Towards a more efficient audit process:

A data-driven approach

by

Joost Vandewal

BSc Industrial Engineering - 2010
Student Number 0745379

In partial fulfillment of the requirements for the degree of

Master of Science
in
Innovation Management

Supervisors:
Dr. Rui Jorge de Almeida – TU/e
Dr. Ir. Remco Dijkman – TU/e
MSc. M. van der Kruijff
MSc. L. Bakker
TU/e Industrial Engineering & Innovation Sciences

Series Master Theses in Innovation Management

Subject Headings: Business Process Management, Classification, Clustering, Data Analytics, Data Science, Data Mining, Design Science, Financial Statement Auditing, Frequent Patterns, Journal Entry Testing, Operational Efficiency
Abstract

A financial statement audit is critical for users of an annual financial statement. Users, like investors, base their decisions on the results of a company. The financial statement audit validates that these results present a true and fair image of the company. One of the operations within the financial statement audit is journal entry testing, which is a business process susceptible to intuitive decision making based on proven business rules and subjective professional judgment. This thesis explores the opportunity for a data-driven approach to journal entry testing that facilitates the current process in a sense that increases operational efficiency. First, several applications have been suggested that improve data-analytic decision making. As a stepping stone towards data-analytical thinking, we introduce automatic bucketing to detect routine behavior in an auditee’s financial transactions. In this way homogeneity is found in a set of transactions that were presumed heterogeneous. Auditors can apply different procedures to the homogeneous buckets of journal entries, which increases the operational efficiency over the entire audit engagement. Although this requires a change of the current way of working, auditors see value in this increased usage of data-analytics.
Management Summary

Introduction

Journal entry testing is a business process in which auditors, i.e. people that are authorized to systematically and independently examine if financial statements represent a true and fair view of the financial position of a firm, provide assurance that management override of control has not occurred. Since management is “...in a unique position to perpetrate fraud because of management’s ability to manipulate accounting records and prepare fraudulent financial statements by overriding controls that otherwise appear to be operating effectively” (ISA, 2009, pp. 162–163), journal entry testing requires auditors to identify and test manually initiated journal entries that are not understood/expected. Currently, this process entails an intuition-based decision making process, in which auditors’ subjective understanding of the client’s business determines which criteria are used to identify high-risk journal entries. Research indicates that there is value in applying a more data-driven approach of utilizing data analytics in an audit setting (Bagga & Singh, 2011; Gray & Debreceny, 2014; Qi Liu & Vasarhelyi, 2014). Applying data analytics goes hand in hand with involving IT advisors in the financial statement audit. Improved involvement and data-analytical thinking occurs if auditors are willing to invest time in this redesigned situation. It is likely that data-analytical thinking leads to increased business understanding over multiple years an auditee remains a client. During the entire engagement, i.e. subsequent years in which an a company remains an audit client, this increased business understanding might lead to efficiency gains.

An emerging way of stimulating data-analytical thinking is to utilize data mining applications to discover value in the data of a company (Provost & Fawcett, 2013). In combination with the diagnosed situation, this thought inspired us to set up the following research question:

*How can a data mining application increase engagement efficiency?*

Research Methodology

Answering the research question will be subdivided in three different parts, in which Figure 1 presents the research design. Firstly, an analysis of literature on data mining journal entries will be conducted guiding the setup of solution directions. The following research methods have been applied:

1. Problem Definition – Unstructured interviews and a document analysis have been conducted to gain understanding of the current situation and detect inefficiencies at a Case Company.
2. Analysis & Diagnosis – Unstructured interviews, semi-structured interviews, a literature review, and a document analysis have been conducted as a way to model the current situation and come up with ideas for solution directions.

Secondly, one of the suggested solution directions that stimulates data-analytic thinking is developed and tested. The following research methods have been applied:

3. Plan of Action – The methodology of CRISP-DM (Shearer, 2000) has been followed to model and test a tool that is able to increase a data-analytical mindset. The tool has been evaluated using the manual journal entries of three different entities to achieve some level of generalizability.
Lastly, auditors at a Case Company evaluated whether a hypothetical redesign involving the tool would lead to increased engagement efficiency. Mechanisms through which the redesign would lead to efficiency are translated in design principles. The following research methods have been applied:

4. **Intervention** – Semi-structured interviews were held in which the tool was linked to specific data of their own client. This enabled us to gain a good understanding of both the auditor’s acceptance as the mechanisms through which this solution direction would be effective.

5. **Evaluation** – Using semi-structured interviews, the mechanisms through which the redesign would be (in)effective have been discussed with auditors in order to link it to engagement efficiency.

---

### Problem Description (Ch. 1)

### Analysis & Diagnosis (Ch. 3 & 4)

### Plan of Action (Ch. 4 & 5)

### Intervention (Ch. 5)

### Evaluation (Ch. 6)

---

**Figure 1: Research design of this master’s thesis**

**As-Is Business Understanding**

Journal entry testing (Figure 2) is an important element of the financial statement audit, since it controls management override of controls. The process is visualized in Figure 2, and it starts with a **Risk Assessment**. In this phase the client’s business risks are discussed among the auditors. After receiving the manual journal entries from the IT advisors (**Journal Entry Extraction**), who are required to understand which population is not yet fully controlled (**IT Control Testing**), auditors discuss the setup of high-risk criteria that would minimize the risk of management override of control. A process called **Filtering** follows as a way to drill down to a population of journal entries falling within the high-risk criteria. The labeled high-risk journal entries are tested using a thorough investigation at the client in the **Testing** phase. The decisions made to drill down to a high-risk population are based on intuition, in which they do not exploit the data to tell what is worth investigating further, i.e. data-analytical thinking.

---

**Solution Direction: Automatic Bucketing**

Applying data mining is an effective way to introduce data-analytical thinking (Provost & Fawcett, 2013). The goal of journal entry testing is to select high-risk manual journal entries, that are not fully understood or controlled, for further analysis. By detecting routine behavior based on previous years’ audits, we have
been able to isolate the known and recurring patterns. Moreover, this leaves us with a “bucket” of journal entries that are heterogeneous by nature and might be susceptible to management override of control. By only incorporating the bucket of heterogeneous manual journal entries as input for the filtering phase, journal entry testing will be more efficient.

In the design phase of this research a tool has been developed that can automatically detect recurring journal entries that are profiled and subsequently bucketed by means of their risk level. This tool is called “Automatic Bucketing”. For this to happen, the behavior evident in journal entries has been modeled. Then, clustering has been applied to find routine behavior in previous year’s manual journal entries. Subsequently, the clusters have been linked to a profile that explains their behavior best. As a next step, we have developed a generic way of detecting profiled journal entries in subsequent year’s data of a client. In this way, auditors would gain a better understanding of the behavior of journal entries which increases their data-analytical mindset and business understanding. Additionally, actively analyzing and discussing the buckets would imply involving IT advisors as being data analysts which increases the communication between the two parties.

**Engagement Efficiency**

The mechanisms through which operational efficiency would be gained are linked to design principles. Design principles are identified as descriptive knowledge that can be transformed into actionable design knowledge (J. E. Van Aken, 2005). Design principles have the ability to link practical knowledge to insights for future research, but also to show practitioners the advantages of data-analytical thinking. Four design principles have been proposed that would increase operational efficiency in the long term.

**IT Advisory Involvement**

In journal entry testing, in which manual journal entries are tested for management override of controls, (automatic) bucketing of journal entries could lead to increased communication between auditors and IT advisors which initiates a cooperative group climate

**Internal Control Implications**

In journal entry testing, in which manual journal entries are tested for management override of controls, (automatic) bucketing of journal entries could lead to identifying control inefficiencies which increases the sharing of internal control implications with the client

**Data-Analytical Mindset**

In journal entry testing, in which manual journal entries are tested for management override of controls, (automatic) bucketing of journal entries leads to an increased understanding of an auditee which increases the quality of an audit

**Effective Sampling**

In journal entry testing, in which manual journal entries are tested for management override of controls, (automatic) bucketing of journal entries could lead to sample or exclude understandable buckets which decreases the throughput time

**Conclusion**

Auditors see value in a situation in which data analytics is part of the journal entry testing process. Especially in the long run, over a multi-year audit engagement, the results will become apparent. The new situation requires a disruptive redesign involving an active involvement of IT advisors as data analysts. The
data-driven approach leads to IT advisors being more involved in a financial statement audit and auditors being able to apply their professional judgment to the most interesting cases. This highly efficient way of understanding journal entries is a stepping stone toward future research.

Revisiting the research question, engagement efficiency can be improved if data mining is used as a way of isolating recurring patterns. Auditors should not treat manual journal entries as a heterogeneous dataset, but profile and bucket them in terms of their behavior. This leaves auditors with the most high-risk journal entries worth investigating using the current testing procedures. Since this data-analytical way of treating journal entries requires the involvement of IT advisors, a breeding ground for increased communication is created which has multiple advantages on other places in the financial statement audit.

Future researchers could continue with this initial prototype by advancing the tool to encompass all additional behavior patterns, only leaving the incidental manual journal entries. In this way profiling “normal” behavior is optimized and semi-supervised anomaly detection has a possibility to enhance journal entry testing. Future practitioners could use the design principles as a way to redesign their financial statement audit using the idea of automatic bucketing suggested in this research.
Preface

This report concludes my Thesis Project developed under the aegis of the Eindhoven University of Technology. I envisioned a research which combines the application of data mining techniques within the context of my MSc program “Innovation Management”. To be honest, I think I did!

Feeling enthusiastic about all the possibilities data analytics can offer, I started my Master’s Thesis. So many challenges ahead which all required me to learn and study research fields not yet encountered before. What is data mining? What is accounting? What is auditing? Not knowing whether the results made sense was quite nerve-racking sometimes, however, the support of people at the case company, friends, and family made this journey worthwhile.

I would like to thank the supervisors of this project Dr. Rui Jorge de Almeida and Dr.Ir. Remco Dijkman. Rui, your broad knowledge base, willingness to guide me, and enthusiasm was really inspiring. I enjoyed the meetings we had where I always walked away with new thoughts. Also Remco, I would like to thank you for the “business process management” view on my thesis. Thank you for the meetings and insightful comments.

Special thanks to the company supervisors for both their time investment and interest. Laurens, thank you for your knowledge on, actually, everything. You guided me through choosing an interesting research topic, helped me contact the right people, and brainstormed with me on results. Thank you very much for this. Maarten, I would thank you for being my supervisor from the approval of my research proposal until the very day of the presentation. Your creative ideas and attention for detail helped me greatly with converging all floating thoughts into one concrete research project. It is a great pleasure working with you, thank you!

Ik wil ook graag mijn vrienden bedanken die mijn studententijd zo geweldig hebben gemaakt. Met vriendengroepen als l’Eon Dix, VC Witj Nemus en Nik(e)s was er voldoende te doen in de studievrije momenten! Ook wil ik graag de mensen die ik bij Industria en Interactie heb leren kennen bedanken. Een belangrijk en leuk project heb ik neergezet met het Industria Congres. Bedankt voor de top samenwerking! Een belangrijke fase in mijn leven was een internationaal semester naar het mooie Porto. Gostaria de agradecer a todos que feziram do meu tempo no Porto inesquecível. As memórias desse tempo fizeram-me continuar durante a tese!

Last, but definitely not least, wil ik graag mijn ouders en vriendin bedanken! Zonder de onvoorwaardelijke steun gedurende mijn studentenleven zou ik het niet succesvol afgerond hebben.

I hope you will enjoy reading this thesis!

Joost Vandewal

Eindhoven, July 2016
Table of Contents

Abstract .......................................................................................................................................................... i
Management Summary ................................................................................................................................ ii
Preface ......................................................................................................................................................... vi
List of Figures ................................................................................................................................................ x
List of Tables ............................................................................................................................................... xii
List of Equations .......................................................................................................................................... xii
List of Abbreviations ................................................................................................................................... xiii
List of Definitions ........................................................................................................................................ xiii

1. Introduction .......................................................................................................................................... 1
   1.1 Problem Statement ....................................................................................................................... 2
   1.2 Research Objectives ...................................................................................................................... 2
   1.3 Research Fields Involved ............................................................................................................... 3
   1.4 Research Questions & Thesis Outline ........................................................................................... 4
   1.5 Contribution to Research & Work Developed .............................................................................. 5

2. Research Methodology ......................................................................................................................... 7
   2.1 Design Focused Business Solving .................................................................................................. 7
   2.2 Research Design ............................................................................................................................ 7
       2.2.1 Problem Definition and Analysis & Diagnosis ....................................................................... 8
       2.2.2 Plan of Action ........................................................................................................................ 9
       2.2.3 Evaluation by Design Principles ............................................................................................ 9
   2.3 Rigor & Relevance ....................................................................................................................... 10

3. As-Is Business Understanding ............................................................................................................. 11
   3.1 Journal Entry ............................................................................................................................... 11
   3.2 Journal Entry Testing ................................................................................................................... 12

4. Solution Direction ............................................................................................................................... 15
   4.1 Data Mining .................................................................................................................................... 15
       4.1.1 What is Data Mining? .......................................................................................................... 15
       4.1.2 Data Mining Functionalities ................................................................................................. 16
   4.2 Data Mining Opportunities in Journal Entry Testing Process ..................................................... 18

5. Developing a Tool for Automatic Bucketing ....................................................................................... 24
   5.1 The Foundation of Automatic Bucketing ....................................................................................... 24
       5.1.1 Data-Analytical Viewpoint .................................................................................................. 25
List of Figures

Figure 1: Research design of this master’s thesis ........................................................................................................ iii
Figure 2: Journal entry testing business process ........................................................................................................ iii
Figure 3: Conceptualization of involved research fields ............................................................................................... 3
Figure 4: Conceptualization of specific research fields ................................................................................................... 4
Figure 5: Thesis outline accompanied by the sub research questions ............................................................................... 5
Figure 6: Regulative cycle for business problem solving (Van Strien, 1997) ................................................................. 7
Figure 7: Research design of this master’s thesis ........................................................................................................... 8
Figure 8: The research design cycle from a science-based design perspective (Van Burg et al., 2008) .................10
Figure 9: Design science research cycles (Hevner, 2007) .............................................................................................. 10
Figure 10: Model of journal entry testing process in business process model notation .............................................13
Figure 11: Relative frequency of research on data-oriented solutions in audit (Gray & Debreceny, 2014) ............16
Figure 12: Conceptual representation of automatic bucketing functionality ............................................................... 24
Figure 13: Conceptualization of profiling & anomaly detection .................................................................................... 25
Figure 14: Subdivision of relevant population "manual" journal entries ......................................................................... 30
Figure 15: Number of journal entry lines (left) and journal entries (right) posted per month in dataset A2 ............30
Figure 16: Total number of manual journal entry lines per category in dataset A2 .................................................... 31
Figure 17: Line items posted over time in dataset A2 .................................................................................................... 31
Figure 18: Accumulated amounts posted per category involving manual journal entries in dataset A2 .............32
Figure 19: Heatmaps of amounts posted to accounts in dataset A2 on specific weekdays (left: 2014, right: 2015) ..32
Figure 20: Association rules in dataset A2 2014 ........................................................................................................... 33
Figure 21: Association rules in dataset A2 2015 ............................................................................................................ 33
Figure 22: VOS matrix visualization of dataset A2 in 2014 ........................................................................................ 34
Figure 23: VOS matrix visualization of dataset A2 in 2015 ........................................................................................ 34
Figure 24: Conceptual representation of automatic bucketing in a fictitious dataset ................................................... 35
Figure 25: Experimental setup for profiling ................................................................................................................ 36
Figure 26: Utilized distance measures ........................................................................................................................ 37
Figure 27: Summary of profiles in 2014 ........................................................................................................................ 39
Figure 28: F-measures of identified clustered profiles 2014 ....................................................................................... 42
Figure 29: Maximum F-measure per clustering technique ........................................................................................ 42
Figure 30: Experimental setup for detection ................................................................................................................ 43
Figure 31: Summary of profiles in 2015 .......................................................................................................................... 46
Figure 32: F-measures of detected profiles 2014 ......................................................................................................... 47
Figure 33: Recall measures of classification models’ ability to learn impure profiles due to FNs ............................48
Figure 34: Classification’s ability to learn on purely predicted clusters .................................................................... .48
Figure 35: The impact of automatic bucketing profiles in entity A 2015 ................................................................. 51
Figure 36: Devil’s quadrangle (Reijers & Mansar, 2005) ........................................................................................... 52
Figure 37: Design principles linking automatic bucketing and engagement efficiency ......................................... 60
Figure 38: Cause-and-effect diagram of encountered problems and their causes within the CPA ..................70
Figure 39: Conceptual presentation of the difference between audit process and engagement process ...............71
Figure 40: Flow from transactions to financial statements (Elder et al., 2012) .......................................................... 75
Figure 41: Risk assessment sub process ...................................................................................................................... 76
Figure 42: IT control test sub process ........................................................................................................................ 76
Figure 43: Journal entry extraction sub process ......................................................................................................... 76
Figure 44: Filtering sub process ...................................................................................................................................... 76
Figure 45: Layering sub process ................................................................................................................................... 77
Figure 46: Testing sub process ....................................................................................................................................... 77
Figure 47: Data mining functionalities (Han et al., 2012; Provost & Fawcett, 2013) .............................................. 80
List of Tables

Table 1: Exemplar journal entry indicating its attributes .............................................................................................................12
Table 2: Research on the application of data mining in the journal entry testing process .................................................................19
Table 3: Ideas for the application of data mining in the journal entry testing process ......................................................................21
Table 4: Conceptual data setup to find routine behavior in journal entries ..................................................................................26
Table 5: Characteristics of entities involved in research ................................................................................................................27
Table 6: Information on subdivided journal entries in entity A, 2014 ........................................................................................29
Table 7: Information on subdivided journal entries in entity A, 2015 ........................................................................................30
Table 8: Parameter setting for applied clustering techniques ........................................................................................................38
Table 9: Confusion matrix ..................................................................................................................................................................40
Table 10: Averaged F-measures on the profiles detected by certain classification techniques ..........................................................49
Table 11: Theoretical grounding for design principle 1 ..........................................................................................................................54
Table 12: Theoretical grounding for design principle 2 ..........................................................................................................................56
Table 13: Theoretical grounding for design principle 3 ..........................................................................................................................57
Table 14: Theoretical grounding for design principle 4 ..........................................................................................................................59
Table 15: List of unstructured interviews held during the first two stages of this research .................................................................69
Table 16: Interviewed researchers for validating journal entry testing process and exploring solution directions ................................72
Table 17: Interviewed auditors for validating journal entry testing process .....................................................................................73
Table 18: Business rules that could be applied if auditors think it is applicable .............................................................................78
Table 19: Decisions based on professional judgment that could be applied if auditors think it is applicable ........................................78
Table 20: List of auditors with whom interviews have been held ........................................................................................................84
Table 21: Summary on interviews regarding: "Validating Profile Selection" ..................................................................................84
Table 22: Summary on interviews regarding: "Bucket of Intra-Category Bookings" ..........................................................................85
Table 23: Summary on interviews regarding: "Automatic Bucketing Exploration" ........................................................................86
Table 24: Summary on interviews regarding: "Profile Analysis" .........................................................................................................87
Table 25: Summary on interviews regarding: "Devil's Quadrangle " .................................................................................................88
Table 26: Summary on interviews regarding: "IT Advisory Involvement" ..........................................................................................89
Table 27: Summary on interviews regarding: "Internal Control Implications" ..................................................................................90
Table 28: Summary on interviews regarding: "Data-Analytical Mindset" .......................................................................................91
Table 29: Summary on interviews regarding: "Effective Sampling" ...................................................................................................92
Table 30: Summary on interviews regarding: "Other Remarks" ......................................................................................................94
Table 31: Attributes involved in the set of journal entries in entity A ..............................................................................................101
Table 32: Attributes involved in the set of journal entries in entity B ..............................................................................................102
Table 33: Attributes involved in the set of journal entries in entity C ...........................................................................................103
Table 34: Data preparation decisions 2014 ........................................................................................................................................105
Table 35: Data preparation decisions 2015 ........................................................................................................................................106
Table 36: Information on subdivided journal entries in entity B, 2014 ..........................................................................................107
Table 37: Information on subdivided journal entries in entity B, 2015 ..........................................................................................107
Table 38: Information on subdivided journal entries in entity C, 2014 ..........................................................................................110
Table 39: Information on subdivided journal entries in entity C, 2015 ..........................................................................................110
Table 40: Cluster models used for labeling profiles of journal entries ............................................................................................126

List of Equations

\[ v_i' = \frac{(v_i - \min A)}{(\max A - \min A)} \] (1) .................................................................................................................................36

\[ \text{Precision} = \frac{TP}{(TP + FP)} \] (2) ........................................................................................................................................40

\[ \text{Recall} = \frac{TP}{(TP + FN)} \] (3) .....................................................................................................................................40

\[ \text{F-measure} = 2 * \left( \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \right) \] (4) ..................................................................................................................................40
List of Abbreviations

A/P: Accounts Payable
ANN: Artificial Neural Network
BPMN: Business Process Model and Notation
CAAT: Computer Assisted Auditing Techniques/Tools
CPA: Certified Public Accounting (Firm)
CR: Credit
CRISP-DM: Cross Industry Standard Process for Data Mining
DBSCAN: Density-Based Spatial Clustering of Applications with Noise
DR: Debit
EM: Expectation-Maximization
FP: False Positive
FN: False Negative
GAAP: Generally Accepted Accounting Principles
GITC: General Information Technology Control
ID: Identifier
ISA: International Standard on Auditing
IT: Information Technology
ITC: IT Control
JE: Journal Entry
$k$NN: $k$-Nearest Neighbors algorithm
SAS: Statement of Auditing Standard
SMOTE: Synthetic Minority Over-Sampling Technique
SOM: Self-Organizing Maps
SOX: Sarbanes-Oxley (Law)
TP: True Positive
TN: True Negative
VOS: Visualization of Similarities

List of Definitions

Accounts: A record in the general ledger that collects and stores debited and credited amounts.
Category: An aggregated term capturing accounts that have accounting-wise a similar behavior.
Cluster: A subset of journal entries that were labeled by a data mining model to be similar to each other.
Data mining application: A structured synthesis of one or several data mining models that aim at solving a data mining problem.
Data mining functionality: The purpose of a data mining application.
Data mining model: A specific parameter setting within a specific data mining technique used to obtain (part of the) results in line with the functionality.
Data mining technique: An algorithm that is proven to enable the mining of data.
Profile: A cluster of journal entries that represents a meaningful behavioral pattern.
Tool: A data mining application that can be integrated in software and/or consisting out of guidelines to prepare and model data for the purpose of solving a data mining problem.
1. Introduction

“It is a capital mistake to theorize before one has data.” - Sherlock Holmes by Arthur Conan Doyle, 1892 -

Imagine a world where an auditor places a flash drive in the client’s computer and uploads a program that scans the entire financial infrastructure of a company. The auditor looks at performance indicators and sees which transactions are red flags. The computer complements the procedure by advising the auditor on applicable follow-up procedures. An ideal financial statement audit like this would strongly decrease the time required and it lets the auditor’s time being spent on the most interesting cases in which its expertise is needed the most. The aforementioned anecdote might entail a futuristic view, but the technology is theoretically ready for this. In practice, some issues cause this evolution to lag behind. One of them is the way a company’s financial activities are tracked by accounting information systems. Besides the automatically managed transactions, there are simply too many financial transactions that could manually be initiated or altered which leaves the accounting information system unaware of the “black box of human intervention”. Since companies are not ready to make every transaction automatic, there remains room for erroneous and fraudulent human activities that needs to be verified by an auditor.

One of the processes executed in a financial statement audit is the manual journal entry testing process. This is a test to provide reasonable assurance that management override of control has not occurred: “Management is in a unique position to perpetrate fraud because of management’s ability to manipulate accounting records and prepare fraudulent financial statements by overriding controls that otherwise appear to be operating effectively” (ISA, 2009, pp. 162–163). Journal entry testing states that auditors should understand the risk environment of a company and are able to filter down a set of manually initiated journal entries to a subset of high-risk journal entries worth investigating. At the moment this process entails an intuition-based decision making process, in which auditor’s subjective understanding of the client’s business determines the high-risk criteria that are set up. Current research indicates that there is value in applying a more data-driven approach of utilizing data analytics in an audit setting (Bagga & Singh, 2011; Gray & Debreceny, 2014; Qi Liu & Vasarhelyi, 2014). In addition, standards like ISA 240 recommend computer-assisted techniques, such as data mining, to find anomalies in the data (IAASB, 2009). Also practitioners envision a situation in which data analytics fully complement the process of auditors (Byrnes, Criste, Stewart, & Vasarhelyi, 2014; Earley, 2015). For this to happen, an important mind shift has to occur. Inspired by this changing environment, a research on the effect of analytical approaches on accounting data is chosen. This research aims at setting a next step towards more efficient and effective journal entry testing by incorporating a more sophisticated computer assisted auditing tool (CAAT). Our aim is to propose a new application that could facilitate auditors in their decision making process accompanied by guidance as to how the changes actually lead to efficiency.

Several types of auditors could be distinguished in practice. The most common are certified public accounting firms (CPAs), government accountability office auditors, revenue agents, and internal auditors (Elder, Beasley, & Arens, 2012). In this Master’s Thesis the emphasis is on certified public accounting firms (CPAs). A CPA is often called an external auditor and has its main focus on auditing publicly traded companies. These are often large companies with a strong financial impact on an economy that legally require an independent audit of their annual financial statement.
1.1 Problem Statement

This research has been conducted at one of the largest CPAs in the world. In this Chapter, the current situation encountered at this CPA is explained. Although current regulations are followed diligently by this CPA, there always remains room for new theoretical developments. An orientation by having 16 unstructured interviews with auditors, IT advisors, and researchers within the CPA and attending two workshops resulted in a cause-and-effect diagram (Appendix 1 & 2) in which a lack of incorporating data analytics and IT advisors in an audit has been seen as the root cause of process inefficiencies in practice. IT advisors are specialized in the extraction of electronic evidence and conducting IT audits, but also in analyzing data (CAQ, 2008; Curtis, Jenkins, Bedard, & Deis, 2009; Yang & Guan, 2004). In practice, they are involved when auditors require an idea of the risks related to an accounting information infrastructure, or need a specific dataset to be extracted. It is exactly this notion, namely IT advisors’ ability to fulfill more roles than solely IT audits and data extraction, that should be noted and is an important starting point in this research.

Currently, auditors apply an intuition-based way of decision making during journal entry testing based on professional judgment and proven business rules. Exemplars of decisions taken can be found in Appendix 7. Since data is not presented intuitively, no insight is gained in the behavior of the total set of manual journal entries. This leads to a vicious circle, in which not knowing what information is in the data, leads to not actively involving IT advisors in parts of the financial statement audit. This situation limits the CPA’s ability to utilize opportunities for improvement in their business process. Some elements auditors require in a novel and more efficient situation are:

- More insight in behavior of journal entries.
- Visualizations that enable them to detect anomalous behavior.
- Outsourcing of non-core accounting activities to IT advisors.

In summary, at the moment auditors do not apply a data-driven approach in their endeavor to check risk of management override and we think that both practical and academic relevance can be gained in doing so. By showing auditors the advantages of incorporating a data-analytical mindset, a more efficient journal entry testing process will arise.

1.2 Research Objectives

A clear definition of the scope of the research is essential. This research aims at developing and introducing a tool novel to the field of journal entry testing. Since this test is required by law for CPAs, the research is focused on these (international) accounting firms. After an exploratory analysis within the Case Company, it is found that auditors and IT advisors lack communication since they do not speak each other’s language. Auditors are not thinking data-analytically on how data might give them insights in terms of anomalous behavior, in which IT advisors’ procedures are limited to executing requests of auditors without thinking along on how auditors could be facilitated. The objective of this research is decreasing this communication gap. A first step towards closing this gap is showing auditors that exploring the data could provide novel insights. Since IT advisors are the designated persons that analyze the data and are involved in discussing its results with auditors, their understanding of the financial statement audit increases. The envisioned situation increases the business understanding of both auditors and IT advisors. Since this entails a quality improvement that might be worth the time investment, this research is put in
light of its impact on operational efficiency. By linking the current situation within the CPA with research gaps in literature, we are able to provide new insights. The objective of this research is fivefold:

1. Review current literature on applying data mining to journal entries
2. Provide ideas on how data mining could increase efficiencies in journal entry testing
3. Develop a tool that sets a subsequent step in this growing field of research
4. Explore auditors’ attitudes towards embedding the tool in a redesigned journal entry testing process
5. Propose design principles for both researchers as practitioners in using data analytics as an ability to increase efficiencies

1.3 Research Fields Involved

The research is based on bringing together three research fields which are barely connected in current literature. This is already indicative for the explorative nature of this research. In this Chapter we shortly introduce the three research fields, visualized in Figure 3, which also facilitates the understanding of definitions used in this report.

Financial Statement Audit

Financial statement auditing implies systematically and independently examining the truthfulness and fairness of the information a company presents regarding its financial situation (Elder et al., 2012). A thorough audit of financial statements fosters confidence in stakeholders regarding both the reliability and accuracy of the business’ administration; and the accordance with external boundaries like the Generally Accepted Accounting Principles (GAAP) and Sarbanes-Oxley (SOX) legislation (Van Der Aalst, Van Hee, Van Der Werf, Kumar, & Verdonk, 2011). An auditor gathers verifiable evidence in a company that has to be sufficient in both quality as well as volume. Gathered evidence consists of electronic and documentary data about transactions, written and electronic communication, observations, and oral testimonies of the auditee (Elder et al., 2012).

By independently auditing a company’s business, business shareholders and other stakeholders can be assured that the financial information a company publishes, i.e. an annual report, either does or does not represent a true and fair view of the company. Besides the legal obligation for publicly traded companies, having an external audit trademark adds value to a company. Managers can feel in control over their systems and place trust in their operational value. Furthermore, when the financial statement of a company is assured by an external auditor the information risk could be minimized. Information risk is important for financial institutions in that it reflects the trueness of the company’s data upon which it had based its interest rate. The reduction of information risk can therefore have a significant effect on the company’s ability to obtain capital at a reasonable cost (Elder et al., 2012).

Computer Assisted Auditing Tools (CAATs)

Information technology (IT) is becoming an integral part of an external audit. Computer Assisted Auditing Tools (CAATs) can strongly enhance the effectiveness of an audit which not only improves the independent assurance that the company is free from material misstatement, but it also gives more insight in the financial infrastructure of the auditee. CAATs utilized in the journal entry testing process are mostly based on simple analytical procedures aimed at “drilling down” into the underlying details (Titera, 2013). It aims at what reasonably can be expected using guidelines, but it does not facilitate the
unsupervised discovering of new insights. Enabling auditors to work with advanced tools could increase the effectivity of an audit (Titera, 2013), which explains the emerging popularity of data-analytic tools in auditing (A. Sharma & Panigrahi, 2012).

**Operational Efficiency**

Facilitating auditors during an audit improves the efficiency of their operations. Inspired by Reider (2002) and Dumas, La Rosa, Mendling, & Reijers (2013) we see operational efficiency as providing a better output value/input value ratio in the engagement process than before. Efficiency in an engagement process could be achieved by eliminating non-value adding or time consuming steps or replacing them by value-adding steps. Mansar & Reijers (2005) link operational efficiency to business process redesign heuristics as a way to generalize practical guidelines. These guidelines could in their turn be used as a way to link findings to operational efficiency gains.

**Sweet Spot of Overlapping Fields**

In the sweet spot of overlapping fields, as can be seen in Figure 4, the three research fields merge in a sweet spot worth adding theory to. A part of the financial statement audit, in which we focus on in this Master’s Thesis is journal entry testing which is discussed in Chapter 3.2. A collection of CAATs that could assist auditors in their procedures are the data mining applications. Chapter 4.1 handles the nature of data mining, whereafter Chapter 4.2 links the research field of data mining to journal entry testing. The research field of operational efficiency could be linked to efficiency gains in specific audit processes. Since the content of audit processes are very dependent on the understanding of auditee’s business, it is thought that looking at the long term would make the redesign initiatives more apparent by means of a learning effect (Beck & Wu, 2006; Earley, 2015). An (audit) engagement is the entire period a company is auditee of a CPA (Appendix 3), which enables us to specify operational efficiency to audit engagement efficiency.

### 1.4 Research Questions & Thesis Outline

A discussed before, literature on this research field is not yet well-tested which limits the ability to actively incorporate current research findings. Knowing this made us conduct a research that is exploratory by nature. We will develop and test a data mining based tool that analyses journal entries and provides auditors with an increased understanding of the auditee’s business. Auditors utilize this understanding throughout the entire engagement period which enables them to focus on the most anomalous journal entries. After taking into account previously mentioned research objectives, the main research question that will be answered in this research is:

*How can a data mining application improve engagement efficiency?*

The research question has been answered by addressing sub research questions. It is important to understand what is journal entry testing to know how auditors go through this process. Next, linking data mining applications to the specific steps in journal entry testing implies ways that data mining lead to efficiency gains in an audit engagement. In addition, we think it is important to answer the research
question by means of a concrete tool. This tool is called “Automatic Bucketing” and it utilizes data mining techniques to find routinely occurring journal entries. Automatic bucketing evokes certain mechanisms that increases engagement efficiency. By highlighting these, ideas for how data mining can improve engagement efficiency are proposed to coherently answer the research question. Every Chapter is built up with explaining important elements needed to answer the sub research questions set up below:

1. How can data mining facilitate the journal entry testing process?
2. What is automatic bucketing?
3. Can data mining be utilized to automatically bucket journal entries?
4. How can automatic bucketing increase engagement efficiency?

Figure 5 presents the thesis outline involving the sub research questions and the elements that contextualize these sub research questions. In addition, in Chapter 1 the introduction is described, In Chapter 2 the research methodology is described, and Chapter 7 is a conclusion of the research involving a discussion, practical recommendations, ideas for future research, limitations to the research, and a note on the ethical considerations.

Figure 5: Thesis outline accompanied by the sub research questions

1.5 Contribution to Research & Work Developed

After reviewing literature on data mining journal entries, two research fields have been noticed. One is related to the detection of fraud. This field is based on trying to find out which transactions have turned out to be fraudulent in the past, and what can be learned from that (i.e. effectivity of fraud detection tools). Another field of research is related to the audit process and the way auditors could utilize data mining techniques as a computer-assisted auditing tool (i.e. efficiency in audit process). In this research we focused on the second field, namely the possible efficiency gains in the engagement process. Essentially, data mining journal entries for audit facilitation is an emerging field (R. S. Debreceny & Gray, 2010; Gray & Debreceny, 2014), and it has many valuable research avenues that require further exploration.

This project contributes to existing research by providing and structuring ideas and reviewing literature on how data mining applications could facilitate auditors in their procedures. Specifically the link to journal entry testing procedures is novel and, because of its importance in a financial statement audit, relevant.

In addition, we suggested the learning effect as a way that previous knowledge on the business could be interesting to enrich decisions without diminishing the effect of professional skepticism. This learning effect enables us and future researchers to let the data tell what is abnormal behavior.
Also, we explored the characteristics of journal entries with data analysis applications. Exploratory data analysis conducted on journal entries has not been applied as a means to detect changing patterns over the years. This research therefore sets an initial step in doing so.

Additionally, accounts have been aggregated to encompassing categories to be able to analyze a more robust unit of analysis over the years. This facilitates in gaining a good insight in transactions over multiple years in an engagement. By looking at journal entries in terms of aggregated categories in the chart of accounts over the years, doors are opened for bucketing based on prior year data, and in an ideal stage a semi-supervised anomaly detection.

Further, in order to concretely find a way data mining could facilitate auditors, we introduced a tool that detects homogeneity within journal entries as a way to improve data-analytical thinking. It is a combination of clustering models to profile homogenous buckets and classification models to detect similar buckets automatically in the subsequent year. By accompanying the tool with guidelines on how the data could be prepared and analyzed, users are able to utilize this research to find buckets in their (manual) journal entries. Research so far did aim to some extent at detecting homogeneity in financial transactions (Argyrou & Andreev, 2011; Thiprungsri, 2012), however it is, to our knowledge, the first model that clusters the behavior of journal entries by means of the debit/credit indicator, the monetary amount transferred, the accounts involved, and the day a journal entry was posted to the general ledger as a way to find homogeneity.

Another contribution to research is the proposition of design principles that can guide practitioners and researchers in their efforts to utilize data-analytical thinking as a way to change the currently intuition based journal entry testing.

The research has been conducted at a CPA. Our legacy to them is more structured insight in the journal entry testing process, and, with that, insight in the inefficiencies that arise. In addition, auditors openness towards data analytics and communication with IT advisors has been mapped. Further, the data analytics research and development team of the CPA found elements mentioned in this research to be very interesting and worth pursuing. And finally, without any previous IT advisory and audit experience we were able to understand the struggle between the two parties but also learn to combine the two mindsets. This must give practitioners hope in their endeavor to bring the two different language together to one generic CPA language.
2. Research Methodology

“Anything that gives us new knowledge gives us an opportunity to be more rational” - Herbert Simon, 1946 -

To answer the research question, a research methodology has been followed which is explained in this Chapter. The research has been conducted at a Case Company, which enabled us to access information that is not accessible by other researchers (Blumberg, Cooper, & Schindler, 2011). Since this research is aimed at proposing how the journal entry testing process ought to be, we introduce the design focused business solving that has been followed (Chapter 2.1). Subsequently, in Chapter 2.2 the research design is explained. Lastly, the fact that both rigorous and relevant research has been conducted is touched upon in Chapter 2.3.

2.1 Design Focused Business Solving

The field of research that focuses on engagement efficiency is emerging, although it lacks clear direction. Due to this scattered approach, it is difficult to find structure in the sweet spot of overlapping research fields regarding data-analytical thinking to increase operational efficiency in engagements. Focusing on journal entry testing enabled us to put the research in context of an existing non-data driven environment. By doing so, research thoughts could immediately be linked to practice.

Since the study required the design of an (IT) artifact that redesigns the unit of analysis, i.e. journal entry testing process, a design science perspective is taken (Hevner, March, Park, & Ram, 2004). The goal of designing an artifact, which would solve the business problem at hand, made us follow the regulative cycle as proposed by Van Aken et al. (2010). The regulative cycle (Figure 6) discusses the steps that should be followed to design an artifact that is specific for a particular business problem. It states that after a problem has been defined in a company, the causes should be analyzed and diagnosed. Subsequently, a plan of action is required to develop a solution. After implementing the solution, and with that intervening in the current way of working, the solution should be evaluated. Since the period of journal entry testing is not simultaneously with our research we were not able to adhere to an actual intervention which diminishes the fourth and fifth step in the regulative cycle, respectively intervention and evaluation.

2.2 Research Design

Since the research field is not well described in scientific literature, and the redesign is dependent on the opinions of auditors in terms of acceptance and effectiveness, it is decided to conduct an exploratory study. Exploratory research suggests the use of qualitative collection methods to gather ideas (Blumberg et al., 2011). The qualitative nature of this research made us utilize interviews as a way to gather and validate information. According to Fontana & Frey (2005), interviews can be subdivided in unstructured interviews, semi-structured interviews, and structured interviews. Unstructured interviews are used to gain insight in the organization, people’s opinions, and solution directions. Semi-structured interviews are used to validate but also to gain insights in decisions made. Structured interviews are used to validate an encountered situation.

The research is built up around three sequential steps guided by the regulative cycle. First the problem is described and validated. As also suggested in literature on business process redesign (Dumas et al., 2013)
this phase is supplemented by an as-is business understanding as a way of diagnosing the current situation. The methods for doing this are discussed in Chapter 2.2.1.

After having identified a problem, solution directions have been explored in terms of data mining being able to facilitate auditors in their way of increasing the engagement efficiency. The methodology followed is discussed in Chapter 2.2.2. The goal of this phase facilitates in reaching the first three objectives:

*Review current literature on applying data mining to journal entries (Chapter 4.2).*

*Provide ideas on how data mining could increase efficiencies in journal entry testing (Chapter 4.2).*

*Develop a tool that sets a subsequent step in this growing field of research (Chapter 5).*

After the solution direction has been chosen, it is decontextualized by linking it to design principles. In Chapter 2.2.3 is explained how this step of the research has been conducted. The goal of this phase is reaching the fourth and fifth objective:

*Explore auditors’ attitudes towards embedding the tool in a redesigned journal entry testing process (Chapter 6).*

*Propose design principles for both researchers as practitioners in using data analytics as an ability to increase efficiencies (Chapter 6).*

When combining the described regulative cycle with the steps that have to be conducted to research our object, a research design can be visualized (Figure 7) and explained in the following paragraphs.

![Figure 7: Research design of this master’s thesis](image-url)

### 2.2.1 Problem Definition and Analysis & Diagnosis

This part follows the first two stages of the regulative cycle which results are discussed throughout Chapter 1, 3, & 4. Unstructured interviews have been held with multiple stakeholders within the Case Company to gain an understanding of the interview’s social reality (Zhang & Wildemuth, 2009) and identify inefficiencies in the current situation (J. Van Aken et al., 2010). In addition two workshops on journal entry testing have been attended. A list of interviewees and topics discussed related to the problem statement can be found in Appendix 1, Table 15. The cause-and-effect diagram that followed from this initial research was validated using four structured interviews (Appendix 2). In this way the inter-
subject agreement, which is shared between at least two different people on the business problem and its causes (J. Van Aken et al., 2010), has been mapped.

Next, in line with the process discovery phase of Dumas et al. (2013), the information of the interviews were combined with a document analysis (Audit Manual, 2015; CAQ, 2008) and reviewed literature as to understand the as-is journal entry testing process (Chapter 3). This triangulation increased our business understanding, whereafter the journal entry testing process has been modeled by an activity diagram. In addition, seven semi-structured interviews have been held to validate the journal entry testing process.

2.2.2 Plan of Action
This part (Chapter 4 & 5) follows the third stage of the Regulative Cycle. A literature review has been conducted to link current research to elements of the journal entry testing process. Semi-structured interviews were held with two researchers in the field of “Data Science” and focused on gather expert opinions as a way to drive explore opportunities on how data mining could facilitate auditors (Appendix 4). A literature review is combined with this analysis to provide ideas for future research, which is discussed in Chapter 4.2.

An important element of our research is designing a tool that increases data-analytical thinking. As discussed by Provost & Fawcett (2013) data mining is a craft, since it involves a substantial amount of (data) science and technology. To get from the initial idea to incorporate data mining in your business to continuously exploit business value, an underlying methodology should be followed allowing reasonable consistency, repeatability, and objectiveness. In this context, data mining is seen from a business perspective in which the CRISP-DM methodology is recommended (Shearer, 2000). The methodology consists of six iterative phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Appendix 9, Figure 48). The initial phase focused on understanding the business and aligning the project objectives and requirements. After a basic idea of the possibilities and goals of applying data mining is obtained it is necessary to understand the data. Data of three different entities have been used to model the tool that facilitates auditors’ procedures. Next, the data can be prepared. Subsequently, it is necessary to have knowledge about the various data mining techniques that can be exploited for developing a tool. Applying the model on three different accounting environments controls to some extent for heterogeneity and strengthens the generic ability of our model to find patterns (Provost & Fawcett, 2013). After selected data mining models have been applied, the results should be evaluated and deployed. To validate the model, quantitative measures have been gathered supplemented by auditor’s opinion of the tool. Six semi-structured interviews have been held to validate the used quantitative measures.

2.2.3 Evaluation by Design Principles
The final part of the research is finding out how the developed tool could increase engagement efficiency. Van Aken’s diagnosis of the management practice as a design science suggests the need for design-oriented knowledge via “grounded and field tested technological rules” (J. E. Van Aken, 2004). These technological rules are often called “Design Principles” (Van Burg, Romme, Gilsing, & Reymen, 2008) or “Design Propositions” (Romme & Endenburg, 2006) in literature. In this research four “Design Principles” are proposed as a way to develop knowledge and design interventions that link a developed data mining tool to engagement efficiency. Design principles follow the CIMO-logic as a way to represent a systematic way of linking science and design (Denyer, Tranfield, & Van Aken, 2008, pp. 395–396): “...
problematic Contexts (C), use this Intervention type (I) to invoke these generative Mechanism(s)(M), to deliver these Outcome(s)(O)”. The methodology used in research should focus on using findings and opinions in practices to drive the creation of design principles (Van Burg, De Jager, Reymen, & Cloodt, 2012; Van Burg et al., 2008). As suggested by Van Burg et al. (2008), design principles developed from practice eventually need a cumulative body of knowledge and practice. As can be seen in Figure 8, when aiming to decontextualize practices to more generic design principles requires theoretical grounding ensuring the most reliable design principles.

Since actual testing the solution is not possible, this part follows a hypothetical questioning of auditors in which the results of the tool have been shared. From every entity two auditors have been interviewed. It is thought that providing auditors with data analysis results on their auditee would facilitate the ability to immerse themselves in a redesigned situation. This has also been noticed during an unstructured interview held before the analysis, in which an auditor stated: “I don’t know what will be the concrete results of this tool” – Auditor (anonymized due to confidentiality).

By semi-structured interviews the mindset and capabilities that the tool produces are linked to practical mindsets. Interviews were favored to all-encompassing cross-CPA surveys in order to gain an exploratory understanding of the ability of the solution direction to increase data-analytic thinking. The output of these interviews are required for discussing the design principles, which made us enclose a more elaborate interview/transcript in Appendix 10. Firstly, a generic interview script was used which started with a number of administrative questions. The subsequent part, was brainstorming on the solution direction and linking its entity-specific results to auditor’s opinion. In this way the auditor is implicitly forced to link the possibilities of the tool to process improvements. In the final part of the interview, explicit mechanisms that increase engagement efficiency were suggested and linked to auditor’s own auditee. This provided a cross-case validation of the developed design principles. The interviews were taped as to handle the data more accurately. Transcribing of these interviews has been conducted word-by-word, after which the qualitative data was selectively coded by means of finding quotes relevant to the questioned object. Selective coding enabled us to validate principles found (Corbin & Strauss, 1990).

2.3 Rigor & Relevance
The setup of the research has a close connection to our goal to gain both rigorous and relevant insights. As discussed by Hevner (2007), design science openly connects these two elements by means of a “Relevance Cycle” and a “Rigor Cycle” (Figure 9). Although we are aware that this research domain has a short legacy, which limits our ability to meet all requirements for rigor, we propose design principles rather than testing them. The developed tool developed provides insights for researchers to increase future knowledge base. By actively working from problems and opportunities found in the Case Company relevance has been gained which enabled us to build and design a tool.
3. As-Is Business Understanding

“To understand where an IT advisor can add value, it is important to understand accountancy.” - Auditor BI, 2016

In this Chapter, the business context is discussed. This leads both to an understanding of the business this research has been conducted in, as an understanding of the current inefficiencies that arise. In Chapter 3.1 the data object under study is, the journal entry, is briefly explained. In Chapter 3.2 journal entry testing, which is an important element of the financial statement audit, is being discussed.

3.1 Journal Entry

Journal entries are the backbone of an accounting infrastructure and record information regarding each financial transaction in a company. Every time a financial transaction occurs, journalizing takes place. Journalizing implies the creation of a journal entry that presents the mutation between accounts in a form that can be captured by the accounting information system (Haber, 2004). Nowadays, most journal entries are created after a transaction has been made by the accounting information system (AIS). The idea of using journal entries has its essence in the double-entry bookkeeping system, which is an embedded system of bookkeeping that allows companies to maintain a complete record of every financial transaction (Carruthers & Espeland, 1991). In practice this system demands of a journal entry that the amount debited equals the amount credited; this “golden rule” leads to a balanced journal (Van Vlimmeren & Fuchs, 2009). Journal entries consist at least of elementary attributes like (an exemplar is shown in Table 1):

- An individual transaction ID – which is identical within one journal entry
- A line item ID
- An account number & description
- Involved currency
- Amount(s) debited (DR) – which should equal the accumulated amount(s) credited within a journal entry
- Amount(s) credited (CR) – which should equal the accumulated amount(s) debited within a journal entry
- The posted date – which is identical within one journal entry
- User – which is identical within one journal entry
- An explanation regarding the reason for initiating the journal entry

Dependent on the accounting information system used by a company, several additional metadata attributes could be attached like more free-text, a profit center, and the number of changes applied.

To specify, a journal entry captures information on the financial transactions between accounts. Examples of accounts are: cash in bank, accounts payable, loans, sales, and salaries. A general taxonomy can be found in the accounting tree in Appendix 12, Figure 50. As discussed by Haber (2004) a journal entry records the transfer of an amount from one account to another account, the journal entries posted are combined in a repository of financial transaction called the general ledger as can be seen in Appendix 5, Figure 40. All information in the general ledger is transferred to a financial statement at the end of an accounting period (2004). Eventually, this financial statement is audited.

The journal entries can generally be classified according to their (non)standard behavior (CAQ, 2008):
• Standard, recurring entries. These support day-to-day activities and are typically automated, posted directly to the general ledger.

• Non-standard entries. These may be recorded during the period-end closing process for purposes of adjusting, accruing, and estimating unusual transactions. These entries are often a result of manual interference.

• Top-side entries. These reflect a manual adjusting entry at a corporate level which generally do not appear as entries to the general ledger. These are most susceptible to fraud by management override, since they are not subject to standard system controls (Lanza & Gilbert, 2007).

In this research the focus is mostly on non-standard entries and top-side entries which are defined as “manual journal entries”. It is generally found that the other standard, recurring entries are controlled by effective IT controls (see Chapter 3.2). In this research IT controls are assumed to capture all controls exploited for labelling “automatic journal entries” as free of management override of controls.

The exemplar in Table 1 represents a journal entry in which suppliers are paid for the received goods. The suppliers have their own account at the company and it is posted to the general ledger by user “HCAMP” on 30-06-2015. Dependent on the IT controls that have been in place at this specific company, this journal entry be labeled as either manually or automatically.

### 3.2 Journal Entry Testing

An important element of an audit is journal entry testing. Although differently organized among CPAs and, due to legislation, between countries, journal entry testing is a well-known method to analyze possible material misstatements (CAQ, 2008). The aim of journal entry testing is gaining insight in the financial process and detect management override of controls. More specifically, the Statement of Auditing Standard (SAS) no. 99, and the international counterpart International Standard on Auditing (ISA) no. 240 requires external auditors to obtain an understanding of the financial reporting process and internal controls over journal entries, identify and select journal entries for testing, determine the timing of the testing, and inquire individuals involved in the financial reporting process (Lanza & Gilbert, 2007).

An audit manual is a CPA’s own translation of legal requirements to audit guidelines within the organization (Audit Manual, 2015). In order to model the journal entry testing process, an audit manual has been followed. The focus of journal entry testing, as a part of auditors’ procedures to mitigate the fraud risk of management override of controls, is identifying and testing the high-risk journal entries.

The choice for analyzing the journal entry testing process in this research is based on the abundance of journal entry data within the Case Company, the willingness to add literature to the limited field of research, and in line with current initiatives at the Case Company.

The journal entry testing process has been analyzed generically, as to not incorporate to many company-specific details. Also, to give an external person a summary of what generally happens a high level of
abstraction is deemed sufficient (Dumas et al., 2013). In Figure 10 an activity diagram of the current journal entry testing process is visualized. It has been modeled in a widely used standard called the “Business Process Model and Notation”, the structure of which is in line with the goal of presenting an easy to use and accessible language for business users (White & Miers, 2008). In the remainder of this Chapter the sub processes are briefly explained.

Risk Assessment Sub Process

Throughout the audit process it is necessary to assess the risks related to the auditee (Appendix 6, Figure 41). The goal of the Risk Assessment phase is understanding the processes underlying the financial transactions and evaluate the degree of controls that mitigate these risks. A better controlled environment means that auditors have to gather less evidence or take into account less high-risk criteria. By understanding the accounting process, auditors aim to assess the risk of material misstatement due to fraud and their effect on the nature and extent of follow-up journal entry testing.

To understand the accounting process of an auditee, the relevant processes for initiation, authorizing, recording, and processing of economic transactions are analyzed by auditors by means of inquiring with employees, analyzing internal documents, and analyzing the accounting information system. After discussions among auditors business understanding is clarified and documented. Subsequently, it is possible to understand and test the design and implementation of both anti-fraud controls and journal entry controls. The goal of control testing during the Risk Assessment phase is to identify the level of control reliance. Are there sufficient controls in place to reduce the risk associated with manual journal entries? One can think of limiting the booking of manual financial transactions to only a certain number of people. Only when a control is thought to be significant enough will it be tested, otherwise this control is not recognized as “effective”.

It is important that auditors constantly apply their knowledge of the industry and professional judgment to develop high-risk criteria that might encompass fraudulent behavior. The process ends with an approach and initial criteria to filter high-risk journal entries. The risk assessment phase is meant to guide the filtering sub process in which the actual high-risk criteria are being specified.
**IT Control Testing Sub Process**

In parallel to auditors’ operations the IT advisors, in the role of IT auditors, analyze the IT controls to be able to control the automated financial processes (Appendix 6, Figure 42). When the processes can be controlled, automatic journal entries can be assured to be free of management override of controls. Despite IT audit being a profession on its own, in terms of journal entry testing it suffices to simplify their operations to this generic level.

**Journal Entry Extraction Sub Process**

In the Journal Entry Extraction phase (Appendix 6, Figure 43), the dataset of journal entries is obtained from the company’s accounting information system. Hereafter a straightforward analysis of the completeness of the journal entry population takes place. By verifying the completeness of the population it is ensured that no information has been withheld. It consists out of rolling forward on the transactions made to see if the results in accounts equal the mutations from transactions (Audit Manual, 2015; Elder et al., 2012). Subsequently, assuming that IT controls are effective, the journal entry dataset can be subdivided into mutually exclusive sets of controlled automated journal entries and manual journal entries that cannot be controlled by the system.

**Filtering Sub Process**

The following sub process is Filtering (Appendix 6, Figure 44). First, the relevant population of journal entries are determined. The relevant population is the set of manual journal entries that could contain high-risk journal entries. After having identified the relevant population, auditors usually investigate accumulated results of journal entries on accounts. After having an idea of the journal entries involved, high-risk criteria have to be selected/developed. This will be done in the layering phase (Appendix 6, Figure 45), where auditors iteratively apply professional judgment to define these high-risk criteria (Appendix 7, Table 19). Further, research on fraud already provides the auditor with a list of standard “business rules” that can be applied, as can be seen in Appendix 7, Table 18. In this iterative process, auditors constantly ask themselves if they understand what transactions have taken place, and if this is in line with their knowledge of the process. If activities occur that are outside their understanding a risk label is attached. All journal entries that fall within the high-risk criteria should be tested integrally, hence the importance of an extensive discussion of auditors to capture only the most high-risk journal entries.

**Testing Sub Process**

In an ideal situation, it is concluded that journal entries are in line with the knowledge of the process. This means that sufficient procedures have been conducted which minimizes the risk on a material misstatement. In other situations all journal entries that match the high-risk criteria should be tested integrally. This process consists of applying professional judgment to find additional evidence to verify the appropriateness of journal entries. Testing (Appendix 6, Figure 46) may consist of inquiring with management and reviewing evidence (Audit Manual, 2015).

If there is an indication of fraud from management override of internal controls, engagement teams should obey to fraud related laws and contact more specialized professionals, namely the forensic accountants (Elder et al., 2012). They investigate possible fraud by, among others, using rigorous accounting skills, and extensive review of documents.
4. Solution Direction

“Decision-making can benefit from even small increases in decision making accuracy based on data analysis”
- Fawcett & Provost, 2013 -

As discussed, journal entry testing could be enhanced by embedding a data-analytical mindset. This Chapter goes into depth on specifying this solution direction. To improve data-analytical thinking, Fawcett & Provost (2013) propose the utilization of concepts under the umbrella term “Data Science”. Data science is an interdisciplinary field focused on analyzing data for the purpose of knowledge discovery in a generalizable way (Dhar, 2012). Its line of thought guides several data-analytic applications that have been around in the auditing field for quite some years (Gray & Debreceny, 2014). As discussed earlier, CAATs have been developed to cope with client’s financial data and thereby facilitate the audit. Established CAATs are, for instance, based on extracting journal entries and filter out red flag entries using predefined rules. This Master’s Thesis explores and utilizes data analytics which differ from applications used so far. One of the emerging fields within the field of financial statement auditing is “Data Mining”. In this Chapter the research field of data mining is discussed in Chapter 4.1, after which it is linked to the research on financial statement auditing (Chapter 4.2).

4.1 Data Mining

In this Chapter, first the concept of data mining is explained in Chapter 4.1.1. Data mining concepts have specific functionalities that provide users with the ability to solve data-rich business problems. In Chapter 4.1.2 the goals that data mining applications could achieve are discussed.

4.1.1 What is Data Mining?

Data mining can be seen as an osmosis of fields like statistics, machine learning, artificial intelligence, data warehousing, and visualization (Kirkos & Manolopoulos, 2004). Data mining has been widely researched and applied since the 1980s in which it aims to discover knowledge in an often large and unstructured set of data (Provost & Fawcett, 2013). Its main quality is its ability to handle unstructured and dispersed data. It analyses this data trying to find patterns that can improve decision-making. Besides the vast amount of techniques to discover business value, reviewing a dataset with the aid of data mining also stimulates a data-analytical way of thinking that enable one to assess whether and how data can improve business performance (Provost & Fawcett, 2013).

The purpose of data mining tools can be split into two separate categories (Provost & Fawcett, 2013). One category is its ability to estimate an unknown value of interest which is called predictive modeling. Examples of predictive applications are classification and regression. Another category is to gain insight in the underlying phenomenon which is called descriptive modeling. Examples of descriptive applications are association learning and clustering. The differences between the two categories are not always strict, since there are also data mining tools that serve both purposes. The choice of the data miner to either focus on prediction or description is important and dependent on the aim of the data miner. It has an impact on the techniques used, requirements of the dataset, and the verification of the outcome.

In addition, the nature of the data mining techniques can be subdivided into two other categories: supervised and unsupervised learners (Provost & Fawcett, 2013). The distinction is important since both types bring about different techniques. Unsupervised analyses requires no specific target as to what the purpose of an analysis is and is often used to discover the data. Clustering and profiling generally are unsupervised data mining applications. Often unsupervised techniques are used for describing rather than
predicting. Classification and regression are generally solved with supervised methods. Supervised analyses are based on predicting an outcome and in order to do that it requires a specific target and a sample set with known outcomes.

Data mining techniques are in an increasing fashion used by auditors to delve deeper in financial statements (Gray & Debreceny, 2014). These involve, among others, more advanced clustering, visualization, anomaly detection, and prediction techniques. By using sophisticated techniques that take into account several factors and a lot of cases, insightful and until now unknown relationships, patterns, anomalies can be discovered (Earley, 2015). In Figure 11 a distinction has been visualized to conceptualize the use of data analytics within audit companies (Gray & Debreceny, 2014). Looking at the relative frequency triangle one can see that data extraction has been given a lot more attention in terms of research and utilization. This leaves ample opportunity for more research on possible data mining applications within auditing.

4.1.2 Data Mining Functionalities

In this Chapter, general data mining functionalities are explained to give an indication of the data mining problems that can be solved. In doing so they are linked to examples involving journal entries to involve the current research area. We distinguish between 5 data mining functionalities (Han, Kamber, & Pei, 2012). Provost & Fawcett (2013) analyze 9 different data mining functionalities, but these can be scoped down to 5 after a careful analysis (see Appendix 8, Figure 47). Note that there remains overlap between these distinctive functionalities so this translation is not exclusive, since Han et al. (2012) categorize less different functionalities, their taxonomy is determined to be leading.

Class Description

Class description is aimed at summarizing the data in concise, yet precise terms. By summarizing the data of the class under study and comparing the target class with a set of comparative classes, an understanding of its nature can be described. Journal entries are constantly being translated to aggregated data to increase business understanding and perform tests on, for instance, the correct representation of balances. In addition, class description techniques aim at comparing the general attributes of cases. The complexity can range from simple data analysis to more sophisticated data mining. CAATs exploit this functionality to a large extent, however the degree of sophistication is very dependent on the quality of the internal audit department and the used accounting information. The functionality is often used in the explorative phase of a data mining project.

Frequent Patterns

Mining frequent patterns might lead to finding undiscovered patterns in the data. These patterns can be subdivided in frequent itemsets, sequential patterns, and structured patterns. A frequent itemset finds, for instance, transactions within a journal entry dataset that often occur together. A sequential pattern is a sequence of transactions that occurs frequently in time. A structured pattern aims at finding substructures that can be combined with itemsets or subsequences. By finding associations and
correlations between several journal entries and/or attributes of journal entries an understanding of the journal entry initiation, authorization, recording, and reporting process could be gained. It can be used to describe and/or predict relations in the population of journal entries. In addition, insights in what events actually influence each other could be gained like a causal relationship between incentives to commit fraud and the journal entry behavior resulting from fraud.

**Classification & Regression**

Data mining applications aimed at classification and regression attempt to predict to which class a specific data point belongs to. This supervised data mining application can be used both for both descriptive as predictive purposes. Developing a set of rules that classifies a journal entry automatically to a specific risk level is interesting for auditors to gain additional insights in the nature of the journal entries. As also discussed by Bay et al. (2006), it can also be used as a detection technique for fraud. In addition, comparing previously classified journal entries to year-after audits could increase engagement efficiency by means of a learning effect (Beck & Wu, 2006). Classification can be used for predicting relatively discrete outcomes, the supervised regression application can be used for continuous variables. The application can be used to estimate a risk score related to a journal entry when incorporating additional internal and external company data (Alles & Gray, 2015; Murphy & Tysiac, 2015).

**Clustering**

The unsupervised and descriptive clustering attempts to group cases in a population together by their similarity, yet not driven by any purpose. While exploring journal entries, clustering is interesting to understand the client’s business processes and identify groups of journal entries that do not fall within a known pattern (Gray & Debreceny, 2014; Thiprungsri & Vasarhelyi, 2011). Isolating a heterogeneous high-risk journal entry population from a homogeneous set of low-risk journal entries could lead to more efficient operations in, for instance, journal entry testing.

**Outlier Analysis**

Outlier analysis attempts to profile the typical behavior of a population after which norms can be established for anomaly detection. By identifying industry standards and/or comparing transactional data over the years anomalous journal entries could be identified that require further investigation. Anomalies in journal entries can fall in three categories (Bay et al., 2006): irregularities due to financial statement fraud, unintentional errors, and unusual entries worth investigating. For instance, by applying process mining on the variables indicating the accounts involved within journal entries, transaction flows can be followed and anomalous transactions can be found (Van Schijndel, 2013). In addition, process mining of journal entries can ensure efficiency, reduce waste, provide assurance for separation of duties, and discover asset misappropriations (M. Jans, Alles, & Vasarhelyi, 2013). Often outlier analysis is conducted in combination with another application (Thiprungsri & Vasarhelyi, 2011). By combining it with clustering, a business and data understanding is gained that eases the anomaly understanding and the ability to detect fraudulent behavior. Overall, fraud or high-risk behavior can be detected by investigating only the suspicious activities. In a different field of research, outlier analysis has been used to profile normal behavior whereafter out-of-the-ordinary activities could be flagged. This line of thought facilitates in finding irregular behavior within auditee’s manual journal entries.
4.2 Data Mining Opportunities in Journal Entry Testing Process

In this Chapter literature and ideas for applying data mining to facilitate decision making are described. In general, three different ways of business decision making approaches could be distinguished (Rupnik & Jaklić, 2009, p. 1):

1. Decision making based on precisely defined business rules.
   - A list of example business rules in journal entry testing can be found in Appendix 7, Table 18.
2. Decision making based on intuition.
   - A list of criteria based solely on professional judgment can be found in Appendix 7, Table 19.
3. Decision making based on the analysis of information.

The first two could be seen as an intuition-based way of decision making, and the third one as a data-driven way of decision making according to the terminology of Provost & Fawcett (2013). Data mining also incorporates the analysis of information for the purpose of facilitating auditors. In Table 2 current research on data mining accounting transactions are linked to implications for journal entry testing. As already said, we focus on the field of research on utilizing data mining techniques as a computer-assisted auditing tool (i.e. efficiency in audit process). Since research so far does not specifically aim at journal entry testing, but only mention that data mining applications could be applied on accounting data, the sub processes of risk assessment and filtering may seem intertwined. In the left column of Table 2 we distinguish between the main steps in journal entry testing, in the second column the taxonomy described in Chapter 4.1.2 is used to identify functionalities of data mining applications, which are explained in the third column. The fourth columns gives a short description on the key findings of each article. More information can be gained by reading the cited article.

In another table (Table 3) ideas for data mining applications within the current journal entry testing process are described. These ideas are based on empirical findings, discussions with researchers at the Case Company, and future research as proposed in literature. The data mining functionalities described in Chapter 4.1.2 were used as a taxonomy and placed in the specific sub steps of journal entry testing as discussed in Chapter 3.2. The idea of identifying ideas for data mining applications are derived from several articles that followed a similar line of thought (Gray & Debreceny, 2014; Q. Liu, 2014; Sirikulvadhana, 2002). These articles, however, describe the entire auditing process, where this analysis has focused solely at journal entry testing as executed by the auditors. The reason for providing thoughts for future research in Table 3 is derived from the lack of research in this area. There are multiple clues in literature that it is beneficial to conduct supplementary research, however there is no research that links the scattered research articles so far to improving the engagement efficiency in journal entry testing. Even Gray & Debreceny (2014), who aimed at structuring the current scattered research on data mining accounting data, do not zoom in or even mention journal entry testing. By doing so, more concrete and operational starting points can be pointed out that eases future research on data mining journal entries.

Activities performed in an IT audit have no formal role within the evidence-based processes conducted by auditors in journal entry testing. That is why only the auditor’s role has been taken into account in this Chapter.
Table 2: Research on the application of data mining in the journal entry testing process

<table>
<thead>
<tr>
<th>Step</th>
<th>Data Mining Functionality</th>
<th>Data Mining Application</th>
<th>Key Research Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk Assessment</td>
<td>Class Descr.</td>
<td>Exploratory Data Analysis (Q. Liu, 2014)</td>
<td>By exploring accounting data, more insights can be gained which facilitates in developing hypothesis on possible data mining applications.</td>
</tr>
<tr>
<td></td>
<td>Class Descr.</td>
<td>Digit Analysis (R. S. Debreceny &amp; Gray, 2010)</td>
<td>By analyzing descriptive statistics in journal entries the existence of fraudulent data could be investigated.</td>
</tr>
<tr>
<td></td>
<td>Frequent Patterns</td>
<td>-</td>
<td>Nothing has been found on frequent patterns so far.</td>
</tr>
<tr>
<td></td>
<td>Class Descr.</td>
<td>EM Algorithm (Thiprungsri, 2012)</td>
<td>By applying clustering methods, like the expectation maximization (EM) algorithm, in an accounting domain exploratory data analysis and anomaly detection could be pursued.</td>
</tr>
<tr>
<td></td>
<td>Clustering</td>
<td>SOM for internal controls (Argyrou &amp; Andreev, 2011)</td>
<td>By applying Self-Organizing-Maps (SOM) on transactions, an accounting database can be clustered in homogeneous clusters that have implications for internal controls.</td>
</tr>
<tr>
<td></td>
<td>Outlier Analysis</td>
<td>Process Mining (M. Jans et al., 2013; Van Schijndel, 2013)</td>
<td>By analyzing process trails, risky accounting transactions can be discovered.</td>
</tr>
<tr>
<td></td>
<td>Classification &amp; Regression</td>
<td>Naïve Bayes Classifier, EM, Logistic Regression, Positive Bayes (Bay et al., 2006) Digit Analysis (R. S. Debreceny &amp; Gray, 2010)</td>
<td>By feeding a dataset with known fraudulent cases and classify these in an early sense in other datasets, fraudulent behavior could be automatically found.</td>
</tr>
<tr>
<td>Filtering</td>
<td>Class Descr.</td>
<td>Exploratory Data Analysis (Q. Liu, 2014)</td>
<td>By exploring accounting data, more insights can be gained which facilitates in developing hypothesis on possible data mining applications.</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>----------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>Class Descr.</td>
<td>Digit Analysis (R. S. Debreceny &amp; Gray, 2010), -</td>
<td>By analyzing descriptive statistics in journal entries the existence of fraudulent data could be investigated.</td>
</tr>
<tr>
<td>Frequent Patterns</td>
<td>Clustering</td>
<td>Demographic Clustering, Neural Clustering (Sirikulvadhana, 2002)</td>
<td>Nothing has been found on frequent patterns so far.</td>
</tr>
<tr>
<td></td>
<td>Class Descr.</td>
<td>EM Algorithm (Thiprungsri, 2012)</td>
<td>By clustering journal entries samples of homogeneous transactions could be identified.</td>
</tr>
<tr>
<td>Clustering</td>
<td>Clustering Outlier Analysis</td>
<td>Process Mining (M. Jans et al., 2013; Van Schijndel, 2013)</td>
<td>By applying clustering methods in an accounting domain, exploratory data analysis and anomaly detection could be pursued.</td>
</tr>
<tr>
<td></td>
<td>Outlier Analysis</td>
<td>Naïve Bayes Classifier &amp; EM, Logistic Regression &amp; Positive Bayes (Bay et al., 2006), Digit Analysis (R. S. Debreceny &amp; Gray, 2010)</td>
<td>By analyzing process trails, risky accounting transactions can be discovered.</td>
</tr>
<tr>
<td></td>
<td>Classification &amp; Regression</td>
<td></td>
<td>By feeding a dataset with known fraudulent cases and classify these in an early sense in other datasets, fraudulent behavior could be automatically found.</td>
</tr>
<tr>
<td>Testing</td>
<td>Class Descr.</td>
<td>Big Data Analysis (Yoon, Hoogduin, &amp; Zhang, 2016)</td>
<td>Journal entries provide only indications for fraud. More elaborate evidence is required to actually detect fraud. Big data audit evidence (Yoon et al., 2016) or text mining emails (Roger S. Debreceny &amp; Gray, 2011) are examples of data mining facilitating auditors.</td>
</tr>
</tbody>
</table>
### Table 3: Ideas for the application of data mining in the journal entry testing process

<table>
<thead>
<tr>
<th>Step</th>
<th>Data Mining Functionality</th>
<th>Idea for Data Mining Application</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Risk Assessment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Understand Manual JE Process</td>
<td>Class Descr.</td>
<td>In journal entry testing, exploring journal entries more extensively using exploratory data analytics could increase the understanding of the process.</td>
</tr>
<tr>
<td></td>
<td>Frequent Patterns</td>
<td>By rolling forward on transactions, monetary transitions can be visualized through the process and anomalous flows can be analyzed on the most granular level. Research so far is still in its infancy (M. Jans et al., 2013; Van Schijndel, 2013).</td>
</tr>
<tr>
<td></td>
<td>Outlier Analysis</td>
<td>If previously known fraudulent journal entry characteristics are incorporated in current years audit, the predicted behavior of journal entries can be put next to the actual behavior to understand the process. In addition, incorporating industry norms in the regression model could facilitate in identifying risky areas.</td>
</tr>
<tr>
<td></td>
<td>Classification &amp; Regression</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clustering</td>
<td>By clustering journal entries, when taking into account multiple attributes, an increased business understanding could be gained by the auditor. By thinking about the nature of these clusters, profiles could be established after which anomalies can be detected.</td>
</tr>
<tr>
<td></td>
<td>Frequent Patterns</td>
<td>One part of the risk assessment phase are the fraud inquiries. Auditors are obliged to identify risks of management override by inquiring with people that are authorized to make transactions. By applying data mining, the most interesting roles to be interviewed could be identified. The outcomes of these inquiries are taken into account for the selection criteria regarding high-risk journal entries.</td>
</tr>
<tr>
<td>Evaluate JE Controls &amp; Evaluate Anti-Fraud Controls</td>
<td>Class Descr. Frequent Patterns Outlier Analysis</td>
<td>Unsupervised descriptive analytics can familiarize the auditor with recurring and anomalous controls (M. Jans, Lybaert, &amp; Vanhoof, 2009).</td>
</tr>
<tr>
<td></td>
<td>Classification &amp; Clustering</td>
<td>If previously known journal entry controls are incorporated in current years audit, the predicted existence of controls can be put next to the actual controls to evaluate the risk and control areas.</td>
</tr>
<tr>
<td>Step</td>
<td>Data Mining Functionality</td>
<td>Idea for Data Mining Application</td>
</tr>
<tr>
<td>------</td>
<td>--------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td></td>
<td>Class Description Frequent Patterns</td>
<td>By comparing industry norms and previously designed and implemented controls, risk areas can be identified. Once a new pattern is found, it could be incorporated as a novel “business rule” that enables auditors to easily detect high-risk behavior.</td>
</tr>
<tr>
<td>Determine High-risk Criteria</td>
<td>Class Description Classification &amp; Clustering</td>
<td>In journal entry testing, exploring journal entries more extensively using exploratory data analytics could increase the understanding of the process. Data mining might be helpful in identifying similar journal entries, outliers, and predicting entity-specific high-risk criteria based on previous years. In addition, the amount of audit evidence needed for determining the completeness and accuracy of journal entries can be predicted by means of incorporating both internal financial indicators as external data. Benchmark transactions between companies in the same industry to identify regular and irregular journal entries. These could increase both the business understanding as the chance of detecting meaningful outliers.</td>
</tr>
<tr>
<td>Filtering</td>
<td>Class Description Classification &amp; Regression</td>
<td>In terms of decision theory, it might be interesting to extend this decision to already filter out more low-risk journal entries. This lowers the time needed in the subsequent layering process. Classifying certain low-risk journal entries based on previous known low-risk and standard journal entries could be worth pursuing.</td>
</tr>
<tr>
<td>Determine &amp; Identify Relevant Population</td>
<td>Class Description Classification &amp; Regression</td>
<td>Auditors apply professional judgment instead of exploring the data for anomalous behavior. The ability to describe the data to understand the most recurring behavior could lead to increased business understanding, and increases the ability to set effective high-risk criteria. Ranking outliers by their relative exceptionality could decrease the number of false positives and provide some evidence on the most exceptional cases. This thought is inspired by Issa (2013), who incorporates a different dataset than journal entries.</td>
</tr>
<tr>
<td>Step</td>
<td>Data Mining Functionality</td>
<td>Idea for Data Mining Application</td>
</tr>
<tr>
<td>------</td>
<td>---------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td></td>
<td>Classification &amp; Regression</td>
<td>Feeding a dataset with anomalous data, whereafter classification learns the ability to find anomalous could facilitate the auditor (Bay et al., 2006). In parallel with a data mining technique, auditors should go through their own process and decide on risky areas. If journal entries found by the auditors match the journal entries found by advanced analytics high confidence can be placed in their professional judgment. If anomalies are to be found, a high level of interest could be placed in pursuing these results. Note that this direction is only feasible if high-risk journal entries that actually were worth investigating, e.g. fraudulent, could be used to train a dataset. If the model would be learned on high-risk journal entries that turned out to be non-fraudulent, it would be equal to “modeling” auditors previous decision process which would provide inefficient results.</td>
</tr>
<tr>
<td></td>
<td>Clustering</td>
<td>For the layering process the goal is detecting high-risk journal entries from a dataset of manual journal entries. The idea of the layering process is detecting anomalous data that are worth investigating. Jiawei, Kamer, &amp; Pei (2012, pp. 543) state that “clustering finds the majority patterns in a data set and organizes the data accordingly, whereas outlier detection tries to capture those exceptional cases that deviate substantially from the majority patterns” (2012, p. 543).</td>
</tr>
<tr>
<td></td>
<td>Outlier Analysis</td>
<td>In general, the introduction of the element of unpredictability during an audit is very subjective. Applying an element of unpredictability using a data mining application, that might even be an element of unpredictability for the auditors, can increase the quality of an audit. If the findings are taken into account over the years as not interesting anymore, some form of learning could arise.</td>
</tr>
<tr>
<td></td>
<td>Multiple functionalities</td>
<td>Finding and classifying journal entries similar to found high-risk journal entries (supervised) can facilitate in the validation/extension of the current list.</td>
</tr>
</tbody>
</table>

| Document JEs | Clustering Classification & Regression | Finding and classifying journal entries similar to found high-risk journal entries (supervised) can facilitate in the validation/extension of the current list. |

| Testing | Class Description Frequent Patterns | Being able to incorporate Big Data and describe data and patterns within and outside the company could facilitate the search for audit evidence (Murphy & Tysiac, 2015). In addition, text mining could facilitate in linking high-risk journal entries to letters/minutes within the company (Roger S. Debreceny & Gray, 2011) or vice versa in which free text in journal entries are mined for its peculiarity. |
|         | Clustering                 | Linking the found high-risk journal entry to a cluster of similar journal entry might improve the understanding. |
5. Developing a Tool for Automatic Bucketing

“A tool that could divide manual journal entries in buckets would facilitate the filtering process”

- Researcher RII, 2016 -

One of the data mining opportunities has been developed further into a tool that utilizes data mining applications to increase engagement efficiency. The scope of this particular application is determined after we zoomed in on the filtering sub process. In the filtering sub process the relevant population is subject to criteria for the cause of drilling down towards high-risk journal entries. We chose this sub process because it is dependent on data, yet immature in terms of utilizing data analytics. In addition, it is in line with strategic data analytics initiatives within the Case Company. The tool developed facilitates auditors by “Automatic Bucketing” of journal entries, in which buckets of recurring manual journal entries are profiled and automatically detected in a new unlabeled dataset. In this way, buckets of homogeneous subsets of manual journal entries open possibilities for a different and possibly more efficient audit approach. The tool utilizes a combination of “Clustering” and “Classification” applications to detect regular behavior or “Frequent Patterns”. A conceptual presentation of the tool can be seen in Figure 12.

In this Chapter the tool that is developed to solve the business problem is explained. In Chapter 5.1 the foundation of automatic bucketing is explained. Subsequently the data utilized for testing the model is explored and prepared, which is discussed in Chapter 5.2. In Chapter 5.3 the tool that enables automatic bucketing is discussed and tested. In Chapter 5.4 we focus on evaluating the tool whereafter in Chapter 5.5 a hypothetical intervention is discussed.

**Figure 12: Conceptual representation of automatic bucketing functionality**

5.1 The Foundation of Automatic Bucketing

The initial phase of developing the tool focuses on understanding the business. As already a lot is discussed on accounting and auditing, we would rather focus on the mind shift that is required for this tool to be relevant. One needs to understand the data-analytical viewpoint that is required as a means to apply automatic bucketing (Chapter 5.1.1). After this has been discussed, the business problem is linked to a problem suitable for data mining (Chapter 5.1.2).
5.1.1 Data-Analytical Viewpoint

In automatic bucketing a change in viewing the dataset of journal entries is required. Current regulations and company guidelines state the procedures that should be conducted when journal entry testing. After following these guidelines auditors find the journal entries that need to be tested elaborately. The underlying goal is not solely following up on these guidelines. For this research, as an independent researcher, current regulations and company guidelines are set aside to find out that the actual purpose of journal entry testing is explaining and understanding manual journal entries that happen differently from “non-fraudulent” and low-risk behavior. At the moment, trying to link professional judgment and known business rules to the entire dataset of manual journal entries are thought to be effective ways of doing so. However, the data itself could provide information on what is anomalous, i.e. different from auditor’s understanding. This viewpoint made us incorporate previous year’s manual journal entries, which were tested to be free of management override of control in that year, as a dataset that captures “normal” behavior. In this way patterns of previous years could be detected in current year’s audit which enables the detection of recurring patterns. This would imply that there are homogeneous buckets of manual journal entries in a presumed heterogeneous dataset. In the long run an ideal situation could arise in which only the manual journal entries that do not act in accordance with known behavior are labeled as possibly risky, hence decreases the relevant population for journal entry testing.

In summary, the main data-analytical viewpoint underlying this research is: In order to know what is anomalous behavior, it is necessary to understand what can be profiled as normal behavior. This semi-supervised anomaly detection requires an understanding of “normal instances” (Chandola, Banerjee, & Kumar, 2009). As conceptualized in Figure 13, this would lead to a situation in which only the protrusions are analyzed in light of journal entry testing and overlapping field have a risk behavior that would be similar to historical transactions.

5.1.2 Data Mining Problem

By bringing together data mining applications and business understanding we were able to translate the aforementioned situation to a data mining problem. The journal entry testing procedures of three entities (A, B, & C) have been analyzed using documented decisions and interviews with auditors. Two years of data per entity has been gathered in the form of historic data (2014) and current data (2015), which enabled us to pioneer in the mindset of incorporating previous year’s patterns.

Regarding journal entry testing, it is required that the auditors within the Case Company try to develop different criteria than previous years in order to stay professionally skeptic every year (Elder et al., 2012). In the current way of working a tunnel vision might be avoided and professional skepticism is triggered every year, but what if essential information is re-occurring over the years? Is simply ‘forgetting’ previous data a good thing to do if it was clearly a significant risk in previous years? Our objective in developing a data mining tool is threefold:
1. Provide evidence that it is possible to find and profile routine behavior in a given year.
2. Provide evidence that it is possible to detect that particular routine behavior in the subsequent year.
3. Link findings to practical data-analytical implications:
   a. If regular routine behavior is detected in 2015, do auditors think this will facilitate their procedures?
   b. If irregular routine behavior is detected in 2015, do auditors think this will provide useful insights?
   c. If routine behavior does not occur in 2015, do auditors think that this will provide useful insights?

Routine behavior is defined as: “Journal entries that have a recurring pattern over a specific amount of time. This could be weekly, monthly, or even annually”. For routine behavior to be profiled and detected, we need to model the behavior of journal entries. This developed concept captures three important elements of a journal entry’s behavior which we think is able to model the behavior of routinely occurring journal entries. The first element is the accounts that are involved. The accounts are connected to the chart of accounts, which is essential to translate the information of journal entries to a specific company context (R. S. Debreceny & Gray, 2010). It is thought that routinely occurring journal entries would involve similar accounts over time. The second element is the amount of money that has been transferred between accounts. It is expected that routinely occurring journal entries would involve similar monetary values over time. A third element interesting to find frequent patterns, is involving either the debit or the credit behavior of the transaction. It is expected that routinely occurring journal entries have a similar debit and credit nature over time.

Subsequently, the data should be prepared as conceptually presented in Table 4. If journal entries are presented as data points and accounts involved as attributes it is possible to determine what monetary amount has been debited (+) or credited (-). As discussed before, credited amounts should be equal to debited amounts which provides us with a sum of amounts equal to 0. Besides these attributes, it is possible to add additional attributes to distinguish journal entries. For example, the day in the month a journal entry has been posted to the general ledger could be added. It is thought that journal entries that are posted regularly, have been posted around the same date or day in the month. This attribute is also taken into account as an extra line of thought when modeling. Adding other attributes can also be done, but are ideas for future research. Chapter 5.2 goes into depth on how this setup is practically approached within the entities involved. A more elaborate conceptual data setup is discussed in Chapter 5.3.1.

Table 4: Conceptual data setup to find routine behavior in journal entries

<table>
<thead>
<tr>
<th>Journal Entry</th>
<th>Liability</th>
<th>Asset</th>
<th>Expense</th>
<th>Revenue</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>JE1</td>
<td>0</td>
<td>+100</td>
<td>0</td>
<td>-100</td>
<td>0</td>
</tr>
<tr>
<td>JE2</td>
<td>-200</td>
<td>0</td>
<td>0</td>
<td>+200</td>
<td>0</td>
</tr>
<tr>
<td>JE3</td>
<td>-400</td>
<td>0</td>
<td>+400</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

5.2 Data Understanding & Data Preparation

The data required to test the tool have been explored and prepared for modeling, which is discussed in this Chapter. First, the entities that have been selected are discussed in Chapter 5.2.1. Next, assumptions that have been taken into account in this research are discussed in Chapter 5.2.2. Since the tool requires preparing the data in a particular form, data preparation is discussed in Chapter 5.2.3. Finally, in Chapter 5.2.4, the data has been explored in order for researchers to understand the data used and for auditors to familiarize with data-analytical insights.
5.2.1 Entity Selection

Clustering of monetary transactions provides us with the tools to find homogeneous groups in a believed heterogeneous dataset of manual journal entries. For this to happen, data has to be collected. We approached auditors within the Case Company that were actively involved in journal entry testing of their engagement. This resulted in the collection of data from three different entities\(^1\) which can be seen as representative for the process of journal entry testing. Applying the model on three different accounting environments controls to some extent for heterogeneity and strengthens the generic ability of our model to find patterns (Provost & Fawcett, 2013). The data collected existed out of:

1. Manual journal entries that were the input for the filtering process, i.e. the relevant population.
   - Two subsequent years have been chosen which have been controlled to be free of management override in the years 2014 and 2015 respectively.
2. Documented decisions during the layering process, i.e. high-risk criteria.
3. Chart of accounts of the entity to understand the behavior of journal entries.

In Table 5, some characteristics of the engagements with the entities are described. It is interesting to see that the number of high-risk labeled journal entries differ over the years. This has to do with a change in regulation at the Case Company. In 2014 auditors set non-specific high-risk criteria that increases the work to be conducted at the testing phase. By setting more specific high-risk criteria the layering phase requires more attention, after which it reduces the amount of work to be conducted during the testing phase. Journal entry testing usually requires two/three people that discuss the audit approach. Next, one/two less experienced auditors apply the high-risk criteria that arise from the discussions after which the other more experienced auditors review the decisions made. The time in manhours for journal entry testing is not documented, which made us rely on auditor’s approximations. The duration of journal entry testing is related to the size of the relevant population manual journal entries and the specificity of high-risk criteria. A short description of the nature of the entities are explored in the subsequent paragraphs.

<table>
<thead>
<tr>
<th>Entity</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry</td>
<td>Electronics Research</td>
<td>Food Production</td>
<td>Waste Service</td>
</tr>
<tr>
<td>Duration of Engagement</td>
<td>7 years</td>
<td>2 years</td>
<td>2 years</td>
</tr>
<tr>
<td>Manual JEs</td>
<td>7,055</td>
<td>8,320</td>
<td>10,904</td>
</tr>
<tr>
<td>High-risk JEs</td>
<td>1,163</td>
<td>370</td>
<td>3476</td>
</tr>
<tr>
<td>Auditors</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Total Manhours Journal Entry Testing</td>
<td>12-18 hours (raw approx.)</td>
<td>12-18 hours (raw approx.)</td>
<td>12-16 hours (raw approx.)</td>
</tr>
</tbody>
</table>

\(^1\) Note that, when referring to an entity in this thesis, the handling and testing of a particular set of manual journal entries is meant. There is not direct link to the entities involved, since it is the journal entry testing process of the Case Company that is subject to research and not the governance of journal entries within the accounting information system of an entity.
5.2.1.1 Entity A
Entity A is active in the electronics research industry and is an auditee for multiple years. Being a “loyal client”, it is very important that auditors deliberately practice professional skepticism to avoid a tunnel vision. The high-risk criteria that were developed in 2014 and 2015 are described in Appendix 13, as well as the dataset of journal entries and their related attributes.

5.2.1.2 Entity B
Entity B is active in the food production industry and is a second-year’s audit. A second-year’s audit is characterized by multiple inquiries at the entity’s location that is required to gain a business understanding. The high-risk criteria that were developed 2014 and 2015 are described in Appendix 14, as well as the dataset of journal entries and their related attributes.

5.2.1.3 Entity C
The third entity, in the waste services industry, is an example of a more efficient data-driven audit. After a survey has been filled in by auditors, a data analytics team analyzes the data and finds the things mentioned to be interesting. It is a pilot within the Case Company to see its effectiveness, but in terms of looking at journal entry testing data-analytically it is a great leap forward. It has drastically increased the number of journal entries that could be analyzed, however, it does not necessarily increase the auditor’s understanding of recurring behavior. A tool like automatic bucketing could still yield interesting results. The high-risk criteria that were developed in 2014 and 2015 are described in Appendix 15, as well as the dataset of journal entries and their related attributes. Different from the previous two entities, there is no distinction made in “automatic” and “manual” journal entries for the purpose of journal entry testing, which increases the dataset, yet for the purpose of finding regular behavior no significant impact occurs.

5.2.2 Assumptions
In order to model the data for bucketing, some assumptions have been made. These are discussed with involved auditors to be legitimate.

- The dataset of entities B & C were not complete and contained double-date journal entries. The double-date population has an integrally different behavior than other journal entries in that they incorporate multiple posted dates in one single journal entry. This behavior is not in line with the characteristic behavior of journal entries and are excluded from this research. Additionally, due to internal decisions, which cannot be discussed due to of confidentiality, only a subset of journal entries are provided. Since our goals is to find some, and not all, routine behavior some freedom in developing a prepared dataset is taken. This assumption indicates that we are aware that not all data under analysis is complete. Decisions made are documented Appendix 16 for future researchers.

- Journal entry testing can be subdivided in three subtests, in which we only focus on the first one since this requires a filtering procedure to come up with the most high-risk journal entries:
  1. Test high-risk journal entries made throughout the period.
  2. Test high-risk journal entries and other adjustments made at the end of the reporting period (this only costs a limited amount of work in the testing sub process)
  3. Examine material post-closing entries made during the financial statement closing process (this only costs a limited amount of work in the testing sub process).

- Rupnik & Jaklić (2009) stated that data mining, as a way to support data-analytical decision making, can be integrated in operational business processes. Creating the model is generally not the end of a project, often one wants to implement the followed steps in a report to facilitate future use (Wegener
Developing buckets is an iterative process in which data mining constantly needs to adapt to the business it is applied to. Since the deployment and business understanding phases in CRISP-DM are very much connected (Rupnik & Jaklič, 2009; Wegener & Rüping, 2010), automatic bucketing requires adaptability and flexibility. The focus is therefore on providing guidelines as to how automatic bucketing could be executed, rather than proving that this tool works generically.

5.2.3 Data Preparation
Firstly, the data has been prepared which facilitates the understanding of the data that are discussed later. Two different types of data have to be prepared. Initially, the chart of accounts of the entities was prepared for modeling. We encountered a situation where account numbers change over the years, accounts are differently impacted by manual journal entries, and accounts are only of temporarily nature. This highly fluctuating behavior limits the ability to analyze account-specific data over the years. In Appendix 12, Figure 50 an exemplar chart of accounts with its child-parent character has been shown (Argyrou & Andreev, 2011). There are several levels on which the aggregated accounts involved can be seen. It is thought that the accounts on a more aggregated category, still have the same generic behavior. An entity’s chart of accounts facilitates in distinguishing between the character of journal entries, in which a Level 1 has been taken into account in Table 4. Aggregating to a Level 3 in this chart captures the specific characters of accounts inside the category. In this way we dealt with the fluctuating nature of accounts without leaving out essential information. In addition, it also reduced the problems that could arise with sparsity in a dataset (Han et al., 2012). In Appendix 12, Figures 51, 52, 53, 54, 55, & 56 the chart of accounts of the specific entities can be found. It is shown until the level 3, i.e. category-level, for confidentiality reasons. After the categories are prepared, the dataset of manual journal entries have been transposed to a dataset as conceptualized in Table 4. Guidelines for doings so can be found in Appendix 16.

5.2.4 Exploratory Data Analysis
After preparing the data, the data has been explored. As discussed by Liu (2014) exploratory data understanding facilitates auditors in detecting risky areas. Since auditors currently do not explore the data before making decisions on high-risk criteria, visualizations and other exploratory data analytics are already a practical implication that enhances journal entry testing. We have decided to combine the data understanding phase of CRISP-DM with an exploratory analysis of the entities to increase auditor’s understanding of the client’s accounting transactions.

This Chapter is solely for understanding the data at a generic level, “it will take deeper data mining to isolate unusual patterns and transactions” (R. S. Debreceny & Gray, 2010, p. 177). Since three entities have been explored and used to build a generic data mining tool, showing all analyses would be overwhelming. We decided to explore the data of the inter-category transactions of one entity (Entity A), however the other visualizations can be found in Appendix 17 & 18. As can be seen in Figure 14, the inter-category transactions (A2) contain both the homogenous sub-population we would like to extract in the profiling phase and the heterogeneous sub-population which is a smaller subset of the initial heterogeneous population. The inter-category transactions are the input for the modeling phase that is discussed in Chapter 5.3. The intra-category transactions have a different risk behavior which will be discussed in Chapter 5.5.1. The double-date population is out of scope in this research. Some preliminary information on the amount of journal entries in the sub-populations can be found in Table 6 & 7.

In summary, although more sub-populations can be found in the relevant population, only the A2 inter-category journal entries are touched upon in the exploratory phase.
Table 6: Information on subdivided journal entries in entity A, 2014

<table>
<thead>
<tr>
<th>Dataset 2014</th>
<th>No. of Transactions</th>
<th>No. of JE Lines</th>
<th>No. of High-risks</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>429</td>
<td>7,045</td>
<td>1,163</td>
</tr>
<tr>
<td>A1</td>
<td>104</td>
<td>588</td>
<td>130</td>
</tr>
<tr>
<td>A2</td>
<td>325</td>
<td>6,457</td>
<td>1,033</td>
</tr>
<tr>
<td>A3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 7: Information on subdivided journal entries in entity A, 2015

<table>
<thead>
<tr>
<th>Dataset 2015</th>
<th>No. of Transactions</th>
<th>No. of JE Lines</th>
<th>No. of High-risks</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>470</td>
<td>8,320</td>
<td>370</td>
</tr>
<tr>
<td>A1</td>
<td>122</td>
<td>841</td>
<td>136</td>
</tr>
<tr>
<td>A2</td>
<td>348</td>
<td>7,479</td>
<td>234</td>
</tr>
<tr>
<td>A3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

5.2.4.1 Visualization of Descriptive Statistics

Descriptive statistics is one way of analyzing a dataset. Debreceny & Gray (2010) propose, for instance, to analyse the number of lines or journal entries that have been posted. We include some visualizations of multiple years, since our idea is that comparing journal entries over the years lead to interesting insights. For confidentiality reasons, no specific amounts are mentioned. In Figure 15 one can see that the number of postings are quite similarly over time, were the slightly larger dataset of 2015 mostly differs throughout the first 11 months.

In Figure 16 the total number of manual bookings in the accounts of entity A could be seen. Positive differences over the years can have multiple reasons. These could be, among others, growth of the company, or new processes that required manual interference of company’s accountants. Negative
differences could arise when a company’s accounting information system is better controlled. The differences over the years might be interesting to auditors, however this is dependent on the risk level of the account and the monetary size of the impact.

In Figure 17, the amounts posted over times has been visualized. The height indicates the amount involved, the + or – place in the graph indicates whether a booking was debited (+) or credited (-). It is interesting to see that in entity A, end-of-the-month transactions arise. The seemingly linear accumulation of journal entry is something worth looking further into. An auditor of this case indicates that: “This end-of-the-month booking is specific for this entity” – Auditor A

However this specific visualization was seen as very interesting by auditors: “It is interesting to see this visualized” – Auditor A

Further, the accounts in combination with the categories involved is of interest. It can be seen that certain categories are utilized with higher accumulated amounts (Figure 18). Initially, it is detectable that some liability accounts have a large accumulated amount debited on Category 8 and that some revenue accounts have a large accumulated amount credited on Category 32. To understand this behavior, some knowledge on accounting is needed. Frankly, this knowledge can be captured in a figure as shown in Appendix 11, Figure 49. In order for revenue to increase, it has to be credited, so Figure 18 does not show unusual behavior in that sense. In the same line of thought, liabilities should also be credited for them to increase. These liabilities are debited, hence the company’s debts has decreased over the years by means of manual financial transactions. It requires understanding of the business to recognize changes over the years as being worth taking into account in the high-risk criteria.
As an explorative analysis of the data for audit purposes, Liu (2014) suggests a heatmap. A heatmap is a two-dimensional representation of data, in which different colors can represent data values. As an example we have used the day in the week and the categories that are involved. In this way the number of involved categories on a specific posting day in the week could be counted and divided by the total amount of transactions. In Figure 19 the heatmaps of dataset A2 in the years 2014 and 2015 are visualized. Only overlapping categories are shown to facilitate comparison. The more reddish a cell is, the less transactions have taken place, while the more greenish a transactions is the more transactions have taken place. In this way it can easily be shown which transactions have taken place and which accounts are the most frequently involved. It is interesting to see that category 8 is touched upon quite frequently throughout the week. In addition, it can be seen that apparently the Sunday has fewer postings in 2015. This might indicate a policy change within the entity’s accounting system. Again it is dependent on the ability of an auditor to explain this change in behavior in which degree this might represent risky behavior.
5.2.4.3 Association Rules

Association rules provide some ideas on the transactions that occur. Providing information on these for audit purposes is suggested by Liu (2014). Association rules involve the categories that are involved in journal entries as a way to see what often occurs together. It requires the dataset to be ordered with a binary indicator on whether accounts are booked within the journal entry. The value “0” (no) “1” (yes) determines whether a category is involved in a journal entry. By looking at bookings that involved similar accounts, these can be isolated as frequent itemsets. A frequent itemset is a collection of items that occur frequently together. Next, association rules could be used to determine that when one transaction occurs, other transactions tend to occur as well. The operators used in RapidMiner can be seen in Appendix 19. In this example a minimum confidence of 0.8 is set, which means that only the transactions have been shown that have a certainty of 80% of occurring. The minimum support is set to 0.2, which means that we are interested in all transactions that occur at least 20% of the time. After doing so, association rules can be provided on the categories that happen together in transactions (Han et al., 2012). It can be seen that when Category 27 is touched upon, there is a high confidence that the transaction involves Category 19. These two categories co-occur in 15.4% of the transactions. Following up on this, in the case Category 27 occurs Category 19 is booked with 94.3% confidence which implies a relation between these two categories. By comparing these over the years it can be seen that, in this case, similar association rules arise. This provides an idea that the most frequently occurring manual transactions co-occur similarly over the years and that there is not direct implication for an “transaction-shift”. It gives different information than heatmaps, since it incorporates combinations of categories rather than single occurrences.

![Figure 20: Association rules in dataset A2 2014](image1)

![Figure 21: Association rules in dataset A2 2015](image2)

5.2.4.4 Visual Similarity Matrix

A final analysis selected for this research is the visualization of similarities (VOS). Visualization of similarities (VOS) is capable of visualizing categories and the closeness to each other (Van Eck & Waltman, 2007). Objects that have a high similarity are located close to each other and objects that are less similar are located further from each other. The similarity is defined as the counts a category occurs simultaneously with other categories in a journal entry. The advantage of this low-dimensional visualization is the ability to compare apparent similarities over the years. This eases the identification of newly added accounts or changes in behavior. Where association rules are able to give an idea on the amount of times transactions occur and which categories often occur together, VOS is able to visualize categories and their similarity.
As can be seen in Figure 22 and 23 category 28 (expense category) and 19 (liability category) have a different behavior than most other accounts. Their transactions are most similar to each other. Firstly, it is interesting to know that these categories have similar behavior which indicates them handling similar transactions. Secondly, changes over the years could be easily detectable due to the intuitive visualization.

Further, it is noticeable that category 4 (asset category) and category 18 (expense category) have become more similar over the year. The categories are used for specific transactions in between the two categories in a lesser extent. Dependent on auditor’s knowledge on the particular transaction, the category is worth investigating further by auditors.

5.3 A Tool that enables Automatic Bucketing

After the data has been analyzed and prepared, it is possible to model and test the tool. This phase consists out of two subsequent parts. The first part is called “Profiling”; it exists out of finding clusters of journal entries that capture similar routine behavior and profile them according to their accounting nature (Chapter 5.4.1). The second part is called “Detection”; it consists out of finding a similar profile in an unlabeled dataset (Chapter 5.4.2). After this profile has been detected, the manual journal entries can be linked to a bucket dependent on their risk level. In this way homogeneous buckets are automatically detected whereafter the nature of the bucket opens up the ability to be treated differently than other (heterogeneous) manual journal entries.
5.3.1 Profiling of Routine Behavior

At first, previous data are profiled on their behavior. Profiling is a term used to characterize typical behavior of a group or population (Provost & Fawcett, 2013). By profiling clusters of journal entries, recurring behavior has been found that increases the business understanding of auditors. Knowing that, for instance, a well-known and low-risk “equipment depreciation”-transaction occurs on a monthly basis initiates different less intensive audit procedures than a more risky single debited transaction on a revenue account. In this Chapter, several techniques have been applied to gather evidence for modeling routine behavior of journal entries as described in Figure 24. This figure is an elaborate presentation of the one presented in Table 4. As can be seen, the attributes in the dataset are the categories involved. The first three journal entries have similar behavior, which might indicate a bucket of routine behavior. As discussed in Chapter 5.1.2, similar behavior is modeled as journal entries involving similar categories, height of amounts, and whether it was credited or debited. Additionally an attribute “GLDay” has been involved in one part of the research. As can be seen in Figure 24, an additional dataset will arise, namely the intra-category bookings. These will be discussed in Chapter 5.5.1.

![Figure 24: Conceptual representation of automatic bucketing in a fictitious dataset](image)

For the analysis an experimental design is followed (Figure 25). For each entity two different paths are followed. The prepared data either involves all overlapping categories (cat) or all overlapping categories plus the day in the month a journal entry is posted to the general ledger (day). It is thought that the attributes involved have the ability to “capture” distinct behavior. Each dataset was then subject to four different clustering techniques that label the raw dataset (raw) with cluster labels. By looking at the raw data, the profiles of the clusters have been determined. This determination existed out of analyzing the journal entries involved and focus on the “explanation” attribute, which was already shown in Table 1. The explanation of a journal entry has the ability to describe the nature of a cluster, since it provides the context and the reason a journal entry has been initiated. Subsequently, a performance metric called the F-measure has been used to determine which clustering technique is favored. This research solely focused on profiling some of all possible routine behavior, as a way to provide evidence that homogeneous journal entries can be detected in a dataset of journal entries. This made us choose four of the developed clusters as profiles based on their ability to describe certain routine behavior.
5.3.1.1 Modeling Techniques

Before the modeling can take place, the impact of the “GLDay” should be made noticeable in the particular dataset (day). We do so by normalizing the data. A technique that allows for this is called “min-max normalization” (Han et al., 2012). It takes the minimum (min\(_A\)) and the maximum (max\(_A\)) value of an attribute and linearly maps all values in between these two values within the range of [0,1]. The following equation is used to compute a new value (\(v'_i\)) for every data point in the current dataset (\(v_i\)):

\[
v'_i = \frac{v_i - \text{min}_A}{\text{max}_A - \text{min}_A}
\]  

(1)

The advantage of this technique is that it preserves the relationships in the data, however it might have problems with an out-of-bound value entering the dataset (Han et al., 2012). It resulted in one dataset that has absolute values for all categories involved in both years’ inter-category transactions (cat), and one dataset that has normalized values for all categories involved in both years’ inter-category transactions and the relative day in the month a value appears (day).

For the “Clustering” part of this research, four different clustering techniques have been used. As discussed, clustering aims to find natural groups in data points, in which data points in one cluster are similar to each other, but dissimilar to data points in other clusters. Clustering techniques can be classified in partitioning methods, hierarchical methods, density-based methods, and grid-based methods (Han et al., 2012). It is chosen to utilize one partition method and one density-based method.

\textit{k-means}

A partition clustering technique partitions data points in non-overlapping clusters in which data within the cluster has small distances in-between each other, and larger distances to data points outside the cluster (Han et al., 2012). One technique suitable for this is \textit{k-means}. \textit{k}-means is a popular and commonly used partition method which is thought to be an interesting stepping stone in the field of automatic bucketing.

The similarity between data points is measured by means of a distance measure. Finding similarity by distance measures is interesting in our dataset since they represent numerical data points in a multi-dimensional area. We are interested in several distance measures since they all might predict different
profiles. Three distance measures have been chosen: Euclidean distance, Manhattan distance, and Cosine distance/similarity (Figure 26). The distance measures used to determine clusters could have an impact on the sort of clusters, which is interesting to find different profiles. Using only the often used Euclidean distance measure might lead to incomplete or missed profiles since it is not always good in handing sparse data (N. Sharma, Bajpai, & Litoriya, 2012). It is thought that Manhattan distance could provide slightly different clusters than the Euclidean distance. We think that the possible errors with the “Curse of Dimensionality” within sparse data could be reduced by incorporating the cosine distance measure (Han et al., 2012). Although it might provide interesting clusters not detected by Euclidean or Manhattan distance, the overall results can be disappointing since it is more appropriate for very high-dimensional data points such as text documents (Tan, Steinbach, & Kumar, 2013). Note that RapidMiner toolbox utilizes the “cosine similarity”. Cosine similarity transforms the base value of cosine distance in a normalized similarity metric (Owen, Anil, Dunning, & Friedman, 2011).

Additionally, k-means requires a proper k value. By visually analyzing the results (see Appendix 22, 23, & 24), we determined k=5 to be sufficiently meaningful. It presented us with transactions involving periodicity which is the element under study. The goal is to find some routine behavior and not all routine behavior, which made us not do a sensitivity analysis on the best k value. The k-means model ran 10 times to account for the randomness of initialization, and it iterated for a maximum number of 100 times for one run of k-means. These are the standard settings in RapidMiner and no errors in relations have been detected in applying those.

**DBSCAN**

In addition, a density-based technique has been chosen. Density-based techniques have the ability to find spherical or arbitrarily shaped clusters that can arise in routine behavior. It is thought that this technique enables the finding of different profiles, which might provide evidence that profiling routine behavior is not solely dependent on distance-based partitioning techniques. The technique chosen in this research is DBSCAN (Density-Based Clustering Based on Connected Regions with High Density). DBSCAN finds dense areas to be clustered together. Any two core points that are close, within a distance ε, are put together. Any border point close enough to a different point is put in a similar cluster as the core point as long as a cluster has a minimal amount of core points (“MinPoints”). In order for this technique to be applied, two parameters have to be set. One possibility of setting the parameters is discussed by Tan, Steinbach, & Kumar (2013). They propose a k-distance graph in which every point in a dataset is linked to its k nearest neighbors. The distances between the points are arranged in increasing order and plotted against the number of distances that have been taken into account. Again, k=5 is chosen as parameter which is said to be similar to the value of MinPoints (Tan et al., 2013). The ε parameter is dependent on the graph, the starting point is where a sharp increase in distances arises. Future research could investigate if and whether different parameter setting yield interesting results in terms of automatic bucketing. In Appendix 21, the DBSCAN k-distance graphs can be found.

**Parameter Setting**

Aforementioned leads to the parameter setting as described in Table 8. The model in RapidMiner to apply the clustering methods is visualized in Appendix 19.
Table 8: Parameter setting for applied clustering techniques

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Distance Measure</th>
<th>K</th>
<th>Number of Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Means</td>
<td>Euclidean</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>K-Means</td>
<td>Manhattan</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>K-Means</td>
<td>Cosine</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>DBSCAN (cat) – Entity A</td>
<td>Euclidean</td>
<td>200,000</td>
<td>5</td>
</tr>
<tr>
<td>DBSCAN (day) – Entity A</td>
<td>Euclidean</td>
<td>0.3</td>
<td>5</td>
</tr>
<tr>
<td>DBSCAN (cat) – Entity B</td>
<td>Euclidean</td>
<td>25,000</td>
<td>5</td>
</tr>
<tr>
<td>DBSCAN (day) – Entity B</td>
<td>Euclidean</td>
<td>0.15</td>
<td>5</td>
</tr>
<tr>
<td>DBSCAN (cat) – Entity C</td>
<td>Euclidean</td>
<td>180,000</td>
<td>5</td>
</tr>
<tr>
<td>DBSCAN (day) – Entity C</td>
<td>Euclidean</td>
<td>0.1</td>
<td>5</td>
</tr>
</tbody>
</table>

5.3.1.2 Profile Description

After analyzing the cluster models in the visualizations (Appendix 22, 23, & 24) and in the raw dataset, four profiles have been chosen. Since multiple clusters have been analyzed, taking all these into account would be an endless summarization of professional judgment. This made us decide to summarize specific profile descriptions. The profiles that have been found are provided in a summarized description in Figure 27. The size of the buckets in relation to the inter-category dataset are provided, as well as a short description on how auditors would label this profile. In addition, whether the journal entry lines fall into high-risk criteria set up by auditors in 2014 has been incorporated. In this way the professional judgment of auditors on its risk-level can be embedded. The connections between the categories and the profiles determine which categories are booked in journal entries. It can be seen that it depends on the entity and the size of the cluster which and how many categories are involved. Note that the categories do not capture the same accounts between entities, this way of visualizing is purely for feasibility reasons.

Seeing the number of journal entries in a profile provides an idea of the periodicity of the routine behavior. Profiles (entity – profile number) A-1, A-2, A-3, A-4, B-2, C-1, C-2, & C-4 have a monthly behavior. Some corrections throughout the year or non-occurring transactions make these transactions not always equal to 12. Knowing about possible different-than-12 occurring transactions could be interesting to auditors. For instance, why would monthly salary payments in B-2 be 8 instead of 12 times per year? The specific behavior has not been discussed in depth with the auditors since the focus was on the changes over the years instead of within the year. In addition, profiles occurring semiweekly (B-1), weekly (B-3), quarterly (C-3), and annually (B-4) have been noticed. The number of journal entry lines provide information on the size of the journal entries incorporated. Routine behavior profiled consists of both large journal entries involving multiple categories (e.g. A-3 & C-1) and smaller journal entries involving less categories (e.g. B-3 & C-3). Some profiles were found to be risky by high-risk criteria set up in 2014, which provides evidence on the previous interest of auditors. Comparing this with the risk level of the following years provides an idea on the impact of a learning effect and professional skepticism that affected the developed high-risk criteria over the years.

In conclusion we found that profiles do capture routine behavior. These profiles found do not only consists out of large often occurring transactions, but are even able to capture routine behavior with less frequencies and smaller sizes. In Chapter 5.3.1.3 the clustering models that best describe these clusters are handled.
Figure 27: Summary of profiles in 2014
5.3.1.3 Performance Metrics

The performance of cluster identification methods can be measured by both internal and external performance metrics (Rendón, Abundez, Arizmendi, & Quiroz, 2011). External performance metrics are based on previous knowledge about the data. Since we are able to dive into the raw dataset, it is possible to link the clusters to the “explanation” attribute of journal entry lines involved in a cluster of journal entries (see Table 1). First, we looked at the journal entries that have been labeled by a specific clustering technique. In the raw dataset containing all journal entry lines, for each cluster of journal entries we analyzed the explanations that are involved. In this way we were able to define the cluster by means of the accounting description involved. In a subsequent step, we checked the raw dataset again for all journal entry lines incorporating the same or similar explanations. This enabled us to find the “true” profile of journal entries that belonged together. As indicated in interviews (Appendix 10, Table 21) this is also the way auditors would label the data, hence we are in the position to evaluate the predicted cluster in comparison to the actual profile. Finally, we linked the profiles found to the ability of a cluster to capture the profile information by means of an external performance metric. The usage of knowledge on the ground truth of the data is suggested by literature (Han et al., 2012; Rendón et al., 2011). The reason for using the explanation as an evaluation is legitimate since it is not chosen as an element that models the journal entry behavior.

External Performance Metrics

A metric that is suitable for external performance measuring is the F-measure (Rendón et al., 2011). This measure is derived from the confusion matrix as discussed by Provost and Fawcett (2013). Its formula links to journal entries in this research by means of the following elements:

- **True positives (TPs)** are journal entries that have correctly been labeled as belonging to a certain profile of routine journal entries by the clustering technique.
- **True negatives (TNs)** are journal entries that have correctly been labeled as not belonging to a certain profile of routine journal entries by the clustering technique.
- **False positives (FPs)** are journal entries that have incorrectly been labeled as belonging to a certain profile of routine journal entries by the clustering technique.
- **False negatives (FNs)** are journal entries that have incorrectly been labeled as not belonging to a certain profile of routine journal entries by the clustering technique.

The F-measure is a harmonic mean between precision and recall (Provost & Fawcett, 2013). It is determined as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (2)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (3)
\]

\[
F\text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)
\]

<table>
<thead>
<tr>
<th>Journal entry belongs to a profile (P)</th>
<th>Journal entry does not belong to a profile (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technique labels a journal entry as belonging to a profile (Y)</td>
<td>Technique labels a journal entry not belonging to a profile (N)</td>
</tr>
<tr>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

A confusion matrix (Table 9) differentiates between decisions made by a data mining application in terms of how they confuse one class with another one. In this case, TPs can be seen as the amount of journal entries that correctly predict a certain profile. In the mind of an auditor, the TPs could be identified different from the TNs which do not belong to that specific profile. The F-measure is preferred to recall or precisions. Although a high precision reduces the chance that FPs are taken into account, it can lead to
reducing the profile to such a low number of journal entries, i.e. a low recall, that putting them in a bucket does not make any sense in terms of operational efficiency. Future research could link the measure to a model involving benefits and costs in a way that maximizes the expected value (Provost & Fawcett, 2013).

Profile Performance Determination

The evaluation of the profiles takes place as follows for every entity:

1. Analyze clusters found using raw data.
   - This will be done by “explanation” and further by analyzing attributes that were the input of the modeling which are account, amounts involved, credit/debit.
2. Assign F-measure to the cluster methods that predicts a meaningful profile.
   - Not all clusters contain journal entries that have a similar behavior, only when journal entries in a cluster have meaningful co-occurrences it can be profiled and subsequently measured. A cluster with an F-measure lower than 0.5 is assumed to be not a meaningful profile.
3. Select 4 profiles in which the maximum F-measure is larger than 0.5.
   - Four profiles will be selected which suffices for our purpose of finding evidence that routine behavior can be detected.
4. End profiling

5.3.1.4 Evaluation of Results

In this Chapter the findings of previous steps are discussed. Remember that the goal of profiling in this research is based on finding evidence that clustering the behavior of journal entries could lead to finding routine transactions. The resulting clusters can be found in Appendix 22, 23, & 24 in which marked circles indicate whether clusters were chosen to be profiles. In Figure 28 the F-measures of the clustering models that profile the journal entries that were discussed in Figure 27 have been shown. The left graph visualizes the Profile-Entity-Number on the horizontal axis, while the right visualizes the Clustering Method/Distance Measure –Attribute set on the horizontal axis. Only cluster models that had a F-measure higher than 0.5 have been taken into account. It can be seen that all models used, were able to profile journal entries. Looking at the maximum F-measure for every clustering technique used, it can be seen that all have the ability to find profiles with an F-measure of at least 94% (Figure 29). This is a strong indication that the techniques involved have the ability to find (near-to-)pure profiles. Especially k-means Euclidean distance and Manhattan distance have the ability to find pure profiles.

Concluding from the cluster models in Appendix 22, 23, & 24 k-means algorithms using Euclidean and Manhattan distances provide purer and more meaningful profiles on the (cat) dataset. This could be explained by the distance measures that are involved. Distance measures on absolute valued attributes tend to be more affected over the total distance. Especially in this case, in which profiles are routine behaviors sticking together, with very distinctive behavior from the rest might lead to better results on (cat) data, rather than being distorted by normalized attributes (day). DBSCAN and k-means using cosine distance measure provides more insightful profiles on normalized data (day). An additional remark is that DBSCAN had the most computational time required, especially for the entity C dataset. In some extent it is therefore interesting to take time as a factor for finding buckets when doing future research.

In summary, the four modeling techniques were able to find meaningful clusters in two different attribute sets, which provides evidence for our tool finding routine behavior. Profiles have distinct behavior which can only be captured in specific tailored ways. Dependent on the profiles to be detected the tool should
incorporate specific cluster models and specific attributes. Within the audit this means that the cluster models can and need to complement each other in finding routine behavior.

![Figure 28: F-measures of identified clustered profiles 2014](image)

![Figure 29: Maximum F-measure per clustering technique](image)

5.3.2 Detection of Routine Behavior

The second step requires the tool to exploit the best technique for each profile and use it to automatically detect that same profile in the subsequent year. The labels provided by this cluster model are used to "train" the classification model. Using outcome labels following the clustering models simulates an environment in which the clustering models selected would automatically find routine behavior. When this is actually the case, this cluster model's labels are a good representation of the routine behavior. The detection of labeled data is possible by classification techniques. As discussed, classification techniques are supervised and predictive data mining applications that can automatically label a data point by basing it on previously learned patterns (Han et al., 2012; Provost & Fawcett, 2013).

For the analysis an experimental design is followed (Figure 30). As discussed, the clustering models that were able to find a cluster that best fits the profile have been used to provide the labeled data. Next, two classification techniques exploit this information to label profiles in a new unlabeled dataset. One of these classification techniques have been researched using two different parameter settings because of the
impact it might have on detecting routine behavior. The unlabeled dataset are the manual journal entries in 2015. By evaluating the profiles, discussed in Chapter 5.5, auditor’s opinion on the risk level of the profiles could be explored which transforms the profiles into meaningful buckets.

First, the classification techniques used are handled (Chapter 5.3.2.1). In Chapter 5.3.2.2 the profiles that have been found are explained. Needless to say that their behavior is similar to the ones found in Chapter 5.3.1. Then, the performance metric, called the F-measure, used to evaluate the techniques has been described in Chapter 5.4.2.3. Finally, in Chapter 5.3.2.4 the techniques are evaluated.

5.3.2.1 Modeling Techniques

Classification works in a two-step process (Han et al., 2012). First, the classification model should be built based on learning a labeled dataset. Subsequently, the model is used to predict the labels in an unlabeled dataset. The previously discussed profiling part of automatic bucketing utilized both distance-based clustering techniques as density-based clustering techniques. A distance-based measure utilizes the distance between data point as a way to compute similarity. Since this similarity measure is also used in some sense by the \( k \)-means cluster technique, a distance-based classification technique is worth investigating. In the previous step both distance-based as density-based clustering technique were able to find routine behavior. Apparently, the normalization in combination with adding the “GLDay” attribute has an impact that is better captured in other ways than solely distance. Because of this, we have reason to believe that a technique different from distance-based measuring could be feasible.

Hoffman & Klinkenberg (2014) indicate that \( k \)-Nearest Neighbor and Artificial Neural Network are two methods capable of working with continuous data, which is the basis of the used dataset. We have therefore chosen to evaluate the distance based \( k \)-Nearest Neighbor (kNN) classification and the Artificial Neural Network (ANN) classification. Since the goal of this research is providing guidelines rather than an all-encompassing tool, this approach is appropriate.

\( k \)-Nearest Neighbor (kNN)

This technique arose from the idea of \( k \)-means’ ability to measure distances between journal entries. By feeding an unlabeled journal entry in an labeled dataset it might be possible to find the nearest neighbor(s) of this unlabeled journal entry in terms of distance. Classifiers like \( k \)NN are based on comparing unlabeled data points with already labeled data points. A data point represents a point in an
In an $n$-dimensional space. By introducing an unlabeled data point to the dataset the similarity with the $k$ nearest neighbors are measured to label the data (Reza, Rezaei, & Minaei-bidgoli, 2009). This similarity is, just as with $k$-means, measured by a certain distance measure. In this research a commonly used distance measure for continuous attributes is used: the Euclidean distance (Figure 26). It is a lazy learner method, which stores information on a dataset only to be used when a new data point enters a dataset (Han et al., 2012; Jiang, Cai, Wang, & Jiang, 2007). After the distances have been measured, the data points closest vote on the most common label at the $k$ nearest neighbors. Since it might be useful to weight the contributions of the neighbors, it is chosen to involve a weighted vote as a way to let the nearer neighbors contribute more than more distant ones.

The choice of $k$ is sensitive to “noise”. Since the data points often involve monthly transactions a $k$ is chosen smaller than 12. To compare two different value measures we have chosen both the uneven $k$ values 3 and 9. In Appendix 20, Figure 81 & 82 the RapidMiner model and its settings has been visualized.

**Artificial Neural Network (ANN)**

Not only the distances between journal entries have to be used when detecting journal entries. A way to incorporate learning on other ways than distance-based is for instance Artificial Neural Network (ANN) classification. It is a set of connected input/output units with weights attached to them. By automatically adjusting the weights in the learning phase, ANN is able to predict the correct class of the input units (Han et al., 2012). In this way ANN can actually learn the computer to think like a person in a multi-attribute environment. In other words, it learns implicit rules that apparently label a journal entry as belonging to a profile or not. It exists out of a network of simple processing elements called neurons (Dunne, 2007). The technique used, learns a model by means of a feed-forward neural network trained by a back propagation algorithm which is one of the most popular neural networks (Svozil, Kvasnicka, & Pospíchal, 1997). The feed-forward neural network connects neurons by moving forward from the input nodes, through the hidden nodes, to the output nodes without any cyclic behavior. Back propagation facilitates in comparing the output values to the correct answers to compute the error. This error is fed back to the model, which can with this information update the weights. After reiterating this process for several training cycles (500 in this research), the network converges to a state of small error calculation. From this point forward the ANN is modeled to be a classifier for new unlabeled data.

Before ANN could be applied it is important to have a normalized input to reduce the impact of each attribute’s absolute value and have a balanced dataset. Normalization has been applied on each attribute by the min-max normalization as discussed in equation 1. In order to normalize, data points over the two years have been appended in order to scale values in both years within the same range (see Appendix 20, Figure 81).

The dataset used for learning ANN is subject to the class imbalance problem (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). In an imbalanced dataset the label values are not approximately equally presented. Having a well-balanced dataset is of utmost importance for training a good prediction model, otherwise models tend to lose generalization. Different approaches to improve the classification accuracy of class-imbalanced data exist (Han et al., 2012). We pursued over-sampling by a technique discussed by Chawla et al. (2002) which is called SMOTE. It is “…an over-sampling approach in which the minority class is over-sampled by creating “synthetic” examples rather than by over-sampling with replacement” (Chawla et al., 2002, p. 328). This tool is not available in RapidMiner, which made us use KNIME. The models were based
on labeling the profiles of interest and let KNIME develop synthetic examples around the vectors involved which provides us with a dataset of 50-50 distribution of “profile x”-“no-profile”. The model used in KNIME can be seen in Appendix 25, Figure 90.

The next step is rearranging the positively labeled data points (journal entries which are labeled as belonging to a certain profile) and negatively labeled data points (journal entries which are labeled as not belonging to a certain profile) to train the model. The rearranged data points have been imputed to a X-validation in RapidMiner. In this cross-validation the dataset, which exist out of all journal entries in 2014 appended with an almost equal synthetic dataset of labeled data points, is distributed in 10 different parts. By training the model on 9/10 part of the dataset and testing in on 1/10 part of the dataset the model is optimized. The optimizing consists out of averaging the 10 different models towards one representative model. The model utilized for cross-validation has been shown in Appendix 20, Figure 84.

Further, the parameter settings within ANN are important when analyzing specific clusters. Since the goal of this research is finding out if neural network suits the detection of routine behavior, rather than finding an optimal parameter setting, the standard setting in RapidMiner is used. These can be found in Appendix 20, Figure 85.

In summary, first ANN needed normalized 2014 data that has been supplemented by synthetic labeled data points. The labels were dependent on the profiles that have been detected in which for each labeled profile a different dataset has been developed. Then, cross-validation has been used to develop an ANN classification model for each dataset labeled by profiles. Subsequently, the model has been applied on a new unlabeled dataset that consisted out of 2015 data points without any information on particular profiles. The outcome of this step was an attribute involving a confidence between 0 and 1 that a particular prediction was a profile. All values above 0.5, i.e. a chance of 50% that it belonged to the profile, were used for measuring the performance of the detection techniques used. The model in RapidMiner can be seen in Appendix 20, Figure 83.

5.3.2.2 Profile Detection

To automatically detect journal entries in 2015, the clustering values of the previous clustering techniques have been used as a label for profile detection. The clustering techniques with the highest F-measure have been used for this, which can be found in Appendix 26, Table 40. After previous detection techniques have been applied, the profiles that already have been discussed should be found. In Figure 31 the profiles have been summarized. In addition to showing the profiles also the behavior changes have been visualized. Over the two years three distinctive behaviors have occurred, which are linked to buckets in Chapter 5.5:

- The routines journal entries proceed as expected. *The routine behavior continuous with a similar periodic pattern throughout the two years.*
- The routine journal entries proceed not as expected. *The routine behavior has a different pattern in the current year.*
- The routine journal entries do not arise anymore. *The routine behavior does not occur anymore in the current year.*
Figure 31: Summary of profiles in 2015
5.3.2.3 Performance Metrics
Similar as discussed in Chapter 5.4.1.3, external indices have been used based on the ground truth of the data in the form of F-measures (Han et al., 2012; Rendón et al., 2011).

5.3.2.4 Evaluation
As can be seen in the F-measures in Figure 32, all three classification models have the possibility to detect the profiles to some extent. As can be noticed, profiles B-4 and C-1 did not occur anymore and can therefore not be detected. When analyzing the relative height of the bar, utilizing a majority vote of 3 nearest neighbors performs slightly better than the 9 nearest neighbors. ANN on its turn has different results in terms of finding profiles. In some cases it is better than in other cases.

As discussed, all three models have the ability to find similar behavior in a subsequent year. For a further analysis, the profiles have been subdivided based on their behavior over two years to see which classification model would be best in detection profiles. It is dependent on the quality of the previous cluster in how far the classification model is possible to detect behavior in the subsequent year. We have developed 3 criteria that should determine which modeling technique is best in particular situations. These will be discussed below.

**Criterion 1:** “If a pattern does not exist anymore, the detection technique should be able to notice that.”

The profiles in question were profiles B-4 and C-1. All three classification models were able to detect the fact that a pattern stopped occurring.

**Criterion 2:** “If a previously used clustering model fails to predict all journal entries of a profile, a classification model should correct for this impurity by learning on a smaller amount of journal entries.”

A clustering model that is presumably the best predictor, fails to predict all journal entries since the F-measure is below 100% caused by at least one FN. A classification technique should be able to correct for this flaw in the clustering model and predict less FNs. The clusters that were predicted but contained more than 0 FN were: A-1, A-2, A-4, B-2, B-3, C-4. The “recall” (Equation 3) is a measure that incorporates the FN value in terms of its ability to measure purity. In Figure 33, the recall measures of the profiles have been visualized. We are aware that not all behavior continues as regularly, which diminishes the idea of comparing with the clustering results in 2014 (yellow “clustering” line). What can be shown however, is that ANN outperforms $k$NN on correcting for impure profiles due to FNs. Its recall is higher or equal in all
cases. Additionally, no conclusive answer can be given comparing 3NN and 9NN. Summarizing, ANN is better in correcting for impure detected profiles due to labeled data having one or more FNs.

\textbf{Criterion 3:} “If a detection technique detects all journal entries of a profile they are preferred.”

There are two profiles that have been classified purely by the cluster models (A-3, C-2), and two that have been detected with some additional FPs (B-1, C-3). Although pure profiles are preferred, if a model has the ability to at least single out all TPs it provides insightful information as well. It cannot be expected that the detection models correct for FPs since models learn on the data you put in, often called “garbage-in-garbage-out”. In Figure 34, the F-measures of these pure profiles are visualized. We are aware that not all behavior continues regularly, which diminishes the idea of comparing with the clustering results in 2014 (yellow “clustering” line). The results are inconclusive in comparing \(k\)NN with ANN. ANN better predicts the pure profiles that had a 100% clustering result. However, \(k\)NN outperforms ANN on profiles that beside all TPs contain some FPs. 3NN performs better or equal to 9NN. It can be concluded that 3NN is suggested when also FPs have been taken into account, and ANN when pure profiles have been clustered.

\textbf{Figure 34: Classification’s ability to learn on purely predicted clusters}

No consistent results on whether to choose 3NN, 9NN or ANN can be found, but it seems that all models have the ability to detect routine behavior. Because of the possible importance of (day) or (cat) attributes in the dataset we compared these. Profiles detected by the attribute set (cat) are the following: A-1, A-2, B-1, B-2, C-2, C-3, C-4. For (day) these are: A-3, A-4, B-3. Although the sample set is small, it might be that
ANN is better with incorporating the specific behavior of clusters that had more features involved which was found based already on normalized data (day). In Table 10 the F-measures of the classification models has been averaged to determine the detection ability. As can be seen, the averaged F-measure is still best detected in the dataset containing all absolute category attributes for all three detection models.

<table>
<thead>
<tr>
<th>Classification Technique Used</th>
<th>Avg. F-Measure (cat)</th>
<th>Avg. F-Measure (day)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3NN</td>
<td>93.55%</td>
<td>86.15%</td>
</tr>
<tr>
<td>9NN</td>
<td>90.29%</td>
<td>85.45%</td>
</tr>
<tr>
<td>ANN</td>
<td>80.90%</td>
<td>70.55%</td>
</tr>
</tbody>
</table>

A final note on this analysis is that ANN requires a significant longer learning time than kNN modeling, however this is not incorporated as input for determining the best detection technique.

In conclusion, all three classification models tried were able to detect journal entries in an unlabeled dataset to some extent. However, no conclusive recommendation on which classification model is better in detecting routine behavior can be provided. Increased sensitivity analysis by altering the clusters and parameters could be future research as a way to best detect routinely occurring journal entries. In the next Chapter, the conclusions of testing the tool will be described.

5.4 Evaluation of the Tool

It can be stated that all clustering models used for profiling can be used to find journal entries. The 12 profiles that have been chosen for further investigation, are predicted to some extent by all three classification models tried. The considerations and conclusions from the analyses in this research provides future researchers an idea of how automatic bucketing could take place and which kind of considerations have to be taken into account to modeling routine behavior of journal entries.

Three conclusions can be made from this analysis. All three lead to evidence saying that automatic bucketing could be applied to profile and detect buckets of manual journal entries.

1. Adding attributes finds different clusters, which implies that a combination of multiple attributes lead to multiple interesting profiles of routine behavior.
2. Four clustering models have been found to cluster routine behavior. The quality and novelty of the clusters is not conclusively assignable to one specific modeling technique, which implies that a combination of multiple cluster models lead to identifying various clusters of routine behavior.
3. Three different classification models have been found to detect routine behavior. The quality of the detecting techniques is not conclusively assignable to one specific detecting technique, which implies that multiple classification models lead to detecting clusters of routine behavior.

This all leads to several ideas for future research. It can be stated that evidence is gathered that favors the usage of a similar methodology in order to automatically bucket journal entries. A tailored approach for every client is required in a pilot study to detect how this could further be exploited. Future research could investigate a hybrid methodology that finds these data and incorporate other parameters and attributes. Practically, the cluster methodologies should be tailored per entity. Every entity needs different parameters, attributes, amounts of buckets. In order to link the ideas found to actual business process improvements, Chapter 5.5 indicates what this way of evaluating journal entries could bring about.
5.5 Buckets Impact on possible Process Redesign (Deployment)

Now a developed tool has been tested to profile and detect routine behavior in manual journal entries, it is possible to turn profiles detected into meaningful buckets. These buckets provide information of the profile involved and the risk level an auditor attaches to it. For every entity two auditors have been presented with the 4 selected profiles. In this Chapter, we shortly discuss auditor’s opinions on seeing these buckets on a regular basis. In Chapter 5.5.1 the impact of intra-category transactions is being discussed. In Chapter 5.5.2 buckets of routine inter-category transactions is put in light of auditor’s thoughts. This results in a smaller dataset of heterogeneous manual journal entries as discussed in Chapter 5.5.3.

5.5.1 Bucket of Intra-Category Transactions

On the bottom row of Figure 24 (Chapter 5.3.1) a specific journal entry can be seen which is said to be an intra-category transaction. It has arisen from transactions that have taken place between accounts within the categories that have been developed. It could be said that by analyzing the journal entries data-analytically, it is possible to filter out such a specific sort of journal entries that could be seen as a bucket.

A special characteristic of intra-category transactions is that they do not have an impact on the net result of a category and on a higher level on the financial statement. Auditors interviewed indicate that the risk-level related to this bucket is indeed lower in terms of management override of controls (Appendix 10, Table 22). This has been indicated by all 6 auditors (AI, AII, BI, BII, CI, CII). Although there are some exceptions to the rule, auditors indicate that getting a bucket of all intra-category transactions to be analyzed before the filtering phase would increase their knowledge of the business.

In conclusion, without any data analysis we have already been able to show auditors that looking at journal entries data-analytically provides us with a bucket of journal entries that have a different risk level than the inter-category counterparts. It show auditors that data-analytical thinking could improve their understanding of the client and is also a stepping stone towards the potential of automatic bucketing.

5.5.2 Buckets of Routine Inter-Category Transactions

During the research 12 profiles have been profiled and detected in 2015. As already explained in Chapter 5.3.2.2, three different situations arose which can be linked to three different buckets. In this Chapter ideas of auditors are shortly discussed as a way to gain insight in risks related to these buckets (Appendix 10, Table 23 & 24).

**Bucket 1: A routine transaction continuous as expected**

Auditors indicate that a routine transactions that proceeds as expected “are not special, therefore are probably better to remain unseen” – (CII). This specific bucket is touched upon later in the Chapter on Engagement Efficiency (Chapter 6.1.4).

**Bucket 2: A routine transaction had a different pattern as expected**

Auditors indicate that if a transaction has a different pattern, it is worth understanding how the behavior changed. However, they are sure that it has a “legitimate reason” – (AII). These buckets would increase the business understanding, but they are not necessarily interesting.
Bucket 3: A routine transaction does not occur anymore

When an initial routine transaction is not occurring anymore it is indicated that: “The client has to explain me, why this specific transaction is not occurring anymore” - (CI). They see this as a real “added value” – (BII) if this information becomes available on-top-off the normal analysis. So it seems that these transactions can be identified more explicitly than that they would have been before. In this way auditors can adjust their high-risk criteria to incorporate these buckets as well.

5.5.3 Bucket of Heterogeneous Inter-Category Journal Entries

The Chapter ends with an indication of the impact such bucketing has on the remaining dataset. If the data in entity A2 would be bucketed as theorized by the automatic bucketing tool, the dataset of heterogeneous manual journal entries decreases with 36.17% (Figure 35). For Entity B and C this leads to a decrease of 21.63% and 3.29% respectively (Appendix 27, Figure 91 & 92). Dependent on the nature of the procedures that have to be conducted on the other developed buckets, the operational efficiency of a journal entry test could increase.

In Chapter 6 the link between applying automatic bucketing and gaining operational efficiency is specified by means of design principles.

Figure 35: The impact of automatic bucketing profiles in entity A 2015
6. Linking Automatic Bucketing to Engagement Efficiency

“It is more efficient to exclude certain routine behavior from your high-risk criteria” - Auditor AI, 2016 -

Suppose automatic bucketing is an integrated part of auditors’ procedures in journal entry testing. It would lead to a redesigned situation in which comparable behavior, both in terms of periodicity and risk-behavior, is captured in buckets and decisions are made dependent on these specific buckets. In this situation, IT advisors are involved in detecting buckets in the dataset and exploring its content together with auditors. This situation would increase engagement efficiency in several ways which can be captured in so called design principles that guide practitioners and researchers in their endeavors to increase operational efficiency.

This Chapter focuses on proposing design principles as evidence that automatic bucketing leads to engagement efficiency. Four design principles have been set up after reviewing literature. Semi-structured interviews with 6 auditors have been used to validate these design principles to some extent. The purpose of this part of the research is twofold. Firstly, ideas on how developing buckets of routinely behaving journal entries could facilitate auditors’ procedures of auditors are handled. Secondly, design principles that are derived from literature can be linked to thoughts of auditors.

6.1 Prescriptive Design Principles

Design principles have been developed by linking them to operational efficiency, and to be more specific, to business process redesign heuristics (Dumas et al., 2013; Reijers & Liman Mansar, 2005). These heuristics are constructed based on a collection of earlier business redesign projects. By incorporating the paradigm of operational efficiency, an evaluation framework known as the “Devil’s Quadrangle” inspired us to ground test the design principles developed (Dumas et al., 2013; Reijers & Liman Mansar, 2005). The devil’s quadrangle has four performance dimensions, which are quality, cost, time, and flexibility (Figure 36). Ideally, a business process redesign improves a situation along all four dimensions, but in practice one improvement might have a weakening effect on another. The following definitions were set up:

**Quality**: The degree of an auditor feeling in control over its performed journal entry testing procedures. It is measured differently among design principles.

**Time**: The throughput time of journal entry testing. Since auditors are the bottleneck in journal entry testing, it is defined as “Time required by auditors to follow up on the procedures of journal entry testing”.

**Cost**: The costs related to journal entry testing. The developed tool is open source and requires involvement of IT advisors, it is defined as “Time required by auditors/IT advisors times hourly wage”.

**Flexibility**: The ability to react to changes (Jansen-Vullers, Looschilder, Kleingeld, & Reijers, 2007). It is inherent to automatic bucketing that more different buckets of journal entries arise that are subject to a different approach. This requires auditors to be more flexible, so we decided to look at it more from the ability of auditors to “Detect changes in the accounting system of the client”.

Auditors indicated that flexibility has a strong correlation with quality; and costs has a strong correlation with time (Appendix 10, Table 25). This made us decide to only discuss the impact a certain design
principle had on the quality and time dimension. The relative trade-off that increased quality should be worth the time investment is thought to be more feasible when a hypothetical solution is discussed.

By semi-structured interviews user’s visions on these scenarios are received and design principles that have been grounded in theory are contextualized in practice. During the meetings, auditors were actively exposed to specific buckets of their audit client to put ideas of them in context and increase the practical relevance of their thoughts. This latent way of discussing buckets was found to be very insightful for the purpose of this research. It is by no means meant to test the best way automatic bucketing takes place, rather an exploration of possible mechanisms that could invoke engagement efficiency. In the following paragraphs four design principles are discussed. The design principles themselves have been found after a good understanding of the auditing, CAAT, and operational efficiency research fields. In every paragraph, first the knowledge base is discussed, whereafter the opinion of auditors are incorporated. This eventually leads to a design principle worth investigating further by actually applying automatic bucketing.

6.1.1 IT Advisory Involvement
The first hypothesis links back to the business problem. It is thought that automatic bucketing leads to a more fertile cooperation between auditor and IT advisors. The business process redesign, in which IT advisors are of great importance to conduct and discuss the buckets found, incorporates a point in the journal entry testing process where IT advisors are structurally more involved. Linking back to the cause-and-effect diagram in Appendix 2, this might lead to auditors having the ability to exploit more tools.

6.1.1.1 Theoretical Grounding
At the moment there is a lack of communication between auditors and IT advisors which limits knowledge sharing (Curtis et al., 2009; Mieke Jans, Lybaert, & Vanhoof, 2010) and limits the exploitation of all CAAT possibilities. When there is an overlapping field of interest, which diminishes this isolation, auditors and IT advisors start speaking the same language. This requires auditors and IT advisors to be positively interdependent on each other as a way to transform group work into team work. The benefits of inter-departmental communication and collaboration are among others (Hansen & Nohria, 2006):

1. Better decision making as a result of advice obtained.
2. Innovation through cross-pollination of ideas.

To create such a cooperative group climate, knowledge sharing is important (Huxham & Vangen, 2005), which on its turn has positive implications for organizational performance and innovativeness (De Dreu, Nijstad, & Van Knippenberg, 2008; Paulus & Nijstad, 2010; Van Wijk, Jansen, & Lyles, 2008). Research implied that the collaboration between these two parties increase mutual understanding and with that knowledge sharing (Bauer & Estep, 2014). Further, O’Donnell, Arnold, & Sutton (2000) suggested that group decision quality improves when auditors and IT advisors work together. However, academic research on the involvement of IT auditors in audit engagement is lacking (Curtis et al., 2009).

When linking this to literature on business process management (Table 11), it is found that the Customer Team heuristic, as a way to compose interdepartmentally work teams, might improve the quality by enhancing the cooperative group climate. This heuristic suggests that the quality of operations could improve, however seeing the Numerical Involvement heuristic it could also lead to coordination problems. It is thought that the cooperative group climate’s increase in knowledge sharing and efficiency improvements throughout the entire audit process, would supersede the coordination problems.
6.1.1.2 Practical Validation

Aforementioned theoretical grounding indicates that discussions on developed buckets, lead to auditors and IT advisors learning from each other’s profession. This has a positive effect on knowledge sharing. One element of this knowledge sharing is IT advisor’s motivation to think along with auditors on how their expert knowledge and developed tools could actively be incorporated in other places during an audit. This might lead to efficiency gains.

Auditors indicated that in a theoretical situation increased communication would increase the quality of an audit (AI, AII, BI, BII, CI, CII), yet it relies on the organization and setting up of good boundaries (AII, BI)(Appendix 10, Table 26). Auditors must remain responsible for the audit, and IT advisors could at the most provide some additional information (BII). Furthermore, the discussed information should always contain the most essential insights (CI). This shows that first confidence in IT advisors’ added value should be gained, in order to establish a more fruitful collaboration. Showing buckets and discussing on its content shows auditors that IT advisors actually understand and are willing to talk in their “language”. Auditors indicated that they are willing to invest time in this to make increased collaboration happen, however they state that there is a long way to go (BI, BII).

6.1.1.3 Design Principle 1

The auditors’ opinions on IT advisory involvement suggests that possibilities for cooperative group climate arise. They are willing to see the advantages of this situation by means of investing time. Actual improvements in efficiency gains might not be apparent in the first few years, they state that first communication should arise in which a mutual understanding is created. In conclusion, automatic bucketing could lead to increased communication as discussed by the auditor; actual efficiency gains are a next step that is hard to imagine on the short term. This leads to proposing a design principle that is in line with our hypothesis, hence is able to facilitate future researchers and practitioners.

In journal entry testing (C), in which manual journal entries are tested for management override of controls, (automatic) bucketing of journal entries (I) could lead to increased communication between auditors and IT advisors (M) which initiates a cooperative group climate (O)

Future (action) research has to test if this also actively leads to increased involvement of IT advisors in other places during an audit. Actually testing these would imply involving all parties and measuring the efficiency gains using an experiment-like situation.

Table 11: Theoretical grounding for design principle 1

<table>
<thead>
<tr>
<th>Literature Review (Other)</th>
<th>Heuristics Operational Efficiency (Dumas et al., 2013)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperative Group Climate (Bauer &amp; Estep, 2014; O’Donnell et al., 2000)</td>
<td>Customer Team - “Consider to compose work teams of people from different departments that will take care of the complete handling of specific sorts of cases” – pp. 268</td>
</tr>
<tr>
<td>Inter-Departmental Collaboration (De Dreu et al., 2008; Hansen &amp; Nohria, 2006; Huxham &amp; Vangen, 2005; Paulus &amp; Nijstad, 2010; Van Wijk et al., 2008)</td>
<td>Numerical Involvement - “Minimize the number of departments, groups and persons involved in a business process” – pp. 268</td>
</tr>
</tbody>
</table>
6.1.2 Internal Control Implications

The second hypothesis involves the positive impact automatic bucketing might have for the client. In this Chapter we focus on implications for internal controls at a client firm. When an accounting information system of a client is controlled and structured, auditors can rely on the (IT) controls which means that less journal entries would be labeled as “manual”.

6.1.2.1 Theoretical Grounding

Having effective internal control environment, gives companies a good understanding of their own accounting environment and could lead to a reliable process (Curtis et al., 2009). Transactions that transcend these internal controls could be labeled as “manual”, hence is input for the journal entry testing process of an external auditor. Argyrou & Andreev (2011) indicate that finding homogeneous buckets of journal entries could supplement internal control procedures. This might lead to less journal entries to take into account by external auditors, which has an impact on the time spent on this element.

In literature, Hogan & Wilkins (2008) stated that auditees pay less for external audits when the internal control environment of a company is more effective. This implies that efficiency gains might arise in a situation in which auditors would discuss the implications for internal control improvements. There is, however, also literature indicating that there is no relation between time invested by an external auditor, and the effectiveness of internal controls at a client (Felix & Gramling, 2001; Hackenbrack & Knechel, 1997). However, these studies were conducted before the Sarbanes-Oxley Act of 2002 which strongly enhanced the regulation of controls. This regulation increased auditor’s sensitivity to rely on these controls (Hogan & Wilkins, 2008). Taking a different position, Hay, Knechel, & Ling (2008) found that strong internal controls are complementary rather than substitutable, which therefore does not impact the audit fees or time required by external auditors. The context of journal entry testing has not been grounded in literature yet, so the link with journal entry testing and internal control reliance could be interesting future research.

When linking this to literature on business process management (Table 12), it is found that the Control Relocation heuristic might have a positive impact when auditors start advising auditees on certain internal controls. This increases customer satisfaction, and takes away time in journal entry testing. Although it might increases the chance of fraudulent behavior at the client’s side, external auditors are still required to test the effectiveness of the novel or improved control which minimizes this side effect.

6.1.2.2 Practical Validation

Aforementioned theoretical grounding is lacking and does not necessarily favor the design principle. However, when incorporating the business process redesign heuristic there is some indication that automatic bucketing could lead to insights in routine behavior that could just as well have been labeled as “automatic” when the correct controls are in place.

The idea of placing some more control at a client could be an outcome of automatic bucketing (A1, AII, B1, & CII), but this placing cannot be forced. It is auditor’s sole responsibility to notice an inefficiency rather than advising or actually implementing an improvement (C1). It is difficult to see this mechanism happening from the view of auditors, since actively identifying inefficiencies has not been done by all auditors until now (BII). When looking at the main line of thought of auditors, it can be seen that possibilities surely arise. The number of implications that are discussed with the client can increase by
automatic bucketing, but there are multiple other elements that have to be in place. The idea is therefore worth investigating further and hard to imagine on the short term (Appendix 10, Table 27).

### 6.1.2.3 Design Principle 2

The auditors’ opinions of internal control implications implies that identified inefficiencies could be advised to the client more actively. Automatic bucketing provides increased business understanding which could be shared with the client. This not only increases the effectiveness of controls on the long term, which is beneficial for external auditors, but also increase the client satisfaction which is a determinant for the duration of an engagement. In conclusion, automatic bucketing could lead to increased internal control implications; actual efficiency gains are a next step that is hard to imagine on the short term. This leads to proposing a design principle that is in line with our hypothesis, hence is able to facilitate future researchers and practitioners.

> In journal entry testing (C), in which manual journal entries are tested for management override of controls, (automatic) bucketing of journal entries (I) could lead to identifying control inefficiencies (M) which increases the sharing of internal control implications with the client (O)

Future (action) research has to test if this leads to active cooperation of the client, and in addition if this implies less work in journal entry testing. Actually testing these would imply involving all parties and measuring the efficiency gains using an experiment-like situation.

### Table 12: Theoretical grounding for design principle 2

<table>
<thead>
<tr>
<th>Literature Review (Other)</th>
<th>Heuristics Operational Efficiency (Dumas et al., 2013)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Internal Control and Audit Fees</strong></td>
<td>Control Relocation - “Move controls towards the customer” – pp. 263</td>
</tr>
<tr>
<td>(Hay et al., 2008; Hogan &amp; Wilkins, 2008)</td>
<td></td>
</tr>
<tr>
<td><strong>Internal Control and Journal Entry Testing</strong></td>
<td></td>
</tr>
<tr>
<td>(Argyrou &amp; Andreev, 2011)</td>
<td></td>
</tr>
</tbody>
</table>

### 6.1.3 Data-Analytical Mindset

The third hypothesis relates back to the advantages of looking at journal entries data-analytically. By doing so, auditors let the data talk about anomalous behavior worth investigating instead of completely relying on professional judgment. In the form of a complementary toolbox auditors have the ability to focus on the most interesting journal entries which requires their professional judgment. The data-analytical mindset itself arises from insights in buckets, but also exposure to the explorative phase such as discussed in Chapter 5.2.4.

#### 6.1.3.1 Theoretical Grounding

Being able to explore buckets of journal entries that would otherwise be hidden knowledge or unknown, would increase the understanding of the client. Given the limitation of the human eye and the amount of journal entries to be analyzed, extra insights in the data could be effective (Lanza & Gilbert, 2007). When linking it to the literature on anomaly detection, knowing what is normal, i.e. profiling, facilitates the detection of what is abnormal (Bolton & Hand, 2002; Provost & Fawcett, 2013). In this way, identified anomalous behavior could be detected more easily which leads to a continuous improving of audit quality and efficiency (Q. Liu, 2014).

Another field of research indicates that data analysis and data mining could be used as a way to find fraud in general. This field of research has been left out of scope in this research, but the ability to use data
mining of journal entries as a control addition for external auditors has surely impact on the effectivity (R. S. Debreceny & Gray, 2010; Gray & Debreceny, 2014).

When linking this to literature on business process management (Table 13), it is found that the Control Addition happens when auditors check additional journal entries to add quality to the process. Although it can be seen as a control addition, which would imply an increase in time effort, over the years the increased business understanding might lead to more focused high-risk criteria which would reduce the work required.

6.1.3.2 Practical Validation
Auditors agreed with the fact that more information on the data, leads to a better understanding of the client (AI, AII, BI, BII, CI, CII). It is worth the time investment to discuss the automatic buckets as a way to gain additional insights (AI, BII, CII). Knowing explicitly in what form high-risk journal entries arise leads to more precise high-risk criteria and therefore a less extensive testing process (AI). Annual insight in the buckets would imply less work on the long term, since accounting transactions are not that susceptible to disruptive changes (CI), this makes it beneficial in the long run (BI) (Appendix 10, Table 28).

In an unstructured interview held before, an auditor indicated that even cost savings can be made when, before an employee is inquired, pre-knowledge has been gathered through the identification of routine behavior. It would namely lead to a more sophisticated discussion since the auditor clearly did some “due diligence”. As a small side note, it could also add value within setting up high-risk criteria by incorporating the element of unpredictability in the light of specific explorative data analysis.

6.1.3.3 Design Principle 3
It might seem a time investment, but gaining more insights from data during journal entry testing leads to benefits on several places during an audit. Understanding the profile of recurring manual journal entries, would ease auditors’ understanding of a client which increases anomaly detection, improves risk assessment, and facilitates in setting up high-risk criteria. This leads to proposing a design principle that is in line with our hypothesis, hence is able to facilitate future researchers and practitioners.

In journal entry testing (C), in which manual journal entries are tested for management override of controls, (automatic) bucketing of journal entries (I) leads to an increased understanding of an auditee (M) which increases the quality of an audit (O)

Future (action) research has to test if this leads to improved incorporation of data analytics in an audit to understand the accounting transactions at a client. Actually testing these would imply involving all parties and measuring the efficiency gains using an experiment-like situation.

<table>
<thead>
<tr>
<th>Literature Review (Other)</th>
<th>Heuristics Operational Efficiency (Dumas et al., 2013)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Analysis increases insights</strong> (R. S. Debreceny &amp; Gray, 2010; Gray &amp; Debreceny, 2014; Q. Liu, 2014)</td>
<td><strong>Control Addition</strong> - “Check the completeness and correctness of incoming materials and check the output before it is sent to customers” – pp. 270</td>
</tr>
<tr>
<td><strong>Profiling increases insights</strong> (Bolton &amp; Hand, 2002; Provost &amp; Fawcett, 2013)</td>
<td></td>
</tr>
</tbody>
</table>
6.1.4 Effective Sampling

The fourth hypothesis is linked to time savings as a means of solely focusing on the most high-risk journal entries. Journal entries that fall within the high-risk criteria are required by regulation to be tested. This leads to a situation in which nonspecific high-risk criteria provide a large set of high-risk labeled journal entries. It is time-inefficient to test all these separately, yet it is required due to regulation. Setting the high-risk criteria more precisely could lead to excluding journal entries that fall within the high-risk criteria, but are known to be low-risk. A way to do so is providing buckets with information on its content. The bucket could be said to be controlled if the routine behavior in it is behaving as expected. If this control is effective, i.e. “green light”, auditor can incorporate this in the high-risk criteria which enables them to leave out these manual journal entries and reduces the population. Reducing the relevant population by effectively leaving out or sample to only representative journal entries we call “effective sampling”.

6.1.4.1 Theoretical Grounding

The basic idea for proposing such a mechanism is simple: when less journal entries have to be tested, less time has to be spent. When a bucket is detected auditors can choose to statistically sample and select a representative subset as a way of representing the entire population within the bucket. This could make the audit more efficient (Bay et al., 2006; Sirikulvadhana, 2002). In this way, the same situation arises as with a homogeneous population, like automatic invoices, that are allowed to be sampled in a different part of the financial statement audit (Elder et al., 2012).

Beside the idea of sampling, another way of having efficiency gains could arise when the bucket is “controlled”. By stating that a control on a bucket is effective, i.e. “green light”, auditors can rely on the continuance of the routine behavior and therefore choose to exclude the entire bucket from the high-risk criteria. This situation is similar as with internal control that determined the nature of the journal entry to be automatic or manual. If such controls are said to be effective, the amount of work to be conducted can be reduced (Elder et al., 2012).

Both mechanisms could be in place simultaneously dependent on the risk level of the bucket and the level of control in the bucket. Some buckets are obviously low-risk where others still have a slightly different pattern worth sampling on.

When linking this to literature on business process management (Table 14), it is found that the Activity Elimination happens when auditors reduce their activities to only the most high-risk journal entries. This business process design heuristic is an exemplar on how data-analytical thinking could increase efficiency by selecting the high-risk criteria more effectively.

6.1.4.2 Practical Validation

Auditors indicated that this mechanism would surely be worth taking into account when setting up high-risk criteria and that it could lead to efficiency gains (AI, AII, BI, BII, CI, CII). However, already some boundaries arose during the meetings. One auditor indicated that it is dependent on the size of the bucket in how far sampling makes sense or that it is legitimate to solely rely on the bucket’s control (BI). Another auditor focused more on the existence of extreme values that are sufficient to be filtered out of a bucket as a way of sampling (CI). So there are multiple factors that should be taken into account when applying effective sampling, all depending on the buckets included (BI, BII, CI).
The executing of the idea can work in two ways. Auditors could choose to rely on the control and leave out the homogeneous bucket (bucket exclusion) or choose to sample their bucket (bucket sampling). When looking at this possibility, one auditor indicated that when you can rely on a control, it is not required anymore to do sampling (BII), however another auditor thought that as long as auditors have the choice to make this decision per bucket, the solution could be combined (AII).

In conclusion, auditors think there are possibilities for automatic bucketing to decrease the amount of journal entries that have to be tested, yet it depends on many factors in what form this would take place (Appendix 10, Table 29).

6.1.4.3 Design Principle 4
The idea of effective sampling does lead to efficiency gains, as could be seen from literature and auditor’s opinions. Automatic bucketing leads to buckets that could be discussed in terms of the risk they entail, after which future controls could be in placed that analyzes and monitors the content of the buckets. If something unusual happens, auditors could choose to take into account the entire bucket. If everything happens in line with the understanding of the bucket, auditors could choose to either exclude those particular journal entries or select a representative sample as a control addition. This leads to proposing a design principle that is in line with our hypothesis, hence is able to facilitate future researchers and practitioners.

> In journal entry testing (C), in which manual journal entries are tested for management override of controls, (automatic) bucketing of journal entries (I) could lead to sample or exclude understandable buckets (M) which decreases the throughput time (O)

Since auditors should first be able to understand how bucketing works and how controlling these buckets could facilitate the process, actually applying this design principle is currently an idea for future research, and dependent on multiple factors.

<table>
<thead>
<tr>
<th>Literature Review (Other)</th>
<th>Heuristics Operational Efficiency (Dumas et al., 2013)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling/Internal Control Reliance (Elder et al., 2012)</td>
<td>Activity Elimination – “Eliminate unnecessary activities from a business process” – pp. 264</td>
</tr>
<tr>
<td>Effective Sampling (Bay et al., 2006; Sirikulvadhana, 2002)</td>
<td></td>
</tr>
</tbody>
</table>

6.2 How can Automatic Bucketing Increase Engagement Efficiency?
After setting up and validating ideas for design principles, it is now possible to answer the sub research question: How can automatic bucketing increase engagement efficiency?

Four design principles have been proposed that elaborate on ways automatic bucketing can increase engagement efficiency. In Figure 37, the design principles are linked to encountered situations to give an overview on these possibilities. As can be seen, these design principles are not mutually exclusive and could even complement each other. The value of design principles is in the prescriptive statements that can guide researchers and practitioners (J. E. Van Aken, 2004), who are strongly encouraged to pursue this research in the form of actually implementing automatic bucketing in a pilot test.
During the meetings also some additional ideas arose that might increase the value of automatic bucketing. Auditors see value for automatic bucketing when comparing buckets from several different clients (AI). In this way generic routine behavior can be found in entire industries, which could then have an impact in making more advanced CAATs that would increase efficiencies. Also in terms of sourcing new clients, the proposition to the client that the Case Company can do more than solely auditing the financial statement, increases the “brand image” of CPAs (BI, BII).

In conclusion, auditors indicate that an automatic bucketing tool is useful when looking over the entire engagement period. Knowing that they have to invest time in discussing analyses with auditors and linking these to their professional judgment, is justifiable and acceptable in the long term. Design principles proposed have the ability to evoke a cooperative group climate, suggestions for internal control improvements, increased business understanding, and less time-intensive testing procedures. Since they indicated that all these principles would increase the quality/time ratio, they are found to increase engagement efficiency.

![Figure 37: Design principles linking automatic bucketing and engagement efficiency](image-url)
7. Conclusion

“All achievement, all earned riches, have their beginning in an idea” - Napoleon Hill, 1938

After analyzing the results, we will conclude on the research and its results in this Chapter. The research first transformed an idea into a tool that has been tested which will be discussed in Chapter 7.1. In Chapter 7.2, we will conclude on the tool’s ability to increase engagement efficiency. The research conducted resulted in practical recommendations for the Case Company and other practitioners, which will be discussed in Chapter 7.3. In Chapter 7.4 idea for future research will be touched upon. Subsequently, in Chapter 7.5, limitations in the current research will be discussed. Finally, a small note on the ethical aspects that have been taken into account will be mentioned in Chapter 7.6.

7.1 Journal Entry Testing using an Automatic Bucketing Tool

Ideally, an auditor would test all manual journal entries to maximize both business understanding and assurance. However, in terms of costs and time this situation is not feasible. An auditor will therefore filter the manual journal entries towards a set of most high-risk journal entries. This process is called journal entry testing and is an important part of the financial statement audit. The process is susceptible to intuitive decision making based on proven business rules and subjective professional judgment. A more data-driven approach to journal entry testing could facilitate auditors in a sense that increases efficiency.

The current journal entry testing process could be improved with two lines of thoughts in mind. Firstly, homogeneous routine behavior could be profiled that is different from the heterogeneous incidental manual journal entries. Secondly, detecting these profiles in subsequent years would automatically bucket certain behavior in line with auditor’s opinion on risk. Over the years a learning effect could arise (Beck & Wu, 2006; Earley, 2015) that ensures auditors to only focus on the most heterogeneous (and likely higher risk) journal entries during journal entry testing.

We developed and tested a tool that provides ideas on how the behavior of journal entries could be clustered into meaningful profiles. Behavior of journal entries is modeled by means of the accounts involved, the height of the account mutation, and whether that amount was credited or debited. In addition, the day in the month a journal entry has been posted to the general ledger contributes to the identification of routine behavior. The tool’s applicability is demonstrated, by showing that recurring and homogeneous journal entries can be clustered together into meaningful profiles. We have found that finding routine transactions requires multiple different approaches. Both the nature of the clustering model used as the attributes that have been incorporated bring about different kind of routine behavior.

The second part of the tool utilizes the labeling possibilities of the “profiling” part as a way to detect the routine behavior over the years. Finding these patterns has been done using three different classification models. All three different classification models were capable of detecting routine behavior in a subsequent year, yet it depends on the situation in how far their effectivity prevails.

In conclusion, an automatic bucketing tool has been applied and tested that facilitates auditors in understanding homogeneous and routine behavior in what was initially presumed a heterogeneous dataset. Auditors indicated that looking at journal entries in terms of buckets could be an interesting complementary analysis, that might in the long term even initiate possibilities for operational efficiency gains. This provided us with the ability to link automatic bucketing to engagement efficiency which will be discussed in Chapter 7.2.
7.2 Increasing Engagement Efficiency by Data Mining

Is it always interesting to investigate buckets that capture routine behavior of journal entries? It certainly increases an auditor’s understanding of the business. Additionally, this research found other advantages of automatic bucketing. From an operational efficiency perspective, automatic bucketing opens up possibilities to reduce the dataset of heterogeneous manual journal entries to be tested by the auditor by effective sampling. This way of data-driven decision making can also increase the involvement of IT advisors which initiates a cooperative group climate. Additionally, auditors indicated that it would increase the implications that would be made regarding the internal control environment of a client. In this way clients are guided in having a better controlled accounting information system that in return has the ability to reduce the procedures that have to be conducted by an auditor.

The main thought from the interviews on these topics was that automatic bucketing could increase the quality of journal entry testing, and maybe even the financial statement audit as a whole, which is worth the time investment. The main investment lies in having more discussions with IT advisors of which is necessary to understand the data and buckets better. Although auspicious, it is found that auditors tend to put new ideas in a conservative context. Which means that auditors have trouble linking the idea of automatic bucketing to a redesigned journal entry testing process. This is understandable and even insightful, since this also opens up ideas for future research to take into account. It is, for instance, not their primary goal to actively provide ideas for improvements to the client. They are open for it, yet it is not something that is being pursued actively. Summarizing, it is found that the data-analytical mindset proposed in the form of automatic bucketing can facilitate auditors which leads to a more efficient journal entry testing process throughout the entire audit engagement.

In conclusion, we would like to return to the main research question:

_How can a data mining application increase engagement efficiency?_

By reviewing literature and introducing ideas on how data mining can increase operational efficiency in an audit (Chapter 5.3.3) an insightful initial answer has been given. One of the ideas have been transformed into a conceptual tool, which has been tested on its ability to increase engagement efficiency in Chapter 6. Since the tool uncovers many opportunities for redesign in the current way of working, four design principles have been proposed that answer the “how” part of the research question (Chapter 7).

7.3 Practical Recommendations

The automatic bucketing tool is thought to increase the communication between auditors and IT advisors. The discussion points that arise for both auditors and IT advisors create a cooperative group climate that is a breeding ground for collaboration. In Appendix 28, Figure 93 a business process redesign has been proposed as an initial thought of a different journal entry testing process. In the redesign there should be more intensive discussion points between auditors and IT advisors. The topic guiding these discussions are initial data explorations which are important for auditors to understand the accounting environment of a client better. This redesign ensures that auditors should start seeing IT advisors as both data analysts as IT auditors in order to fully pursue efficiency initiatives. Although they were not the main focus of this research, also IT advisors could from this point forward more actively propose developed tools and analyses during the audit. A simple first step is transforming the point where manual journal entries (or buckets) are provided to auditors to a discussion point. If IT advisors proactively come up with ideas, auditors are more willing to discuss the possibilities of embedding their efforts.
In addition, the automatic bucketing tool could be further tailored for specific clients, and the data mining applications mentioned in Chapter 4.2 could be pursued to increase engagement efficiency. A fully integrated automatic bucketing might be a bit farfetched at the moment, but it could initially be used as an element of unpredictability in which buckets that are unexpected from previous behavior are added to the list of high-risk criteria.

A third recommendation is that high-risk criteria could be specified when incorporating the idea of buckets. If the nature of low-risk buckets would be incorporated in setting up high-risk criteria this decreases the number of high-risk journal entries that have to be tested integrally.

A fourth recommendation is that within journal entry testing, previous journal entries provide information that is interesting for setting up high-risk criteria in a specific year. Higher quality high-risk criteria can be developed when incorporating previous high-risk behavior and their changing patterns.

A final recommendation is both a summarization as a statement: Thinking data-analytically could facilitate the decision making during journal entry testing which brings about efficiency gains on the long term.

7.4 Future Research

This is the first research, as far as we know, that models the behavior of journal entries in a general ledger for the purpose of detecting routine behavior. The automatic bucketing tool developed could benefit researchers and practitioners in its ability to detect routine behavior in a presumed heterogeneous dataset. By doing so we add new literature in utilizing data mining on frequent patterns to facilitate auditors. As can be seen in Table 2, there is no research so far on the functionality of “Frequent Patterns”, which makes it at least a novel research initiative. Furthermore, it combines clustering and classification of journal entries as a way to look differently at journal entries than has currently been done. The goal of automatic bucketing is to detect routine behavior, which in their turn increases the likelihood of finding representative red flags. This already indicates the relatedness between outlier detection and clustering.

We hope that both researchers and practitioners continue our work. This could lead to both more research on the topic, but also an entity-specific automatic bucketing tool that finds all routine behavior in a particular entity. Future research could involve an experiment like situation, in which one pilot group experiments with the automatic bucketing and is compared to one control group.

The most efficient and effective financial audit can be achieved by analyzing several data sources: all journal entries, previous financial statements, sector information, etc. This research only takes into account journal entries and charts of accounts as input data to set an insightful and feasible next step in this young field of research. In the long term more research involving multiple data sources could and should be pursued.

Future research could explore the best way a business process redesign could take place when taking into account a fully automated bucketing procedure. The design principles are proposed, but not fully tested. The current research field of data mining to increase engagement efficiency lacks elaborate research, so even though research findings implies theoretical grounding for design principles, we suggest future research to decontextualize the results even further.

Finally, auditors foresee applications for automatic bucketing outside the scope of journal entry testing. Understanding and auditing journal entries are essential for the entire financial statement audit, so
conducting future research on applications of automatic bucketing outside the scope of journal entry testing is recommended.

7.5 Limitations
In this research, the usual caveats of case study research apply, although the single case study might give an idea on how automatic bucketing would need to be designed and what its impact might be, the results cannot be generalized. Since this research is explorative by nature we can only make correlational claims regarding the outcome (Blumberg et al., 2011).

The data-analytical mindset introduced assumes that the journal entries in previous years are known to be free of management override of control. This assumption is based on stating that all previously taken decisions reflect a good understanding of anomalous behavior. If this is actually the case in the proposed redesigned solution is debatable.

For testing the automatic bucketing tool, only 12 different profiles have been investigated. This limits the validation of the guidelines that bring about the automatic bucketing tool. In addition, it reduced the ability to validate the specific detection models since the scenarios under study were dependent on only a small portion of the 12 different profiles.

The profiles identified did not all contain pure and consistent recurring journal entry lines. This led to some journal entry lines that did not occur as routine as the main behavior of the journal entry. By checking the most frequent occurring explanations throughout the year this limitation does not cause the results to be misstated.

When linking automatic bucketing to operational efficiency a hypothetical redesign has been discussed with auditors. Auditors could have different interpretations on how this redesign could take place. The design principles were not interpreted similar over auditors. Some auditors placed it directly in the context of current procedures, while others directly linked it to a theoretical “perfect” situation. Because of the limited time that was available in interviews, it was not always possible to converge to a similar viewpoint for all auditors. Although it limits the validation of design principles, it creates extra understanding on the parameters that should be in place for automatic bucketing to flourish. Further, as qualitative interviews are used, some level of subjectivity and a lack clear conceptualization might be issues for misinterpretation of the results. Also, this research focused on auditors in particular, in which IT advisors have not been incorporated in validating the design principles which limits the triangulation of results. In addition, other CPAs could have been involved when interviewing to account for journal entry testing governance in other environments. Finally, when validating the design principles, other data collection sources than semi-structured interviews have not been used. Having surveys to analyze this matter increases the validation capability of the design principles since more auditors could be questioned (Corbin & Strauss, 1990).

7.6 Ethics
This Chapter captures a small note on ethical considerations. In this research all personal identifiers from data used has been aggregated or removed. This situation might lead to future research having trouble redoing the specific research, however the supervisors of this project had the ability to analyze the controllability.
References


Issa, H. (2013). *Exceptional Exceptions*. Graduate School of Rutgers, Newark, USA.


Liu, Q. (2014). *The Application of Exploratory Data Analysis in Auditing*. Doctoral Dissertation - Graduate School of Rutgers, Newark, USA.


## Appendix 1: List of Unstructured Interviews – Problem Statement & Diagnosis

*Table 15: List of unstructured interviews held during the first two stages of this research*

<table>
<thead>
<tr>
<th>Function Interviewee</th>
<th>Number of people interviewed (between 30 – 60 minutes)</th>
<th>Predominant topic of unstructured interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auditor</td>
<td>5</td>
<td>What is the function of a CPA?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is auditing?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is accounting?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How is the communication with IT advisory?</td>
</tr>
<tr>
<td>Researcher</td>
<td>4</td>
<td>What is currently being developed within the CPA in the field of data analytics?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What are tools used in practice?</td>
</tr>
<tr>
<td>IT Advisor</td>
<td>7</td>
<td>How is the communication with audit?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What tools are used in practice?</td>
</tr>
<tr>
<td>Workshop</td>
<td>2</td>
<td>Data &amp; analytics in audit</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data-driven journal entry testing</td>
</tr>
</tbody>
</table>
Appendix 2: Cause-and-Effect Diagram

Figure 38: Cause-and-effect diagram of encountered problems and their causes within the CPA

This cause-and-effect diagram has been validated by 2 auditors and 2 IT advisors, after it has been presented. All 4 interviewees answered “Yes” to the following questions.

Questions to Auditors:

Do you think auditors are not always willing to work with IT (lack of idle time or motivation)?

Do you think that predefined rules are sufficient to follow and there is no need to incorporate IT advisors more?

Do you think there is a lack of collaboration and communication between auditors and IT advisors?

Could your ideas on IT advisory now lead to more added value at some places in the audit process?

Do you think auditors apply repetitive tasks?

Do you think there are many decisions you make that are different over time and subjective in essence?

Questions to IT Advisors:

Do you think you could be more involved in an audit?

Do you think there is a lack of collaboration with auditors during an audit?

Do you think there is a lack of communication with auditors during an audit?

Do you think you lack experience with auditing which limits the ease of collaboration?

Do you think that you are able to do more than what auditors request?

Do you think being more involved in an audit would lead to ideas for improvements?
Appendix 3: Engagement Process

An audit is governed as follows (Elder et al., 2012):

The first phase is to set up an engagement with a firm and design an audit approach (Phase I). This phase entails setting up an engagement team, gaining understanding of the business, and planning and designing an audit approach. It is important to get an overview of the controls that are in place within the firm. The more effective the internal controls set up by a company, the fewer evidence has to be accumulated.

Phase II, comprises out of a risk assessment. To test the control risk, several procedures could be conducted. For example, by examining a single transaction within a process and simultaneously inquiring with employees, the internal control risks in that process could be assessed. In addition, monetary amounts of transactions could be verified, which is called substantive test of transactions. Both tests assess the accuracy of company’s processes, which indicate the level of risk within a company.

In phase III analytical procedures and details of balances are being tested. Analytical procedures are used to assess whether account balances and other data appear reasonable. This test could, for instance, check for unusually large amounts that might indicate a risk of fraud. The test of details of balances are procedures intended to test monetary misstatements in the balances. The extensiveness of tests that have to conducted, depend on the trust an auditor places in the controls of a company.

The final phase IV is about completing the audit and issuing an audit report. This report must entail the auditor’s findings and inform users of the degree of correspondence between the information audited and the established criteria.
Appendix 4: List of Semi-Structured Interviews – Journal Entry Testing and Data Mining Questions/Transcript of Semi-structured interview with Researchers

Table 16: Interviewed researchers for validating journal entry testing process and exploring solution directions

<table>
<thead>
<tr>
<th>Function - Abbreviation</th>
<th>Theme</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researcher – R1</td>
<td>Matching of data mining possibilities with filter process</td>
<td>23-02-16</td>
</tr>
<tr>
<td>Researcher – R2</td>
<td>Matching of data mining possibilities with filter process</td>
<td>26-02-16.</td>
</tr>
</tbody>
</table>

Two researchers were interviewed separately one hour each. One interview was taped (R2), the other was not (R1). Only a small summary of research thoughts have been included in this Appendix to indicate the impact their decisions had on the research. A summary of the answers of R1 (Researcher 1) on 23-02-2016 and R2 (Researcher 2) on 26-02-2016 is included in this transcript. Since many insights gain during this interview lead to assumptions and iterations otherwise during the Master’s Thesis, no interest is attached to a more elaborate transcript.

Research motive:

My name is Joost Vandewal and I am graduating at the Eindhoven University of Technology within the field of industrial engineering. This research is the concluding part of my master Innovation Management. The research is about facilitating the decision making process of auditors.

Research objective:

The aim of the study is to increase insights in the journal entries of an auditee, which improves the engagement efficiency. Auditors can use this information over all years the auditee is a client which strongly enhances their decision making process. I hypothesize that when IT Advisors are included in the current journal entry testing and in the same sense provide an advanced analysis for auditors, a new situation arises that is beneficial for both IT advisors and auditors. By finding a methodology to apply data mining, auditors have guidelines to incorporate data mining.

>> This interview is facilitated by a Powerpoint of descriptions. This Powerpoint can be obtained from the researcher if requested.

>> This interview was in Dutch, but the questions have been translated to English in this Appendix.

Could you indicate what your functions within the CPA is?

R1: Researcher within a development team in the USA. He does research in applying meta-data in financial statement (audits)

R2: Researcher to develop a tool for audit facilitation. He does research in developing a data analytics tool for auditors

>> Explanation of research setup. Explanation of journal entry test

Do you agree with this view on journal entry testing?

R1: Yes

R2: Yes

Do you miss elements?
R1: No
R2: No

Did you ever consider applying data mining on this process?

R1: Another part of our research team is involved with this

R2: He is doing this during the risk assessment, in which accounts are summarized as a way to get insight in the flow of financial transactions

>> We explain the terminology used for the different data mining applications, and I trigger ideas from the researcher. Per single definition we walk through the list of applications.

This is my definition of Data Mining, what is your relation to the subject?

R1: He uses metadata to assess risks in terms of the financial statement
R2: He uses data mining on an account level, but not on a journal entry level

Are these applications available for journal entry testing?

R1: He sees value in all application fields. His focus goes out the Classification as a way to label high-risk journal entries.

R2: He understands the applications fields, and with that he sees value in the application fields. His main emphasis goes out to Profiling.

Do you have ideas on how these could be applied on journal entries?

R1: At the moment this is being piloted at the risk-assessment phase.
R2: He indicates that the risk assessment is a place to understand the processes and the bookings made, during the layering only the most peculiar elements should be filtered, which is a process that can be facilitated using data analytics

>> I indicate that the part-by-part evaluation of data mining techniques is over and I ask:

In which application do you see most value?

R1: The classification of high-risk entries based on predefined rules
R2: Profiling as a way to indicate normal behavior

Questions/Transcript of Semi-Structured Interview with Auditors

Table 17: Interviewed auditors for validating journal entry testing process

<table>
<thead>
<tr>
<th>Function – Entity - Abbreviation</th>
<th>Theme</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auditor General – Aud1</td>
<td>Validity journal entry test</td>
<td>20-04-16</td>
</tr>
<tr>
<td></td>
<td>Explore Requirements</td>
<td></td>
</tr>
<tr>
<td>Auditor General – Aud2</td>
<td>Validity journal entry test</td>
<td>Validation Journal entry test on 20-02-16</td>
</tr>
<tr>
<td></td>
<td>Explore Requirements</td>
<td>Requirements Audit on 06-04-16</td>
</tr>
<tr>
<td>Auditor Entity A – Aud3</td>
<td>Validity journal entry test</td>
<td>25-04-16</td>
</tr>
<tr>
<td></td>
<td>Explore Requirements</td>
<td></td>
</tr>
<tr>
<td>Auditor Entity B – Aud4</td>
<td>Validity journal entry test</td>
<td>21-03-16</td>
</tr>
</tbody>
</table>
Explore Requirements

| Auditor Entity C – Aud5 | Validity journal entry test | 31-03-16 |

Five auditors were interviewed separately half an hour each. The two interviews with general auditors have not been taped, the three interviews that are entity specific have been taped. General auditors imply auditors not linked to the entities researched, but are chosen for their knowledge on auditing in general. Only a answered to the validation part of the interviews have been included in this Appendix. Since many insights gain during this interview lead to assumptions, validations, and ideas for the next round of interviews, no interest is attached to a more elaborated transcript.

>> Explain my model of journal entry testing and explain assumptions

What is your opinion on its validity?

Aud1: This is a good representation of journal entry testing
Aud2: This is a good representation of journal entry testing
Aud3: This is a good representation of journal entry testing
Aud4: This is a good representation of journal entry testing
Aud5: This is a good representation of journal entry testing

Can you incorporate history in your risk assessment?

Aud1: If it us used to detect anomalies, yes
Aud2: Yes, but the data should be available quickly
Aud3: Yes, it is not interesting to look to the same elements every year
Aud4: Yes, you can incorporate learning points of the clients as to set up high-risk criteria
Aud5: Yes, if you can find anomalies with it

If clustering is applied to find buckets of related journal entries, would it facilitate your procedures?

Aud1: It would be interesting to detect patterns
Aud2: Information on clusters but also anomalous could be interesting.
Aud3: He is open to include clusters with patterns as a way to increase quality
Aud4: He thinks it adds value when input for meetings in terms of clusters can be analyzed
Aud5: He thinks that the results could be insightful as a control addition

What are elements that should be improved using this analysis?

Aud1: It should improve communication, insight in anomalies, and the possibility to change future audits according to findings
Aud2: Theoretically, we would like to understand the client as much as possible, every time dimension that could be saved with this would be added value
Aud3: A more directed approach aimed at filtering only the highest criteria, and not, how it is now, a set of journal entries subject to specific rules, that are in essence not interesting
Aud4: Having an analysis that provides outliers is probably the way auditing should be done in the future, which should therefore be included in the analysis
Aud5: We already have an efficient way of working with the data, the new analysis, should probably not give any groundbreaking results
Appendix 5: Transaction Flow from Journals to Financial Statements

Figure 40: Flow from transactions to financial statements (Elder et al., 2012)
Appendix 6: Business Processes

Figure 41: Risk assessment sub process

Figure 42: IT control test sub process

Figure 43: Journal entry extraction sub process

Figure 44: Filtering sub process
Figure 45: Layering sub process

Figure 46: Testing sub process
## Appendix 7: Exemplar Business Rules & Professional Judgment

### Table 18: Business rules that could be applied if auditors think it is applicable

<table>
<thead>
<tr>
<th>Collection</th>
<th>Examples</th>
</tr>
</thead>
</table>
| **High-risk Criteria on the Characteristics of Fraudulent Journal Entries** | Inappropriate journal entries or other adjustments often have unique identifying characteristics and may include entries:  
• made to unrelated, unusual, or seldom-used accounts.  
• made by individuals who typically do not make journal entries, that we would not expect to make journal entries or are not authorized to post journal entries.  
• recorded at the end of the period or as post-closing entries that have little or no explanation or description.  
• made either before or during the preparation of financial statements that do not have account numbers.  
• containing round numbers or consistent ending numbers or are just below an authorization or review limit.  
• posted at specific times. |
| **High-risk Criteria by Erroneous Journal Entries** | • journal entries that have negative values (a negative debit or a negative credit).  
• unbalanced journal entries.  
• journal entries that have non-standard formats for data such as posting date.  
• duplicate entries.  
• journal entries with invalid effective dates.  
• journal entries with lines having zero monetary value.  
• journal entries containing the word "restatement".  
• journal entries containing the word "reversal".  
• journal entries containing the word "reclass" or "reclassification".  
• journal entries posted without a user ID.  
• journal entries posted and approved by the same user. |
| **High-risk Criteria by combining characteristics** | Criteria discussed above can be combined to be considered high-risk |

### Table 19: Decisions based on professional judgment that could be applied if auditors think it is applicable

<table>
<thead>
<tr>
<th>Collection</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High-risk Criteria on Assertions for Fraud Risk</strong></td>
<td>An audit team applies professional judgment to find specific fraud risks like revenue recognition and management override. The presence of fraud risk factors and other information obtained during an assessment of the risks of material misstatement due to fraud may indicate increased risk of management override of controls over journal entries and other adjustments.</td>
</tr>
</tbody>
</table>
| **High-risk Criteria on the nature and complexity of Accounts** | An audit team considers whether certain accounts are more likely to contain inappropriate journal entries. Those accounts might include the following:  
• accounts that contain transactions that are complex or unusual in nature.  
• accounts used for post-closing entries and period-end adjustments. |
- accounts that are seldom-used or that may include journal entries or other adjustments processed outside the normal course of business.
- accounts that contain significant estimates.
- accounts that have been susceptible to errors in the past.
- accounts that have not been reconciled on a timely basis or contain unreconciled differences.
- clearing accounts that have not been aged.
- accounts that are otherwise associated with a fraud risk.

| High-risk Criteria as an element of unpredictability | An audit team can select a particular account or characteristic of a journal entry as an element which brings in some unpredictability within journal entry testing. |
Appendix 8: Data Mining Functionalities

Characterization & Discrimination

Frequent Patterns, Associations, & Correlations

Classification & Regression

Clustering Analysis

Outlier Analysis

Han, Kamber, & Pei, 2012

Data Reduction

Co-occurrence Grouping

Classification

Clustering

Profiling

Provost & Fawcett, 2013

Similarity Matching

Link Prediction

Regression

Causal Modeling

Figure 47: Data mining functionalities (Han et al., 2012; Provost & Fawcett, 2013)
Appendix 9: CRISP-DM Process

Figure 48: The CRISP-DM process (Provost & Fawcett, 2013)
Appendix 10: Coding of Interviews – Engagement Efficiency

In this interview round, we were interested in auditors opinion on 3 different fields:

1. Validate results of the automatic bucketing tool (by means of the profiling ability)
2. Explore applications of automatic bucketing Tool
3. Validate design principles

Every interview is taped, whereafter a transcript is written down. A summary of this transcript is added in this Appendix. Specific Quotes are taken into account to discuss auditor’s thoughts in this report. Auditors are guided in their line of thought by means of a Likert scale. The idea is that auditors think about the application of automatic bucketing in terms of quality vs. costs. Discussing auditors thoughts on how automatic bucketing increases the quality of an audit in light of the time investment, proves as a strong indication for the validity of the design principle. Since actual testing the design principles takes multiple years of application, in addition with the lack of current research a validation of the design principle is not possible. Auditors have been linked to the case they belong (Case A, B or C) to, and concatenated with a Roman numeral (I or II). The script of the interviews was the following:

**Interviewee demographics (+/- 5 mins)**
- What is your function within the company?
- For how long have you been doing this job?
- Do you have any previous experience with data analytics in journal entry testing?
- What are your responsibilities within the company?

**Company Description (+/- 5 mins)**
- In general, how many auditors are involved in a journal entry test?
- In general, what is the composition of the audit team in terms of experience of auditors?
- How long does a journal entry test take in general?

**Brainstorm on Automatic Bucketing (+/- 10 mins)**
>> Explain idea of automatic bucketing by means of pictures
- Do you think this idea could facilitate you, as auditors?
- If yes, how can this idea facilitate your procedures?
- If not, why not? What are the restrictions?
- An example of a different way of looking at data are the intra-category transactions, what do you think about this bucket in terms of risk? Would knowing this be able to conduct different procedures?

**Single Case Analysis (+/- 20 mins)**
>> Provide auditors with information on found buckets in 2014 and 2015.

**Iterate 4 times for each bucket**
- Did you know that these journal entries are related?
- How do you think these journal entries relate?
- Detecting this routine behavior on an annual basis might have an impact on your operations. How could your operations be facilitated?

**Cross Case Analysis (+/- 20 mins)**
We propose 4 impacts that redesigns can have (Design Principles) which we would like to validate and ‘score’ on a Likert scale on the 4 dimensions of the Devil’s Quadrangle. In Appendix A one can find the measure schemes of the interviews. Since our aim is comparing the redesign with the current situation and not measuring the performance, the opinions of the auditor are important and will be leading in this interview. By linking back to the profiles detected, we are able to provide tailored examples of these 4 design principles.

>> Explain the definitions of Quality, Time, Cost, Flexibility
1. Quality is defined as the change in business understanding and one main characteristic of each design principle.

2. Time is defined as the change in throughput time related to journal entry testing. Throughput time could be split up in savings through more efficient, often parallel, work and time costs by introducing contact points.

3. Costs is defined as the wage*time related to including IT advisors in the test.

4. Flexibility is defined as the ability to detect changes in the company’s accounting system.

>> Explain scaling measures

1: This will greatly deteriorate when comparing it to the current process
2: This will moderately deteriorate when comparing it to the current process
3: This will slightly deteriorate when comparing it to the current process
4: This will not change the current process
5: This will slightly improve when comparing it to the current process
6: This will moderately improve when comparing it to the current process
7: This will greatly improve when comparing it to the current process

Every design principle has different questions related to its impact. For every design principle we will ask different questions. The redesign ideas that are additionally mentioned by auditors, will be added in the exploratory part of this research.

Start Question: Do you understand the 4 dimensions?
General Question: As an auditor, can you put these elements in an order that indicates which dimension you attach the most value to in a journal entry test (X=Most, Much, Lower, Lowest):
X: Quality
X: Time
X: Cost
X: Flexibility

**Design Principle 1: IT Advisory Involvement**

**Quality**
The new situation changes the number of contact points between auditors and IT Advisors 1 2 3 4 5 6 7

Why?

**Time**
The new situation changes the time it takes to conduct a journal entry test 1 2 3 4 5 6 7

Time vs Quality
I am prepared to invest time in a journal entry test if it leads to a higher quality 1 2 3 4 5 6 7

Why?

**Design Principle 2: Control Addition**

**Quality**
The new situation changes the understanding of financial transactions of a client 1 2 3 4 5 6 7

Why?

**Time**
The new situation changes the time it takes to conduct a journal entry test 1 2 3 4 5 6 7

Time vs Quality
I am prepared to invest time in a journal entry test if it leads to a higher quality 1 2 3 4 5 6 7

Why?

**Design Principle 3: Internal Control**

**Quality**
The new situation changes the number of implications for internal controls we suggest 1 2 3 4 5 6 7

Why?

**Time**
The new situation changes the time it takes to conduct a journal entry test 1 2 3 4 5 6 7

Time vs Quality
I am prepared to invest time in a journal entry test if it leads to a higher quality 1 2 3 4 5 6 7

Why?

**Design Principle 4: Effective Sampling**
Quality
The new situation changes the amount of different journal entries that can be analyzed 1 2 3 4 5 6 7
Why?

The new situation changes the amount of controls we put on journal entries 1 2 3 4 5 6 7
Why?

Time
The new situation changes the time it takes to conduct a journal entry test 1 2 3 4 5 6 7
Why?
Time vs Quality
I am prepared to invest time in a journal entry test if it leads to a higher quality 1 2 3 4 5 6 7
Why?

Table 20: List of auditors with whom interviews have been held

<table>
<thead>
<tr>
<th>Auditor</th>
<th>Function</th>
<th>Audit Experience</th>
<th>Openness towards Data Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>A I</td>
<td>Supervisor</td>
<td>6 years</td>
<td>Since recently in an active sense. Before, Excel analyses</td>
</tr>
<tr>
<td>A II</td>
<td>Junior Trainee</td>
<td>2 years</td>
<td>Operationally applying data analytics in journal entry test</td>
</tr>
<tr>
<td>B I</td>
<td>Supervisor</td>
<td>6 years</td>
<td>The auditor is open for the current data analytic initiative which would make auditors work different, but also more efficient.</td>
</tr>
<tr>
<td>B II</td>
<td>Senior</td>
<td>3 years</td>
<td>The auditor has conducted some analytics and understands data and analytics as the way things are heading</td>
</tr>
<tr>
<td>C I</td>
<td>Supervisor</td>
<td>5 years</td>
<td>The auditor is interested and already involved in applying some basic form of data analytics, but sees opportunities for improvements.</td>
</tr>
<tr>
<td>C II</td>
<td>Assistant Manager</td>
<td>8 years</td>
<td>The auditor is active as a data analytics motivator within the CPA. This leads to actively encouraging auditors to utilize it in their audit. In terms of applying it, he often had a different party to discuss extra analysis on top off his current thoughts as a way to add control during the risk assessment.</td>
</tr>
</tbody>
</table>

Validating Profile Selection

Table 21: Summary on interviews regarding: “Validating Profile Selection”

<table>
<thead>
<tr>
<th>Case-Auditor</th>
<th>Summary</th>
<th>Quotes</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>A I</td>
<td>The auditor indicates that profiling clusters on behalf of the explanation name, is something he would do himself as well.</td>
<td>“Ja”</td>
<td></td>
</tr>
<tr>
<td>A II</td>
<td>The auditor indicates that profiling clusters on behalf of the explanation name, is something he would do himself as well.</td>
<td>“Lijkt me wel de meest logische manier.”</td>
<td></td>
</tr>
<tr>
<td>B I</td>
<td>The auditor indicates that profiling clusters on behalf of the explanation name, is something he would do himself as well.</td>
<td>“Ja”</td>
<td></td>
</tr>
<tr>
<td>B II</td>
<td>The auditor indicates that profiling clusters on behalf of the explanation name, is something he would do himself as well. The auditor indicates that on a granular level, the profiles could be named according to the explanation in the</td>
<td>“.. ik zou toch voornamelijk naar account niveau kijken, want dat is hetgeen dat op grootboekniveau wordt geboekt. Pas als ik naar frauderisico zou kijken dan zou je kunnen gaan kijken naar omschrijving”</td>
<td></td>
</tr>
</tbody>
</table>
journal entry. This is namely the element that would be utilized when fraud should be detected. All other ways of determining whether they are a profile is the cluster criteria already taken into account.

C

The auditor indicates that profiling clusters on behalf of the explanation name, is something he would do himself as well.

“Wat ik zelf altijd doe is kijken naar welke grootboekrekening wordt gebruikt, waarin je telkens een stapje dieper gaat (Step 1). Dan is de beschrijving het laatste dingetje waar ik op terug val (Step 2).”

CII

The auditor indicates that the pivot table involving the accounts, debit/credit and net amount is something he would do to indicate the coherence of the journal entries. He does not go into depth on the explanation attribute as a way to label the cluster.

“als ik hier een pivot op zou draaien, dan zou ik weten, van is dat altemaal ongeveer dezelfde boeking die wordt gemaakt. Als ik honderd verschillende rekeningen krijg, dan is het niet een soort boeking die wordt gemaakt, als het 5 verschillende rekeningen zijn dan krijg ik wel vertrouwen dat het goed gaat.”

Not mentioned as a way to “name” the cluster.

Bucket of Intra-Category Bookings

Table 22: Summary on interviews regarding: “Bucket of Intra-Category Bookings”

<table>
<thead>
<tr>
<th>Case - Auditor</th>
<th>Summary</th>
<th>Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>A bucket like intra-category transactions are often low-risk, since they have no impact on financial ratios used by users of the annual report to make decisions, however there are exceptions on this rule.</td>
<td>“Als iets binnen één categorie valt, zit er meestal minder risico aan.” “Ratio’s worden niet beïnvloed als er binnen één zo’n account geboekt wordt.” “Er zijn wat uitzonderingen op de regel.”</td>
</tr>
<tr>
<td>AII</td>
<td>The auditor sees less risk in buckets, but sounds doubtful when looking at the risk level/feasibility.</td>
<td>“Dat is wel, ik zou zeggen, A1 hebben minder impact, en ga je dus niet controleren.”</td>
</tr>
<tr>
<td>BI</td>
<td>The auditor thinks that an intra-category transaction could be interesting to know, since it is part of ‘understanding the entity’ to know if this bucket has a low or high-risk. Especially during a first year’s audit, this information could facilitate the risk-assessment.</td>
<td>“Aan de ene kant zou ik zeggen: Als het geen impact heeft, heeft het geen impact. Aan de andere kant van het zelfde geldt de debet boeking is goed en de credit had niet gemoeten. Maar hij had niet nul moeten zijn maar gelijk aan het credit bedrag. Dit is wel volledig afhankelijk van de aard van de boeking.” “ik denk dat het gewoon heel belangrijk is “understanding the entity”, hoe gaan die boekingsgangen bij een klant. En ik denk dat het daarom voor het eerste jaar goed is om te kijken of die boekingen die je doet met een klant te bespreken.”</td>
</tr>
<tr>
<td>BII</td>
<td>The auditor indicates that there is less risk in a bucket of intra-category bookings, especially when having in mind the goal of detecting management override of controls.</td>
<td>“...het zijn eigenlijk manual entries, maar ik zie daar geen groot risico in.” “En juist ook niet in het kader van het fraude-aspect.”</td>
</tr>
</tbody>
</table>
| C1             | The auditor thinks there is value in making buckets of intra-category transactions which have a risk-label that is lower than other transactions, and this could reduce the dataset you will take into account. | “Ik denk wel dat het belangrijk is om even te weten: Wat gebeurd er nou. Is het logisch wat er gebeurd? Is het logisch wat er gebeurd, maar als je denkt zit er een heel groot risico aan? Dan denk
The auditor thinks that an intra-category booking has minimal risk, which he implicitly sees as adding value in the risk assessment phase by introducing experience he has with a different tool.

**Automatic Bucketing Exploration**

Table 23: Summary on interviews regarding: “Automatic Bucketing Exploration”

<table>
<thead>
<tr>
<th>Case</th>
<th>Summary</th>
<th>Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>This auditor indicates that decisions within journal entry tests are, implicitly, subject to buckets. Clarifying these buckets in terms of risks is deemed interesting to find buckets that should be analyzed for sure.</td>
<td>“[een bucket] zou je op de voorkant kunnen afvangen, mits je goed rekening houdt met de uitzonderingen die erin zitten.” “Wij kunnen op totaalniveau controleren of [routinematige transacties] juist zijn, dan zit er ook weinig risico in de boekingen die er nog onder liggen. In die zin is zo’n onderscheid wel handig.”</td>
</tr>
<tr>
<td>AII</td>
<td>The auditor sees value in bucketing, but thinks of this process as not to be placed at the end of the audit (manual journal entry test), but more at the beginning to gain insight in what journal entries to take into account.</td>
<td>“maar ik denk wel, dat het journal entry testing proces an sich wel zeker kan bijdragen aan de audit. Ik vind het een goede insteek. We zijn vaak geneigd om te veel te doen, ik denk door de insteek die jij hebt gekozen, efficiënter kan werken”</td>
</tr>
<tr>
<td>B1</td>
<td>The auditor sees value in the automatic bucketing, where he thinks that this is currently done, but on an inefficient way, namely at the same time as the layering process, in which you would expect this before.</td>
<td>“Ik denk dat het goed is. Ik denk dat het voor een deel al gebeurd, maar op een inefficiënte manier. Dat wij wel elk jaar zeggen, vanuit de hele populatie gaan en dat we een risicovolle populatie selecteren waarop we denken, hmm deze zijn misschien niet risicovol, dat snappen we, daar gaan we verder niets meer aan doen. Terwijl je eigenlijk er uit moet filteren voordat je je analyse gaat doen.”</td>
</tr>
<tr>
<td>BII</td>
<td>The auditor indicates that it is to abstract at the moment, in terms of knowing what could be received from the dataset. But the initiative in essence could be interesting. In terms of covering fraud, one should be aware of the fraudulent behavior that might appear in periodic transactions.</td>
<td>“Ik weet niet wat er uitkomt, dat is voor mij lastig in te schatten. Ik denk dat dit soort initiatieven altijd welkom zijn, altijd goed om uit te voeren.” “waar je het ook voor doet, is voor je fraude risico of af te dekken, bij de manual journal entries. En daarvan is het denk ik juist dat wekelijkse of periodieke transacties juist ook een heel hoog fraude-risico hebben”</td>
</tr>
<tr>
<td>C1</td>
<td>The auditors sees value in bucketing since you explicitly label journal entries in relation to their risk level which impacts the relevant population.</td>
<td>“[waarde als] .. standaard processen eruit halen, en je processen eruitslopen en dan hou je de manual of bijzondere entries over”</td>
</tr>
<tr>
<td>CII</td>
<td>The auditor sees a lot of value in automatic bucketing. The idea of bucketing itself is, sort of, executed in a sense that auditors explain buckets</td>
<td>“Kijk, je doet het [implicit buckets maken] wel, maar je gaat ze wel nog steeds allemaal af”</td>
</tr>
</tbody>
</table>
by using the knowledge of previous years. This actually results from inefficiently going through the journal entries all over again. An automatic way of detecting buckets would therefore be more efficient. He would place this before the testing takes place and dependent on the size of the client this could even be during an interim phase.

"als je de boekingen bekijkt: heel veel dingen herken je nog van voorgaande jaren, en dan zeg je: ja die ken ik. Dat deden ze vorig jaar ook."

"op het begin, de hele populatie knal je erin. Dan ga je ze in hokjes stoppen en daar wil je eigenlijk zeggen, deze hokjes ken ik en die vind ik normaal, en die hokjes vind ik wel spannend en die ga ik testen. Dus eigenlijk voordat je gaat testen, dus eigenlijk bij een grote klant eerder dan jaareinde, maar anders gewoon per jaareinde."

Profile Analysis

Table 24: Summary on interviews regarding: “Profile Analysis”

<table>
<thead>
<tr>
<th>Case - Auditor</th>
<th>Summary</th>
<th>Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>The auditor actively links the profiles found to experiences he has at the Case Company. Responses to risks related to buckets differ from low-risks related to explanation name changes, low-risks to an initial incorporation of a single category, to higher risks when over the years a pattern is not as expected to be. The pattern and continuity within an audit, is something which is conducted manually at places in an audit, which he sees as an idea for automation.</td>
<td>“Maar het is op het eerste gezicht vreemd, dus als ik deze informatie zou hebben, dan is dat iets wat ik zou navragen, zou onderzoeken.” “Het zou bijvoorbeeld wel gek zijn, als ze een bepaalde maand, geen xxxkosten hebben geboekt.”</td>
</tr>
<tr>
<td>AII</td>
<td>The auditor has trouble identifying transactions controlled, as him having less experience. He looks at clusters in terms of riskiness of profiles and tries to link it to his knowledge of the client. Often it is indicated that “knowing this, would have influenced the journal entries we analyzed”</td>
<td>“Misschien wel interessant om het eruit te laten springen als journaal entry testing. In ieder geval, de populatie die je gaat testen, dat zou wel handig zijn. Want het is inderdaad, 1x in de 2 jaar, dat is wel iets om op te letten, het zal wel een hele logische onderbouwing hebben.”</td>
</tr>
<tr>
<td>B1</td>
<td>The auditor is an active advocate of incorporating the knowledge of changing patterns in accounting processes of an auditee. In terms of determining the relevant population, he indicates that this should be done more strict and scoped towards “actual” incidental manual journal entries. As an extra thought, he would like to include as a control addition, that if a previously marked high-risk journal entry has occurred this year.</td>
<td>“Wat je bij ons wel eens ziet, wat moeten we nu precies bij het journal entry testing doen, doe je nou genoeg? Of doe je te veel? Dit geeft wel heel gericht aan, zie ik nu hele gekke dingen.” “Dit [een minder frequent patroon] lijkt er gewoon op te duiden dat ze een andere manier van boeken hebben. Ook hier weer, ik vind dat dit zeker een vraag waard is.” “Dat iets met een bepaalde code wordt geboekt, of wat wij zien als een manual journal entry in hun systeem JD Edwards, dat dat omschrijving heeft manual journal entry, dat kan, maar uiteindelijk gaat het erom, of dat het een procesmatige boeking is ja of nee.” “waar we ons wel in zouden kunnen verbeteren is ook vast te leggen dat de high-risk entry die we vorig jaar hadden, dat die wel of niet dit jaar terugkomen.”</td>
</tr>
<tr>
<td>BII</td>
<td>The auditor thinks that applying data analytics would lead to insights during the planning stage of an audit that will increasingly more be incorporated.</td>
<td>“[analyses op de data] Ik denk dit is wel de toekomst van het controleren.”</td>
</tr>
</tbody>
</table>
The auditor indicates that it is important to know when a transaction is not occurring anymore. The auditor sees value in providing profiles and comparing them over the years, but these understandings should be presented earlier in the process than in journal entry testing. Explorations like, routine transactions that stopped occurring and outliers are very interesting. The fact that description names change has not impact on their business understanding.

“[als iets het jaar erna niet terug komt] Ik denk dat de klant moet uitleggen waarom ze deze boeking wel in 2014 maken en niet in 2015.” “…er zit een wijziging in het proces, daar moeten we iets mee. Daar willen we zeker eerder in de controle adresseren dan in het journal entry testing.” “Het feit dat ze november of december niet noemen vind ik totaal niet spannend.”

The auditor focuses on journal entries that stopped occurring over the years or throughout a time period since these might be interesting to add to your control. This is however only interesting/efficient, if it requires on the most biggest journal entries in terms of amount.

“ik heb vorig jaar zo’n staartje gezien. En toen zag ik iets heel gek, en dat was een van die dingen die ik uit heb laten zoeken. Het is dus altijd nuttig om inzichtelijk te krijgen” “Ik vind de risico meer als ze iets nieuws ineens gaan boeken. Maar vaak zijn het incidenten.” “bij de financiële administratie hebben ze een lijstje met alle boekingen die ze maandelijks terugkrijgen van memoriaal boekingen.” “Kijk, dat boekingen er niet meer zijn, is wel handig om te weten, voor grote stromen” 

Devil’s Quadrangle

Table 25: Summary on interviews regarding: “Devil’s Quadrangle ”

<table>
<thead>
<tr>
<th>Case - Auditor</th>
<th>Summary</th>
<th>Setup</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>The auditor indicates that quality and the ability to react to changes in a client’s accounting system are of most important interest by an auditor. The impact on time and costs are subservient to these. This can also be linked to this auditor’s openness to data analyses.</td>
<td>1. Quality/Flexibility</td>
</tr>
<tr>
<td>AII</td>
<td>The auditor indicates that he attaches most value to quality. Then flexibility. He sees time and costs as coherent, but needing to make a decision, as an auditor he thinks time is a bit more important to him.</td>
<td>1. Quality</td>
</tr>
<tr>
<td>B1</td>
<td>The auditor indicates that he attaches most value to quality, and in a similar sense he argues flexibility to be related. Time is a bit more important to him that the costs for the CPA.</td>
<td>1. Quality/Flexibility</td>
</tr>
<tr>
<td>BII</td>
<td>The auditor indicates that quality is most important, and should be seen as questioning the ‘right’ entries. Then, flexibility and thereafter costs. Time, he sees as something that should be of minimal impact, since knowing what you want to ask the client enables you to ask the right questions in limited time.</td>
<td>1. Quality</td>
</tr>
<tr>
<td>C1</td>
<td>The auditor thinks quality/flexibility is most important. However, a redesign or tool should not aim at providing analyses with the purpose of</td>
<td>1. Quality/Flexibility</td>
</tr>
</tbody>
</table>
providing analyses. It should provide added value. He would invest more time for more quality, but with something like automatic bucketing he thinks that this will decrease the time required in the long term.

CII

The auditor thinks that quality should be good. The impact of the lead time is dependent on the client’s accounting environment. Although costs are important, quality should be leading. He has trouble placing time and flexibility because they are too dependent on the context.

1. Quality
2. Cost
3. Time & Flexibility

IT Advisory Involvement

Table 26: Summary on interviews regarding: “IT Advisory Involvement”

<table>
<thead>
<tr>
<th>Case - Auditor</th>
<th>Summary</th>
<th>Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>The auditor thinks that involving IT advisory and data analysis in journal entry testing would lead to additional client satisfaction and perhaps extra advisory projects.</td>
<td>“ik denk dat je daar met name voordeel bij haalt op het gebied van klanttevredenheid en het scoren van extra adviesopdrachten bij klanten.”</td>
</tr>
</tbody>
</table>
| AII | The auditor was linking the idea of automatic bucketing to the current situation, involving the current restrictions that might arise, which made him skeptic in terms of actual contact points that would arise. But with regard to possibilities, he thought that it would benefit the audit. | “maar als je kijkt naar de realiteit is de journal entry test, vooral bij de national practice, dat het vooral een verantwoordelijkheid blijft van de auditor zelf en niet van de IT auditor.”
“Vooral bij de kleinere klanten zal de toegevoegde waarde van ITA minimaal zijn.”
“ik denk inderdaad als het gratis en voor niks is he handig, ik zou het toch aannemen als controle” |
| B | The auditor sees value in involving ITA more in the audit process, and thinks that, if this involvement is structured in the right way, automatic bucketing could lead to such a change. He values communication in this way, since this is the enable of involvement. | “…iemand die bij ITA werkt, heeft geen accounting achtergrond. Terwijl om te kunnen snappen, waar je mee bezig bent, is wel van belang om een accounting achtergrond te hebben. Andersom precies hetzelfde. Willen wij snappen wat ITA aan het doen is moeten we eigenlijk ook een technische achtergrond hebben. Deze situatie is niet, vandaar dat communicatie ook belangrijk is.”
“Ik denk dus dat in een theoretische situatie dat dat de kwaliteit ten goede komt, zolang je organiseert, en dat praktisch gezien er eerst randvoorwaarden moeten komen om ook de kwaliteit te kunnen verhogen.”
“daar is weer van belang dat je blijft communiceren, blijft overleggen van oke, dat ITA de auditor mee in het process neemt van ok wat hebben we nu gedaan. En dat het niet wordt van, hier komt een rapportje en van, volg maar op en klaar. Dus ik denk dat zeker zo’n bucketing daarbij kan helpen” |
| BII | The auditor is skeptic on the influence auditors could have on other places. The different languages that are spoken is quite a barrier to break. It might be worth a try if an IT advisor will be linked to a specific company to give specific analyses. | “ITA in mijn ogen heel weinig verstand heeft van boekingen, als ik het met ITA erover heb, krijg ik vaak de vraag: Hoe werkt het dat nu debit en credit, en dan denk ik van ja: er is een lange weg te gaan voordat ITA ons inhoudelijk kan helpen en inhoudelijk een paar slagen kan maken voor het journal entry testing”
“Weet je, ik vind het belangrijk dat ITA zoveel mogelijk aanhaakt bij het gewone audit team. In hoeverre zij ook daadwerkelijk inhoudelijk bijdrage leveren aan de audit weet ik niet.” |
“Belangrijk is gewoon dat de juiste analyses worden uitgevoerd”
“vooral als je het on top off doet, waarbij je een aanvullende analyse door ITA op deze manier laat doen, zou ik alleen maar goed vinden”

CI The auditor thinks that involving ITA would lead to an increased communication in terms of Quality, which impact is mostly on getting more effective information.

“Ik denk met name de inhoud. Ik denk het aantal dat dat meevalt.”
“Wat ik denk in het eerste jaar zal het meer tijd kosten, in de jaren daarna is het steeds minder. Ik denk dat iedereen daartoe bereid is. Of in ieder geval zou moeten zijn, ik ben er in ieder geval zeker toe bereid. Maar we ook niet moeten vergeten, dat het een piek weghaalt per jaarreinde en de tijd die je jaarreinde besteedt verdeelt over de hele controle. In the end het zou meer tijd kosten, en daar ben ik bereid toe.”

CII The auditor thinks that thinking along could arise, but before any positive interdependence could arise audit and IT advisory should be more connected. Which requires a time investments on both sides.

“Dat zou heel erg een combinatie moeten zijn van ITA en Audit, anders krijg je dat nooit van elkaar. Audit kan het niet doen en ITA moet heel goed weten wat wij precies aan het doen zijn om dat goed te kunnen onderwerpen.”
“[automatic bucketing] Vanuit dat idee denk ik dat je best wel ver kan komen. Dat je er echt stoppen in kan maken. Daar zitten momenteel heel veel IT’ers achter. Ik denk wel dat er vanuit de audit ook tijd in moet worden gestopt.”
“De communication is echt heel erg slecht. Dat kan zeker helpen, maar wat we bijvoorbeeld bij een andere grote opdracht doen….., twee weken gaat iemand van ons met hun meelopen en wij krijgen twee weken iemand van ITA in het team…. ik denk dat dat soort uitwisselingen erg waardevol zullen zijn”

Internal Control Implications

Table 27: Summary on interviews regarding: “Internal Control Implications”

<table>
<thead>
<tr>
<th>Case - Auditor</th>
<th>Summary</th>
<th>Quotes</th>
</tr>
</thead>
</table>
| AI             | The auditor thinks that an implication like internal control could increase the quality which is worth the time investment. | “Kijk als je heel goed inzichtelijk hebt welke stromen er precies allemaal lopen, is daarna ook makkelijk een stap die je kunt maken, welke controls zou je hier allemaal op verwachten.”
“[extra contact met ITA] als me dat iets oplevert, waar ik ook mee terug naar de klant kan; zeker.” |
| AII            | The auditor thinks that advising a client on the possibilities for internal control improvements when a dataset has been seen as standard could be efficient. | “Het kan geïdentificeerd worden, maar er is vaak een makkelijke reden voor. Bijvoorbeeld bij de huurcontracten. Specifiek voor de huurkosten vind ik het niets toevoegen, want dit is een standaard boeking.”
“we halen hem eruit, we nemen hem niet mee in de controle, dan ben ik het met je eens.” |
The auditor indicates that automatic bucketing would lead to more discussion with the client, in which the definition of manual journal entry will be an important topic of conversation.

“Opzich wel, ik denk dat daarbij wel van belang is dat wij als accountant manuele boekingen, de definitie van de manuele journal entries gelijk stellen aan die van de klant.... want als er in het manuele journal entries procesmatig eigenlijk gebeurd, dan is het lastig, dan heeft het weinig toegevoegde waarde.”

At the moment, auditors do not actively advise the client on possible improvements. This is linked to the control risks within the entity which is again linked to the maturity of the systems used.

“het feit dat een wekelijkse of maandelijkse controle is hoeft niet te betekenen dat het geen manual control meer moet zijn,”
“Het hangt van het interne controle systeem af in hoeverre die manuele boekingen of daar een review op plaats vindt.”

Although advising clients on controls is not an auditor’s main business, it would lead to a better understanding when the existence of control effectivity/existence are discussed.

“Nou ja, de vraag of ze er meer controle op kunnen zetten is niet iets van ons. Wat wij doen is dat wij gewoon signaleren, je hebt geen controle en het is aan de klant om een overweging te maken om daar een controle op te zetten of niet.”

The auditor thinks that automatic bucketing would actively lead to advising the client on anomalous behavior, since this is an activity that is mutually beneficial. He is, however, skeptical on the term this would actively take place, this might take some years to develop (over the engagement), but small wins are oftentimes possible.

“wat we eigenlijk doen is echt op uitzonderingen en principe risk of management override. Dat is het enige dat wij moeten doen. Dus ja, als je daar kunt zeggen: Deze boekingen vind ik normaal. Dat vind ik juist heel goed, daar denk ik dat we naar toe moeten.”
“Het idee is geloof ik ook dat we dat bij de klant neer gaan leggen van kijk dit loopt allemaal normaal en dit niet. Doe er eens iets aan, want dit kost jullie bakken met geld. Want bij een klant is het ook, alles wat anders loopt dan normaal kost manuele handelingen. Dus ja ik denk zeker dat je de klant daarmee iets kan leveren, en ons zelf dan ook weer.”

### Data-Analytical Mindset

<table>
<thead>
<tr>
<th>Case - Auditor</th>
<th>Summary</th>
<th>Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>The auditor thinks that data analytics could lead to understanding the client which directs journal entry testing more specifically to the most interesting cases.</td>
<td>“Ja ik denk inderdaad het heeft gewoon nodig dat je al een aantal jaren de klant kent, als dat zo is, dan kun je veel preciezer de high-risk journal entries bepalen, wat leidt tot, verwacht ik, minder boekingen hoeft te zien.”</td>
</tr>
<tr>
<td>AII</td>
<td>The auditor indicates that the value of data analysis for small clients is less of added value. But besides the practicalities, he thinks that getting an increased understanding of the data, adds value to journal entry testing which is worth the investment.</td>
<td>“Maar ik denk vooral de grotere klanten, de corporates, daar is heel veel toegevoegde waarde zijn. De national practice, mkb, daar zijn de processen redelijk eenvoudig. Dus het ligt aan welke klanten je bekijs.” “Meer kennis he: Meten is weten zeggen ze”</td>
</tr>
</tbody>
</table>
The auditor indicates that he believes that automatic bucketing leads to interesting insights, but it only leads to efficiencies, if these insights lead to less “work” on other places during the audit process in the long term.

"Zo’n analyse geeft heel veel toegevoegde waarde, maar als jij meer energie steekt in het IT gedeelte, of de data analyse, betekent dat aan de andere kant, gegevensgericht, binnen onze werkzaamheden dingen vervallen, want anders ben je alleen maar meer aan het doen en alleen maar de kwaliteit aan het verhogen, of in ieder geval meer informatie in je dossier aan het stoppen, waarbij in the end je resultaten naar beneden gaan, en dat blijft toch altijd wel een onderwerp van gesprek, dat je ook je budget gehaald worden.”

"Op de korte termijn gaat je dit waarschijnlijk niet lukken en zul je extra energie erin moeten steken.”

The auditor thinks that automatic bucketing could lead to increased understanding of the data, in which if it actually adds value, time should be spend.

"Het ligt er een beetje aan hoeveel tijd. Alles wat kwaliteit toevoegt aan een audit, daar moet je tijd voor maken. Dus hiervoor ook.”

"ik zou nog altijd de afweging maken, van vind ik het interessant ja of nee.”

The auditor thinks that analyzing the data by means of getting a better understanding of the client would lead to an initial time investment and a small update every year, since a client does not change its entire transactional atmosphere at once.

"Ja, dat is eenmalig tijd erin steken en dan is het ieder jaar updaten. De klant verandert niet in een keer zijn hele boekingsgang.”

The auditor thinks that quality improvement in terms of data analytics should be done, in which the quality it improves is more important than the time investment on the short term. On the long term efficiency gains should be incorporated.

"Met zijn allen moeten we gewoon een kwaliteitsslag maken, dat moet gewoon nog. Aan de andere kant hebben we een zwaar tijd tekort allemaal, maar ik denk dat je eerst gewoon het perfecter moet gaan doen, daarna kun je kijken wat leren we ervan en hoe kunnen we het efficiënter doen ik denk wel dat we daarin moeten investeren.”

Effective Sampling

Table 29: Summary on interviews regarding: “Effective Sampling”

<table>
<thead>
<tr>
<th>Case - Auditor</th>
<th>Summary</th>
<th>Quotes</th>
</tr>
</thead>
</table>
| A1            | The auditor thinks that effective sampling has multiple applications that lead to a decrease in time, since less journal entries have to be taken into account. | “Ik zou ook voorstellen dat je niet elk jaar zo’n bucket hoeft te testen. Als je zo’n bucket eenmaal gezien hebt, dan zou je het jaar daarna misschien alleen de nieuwe te selecteren.”
|               |         | “dit zou er toe kunnen leiden dat je je high-risk criteria aanpast.” |
| A11           | The auditor strongly agrees that some transactions should not be taken into account in the testing process, even if they are manual journal entries, which is related to an auditor’s ability to make the risk criteria as precise as possible. Upfront, during the risk assessment, it should be known which decisions are to be taken and this is pretty | “...ik zou er geen uit pakken”
|               |         | “alles moet integraal getest worden. Ik vind dan [als herkenbare manuele journal entries mee zijn genomen] is er aan de voorkant iets niet goed gegaan”
|               |         | “...ik zou er nul meenemen. [en zou je dan in combinatie met de control willen werken?] en als het afwijkt zou ik dat [een manuele journal entry van de bucket] eruit willen halen” |
much related to your knowledge of the audit process.

B1 | The auditor indicates that whether sampling will/can be applied is dependent of the population of journal entries in terms of size, amount and other elements, which makes this decisions also dependent on the kind of bucket we are talking about. Having the ability to do so could be interesting and can therefore make the process flexible.

In terms of control internally for the CPA, it could be effective, but only if the definition for “green light” is consistent.

BII | The auditor sees efficiency possibilities for effective sampling, however if a control on this bucket is effective, there is no need to element of effective sampling.

The auditor thinks that control might be efficient, but this is highly dependent on the setup of the control, which also determines the time savings related to it.

C1 | The auditor sees possibilities for effective sampling, however he leans towards a situation in which only if extreme values arise, this kind of sampling might arise, but if the control is effective, this might be redundant.

The auditor sees the ability to use insights of cumulative audit evidence as a way of do very specific data oriented procedures. Which he is interested to take extra effort in.

| “Volledig afhankelijk van de omvang van de populatie. Waarom kiezen we vaak voor sampling? Omdat je met relatief weinig selectie items een grote populatie kunt controleren, dus dat is wel, alleen dan zou het efficiënt kunnen zijn, de vraag is van ja, als het echt een homogene populatie is en het gaat om een grote populatie, hoe meer we samplen hoe beter…. het gaat bij grote populatie, zijn factoren die ik meenemen, wat is de materialiteit en wat is de omvang van de individuele transacties”
| “..ik vind het lastig om te bepalen, ook daar weer is het van belang, over wat voor populatie spreek ik…. dan zou ik best durven te zeggen daar doen we verder niets mee. Maar dat hangt van heel veel verschillende factoren af.”
| “Ik denk dat dat wel goed zou kunnen zijn, waarbij het dan wel heel goed van belang is, dat je vooraf duidelijk definieert, wat dan de kenmerken zijn van de transacties die groen worden markeerd”
| “Daar zie ik wel mogelijkheden voor. Het moet alleen nog onderbouwd worden. Waarom maak je bepaalde keuzes, waarom vinden we dat wel of geen risico. En waarom vinden we de werkzaamheden die we eraan hebben gedaan voldoende om het risico af te dekken.”
| “Denk het wel, als je minder hoeft te testen levert het waarschijnlijk tijd op. Het kost misschien niet extreem veel tijd op dit moment, maar het kan wel tijd opleveren”
| “ik sta altijd open voor dit soort analyses, en dan moet je altijd samen bekijken, dan zouden we er kritisch naar moeten kijken en zeggen is dit echt een beperkt risico of niet.”
| “En of dat dan uiteindelijk minder tijd gaat opleveren, dat weet ik niet.”
| “[internal control vs sampling] Ik vind, als je er op kunt steunen, dan moet je er op steunen. Dat is puur angst van, ik ga er toch maar 1 testen, dat heeft dan eigenlijk in mijn ogen geen zin.”
| “dat we [de definitie van high-risk omzetten naar een cluster] we één cluster aan entries of daar maar 1 van controleren, dan [mogelijkheden voor effective sampling] zeker denk ik.”
| “Tenzij ik zeg dat een high-risk entry is geen high-risk als ik hem vorig jaar ook al heb gezien”
| “[journal entries buiten eigen cumulatief audit evidence] die het niet gaat halen en daar ga ik specifiek gegevensgerichte werkzaamheden op doen.”
“Kijk, audit verandert continu he. Wat we vorig jaar deden was nu al weer oud, dus we stoppen continu tijd in het ontwikkelen van een controle, dus ja zeker.”

The auditor indicates that during the layering phase, which is used to set up the high-risk criteria, understanding more of the client would lead to more strict high-risk criteria which would decrease the time required for testing.
The auditor thinks that leaving out buckets of journal entries that are known to be normal, could and would be good for the future of journal entry testing.

“We moeten de high-risk entries noteren, maar dat doen we op basis van een bakboek die best wel groot is, en daarop gaan we filteren. En daarna moeten we ze inderdaad allemaal pakken. Die binnen je filters passen.”

“De filter zetten we meestal best wel krap, dat we niet heel veel overhouden [omdat je als auditor voldoende van de data weet].”

“maar voordat je dat bereikt hebt en je met de klant weet ik veel hoeveel jaren verder bent, misschien niet meteen. Aan de andere kant, als je de klant heel veel toegevoegde waarde gaat leveren op zo’n, manier, dan krijg je er ook gewoon voor betaald, is het ook weer goed”

Other Remarks

Table 30: Summary on interviews regarding: “Other Remarks”

<table>
<thead>
<tr>
<th>Case - Auditor</th>
<th>Summary</th>
<th>Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AII</td>
<td>The auditor sees the advantages of automatic bucketing beyond journal entry testing, and might link it to other places in an audit.</td>
<td>“Ik denk dat dat heel goed is om een risico-inschatting te kunnen doen/maken.” “manual journal entry testing insteken, zit het helemaal aan het einde van het proces. Het zit ook in het dossier helemaal achteraan, het is iets dat we als allerallerlaatst oppakken, terwijl de waarde die me dit biedt, met name zit op het gebied van risico-inschatting. Dat zou je dus eigenlijk helemaal aan het begin van het proces verwachten” “Kijk als je eenmaal het aantal wijzigingen bij een klant hebt, het aantal wijzigingen bij een klant is niet zo heel groot. Het aantal wijzigingen bij een klant, ITA echt aannemen en input ervan verwachten 1x in de zoveel jaren te doen. Dat je het ook tijd geeft dat er genoeg verandert is, om zinvol er iets over te zeggen.” “ik denk dat het heel interessant zal zijn als je hiermee aan nog analyses kunt toevoegen over klanten heen.”</td>
</tr>
<tr>
<td>BII</td>
<td>The auditor thinks that data analytics is a good medium for the CPA in sourcing new clients.</td>
<td>“volgens mij wordt dit ook nog wel eens meegenomen in proposals naar klanten, waarin je aangeeft dat data-analyse hot is, en daarmee kun je jezelf profileren.”</td>
</tr>
<tr>
<td>BIII</td>
<td>The auditor thinks that data analytics is a good medium for the CPA in sourcing new clients.</td>
<td>“Ik denk dat het iets is, als je echt een goede methode hebt, dat je dat ook kunt laten zien en kunt presenteren tijdens je proposal.”</td>
</tr>
<tr>
<td>CI</td>
<td>The auditor gains insights when looking at the data. The auditor thinks automatic bucketing could provide advantages on other places during an audit. The automatic bucketing should be initiated at first year companies to gain the best efficiencies. It provides information which facilitates the meetings with the client.</td>
<td>“Je ziet ook dat ze in 2015 iets gestructureerder zijn geweest dan in 2014.” “Kijk, het eerste jaar wat je doet, alles is nieuw wat je ziet, maar wat ik dan graag zou zien is dat we dan dat in de planningsfase al doen, dat we dan al kijken naar de bucketing. Want dat we dan al gaan kijken, wat zijn dan de issues, waar moeten we op focussen.” “[automatic bucketing] ..zeker in het eerste jaar. Het is veel efficiënter om een controle al in het eerste jaar goed in te richten.” “[momenteel] we stappen blank de controle in” “[de issues] daar moeten we ons op focussen tijdens gesprekken”</td>
</tr>
<tr>
<td>CII</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
## Appendix 11: Accounting Rules

### DEBITS (DR) & CREDITS (CR)

<table>
<thead>
<tr>
<th>ACCOUNT TYPE</th>
<th>ACCOUNT NORM</th>
<th>INCREASE</th>
<th>DECREASE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ASSET</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e. Bank Accounts</td>
<td>DR</td>
<td>DR</td>
<td>CR</td>
</tr>
<tr>
<td>Accounts Receivable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LIABILITY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e. Bank Loan</td>
<td>CR</td>
<td>CR</td>
<td>DR</td>
</tr>
<tr>
<td>Accounts Payable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Government Payables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EQUITY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e. Capital / Draw</td>
<td>CR</td>
<td>CR</td>
<td>DR</td>
</tr>
<tr>
<td>Current &amp; Retained Earnings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>REVENUE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e. Sales Products</td>
<td>CR</td>
<td>CR</td>
<td>DR</td>
</tr>
<tr>
<td><strong>EXPENSES</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e. Productions costs</td>
<td>DR</td>
<td>DR</td>
<td>CR</td>
</tr>
<tr>
<td>Administrative costs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overhead, etc</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**NOTE:** GST on Sales Norm is a credit

GST on Purchases Norm is a debit

---

Figure 49: Basic rules of accounting (retrieved 13-03-2016 from [http://accountrain.com/debit_credit.html](http://accountrain.com/debit_credit.html))
Figure 50: The accounting tree (Argyrou & Andreev, 2011)

Figure 51: Chart of accounts involving manual journal entries in entity A, 2014
Figure 52: Chart of accounts involving manual journal entries in entity A, 2015

Figure 53: Chart of accounts involving manual journal entries in entity B, 2014
Figure 54: Chart of accounts involving manual journal entries in entity B, 2015

Figure 55: Chart of accounts involving manual journal entries in entity C, 2014
Figure 56: Chart of accounts involving manual journal entries in entity C, 2015
Appendix 13: Journal Entry Testing Information Entity A

We only described the attributes and journal entry types taken into account for this research. Other attributes/journal entries have not been of interest to auditors and therefore also not to us.

Table 31: Attributes involved in the set of journal entries in entity A

<table>
<thead>
<tr>
<th>Name</th>
<th>Meaning</th>
<th>Descriptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>DocumentNo</td>
<td>A unique identifier for a journal entry.</td>
<td>Integer (3000001661-3000002115) – 2014</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Integer (3000002085-3000002589) - 2015</td>
</tr>
<tr>
<td>Itm</td>
<td>A value that identifies each line of the journal entry.</td>
<td>Integer (1-156) – 2014</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Integer (1-147) - 2015</td>
</tr>
<tr>
<td>Type</td>
<td>Type indicates what kind of journal entry we are dealing with.</td>
<td>String (“DJ”)</td>
</tr>
<tr>
<td></td>
<td>After inquiries, auditors find out which stand for “Manual”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>journal entry.</td>
<td></td>
</tr>
<tr>
<td>Type Omschrijving</td>
<td>Description of “Type”.</td>
<td>String (“Divers Journaal”)</td>
</tr>
<tr>
<td>Account</td>
<td>Accounts involved (see Appendix 12).</td>
<td>Integer</td>
</tr>
<tr>
<td>Account Omschrijving</td>
<td>Description on the accounts involved, aggregated in chart of accounts</td>
<td>String</td>
</tr>
<tr>
<td></td>
<td>(see Appendix 12).</td>
<td></td>
</tr>
<tr>
<td>User Name</td>
<td>The person that entered the batch.</td>
<td>String (6 unique values) – in 2014</td>
</tr>
<tr>
<td></td>
<td></td>
<td>String (7 unique values) – in 2015</td>
</tr>
<tr>
<td>Assignment</td>
<td>A name that is (manually) provided by the auditor to every JE line,</td>
<td>String (523 unique values) – in 2014</td>
</tr>
<tr>
<td></td>
<td>as a way to understand the reason for booking.</td>
<td>String (559 unique values) – in 2015</td>
</tr>
<tr>
<td>Posting Date</td>
<td>The date that the entry is posted to the ledger.</td>
<td>Date</td>
</tr>
<tr>
<td>GLDay</td>
<td>Custom built Attribute to indicate the day in the month the JE</td>
<td>Integer (1-31)</td>
</tr>
<tr>
<td></td>
<td>has been posted.</td>
<td></td>
</tr>
<tr>
<td>Month</td>
<td>Custom built Attribute to indicate the month in the year the JE</td>
<td>Integer (1-12)</td>
</tr>
<tr>
<td></td>
<td>has been posted.</td>
<td></td>
</tr>
<tr>
<td>LCurr</td>
<td>The currency in which the amount has been posted.</td>
<td>String (“EUR” in all transactions)</td>
</tr>
<tr>
<td>Amount in LC</td>
<td>Enter a value that represents the amount of the transaction.</td>
<td>Real</td>
</tr>
<tr>
<td>RULES14</td>
<td>Custom built Attribute to indicate whether the high-risk criteria</td>
<td>Binary (0-1)</td>
</tr>
<tr>
<td>RULES15</td>
<td>did (1) or did not (0) filter out a JE.</td>
<td></td>
</tr>
<tr>
<td>Cat</td>
<td>Custom built Attribute that categorized accounts based on the</td>
<td>Integer (1-34) – 2014</td>
</tr>
<tr>
<td></td>
<td>Parent&gt;Child connections.</td>
<td>Integer (1-34) – 2015</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Integer (1-429) – 2014</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Integer (1-470) – 2015</td>
</tr>
<tr>
<td>JE</td>
<td>A unique identifier directly related to the DocumentNo.</td>
<td></td>
</tr>
</tbody>
</table>

Relevant Population

2014: All journal entries that involve journal type “DJ” (Divers Journal) – 7045 JE lines
2015: All journal entries that involve journal type “DJ” (Divers Journal) – 8320 JE lines

High-risk Criteria 2014

- Journal Entries involved in Net Account Amounts > Confidential Amount (Professional Judgment)
- Account combinations of specific confidential accounts (Professional Judgment)

High-risk Criteria 2015

- Out of Balance JEs (Business Rule)
- Duplicate JEs (Business Rule)
- Missing JEs (Business Rule)
- JE in the weekend (Business Rule)
- JE on specific dates (Business Rule)
- JE by specific user (Business Rule)
- Account combinations of accounts - confidential (Professional Judgment)
- JE Rounded Amounts (Business Rule)
- JE ending in 999 (Business Rule)
- JE with specific comments (Professional Judgment)
Appendix 14: Journal Entry Testing Information Entity B

We only described the attributes and journal entry types taken into account for this research. Other attributes/journal entries have not been of interest to auditors and therefore also not to us.

Table 32: Attributes involved in the set of journal entries in entity B

<table>
<thead>
<tr>
<th>Name</th>
<th>Meaning</th>
<th>Descriptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>JE Line</td>
<td>A value that identifies each line of the journal entry.</td>
<td>Integer (1-1351) – 2014 Integer (1-454) – 2015</td>
</tr>
<tr>
<td>Do Ty</td>
<td>Type indicates what kind of journal entry we are dealing with. After inquiries, auditors find out which stand for “Manual” journal entry.</td>
<td>String (&quot;AF&quot;, “JE”, “JX”) – 2014 String (&quot;AF&quot;, “JE”, “JH”) – 2015 <strong>&quot;JX&quot; is a new name for “JH”</strong></td>
</tr>
<tr>
<td>Obj Acct</td>
<td>Accounts involved (see Appendix 12).</td>
<td>Integer</td>
</tr>
<tr>
<td>Transaction Originator</td>
<td>The person that entered the batch.</td>
<td>String (10 unique values) – in 2014 String (15 unique values) – in 2015</td>
</tr>
<tr>
<td>USER ID</td>
<td>The person that uploaded the batch.</td>
<td>String (10 unique values) – in 2014 String (9 unique values) – in 2015</td>
</tr>
<tr>
<td>Explanation Alpha Name</td>
<td>A name that is (manually) provided by the auditor to every JE line, as a way to understand the reason for booking.</td>
<td>String (232 unique values) – in 2014 String (239 unique values) – in 2015</td>
</tr>
<tr>
<td>Explanation – Remark -</td>
<td>Additional Information to the Explanation Alpha Name.</td>
<td>String (2313 unique values) – in 2014 String (1602 unique values) – in 2015</td>
</tr>
<tr>
<td>G/L Date</td>
<td>The date that the entry is posted to the ledger.</td>
<td>Date</td>
</tr>
<tr>
<td>GLDay</td>
<td>Custom built Attribute to indicate the day in the month the JE has been posted.</td>
<td>Integer (1-31)</td>
</tr>
<tr>
<td>Month</td>
<td>Custom built Attribute to indicate the month in the year the JE has been posted.</td>
<td>Integer (1-12)</td>
</tr>
<tr>
<td>Cur Cod</td>
<td>The currency in which the amount has been posted.</td>
<td>String (&quot;EUR&quot; in all transactions)</td>
</tr>
<tr>
<td>Amount</td>
<td>Enter a value that represents the amount of the transaction.</td>
<td>Real</td>
</tr>
<tr>
<td>RULES14 RULES15</td>
<td>Custom built Attribute to indicate whether the high-risk criteria did (1) or did not (0) filter out a JE.</td>
<td>Binary (0-1)</td>
</tr>
</tbody>
</table>

**Relevant Population**

2014: All journal entries that involve journal type “AF”, “JE”, “JX” – 10,904 JE lines
2015: All journal entries that involve journal type “AF”, “JE”, “JH” (Divers Journal) – 7,765 JE lines

**High-risk Criteria 2014**

- Journal Entries involved in Net Account Amounts > Confidential Amount *(Professional Judgment)*
- Element of unpredictability: chose 1 item based at random from a high-risk account, to verify that supporting documentation has been documented, as mentioned during the inquiry of the controller *(Professional Judgment)*

**High-risk Criteria 2015**

- JE by specific user *(Business Rule)*
- Journal Entries involved in Net Account Amounts > Confidential Amount *(Professional Judgment)*
Appendix 15: Journal Entry Testing Information Entity C

We only described the attributes and journal entry types taken into account for this research. Other attributes/journal entries have not been of interest to auditors and therefore also not to us.

Table 33: Attributes involved in the set of journal entries in entity C

<table>
<thead>
<tr>
<th>Name</th>
<th>Meaning</th>
<th>Descriptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>JOURNAL_NR</td>
<td>A identifier for a journal entry.</td>
<td>Integer (2-990) – 2014 Integer (0-990) – 2015</td>
</tr>
<tr>
<td>ENTRY_NR</td>
<td>A value that identifies each line of the journal entry.</td>
<td>Integer (90136-3400089) – 2014 Integer (1-34000047) – 2015</td>
</tr>
<tr>
<td>ACC_NR</td>
<td>Accounts involved (see Appendix 12).</td>
<td>Integer</td>
</tr>
<tr>
<td>USERID_LC</td>
<td>The person that uploaded the batch.</td>
<td>Not Provided – in 2014 String (30 unique values) – in 2015</td>
</tr>
<tr>
<td>DESCR</td>
<td>A name that is (manually) provided by the auditor to every JE line, as a way to understand the reason for booking.</td>
<td>String (102,385 unique values) – in 2014 String (102,159 unique values) – in 2015</td>
</tr>
<tr>
<td>DATE_</td>
<td>The date that the entry is posted to the ledger.</td>
<td>Date</td>
</tr>
<tr>
<td>GLDay</td>
<td>Custom built Attribute to indicate the day in the month the JE has been posted.</td>
<td>Integer (1-31)</td>
</tr>
<tr>
<td>Month</td>
<td>Custom built Attribute to indicate the month in the year the JE has been posted.</td>
<td>Integer (1-12)</td>
</tr>
<tr>
<td>CURR_CODE</td>
<td>The currency in which the amount has been posted.</td>
<td>String (&quot;EUR&quot; in all transactions)</td>
</tr>
<tr>
<td>AMT_DEF_CUR</td>
<td>Enter a value that represents the amount of the transaction.</td>
<td>Real</td>
</tr>
<tr>
<td>RULES14 RULES15</td>
<td>Custom built Attribute to indicate whether the high-risk criteria did (1) or did not (0) filter out a JE.</td>
<td>Binary (0-1)</td>
</tr>
</tbody>
</table>

Relevant Population

This journal entry test followed a different approach, in which all journal entries are subjected to a query of High-risk criteria and the JEs with the most hits on these criteria are subject for testing. This implies an efficient layering process, but not necessarily increases the business understanding of an auditor.

2014: All journal entries of the entity – 285,674 JE lines
2015: All journal entries of the entity – 771,223 JE lines

High-risk Criteria 2014

- Professional Judgment and Business Rules were combined as setup for parameters in query

High-risk Criteria 2015

- Professional Judgment and Business Rules were combined as setup for parameters in query
Appendix 16: Guidelines for Data Preparation

Pre-process

2. Label Journal Entries that are considered to be high-risk in a particular year.
   - Develop attributes “RULES14” and “RULES15” in Excel.
3. Label the GLDay and Month the journal entries have taken place.
   - Develop attributes “GLDay” and “Month” in Excel.
4. Exclude journal entries that do not add up to 0 in Excel.
   - Reasons for this are not known, auditors think that standard journal entries are added with ‘manual’ journal entries. This phenomenon does not affect the automatic bucketing, but might be an attention point when perfecting the technique.
5. Exclude journal entries that do not have a personal identifier in Excel.
   - Reasons for this are related to the manual nature of these journal entries.
6. Run “Accounts Summarization” using IDEA Smart Analyzer\(^2\) or Pivot Accounts using Excel.
   - IDEA Smart Analyzer extracts the accounts involved in the journal entries and provide information on the amounts transferred and activated accounts.
   - Reasons for having accounts which are not incorporated in the chart of accounts has no influence on the result of the company, which is a reason for auditors to not attach risk to. Communicating with the client on a recurring base would increase this knowledge.
8. Align JE numbers chronologically in Excel.
   - Sometimes manual journal entries are allowed to be uploaded on different times to the general ledger. If issues arise with journal entries that have more dates involved, it is decided to choose the document number (journal entry number) with the most journal entry transactions. That is the case when adding values, may lead to more than the expected number of transactions.
   - Label accounts with an category. Develop an attribute “Category” in which similar categories over the years should have the same value.
10. Visualize chart of accounts.
    - See Appendix 12 for the Chart of Accounts only involving the accounts involved in manual journal entries.
11. Special operation.
    - Entity C has more transactions than Excel can normally handle without having to be restarted all over again. It is chosen to delete all transactions between and equal to -50€ and 50€. Then all journal entries that are not 0 are eliminated.
12. Label journal entry identifier with unique “JE” starting from 1 every year.
13. Select interesting Attributes for clustering of amounts.
    - We developed two different Attribute set. One includes all category values in both years (cat). The other includes all category values in both years and the attribute “GLDay” (day).
14. Bucket Intra-Category Transactions

\(^2\) IDEA Smart Analyzer is a CAAT used to provide filter queries in a user-friendly environment
To take entity A as an example, we see already an A1 dataset that consists out of journal entries that book transactions between accounts that only occur within one category. In the dataset these are journal entries that have a zero value for all attributes. These intra-category transactions can be seen as a bucket that have a different less-risky behavior. In Chapter 5.5.1 these intra-category transactions are analyzed further. See Appendix 20, Figure 76 for the RapidMiner model that is used to filter these buckets.

15. Bucket double-date journal entries
   - Another bucket could be made by journal entries that have multiple dates in a journal entry. This bucket has a risk level dependent on the entity. It is chosen to not go into depth on this bucket. See Appendix 20, Figure 77 for the RapidMiner model that is used to filter these buckets.

16. Explore inter-category transactions
   - This logically leads to a dataset of inter-category transactions A2. This is input for the profiling phase, in which the homogeneous sub-population of journal entries with routine behavior is identified. In Chapter 5.2.4 the exploration of the A2 dataset has been conducted.

17. Exclude overlapping categories
   - For the actual modeling to take place, only categories that are similar over both years need to be included.

18. Exclude 0 categories in both years
   - Some categories do not have any impact in the remaining dataset, i.e. a 0 booking on all journal entries involved, these should be excluded.

19. Resulting population for modeling
   - Sub-population A2, B2, and C2.

20. Set Parameters for modeling

*Multiple errors arose with preparing the data. This limitation is unrelated to the goal of this research, but future data collection method should gather the data as pure as possible. Entity A could therefore be seen as an examplar dataset, while entity B’s & C’s data is different from normal understanding. This does not mean that journal entry testing has been conducted incorrect, however more structure in gathering the data might facilitate future data-analytics endeavors.

Table 34: Data preparation decisions 2014

<table>
<thead>
<tr>
<th>Steps 2014</th>
<th>Summary Entity A</th>
<th>Summary Entity B</th>
<th>Summary Entity C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Total JE lines</td>
<td>All “DJ” JE lines. Which leaves 7,055 JE lines</td>
<td>All “AF”, “JE”, “JX JE Lines. Which leaves 12,886 JE lines</td>
<td>All JEs (except 901, 902, 903, 904, 905). Which leaves 197,027 JE lines</td>
</tr>
<tr>
<td>2. No. of High-risks</td>
<td>1,163</td>
<td>3,476</td>
<td>10</td>
</tr>
<tr>
<td>11. Special Operation</td>
<td>-</td>
<td>-</td>
<td>Delete all JEs with monetary values between -50€ and 50€, finally delete all incomplete JE lines.</td>
</tr>
<tr>
<td>16. Relevant Population</td>
<td>429 JEs 7,045 JE lines</td>
<td>644 JEs 10,904 JE lines</td>
<td>21,965 JEs 88,045 JE lines</td>
</tr>
<tr>
<td>Steps 2015</td>
<td>Summary Entity A</td>
<td>Summary Entity B</td>
<td>Summary Entity C</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1. Total JE lines</td>
<td>All &quot;DJ” JE lines. Which leaves 8,320 JE lines</td>
<td>All “AF”, “JE”, “JX JE Lines. Which leaves 8,845 JE lines</td>
<td>All JEs (except 901, 902, 903, 904, 905). Which leaves 545,414 JE lines</td>
</tr>
<tr>
<td>2. No. of High-risks</td>
<td>370</td>
<td>37</td>
<td>12</td>
</tr>
<tr>
<td>11. Special Operation</td>
<td>-</td>
<td>-</td>
<td>Delete all JEs with monetary values between -50€ and 50€, afinally delete all incomplete JE lines</td>
</tr>
<tr>
<td>16. Relevant Population</td>
<td>470 JEs 8,320 JE lines</td>
<td>845 JEs 7,765 JE lines</td>
<td>24,324 JEs 93,520 JE lines</td>
</tr>
<tr>
<td>17. Exclude non-overlapping categories</td>
<td>Cat 31</td>
<td>Cat 1, 8, 9, 17, 18, 22, 32, 33</td>
<td>Cat 2, 3, 32, 34, 37</td>
</tr>
<tr>
<td>18. Exclude zero-categories</td>
<td>Cat 11, 14</td>
<td>Cat 5, 35, 38</td>
<td>Cat 10</td>
</tr>
<tr>
<td>19. Resulting Attributes</td>
<td>Cat 1, 2, 3, 4, 5, 7, 8, 9, 10, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 32, 33, 34, GLDay, JE</td>
<td>Cat 4, 6, 7, 10, 11, 12, 13, 14, 15, 16, 19, 20, 21, 23, 24, 25, 26, 27, 28, 31, 34, 37, GLDay, JE</td>
<td>Cat 1, 4, 5, 6, 7, 8, 9, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 35, GLDay, JE</td>
</tr>
</tbody>
</table>
Appendix 17: Exploratory Data Analysis Entity B

Table 36: Information on subdivided journal entries in entity B, 2014

<table>
<thead>
<tr>
<th>Dataset 2014</th>
<th>No. of Transactions</th>
<th>No. of JE Lines</th>
<th>No. of High-risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>698</td>
<td>10,904</td>
<td>3,235</td>
</tr>
<tr>
<td>B1</td>
<td>42</td>
<td>4,374</td>
<td>1,441</td>
</tr>
<tr>
<td>B2</td>
<td>644</td>
<td>3,165</td>
<td>529</td>
</tr>
<tr>
<td>B3</td>
<td>12</td>
<td>3,365</td>
<td>1,265</td>
</tr>
</tbody>
</table>

Table 37: Information on subdivided journal entries in entity B, 2015

<table>
<thead>
<tr>
<th>Dataset 2015</th>
<th>No. of Transactions</th>
<th>No. of JE Lines</th>
<th>No. of High-risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>1,017</td>
<td>7,765</td>
<td>37</td>
</tr>
<tr>
<td>B1</td>
<td>163</td>
<td>5,074</td>
<td>0</td>
</tr>
<tr>
<td>B2</td>
<td>845</td>
<td>2,263</td>
<td>8</td>
</tr>
<tr>
<td>B3</td>
<td>9</td>
<td>428</td>
<td>29</td>
</tr>
</tbody>
</table>

Figure 57: Number of journal entry lines (left) and journal entries (right) posted per month in dataset B2

Figure 58: Total number of manual journal entry lines per category in dataset B2
Figure 59: Line items posted over time in dataset B2

Figure 60: Accumulated amounts posted per category involving manual journal entries in dataset B2
Figure 61: Heatmaps of amounts posted to accounts in dataset B2 on specific weekdays (left: 2014, right: 2015)

<table>
<thead>
<tr>
<th>Category</th>
<th>Sunday</th>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
<th>Thursday</th>
<th>Friday</th>
<th>Saturday</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.2%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.0%</td>
</tr>
<tr>
<td>6</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>7</td>
<td>0.3%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.3%</td>
<td>0.0%</td>
</tr>
<tr>
<td>10</td>
<td>0.4%</td>
<td>0.0%</td>
<td>0.3%</td>
<td>3.2%</td>
<td>4.3%</td>
<td>0.7%</td>
<td>0.0%</td>
</tr>
<tr>
<td>11</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>12</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.0%</td>
</tr>
<tr>
<td>13</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.4%</td>
<td>12.7%</td>
<td>26.1%</td>
<td>2.5%</td>
<td>0.10%</td>
</tr>
<tr>
<td>14</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.4%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>15</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.4%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.3%</td>
<td>0.0%</td>
</tr>
<tr>
<td>16</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.7%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>19</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.8%</td>
<td>9.3%</td>
<td>22.0%</td>
<td>1.8%</td>
<td>1.0%</td>
</tr>
<tr>
<td>20</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.6%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>21</td>
<td>0.3%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.4%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>24</td>
<td>0.4%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.2%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>25</td>
<td>0.3%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.3%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>26</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>27</td>
<td>0.2%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.4%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.0%</td>
</tr>
<tr>
<td>28</td>
<td>0.7%</td>
<td>0.1%</td>
<td>0.5%</td>
<td>0.1%</td>
<td>0.7%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>31</td>
<td>0.3%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>34</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>37</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.2%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Figure 62: Association rules in dataset B2 in 2014

<table>
<thead>
<tr>
<th>No.</th>
<th>Premises</th>
<th>Conclusion</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>A$10</td>
<td>Lab13</td>
<td>0.168</td>
<td>0.915</td>
</tr>
<tr>
<td>3</td>
<td>Exp19</td>
<td>Lab13</td>
<td>0.745</td>
<td>0.998</td>
</tr>
</tbody>
</table>

Figure 63: Association rules in dataset B2 in 2015

Figure 64: VOS matrix visualization of dataset B2 in 2014

Figure 65: VOS matrix visualization of dataset B2 in 2015
Appendix 18: Exploratory Data Analysis Entity C

Table 38: Information on subdivided journal entries in entity C, 2014

<table>
<thead>
<tr>
<th>Dataset 2014</th>
<th>No. of Transactions</th>
<th>No. of JE Lines</th>
<th>No. of High-risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>21,965</td>
<td>88,045</td>
<td>10</td>
</tr>
<tr>
<td>C1</td>
<td>398</td>
<td>11,537</td>
<td>10</td>
</tr>
<tr>
<td>C2</td>
<td>21,535</td>
<td>75,934</td>
<td>10</td>
</tr>
<tr>
<td>C3</td>
<td>32</td>
<td>574</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 39: Information on subdivided journal entries in entity C, 2015

<table>
<thead>
<tr>
<th>Dataset 2015</th>
<th>No. of Transactions</th>
<th>No. of JE Lines</th>
<th>No. of High-risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>24,324</td>
<td>93,520</td>
<td>12</td>
</tr>
<tr>
<td>C1</td>
<td>776</td>
<td>11,368</td>
<td>12</td>
</tr>
<tr>
<td>C2</td>
<td>23,289</td>
<td>80,605</td>
<td>0</td>
</tr>
<tr>
<td>C3</td>
<td>259</td>
<td>1,547</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 66: Number of journal entry lines (left) and journal entries (right) posted per month in dataset C2

Figure 67: Total number of manual journal entry lines per category in dataset C2
Figure 68: Line Items posted over time in dataset C2

Figure 69: Accumulated amounts posted per category involving manual journal entries in dataset C2
Figure 70: Heatmaps of amounts posted to accounts in dataset C2 on specific weekdays (left: 2014, right: 2015)

Figure 71: Association rules in dataset C2 in 2014

Figure 72: Association rules in dataset C2 in 2015
Figure 73: VOS matrix visualization of dataset C2 in 2014

Figure 74: VOS matrix visualization of dataset C2 in 2015
Appendix 19: RapidMiner Association Rules

Figure 75: RapidMiner model for finding association rules
Appendix 20: RapidMiner - Modeling

Figure 76: RapidMiner model required to split intra-category transactions from inter-category transactions
Figure 77: RapidMiner model required to filter out B3 (Double-Date Bookings)
Figure 78: RapidMiner model to apply clustering and extract best profile predictor

Figure 79: RapidMiner operator used for measuring k-distance graph
Figure 80: Parameter settings of DBSCAN clustering operators

Figure 81: RapidMiner model of applying kNN and prepare the data for ANN
Figure 82: Parameter setting kNN (similar for k=3)

Figure 83: RapidMiner model that for ANN preparation on SMOTE model in KNIME
Figure 84: RapidMiner model of 10-fold validation of training ANN

Figure 85: Parameter Setting of ANN for training
Appendix 21: DBSCAN Parameter Setting

Figure 86: k-distance plot for different entities and different attributes involved
Appendix 22: Clustering Results Entity A

Figure 87: Cluster output from RapidMiner labeled with found profile (entity A)
Appendix 23: Clustering Results Entity B

Figure 88: Cluster output from RapidMiner labeled with found profile (entity B)
Appendix 24: Clustering Results Entity C

Figure 89: Cluster output from RapidMiner labeled with found profile (entity C)
Appendix 25: Model used for SMOTE

SMOTE has been applied in KNIME using the following format. The (cat) or (day) attributes of every profile (1, 2, 3, 4) have been read in which after the “profile” predictor has been converted to a string. This is needed for SMOTE to label “profile” as the target attribute that has to be oversampled. The outcome is an Excel file which is again read by RapidMiner as can be seen in Appendix 20, Figure 83.

Figure 90: SMOTE operator design in KNIME (for one entity)
## Appendix 26: Clustering Methods with Highest F-measures

*Table 40: Cluster models used for labeling profiles of journal entries*

<table>
<thead>
<tr>
<th>Profile</th>
<th>Entity A</th>
<th>Entity B</th>
<th>Entity C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profile 1</td>
<td>Euc-Cat Cl2&amp;3</td>
<td>Cos-Cat Cl1</td>
<td>Euc-Cat Cl2</td>
</tr>
<tr>
<td>Profile 2</td>
<td>Euc-Cat Cl1</td>
<td>DBSCAN-Cat Cl3</td>
<td>Man-Cat Cl3</td>
</tr>
<tr>
<td>Profile 3</td>
<td>Euc-Day Cl4</td>
<td>Euc-Day Cl0</td>
<td>Man-Cat Cl2</td>
</tr>
<tr>
<td>Profile 4</td>
<td>DBSCAN-Day Cl2</td>
<td>Euc-Cat Cl1</td>
<td>DBSCAN-Cat Cl3</td>
</tr>
</tbody>
</table>
Appendix 27: Theoretical impact of Automatic Bucketing

**Figure 91:** The impact of automatic bucketing profiles in entity B 2015

**Figure 92:** The impact of automatic bucketing profiles in entity C 2015
Appendix 28: Business Process Redesign

Figure 93: Proposed redesign for case company