166-1 Alpha 2 or ISO 3166-1 Alpha 3 format (XX or XXX).
17	Active_Date	2	DATE		Date the Client declared active.
18	Inactive_Date	2	DATE		Date the Client was declared inactive.
19	Transaction_Credit_Limit	2	NUM		The per invoice credit limit established for this client.
20	Overall_Credit_Limit	2	NUM		The credit limit for this client's total outstanding balance.

Field #		Field Name	Level	Flat File Data		Description
				DataType	Length	
21	**2**	Terms_Discount_Percentage	2	NUMERIC		The discount percentage the client may take if an invoice is paid before a certain number of days. In the flat file, terms are represented as digits to one decimal place (for example, 10% would be represented as 10.0). In extensible business reporting language global ledger taxonomy framework (XBRL GL), the three fields Terms_Discount_Percentage, Terms_Discount_Days and Terms_Due_Days would be entered in the form "xx.x% dd Net dd," such as 2% 10 Net 30 for 2% discount if paid within 10 days, with the net due in 30 days.
22	**2**	Terms_Discount_Days	2	NUMERIC		The number of days from the invoice date the client has to take advantage of discounted terms. Terms are represented as digits with no decimal places (for example, nnn).
23		Terms_Due_Days	2	NUMERIC		The number of days allowed to meet the obligation before an invoice becomes overdue.
24	**2**	Entered_By	1	TEXT	100	User_ID (from User_Listing file) for person who created the record.
25	**2**	Entered_Date	2	DATE		Date the client was entered into the system. This is sometimes referred to as the creation date. This should be a system-generated date (rather than user-entered date), when possible.
26		Entered_Time	2	TIME		The time this client was entered into the system. ISO 8601 representing time in 24-hour time (hhmm) (for example, 1:00 PM = 1300).
27		Approved_By	2	TEXT	100	User_ID (from User_Listing file) for person who approved client master additions or changes.

Field #	Field Name	Level	Flat File Data		Description
			DataType	Length	
28	Approved_Date	2	DATE		Date the client master additions or changes were approved.
29	Approved_Time	2	TIME		The time the entry was approved. ISO 8601 representing time in 24-hour time (hhmm) (for example, 1:00 PM = 1300).
30	Last_Modified_By	2	TEXT	100	User_ID (from User_Listing file) for the last person modifying this entry.
31	Last_Modified_Date	2	DATE		The date the client record was last modified.
32	Last_Modified_Time	2	TIME		The time the entry was last modified. ISO 8601 representing time in 24-hour time (hhmm) (for example, 1:00 PM = 1300).
33	PrimaryContact_Name	2	TEXT	100	Name of the primary contact at the client.
34	PrimaryContact_Phone	2	NUMERIC		Phone number of the primary contact at the client.
35	PrimaryContact_Email	2	TEXT	100	Email address of the primary contact at the client.

Contracts_Master__YYYYMMDD

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
1	**Contract_ID**	1	TEXT	100	Unique identifier for the contract. May require concatenation of different fields.
2	**Issuance_Date**	1	DATE		The date the policy (Contract) was issued
3	**Beg_Effective_Date**	1	DATE		The starting date of the contract.

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
4	**End_Effective_Date**	1	DATE		The ending date of the contract.
5	**Status**	1	TEXT	25	The current status of the contract (active, suspended, canceled, etc.)
6	**Insurance_Type**	1	TEXT	100	Type of insurance (Life, Disability, Health, Auto, Home, etc.)
7	**Face_Value**	1	Num	-	The total face value of the contract in local currency.
8	**Insured**	1	TEXT	100	Unique identifier of the insured person in the contract (a foreign key from the table Client_Master)
9	**Approved_By**	1	TEXT	25	User ID (from User_Listing file) for person who approved the terms of the contract. This field is for keeping up with the already published ADS.
10	**Segment01**	2	TEXT	25	Reserved segment field that can be used for profit center, division, fund, program, branch, project, and so on.
11	**Segment02**	2	TEXT	25	Same as above.
12	**Segment03**	2	TEXT	25	Same as above.
13	**Segment04**	2	TEXT	25	Same as above.
14	**Segment05**	2	TEXT	25	Same as above.

Contracts_Coverage (this table is created because one single contract can cover more than one type of event).

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
1	**Contract_ID**	1	TEXT	100	Unique identifier for the contract. May require concatenation of different fields.

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
2	**Coverage**	1	TEXT	100	The events that the insurance contract cover against (death, accidents, hospitalizations, etc.)
3	**Face_Value**	1	Num	-	The face value per each type of coverage
4	**Entered_By**	1	TEXT	100	User ID (from User_Listing file) for person who entered the terms of the contract. This field is for keeping up with the already published ADS.
5	**Entered_Date**	2	DATE	-	Date of data entry

Contracts_Beneficiary (this table is created because one single contract-coverage can have more than one Beneficiary).

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
1	**Contract_ID**	1	TEXT	100	Unique identifier for the contract. May require concatenation of different fields.
2	**Coverage**	1	TEXT	100	The events that the insurance contract cover against (death, accidents, hospitalizations, etc.)
3	**Beneficiary**	1	TEXT	100	Unique identifier of the Beneficiary person in the contract (a foreign key from the table Client_Master)
4	**Beneficiary_Percentage**	1	Num	-	The beneficiary's share of the total Face_Value
5	**Entered_By**	1	TEXT	100	User ID (from User_Listing file) for person who entered the terms of the contract. This field is for keeping up with the already published ADS.
6	**Entered_Date**	2	DATE	-	Date of data entry

Claim_Master__YYYYMMDD_YYYYMMDD

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
1	**Claim_ID**	1	TEXT	100	Unique identifier for the claim. May require concatenation of different fields.
2	**Contract_ID**	1	TEXT	100	Unique identifier for the contract. Foreign key from the Contract_Master).
3	**Coverage**	1	TEXT	100	The events that the insurance contract cover against (death, accidents, hospitalizations, etc.)
4	**Claim_Reason**	1	Text	100	The event causing the claim (Date of death, date of accident, etc.)
5	**Reason_Details**	1	TEXT	256	The details of the reason of the claim (if the reason is death, the details would be the cause of death, if the reason is disability, the details will mention the exact type of disability, etc.)
6	**Event_Person**	1	TEXT	100	The person subjected to the claim event and caused the claim.
7	**Event_Date**	1	DATE		The date of the event causing the claim (Date of death, date of accident, etc.)
8	**Approved_Amount**	1	Num	-	Total amount approved for this specific claim.
9	**Approved_By**	1	TEXT	25	User ID (from User_Listing file) for person who approved the entry.
10	**Approved_Date**	1	DATE		The date the entry was approved.
11	**Documents_Collected**	1	TEXT	256	The type of physical documents the insurance company collected to support the claim.
12	**Contact_Date**	1	DATE		The date the client first contacted the company regarding a claim.
13	**Filing_Date**	2	DATE		The date the client filed all the required documents with the company, in case of a claim.
14	**Daily_Interest_Rate**	1	NUM		The daily interest rate applicable in case the insurance company doesn't pay the claim amount in time.

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
15	**Segment01**	2	TEXT	25	Reserved segment field that can be used for profit center, division, fund, program, branch, project, and so on.
16	**Segment02**	2	TEXT	25	Same as above.
17	**Segment03**	2	TEXT	25	Same as above.
18	**Segment04**	2	TEXT	25	Same as above.
19	**Segment05**	2	TEXT	25	Same as above.

Butcher_Table

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
1	**Insurance_Type**	1	TEXT	100	Type of insurance (Life, Disability, Health, Auto, Home, etc.)
2	**Coverage**	1	TEXT	100	The events that the insurance contract cover against (death, accidents, hospitalizations, etc.)
3	**Coverage_Details**	1	TEXT	256	The details of the reason of the claim (if the reason is death, the details would be the cause of death, if the reason is disability, the details will mention the exact type of disability, etc.)
4	**Payable_Percentage**	1	Num	-	The percentage of the face value of the contract payable upon the occurrence of the specified event based on the terms of the contract.

Claim_Payment_YYYYMMDD_YYYYMMDD

Field #	Field Name	Level	Flat File Data		Description
			Data Type	Length	
1	**Claim_ID**	1	TEXT	100	Unique identifier for the claim. May require concatenation of different fields.
2	**Payment_Date**	1	DATE		The date the entry was approved.
3	**Approved_By**	1	TEXT	25	User ID (from User_Listing file) for person who approved the entry.
4	**Payee**	1	TEXT	100	Unique identifier of the Person we paid to person in the contract (a foreign key from the table Client_Master)

Figure 24 below shows the table created for this example and their relationships. The tables benefit from the base standard's "User_List" table. It is connected to the "Employees" table for more information about the employees and their authorities. The clients table have all information about the clients of the insurance company whether they are insured, beneficiaries, or payees. Since one contract can have more than one coverage, we created the contract_coverage table. In the same manner, one contract_coverage can have more than one beneficiary so we created contract_beneficiary table. The claims' master file will be connected to the contract-coverage table as the claim is filed against a specific coverage within a contract. When an approved amount is set in the claims table, the claim_payments will track the actual payments, their dates and their receivers. An optional butcher table sets the percentage of payments applicable to each type of disability if needed.

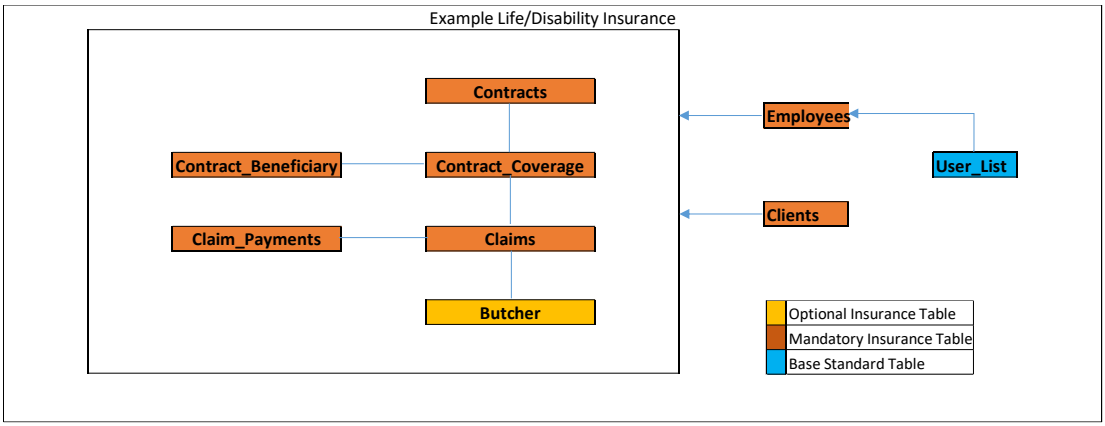

Figure 24: Example Life/Disability Insurance

The Interactive Auditor Dashboard Model

Following what we did in the ADS generation model, to build a dashboard we start with determining the specific industry the dashboard will be generated for. This industry is then broken down into its different business cycles (see Figure 25).

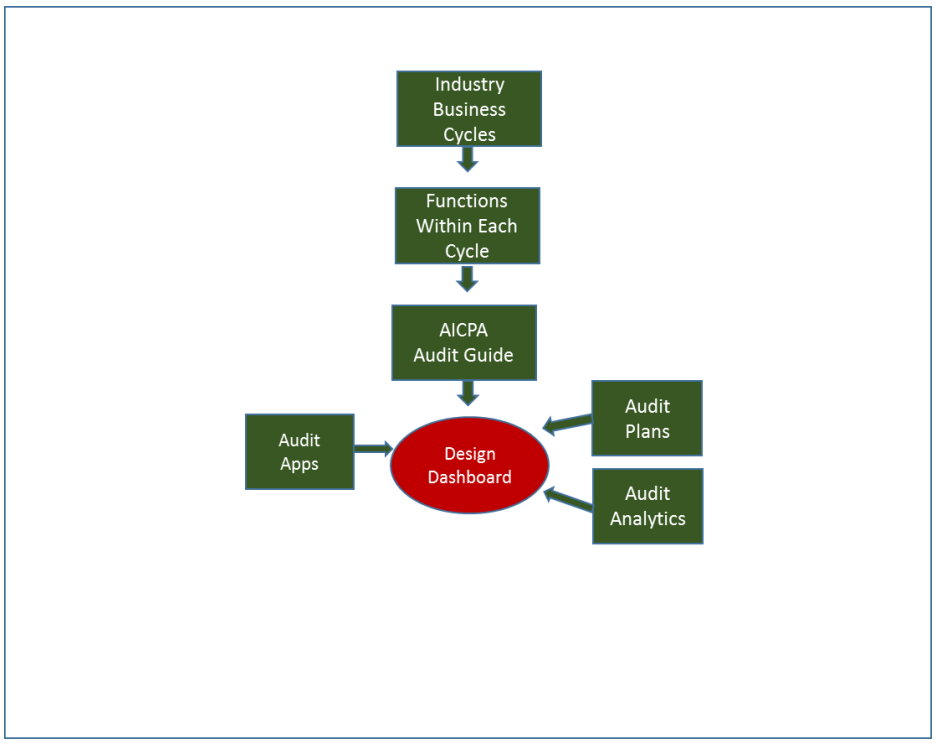

Figure 25: Building A Dashboard

The second step would be to focus on a specific business cycle and break it down into the basic functions or procedures that are carried out within this cycle. We then go to the published AICPA audit guide. In their audit guide they provide industry-specific examples of audit procedures the auditor should perform in his audit. But since they are “examples”, these audit procedures do not constitute a full comprehensive audit plan. We also incorporate actual industry-specific audit plans generated by experienced auditors and audit researchers’ developments and reviews of audit analytics by reviewing the literature. As another developers of audit analytics, the work of major providers of audit apps can also be incorporated into the dashboard. After that, we end up with a set of audit objectives for each functions within the specific business cycle. At this point we start creating the auditor dashboard.

The Auditor Dashboard

For the purpose of this paper, we are creating the auditor dashboard for an insurance company. In specific, we are creating it for two business cycles; the expense cycle (claims payments) and the revenue cycle (Premium collections). The following two subsections describe the dashboard created.

The Expense Cycle (Claims Payments)

We have broken the expense cycle into different functions and assigned different audit tasks to each function. The Figure 26 summarizes the expense cycle (claims payment) for our company.

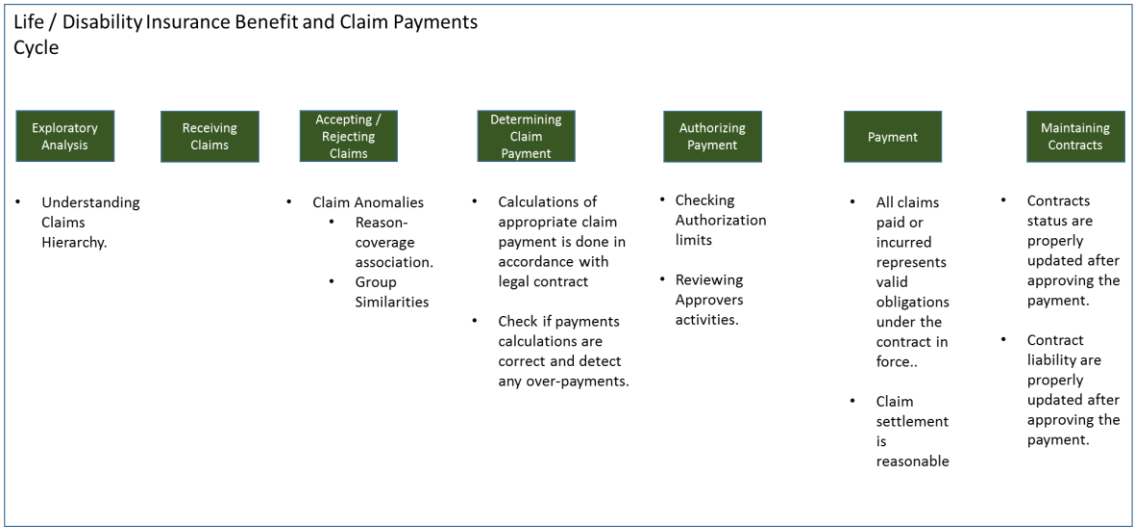

Figure 26: Expense Cycle - Dashboard

Exploratory Analysis

As shown in Figure 26, the dashboard starts with exploratory analysis. The auditor is able to look at the claims hierarchy. The claim hierarchy shows the total paid claims first then it can break it down to see the claims paid by each branch of the

company, by each type of insurance, by each insurance product, by each policy, and finally down to by each single claim. The auditor can also choose to see only claims within a specific range of monetary value. We believe this is important for the auditor to understand the population of the claims and set his audit plan.

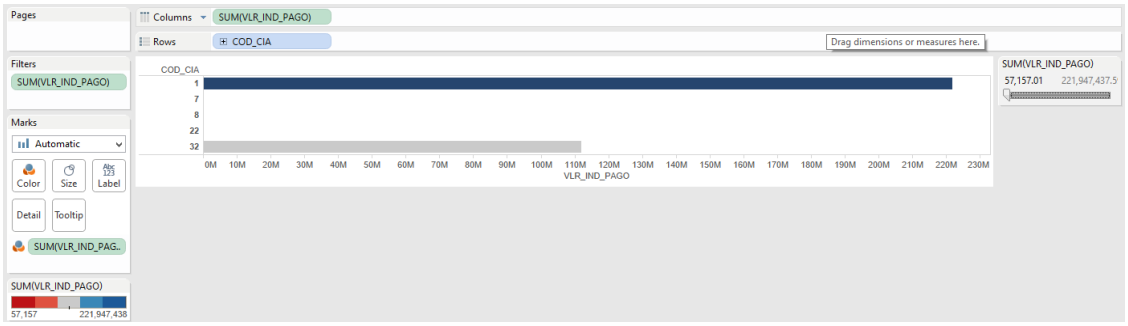

Figure 27: Claim Hierarchy

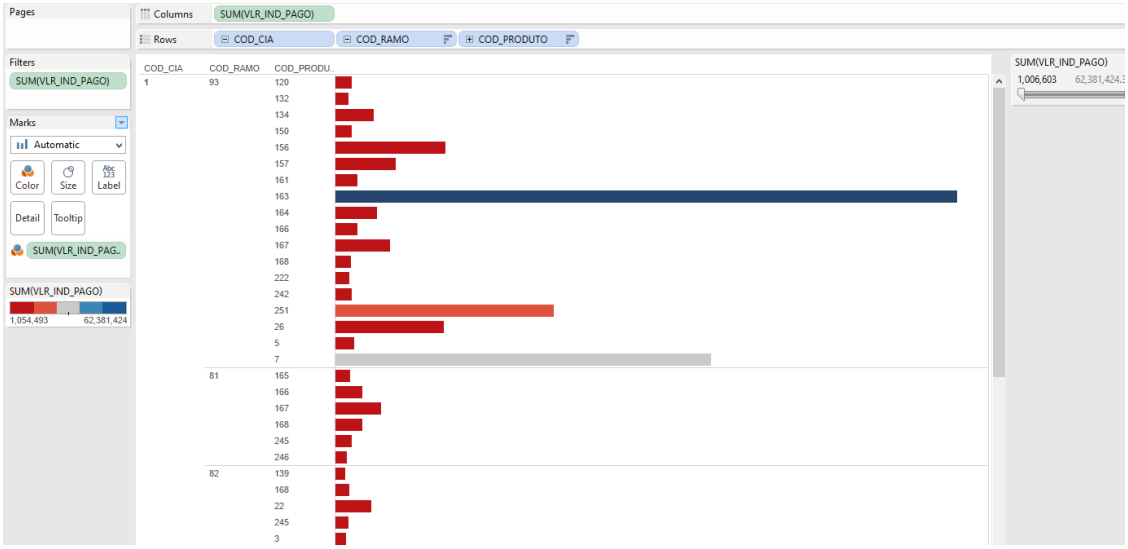

Figure 28: Claim Hierarchy - Drill Down

Another graph that can be used by the auditor in his exploratory analysis is Figure 29.

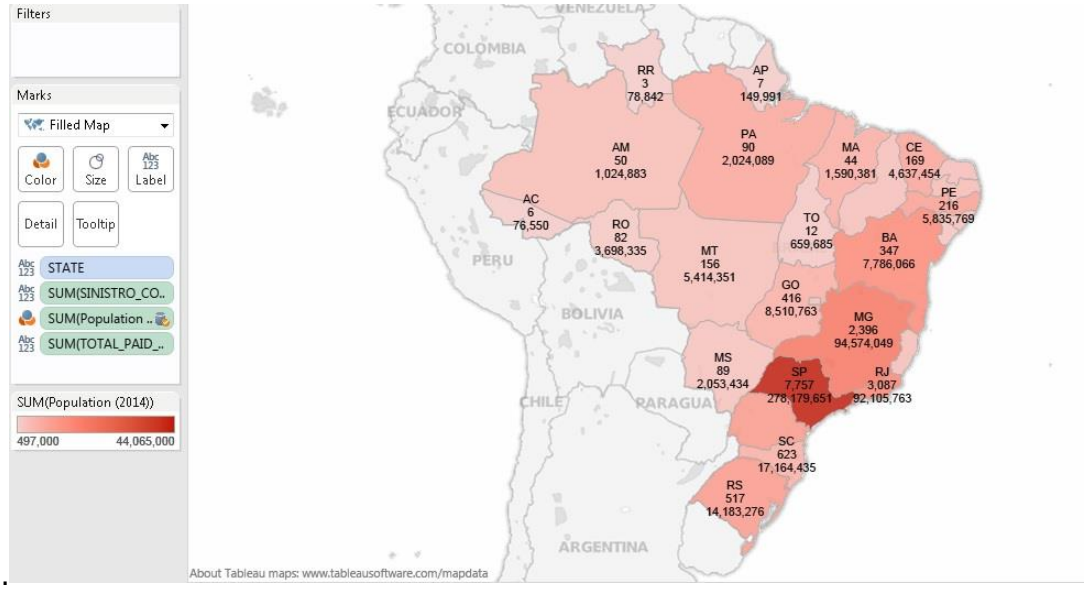

Figure 29: Graphical Distribution of the Claims

The graph shows the geographical distribution of the claims. The color intensity represents the population of the state. The darker the color, the bigger the population. The small numbers shown in each state are the number of claims originated from this state. The big numbers are the total amount paid against these claims. By examining Figure 29, the auditor will be able to determine where most of the activities are. He would then be able to decide if he needs to send auditors to specific locations.

Figure 30 gives more explanation of the geographical distribution of the claims.

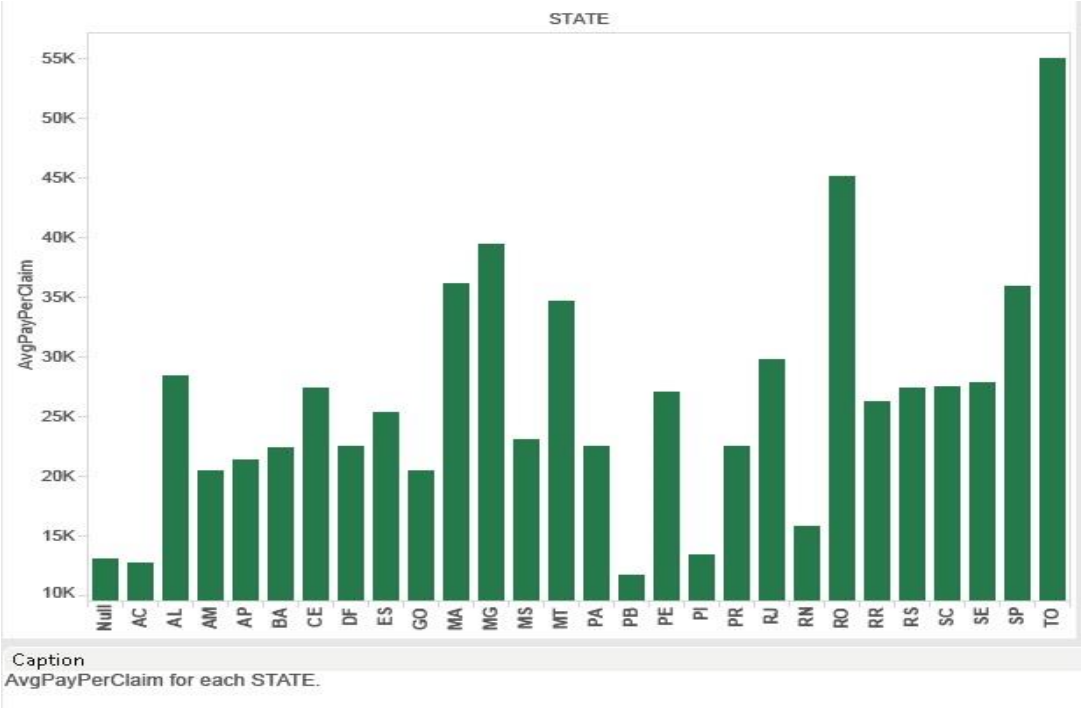

Figure 30: Avg Pay Per Claim for Each State

The graph shows the average pay per claim rate for each state. The average pay per claim is calculated by dividing the total claim payments for each state over the total count of claims in this state. By examining this graph, the auditor would be able to determine the state or region with higher average pay per claim. He might want to investigate why, for example, “TO” has a higher rate than all of the other states. He might also want to investigate the significant difference in the population characteristics between PB state (the lowest rate) and TO state (the highest rate) that causes this huge difference in the average pay per claim rate.

Figure 31 further investigate the geographical property of the claims.

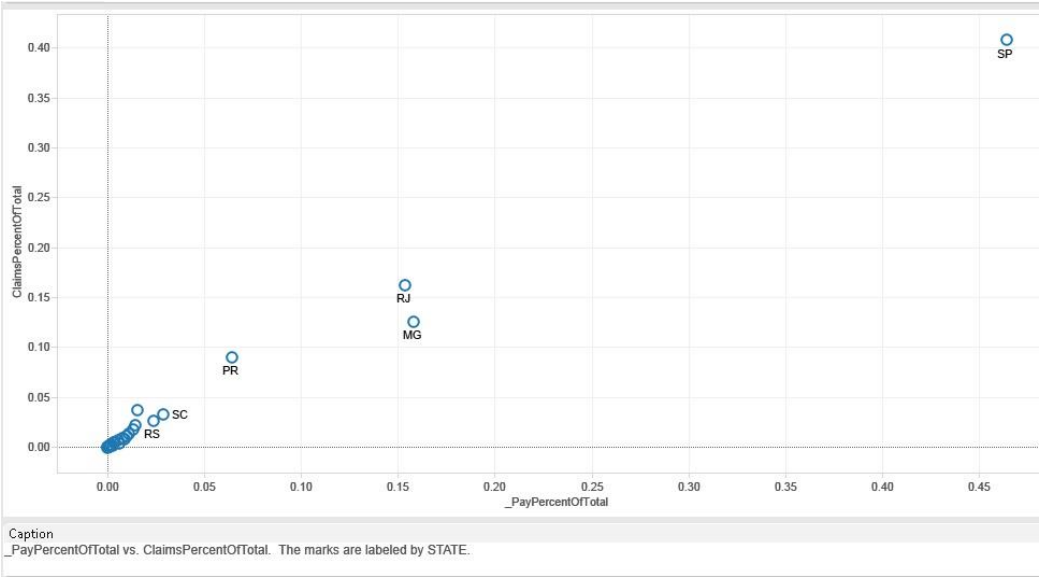

Figure 31: Share of Total Claim Payment VS. Share of Total Claim Count

The graph shows the relationship between each state's percentage share of total claim payments and its percentage share of total counts of claims. For example, let's say that the total payments of claims for all the states is BR 100,000 and the total number of claims for all the states is 400 claims. If a specific state, say PR has a total of 50 claims and a total payments of BR 20,000, then PR's percentage share of total claim payments is 20% (20,000 / 100,000) and its percentage share of total number of claims is 12.5%.

Authorizing Payments

Figure 32 provides the auditor with a comparison of the authorization limit of each claim approver during a period of time (shown in green) and what he actually approved during that period of time (shown in orange).

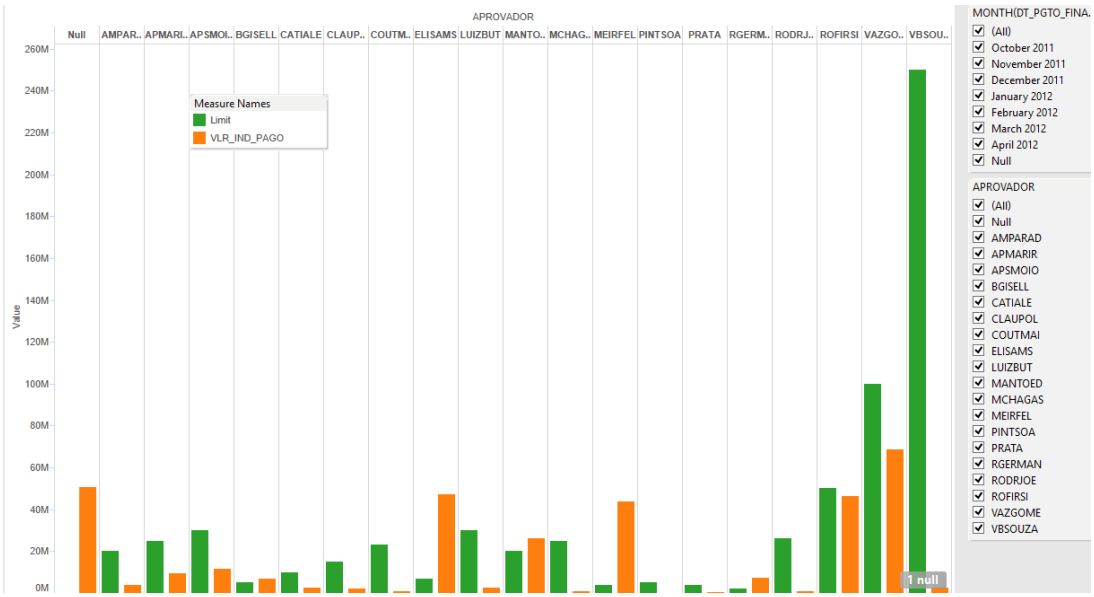

Figure 32:Payment Authorization - Actual Vs. Limits

The auditor can also manipulate which provider(s) to see or which month(s) of a year he wants to review. If the auditor sees something he wants to investigate more, he can right click any approver to view the actual data related to this approver.

Another view that helps the auditor understands the approvers' activities is shown in Figure 33.

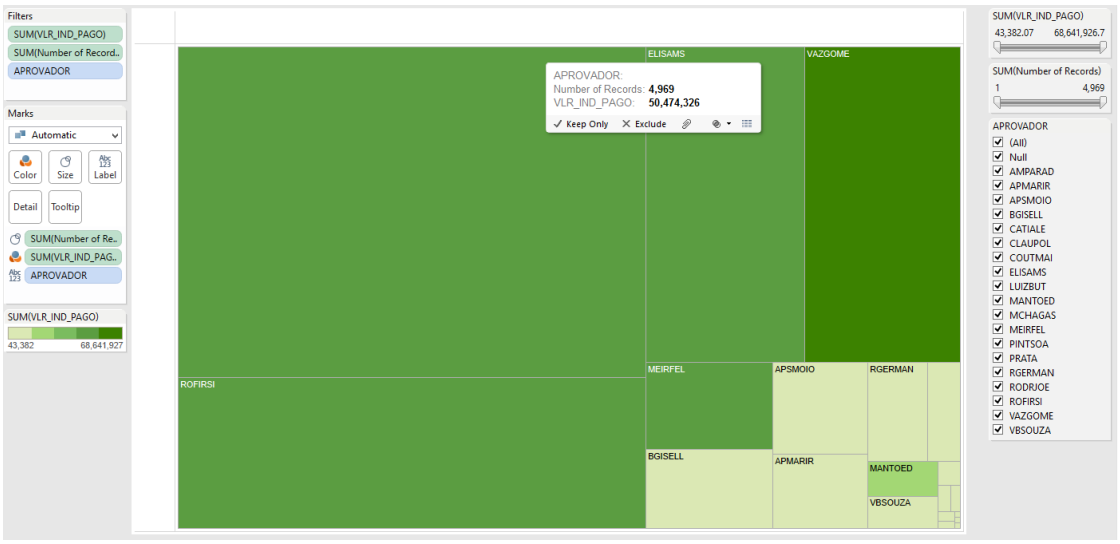

Figure 33: Approvers' Activity

The graph shows both the number of transactions authorized by a specific approver (The higher the number of claims, the bigger the size of the square) and the total monetary value he authorized as claim payments (the higher the value, the darker the green). The auditor can also filter by claim monetary value and/or approver's name. What is interesting about this graph is that it summarizes the whole activities of the approvers. An auditor might be interested in the approver who authorized the most number of claims. Or, he might be interested in the one who approved the highest payments. The graph will also show the auditor the approver who approved the least number of claims. If we looked in the lower right corner, we will see an approver who approved only one claim. An auditor might be interested in knowing why ONLY this claim.

Figure 34 shows a different type of charts an auditor can find useful.

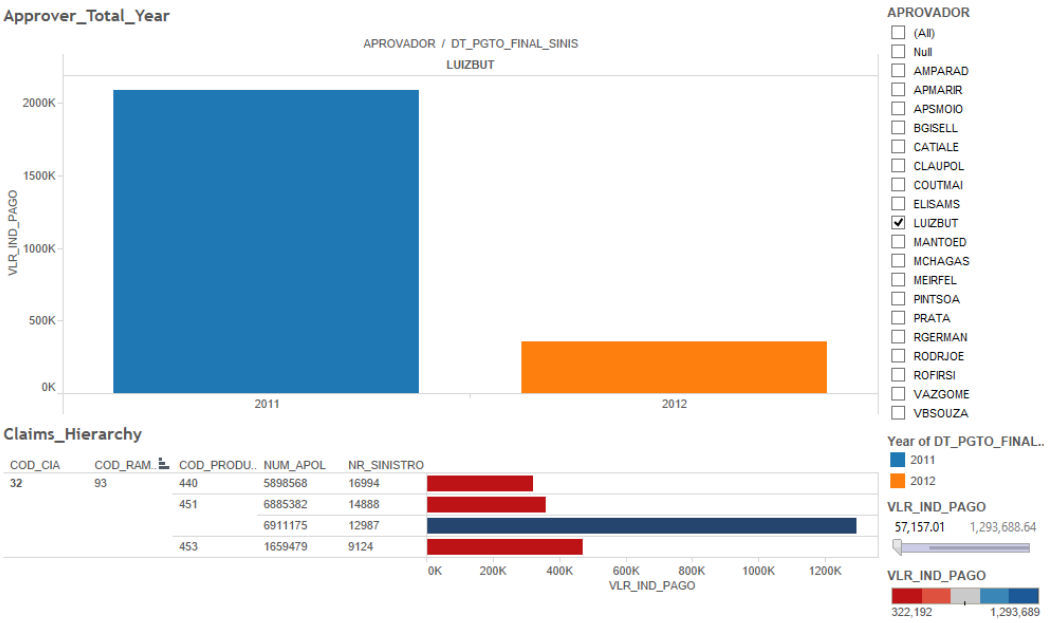

Figure 34: Approvers' Activity - Detailed by Split View

Figure 34 is split into two sections. When an auditor selects a specific approver, the top section of the chart shows total claim payments this approver has authorized over the years. The lower section shows the details of these claims in a claim hierarchy way (starting from the branch of the company who paid the claim and broken down into the type of insurance, the type of product, the specific policy, down to the specific claim). The auditor can choose not to see all the claims but only those which falls within a specific monetary value or happens in a specific year. Figure 34 can help the auditor understand which claim (or product) contributed the most to the approver's total activity. He can also identify if an approver is working on a product (or in a branch) that he shouldn't be working on (in).

Payment

The policies and claims we are working on are Disability and Life policies. The reason of the claims could be either Death, in this case the company is expected to pay the full amount of the face value. Or, the reason might be a disability, in this case the company will only have to pay a portion of the face value. If the company paid the claim after 30 days from the day they received all the required documents, they will have to pay an interest for each day beyond the allowed 30 days. That is why we thought Figure 35 is going to be useful to the auditor.

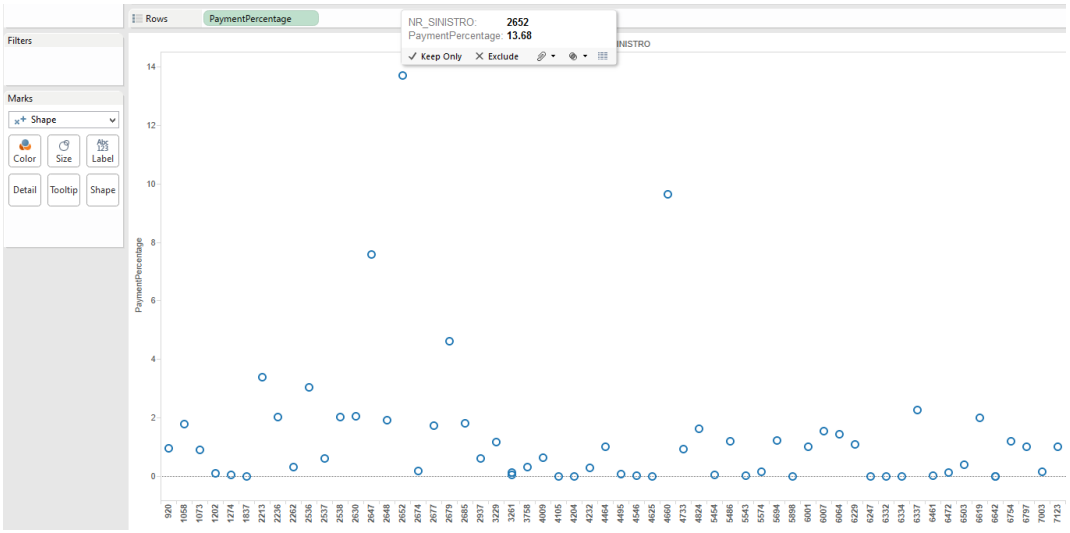

Figure 35: Claim Payment Percentage of Face Value

Figure 35 shows each claim's payment percentage of the face value (amount paid / face value). From Figure 35, an auditor can easily spot the claims he wants to investigate more. The highest percentage shown in this graph represent a payment of thirteen times of actual face value. As been said before, the auditor can easily see the actual dataset of the claim by a right click.

Another chart an auditor can find useful in studying the reasonability of the payments of claims is shown next.

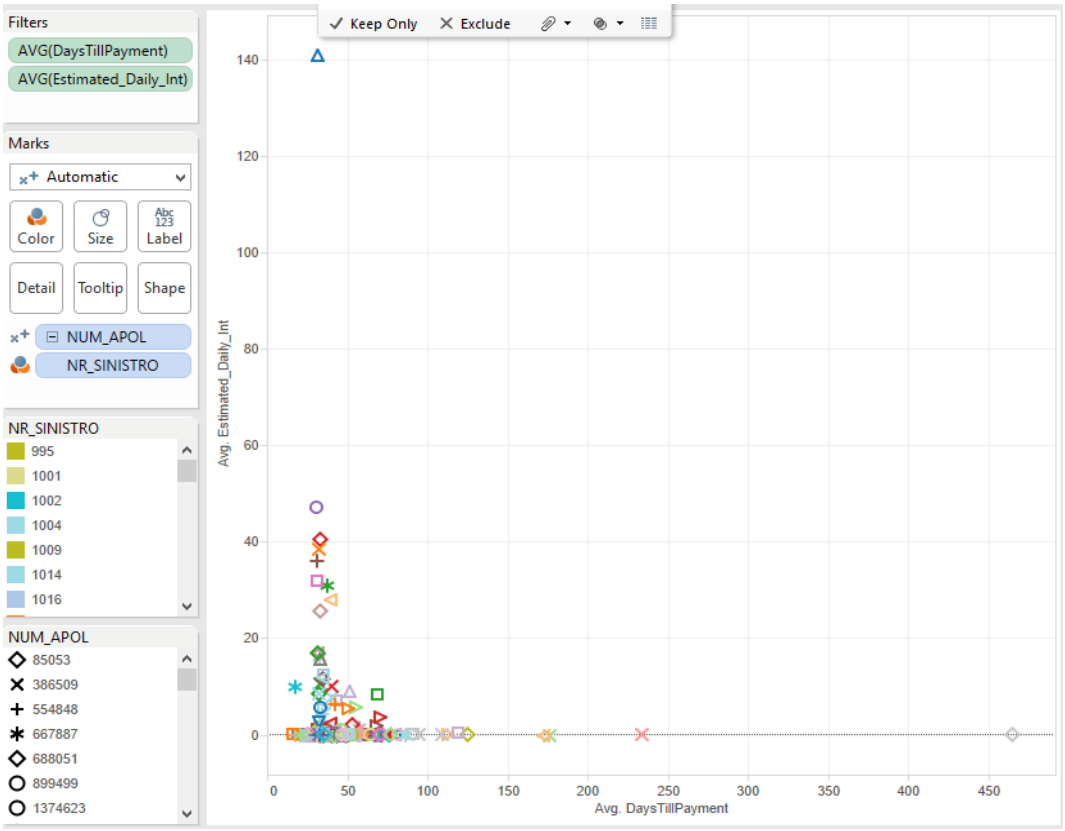

Figure 36: Estimated Daily Interest and Days till Payment

Figure 36 shows the relationship between the estimated daily interest ([(amount paid – face value)/Face value]/days till payment – 30) and the number of days the company took to pay the claim beyond the allowed 30 days. From the chart an auditor can easily see where the relationship does not make sense (higher rate with fewer days till payments). The auditor can filter by number of days or by interest percentage. If he wants to investigate a specific claim, he can see the actual dataset by right clicking the specific node that he wants to see the data for.

Figure 37 shows the auditor all the approvers who approved claims and the maximum interest rate each of them authorized. The higher the rate the bigger the circle.

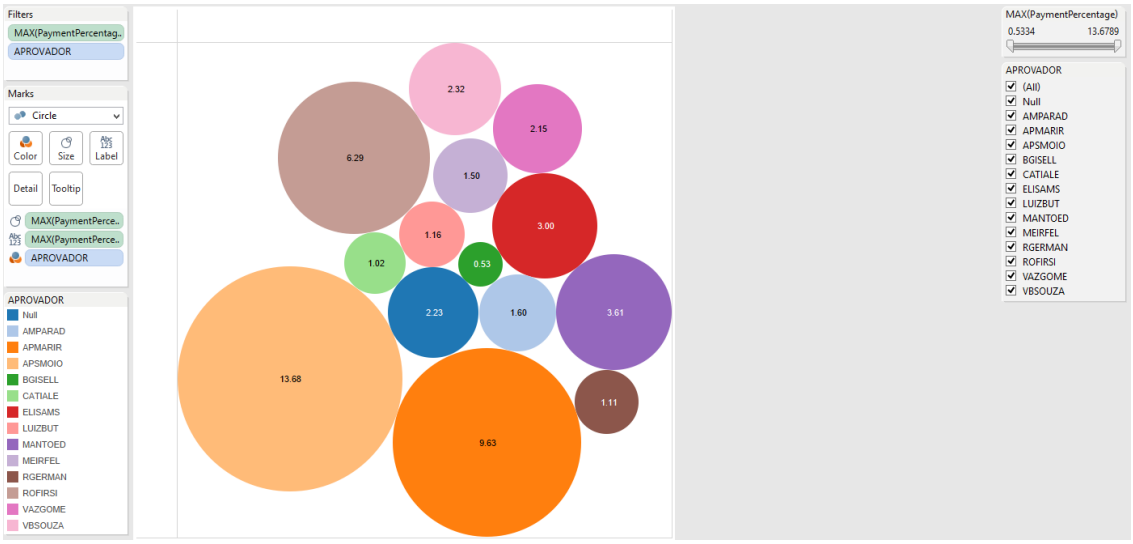

Figure 37: Approvers Interest Rates

From Figure 37, the auditor can spot the approver(s) who authorizes unreasonable interest rates.

Conclusion and Limitations

This chapter is a design science research. We proposed a framework to develop the Audit Data Standards (ADSs) for specific industries. So far, the standards that has been issued do not follow any model and are not industry specific. As an example, we applied this framework on the insurance industry. In specific, we applied it on the audit objectives related to the approval, valuation, and payment of Life/Disability insurance claims. When applying the model, we noticed that sometimes we are either going to need to update the already published ADS or to

sacrifice some of the data normalization rules. For example, the base ADS includes a table for system users where the User_ID, User_Name, Title and responsibilities are defined for every employee who has access to the system. When it comes to the insurance company audit, the auditor needs to gain understanding of the education level and expertise of the employees regardless of whether they have access to the system or not. In a perfect database world, a table would be generated for all the employees and their details along with a User_ID field for those employees who can access the system. But since the Base standard is already published we had to sacrifice the normalization rules and ended up with two tables describing employees. After generating the ADS, we used them in creating an interactive audit dashboard that helps the auditor in performing his audit.

We need to mention some limitations of this research. First of all, we did not have any life/disability insurance audit plans to integrate into the model. Secondly, to keep this chapter in an appropriate size we had to limit the scope to only the audit objectives related to the approval, valuation, and payment of Life/Disability insurance claims instead of the whole business cycle.

Chapter Four: USING MACRO ECONOMIC FACTORS TO IMPROVE AUDIT ANALYTICAL PROCEDURES

Introduction

Recently, there has been much of pressure exerted over the auditing profession to improve and to keep up with the current technology and the changes in the investors' needs. Advocates for continuous auditing and continuous assurance emphasis that the archival audit where the auditor go to the audited entity at the end of the year to examine financial statements and issue an ex-post opinion on these statements should, and will, be replaced by a more timely assurance services (Alles et al. 2006, 2002; Vasarhelyi et al. 2004). The idea of the continuous auditing and assurance has emerged, in part, from the availability of the required technology. Thanks to the current technology, companies are able to collect transactional data almost instantly making the availability of the data almost continuous (Alles et al. 2002). But, the auditor's problem in this case is not the data availability, it is the data's accessibility. Auditors face a lot of challenges in accessing the data they need to fulfill their duties and form their opinion even when their clients are fully digitalized and technologically capable of providing the needed data. (Zhang et al. 2012) states that without open access to data, innovative audit tools and techniques might be disregarded. Researchers argued that there is a disparate need to standardize the data that should be available to the auditor (Moffitt and Vasarhelyi 2013; Vasarhelyi 2013; Zhang et al. 2012). The standardized data should facilitate the auditor's work by giving him access to the data he needs and also by paving the road to the standardized audit applications.

Efforts toward issuing data standards have been already launched. The first data standards to be issued were standards for general accounts that most if not all the companies have in common like the General ledger and Accounts Receivables. The standards still face challenges. One of the main challenges is the enforcement of such standards. So far the standards are being issued as recommendations that are not enforced by the GASB. The hope is that in time as professionals grow more accustomed to the idea and as the standards are better established there will be some kind of enforcement.

While the data standards might take some time before auditors can use them, researchers are trying to find other ways to enhance the performance of their analytical procedures even with the restricted access to clients' data. Analytical reviews are supposed to help the auditor reach a reasonable expectations of what the account balances should be so he can determine the extent to which the actual balances deviate from these expectations (Lev 1980). The main stream of research focused on improving the performance of statistical models used in analytical procedures to predict the account balances and detect any discrepancies. Since firms do not operate in vacuum, it is expected that economy wide factors and industry wide factors affect the operations of companies operating in such economy and in the specific industry (Lev 1980). One of the research studies that aimed to improve the performance of the statistical models used in analytical procedures proposed the use of Gross National Product "GNP" and Total Corporate Profits after tax "TCP" to improve the prediction models of a company's sales, operating income and net income (Lev 1980). Another research proposed

the use of data from peer companies to improve the prediction and error detection performance of statistical models (Hoitash et al. 2006).

In this research, we propose using macro-economic indicators to improve the prediction and error detection performance of the statistical models. We add to Lev's 2008 research the use of monthly data instead of the annual data and the use of multiple macro indicators. We test the effectiveness of the macroeconomic indicators in detecting coordinated errors and in mitigating the effects of misstated accounts. We also test the effectiveness of using macroeconomic indicators with peer data.

In the following sections we first discuss previous research, then we set our research questions, we then talk about the data and the models, and in the last part of the paper we will discuss the results and conclude.

Related Literature

There are several research studies investigating different ways to improve the performance of the analytical review models in detecting errors in the account balances. They can be grouped in three different groups; a group of research studies focused on the choice of statistical techniques, another group focused on the aggregation level of the data used in the models, and a third group focused on the use of external data sources to improve the performance of the models.

For the aggregation level of the data, there has been a lot of prior research studies that compared between the use of monthly data and quarterly data (Chen and Leitch 1998, 1999; Cogger 1981; Dzeng 1994; Kinney and McDaniel 1989;

Knechel 1988). Their results suggested the superiority of the disaggregated monthly data over the quarterly data as it increased the effectiveness of their analytical procedures. Chen and Leitch (1998) explained in their paper that the increase of effectiveness was due to three reasons. The first reason is that the use of monthly data provides a larger sample size. The second reason is that because the monthly data with its larger sample size allows the researcher to use a shorter time span. The shorter time span provides more stable results as it is less influenced by structural changes in the organization. The third reason according to Chen and Leitch (1998) is that the monthly data will be more correlated than the quarterly data.

For the use of external data, (Lev 1980) proposed the use of annual Gross National Product “GNP” and annual Total Corporate Profits after tax “TCP” to improve the prediction models of annual firm-specific variables represented in the company’s sales, operating income and net income. He explained that index models are capable of characterizing some important systematic attributes of accounting numbers reflected by the relationships between the firm and the environment (industry and economy) within which it operates and so he expected that this relationship would be stable over a period of time and allow him to better predict the firm-specific variables. His research concluded that the macroeconomic indicators resulted in lower Mean Absolute Percentage Error (MAPE) than this resulted from the single use of firm-specific variables. Other researchers used external data from the same industry (industry-specific) data. For example, many studies (Hogan and Jeter 1999; Kwon 1996; Taylor 2000; Wright and Wright 1997)

look at industry specialization and the effects of such specialization on audit performance, audit fees, and economies of scale among other factors. Another research explored a different angle of the industry effect. Hoitash et al proposed the use of data from peer companies to improve the prediction and error detection performance of statistical models (Hoitash et al. 2006). They defined the peers to be companies in the same industry (4-digit SIC code) with proximity in both size (Revenues) and growth (growth in revenues). They tested the effectiveness of using peer data to improve the performance of the statistical models used in auditing. Their results showed an increase in the effectiveness of the models using the peers' data. In their study they argue that the use of peers data is feasible given the increasing industry/auditor concentration (Hogan and Jeter 1999), and the consolidation among the large public accounting companies. They assume that in these situations auditors that specialize in certain industries can transfer information from one audit to the next (from one peer to the other) and consequently improve the effectiveness of their analytical procedures. In this paper, we argue that sharing peers or competitors' information across auditing firms is hardly feasible. As stated by the authors of the Hoitash paper themselves "… is only theoretically feasible". In this paper we propose using multiple economic indicators to improve the performance of the statistical models used by the auditors. We also test the effectiveness of using both the economic data and peers' data (if available to the auditors).

Research Questions

Following the Hoitash et al. (2006)'s paper, this paper investigates two research questions. Each research question is divided into three parts. The first part investigates the effect of the macroeconomic indicators individually. The second part investigates the effect of the macroeconomic indicators collectively. The third part investigates the effect of combining the macroeconomic indicators' and the peers' data.

The first research question investigates the effect of using macroeconomic indicators in the prediction models by comparing the Mean Absolute Percentage Error (MAPE) with and without the use of the macroeconomic indicators.

Research Question 1.1: Do Models that incorporate Macroeconomic indicators individually at least match the prediction performance of models incorporating Peers?

Research Question 1.2: Do Models that incorporate Macroeconomic indicators collectively at least match the prediction performance of models incorporating Peers?

Research Question 1.3: Does adding macroeconomic indicators to peers models enhance their predictive performance?

Sometimes the auditors face even greater difficulty using an independent variable to predict the dependent variable when the independent variable itself contains undetected errors. So, our second research question investigates the effect of

using macroeconomic indicators in the prediction models when the independent variables contain undetected errors.

Research Question 2.1: Can models incorporating Macroeconomic data moderate the impact of materially misstated account balances on the prediction accuracy of related accounts when used individually relative to Models incorporating Peers?

Research Question 2.2: Can models incorporating Macroeconomic data moderate the impact of materially misstated account balances on the prediction accuracy of related accounts when used collectively relative to Models incorporating Peers?

Research Question 2.3: Can models incorporating both Macroeconomic data and peers model moderate the impact of materially misstated account balances on the prediction accuracy of related accounts relative to Models incorporating Peers only?

Data

USA Data

We extracted 1,277,227 records of quarterly financial data for USA from COMPUTSTAT. The data covers the period from January 1986 to December 2014. The financial variables that we are using are cost of goods sold, revenues, accounts receivables, and accounts payable. The data include eight different major industries. Figure 38 shows the industry distribution in the USA financial data.

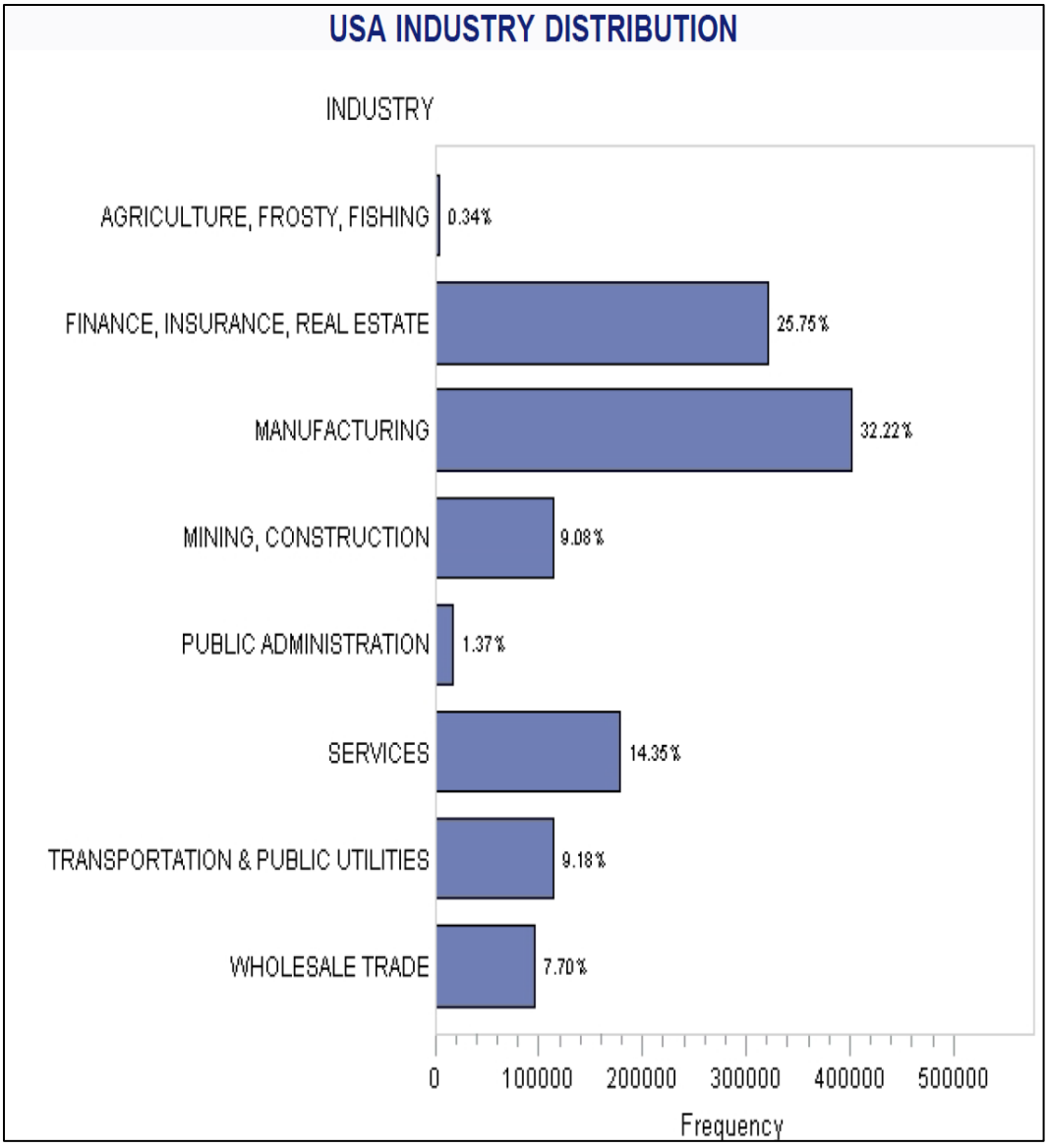

Figure 38: USA Industry Distribution

The data included 31,739 different companies distributed among the major industries. Table 16 shows the company counts and some descriptive statistics of the USA financial variables per industry.

Table 16: USA Financial Variables Per Industries

USA FINANCIAL VARIABLES PER INDUSTRY

INDUSTRY	COMPANY_COUNT	MEAN_APQ	MEAN_COGSQ	MEAN_RECTQ	MEAN_REVTQ	MEAN_SALESQ
AGRICULTURE, FROSTY, FISHING	110.00	47.44	97.50	101.21	139.95	139.95
FINANCE, INSURANCE, REAL ESTATE	8674.00	7402.92	274.21	8616.09	470.26	453.58
MANUFACTURING	9222.00	212.62	374.09	358.87	527.46	527.46
MINING, CONSTRUCTION	3437.00	73.08	117.79	95.39	171.95	171.95
PUBLIC ADMINISTRATION	454.00	340.71	438.14	1776.27	603.21	603.21
SERVICES	4987.00	57.42	87.18	116.36	144.12	144.12
TRANSPORTATION & PUBLIC UTILITIES	2490.00	304.86	482.03	369.85	714.55	714.55
WHOLESALE TRADE	2365.00	230.98	540.61	175.32	680.39	680.39

We deleted the records of all missing financial data. We ended up with 886,831 records of quarterly financial data out of the original 1,277,227 records.

We also got monthly Macro-Economic data for the USA. The data covers the period from January 1993 till December 2014 (264 months). The data includes thirty-three different indicators.

Sample Selection

In choosing our sample from the financial data, we decided the following;

- Matching the financial data period to the macroeconomic indicators period. Even though the financial data that we have covered the period from 1986 to 2014, the economic indicators data covered only the period from 1993 to 2014. Since we need both financial and macroeconomic data to test our research questions, we decided to only keep the financial data covers the period between 1993 and 2014 (22 years = 88 quarters).
- Allowing for companies with different fiscal year reporting.

Companies that use different fiscal year reporting can be missing data points at the beginning of 1993 and the ending of 2014 just because of the different reporting dates. For this reason, we decided to use the data between the third quarter of 1993 and the second quarter of 2014 (21 years = 84 quarters).

- Keeping only uninterrupted data with non-zero values.

  To keep a specific company in our sample, the company has to have uninterrupted data for at least 5 years starting from the third quarter of 1993. We chose 5 years to account for one year of lagged revenues and cost of goods sold, plus three years to run the models, plus one year for predictions. We also selected only companies with a non-zero value for both cost of goods sold and revenues. In our dataset, we have 1,935 companies with these characteristics.

- Keeping only 4-digit SIC codes if it has more than one company.

  Among the 1,935 companies mentioned above, there were 44 companies that each belonged to different SIC codes. Since our study includes peer selection within same 4-digit SIC code, we deleted those 44 companies from the sample. We ended up with 1,891 companies. Figure 39 shows the industry distribution of these companies.

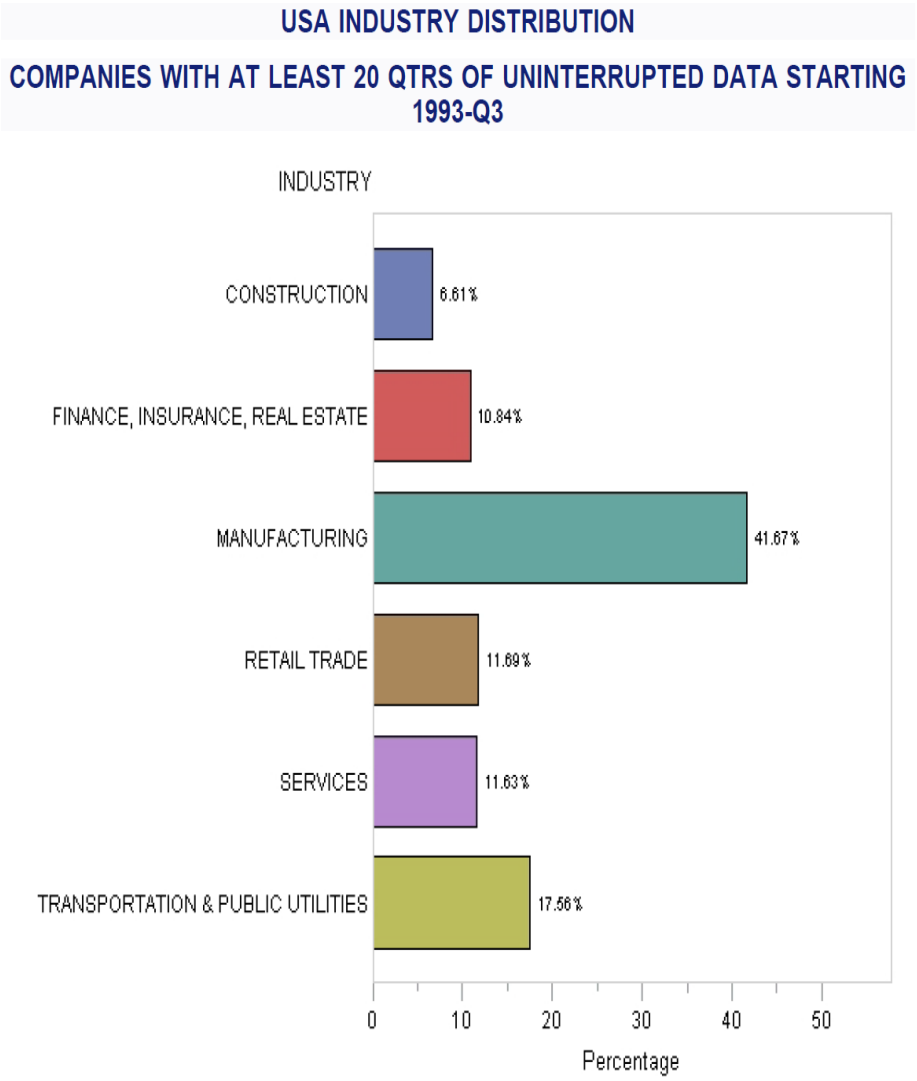

Figure 39: USA INDUSTRY DISTRIBUTION - 20 OR MORE QTRS

- Keeping only companies with matched peers.

  Since our research questions involve the use of peers, we only keep those companies that can be matched with at least one peer for each company year (please see details on how we select company peers in the “Peer Selection” section below). For the period of 84 quarters we ended up with 1,069 companies with uninterrupted data for at least 20 quarters (5 years)

starting from 1993-Q3 and matched peers for every company year. Table 17 shows the frequencies of the number of companies per each SIC code.

Table 17: PEERS - COMPANIES PER SIC CODE DISTRIBUTION

COMPANIES	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	12	14.29	12	14.29
2	10	11.90	22	26.19
3	7	8.33	29	34.52
4	3	3.57	32	38.10
5	7	8.33	39	46.43
6	4	4.76	43	51.19
7	4	4.76	47	55.95
8	6	7.14	53	63.10
9	2	2.38	55	65.48
10	1	1.19	56	66.67
12	3	3.57	59	70.24
13	1	1.19	60	71.43
15	2	2.38	62	73.81
16	1	1.19	63	75.00
17	2	2.38	65	77.38
18	2	2.38	67	79.76
20	2	2.38	69	82.14
21	2	2.38	71	84.52
26	1	1.19	72	85.71
27	1	1.19	73	86.90
29	1	1.19	74	88.10
34	1	1.19	75	89.29
36	1	1.19	76	90.48
37	1	1.19	77	91.67
41	1	1.19	78	92.86
45	2	2.38	80	95.24
58	1	1.19	81	96.43
59	1	1.19	82	97.62
70	1	1.19	83	98.81
87	1	1.19	84	100.00

- Working on only 4-digit SIC codes with at least 20 companies.

In our final sample we decided on keeping only those SIC codes with at least 20 companies. Leaving us with a final count of 17 different 4-digit SIC codes and 676 companies.

- Allowing one year for the use of “pervious year data points”.

  Since our models involve the use of the previous year data points (to predict Sales of this year, we need the sales of last year. Same goes for predicting cost of goods sold), we actually run our models for the period from third quarter of 1994 to the second quarter of 2014 (20 years).

Figure 40 shows the industry distribution of the final 676 companies we use in our sample.

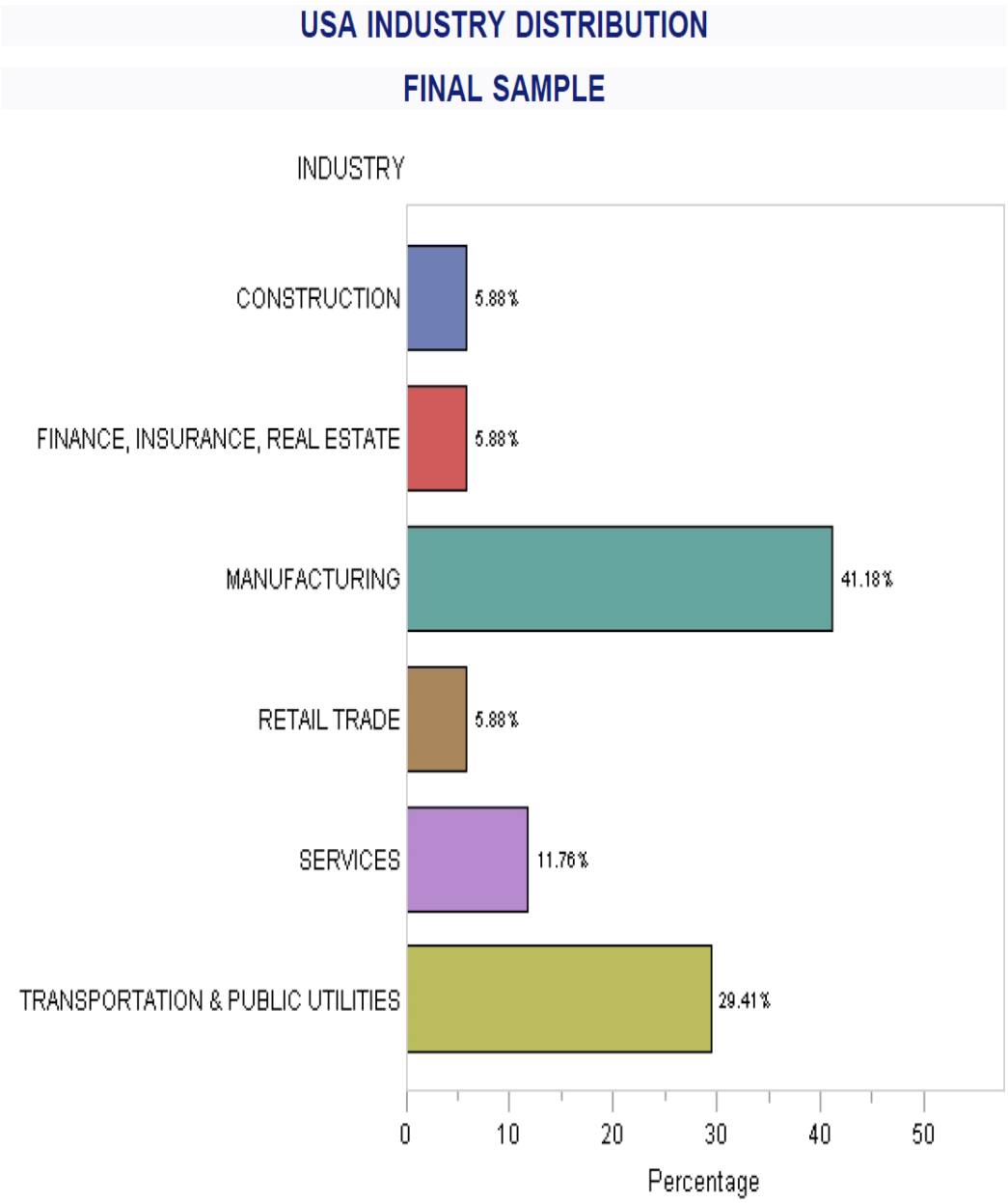

Figure 40: USA Final Sample: Industry Distribution

Only one macroeconomic indicator (Discount rate) had missing values for the same period (1994-Q3 to 2014-Q2). We deleted this indicator and used only the indicators with no missing values.

Table 18 shows some statistics of USA macroeconomic indicators used.

Table 18: USA MACROECONOMIC MONTHLY INDICATORS STATS

USA ECONOMIC INDICATORS STATISTICS

Variable	N Miss	N	Mean	Minimum	Maximum
GOODS & SERVICES BALANCE ON A BA	0	240	-35957.05	-67823.00	-5924.00
COMMERCIAL BANK ASSETS - LOANS &	0	240	4957.48	2243.82	7675.46
PRIME RATE CHARGED BY BANKS (MON	0	240	6.01	3.25	9.50
COMMERCIAL BANK ASSETS - COMMERC	0	240	1091.39	616.18	1689.00
CPI - ALL URBAN: ALL ITEMS SADJ	0	240	191.68	148.40	237.69
CPI - ALL ITEMS LESS FOOD & ENER	0	240	197.69	156.70	238.08
EMPLOYED - NONFARM INDUSTRIES TO	0	240	130421.10	114624.00	138764.00
TOTAL CIVILIAN EMPLOYMENT VOLA	0	240	137551.97	122706.00	146595.00
EXPORTS OF GOODS ON A BALANCE OF	0	240	81313.22	41475.00	136933.00
EXPORT PRICE INDEX - ALL COMMODI	0	240	111.18	97.30	135.30
TREASURY BILL RATE - 3 MONTH (EP	0	240	2.78	0.01	6.37
TREASURY YIELD ADJUSTED TO CONST	0	240	5.08	2.22	8.20
NEW PRIVATE HOUSING UNITS STARTE	0	240	1353.07	478.00	2273.00
IMPORTS OF GOODS ON A BALANCE OF	0	240	126210.35	56107.00	200695.00
IMPORTS F.A.S. CURA	0	240	124669.80	56096.00	198566.00
IMPORT PRICE INDEX - ALL COMMODI	0	240	112.14	91.00	147.50
WHOLESALE TRADE INVENTORIES - TO	0	240	350364.18	211902.00	533021.00
INDUSTRIAL PRODUCTION - TOTAL IN	0	240	89.85	68.57	104.11
FOREIGN NET LONG TERM FLOWS IN S	0	240	39253.97	-72882.00	139697.00
NEW ORDERS - MANUFACTURING, DURA	0	240	190645.33	144167.00	244841.00
NEW ORDERS - MANUFACTURING, EXCL	0	240	318955.92	227522.00	428809.00
PPI - FINISHED GOODS LESS FOODS	0	240	158.14	137.30	188.60
DISCOUNT RATE-WIND.BORR-NY FED,P	102	138	2.32	0.50	6.25
PPI - FINISHED GOODS SADJ	0	240	156.05	125.50	202.30
DOW JONES INDUSTRIALS SHARE PRIC	0	240	10120.54	3739.23	16826.60
UNEMPLOYMENT RATE SADJ	0	240	6.01	3.80	10.00
UNEMPLOYED (16 YRS & OVER) VOLA	0	240	8872.53	5481.00	15352.00
GOODS TRADE BALANCE ON A BALANCE	0	240	-44897.12	-77628.00	-12861.00
CONSUMER CREDIT OUTSTANDING CURA	0	240	2079.15	931.13	3214.24
EXPORTS F.A.S. CURA	0	240	80808.00	41955.00	135667.00
TRADE-WEIGHTED VALUE OF US DOLLA	0	240	86.41	69.02	112.20
INDUSTRIAL PRODUCTION - MANUFACT	0	240	87.39	63.87	101.58
NEW ORDERS - ALL MANUFACTURING I	0	240	375003.30	267694.00	505210.00

We then merged the monthly macroeconomic indicators with financial variables after converting the quarterly financial data into monthly data (please see below section for details on converting quarterly data to monthly data). We used the monthly date as a basis of the merge. Table 19 shows some descriptive statistics of the final sample used.

Table 19: USA FINAL SAMPLE STATS

FINAL SAMPLE DESCRIPTIVE STATISTICS

SIC	NAME	COMPANIES	AP	AR	COGS	REVENUE	ASSETS
1311	Crude Petroleum and Natural Gas	87	329.45687	287.72472	156.00543	225.49054	4025.7424
2834	Pharmaceutical Preparations	37	68.018439	218.09817	25.289756	111.72504	2033.6748
3661	Telephone and Telegraph Apparatus	20	17.48035	36.098211	8.6086858	15.333769	167.71691
3663	Radio and Television Broadcasting and Communications Equipment	26	182.04386	350.77728	108.77603	169.35904	2150.1118
3674	Semiconductors and Related Devices	45	117.92636	215.14959	55.366564	158.37303	2801.7665
3714	Motor Vehicle Parts and Accessories	21	87.62419	167.1549	63.668338	82.641347	1036.0479
3841	Surgical and Medical Instruments and Apparatus	20	70.094245	95.348956	20.136381	46.184635	681.70518
3845	Electromedical and Electrotherapeutic Apparatus	29	5.3684753	17.666388	3.5131242	7.4000446	101.98103
4213	Trucking, Except Local	21	54.233019	146.42274	96.590983	108.9195	702.82423
4813	Telephone Communications, Except Radiotelephone	41	492.01989	980.77614	204.58755	404.46413	9625.6755
4911	Electric Services	70	328.95016	421.74418	186.11014	265.24557	9064.1056
4924	Natural Gas Distribution	27	100.99492	124.6471	64.820085	78.363363	1636.2132
4931	Electric and Other Services Combined	36	243.30213	329.63547	184.30802	245.91101	7499.757
5812	Eating Places	45	46.417001	47.880021	74.030347	111.32795	1374.487
6798	Real Estate Investment Trusts	58	28.476762	80.854453	9.7887708	16.824131	1418.7739
7372	Prepackaged Software	59	82.669217	265.70946	26.427358	132.25969	2528.4925
7373	Computer Integrated Systems Design	34	38.960384	113.30139	28.99688	48.380839	542.67858
Total		676					

Peers Selection

For the selection of the peers we start with the 1,891 companies which fulfilled all our requirements discussed above. Then we followed the following steps to generate the peers for each company. Our peers selection process is similar to the one employed by previous research study (Hoitash et al. 2006);

- For the finance industry we used the Assets as the size proxy. For all other industries in the final sample we used the revenues. Either way, we use the selected account's value of last quarter of each company year as a proxy of the size of this company in that particular year.
- For the finance industry we used the assets growth (assets of last quarter of current year less assets of last quarter of previous year over assets of last quarter of previous year) as a proxy for the growth of the company. For all other industries we used revenue growth (revenue of last quarter of current year less revenue of last quarter of previous year over revenue of last quarter of previous year) of each company year as a proxy of the growth of this company in that particular year.
- We rank all companies within each 4-digit SIC code once based on their size proxy and another based on their growth proxy for each year.
- We apply the following rules when selecting peers for each company year;
    - Peers of a company year have to be within same 4-Digit code as this company.

  - Size proxy of the peer for a certain year has to be within (n/5) steps of the company's size proxy of the same year. 'n' being the company count within the 4-digit SIC code.
  - Growth proxy of the peer for a certain year has to be within (n/4) steps of the company's growth proxy of the same year. 'n' being the company count within the 4-digit SIC code.
  - A certain company cannot be selected to be its own peer.
- If a company has more than one peer for a certain year, we use the average values of the peers' standardized revenues (when predicting revenues) and COGS (when predicting COGS).
- If a company could not be matched with any peers for at least one company-year, we delete the company from our sample.

As mentioned before, we ended up with only 676 companies with uninterrupted data for at least 5 years (20 quarter) starting 1993-Q3 and peer selection for every single year.

Converting Quarterly Data to Monthly Data

There has been much of prior research studies that compared between the use of monthly data and quarterly data (Cogger 1981, Knechel 1988, Dzeng 1994, Kinney 1987, Chen & Leitch 1998, Chen & Leitch 1999). Their results suggested the superiority of the disaggregated monthly data over the quarterly data as it increased the effectiveness of their analytical procedures. Chen and Leitch (1998) explained in their paper that the increase of effectiveness was due to three reasons. The first reason is that the use of monthly data provides a larger sample

size. The second reason is that because the monthly data with its larger sample size allows the researcher to use a shorter time span. The shorter time span provides more stable results as it is less influenced by structural changes in the organization. The third reason according to Chen and Leitch (1998) is that the monthly data will be more correlated than the quarterly data. But since the published financial data is either annual data or quarterly data, there was a need to develop a methodology to simulate or generate monthly data from quarterly or annual data. In their 1998 and 1999 papers, Chen and Leitch developed a methodology to generate monthly data from quarterly data using curve fitting. The cubic splines interpolation method they developed interpolate the monthly data from quarterly data through curve fitting. Figure 41 is a graph from the Chen & Leitch 1998 paper showing an example of how they used curve fitting to interpolate monthly data from quarterly data.

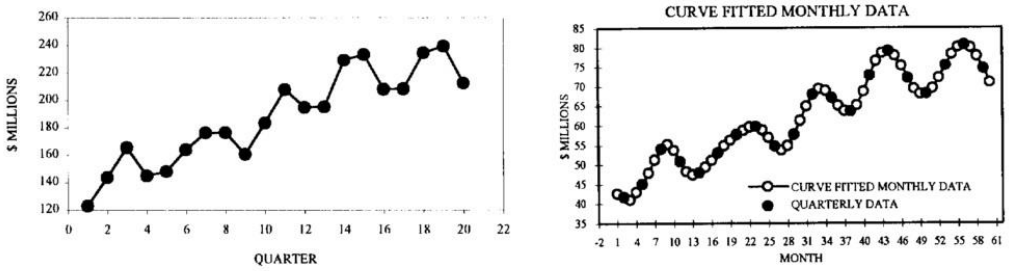

Figure 41: Chen & Leitch 1998's Curve Fitting

Different financial variables need to be treated differently when interpolating the monthly data (Chen and Leitch 1998, 1999; Hoitash et al. 2006). For example, when the variable is an income statement variable such as “Sales”, the quarterly value represents the total value of the whole period. While, if the variable is a balance sheet variable such as “Accounts Receivables”, the quarterly value

represents the value at the end of the period (the value at the end equal to the value at the beginning plus and minus changes throughout the period). Since the two types of variables represents different values, they have to be treated differently when generating the monthly data.

The Method

Model Specification

We start with examining the prediction value of the macroeconomic indicators relative to the peers model (Hoitash et al. 2006). In doing that, we compare the predictive performance of three models. The first model (Model S1) predicts the sales of one month based on the sales of the same month in the previous year (t-12). The second model (Model S2), which is the peer model, predicts the sales of a certain month based on the sales of the same month in the previous year (t-12), and the average peers' standardized sales value of the same predicted month. The peers' standardized sales values are used to avoid the impact of company size on the peer average. The mean and standard deviation of the sales are calculated during the estimation period (3 years) and then used to calculate the standardized value for each data point in the same period based on the formula $Z = \frac{Y - \mu}{\sigma}$ where Z is the standardized value of sales, Y is the sale balance, μ is the mean of the sales account balances over the three-years period, and σ is the standard deviation of the sales account balances over the same three-years period. The third model (Model S3), which is the macroeconomic indicator model, predicts the sales of a certain month based on the sales of the same month in the previous year (t-12), and the macroeconomic indicator value of the month before

the predicted month. The forth model (Model S4), which is the mixed model (both macroeconomic indicators and peers), predicts the sales of a certain month based on the sales of the same month in the previous year (t-12), the average peers' standardized sales value of the same predicted month, and the macroeconomic indicator value of the month before the predicted month.

We then examine the ability of the macroeconomic indicators to moderate the impact of materially misstated account balances on the prediction accuracy of related revenue account. For this, we compare the predictive performance of another three models. The first model (Model S5) predicts the sales of one month based on the sales of the same month in the previous year (t-12), and the balance of the accounts receivables for the same predicted month. The second model (Model S6), which is the peer model, predicts the sales of a certain month based on the sales of the same month in the previous year (t-12), the balance of the accounts receivables for the same predicted month, and the average peers' standardized sales value of the same predicted month. The third model (Model S7), which is the macroeconomic indicator model, predicts the sales of a certain month based on the sales of the same month in the previous year (t-12), the balance of the accounts receivables for the same predicted month, and the macroeconomic indicator value of the month before the predicted month. The forth model (Model S8), which is the mixed model (both macroeconomic indicators and peers), predicts the sales of a certain month based on the sales of the same month in the previous year (t-12), the balance of the accounts receivables for the same predicted month, the average peers' standardized sales value of the same

predicted month, and the macroeconomic indicator value of the month before the predicted month.

Table 20: Models Specification

$$SALES_t = \alpha + \beta_1 SALES_{t-12} + \varepsilon_t \quad (S1)$$

$$SALES_t = \alpha + \beta_1 SALES_{t-12} + \beta_2 PEER_t + \varepsilon_t \quad (S2)$$

$$SALES_t = \alpha + \beta_1 SALES_{t-12} + \beta_2 ECO_{t-1} + \varepsilon_t \quad (S3)$$

$$SALES_t = \alpha + \beta_1 SALES_{t-12} + \beta_2 PEER_t + \beta_3 ECO_{t-1} + \varepsilon_t \quad (S4)$$

$$SALES_t = \alpha + \beta_1 SALES_{t-12} + \beta_2 AR_t + \varepsilon_t \quad (S5)$$

$$SALES_t = \alpha + \beta_1 SALES_{t-12} + \beta_2 AR_t + \beta_2 PEER_t + \varepsilon_t \quad (S6)$$

$$SALES_t = \alpha + \beta_1 SALES_{t-12} + \beta_2 AR_t + \beta_3 ECO_{t-1} + \varepsilon_t \quad (S7)$$

$$SALES_t = \alpha + \beta_1 SALES_{t-12} + \beta_2 AR_t + \beta_3 PEER_t + \beta_4 ECO_{t-1} + \varepsilon_t \quad (S8)$$

$$COGS_t = \alpha + \beta_1 COGS_{t-12} + \varepsilon_t \quad (C1)$$

$$COGS_t = \alpha + \beta_1 COGS_{t-12} + \beta_2 PEER_t + \varepsilon_t \quad (C2)$$

$$COGS_t = \alpha + \beta_1 COGS_{t-12} + \beta_2 ECO_{t-1} + \varepsilon_t \quad (C3)$$

$$COGS_t = \alpha + \beta_1 COGS_{t-12} + \beta_2 PEER_t + \beta_3 ECO_{t-1} + \varepsilon_t \quad (C4)$$

$$COGS_t = \alpha + \beta_1 COGS_{t-12} + \beta_2 AP_t + \varepsilon_t \quad (C5)$$

$$COGS_t = \alpha + \beta_1 COGS_{t-12} + \beta_2 AP_t + \beta_2 PEER_t + \varepsilon_t \quad (C6)$$

$$COGS_t = \alpha + \beta_1 COGS_{t-12} + \beta_2 AP_t + \beta_3 ECO_{t-1} + \varepsilon_t \quad (C7)$$

$$COGS_t = \alpha + \beta_1 COGS_{t-12} + \beta_2 AP_t + \beta_3 PEER_t + \beta_4 ECO_{t-1} + \varepsilon_t \quad (C8)$$

We repeat the same process for the cost of goods sold (COGS). We compare the predictive performance of three models. The first model (Model C1) predicts the COGS of one month based on the COGS of the same month in the previous year (t-12). The second model (Model C2), which is the peer model, predicts the COGS

of a certain month based on the COGS of the same month in the previous year (t-12), and the average peers' standardized COGS value of the same predicted month. The third model (Model C3), which is the macroeconomic indicator model, predicts the COGS of a certain month based on the COGS of the same month in the previous year (t-12), and the macroeconomic indicator value of the month before the predicted month. The forth model (Model C4), which is the mixed model (both macroeconomic indicators and peers), predicts the COGS of a certain month based on the COGS of the same month in the previous year (t-12), the average peers' standardized COGS value of the same predicted month, and the macroeconomic indicator value of the month before the predicted month.

To examine the ability of the macroeconomic indicators to moderate the impact of materially misstated account balances on the prediction accuracy of related COGS account, we compare the predictive performance of another three models. The first model (Model C5) predicts the COGS of one month based on the COGS of the same month in the previous year (t-12), and the balance of the accounts payables for the same predicted month. The second model (Model C6), which is the peer model, predicts the COGS of a certain month based on the COGS of the same month in the previous year (t-12), the balance of the accounts payables for the same predicted month, and the average peers' standardized COGS value of the same predicted month. The third model (Model C7), which is the macroeconomic indicator model, predicts the COGS of a certain month based on the COGS of the same month in the previous year (t-12), the balance of the accounts payables for the same predicted month, and the macroeconomic indicator value of the same

predicted month. The forth model (Model C8), which is the mixed model (both macroeconomic indicators and peers), predicts the COGS of a certain month based on the COGS of the same month in the previous year (t-12), the balance of the accounts payables for the same predicted month, the average peers' standardized COGS value of the same predicted month, and the macroeconomic indicator value of the month before the predicted month.

We then use the Mean Absolute Percentage Error (MAPE) as follows:

$$MAPE = \frac{1}{i * 12} \sum_{i=1}^{i} \sum_{j=1}^{12} \frac{|P_{ij} - A_{ij}|}{A_{ij}}$$

Where "P" represents the predicted value, "A" represents the actual value, "i" is the number of companies in each industry and "j" is the number of months predicted.

Collective Macroeconomic Indicators

We are testing the effect of the macroeconomic indicators both individually and collectively. When testing their effect individually, we pick the macroeconomic indicator with the highest correlation with the dependent variable in each industry. When testing the effect of the macroeconomic indicators collectively, we pick the

three macroeconomic indicators with the highest correlation with the dependent variable in each industry. We then use them together in the model as shown below.

$$SALES_t = \alpha + \beta_1 SALES_{t-12} + \beta_2 ECO1_{t-1} + \beta_3 ECO2_{t-1} + \beta_4 ECO3_{t-1} + \varepsilon_t \quad \text{(S3c)}$$

$$SALES_t = \alpha + \beta_1 SALES_{t-12} + \beta_2 PEER_t + \beta_3 ECO1_{t-1} + \beta_4 ECO2_{t-1} + \beta_5 ECO3_{t-1} + \varepsilon_t \quad \text{(S4c)}$$

$$SALES_t = \alpha + \beta_1 SALES_{t-12} + \beta_2 AR_t + \beta_3 ECO1_{t-1} + \beta_4 ECO2_{t-1} + \beta_5 ECO3_{t-1} + \varepsilon_t \quad \text{(S7c)}$$

$$SALES_t = \alpha + \beta_1 SALES_{t-12} + \beta_2 AR_t + \beta_3 PEER_t + \beta_4 ECO1_{t-1} + \beta_5 ECO2_{t-1} + \beta_6 ECO3_{t-1} + \varepsilon_t \quad \text{(S8c)}$$

$$COGS_t = \alpha + \beta_1 COGS_{t-12} + \beta_2 ECO1_{t-1} + \beta_3 ECO2_{t-1} + \beta_4 ECO3_{t-1} + \varepsilon_t \quad \text{(C3c)}$$

$$COGS_t = \alpha + \beta_1 COGS_{t-12} + \beta_2 PEER_t + \beta_3 ECO1_{t-1} + \beta_4 ECO2_{t-1} + \beta_5 ECO3_{t-1} + \varepsilon_t \quad \text{(C4c)}$$

$$COGS_t = \alpha + \beta_1 COGS_{t-12} + \beta_2 AP_t + \beta_3 ECO1_{t-1} + \beta_4 ECO2_{t-1} + \beta_5 ECO3_{t-1} + \varepsilon_t \quad \text{(C7c)}$$

$$COGS_t = \alpha + \beta_1 COGS_{t-12} + \beta_2 AP_t + \beta_3 PEER_t + \beta_4 ECO1_{t-1} + \beta_5 ECO2_{t-1} + \beta_6 ECO3_{t-1} + \varepsilon_t \quad \text{(C8c)}$$

Rolling Window Regression

In this research study, we are proposing the use of macro-economic indicators to improve the prediction and error detection performance of the statistical models. We are basing our analysis on the notion that using a regression model on a time series data can capture systematic changes in a company's account balances (Chen and Leitch 1998; Hoitash et al. 2006). However, over an extended period of time the relationship between the dependent variable and the independent variables might change due to different reasons that could be related to the company itself or to the economy as a whole (Hoitash et al. 2006). For that reason,

we are using the rolling window regression. In a rolling window regression, we create partially overlapping subsamples from the time series dataset. Each subsample has a different starting and ending dates while keeping the length of the time period fixed. We then run the regression model on each subsample. In this research, we predict one year's dependent variable based on the previous three-year period (36 months).

Test of Research Questions

Each regression model mentioned in Table 20 above is estimated over 36 consecutive months. Each model is then tested over the subsequent 12 months. Every model is estimated separately for each company based on its unique set of peer companies and the macroeconomic indicators. When a company has more than one assigned peer, we use the average value of the standardized revenues and cost of goods sold for all assigned peers. The selection of the 36 months as the training period and the 12 months as the testing period is similar to other studies' research design (Hoitash et al. 2006).

Prediction performance is evaluated based on examining the mean absolute percentage error (MAPE) for each account-model (Sales models S1 through S8, and COGS models C1 through C8). The MAPE is calculated for the out-of-sample prediction for each account-company-month. The MAPEs for the 12-month period are aggregated over company-year resulting in an aggregated measure of MAPE for each company-account-model. To evaluate the prediction performance of each model, results are aggregated over each account-industry, resulting in one MAPE for every account-model industry (industry is defined as per 4-digit SIC CODE).

This process is similar to the one employed by a previous study (Hoitash et al. 2006).

To test the first research question we use account-models 1 through 4. We estimate the coefficients of the independent variables for each three firm-years and then we use the coefficients to predict the dependent account balance for the forth firm-year. We calculate a firm-monthly MAPE and then average it over the entire industry. We then compare the MAPEs generated across the different models.

To test the second research question, we use account models 5 through 8. We follow the same process as per the first research question except that in the prediction year we use misstated independent account balance (overstated AR when predicting revenues, understated AP when predicting COGS) in the prediction of the dependent account balance. The objective is to test the models under the impact of materially misstated independent variables. For the materiality definition we use a 2% increase or decrease in the account balance (Hoitash et al. 2006; Knechel 1988).

The Results

First Research Question

In our first research question, we anticipate that the models incorporating the economic indicators will have at least the same prediction performance as those incorporating only the peers' standardized values. For that purpose, we evaluate the predictions from models 1 through 4 for each account (sales and COGS).

Models S3c and S4c are similar to models S3 and S4 except that they use multiple macroeconomic indicators instead of just one indicator. The Prediction performance is evaluated by using a nonparametric test (Wilcoxon Rank-Sum test) to evaluate the differences between the mean absolute percentage errors of the benchmark, peers, macroeconomic indicators and the mixed models. Table 21 contain the results for the test of the first research question for revenues account. As seen in the tables, results suggest that macroeconomic indicators improve the prediction. As seen in the table, model S3 (macroeconomic indicators) resulted in better MAPE in 9 of the 17 industries used (shown in blue). Model S4 (mixed model), resulted in better MAPE than the peers' model across all 17^{th} industries. The best model across each industry is shown in green. The models with multiple economic indicators generated better results in general than the models using only one economic indicators. The mixed model S4c had the best performance of all the models used. Table 22 shows the statistical significance of these results. It shows that all the models used were significantly better than the benchmark model S1. Models S3c, S4, and S4c were significantly better than the Peers model S2 in at least 13 out of the 17 industries.

Table 21: Results - 1st RQ - Revenues

SIC	OBS	S1	S2	S3	S3c	S4	S4c
1311	9158	0.351	0.288	0.266	0.215	0.248	0.234
2834	3221	0.47	0.389	0.301	0.283	0.276	0.243
3661	1385	0.193	0.146	0.151	0.144	0.135	0.135
3663	2635	0.267	0.221	0.208	0.201	0.192	0.188
3674	6439	0.238	0.188	0.195	0.178	0.171	0.165
3714	1609	0.141	0.115	0.121	0.105	0.105	0.103
3841	1947	0.117	0.111	0.098	0.093	0.097	0.093
3845	2856	0.155	0.248	0.138	0.144	0.179	0.181
4213	2204	0.093	0.07	0.072	0.07	0.066	0.063
4813	2408	0.105	0.085	0.087	0.074	0.076	0.07
4911	10172	0.121	0.103	0.113	0.112	0.103	0.104
4924	2479	0.239	0.244	0.235	0.244	0.232	0.252
4931	4963	0.128	0.123	0.125	0.126	0.121	0.119
5812	4486	0.077	0.07	0.071	0.066	0.066	0.063
6798	7212	0.119	0.145	0.087	0.084	0.089	0.114
7372	5145	0.322	0.282	0.273	0.239	0.249	0.217
7373	3493	0.233	0.242	0.188	0.149	0.2	0.19

Best across industry
3 better than 2

Table 22: RQ1 - Revenues -Statistical Significance

		S2 VS. S1		S3 VS. S1		S3 VS. S2		S3c Vs. S1		S3c VS. S2		S4 VS. S1		S4 VS. S2		S4c VS. S1		S4c VS. S2	
SIC	OBS	S2	S1	S3	S1	S3	S2	S3c	S1	S3c	S2	S4	S1	S4	S2	S4c	S1	S4c	S2
1311	9158	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
2834	3221	0.000	-----	0.000	-----	-----	0.024	0.000	-----	0.000	-----	0.000	-----	0.033	-----	0.000	-----	0.000	-----
3661	1385	0.000	-----	0.000	-----	-----	0.166	0.000	-----	0.073	-----	0.000	-----	0.098	-----	0.000	-----	0.003	-----
3663	2635	0.000	-----	0.000	-----	0.113	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
3674	6439	0.000	-----	0.000	-----	-----	0.000	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
3714	1609	0.000	-----	0.000	-----	-----	0.159	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
3841	1947	0.001	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
3845	2856	0.000	-----	0.000	-----	-----	0.320	0.000	-----	0.014	-----	0.000	-----	0.000	-----	0.000	-----	0.003	-----
4213	2204	0.000	-----	0.000	-----	-----	0.001	0.000	-----	0.407	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
4813	2408	0.000	-----	0.000	-----	-----	0.068	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
4911	10172	0.000	-----	0.000	-----	-----	0.000	0.000	-----	-----	0.000	0.000	-----	0.409	-----	0.000	-----	-----	0.279
4924	2479	0.068	-----	-----	0.000	-----	0.018	-----	0.000	-----	0.000	-----	0.286	-----	0.114	-----	0.000	-----	0.000
4931	4963	0.000	-----	0.015	-----	-----	0.007	0.001	-----	-----	0.073	0.000	-----	0.044	-----	0.000	-----	0.111	-----
5812	4486	0.000	-----	0.000	-----	-----	0.021	0.000	-----	0.000	-----	0.000	-----	0.125	-----	0.000	-----	0.000	-----
6798	7212	0.000	-----	0.000	-----	-----	0.357	0.000	-----	0.005	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
7372	5145	0.000	-----	0.000	-----	-----	0.022	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
7373	3493	0.000	-----	0.000	-----	-----	0.189	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----

Table 23 contain the results for the test of the first research question for COGS account. As seen in the tables, results suggest that macroeconomic indicators improve the prediction. As seen in the table, model C3 (macroeconomic indicators) resulted in better MAPE in 12 of the 17 industries used (shown in blue/green). Model C4 (mixed model), resulted in better MAPE than the peers' model across all 17th industries. The best model across each industry is shown in green. The models with multiple economic indicators generated better results in general than the models using only one economic indicators. The mixed model C4c had the best performance of all the models used. Table 24 shows the statistical significance of these results. It shows that all the models used were significantly better than the benchmark model C1. Models C3c, C4, and C4c were significantly better than the Peers model C2 in at least 16 out of the 17 industries.

Table 23: Results - 1st RQ - COGS

SIC	OBS	C1	C2	C3	C3c	C4	C4c
1311	**9158**	0.538	0.502	0.518	0.499	0.494	0.544
2834	**3221**	0.475	0.436	0.331	0.295	0.338	0.287
3661	**1385**	0.297	0.231	0.159	0.146	0.128	0.120
3663	**2635**	0.287	0.241	0.226	0.206	0.211	0.207
3674	**6439**	0.278	0.230	0.221	0.189	0.194	0.173
3714	**1609**	0.138	0.113	0.120	0.112	0.105	0.102
3841	**1947**	0.178	0.169	0.137	0.127	0.131	0.130
3845	**2856**	0.822	0.368	0.478	0.205	0.299	0.265
4213	**2204**	0.097	0.078	0.075	0.070	0.071	0.067
4813	**2408**	0.154	0.138	0.117	0.105	0.118	0.111
4911	**10172**	0.162	0.137	0.141	0.129	0.127	0.121
4924	**2479**	0.340	0.374	0.299	0.299	0.324	0.327
4931	**4963**	0.163	0.159	0.154	0.155	0.154	0.152
5812	**4486**	0.112	0.099	0.100	0.091	0.091	0.088
6798	**7212**	0.164	0.168	0.144	0.144	0.148	0.148
7372	**5145**	1.021	0.614	0.374	0.351	0.429	0.378
7373	**3493**	0.078	0.174	0.058	0.063	0.142	0.131

Best across industry
3 better than 2

Table 24: RQ1 - COGS -Statistical Significance

		C2 VS. C1		C3 VS. C1		C3 VS. C2		C3c Vs. C1		C3c VS. C2		C4 VS. C1		C4 VS. C2		C4c VS. C1		C4c VS. C2	
SIC	OBS	C2	C1	C3	C1	C3	C2	C3c	C1	C3c	C2	C4	C1	C4	C2	C4c	C1	C4c	C2
1311	9158	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
2834	3221	0.000	-----	0.040	-----	-----	0.042	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
3661	1385	0.000	-----	0.000	-----	0.099	-----	0.002	-----	0.000	-----	0.000	-----	0.000	-----	0.002	-----	0.000	-----
3663	2635	0.000	-----	0.000	-----	0.034	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
3674	6439	0.000	-----	0.000	-----	-----	0.000	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
3714	1609	0.000	-----	0.000	-----	0.116	-----	0.014	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
3841	1947	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
3845	2856	0.000	-----	0.000	-----	-----	0.070	0.007	-----	0.000	------	0.000	-----	0.067	-----	0.000	-----	0.000	-----
4213	2204	0.000	-----	0.000	-----	0.424	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
4813	2408	0.000	-----	0.000	-----	0.018	------	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
4911	10172	0.000	-----	0.003	-----	-----	0.000	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
4924	2479	0.006	-----	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.301	-----	0.000	-----	0.300	-----	0.139
4931	4963	0.000	-----	0.000	-----	-----	0.466	0.082	-----	0.000	-----	0.000	-----	0.042	-----	0.000	-----	0.001	-----
5812	4486	0.000	-----	0.000	-----	-----	0.016	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
6798	7212	0.000	-----	0.000	-----	0.000	------	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
7372	5145	0.000	-----	0.000	-----	0.016	------	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
7373	3493	0.000	-----	0.000	-----	-----	0.249	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----

Second Research Question

In our second research question, we anticipate that the models incorporating the economic indicators will have at least the same prediction performance as those incorporating only the peers' standardized values. For that purpose, we evaluate the predictions from models 5 through 8 for each account (sales and COGS). Prediction performance is evaluated by using a nonparametric test (Wilcoxon Rank-Sum test) to evaluate the differences between the mean absolute percentage errors of the benchmark, peers, macroeconomic indicators and the mixed models. Table 25 contain the results for the test of the second research question for revenues account. As seen in the tables, results suggest that macroeconomic indicators improve the prediction. As seen in the table, model S7 (macroeconomic indicators) resulted in better MAPE in 13 of the 17 industries used (shown in blue/green). Model S8 (mixed model), resulted in better MAPE than the

peers' model in 15 out of the 17 industries. The best model across each industry is shown in green. The models with multiple economic indicators generated better results in general than the models using only one economic indicators. The mixed model S4c had the best performance of all the models used. Table 26 shows the statistical significance of these results. It shows that all the models used were significantly better than the benchmark model S1 in at least 16 out of the 17 industries. Models S7c, S8, and S8c were significantly better than the Peers model S6 in at least 11 out of the 17 industries.

Table 25: Results – 2ndt RQ - Revenues

SIC	OBS	S5	S6	S7	S7c	S8	S8c
1311	**9158**	0.253	0.23	0.22	0.193	0.215	0.202
2834	**3221**	0.469	0.491	0.329	0.294	0.3	0.265
3661	**1385**	0.106	0.1	0.088	0.086	0.088	0.089
3663	**2635**	0.199	0.183	0.172	0.166	0.166	0.165
3674	**6439**	0.152	0.146	0.142	0.138	0.138	0.136
3714	**1609**	0.101	0.097	0.094	0.092	0.085	0.083
3841	**1947**	0.098	0.101	0.085	0.082	0.086	0.084
3845	**2856**	0.119	0.141	0.1	0.104	0.161	0.17
4213	**2204**	0.064	0.058	0.058	0.056	0.054	0.054
4813	**2408**	0.077	0.07	0.069	0.063	0.063	0.063
4911	**10172**	0.11	0.106	0.104	0.101	0.098	0.099
4924	**2479**	0.245	0.235	0.226	0.226	0.219	0.223
4931	**4963**	0.12	0.118	0.116	0.114	0.114	0.114
5812	**4486**	0.071	0.068	0.069	0.065	0.065	0.063
6798	**7212**	0.141	0.137	0.117	0.112	0.116	0.123
7372	**5145**	0.234	0.205	0.209	0.2	0.196	0.183
7373	**3493**	0.141	0.112	0.129	0.127	0.174	0.17

Best across industry
3 better than 2

Table 26: RQ2 - Revenues -Statistical Significance

		S6 VS. S5		S7 VS. S5		S7 VS. S6		S7c Vs. S5		S7c VS. S6		S8 VS. S5		S8 VS. S6		S8c VS. S5		S8c VS. S6	
SIC	OBS	S6	S5	S7	S5	S7	S6	S7c	S5	S7c	S6	S8	S5	S8	S6	S8c	S5	S8c	S6
1311	9158	0.000	-----	0.000	-----	0.003	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
2834	3221	0.000	-----	0.000	-----	0.109	-----	0.000	-----	0.003	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
3661	1385	0.016	-----	0.008	-----	0.015	-----	0.002	-----	0.009	-----	0.001	-----	0.027	-----	0.002	-----	0.038	-----
3663	2635	0.000	-----	0.000	-----	-----	0.348	0.000	-----	0.028	-----	0.000	-----	0.129	-----	0.000	-----	0.013	-----
3674	6439	0.000	-----	0.000	-----	-----	0.317	0.000	-----	0.003	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
3714	1609	0.011	-----	0.036	-----	0.089	-----	0.014	-----	0.275	-----	0.000	-----	0.000	-----	0.000	-----	0.001	-----
3841	1947	0.042	-----	0.001	-----	0.061	-----	0.000	-----	0.006	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
3845	2856	0.005	-----	0.000	-----	-----	0.361	0.007	-----	-----	0.098	0.006	-----	0.492	-----	0.375	-----	-----	0.027
4213	2204	0.000	-----	0.000	-----	0.203	-----	0.000	-----	0.031	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
4813	2408	0.000	-----	0.000	-----	-----	0.395	0.000	-----	0.009	-----	0.000	-----	0.000	-----	0.000	-----	0.002	-----
4911	10172	0.000	-----	0.058	-----	-----	0.001	0.000	-----	0.496	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
4924	2479	-----	0.130	-----	0.003	-----	0.300	-----	0.000	-----	0.035	-----	0.045	-----	0.452	-----	0.024	-----	0.167
4931	4963	0.150	-----	0.147	-----	-----	0.356	0.082	-----	0.345	-----	-----	0.488	0.294	-----	0.500	-----	0.174	-----
5812	4486	0.000	-----	0.000	-----	-----	0.290	0.000	-----	0.002	-----	0.000	-----	0.000	-----	0.000	-----	0.001	-----
6798	7212	0.000	-----	0.000	-----	-----	0.030	0.000	-----	0.002	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
7372	5145	0.000	-----	0.000	-----	-----	0.348	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
7373	3493	0.000	-----	0.000	-----	-----	0.238	0.000	-----	0.330	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----

Table 27 contains the results for the test of the second research question for COGS account. As seen in the tables, results suggest that macroeconomic indicators improve the prediction. As seen in the tables, model C7 (macroeconomic indicators) resulted in better MAPE in 16 of the 17 industries used. Model C8 (mixed model), resulted in better MAPE than the peers' model in 16 out of the 17 industries. The best model across each industry is shown in green. The models with multiple economic indicators generated better results in general than the models using only one economic indicators. The mixed model C7c had the best performance of all the models used. Table 28 shows the statistical significance of these results. It shows that all the models used were significantly better than the benchmark model C5 except in one industry. Models C7c, C8, and C8c were significantly better than the Peers model C6 in at least 10 out of the 17 industries.

Table 27: Results - 2nd RQ - COGS

SIC	OBS	C5	C6	C7	C7c	C8	C8c
1311	**9158**	0.533	0.530	0.566	0.564	0.562	0.620
2834	**3221**	0.432	0.450	0.304	0.293	0.324	0.286
3661	**1385**	0.292	0.265	0.142	0.193	0.130	0.124
3663	**2635**	0.229	0.218	0.199	0.191	0.197	0.192
3674	**6439**	0.209	0.196	0.184	0.183	0.174	0.164
3714	**1609**	0.110	0.112	0.104	0.100	0.098	0.096
3841	**1947**	0.174	0.155	0.130	0.122	0.128	0.129
3845	**2856**	0.785	0.524	0.469	0.224	0.305	0.253
4213	**2204**	0.079	0.071	0.068	0.065	0.065	0.064
4813	**2408**	0.127	0.112	0.101	0.093	0.109	0.104
4911	**10172**	0.135	0.126	0.124	0.119	0.120	0.122
4924	**2479**	0.290	0.323	0.271	0.273	0.297	0.291
4931	**4963**	0.144	0.147	0.136	0.138	0.140	0.138
5812	**4486**	0.102	0.098	0.093	0.089	0.086	0.085
6798	**7212**	0.171	0.163	0.154	0.151	0.151	0.155
7372	**5145**	0.890	0.710	0.308	0.326	0.386	0.380
7373	**3493**	0.140	0.162	0.113	0.097	0.119	0.118

Best across industry
3 better than 2

Table 28: RQ2 -COGS -Statistical Significance

		C6 VS. C5		**C7 VS. C5**		**C7 VS. C6**		**C7c Vs. C5**		**C7c VS. C6**		**C8 VS. C5**		**C8 VS. C6**		**C8c VS. C5**		**C8c VS. C6**	
SIC	OBS	C6	C5	C7	C5	C7	C6	C7c	C5	C7c	C6	C8	C5	C8	C6	C8c	C5	C8c	C6
1311	9158	0.000	-----	0.000	-----	0.002	-----	0.000	-----	0.155	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
2834	3221	0.000	-----	0.073	-----	-----	0.022	0.000	-----	0.017	-----	0.000	-----	0.085	-----	0.000	-----	0.033	-----
3661	1385	0.043	-----	0.023	-----	0.394	-----	0.181	-----	0.394	-----	0.000	-----	0.001	-----	0.000	-----	0.023	-----
3663	2635	0.000	-----	0.000	-----	0.002	-----	0.000	-----	0.000	-----	0.000	-----	0.003	-----	0.000	-----	0.000	-----
3674	6439	0.000	-----	0.000	-----	-----	0.373	0.000	-----	0.014	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
3714	1609	0.185	-----	0.036	-----	0.023	-----	0.016	-----	0.002	-----	0.000	-----	0.045	-----	0.007	-----	0.006	-----
3841	1947	0.006	-----	0.000	-----	0.026	-----	0.000	-----	0.062	-----	0.000	-----	0.012	-----	0.000	-----	0.003	-----
3845	2856	0.000	-----	0.000	-----	0.105	-----	0.000	-----	0.161	------	0.000	-----	0.112	-----	0.000	-----	0.256	-----
4213	2204	0.000	-----	0.000	-----	0.375	-----	0.000	-----	0.003	-----	0.000	-----	0.010	-----	0.000	-----	0.002	-----
4813	2408	0.006	-----	0.000	-----	0.024	------	0.000	-----	0.180	-----	0.000	-----	0.002	-----	0.000	-----	0.000	-----
4911	10172	0.000	-----	0.500	-----	-----	0.000	0.000	-----	-----	0.395	0.000	-----	0.000	-----	0.000	-----	0.000	-----
4924	2479	-----	0.015	-----	0.139	0.039	-----	-----	0.092	0.148	-----	0.374	-----	0.178	-----	-----	0.374	0.260	-----
4931	4963	0.174	-----	0.140	-----	0.346	-----	0.134	-----	0.421	-----	0.087	-----	-----	0.101	0.002	-----	0.213	-----
5812	4486	0.000	-----	0.000	-----	0.321	-----	0.000	-----	0.040	-----	0.000	-----	0.002	-----	0.000	-----	0.000	-----
6798	7212	0.000	-----	0.000	-----	0.439	------	0.000	-----	0.029	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
7372	5145	0.000	-----	0.000	-----	0.390	------	0.000	-----	0.005	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----
7373	3493	0.000	-----	0.000	-----	-----	0.306	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----	0.000	-----

Conclusion and Limitations

Our research examines the potential benefits of using macroeconomic indicators in performing audit analytical procedures as opposed to using peer data. The research also examines the potential benefits of using both types of data together. Our results strongly indicate that using macroeconomic indicators' data is beneficial for improving the overall performance of the analytical procedures when the peers' data is not included. In addition, the results show that adding the macroeconomic indicators data to the models already using the peers' data have significantly improved their performance. The results are consistent across most of the industries included in the research. But, as expected, the effect of including the macroeconomic indicators differs from one industry to the other as different industries react differently to different macroeconomic indicators.

We need to mention some limitations of this research. First of all, the peers' selection process is a subjective process. Our results may have been different given a different peers selection process. Secondly, our research didn't take into consideration the interrelationship between the different macroeconomic indicators when used collectively in the models. Lastly, the research didn't take into consideration the interrelationship between peers' data and the macroeconomic indicators' as it may be expected the peers' data already reflects the changes in the macroeconomic indicators.

Chapter Five: Conclusion

In this thesis we worked on three different essays. In our first essay, we proposed different methodologies to detect anomalies in life / disability insurance data. we worked on both benefit/claim payment business cycle and Premium collection business cycle. We used life/disability insurance data provided to us by a leading international insurance company. To detect anomalies in the claims payment cycle, we used a weighted multi-dimensional approach in which we divide the attributes we have into different groups (dimensions). We used each dimension to logically find insurance claim anomalies. then we prioritized the anomalies based on the weighted average of the dimensions that triggered this claim as an anomaly. As an additional way of prioritizing the outliers, we used the belief function to give a “Risk Score” to the different branches within the insurance company. The anomalies detected will then be weighted by the risk score of the company branch that generated it. To detect anomalies in premium collections cycle we used a robust regression model.

While working on our first essay, we faced a lot of difficulties in collecting the data from the insurance company. These difficulties caused to turn the first essay from an empirical study to a design science with examples to show case our methodology. These difficulties also made us realize that the auditor’s problem is not the data availability anymore, it is the data’s accessibility. Auditors face many challenges in accessing the data they need to fulfill their duties and form their opinion even when their clients are fully digitalized and technologically capable of

providing the needed data. This was our motivation for the next two chapters; trying to find solutions to give the auditor better accessibility to the needed data or to give him other sources of data to be able to perform his duties.

In the second essay, we discussed one of the solutions for the auditor's data accessibility problem which is the Audit Data Standards (ADS) issued by the AICPA Assurance Services Executive Committee. Five standards have been issued till now; Base, General Ledger, Accounts Receivables, Order to Cash and Procure to pay. So far, there hasn't been a specific model to follow in designing the ADS. Also, there hasn't been any industry specific ADS yet. In the second essay, we propose a model for generating industry specific ADS based on AICPA audit guide, audit analytics, and audit plans. We give an example by generating ADS for life/disability insurance for the audit objectives related to the authenticity and valuation of paid claims. We concluded that sometimes we are either going to need to update the already published ADS or to sacrifice some of the data normalization rules. For example, the base ADS includes a table for system users where the User_ID, User_Name, Title and responsibilities are defined for every employee who has access to the system. When it comes to the insurance company audit, the auditor needs to gain understanding of the education level and expertise of the employees regardless of whether they have access to the system or not. In a perfect database world, a table would be generated for all the employees and their details along with a User_ID field for those employees who can access the system. But since the Base standard is already published we had to sacrifice the

normalization rules and ended up with two tables describing employees. After generating the ADS, we used them in creating an interactive audit dashboard that helps the auditor in performing his audit. For future research related to this essay, we would like to automate the model.

The third essay, discuss another possible solution for the auditor; the use of external public data to assist in the audit process. In this essay we proposed using macro-economic indicators to improve the prediction and error detection performance of the statistical models. We tested the effectiveness of the macroeconomic indicators in detecting coordinated errors and in mitigating the effects of misstated accounts. We also tested the effectiveness of using macroeconomic indicators with peer data. our first research question investigated the effect of using macroeconomic indicators in the prediction models by comparing the Mean Absolute Percentage Error (MAPE) with and without the use of the macroeconomic indicators. our second research question investigated the effect of using macroeconomic indicators in the prediction models when the independent variables contain undetected errors. We tested each research question in three different situations; the macroeconomic variable is used individually, collectively, or with the peer data. our results came in favor of the macroeconomic indicator's use, specifically when used along with the peer data. The limitations we faced in the third essay were that firstly, the peers' selection process is a subjective process. Our results may have been different given a different peers selection process. secondly, our research didn't take into

consideration the interrelationship between the different macroeconomic indicators when used collectively in the models. Lastly, the research didn't take into consideration the interrelationship between peers' data and the macroeconomic indicators' as it may be expected the peers' data already reflects the changes in the macroeconomic indicators. For future research we would want to test of the use of more publicly available data like geographical data or weather related data for example.

REFERENCES

Alles, Michael, Gerard Brennan, Alexander Kogan, and Miklos a. Vasarhelyi. 2006. "Continuous Monitoring of Business Process Controls: A Pilot Implementation of a Continuous Auditing System at Siemens." *International Journal of Accounting Information Systems* 7(2):137–61. Retrieved March 30, 2012 (http://linkinghub.elsevier.com/retrieve/pii/S1467089506000273).

Alles, Michael G., Alexander Kogan, and Miklos A. Vasarhelyi. 2002. "Feasibility and Economics of Continuous Assurance." *Auditing: A Journal of Practice & Theory* 21(1).

Anderson, JC and JM Mueller. 2005. "The Effects of Experience and Data Presentation Format on an Auditing Judgment." *Journal of Applied Business Research* 21(1):53–63. Retrieved May 10, 2014 (http://cluteonline.com/journals/index.php/JABR/article/viewArticle/1500).

Artís, M., Mercedes Ayuso, and M. Guillen. 1999. "Modelling Different Types of Automobile Insurance Fraud Behaviour in the Spanish Market." *Insurance: Mathematics and Economics* 24:67–81. Retrieved April 18, 2015 (http://www.sciencedirect.com/science/article/pii/S0167668798000389).

Artís, Manuel, M. Ayuso, and M. Guillén. 2002. "Detection of Automobile Insurance Fraud with Discrete Choice Models and Misclassified Claims." *Journal of Risk and Insurance* 69(3):325–40. Retrieved April 18, 2015 (http://onlinelibrary.wiley.com/doi/10.1111/1539-6975.00022/full).

Bakar, Zuriana Abu, Rosmayati Mohemad, Akbar Ahmad, and Mustafa Mat Deris. 2006. "A Comparative Study for Outlier Detection Techniques in Data Mining." *IEEE Conference on Cybernetics and Intelligent Systems* 1–6. Retrieved (http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4017846).

Benbasat, Izak and Albert S. Dexter. 1986. "An Investigation of the Effectiveness of Color and Graphical Information Presentation Under Varying Time Constraints." *MIS Quarterly.* 10(1):59–83. 25p. 7 Charts.

Benbasat, Izak and Roger G. Schroeder. 1977. "An Experimental Investigation of Some MIS Design Variables." *MIS Quarterly* 1(1):37–49.

Bills, Kenneth L., Debra C. Jeter, and Sarah E. Stein. 2015. "Auditor Industry Specialization and Evidence of Cost Efficiencies in Homogenous Industries." *Accounting Review* 90(5):1721–54.

Bolton, Richard J. and David J. Hand. 2002. "Statistical Fraud Detection: A Review." *Statistical Science* 17(3):235–55.

Breunig, Mm, Hp Kriegel, Rt Ng, and Jörg Sander. 2000. "LOF: Identifying Density-Based Local Outliers." *ACM Sigmod Record* 1–12. Retrieved (http://dl.acm.org/citation.cfm?id=335388).

Cahan, Steven F., Debra C. Jeter, and Vic Naiker. 2011. "Are All Industry Specialist Auditors the Same?" *Auditing* 30(4):191–222.

Cairney, Timothy D. and George R. Young. 2006. "Homogenous Industries and Auditor Specialization: An Indication of Production Economies." *Auditing* 25(1):49–67.

Campbell, Colin and Kristin P. Bennett. 2001. "A Linear Programming Approach to Novelty Detection." *Proceedings of the Conference on Advances in Neural Information Processing* 14.

Chan, David Y. and Miklos a. Vasarhelyi. 2011. "Innovation and Practice of Continuous Auditing." *International Journal of Accounting Information Systems* 12(2):152–60. Retrieved April 18, 2015 (http://linkinghub.elsevier.com/retrieve/pii/S1467089511000029).

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. 2009. "Anomaly Detection: A Survey." *ACM Computing Surveys* 41(3):1–6. Retrieved (http://portal.acm.org/citation.cfm?doid=1541880.1541882).

Chen, Haifeng Chen Haifeng and P. Meer. 2003. "Robust Regression with Projection Based M-Estimators." *Proceedings Ninth IEEE International Conference on Computer Vision* (Iccv).

Chen, Hsinchun, Roger Chiang, and Veda C. Storey. 2012. "Business Intelligence and Analytics : From Big Data To Big Impact." *Mis Quarterly* 36(4):1165–88. Retrieved (http://web.a.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=c72752a6-fd0c-4184-ad0b-39ae0c9c16d8@sessionmgr4003&vid=1&hid=4209).

Chen, Y. and RA Leitch. 1998. “The Error Detection of Structural Analytical Procedures: A Simulation Study.” *Auditing: A Journal of Practice & Theory* 17(2 Fall). Retrieved February 27, 2015 (http://www.questia.com/library/journal/1G1-21237589/the-error-detection-of-structural-analytical-procedures).

Chen, Y. and RA Leitch. 1999. “An Analysis of the Relative Power Characteristics of Analytical Procedures.” *Auditing: A Journal of Practice & Theory* 18(2 Fall). Retrieved February 27, 2015 (http://aaajournals.org/doi/abs/10.2308/aud.1999.18.2.35).

Cogger, Kenneth O. 1981. “A Time-Series Analytic Approach to Aggregation Issues in Accounting Data.” *Journal of Accounting Research* 19(2):285–98. Retrieved (http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=6406113&site=ehost-live&scope=site).

Committee, AICPA Assurance Services Exectuve. 2013a. “Accounts Receivable Subledger Standard.” *AICPA* (August).

Committee, AICPA Assurance Services Exectuve. 2013b. “General Ledger Standard.” *AICPA* (August).

Committee, AICPA Assurance Services Executive. 2015a. *Audit Data Standards: Base Standards*. Retrieved (aicpa.org/FRC).

Committee, AICPA Assurance Services Executive. 2015b. *Audit Data Standards: General Ledger Standard*. Retrieved (aicpa.org/FRC).

Committee, AICPA Assurance Services Executive. 2015c. *Audit Data Standards: Procure to Pay Subledger Standard*. Retrieved (aicpa.org/FRC).

Committee, AICPA Assurance Services Executive. 2015d. *Audit Data Standards - O2C: Subledger Standard*. Retrieved (aicpa.org/FRC).

Dai, Jun, Qaio Li, and M. Vasarhelyi. 2016. *Designing Audit Apps for Armchair Auditors to Analyze Government Procurement Contracts*.

Danos, Paul and John W. Eichenseher. 1982. “Audit Industry Dynamics: Factors Affecting Changes in Client-Industry Market Shares.” *Journal of Accounting Research*

20(2):604–16. Retrieved (http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=6406179&site=ehost-live&scope=site).

Danos, Paul and John W. Eichenseher. 1986. “Long-Term Trends Toward Seller Concentration in the U.S. Audit Market.” *The Accounting Review* 61(4):633. Retrieved (http://proquest.umi.com/pqdweb?did=925444&Fmt=7&clientId=47941&RQT=309&VName=PQD).

DeSanctis, Gerardine. 1984. “COMPUTER GRAPHICS AS DECISION AIDS: DIRECTIONS FOR RESEARCH.” *Decision Sciences* 15(4):463–87. Retrieved May 13, 2014 (http://doi.wiley.com/10.1111/j.1540-5915.1984.tb01236.x).

Dickson, Gary W., Gerardine DeSanctis, and D. J. McBride. 1986. “Understanding the Effectiveness of Computer Graphics for Decision Support: A Cumulative Experimental Approach.” *Communications of the ACM* 29(1):40–47. Retrieved May 13, 2014 (http://dl.acm.org/citation.cfm?id=5465.5469).

Dzeng, Simon C. 1994. “A Comparison of Analytical Procedure Expectation Models Using Both Aggregate and Disaggregate Data.” *AUDITING: A Journal of Practice & Theory* 13(2).

Eichenseher, John W. and Paul Danos. 1981. “The Analysis of Industry-Specific Auditor Concentration: Towards an Explanatory Model.” *Accounting Review* 56(3):479–92. Retrieved (http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=4491911&site=ehost-live&scope=site).

Eskin, Eleazar. 2000. “Anomaly Detection over Noisy Data Using Learned Probability Distributions.” *In Proceedings of the International Conference on Machine Learning* 255–62. Retrieved (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.32.9761).

Feliciano, Gloria D., Richard D. Powers, and Bryant E. Kearl. 1963. “The Presentation of Statistical Information.” *Audio Visual communication review* 11(3):32–39. Retrieved May 13, 2014 (http://link.springer.com/article/10.1007/BF02768404).

Frownfelter-Lohrke, Cynthia. 1998. “The Effects of Differing Information Presentations of

General Purpose Financial Statements on Users' Decisions." *Journal of Information Systems* 12(2):99–107. Retrieved May 10, 2014 (http://dialnet.unirioja.es/servlet/articulo?codigo=452387).

Fung, Simon Yu Kit, Ferdinand A. Gul, and Jagan Krishnan. 2012. "City-Level Auditor Industry Specialization, Economies of Scale, and Audit Pricing." *Accounting Review* 87(4):1281–1307.

Ghosh and Reilly. 1994. "Credit Card Fraud Detection with a Neural-Network." *Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences* 3:621–30.

Hawkins, Simon, Hongxing He, Graham Williams, and Ra Baxter. 2002. "Outlier Detection Using Replicator Neural Networks." *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery* 170–80. Retrieved (http://www.springerlink.com/index/6WVYJCMEGHDD4FBN.pdf).

Henry C. Lucas, Jr. 1981. "An Experimental Investigation of the Use of Computer-Based Graphics in Decision Making." *Management Science* 27(7):757–68. Retrieved May 10, 2014 (http://pubsonline.informs.org/doi/abs/10.1287/mnsc.27.7.757).

Hodge, FD. 2001. "Hyperlinking Unaudited Information to Audited Financial Statements: Effects on Investor Judgments." *The Accounting Review* 76(4):675–91. Retrieved May 10, 2014 (http://www.aaajournals.org/doi/abs/10.2308/accr.2001.76.4.675).

Hogan, Chris E. and Debra C. Jeter. 1999. "Industry Specialization by Auditors." *Auditing* 18(1):1–17.

Hoitash, Rani, Alexander Kogan, and MA Vasarhelyi. 2006. "Peer-Based Approach for Analytical Procedures." *Auditing: A Journal of Practice & Theory* 25(2):53–84. Retrieved February 2, 2015 (http://aaajournals.org/doi/abs/10.2308/aud.2006.25.2.53).

Hu, Wenjie, Yihua Liao, and V.Rao Vemuri. 2003. "Robust Support Vector Machines for Anomaly Detection in Computer Security." *Icmla* 168–74.

Huber, Peter. 1973. "Robust Regression: Asymptotics, Conjectures and Monte Carlo." *The Annals of Statistics* 1(5):799–821.

Issa, Hussein. 2013. “Exceptional Exceptions.” Rutgers Business School.

Jarvenpaa, SL and GW Dickson. 1988. “Graphics and Managerial Decision Making: Research-Based Guidelines.” *Communications of the ACM* 37(6). Retrieved May 10, 2014 (http://dl.acm.org/citation.cfm?id=62971).

Jarvenpaa, SL, GW Dickson, and G. DeSanctis. 1985. “Methodological Issues in Experimental IS Research: Experiences and Recommendations.” *MIS quarterly* 9(June):141–57. Retrieved May 10, 2014 (http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=02767783&AN=4677438&h=YYJ8gBxDZsV0yaBMUhs6dnP/8oKhbtae+l+jUKvUnt/GJwPWNVkS/XTdjMavzi5PpvXbc8tjFpkYvVLL/CQtMA==&crl=c).

Kaplan, Steven E. 1988. “An Examination of the Effect of Presentation Format on Auditors ’ Expected Value Judgments.” *Accounting Horizons* (September):90–95.

Kelton, Andrea Seaton, Robin R. Pennington, and Brad M. Tuttle. 2010. “The Effects of Information Presentation Format on Judgment and Decision Making: A Review of the Information Systems Research.” *Journal of Information Systems* 24(2):79–105. Retrieved May 10, 2014 (http://aaajournals.org/doi/abs/10.2308/jis.2010.24.2.79).

Kinney, William R. and Linda S. McDaniel. 1989. “Characteristics of Firms Correcting Previously Reported Quarterly Earnings.” *Journal of Accounting and Economics* 11(1):71–93.

Knechel, W.Robert. 1988. “The Effectiveness of Statistical Analytical Review as a Substantive Auditing Procedure: A Simulation Analysis.” *Accounting Review* Vol. 63(1):74. 22p. </dd><dt>Document Type:</dt><dd>Article</. Retrieved (http://libaccess.mcmaster.ca.libaccess.lib.mcmaster.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=4482249&site=ehost-live&scope=site).

Knorr, Edwin M., Raymond T. Ng, and Vladimir Tucakov. 2000. “Distance-Based Outliers: Algorithms and Applications.” *The International Journal on Very Large Data Bases (The VLDB)* 8(3–4):237–53.

Kwon, Sooyoung. 1996. “The Impact of Competition within the Client ’S Industry on the Auditor Selection Decision.” *Auditing: A Journal of Practice & Theory* 15(1).

Lee, Wenke, Salvatore J. Stolfo, and K. U. I. W. Mok. 2000. “Adaptive Intrusion Detection: A Data Mining Approach*.” *Artificial Intelligence Review* 14(6):533–67.

Lev, Baruch. 1980. “On the Use of Index Models in Analytical Reviews by Auditors.” *Journal of Accounting Research* 18(2 (Autumn)):524–50. Retrieved March 6, 2015 (http://www.jstor.org/stable/2490591).

Lili, Sun, Rajendra P. Srivastava, Theodore J. Mock, and Lili Sun. 2006. “An Information Systems Security Risk Assessment Model Under the Dempster-Shafer Theory of Belief Functions.” *Journal of Management Information Systems* 22(4):109–42. Retrieved (http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=20597994&site=ehost-live&scope=site\nhttp://ezproxy.library.capella.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=20597994&site=ehost-live&scope=site).

Lucas, Henry C. and Norman R. Nielsen. 1980. “The Impact of the Mode of Information Presentation on Learning and Performance.” *Management Science* 26(10):982–93. Retrieved May 10, 2014 (http://pubsonline.informs.org/doi/abs/10.1287/mnsc.26.10.982).

Lusk, Edward J. and Michael Kersnick. 1979. “The Effect of Cognitive Style and Report Format on Task Performance: The MIS Design Consequences.” *Management Science* 25(8):787–98. Retrieved May 10, 2014 (http://pubsonline.informs.org/doi/abs/10.1287/mnsc.25.8.787).

Major, John A. and Dan R. Riedinger. 2002. “EFD : A HYBRID KNOWLEDGE / STATISTICAL - BASED SYSTEM FOR THE DETECTION OF FRAUD.” *The Journal of Risk and Insurance* 69(3):309–24.

Mock, Theodore J., Lili Sun, Rajendra P. Srivastava, and Miklos Vasarhelyi. 2009. “An Evidential Reasoning Approach to Sarbanes-Oxley Mandated Internal Control Risk Assessment.” *International Journal of Accounting Information Systems* 10(2):65–78.

Moffitt, Kevin C. and Miklos a. Vasarhelyi. 2013. “AIS in an Age of Big Data.” *Journal of Information Systems* 27(2):1–19. Retrieved January 30, 2015 (http://aaajournals.org/doi/abs/10.2308/isys-10372).

Patcha, Animesh and Jung Min Park. 2007a. “An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends.” *Computer Networks* 51(12):3448–70.

Patcha, Animesh and Jung Min Park. 2007b. “Network Anomaly Detection with Incomplete Audit Data.” *Computer Networks* 51(13):3935–55.

Pathak, Jagdish, Navneet Vidyarthi, and Scott L. Summers. 2005. “A Fuzzy-Based Algorithm For Auditors To Detect Elements Of Fraud In Settled Insurance Claims.” *Managerial Auditing Journal* 20(6):632–44.

Rousseeuw, Peter. 1984. “Least Median of Squares Regression.” *Journal of the American Statistical Association* 79(388):871–80. Retrieved (http://www.tandfonline.com/doi/abs/10.1080/01621459.1984.10477105).

Rousseeuw, Peter J. and Katrien Van Driessen. 2006. “Computing LTS Regression for Large Data Sets.” *Data Mining and Knowledge Discovery* 12(1):29–45.

Rousseeuw, Peter and V. Yohai. 1984. “Robust Regression by Means of S-Estimators.” *Lecture Notes on Statistics* 26:256–72.

Schulz, AKD and P. Booth. 1995. “THE EFFECTS OF PRESENTATION FORMAT ON THE EFFECTIVENESS AND EFFICIENCY OF AUDITORS ’ANALYTICAL REVIEW JUDGMENTS.” *Accounting & Finance* 107–31. Retrieved May 10, 2014 (http://onlinelibrary.wiley.com/doi/10.1111/j.1467-629X.1995.tb00279.x/abstract).

Shafer, Glenn R. 1996. *The Art of Causal Conjecture.* MIT press. Retrieved (https://books.google.com/books?hl=en&lr=&id=sY7os7OCykUC&oi=fnd&pg=PR13&dq=Shafer,+G.+(1996)+The+Art+of+Causal+Conjecture.+MIT+Press&ots=kphvufgezr&sig=n_V3cCw0Xgs1691asTfZpt0AYSo#v=onepage&q=Shafer, G. (1996) The Art of Causal Conjecture. MIT Press&f=f).

Shafer, Glenn R. and Rajendra P. Srivastava. 1990. “The Bayesian and Belief-Function Formalisms: A General Perspective for Auditing.” *Auditing: A Journal of Practice & Theory* 9:110–37. Retrieved (http://search.ebscohost.com/login.aspx?direct=true&db=lgh&AN=15116617&site=ehost-live).

Solberg, Helge Erik and Ari Lahti. 2005. “Detection of Outliers in Reference Distributions: Performance of Horn’s Algorithm.” *Clinical Chemistry* 51(12):2326–32.

Speier, Cheri. 2006. “The Influence of Information Presentation Formats on Complex Task Decision-Making Performance.” *International Journal of Human Computer Studies* 64(11):1115–31.

Srivastava, Rajendra P. 1993. “Belief Functions and Audit Decisions.” *Auditors Report* 17(1):8–12.

Srivastava, Rajendra P. and Theodore J. Mock. 2000. “Belief Functions in Accounting Behavioral Research.” *Advances in Accounting Behavioral Research* 3(213):225–42.

Srivastava, Rajendra P. and Theodore J. Mock. 2000. “Introduction to Belief Functions.” *Advances in Accounting Behavioral Research* 3(1990):225–42. Retrieved (http://www.gipsa-lab.fr/summerschool/bfta/includes/Denoeux_introduction_belief_functions.pdf).

Srivastava, Rajendra P. and Theodore J. Mock. 2005. “Why We Should Consider Belief Functions in Auditing Research and Practice.” *Auditors Report* 28(2):1–8.

Srivastava, Rajendra P. and Theodore J. Mock. 2011. “The Dempster-Shafer Theory of Belief Functions for Managing Uncertainties : An Introduction and Fraud Risk Assessment Illustration.” *Australian Accounting Review* 21(3):282–91.

Srivastava, Rajendra P., Theodore J. Mock, and Jerry L. Turner. 2007. “Analytical Formulas for Risk Assessment for a Class of Problems Where Risk Depends on Three Interrelated Variables.” *International Journal of Approximate Reasoning* 45(1):123–51.

Srivastava, Rajendra P., Sunita S. Rao, and Theodore J. Mock. 2013. “Planning and Evaluation of Assurance Services for Sustainability Reporting: An Evidential Reasoning Approach.” *Journal of Information Systems* 27(2):107–26. Retrieved (http://aaajournals.org/doi/abs/10.2308/isys-50564).

Srivastava, Rajendra P. and Glenn R. Shafer. 1992. “Belief-Function Formulas for Audit Risk.” *The Accounting Review* 67(2):249–2833.

Taylor, BG and LK Anderson. 1986. “MISLEADING GRAPHS-GUIDELINES FOR THE ACCOUNTANT.” *Journal of Accountancy* 162(October):126–35. Retrieved May 10, 2014 (http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:MISLEADING+GRAPHS+:+GUIDELINES+FOR+The+Accountant#0).

Taylor, Mark H. 2000. “The Effects of Industry Specialization on Auditors’ Inherent Risk Assessments and Confidence Judgements’.” *Contemporary Accounting Research* 17(4):713–15.

Tuttle, Brad M. and Russell Kershaw. 1998. “Information Presentation and Judgment Strategy from a Cognitive Fit Perspective.” *Journal of Information Systems* 12(1):1. Retrieved (http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=2489632&site=ehost-live).

Vasarhelyi, M., Michael Alles, and K. Williams. 2010. “Continuous Assurance for the Now Economy. A Thought Leadership Paper for the Institute of Chartered Accountants in Australia.” (February). Retrieved January 30, 2015 (http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Continuous+Assurance+for+the+Now+Economy+A+Thought+Leadership+Paper+for+the+Institute+of+Chartered+Accountants+in+Australia#0).

Vasarhelyi, MA, MG Alles, and A. Kogan. 2004. “Principles of Analytic Monitoring for Continuous Assurance.” *Journal of emerging Technologies in Accounting* 1:1–21. Retrieved January 30, 2015 (http://aaajournals.org/doi/abs/10.2308/jeta.2004.1.1.1).

Vasarhelyi, Miklos a. 2013. “Formalization of Standards, Automation, Robots, and IT Governance.” *Journal of Information Systems* 27(1):1–11. Retrieved January 30, 2015 (http://aaajournals.org/doi/abs/10.2308/isys-10347).

Vessey, Iris. 1991. “Cognitive Fit: A Theory-Based Analysis of the Graphs Versus Tables Literature.” *Decision Sciences* 22(2):219–40. Retrieved May 10, 2014 (http://doi.wiley.com/10.1111/j.1540-5915.1991.tb00344.x).

Vessey, Iris and Dennis Galletta. 1991. “Cognitive Fit : An Empirical Study of Information Acquisition.” *Information Systems Research* 2(1):63–84.

Viaene, S., G. Dedene, and R. Derrig. 2005. “Auto Claim Fraud Detection Using Bayesian Learning Neural Networks.” *Expert Systems with Applications* 29(3):653–66. Retrieved April 16, 2012 (http://linkinghub.elsevier.com/retrieve/pii/S0957417405000825).

Viaene, Stijn, Mercedes Ayuso, Montserrat Guillen, Dirk Van Gheel, and Guido Dedene. 2007. “Strategies for Detecting Fraudulent Claims in the Automobile Insurance Industry.” *European Journal of Operational Research* 176(1):565–83. Retrieved April 18, 2015 (http://linkinghub.elsevier.com/retrieve/pii/S0377221705006405).

Watson, Collin J. and Russell W. Driver. 1983. “The Influence of Computer Graphics on the Recall of Information.” *MIS Quarterly* 7(1):45. Retrieved May 10, 2014 (http://dl.acm.org/citation.cfm?id=2017600.2017604).

Williams, Graham and Rohan Baxter. 2002. “A Comparative Study of RNN for Outlier Detection in Data Mining.” *IEEE International Conference on Data Mining* (December 2002):1–16. Retrieved (http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1184035).

Wong, Wk, a Moore, G. Cooper, and M. Wagner. 2003. “Bayesian Network Anomaly Pattern Detection for Disease Outbreaks.” *Icml* 808–15. Retrieved (http://www.aaai.org/Papers/ICML/2003/ICML03-105.pdf).

Wright, Sally and Arnold M. Wright. 1997. “The Effect of Industry Experience on Hypothesis Generation and Audit Planning Decisions.” *Behavioral Research in Accounting* 9:273. Retrieved (http://proquest.umi.com/pqdweb?did=13424383&Fmt=7&clientId=47297&RQT=309&VName=PQD).

Yamanishi, K., JI Takeuchi, Graham Williams, and Peter Milne. 2004. “On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms.” *Data Mining and Knowledge Discovery* 8:275–300. Retrieved April 18, 2012 (http://dl.acm.org/citation.cfm?id=347160).

Yohai, Victor J. 1987. “High Breakdown-Point and High Efficiency Robust Estimates for Regression.” *The Annals of Statistics* 15(2):642–56.

Yu, Dantong, Gholamhosein Sheikholeslami, and Aidong Zhang. 2002. “FindOut : Finding Outliers in Very Large Datasets.” *Knowledge and Information Systems* 4(4):387–412.

Zhang, Li, Amy R. Pawlicki, Dorothy McQuilken, and William R. Titera. 2012. “The AICPA Assurance Services Executive Committee Emerging Assurance Technologies Task Force: The Audit Data Standards (ADS) Initiative.” *Journal of Information Systems* 26(1):199–205. Retrieved January 30, 2015 (http://aaajournals.org/doi/abs/10.2308/isys-10277).

Zmud, Robert W. 1978. “AN EMPIRICAL INVESTIGATION OF THE DIMENSIONALITY OF THE CONCEPT OF INFORMATION.” *Decision Sciences* 9(2):187–95. Retrieved May 13, 2014 (http://doi.wiley.com/10.1111/j.1540-5915.1978.tb01378.x).