Threshold tuning of transaction monitoring models

A risk-based approach to combat money laundering

S. Vis

Delft University of Technology

TU Delft

Threshold tuning of transaction monitoring models

A risk-based approach to combat money laundering

by

S. Vis

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday November 9th, 2022 at 14:00 PM.

Student number:	4603990	
Project duration:	February 7th, 2022 - November 2nd, 2022	
Thesis committee:	Dr. D. Kurowicka	TU Delft, supervisor
	Dr. J. Goudsmit	Rabobank, supervisor
	Dr. W. van Willigen	Rabobank, supervisor
	Dr. K.S. Postek	TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.
Cover image adapted from https://www.investopedia.com [17, 42].

TU Delft

Abstract

Money laundering is an increasing problem for the global economy. To combat money laundering, banks use transaction monitoring models with particular thresholds to detect unusual transaction behaviour. However, it is a challenge to determine and evaluate the suitability of a threshold level to ensure that the risk of misclassification of transactions falls within the bank's risk appetite. In the threshold tuning process, the suitability of a threshold level can be evaluated with a sample of the transactions below or above a threshold level which are reviewed by an analyst.
One problem is that the review process of transactions during the threshold tuning process is time-consuming. In addition, banks want to be able to quantify the risk of misclassification of transactions to determine whether this falls within their risk appetite.

This underlines the need to develop a threshold tuning strategy to accelerate the threshold tuning process in which the risk of misclassification of transactions can be quantified to determine whether it falls within the bank's risk appetite. To accelerate the threshold tuning process, a framework was developed and five threshold tuning strategies were established which evaluate the suitability of different threshold levels with a given strategy. In addition, several methods to determine a confidence interval were examined to quantify the risk of misclassification and to ensure that it falls within the bank's risk appetite.

The threshold tuning strategies were compared and evaluated on the required amount of reviews of transactions and the difference between the found and true threshold level using synthetic data sets. Overall, the bisection threshold tuning strategy is recommended, since this strategy resulted in the lowest number of required reviews of transactions and resulted in a small difference between the found and true threshold level.
The results of the synthetic data sets were promising, but more experiments with preferably real transaction data or other distributions are required to further evolve and fully validate the framework and proposed bisection strategy.
The work presented in this thesis contributed to a more risk-based approach to enhance the efficiency and effectiveness of the threshold tuning process of transaction monitoring models.

Preface

The last nine months have been an interesting and educational experience. This thesis is my final project for the Master Applied Mathematics at the Delft University of Technology and marks the end of my study time in Delft. I am immensely grateful for all the opportunities I have had over the past six years in Delft both within my studies and outside them, which have allowed me to develop myself, discover my interests and made me the person I am today.

I would like to take this opportunity to express my sincere gratitude to the people that have helped me during my research journey. I am thankful for this interesting applied research project and the opportunity to collaborate with the Compliance Model Validation department within Rabobank. I would like to thank my supervisors Jeroen Goudsmit and Willem van Willigen from Rabobank for their support, guidance and sharing their knowledge on this interesting topic. Through this fruitful collaboration, I have had the opportunity to learn from domain experts in the field of money laundering, model validation and quantifying risk, for which I am extremely grateful. Additionally, I would like to thank Dorota Kurowicka who supervised me from the TU Delft for her suggestions, feedback and kind conversations.

Furthermore, I would like to thank my family and friends for their support and motivation during my studies. First of all, I would like to thank my parents and sister for their infinite love and support my whole life. You are always there for me and have given me the confidence to choose my own path.
I would also like to specifically thank Frank, Eva, Anna W, Anna N and Maricke that have joined me for lunches, coffee breaks, walks over campus and study sessions during my thesis.
Furthermore, I would like to thank my housemates where it was always nice to come home, my study friends for our study time together, my friends of the 'TU Delft Solar Boat Team 2021' with whom I had an unforgettable gap year and my rowing team 'the Potvissen' for our dear friendship I cherish for the rest of my life.
Finally, I would like to thank my boyfriend for his kindness and his willingness to listen to my challenges in the past nine months. Thank you all!

Simone Vis
Delft, November 2022

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

Abbreviation	Definition
AML	Anti-Money Laundering
ATL	Above-the-line
BTL	Below-the-line
CDD	Customer Due Diligence
CI	Confidence Interval
CP	Coverage Probability
DNB	De Nederlandse Bank
EL	Expected Length
FATF	Financial Action Task Force
FIU	Financial Intelligence Unit
FN	False Negative
FP	False Postive
FPC	Finite Population Correction
MOT	Disclosure of Unusual Transactions Act
RTL	Risk Tolerance Level
SIRA	Systematic Integrity Risk Analysis
Sw	Sanction Act
TN	True Negative
TP	True Positive
Wwft	Anti-Money Laundering and Terrorist Financing Act

Symbols

Symbol	Definition
α	The significance level of a confidence interval
B	The amount of of bootstrap samples
e	The Margin of Error of a confidence interval
k	Amount of interesting transactions in a sample
K	The amount of threshold levels
n	The size of the sample
N	The size of the population
p	Proportion of interesting transactions in the population
$\hat{p}$	Proportion of interesting transactions in the sample
p_0	Preliminary point estimate for the proportion of interesting transactions in the population
$R_i(t)$	The reward for threshold level i at time t
$\hat{\mu}_i(t)$	The average reward for threshold level i at time t
τ	The temperature factor controlling exploration versus exploitation
θ^*	The bootstrap estimate

1

Introduction

Money laundering is an increasing problem for the global economy with sums estimated between 2 and 5 per cent of global GDP [6]. Nonetheless, the penalties financial institutions receive for failure of compliance continued to rise with an estimated amount of $42 billion globally in 2016 [15]. On top of that, the Scientific Research and Documentation Centre (WODC) has estimated that around 16 billion euros of criminal money was laundered in the Netherlands in 2014. This amount consists of 6.9 billion euros of domestic criminally obtained money that is laundered in the Netherlands and the inflow of 9.1 billion euros of laundered money from other countries, reflecting the international nature of money laundering in the Netherlands [30]. Therefore, regulators around the world encourage innovative approaches to combat money laundering and protect the integrity and stability of financial institutions.

Money laundering is an illegal process in which criminals make it appear that amounts of money come from a legitimate source. Financial institutions have an important role as gatekeepers of the stability and integrity of the global financial system and are required to monitor all transactions passing through their system and to classify each transaction as unusual or not. Therefore, Anti-Money Laundering (AML) models, which currently rely mainly on rule-based systems with fixed derived thresholds, aim to detect unusual behaviour in transaction data [19].

The following example explains how banks could detect money laundering through transaction monitoring. A well-known money laundering method is smurfing, also known as structuring, when large amounts of money are split up into several small transactions at banks. Suppose that each criminal in a group of X members in country A has an account at a bank, which may be at different banks. All members deposit a certain amount of money M in the first few days of the week into their accounts, which at first will go unnoticed by the bank. After a few days at the end of the week, each member transfers that certain amount of money to the same account in country B which results in an amount of M*X flowing into an account in country B every week [29].

To detect smurfing, the bank can develop a transaction monitoring scenario. For example, if more than 90 per cent of deposited money is transferred to another account within 7 days, the transaction should generate an alert. An alert refers to a signal indicating a potentially unusual transaction [10]. With such scenarios, the bank tries to detect unusual transaction behaviour. The transactions will generate an alert when the criminals forward their deposit within a week. The alerts will be investigated and institutions are obligated to report unusual transactions to the Financial Intelligence Unit (FIU) immediately after the unusual nature of the transactions becomes known, as described in Article 16 of the Anti-Money Laundering and Terrorist Financing Act (Wwft) [11]. Then, the FIU analyses the report and sends it to law enforcement if the analysis is confirmed [13]. In this case, money laundering occurred and the AML model correctly classified the transactions as unusual referred to as true positives. On the other hand, an AML model can also generate alerts for transactions that did not turn out to be money laundering. The misclassified alerts are referred to as false positives.

Unfortunately, not all unusual transactions are detected by an AML model. In the previous exam-

ple, this could happen if the criminals forward the deposit after 8 days while the scenario only generates an alert for deposits forwarded within 7 days, which has the consequence that the transactions do not generate an alert. The unusual transactions that are not detected by the AML model are misclassified and referred to as false negatives, whereas the not detected normal transactions are correctly classified as true negatives.

How many transactions will be misclassified as false negatives below the threshold level or false positives above the threshold level depends on the choice of the threshold level. To determine which threshold level is most appropriate for an AML model, a sample of transactions from above and/or below the threshold level is taken. This makes it possible to evaluate the effectiveness of a threshold level of an AML model with a certain confidence level. However, the threshold tuning process is very time-consuming since all sampled transactions have to be investigated by an analyst.
Moreover, transaction monitoring often presents a significant compliance challenge and AML model developers have to be careful with the choice for a threshold level. If the thresholds are too sensitive to unusual transaction behaviour, the thresholds are likely to generate a large number of not interesting alerted transactions (false positives) which take substantial time to investigate. On the other hand, thresholds that are not sensitive enough, present the risk of missing interesting transactions (false negatives) that pose a risk to the bank.

Therefore, the first goal of this thesis is to research possibilities to advance the threshold tuning process. The objective is to define an efficient threshold tuning strategy to accelerate the threshold tuning process to determine the most suitable threshold level. The effectiveness and limitations of the proposed threshold tuning strategy and the relative and absolute performance will be validated.
Furthermore, this thesis will discuss how the uncertainty that arises from sampling in the threshold tuning process can be quantified. This quantification makes it possible to make statements about the population of transactions with a certain confidence level after completing the threshold tuning process.
In this thesis, the potential of various threshold tuning strategies will be evaluated using three synthetically generated data sets. This makes it possible to control experiments using prior knowledge of unusual transactions.

1.1 Research objective

After introducing the thesis subject, it is important to state the research objectives. The research objectives of this thesis are:

1. How can a threshold tuning strategy be incorporated to accelerate the threshold tuning process of transaction monitoring models?
2. How can the uncertainty in the threshold tuning process for the chosen threshold level be quantified?

1.2 Thesis structure

This section introduces the structure of the thesis. Before we dive deeper into threshold tuning strategies to accelerate the threshold tuning process, Chapters 2 and 3 introduce an overview of the necessary background literature. First, Chapter 2 describes money laundering and the current Anti-Money Laundering validation practices in place at financial institutions. Then, the problem setting is formulated. This provides the background as to why this thesis focuses on threshold tuning strategies. Chapter 3 describes various confidence interval methods to quantify the uncertainty with sampling and how the required sample size can be determined. Chapter 4 describes various threshold tuning strategies that are investigated in more detail in this thesis. Experiments using different synthetic data sets are described in Chapter 5. These experiments are conducted to gain a more thorough understanding of threshold tuning strategies' performances. The empirical results of the experiments are presented in Chapter 6. Chapter 7 concludes this thesis by providing a conclusion of the results derived in this thesis. Moreover, recommendations are presented for further research directions.

2

Threshold tuning background

This chapter provides an overview of relevant background information about transaction monitoring within banks and the threshold tuning process in more detail. This chapter starts with an introduction to anti-money laundering practices and the transaction monitoring process within banks in Section 2.1.1. Subsequently, Section 2.2 discusses the threshold tuning process of transaction monitoring models. Furthermore, the problem setting for this thesis is described in Section 2.3. Lastly, Section 2.4 summarises the most important aspects of the threshold tuning process.

2.1 Anti-money laundering

As briefly introduced in the introduction, money laundering can be described as an illegal process in which criminals make it appear that amounts of money come from a legitimate source. Criminals often follow three steps when laundering money. The first step is to secretly inject 'dirty money' into the legitimate financial system, also known as the placement step. This is followed by the layering step, where criminals try to hide the origin of the money through transactions and bookkeeping tricks. Finally, criminals can now withdraw the laundered money from a legitimate account and it is integrated into the system [25].

Money laundering and terrorism funding is a growing problem for society. To address the international fight against money laundering and terrorist financing, the European Parliament and the Council issued directives to prevent the use of the financial system for money laundering or terrorist financing based on the recommendation of the Financial Action Task Force (FATF). The Dutch Central Bank (DNB) supervises the integrity of Dutch financial institutions considering among others the Wwft, which replaces the Disclosure of Unusual Transactions Act (MOT), and the Sanctions Act (Sw), which was established in 1977 [9]. On the recommendation of the FATF, DNB set up a guideline to provide financial institutions with tools to adequately perform their role as gatekeepers of the financial system in 2011. As a result, in recent years banks have invested heavily in customer due diligence (CDD) and transaction monitoring processes to identify unusual transactions [7].

2.1.1 Anti-money laundering transaction monitoring process within banks

The DNB has published a guidance report on the post-event transaction monitoring process at banks [10]. The key aspects of this report are further explained in this section. Many banks use the so-called 'three lines of defence' model which is a framework designed to facilitate an effective risk management system (risk owners, risk oversight and risk assurance). To translate this framework to the transaction monitoring process, a clear separation of functions is important because inadequate governance of model development, implementation, use and validation can create an increased risk for the bank. In the transaction monitoring process, there are often three functions distinguished:

1. **The 1^{st} line (business)**: responsible for transaction monitoring.

2. **The 2^{nd} line (compliance)**: responsible for quality assurance through an independent advisory and controlling role and periodically challenging first-line model risk management processes.
3. **The 3^{rd} line (audit)**: responsible for an independent check on the functioning of the 1^{st} and 2^{nd} line.

The functions are separated to minimise model risk as much as possible. Additionally, periodic training of employees is essential to create awareness of money laundering and terrorist financing risks. Besides the separation of functions and training, the AML and transaction monitoring process can be divided into several steps as described in the guidance report by DNB on the post-event transaction monitoring process at banks and presented in Figure 2.1 [10].

- Customer screening
- Risk identification
- Policies and procedures
- Transaction monitoring
- Model validation

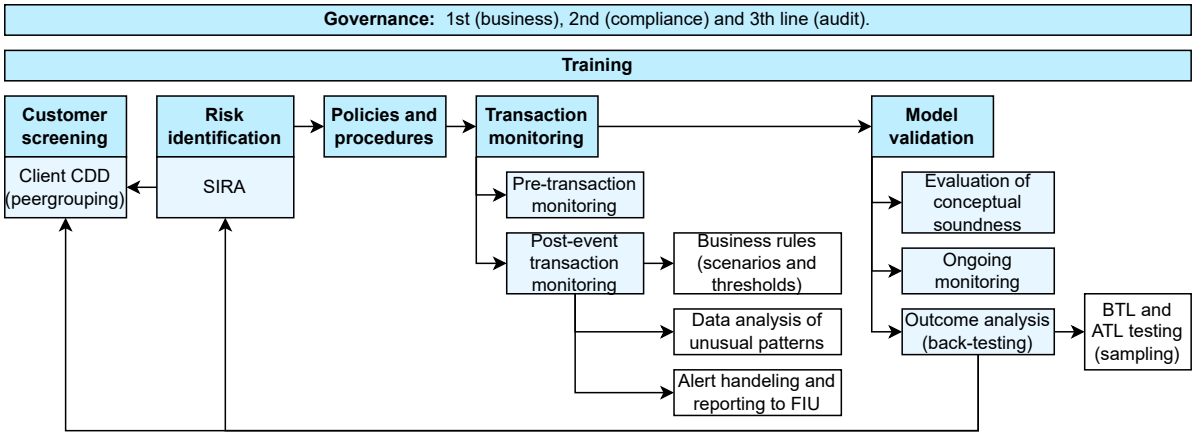


Figure 2.1: The anti-money laundering transaction monitoring process within banks [10].

Customer screening
First, under the Wwft, banks are required to conduct client screening and draw up a client risk profile. When determining a Customer's Due Diligence (CDD) risk rating (low, medium, high), the bank must establish the expected transaction behaviour through peer grouping. Peer grouping allows the bank to determine an expected transaction profile to detect unusual transactions. This is necessary so that banks can ensure that the transactions carried out are consistent with the bank's knowledge and the risk profile of the customer. This allows banks to detect unusual transactions.

Risk identification
As a second step, banks need to make a risk identification of their types of customers, products, distribution panels and transactions. This risk analysis should be reported through a Systematic Integrity Risk Analysis (SIRA). The SIRA consists of a cycle of four aspects [8]:

1. Risk identification based on several factors.
2. Risk analysis where the risk is equal to the probability of the risk multiplied by its impact.
3. Risk management through policies, procedures, systems and controls.
4. Risk monitoring and review.

The actions that follow from the risk analysis depend on the bank's risk appetite.

Policies and procedures
Thirdly, banks should make a visible translation of the SIRA into policies and procedures regarding the transaction monitoring process.

Transaction monitoring
The penultimate step is the transaction monitoring process itself which can be divided into pre-transaction monitoring and post-event transaction monitoring. Pre-transaction monitoring refers to situations where there is face-to-face contact between the customer and the bank employee. Examples are when a customer wants to exchange foreign currency at a bank office or deposit large amounts of cash. Post-event transaction monitoring refers to the detection of unusual non-cash transactions and can be divided into three aspects: converting policies and procedures into a set of business rules, detecting unusual transactions through data analysis and the alert handling and reporting process.

1. **Set up of business rules**: The set of business rules are detection rules consisting of scenarios and certain thresholds such as amounts and numbers of transactions or combinations thereof. The business rules are used to detect unusual transaction patterns that may indicate money laundering or terrorist financing.
2. **Transaction detection through data analysis**: Data analysis is used to detect potentially unusual transaction behaviour whereby an alert is generated. The alerts are then investigated and reported to FIU-NL immediately after the unusual nature of the transactions becomes known.
3. **Alert handling**: The alert handling and reporting process involves assessing and reporting the consequences of the report to FIU-NL and whether additional control measures are necessary. The considerations and conclusions must be recorded to close an alert or report it to the FIU-NL.

Model validation
The final step is the continuous process of model validation to verify if the models perform as expected and if the current validation activities are sufficient to reduce model risk. This section about model validation is largely based on the Model Risk Management report from Comptroller's Handbook [23]. The goal of validation is to challenge developers and users and to note their findings about shortcomings that need to be revised. For an effective validation framework, three elements are important: evaluation of conceptual soundness, ongoing monitoring and outcome analysis.

- **Evaluation of conceptual soundness**: In the evaluation of conceptual soundness an assessment is made of the quality of the design and construction of the model. This requires documentation that supports all model choices, model assumptions, data and theoretical construction.
- **Ongoing monitoring**: Ongoing monitoring is executed to confirm that a model has been implemented and used correctly, and is performing as intended. In addition, evaluations of changes in products, customers or market conditions indicate that a model needs to be adjusted.
- **Outcome analysis**: In outcome analysis, a comparison is made between the expected model output and the actual outcomes which can be achieved by statistical tests or other quantitative measures. One form of outcome analysis involves back-testing in which outcomes are compared to the model forecasts during a sample period that is different from the data used in the model development with statistical confidence intervals.

Back-testing for many AML models can be difficult or even impossible because there is no specific manner to compare truly unusual transactions to those identified by the system as unusual. One way around this is to evaluate AML models via above-the-line (ATL) and below-the-line (BTL) testing in a given period.

- **Above-the-line (ATL)** testing involves alert productivity metrics where it takes into account interesting transactions above a threshold (alerted transactions). The goal of ATL testing is to reduce the number of not interesting alerts to increase the overall effectiveness of scenarios.
- **Below-the-line (BTL)** testing focuses on transactions below a threshold, which a model would not identify as interesting and therefore did not generate an alert. A sample of these transactions is taken to determine the missed interesting transactions which indicate detection failure and validate whether this risk falls within the bank's risk appetite. Unlike ATL testing, the goal of

BTL testing is not to reduce the volume of not interesting alerts, but to validate that the set threshold level is in line with the bank's risk appetite.

Banks need to substantiate and document model choices, choices of tests and interpretation of results. Back-testing requires statistical tests but it is often a challenge to substantiate which test is most appropriate in a given setting or how to interpret the results.

Therefore, this thesis will investigate threshold tuning strategies to accelerate the back-testing and threshold tuning process.

2.1.2 Transaction monitoring stages

Transactions can go through three stages during the transaction monitoring process at financial institutions; the alert stage, the case stage, and the reporting stage as shown in Figure 2.2 [19]. All transactions go through the alert stage where the transactions are compared to a set of rules. Alerted transactions are examined with simple initial manual checks. The alerted transactions are not investigated further if the transactions appear legitimate. The remaining alerts go to the case stage and the cases are extensively investigated by experienced analysts and labelled as unusual or not unusual. The cases that are labelled as unusual go to the reporting stage, where the final decision is taken as to whether the case will be reported to the FIU.

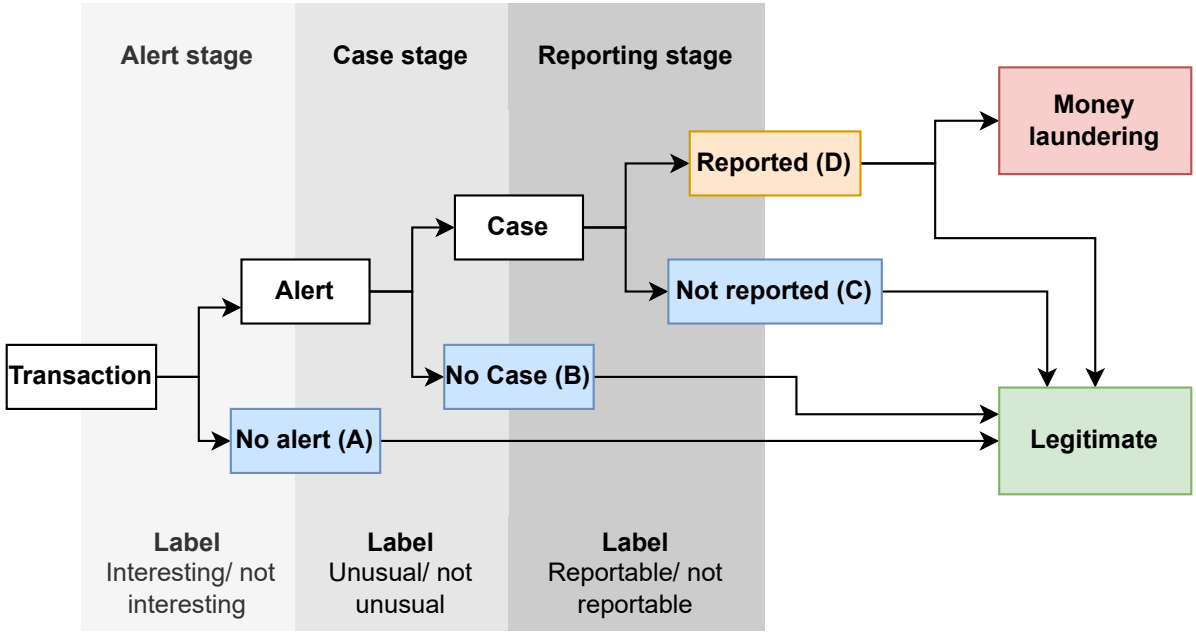


Figure 2.2: Stages of the transaction monitoring process and labels for transaction types [19].

2.2 Threshold tuning process of transaction monitoring models

In section 2.1.1, the different aspects of the transaction monitoring process are briefly introduced. In this process, data analysis is performed to detect unusual transaction patterns by developing different scenarios that characterise the expected behaviour of criminals. Returning to the smurfing example from the introduction, a scenario can be defined as follows:

To detect smurfing, a transaction monitoring model should detect whether large sums of money are structurally split and received at a certain account (the receiving party) or whether sums of money are structurally transferred that are just below a publicly known threshold (the sending party). This scenario should characterise criminal behaviour that may indicate smurfing.

Subsequently, it is stated in the guidance report by DNB that banks should develop for each scenario one or more (business) rules to mitigate this scenario [10]. For example, the following rules can be defined for the smurfing scenario example:

- Sending party rule: repeated transactions to a foreign account in a country that is classified as high risk for money laundering during a certain period, in case the sending party is a customer of the bank.
- Receiving party rule: repeated transactions from a foreign account in a country that is classified as high risk for money laundering during a certain period, in case the receiving party is a customer of the bank.

Thereafter, for each rule, one or more thresholds should be established. In this example, the following thresholds could be defined:

- The period during which transactions are monitored, for example, one month, two months etc.
- The frequency of money transfers to/from a foreign account, for example, every five days, every week, every two weeks etc.
- The amount of money that is repeatedly transferred to/from a foreign account, for example, 1000 euros, 2000 euros etc. This may also include monitoring the average transfer amount or the variance of the transfer amount over the period.

The process of data analysis of unusual transaction patterns to determine a suitable threshold level is illustrated in Figure 2.3.

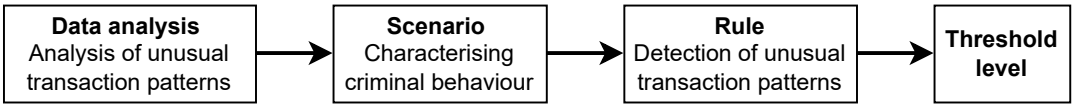


Figure 2.3: The process from data analysis to a threshold level for transaction monitoring models.

To explain how the most suitable threshold level can be determined, it is important to first explain how currently different threshold levels are compared which is described in the Model Risk Management report from Comptrollers Handbook [23]. Afterwards, different threshold tuning strategies will be introduced to determine the most appropriate threshold level while the underlying probabilities of unusual transactions for different rules are unknown.

2.2.1 Back-testing process of rules

As described in the previous section, thresholds of transaction monitoring models are used to detect unusual transaction behaviour. These thresholds are determined during the threshold tuning process through ATL and BTL testing in a testing environment whereby thresholds are increased or decreased to identify the best possible threshold levels and parameters. There are two situations to consider when the threshold tuning process is performed:

- **Initial threshold tuning for a new rule**: to assess and evaluate different threshold levels to set a good initial threshold level when no threshold level has been set before concerning a set of specified performance metrics. In the situation of initial threshold tuning, there is no prior knowledge about unusual transaction behaviour for a certain rule.
- **Periodic evaluation of a threshold for an existing rule**: it is stated in the Model Risk Management report from Comptrollers Handbook [23] that banks should periodically evaluate models to determine whether the current threshold level results in the intended performance and if the performance is sufficient concerning a set of specified performance metrics. In the situation of periodic evaluation of a threshold for an existing rule, there is usually some historical prior knowledge about unusual transaction behaviour for a certain rule.

Of the two situations mentioned above, the main focus of this thesis will be on initial threshold tuning. It is important to understand that although initial threshold tuning happens less often than periodic evaluation of a current threshold for an existing rule, the risk of an incorrect initial threshold level is higher. An incorrect initial threshold level could result in large numbers of incorrectly alerted transactions, causing a higher workload than necessary, or could result in a high risk for the bank in case too many unusual transactions are missed. The risk of an incorrect threshold level is lower with periodic evaluation of a current threshold of an existing rule whereby the risk of the current threshold level is known and accepted. Therefore, the added value for initial threshold tuning could be considered higher.

Back-testing of a threshold can be performed through ATL and BTL testing. A sample is taken of the transactions below and/or above the tested threshold level. Subsequently, an analyst evaluates the sampled transactions from below and/or above the tested threshold level and labels the transactions as interesting or not. This is illustrated in Figure 2.4 where each short black line represents a threshold level.

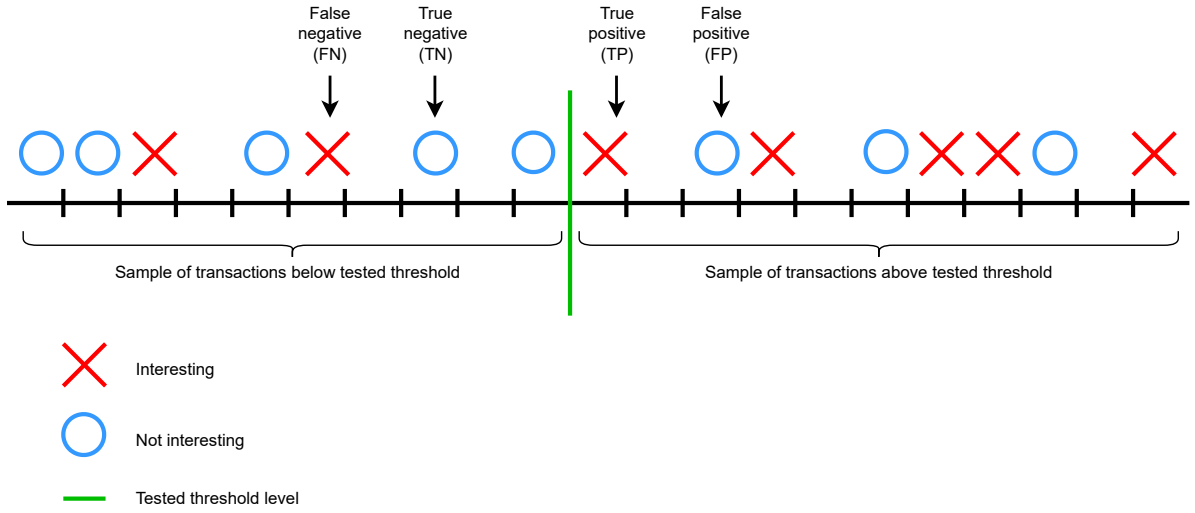


Figure 2.4: Illustration of the trade-off for setting the threshold level.

A confusion matrix can be created after all transactions have been evaluated and labelled. All sampled transactions below the threshold labelled as interesting are false negatives, whereas the sampled transactions labelled as not interesting are true negatives. In addition, all sampled transactions above the threshold labelled as interesting are true positives, whereas sampled transactions labelled as not interesting are false positives. An example of a confusion matrix is shown in Figure 2.5.

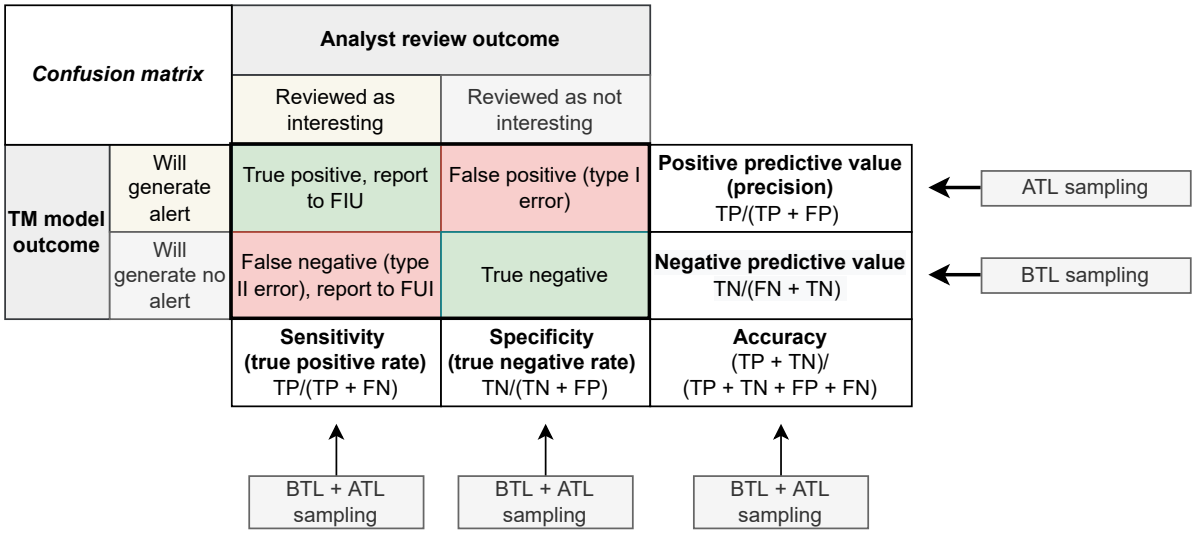


Figure 2.5: Example of a confusion matrix including various performance metrics.

With the confusion matrix, the following five performance metrics and their complements can be determined; positive predictive value, negative predictive value, sensitivity, specificity and accuracy. To determine the positive predictive value, which is presented by the top row of the confusion matrix, it is only necessary to take a sample of the transactions above the tested threshold. To determine the negative predictive value, which is given by the bottom row of the confusion matrix, it is only necessary to take a sample of the transactions below the tested threshold. However, to determine the sensitivity, specificity or accuracy, it is necessary to take both a sample of the transactions below and above the tested threshold level. The terminology and derivations from a confusion matrix are summarised in Appendix A.

Before it is possible to determine the most appropriate threshold level in the threshold tuning process, a performance measure must be chosen in conjunction with a criterion, which serves as a stopping condition for the threshold tuning process. In this thesis, the main focus is on the false omission rate $\frac{FN}{FN+TN}$ as a measure, which is 1 minus the negative predictive value, since the risk of detection failure is often an important aspect in the bank's risk appetite. Furthermore, in this thesis, our own performance criterion is defined as the Risk Tolerance Level (RTL):

Definition 1. ***Risk Tolerance level (RTL)***. *The Risk Tolerance Level is the maximum permissible percentage X of missed interesting transactions (false negatives) of the sampled transitions below the tested threshold with a $(1-\alpha)\times 100\%$ confidence level.*

where α is the significance level of a confidence interval, often chosen as 0.05. The back-testing process is illustrated in Figure 2.6. The only block that changes in the threshold tuning process, when a different performance measure is chosen, is the yellow block with the possible consequence of sampling transactions below and/or above the tested threshold. Sections 3.3 and 6.5 will discuss in more detail how the choice for a performance metric other than the false omission rate affects the threshold tuning process to determine the most appropriate threshold level which depends on the chosen performance metric.

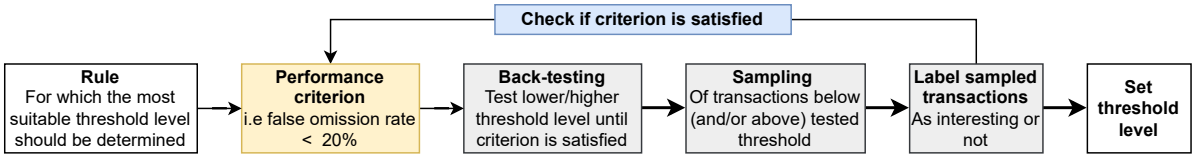


Figure 2.6: Illustration of the back-testing process.

2.3 Problem formulation

As briefly stated in the introduction, the objective of this thesis is to define a threshold tuning strategy to test threshold levels in an ordered and efficient manner to determine the most appropriate threshold that satisfies a certain performance criterion. Suppose that the performance of a threshold level is measured in the false omission rate of a tested threshold level with as a criterion the RTL.

The exploration-exploitation trade-off
The dilemma of finding the most appropriate threshold level, where the risk falls within the bank's risk appetite, can be described as a decision problem involving an exploration-exploitation trade-off. The exploration-exploitation trade-off describes how, when faced with several competing alternatives, a decision maker has to choose between a known good alternative, defined as exploitation, and one or more unknown but potentially more rewarding alternatives, defined as exploration. In the situation of threshold tuning, a choice has to be made between reviewing transactions from another threshold level than the threshold level that has the most promising performance at that moment, defined as exploration. This takes more time but potentially results in a more appropriate threshold level. The other choice is to continue reviewing transactions from the threshold that has the most promising performance so far, defined as exploitation. This problem setting, which requires a trade-off between exploration, gathering enough information about all threshold levels to learn which threshold level is the most suitable, and exploitation, using the known information to avoid underperforming threshold levels, is reflected by the multi-armed bandit problem that is introduced by Robbins in 1952 [21].

The multi-armed bandit problem is formulated as follows. A gambler visits a casino and chooses to play a row of slot machines, each producing a random payout according to some distribution which is unknown to the gambler. The gambler can learn about the distributions of the slot machines by pulling different arms, each giving a certain payout. The gambler needs to make a trade-off between exploiting his existing knowledge and pulling arms that gave the highest payout previously and earning more in the short term or exploring the other arms to learn the distributions of all arms to receive the highest payout in the long term. The two-fold goal of the gambler is to develop a sequential strategy to discover which arm has the highest expected payout and on the other hand maximise his payout while playing. Translating the two-fold goal to the threshold tuning process would imply the following:

Two-fold threshold tuning objective

- To develop a threshold tuning strategy to determine the most suitable threshold level that satisfies the performance criterion set beforehand.
- To minimise the amount of sampled transactions that an analyst needs to review to determine the most suitable threshold level with a $(1 - \alpha) \times 100\%$ confidence level.

Before diving deeper into various threshold tuning strategies to achieve this two-fold objective, Chapter 3 will first describe how the uncertainty that arises during sampling in these threshold tuning strategies can be quantified.

2.4 Chapter summary

This chapter explained the anti-money laundering transaction monitoring process within banks and how AML models can be validated. AML models can be evaluated via above-the-line (ATL) and below-the-line (BTL) testing, taking a sample of the transactions above and/or below the tested threshold level. To detect money laundering, a scenario and rule is established that should characterise the criminal behaviour. Depending on the chosen performance metric, a performance criterion can be defined. Thereafter, in the back-testing process, sampled transactions are reviewed by an analyst with which a confusion matrix can be created. Depending on the outcome of the reviewed transactions, a higher or lower threshold level is tested until the most appropriate threshold level that satisfies the performance criterion is determined.

3

Sampling and confidence interval methods

This chapter describes how the uncertainty that arises during sampling in the threshold tuning process can be quantified. The back-testing process via BTL and ATL testing discussed in Chapter 2 forms the basis for the aspects discussed in this chapter.

This chapter is divided into four sections. An important aspect of back-testing threshold levels is sampling of the transactions below and/or above the tested threshold level. Therefore, Section 3.1 discusses different random sampling methods. These methods describe different practices on how a sample can be taken from a population.
Subsequently, Section 3.2 describes how the uncertainty in the sample estimate can be quantified using a confidence interval since the underlying distribution of the number of usual and unusual transactions in the population can be described by a binomial distribution. This section provides two performance criteria to compare five well-known confidence interval methods and explains how the required sample size can be determined using the confidence interval methods for the binomial distribution.
However, the chosen confidence method described in Section 3.2 is not suitable to determine confidence interval limits for the sensitivity, specificity, or accuracy as measure since these measures express a ratio of two random variables with unknown distributions. Therefore, Section 3.3 describes a method to construct a confidence interval for the sample estimate of a ratio.
Finally, a summary of the important choices that are made in this chapter which will be used further in this thesis is provided in Section 3.4.

3.1 Random sampling

In Section 2.2, it is described how threshold tuning can be performed by taking a sample from the transactions above and/or below the tested threshold level which is referred to as BTL and ATL testing. In this section, different techniques how to take a sample from a population of transactions below and/or above the tested threshold level will be discussed in more detail.

Sampling is the selection of a subset of individual observations from within a statistical population to estimate the characteristics of the whole population. In this case, the objective of sampling is to collect samples that are representative of the population. When sampling, it is important to define the population from which the sample is drawn. A population can be defined as including all observations with a certain characteristic.

The sampling techniques that will be discussed are probabilistic in nature, implying that each observation in the population has a specified probability of being included in the sample and that the actual composition of the sample is random [18]. There are various methods for conducting a random sample. Common methods of random sampling are simple random sampling, systematic sampling,

stratified sampling and cluster sampling. The difference in sample composition between the different sampling methods is illustrated in Figure 3.1.

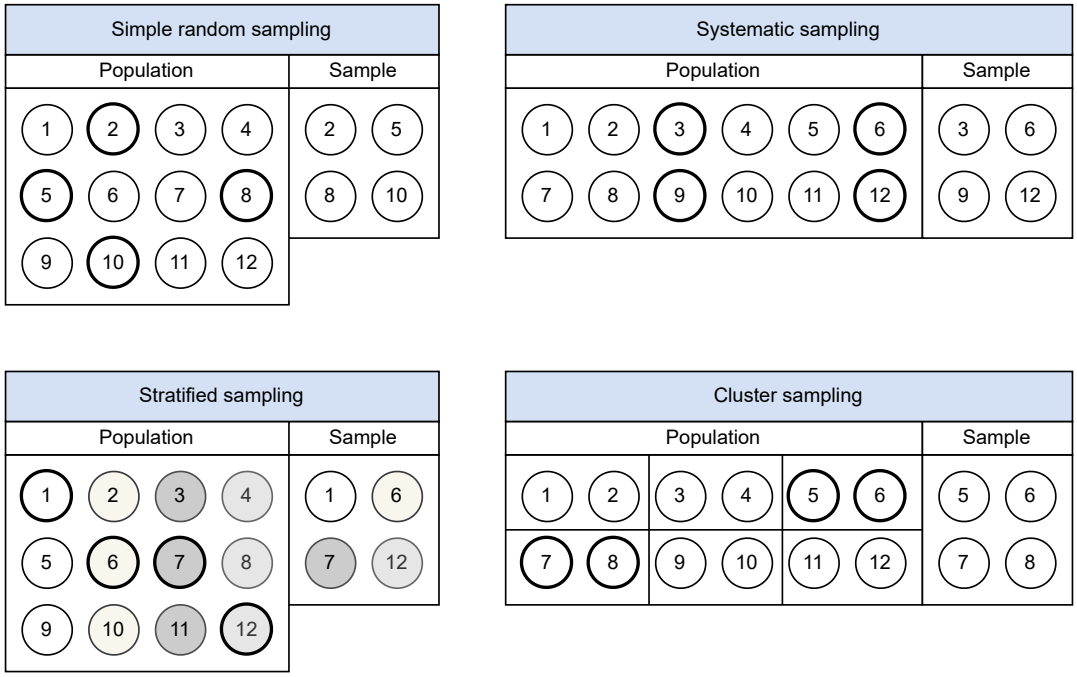


Figure 3.1: Illustration of sample composition of different probability random sampling methods.

With simple random sampling, each observation has an equal chance of being selected without any subgroups in the population. In systematic sampling, the population is ordered according to a certain characteristic. Then observations are selected at a certain interval with a random starting point. By stratified sampling, the observations are divided into a number of distinct categories (strata) and a random sample is taken from each stratum as an independent homogeneous sub-population. This method can be used if one is interested in information about each of a number of subpopulations in addition to information about the population as a whole [18]. With cluster sampling, the observations are selected in a certain group (cluster). The difference between cluster and stratified sampling is that cluster sampling divides a population into groups and then includes all observations of some randomly chosen groups, whereas stratified sampling divides a population into groups but only includes some observations of each group. An advantage of cluster sampling is that it can be more cost-effective and is very suitable for large populations [27].

Overall, the performance and choice of sampling method are mainly dependent on the properties of the population. In general, stratified and systematic random sampling are more effective than simple random sampling [12]. However, a major advantage of simple random sampling is its simplicity and that the technique requires no assumptions about the population.

Therefore, since it is difficult to make assumptions about the unknown population of transactions, especially concerning the proportion of unusual transactions, the method of simple random sampling is chosen for back-testing different threshold tuning strategies.

3.2 Confidence interval and sample size methods for the binomial distribution

In the previous section, it is explained that the transactions below and/or above the tested threshold level are sampled using ‘simple random sampling’. In this section, different confidence interval methods for the binomial distribution will be discussed. The advantages and limitations of each method will

be explained through a simulation study. Afterwards, it will be discussed how the sample size can be determined from these confidence intervals.

There are various statistical methods to determine the confidence interval for the population proportion of the binomial distribution. The sample size can be determined using these confidence interval methods by rewriting the confidence interval formulas. Before discussing the different confidence interval methods, the definition of a confidence interval is stated.

Definition 2. ***Confidence Interval (CI)***. *A confidence interval for a population parameter, θ, is a random interval, calculated from the sample, that contains θ with some specified probability [18].*

For example, a 95% confidence interval for a population parameter θ is a random interval that contains θ with a probability of 0.95. In other words, if one takes many random samples and forms a confidence interval around each one, about 95% of these intervals would contain θ [18]. To illustrate the concept of a confidence interval, suppose one draws 20 samples from a population to approximate a 95% confidence interval for θ. This would mean that on average 5% of the 95% confidence intervals, or 1 out of 20, would not include θ [18].
Lastly, an interval is called a $(1-\alpha)\times 100\%$ confidence interval if the coverage probability is $(1-\alpha)\times 100\%$. Three factors influence the width of a confidence interval given a level of confidence [28]:

- The width of the confidence interval depends on the variance of the sample. If the sample has a larger variance, the confidence interval will be wider.
- The size of the sample influences the width of the confidence interval. A larger sample size gives a higher precision and thus a smaller confidence interval, whereby the precision is defined as half of the expected length of a confidence interval.
- A 99% confidence interval must be wider than a 95% confidence interval, because the interval must contain the population parameter with more certainty.

3.2.1 CI performance criteria: Coverage Probability and Expected Length

Much research on constructing confidence intervals for a binomial distribution is available in literature. The two criteria often used to assess the performance of these methods are the Coverage Probability (CP) and the Expected Length (EL). The Coverage Probability is the actual probability that the interval contains the true population parameter whereas the 'nominal coverage probability' is the confidence level of a constructed confidence interval which is often set at 95%. The Expected Length is the expected width of a confidence interval. The binomial distribution probability density function is defined as follows:

$$f(k,n,p) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.1}$$

for $k = 0, 1, 2, ..., n$, where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

where k is the number of successes in n trails. Consider a large or infinite population, in which X is defined as the number of successes, from which a random sample of size n is drawn. Given p as an unknown population proportion, a two-sided confidence interval with nominal confidence level $(1-\alpha)\times 100\%$ can be represented by $[L(X), U(X)]$. Given a sample size n and population proportion p the CP and EL are defined as follows [24, 16]:

$$CP(n,p) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} I_{[L(k),U(k)]}(p) \tag{3.2}$$

(with $I_{[a,b]}(x) = 1$ if $x \in [a,b]$ and $I_{[a,b]}(x) = 0$, otherwise) and

$$EL(n,p) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} (U(k) - L(k)) \tag{3.3}$$

As mentioned before, a confidence interval does not always contain the actual value of the parameter, but the CP of a random interval $[L(X), U(X)]$ as a solution of equation 3.2 should be $(1-\alpha)$. Furthermore, there is a preference for methods with the smallest EL when both methods have approximately the same CP [16].

For confidence interval methods for the binomial distribution, a clear distinction can be made between methods based on a normal approximation and an exact method. The following methods will be compared:

- Exact method: Clopper-Pearson
- Methods based on normal approximation
  - Wald (with a finite population and Yates' continuity correction)
  - Agresti-Coull
  - Wilson Score (with a finite population and Yates' continuity correction)
  - Arcsine (with a finite population and Yates' continuity correction)

These different confidence interval methods are compared because they have been identified in the literature as having several good properties [16]. In case the CP is taken as criteria to compare the different confidence interval methods, the methods can be classified into three different groups [24]:

- **1st group: strictly conservative methods**, for which the minimum coverage probability is greater or equal to $1-\alpha-0.005$ for all $n \geq 10$ and all p:
$$\min_p CP(p,n) \geq 1-\alpha-0.005, \forall n \geq 10$$
- **2nd group: on average correct methods**, for which the mean coverage probability is greater or equal to $1-\alpha-0.005$ for all $n \geq 10$:
$$\int_0^1 CP(p,n)dp \geq 1-\alpha-0.005, \forall n \geq 10$$
- **3rd group: other methods**, which do not belong to group 1 or 2.

Only the first two groups are considered acceptable confidence interval methods. Before delving deeper into the formulas for the different confidence intervals, the following advantages and disadvantages are often mentioned in literature for using the exact confidence interval method or the confidence interval methods based on a normal approximation [16, 4].

Exact method (Clopper-Pearson)
Advantages

- This method is accurate when $np < 5$ or $n(1-p) < 5$.
- The calculation of the confidence interval is possible when $p = 0$ or $p = 1$.
- No assumption or approximation has to be made about the underlying distribution.

Disadvantage

- The formulas to determine the upper and lower bound of the confidence interval are more complex and a computer is required to calculate the upper and lower bound.

Normal approximation methods
Advantages

- These methods are in general easy to understand.
- The upper and lower bounds of these methods are easier to calculate by hand.

Disadvantages

- In general, the accuracy suffers when $np < 5$ or $n(1-p) < 5$.
- The calculation of the confidence interval is not possible when $p = 0$ or $p = 1$.

3.2.2 Exact method

Clopper-Pearson

The Clopper-Pearson method is considered the golden standard for determining an exact confidence interval of the probability, based on the binomial probability function [39]. Given α and the sample proportion $\hat{p} = \frac{k}{n}$, the exact confidence limits can be determined by solving the following equations [24]:

$$\sum_{i=k}^{n} \binom{n}{i} p_{lower}^{i}(1-p_{lower})^{n-i} = \frac{\alpha}{2} \tag{3.4}$$

$$\sum_{i=0}^{k} \binom{n}{i} p_{upper}^{i}(1-p_{upper})^{n-i} = \frac{\alpha}{2} \tag{3.5}$$

where p_{lower} corresponds to $L(X)$ and p_{upper} corresponds to $U(X)$. For $k = 0$ and $k = n$, the solutions to the equations are explicit:

$$k = 0; p_{lower} = 0, p_{upper} = 1 - (\alpha/2)^{1/n} \tag{3.6}$$

$$k = n; p_{lower} = (\alpha/2)^{1/n}, p_{upper} = 1 \tag{3.7}$$

For the other cases, the solutions can be determined by the relation [24]:

$$1 - P(X \leq k-1) = P(X \geq k) \tag{3.8}$$

$$= \sum_{i=k}^{n} \binom{n}{i} p^{i}(1-p)^{n-i} = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} \int_{0}^{p} t^{k-1}(1-t)^{n-k}dt \tag{3.9}$$

$$= I_p(k, n-k+1) = P(X \leq p) \tag{3.10}$$

$$\rightarrow P(X \geq k) = I_p(k, n-k+1) \tag{3.11}$$

where $I_y(a, b)$, also called the regularised incomplete beta function, denotes the cumulative distribution function (CDF) of a beta random variable X with parameters $a > 0$ and $b > 0$. The limits of the interval are equal to the quantiles of the Beta distribution. As result, the Clopper-Pearson confidence interval is given by [24]:

$$(p_{lower}, p_{upper}) = (\beta_{\alpha/2}(k, n-k+1), \beta_{1-\alpha/2}(k+1, n-k)) \tag{3.12}$$

It is not possible to obtain a confidence interval with the exactly specified confidence level, because there is no closed-form solution, but it is possible to construct a confidence interval that has a coverage probability of at least $(1-\alpha)$ [24].

A proof of the relation in equation 3.11 is described in the report of Scholz [26] in which two facts are proven:

1. Let $x(p) = P(X \geq k)$ and $y(p) = I_p(k, n-k+1)$. Firstly, it is proven that:

$$x'(p) = \frac{\partial P(X \geq k)}{\partial p} = \frac{\partial I_p(k, n-k+1)}{\partial p} = y'(p) \; \forall p \geq 0. \tag{3.13}$$

2. In addition, it is proven that:

$$x(p) = P(X \geq k) = I_p(k, n-k+1) = y(p) \text{ for } \; p = 0. \tag{3.14}$$

3. From 1 and 2, it can be concluded that $P(X \geq k) = I_p(k, n-k+1)$ for all values of $p \geq 0$ which proves the relation in equation 3.11.

The conclusion in point 3 follows from the fact that $x'(p) = y'(p) = f(p)$ which results in $x(p) = y(p)+C$. Furthermore, it is proven that $x(0) = y(0)$ from which follows that $C = 0$. Therefore, it can be concluded that $x(p) = y(p) \; \forall p \geq 0$ since the functions are continuous which implies the uniqueness of the solution. A proof of equation 3.11 from which equation 3.12 follows, is given in Appendix C.

3.2.3 Normal approximation methods

Methods often referred to in literature to determine confidence interval are based on the approximation of the $Bin(n,p)$ by the $N(np, np(1-p))$ distribution [24]:

$$\frac{k-np}{\sqrt{np(1-p)}} = \frac{\frac{k}{n}-p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{d} N(0,1) \tag{3.15}$$

and

$$P\left(-z_{1-\alpha/2} \leq \frac{\frac{k}{n}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{1-\alpha/2}\right) \approx 1-\alpha \tag{3.16}$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ represent the z scores for the standard normal distribution which describe how far the score is apart from the mean in units of standard deviations. This method of approximating the binomial distribution, which is discrete, with the normal distribution, which is continuous, is based on the Central Limit Theorem (CLT) and is unreliable when the sample size is not sufficiently large or when the population proportion p is close to 0 or 1. A rule of thumb when these methods may be used is when $np > 5$ and $n(1-p) > 5$ [4]. A well-known normal approximation method is Wald's method.

Wald's method
Using $\hat{p} = \frac{k}{n}$, the Wald's confidence interval is given by [24]:

$$(p_{lower}, p_{upper}) = \hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{k}{n} \pm z_{1-\alpha/2}\sqrt{\frac{k}{n^2}\left(1-\frac{k}{n}\right)} \tag{3.17}$$

It is important to note that equation 3.17 results in values outside the interval [0,1] which is not possible for proportions [16]. To avoid this, Wald's lower and upper bounds are adapted to the formulas in Table 3.2. The lower and upper bound of Wald's interval can be improved by introducing Yates' continuity correction, replacing $\hat{p}$ by $\tilde{p} = \hat{p} \pm \frac{1}{2n}$ [24]. A Continuity Correction has to be applied when a continuous distribution is used to approximate a discrete distribution. When the sample size is larger than 5% of the total population when sampling without replacement from a finite population, the estimate of the standard error must be corrected by multiplying by a Finite Population Correction, $FPC = \sqrt{\frac{N-n}{N-1}}$ to scale the variance $\frac{\hat{p}(1-\hat{p})}{n} * FPC^2 = \frac{\hat{p}(1-\hat{p})}{n}\frac{N-n}{N-1} = \frac{\hat{p}(1-\hat{p})}{n\frac{N-1}{N-n}}$, replacing n by $\tilde{n} = n\frac{N-1}{N-n}$ [41]. The Finite Population Correction captures the difference between sampling with and without replacement since the correction term is close to 1 unless the sample is a significant part, i.e. larger than 5%, of the population which is illustrated in Table 3.1 [20].

n	**FPC**
10	1.000
50	0.998
100	0.995
250	0.987
500	0.975
1000	0.949
5000	0.707
8000	0.447

Table 3.1: The Finite Population Correction term for different sample sizes n for a population size $N = 10,000$.

Wilson Score method
The Wilson score method was proposed to determine an asymmetric confidence interval for the proportion $\hat{p} = \frac{k}{n}$ and is given by [39]:

$$(p_{lower}, p_{upper}) = \frac{2n\hat{p} + z^2_{1-\alpha/2} \pm z_{1-\alpha/2}\sqrt{4n\hat{p}(1-\hat{p}) + z^2_{1-\alpha/2}}}{2(n + z^2_{1-\alpha/2})} \tag{3.18}$$

$$= \frac{2k + z^2_{1-\alpha/2} \pm z_{1-\alpha/2}\sqrt{4k(1-k/n) + z^2_{1-\alpha/2}}}{2(n + z^2_{1-\alpha/2})} \tag{3.19}$$

Yates' continuity correction and the finite population correction can also be applied to the Wilson score interval.

Agresti-Coull method
For a 95% confidence interval, $z^2 = 1.96^2 \approx 4$ and the midpoint of the Wilson Score interval is equal to $(k + z^2/2)/(n + z^2) \approx (k + 2)/(n + 4)$. Agresti-Coull suggested a simple but effective method of adding 4 observations to the sample, $\tilde{n} = n + z^2$ of which 2 are successes and 2 failures, $\tilde{k} = k + \frac{z^2}{2}$, then $\tilde{p} = \frac{\tilde{k}}{\tilde{n}} = \frac{1}{n+z^2}(k + \frac{z^2}{2})$ and then applying Wald's confidence interval [24]:

$$(p_{lower}, p_{upper}) = \frac{k+2}{n+4} \pm z_{1-\alpha/2}\sqrt{\frac{k+2}{(n+4)^2}\left(1 - \frac{k+2}{n+4}\right)} \tag{3.20}$$

Arcsine method
The Arcsine method is based on the approximate normal distribution after a stabilising transformation has been applied to the variance [24]:

$$arcsin\sqrt{\frac{k}{n}} \xrightarrow{d} N\left(arcsin\sqrt{p}, \frac{1}{4n}\right) \tag{3.21}$$

Applying Yates' continuity correction improves the performance resulting in higher coverage probabilities [24, 2]. The finite population correction can also be applied to the Arcsine interval.

The lower and upper bounds of the confidence interval for the proportion p of the discussed methods are summarised in Table 3.2, which are also presented in the paper [24].

Methods	**Lower bound L(X)**	**Upper bound U(X)**
Clopper-Pearson		
$k = 0$	0	$1 - (\alpha/2)^{1/n}$
$0 < k < n$	$\beta_{\alpha/2}(k, n-k+1)$	$\beta_{1-\alpha/2}(k+1, n-k)$
$k = n$	$(\alpha/2)^{1/n}$	1
Wald	$max\left(\frac{k}{n} - z_{1-\alpha/2}\sqrt{\frac{k}{n^2}(1-\frac{k}{n})}; 0\right)$	$min\left(\frac{k}{n} + z_{1-\alpha/2}\sqrt{\frac{k}{n^2}(1-\frac{k}{n})}; 1\right)$
Agresti-Coull	$max\left(\frac{k+2}{n+4} - z_{1-\alpha/2}\sqrt{\frac{k+2}{(n+4)^2}\left(1-\frac{k+2}{n+4}\right)}; 0\right)$	$min\left(\frac{k+2}{n+4} + z_{1-\alpha/2}\sqrt{\frac{k+2}{(n+4)^2}\left(1-\frac{k+2}{n+4}\right)}; 1\right)$
Wilson Score	$\frac{2k+z^2_{1-\alpha/2} - z_{1-\alpha/2}\sqrt{4k(1-k/n)+z^2_{1-\alpha/2}}}{2(n+z^2_{1-\alpha/2})}$	$\frac{2k+z^2_{1-\alpha/2} + z_{1-\alpha/2}\sqrt{4k(1-k/n)+z^2_{1-\alpha/2}}}{2(n+z^2_{1-\alpha/2})}$
Arcsine		
$k = 0$	0	$sin^2\left(min\left(arcsin(\sqrt{k/n}) + \frac{z_{1-\alpha/2}}{2\sqrt{n}}; \frac{\pi}{2}\right)\right)$
$0 < k < n$	$sin^2\left(max\left(arcsin(\sqrt{k/n}) - \frac{z_{1-\alpha/2}}{2\sqrt{n}}; 0\right)\right)$	$sin^2\left(min\left(arcsin(\sqrt{k/n}) + \frac{z_{1-\alpha/2}}{2\sqrt{n}}; \frac{\pi}{2}\right)\right)$
$k = n$	$sin^2\left(max\left(arcsin(\sqrt{k/n}) - \frac{z_{1-\alpha/2}}{2\sqrt{n}}; 0\right)\right)$	1

Table 3.2: The lower and upper bounds $L(X)$ and $U(X)$ of a $(1-\alpha) \times 100\%$ confidence level for a two-sided confidence interval for p for different confidence interval methods.

3.2.4 Simulation study on Coverage Probability and Expected Length

To compare the different methods to determine a confidence interval, a simulation study is conducted in which the different confidence interval methods are compared on the following two criteria:

- Coverage Probability, as defined in equation 3.2.
- Expected Length, as defined in equation 3.3.

To determine which confidence interval method is most appropriate to use in this thesis, three requirements are formulated which are as follows:

Requirements for confidence interval method

- Firstly, the confidence interval method should guarantee a Coverage Probability at least equal to the nominal Coverage Probability.
- In addition, there is a preference for a confidence interval method that has the smallest EL when more methods have approximately the same Coverage Probability.
- Lastly, it is preferred that the confidence interval method can always be used, independent of the population proportion of interesting transactions.

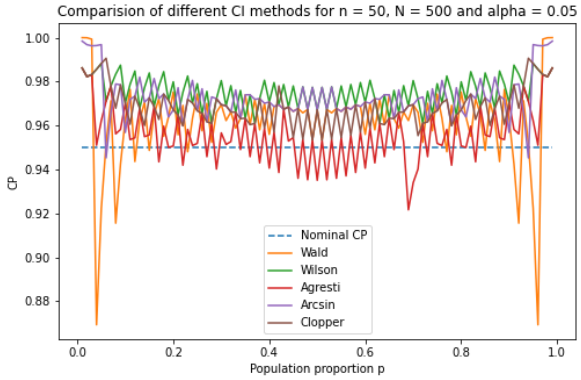
Simulation study

For the simulation study, the Coverage Probability for the five confidence interval methods, Wald, Wilson Score, Agresti-Coull, Arcsine and Clopper-Pearson, is determined via equation 3.2. Subsequently, the Expected Length was also determined for the five confidence interval methods via equation 3.3. The simulation study is conducted for a small, medium, and large sample size, $n = 50, n = 250$, and $n = 500$, to assess the influence of a larger sample size on the accuracy of the Coverage Probability and the influence on the size of the Expected Length of a confidence interval.
In addition, the simulation study is conducted for a small, medium, and large population, $N = 500, N = 1000$, and $N = 10,000$, to assess the influence of a larger population on the accuracy of the Coverage Probability and the influence on the size of the Expected Length of a confidence interval. Yates' continuity correction and the finite population correction are applied to the confidence interval based on the normal approximation to ensure that the Coverage Probability should not become less accurate for a smaller population N. The results for the Coverage Probability for samples size 50, 250, and 500 and population size 500, 1000, and 10,000 for a 95% confidence interval are illustrated in Figure 3.2. The Coverage Probability (CP) is shown on the y-axis and the population proportion p on the x-axis for $p = [0.01, 0.02, ..., 0.98, 0.99]$.

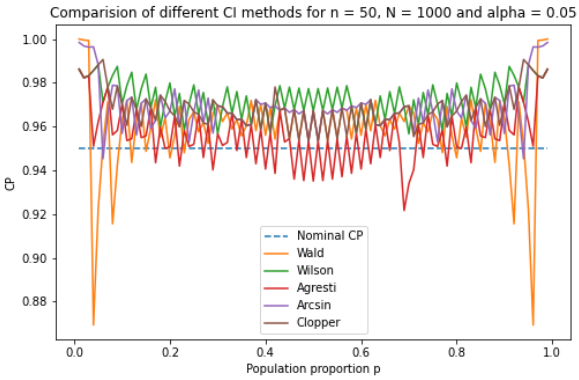
Simulation study results: based on the Coverage Probability (CP)
Figure 3.2 shows that the CP of the Wald confidence interval with continuity and finite population correction for values close to $p = 0$ and $p = 1$ is reasonably smaller than the nominal CP for a smaller sample size, which is visible in Figures 3.2a, 3.2b and 3.2c. Furthermore, the CP of the Agresti-Coull confidence interval with continuity and finite population correction is around the nominal CP but is regularly smaller than the nominal CP. In addition, the CP of the Arcsine confidence interval with continuity and finite population correction is often above the nominal CP but can be smaller than the nominal CP. Finally, an analysis of the CP for the Wilson Score with continuity and finite population correction and Clopper-Pearson confidence interval shows that the minimal CP is always larger than the nominal CP. Based on the results of Figure 3.2, the five confidence interval methods can be classified as follows:

- **1st group - strictly conservative methods:** the Wilson Score with continuity and finite population correction and the Clopper-Pearson methods belong to this group for a 95% confidence interval. It is also stated in literature that the Clopper-Pearson confidence interval is a strictly conservative method [1].
- **2nd group - on average correct method:** the Arcsine and Agresti-Coull with continuity and finite population correction methods belong to this group for a 95% confidence interval. The Agresti-Coull confidence interval is currently mainly recommended for large samples [16].
- **3rd group - other:** the Wald with continuity and finite population correction method belongs to this group for a 95% confidence interval. Furthermore, it is important to note that for $k = 0$ or $k = n$, Wald's interval has zero length. In these two cases, it is advisable to use the Clopper-Pearson equations. Finally, Wald's confidence interval is unsuitable to use for small samples and skewed data [41].

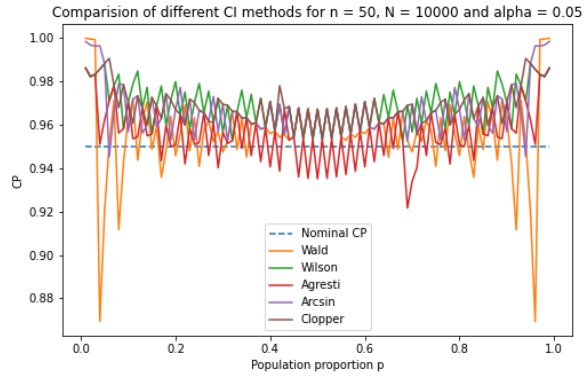
These results are in line with the results in the papers [24, 16] in which the Wilson Score with continuity correction and Clopper-Pearson are classified as strictly conservative methods, the Arcsine transform with continuity correction and Agresti-Coull are classified as on average correct methods and Wald's method is classified as the third group for a 95% confidence interval.
Based on the results of the CP, the Wilson Score with continuity and finite population correction and the Clopper-Pearson methods are the most reliable and therefore the most suitable to be used to determine a confidence interval for the proportion of interesting transactions. Since the CP of both methods is comparable, an analysis of the Expected Length, i.e the expected width of a confidence interval, is performed, in which there is a preference for the confidence interval method with the smallest EL. The results for the Expected Length for samples size 50, 250, and 500 and population size 500, 1000, and 10,000 for a 95% confidence interval are illustrated in Figure 3.3.



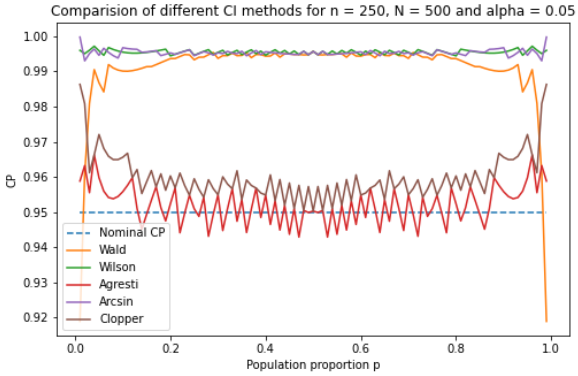
(a) n = 50, N = 500



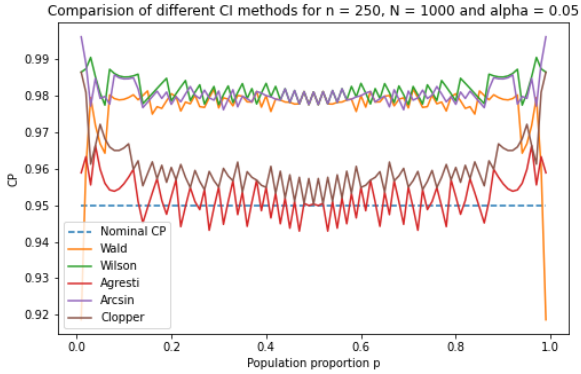
(b) n = 50, N = 1000



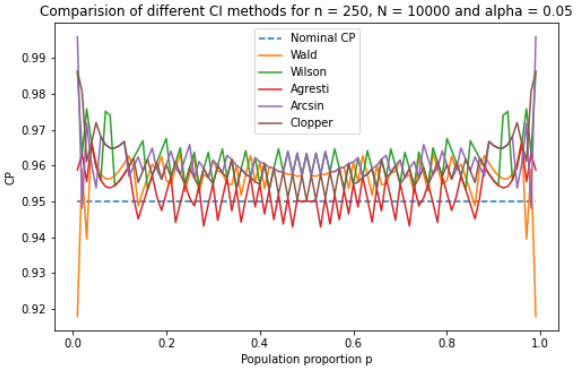
(c) n = 50, N = 10000



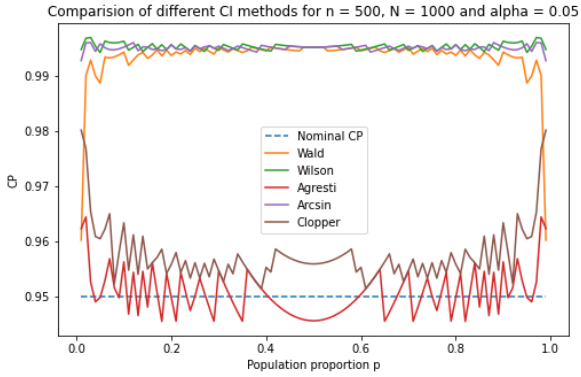
(d) n = 250, N = 500



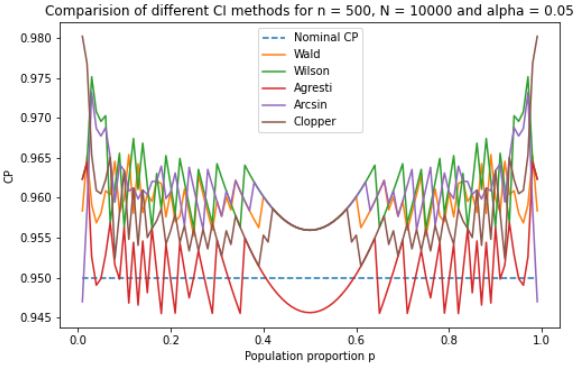
(e) n = 250, N = 1000



(f) n = 250, N = 10000



(g) n = 500, N = 1000



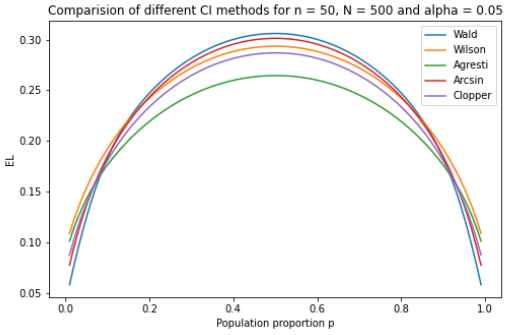
(h) n = 500, N = 10000

Figure 3.2: The Coverage Probabilities (CP) of Wald, Agresti, Wilson Score, Arcsine and Clopper-Pearson confidence interval methods for $\alpha = 0.05$.

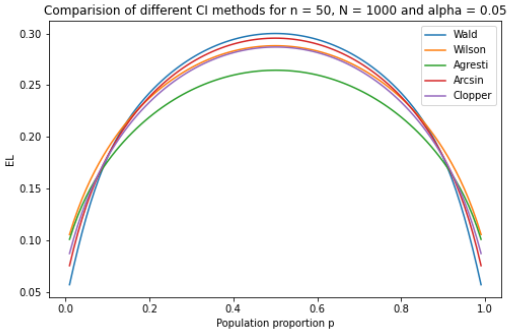
Simulation study results: based on the Expected Length (EL)
From an analysis of Figure 3.3, it is visible that a larger sample size results in a smaller EL, comparing Figures 3.3a, 3.3b, and 3.3c with Figures 3.3g and 3.3h.
In addition, it becomes clear that the EL is comparable for the Clopper-Pearson and the Wilson Score confidence interval with continuity and finite population correction when the sample size is a small fraction of the population, i.e. the sample size is less than or equal to 5 per cent of the population, which is the case in Figures 3.3b, 3.3c, 3.3f, and 3.3h. However, in the case that the sample size is a large fraction of the population, i.e. the sample size is more than 5 per cent of the population, the difference in the EL for the Clopper-Pearson and the Wilson Score confidence interval with continuity and finite population correction is clearly visible in Figures 3.3d, 3.3e, and 3.3g, whereby the EL for the Clopper-Pearson confidence interval is a reasonably smaller than the EL for the Wilson Score confidence interval.
Lastly, the EL of the Clopper-Pearson confidence interval is smaller than the EL for the Wilson Score confidence interval for a sample size of 250 or larger, even though it is negligible when the sample size is less than 5 per cent of the population.

As a final remark, the choice of the confidence interval method depends on the desired degree of conservativeness and which population proportions are expected. If it is expected that the population proportion will not be close to 0 or 1, then the Wilson Score confidence interval could be used. If for instance $CP(n, p) \geq 1 - \alpha$ is mandatory, Clopper-Pearson is an appropriate method.

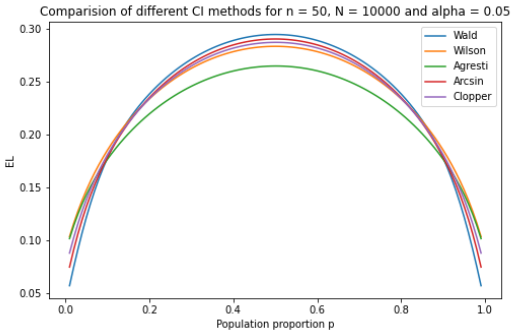
Conclusion
To summarise, the requirements for the most suitable confidence interval method were described at the beginning of Section 3.2.4. In this thesis, the Clopper-Pearson confidence interval method is chosen to determine the confidence interval for the population proportion of interesting transactions, because this method always guarantees a CP that is at least equal to the nominal CP. Secondly, the Clopper-Pearson method has the smallest EL. Lastly, a considerable advantage is that the Clopper-Pearson method can always be used, regardless of the proportion of interesting transactions in the population. Although it is unlikely that the proportion of interesting transactions in a population will be close to 1, proportions close to 0 cannot be excluded and the conditions $np > 5$ and $n(1 - p) > 5$ cannot be guaranteed to be satisfied.



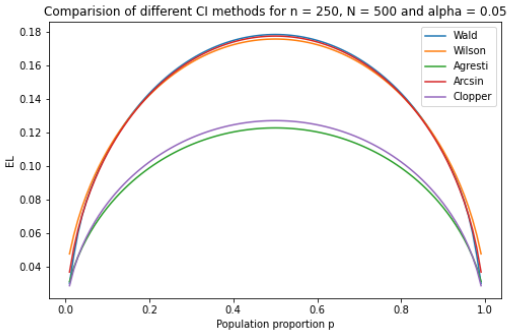
(a) n = 50, N = 500



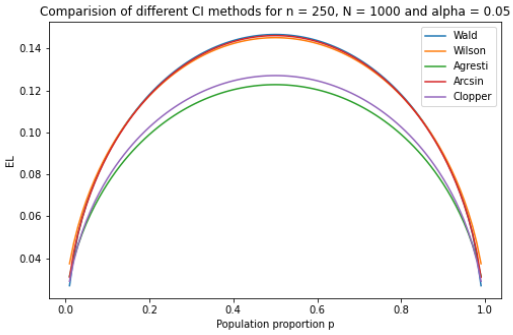
(b) n = 50, N = 1000



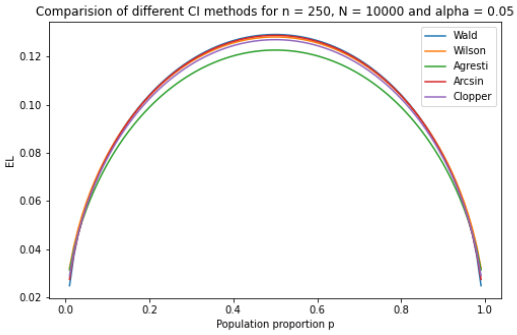
(c) n = 50, N = 10000



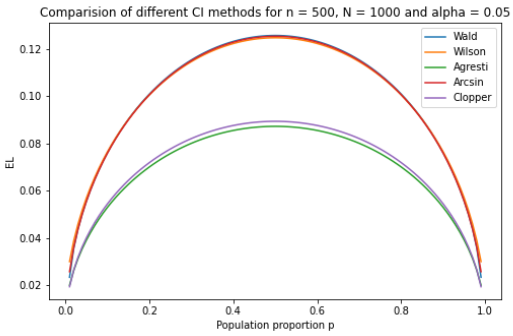
(d) n = 250, N = 500



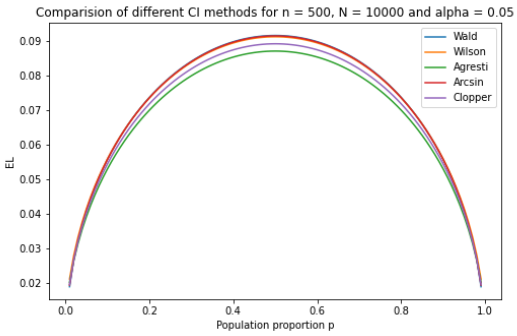
(e) n = 250, N = 1000



(f) n = 250, N = 10000



(g) n = 500, N = 1000



(h) n = 500, N = 10000

Figure 3.3: The Expected Length (EL) for Wilson Score and Clopper-Pearson confidence interval methods for $\alpha = 0.05$.

3.2.5 Translation to sample size methods

Before the confidence interval limits of the sample estimate can be determined, a sample has to be taken from the population of transactions. The required sample size n, which is the number of transactions that need to be reviewed by an analyst, can be determined using the confidence intervals described in Table 3.2. A well-known method to determine the required sample size is Wald's confidence interval.

Wald's sample size

The required sample size can be determined by inverting Wald's confidence interval where the margin of error e represented by equation 3.23 is defined as one-half the length of the confidence interval. To determine the required sample size, with certain Coverage Probability and with a desired margin of error e, it is necessary to replace the population approximation $\hat{p}$ with a preliminary point estimate for p_0 for the population proportion of interesting transactions. Rewriting equation 3.23, whereby, the population approximation $\hat{p}$ is replaced by preliminary point estimate p_0 results in the required sample size n presented in equation 3.25, rounded up to the nearest integer.

$$(p_{lower}, p_{upper}) = p_0 \pm z_{1-\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}} \tag{3.22}$$

$$e = z_{1-\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}} \tag{3.23}$$

$$n = \left\lceil \frac{(z_{1-\alpha/2})^2 p_0(1-p_0)}{e^2} \right\rceil \tag{3.24}$$

When it is difficult to make a preliminary point estimate p_0 because no previous study on the population proportion is available, the conservative $p_0 = 0.5$ can be chosen, which maximises equation 3.24.

The sample size formula that follows from Wald's adapted upper and lower bounds in Table 3.2 is given by [16]:

$$n = \begin{cases} \left\lceil \frac{(z_{1-\alpha/2})^2 p_0(1-p_0)}{(e-p_0)^2} \right\rceil, 0 \leq p_0 < \frac{e}{2} \\ \left\lceil \frac{(z_{1-\alpha/2})^2 p_0(1-p_0)}{e^2} \right\rceil, \frac{e}{2} \leq p_0 \leq 1 - \frac{e}{2} \\ \left\lceil \frac{(z_{1-\alpha/2})^2 p_0(1-p_0)}{(e-(1-p_0))^2} \right\rceil, 1 - \frac{e}{2} < p_0 \leq 1 \end{cases} \tag{3.25}$$

Agresti-Coull sample size

Applying this approach to the Agresti-Coull confidence interval to determine the required sample size results in the same required sample size as when Wald's confidence interval method is used, because the Agresti-Coull confidence interval is the same as Wald's confidence interval, only with a modified proportion $\tilde{p} = \frac{\tilde{k}}{\tilde{n}} = \frac{1}{n+z^2}(k + \frac{z^2}{2})$.

Wilson Score sample size

It is not possible to invert the Wilson Score confidence interval to get an analytic solution for the required sample size. However, the sample size can be determined with another approach. Given the Wilson Score confidence interval in equation 3.18 in which the population approximation $\hat{p}$ is replaced with a preliminary point estimate p_0:

$$(p_{lower}, p_{upper}) = \frac{2np_0 + z^2_{1-\alpha/2} \pm z_{1-\alpha/2}\sqrt{4np_0(1-p_0) + z^2_{1-\alpha/2}}}{2(n + z^2_{1-\alpha/2})} \tag{3.26}$$

The required sample size n to get a certain Coverage Probability and with a desired margin of error e, can be determined by solving the following system of non-linear equations [39]:

$$\begin{cases} p_{upper} \leq \frac{2np_0+z^2_{1-\alpha/2}+z_{1-\alpha/2}\sqrt{4np_0(1-p_0)+z^2_{1-\alpha/2}}}{2(n+z^2_{1-\alpha/2})} \\ p_{lower} \geq \frac{2np_0+z^2_{1-\alpha/2}-z_{1-\alpha/2}\sqrt{4np_0(1-p_0)+z^2_{1-\alpha/2}}}{2(n+z^2_{1-\alpha/2})} \\ p_{upper} - p_{lower} < 2e \end{cases} \tag{3.27}$$

This system of non-linear equations can be solved using a solver function such as $scipy.optimize.fsolve()$ in Python. The sample size that follows from solving the system of non-linear equations will be rounded up to the nearest integer.

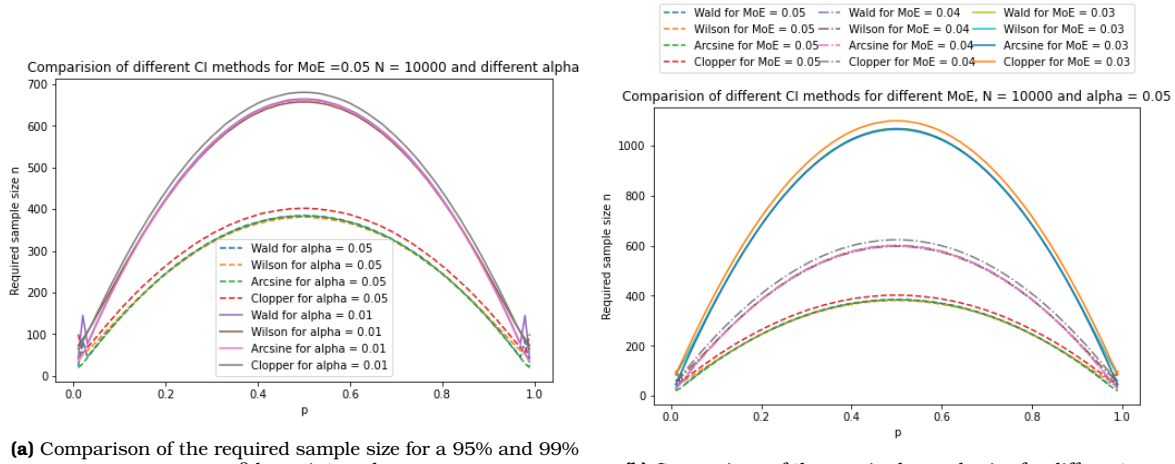
Arcsine and Clopper-Pearson sample size

This approach of solving a system of non-linear equations can also be applied to the Arcsine and Clopper-Pearson confidence interval to determine the required sample size by replacing the right-hand side of the upper two equations with the lower and upper bounds of the Arcsine and Clopper-Pearson confidence interval as presented in Table 3.2. This approach results in three systems of non-linear equations for the Arcsine and Clopper-Pearson confidence interval, each system for one of the three cases. The three systems of non-linear equations applied to the Arcsine and Clopper-Pearson confidence interval are described in Appendix B.

It is important to note that the Wald, Agresti-Coull, Arcsine and Wilson Score confidence interval methods are based on the normal approximation of the binomial distribution. The rule of thumb for using these methods is when $np > 5$ and $n(1-p) > 5$ and p is not close to 0 or 1. Therefore, it is recommended not to use these methods to determine the sample size when it is expected that the population proportion of interesting transactions could be close to 0 or 1.

Sample size results

For $\alpha = 0.05$ and $\alpha = 0.01$ and a margin of error equal to $0.05, 0.04$ and 0.03 the required sample sizes that follow from equation 3.25 and equation, 3.27 in which the upper and lower bounds of a chosen confidence interval method are applied to the right-hand side of the upper two equations for preliminary point estimates $0.01 < p_0 < 0.99$ are presented in Figure 3.4. The sample size using the Agresti-Coull method is not shown in the figure, since this method results in the same sample size as Wald's method.



(a) Comparison of the required sample size for a 95% and 99% confidence interval.

(b) Comparison of the required sample size for different e.

Figure 3.4: A comparison of the required sample size with different confidence interval methods for a 95% or 99% confidence interval with a margin of error $e = 0.05, 0.04$ and 0.03

First of all, Figure 3.4a shows that the required sample size increases for a confidence interval with a higher confidence level, $1 - \alpha$. The required sample size approximately doubles for a 99% confidence

interval compared to a 95% confidence interval. This is as expected since the confidence interval must in this case contain the population parameter 99 out of 100 times instead of 95 out of 100 times, which results in a larger required sample size.

In addition, Figure 3.4b shows that the required sample size increases for a smaller margin of error which is equivalent to a narrower confidence interval. This is also as expected since the confidence interval becomes less wide when a larger sample of the population is taken.
Figures 3.4a and 3.4b are presented separately for each confidence interval method in Figures B.1 and B.2 in Appendix B.1.

The required sample sizes that are determined with different confidence interval methods for preliminary point estimates between $0.05 < p_0 < 0.5$ are presented in Table 3.3 for a 95% confidence level and in Table 3.4 for a 99% confidence with a margin of error equal to $e = 0.05$. The required sample size determined with the Agresti-Coull method is not presented in the table, since this method results in the same sample size as Walds method. Tables 3.3 and 3.4 show that the required sample size determined with Wald's, Wilson Score or the Arcsine method, results in a comparable sample size and the required sample size determined with Clopper-Pearson confidence interval results in the largest sample size.

p_0	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**	**0.5**
Wald	73	139	196	246	289	323	350	369	381	385
Wilson	83	141	196	245	286	320	347	366	377	381
Arcsine	72	138	195	245	287	322	349	368	380	383
Clopper-Pearson	94	158	215	264	306	341	367	387	398	402

Table 3.3: The required sample size with different confidence interval methods for $\alpha = 0.05$ and $e = 0.05$ for preliminary point estimates p_0.

p_0	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**	**0.5**
Wald	127	239	339	425	498	558	604	637	657	664
Wilson	144	244	339	422	494	552	598	631	651	657
Arcsine	124	237	337	423	496	556	602	635	655	662
Clopper	149	258	356	442	515	574	620	654	673	680

Table 3.4: The required sample size with different confidence interval methods for $\alpha = 0.01$ and $e = 0.05$ for preliminary point estimates p_0.

Conclusion
In this thesis, the Clopper-Pearson confidence interval is chosen to determine the confidence interval upper bound for the unknown population proportion of interesting transactions. Therefore, in this thesis, the Clopper-Pearson confidence interval will also be used to determine the required sample size because the determined sample size with the Clopper-Pearson confidence interval guarantees a $(1 - \alpha) \times 100\%$ confidence level only for the Clopper-Pearson confidence interval. To determine the sample size for BTL testing, it is chosen to set the preliminary point estimate p_0 equal to the RTL. With BTL threshold tuning, different threshold levels are tested to determine the threshold level for which the upper bound of the false omission rate (proportion of FN transactions) is smaller than the RTL, i.e $p_{upper} < RTL$. By choosing $p_0 = RTL$, the true population proportion p, which will be smaller than p_{upper} with $(1 - \alpha) \times 100\%$ certainty, will also be smaller than p_0 with $(1 - \alpha) \times 100\%$ certainty. As consequence, the sample size determined with p_0 results in a confidence interval with a width that is at most two times the margin of error. Lastly, the same sample size will be chosen for ATL and BTL testing in this thesis.

3.3 Confidence interval using bootstrapping

The Clopper-Pearson method is not suitable to determine the lower and upper confidence interval limits for the measures sensitivity, specificity and accuracy. The reason for this is that these measures express a ratio of two random variables with unknown distributions. Therefore, a different approach is required to determine a confidence interval for these measures. In literature, various parametric approximations for constructing confidence intervals around a ratio estimator have been proposed. However, the sampling distribution of these measures is unknown which is a reason to be cautious with parametric techniques [3]. A non-parametric bootstrap confidence interval is potentially a more suitable method to construct a confidence interval for these measures since it does not depend on parametric assumptions of the sampling distribution [3].

3.3.1 Bootstrap estimation of the sampling distribution

Bootstrapping, a computational non-parametric technique for 're-sampling', makes it possible to draw a conclusion about the characteristics of a population strictly from the existing sample rather than by making parametric assumptions about the estimator.

Suppose a particular population which has an unobserved probability distribution F from which a random sample S of size n is taken. The aim is to determine the statistic of interest to make statements about population parameter θ based on the sample from the population which is illustrated in the first three steps in Figure 3.5. The bootstrap method considers the observed random sample as an empirical estimate of the probability distribution [3]. Afterwards, B random samples of size n are drawn from S with replacement which results in the bootstrap samples as illustrated in steps four and five in Figure 3.5. Subsequently, the statistic of interest θ is calculated for each bootstrap sample. The empirical distribution of the resulting values $\theta_1^*, \theta_1^*, ..., \theta_B^*$ is a good approximation of the sampling distribution $\hat{\theta}$ if B is large as illustrated in steps six and seven in Figure 3.5. A confidence interval can then be constructed using the sampling distribution.

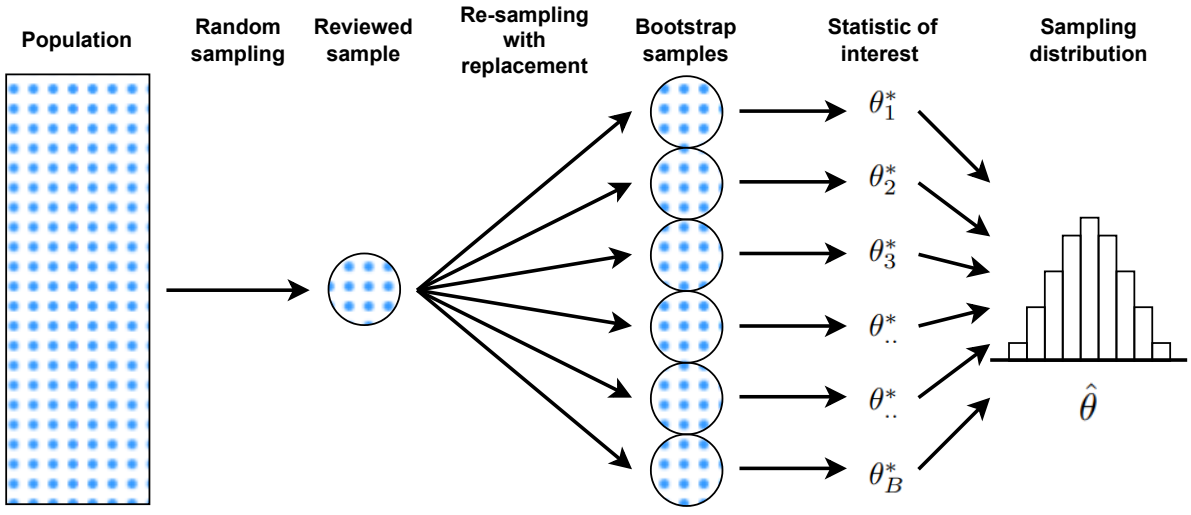


Figure 3.5: Illustration of the bootstrap method for a tested threshold level to determine the sampling population of the statistic of interest.

The sample estimate for sensitivity, also known as true positive rate (TPR), specificity, also known as true negative rate (TNR), and accuracy (ACC) can be determined via the following formulas, as already briefly described in Figure 2.5:

$$\hat{TPR} = \frac{TP}{TP+FN} = \frac{\hat{p_a}n}{\hat{p_a}n + \hat{p_b}n} = \frac{\hat{p_a}}{\hat{p_a} + \hat{p_b}} \quad \text{sensitivity} \tag{3.28}$$

$$\hat{TNR} = \frac{TN}{TN+FP} = \frac{(1-\hat{p_b})n}{(1-\hat{p_b})n + (1-\hat{p_a})n} = \frac{(1-\hat{p_b})}{(1-\hat{p_b}) + (1-\hat{p_a})} \quad \text{specificity} \tag{3.29}$$

$$\begin{aligned} \hat{ACC} &= \frac{TP+TN}{TP+TN+FP+FN} = \frac{\hat{p_a}n + (1-\hat{p_b})n}{\hat{p_a}n + (1-\hat{p_b})n + (1-\hat{p_a})n + \hat{p_b}n} \\ &= \frac{\hat{p_a} + (1-\hat{p_b})}{\hat{p_a} + (1-\hat{p_b}) + (1-\hat{p_a}) + \hat{p_b}} \quad \text{accuracy} \end{aligned} \tag{3.30}$$

where $\hat{p_a}$ is the proportion of interesting transactions in the sample above the tested threshold, i.e. the positive predictive value, and $\hat{p_b}$ the proportion of interesting transactions in the sample below the tested threshold, i.e. the false omission rate.

Suppose that the sensitivity is the measure of interest, i.e. $\theta = TPR$. Since the sample estimate for sensitivity is estimated using two statistics $\hat{p_a}$ and $\hat{p_b}$ from two samples, one sample below the tested threshold level S_b and one sample above the tested threshold level S_a, the bootstrap estimate of the sampling distribution of the sensitivity can be obtained using the following procedure:

Two-stage bootstrap process:

1. Sample with replacement n observations from both the BTL and ATL samples S_b and S_a to calculate p_a^* and p_b^* which are the bootstrap estimates of $\hat{p_a}$ and $\hat{p_b}$.
2. The bootstrap estimate of the sensitivity is then given by:

$$\theta^* = \frac{p_a^*}{p_a^* + p_b^*} \tag{3.31}$$

Repeating this two-stage process B times results in a vector of bootstrap estimates which is the empirical sampling distribution of the sensitivity statistic $\hat{\theta}$. The *expectation* of the empirical sampling distribution is denoted by [3]:

$$\bar{\theta}^* = \frac{1}{B}\sum_{i=1}^{B} \theta^{*i} \tag{3.32}$$

The bootstrap estimate of the *standard error* of the estimator is given by the standard deviation of the empirical sampling distribution [3]:

$$\hat{\sigma}^* = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\bar{\theta}^* - \theta^{*i})^2} \tag{3.33}$$

The approximation of the sampling distribution becomes arbitrarily accurate by taking B arbitrarily large [18]. In practice, there are no formal rules regarding the number of required bootstrap replications for a reliable estimation of the sampling distribution.

3.3.2 Bootstrap confidence interval

After determining the sampling distribution for the estimator expressed as a ratio, e.g. the sensitivity, of a tested threshold level, the paper of Briggs [3] describes four commonly used methods to determine a confidence interval for the estimator.

1. **Normal approximation method:** the idea of this method is to take the bootstrap estimate of standard error and assume that the sampling distribution of the estimator is normal. The Wald confidence interval then results in a $(1-\alpha)\times 100\%$ confidence interval for the estimator. However, this method may be misleading if the sampling distribution of the estimator is not normal [3].

This results in the $(1-\alpha)\times 100\%$ confidence interval:

$$\left(\hat{\theta} - z_{\alpha/2}\hat{\sigma^*}, \hat{\theta} + z_{\alpha/2}\hat{\sigma^*}\right) \tag{3.34}$$

2. **Percentile method:** the idea of this method is to use the quantiles of the sampling distribution $\hat{\theta}$. The $(\frac{\alpha}{2})\times 100$ and $(1-\frac{\alpha}{2})\times 100$ quantile values of the bootstrap sampling distribution $\hat{\theta}$ are used as upper and lower confidence interval limits:

$$\left(\theta^*_{\alpha/2}, \theta^*_{1-\alpha/2}\right) \tag{3.35}$$

The advantage of this method is its simplicity. However, the percentile method assumes that the bootstrap replicates of the estimator are unbiased [3], making this method not suitable if the ratio estimators $\theta^*_1, \theta^*_1, ..., \theta^*_B$ are biased.

3. **Bias corrected and accelerated (BCa) percentile method:** this is a modification of the percentile method taking into account a possible bias and skewness correction in the sampling distribution of $\hat{\theta}$. The adjusted percentiles are given by [3]:

$$\alpha_1 = \Phi\left(\hat{z} + \frac{\hat{z} + z_{\alpha/2}}{1 - \hat{a}(\hat{z} + z_{\alpha/2})}\right) \tag{3.36}$$

$$\alpha_2 = \Phi\left(\hat{z} + \frac{\hat{z} + z_{1-\alpha/2}}{1 - \hat{a}(\hat{z} + z_{1-\alpha/2})}\right) \tag{3.37}$$

where Φ is the standard normal cumulative distribution function and z_α the 100α percentile point of the standard normal distribution. In addition, $\hat{z} = \Phi(Q)$ adjusts the sampling distribution for the bias of the estimator, where Q is the proportion of bootstrap replicates which are less than the sample estimate $\hat{\theta}$ [3]. Moreover, $\hat{a}$ adjusts for the skewness of the sampling distribution $\hat{\theta}$. It is recommended to use a jackknife estimate for $\hat{a}$ [3]:

$$\hat{a} = \frac{\sum_{i=1}^{n}(\bar{\theta}^{**} - \hat{\theta}^{**}_i)^3}{6[\sum_{i=1}^{n}(\bar{\theta}^{**} - \hat{\theta}^{**}_i)^2]^{3/2}} \tag{3.38}$$

with $\bar{\theta}^{**} = (\sum_{i=1}^{n}\hat{\theta}^{**}_i)/n$ and $\hat{\theta}^{**}_i$ jackknife replicate of $\hat{\theta}$ without the i^{th} observation. The percentiles from equations 3.36 and 3.37 correspond to those from the percentile method if $\hat{a}$ and $\hat{z}$ are zero. Although the BCa percentile method does not make assumptions on the sampling distribution, it does make parametric assumptions on the distribution of the observed bias, which is a potential weakness of this method [3].

4. **Percentile-t method:** in this case is each bootstrap replicate θ^* of the estimator transformed into a standardised variable t^*, given by:

$$t^{*i} = \frac{\theta^{*i} - \hat{\theta}}{\hat{\sigma}^{*i}} \tag{3.39}$$

where $\hat{\sigma}^{*i}$ is calculated by another round of bootstrapping, requiring substantially more computations which is a considerable drawback of this method. This results in the percentile-t $(1-\alpha)\times 100\%$ confidence interval:

$$\left(\hat{\theta} - t^*_{1-\alpha/2}\hat{\sigma^*}, \hat{\theta} + t^*_{1-\alpha/2}\hat{\sigma^*}\right) \tag{3.40}$$

To summarise, if it is preferred to make a statement about the sensitivity, specificity, or accuracy of a tested threshold level with a $(1-\alpha)\times 100\%$ confidence level, bootstrapping and one of the above-mentioned bootstrap confidence interval methods can be used.

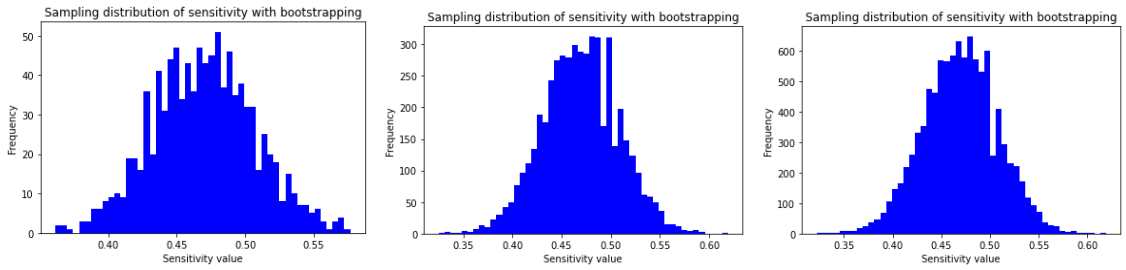
However, bootstrapping to determine the confidence interval bounds of the sampling distribution of

the sensitivity, specificity, or accuracy for each tested threshold level is computationally more intensive than for the false omission rate or positive predictive value with the Clopper-Pearson confidence interval. The difference in computation time is demonstrated in Section 3.3.3. Once the sampling distribution is determined, the confidence interval limits can be determined relatively easily if the sampling distribution of one of the chosen measures appears to have a normal distribution, allowing the application of the normal approximation method. It is also relatively simple to determine the confidence interval limits if the bootstrap replicates of the estimator are unbiased, making it possible to use the percentile method. However, if this is not the case, the BCa percentile method can be used, but the parameters $\hat{z}$ and $\hat{a}$ should be determined and this method has the drawback that it makes parametric assumptions on the distribution of the observed bias.

3.3.3 Simulation of computational time for bootstrap CI

A short simulation is performed to demonstrate the relative intensive computational time of bootstrapping to determine a confidence interval for the sensitivity, accuracy or specificity in comparison to the required computational time to determine a confidence interval for the false omission rate or positive predictive value using the Clopper confidence interval. In this simulation, a confidence interval for sensitivity is determined for case study 1, which is described in Chapter 5, for threshold level 10. First, a BTL and ATL sample is taken of sample size n and the transactions in the two samples were reviewed as interesting (1) or not (0) resulting in two labelled samples. The simulation is performed with $B = [1000, 5000, 10, 000]$ bootstrap samples.
Then, through the two-stage bootstrap process, the sampling distribution of the sensitivity can be determined as illustrated in Figure 3.6. It appears that the sampling distribution for sensitivity for case study 1 of threshold level 10 has a normal distribution which makes it possible to use the normal approximation method to determine a confidence interval for the sensitivity. Finally, a confidence interval can be determined for the sensitivity using equation 3.34 with a $(1 - \alpha) \times 100\%$ confidence level. The time required to determine the confidence interval for the sensitivity, the false omission rate and the positive predictive value are presented in Table 3.6. It can be concluded that the time required to determine the confidence interval for the false omission rate and positive predictive value with the Clopper-Pearson confidence interval is significantly shorter than the time required to determine the confidence interval for the sensitivity with bootstrapping. It is worth noting that the time depends on the exact algorithm, the computer used and whether other programmes are active, but these results do show a distinct difference in the computational time.

However, this computational time to determine a confidence interval for sensitivity is negligible in comparison to the time it takes an analyst to review the BTL and ATL samples of $n = 264$ transactions for one threshold level. Assuming an analyst takes at least 10 minutes to review one transaction, reviewing 2*264 transactions takes about 88 hours, compared to the 70 seconds to determine the confidence interval for sensitivity with bootstrapping with 10,000 bootstrap samples.



(a) B=1000 **(b)** B=5000 **(c)** B=10000

Figure 3.6: The sampling distribution of the sensitivity of case study 1 of threshold level 10 with $n = 264$.

B	θ	$\hat{\theta}^*$	**CI for** θ	p_b	$\hat{p_b}$	**CI for** p_b	p_a	$\hat{p_a}$	**CI for** p_a
1000	0.4885	0.4697	(0.3968, 0.5426)	0.2367	0.2614	(0.2094, 0.3188)	0.2261	0.2311	(0.1816, 0.2867)
5000	0.4885	0.4691	(0.3938, 0.5444)	0.2367	0.2614	(0.2094, 0.3188)	0.2261	0.2311	(0.1816, 0.2867)
10000	0.4885	0.4695	(0.3942, 0.5448)	0.2367	0.2614	(0.2094, 0.3188)	0.2261	0.2311	(0.1816, 0.2867)

Table 3.5: Confidence intervals determined with Clopper Pearson for p_a and p_b and with bootstrapping for θ, with $\theta = TPR$ (sensitivity) and $n = 264$.

B	**CI for** θ	**CI for** p_b	**CI for** p_a
1000	5.26868	0.00092	0.00258
5000	33.96111	0.00059	0.00065
10000	70.77295	0.00049	0.00035

Table 3.6: Required time in seconds to compute the confidence interval with Clopper Pearson for p_a and p_b and with bootstrapping for θ, with $\theta = TPR$ (sensitivity) and $n = 264$.

3.4 Chapter conclusions

In this section, a summary of the important choices that were made in this chapter is provided. The choices that were made, which will be used further in this thesis, are the following:

1. Simple random sampling is chosen as random sampling method.
2. The Clopper-Pearson confidence interval method is chosen as the confidence interval method to determine the confidence interval for the measures false omission rate or positive predictive value.
3. The required sample size is determined with the Clopper-Pearson method.
4. Bootstrapping with one of the four appropriate bootstrap confidence interval methods is chosen as the method to determine the confidence interval for the measures sensitivity, accuracy or specificity.

4

Methodology

In this chapter, different threshold tuning strategies used in this thesis are described in more detail. The current transaction monitoring process and back-testing via BTL and ATL testing discussed in Chapter 2 form the starting point for the tuning strategies discussed in this chapter. In this thesis, the main focus is on the false omission rate as measure to quantify the risk of missing interesting transactions below a threshold level, as stated at the end of Section 2.2.

This chapter is divided into several sections. First, the model framework, for the decision problem that is briefly introduced in Section 2.3, will be described in more detail in Section 4.1 and forms the basis for the different threshold tuning strategies.
To address the second research objective, two threshold evaluation approaches to quantify the risk of missing interesting transactions below a threshold level are discussed in Section 4.2 .
To address the first research objective, different tuning strategies are discussed in Sections 4.3 and 4.4. Two static threshold tuning strategies are discussed in Section 4.3. For these strategies, there is no exploration phase but only an exploitation phase to determine the most appropriate threshold level. Subsequently, in Section 4.4, two dynamic threshold tuning strategies are discussed, combining an exploration and exploitation phase to determine the most appropriate threshold level.
Furthermore, to accelerate the threshold tuning process, Section 4.5 discusses two methods to reduce the number of transactions that need to be reviewed. Finally, the evaluation metrics used throughout the thesis experiments are discussed in Section 4.6.

4.1 Model framework

The decision problem introduced in Section 2.3 involving an exploration-exploitation trade-off is reflected in the multi-armed bandit problem. An analyst can choose between two options while reviewing transactions. As a first option, an analyst can choose to review transactions from another threshold level than the threshold level that has the most promising performance at that moment which is defined as exploration. On the other hand, an analyst can choose to continue reviewing transactions from the threshold that has the most promising performance so far which is defined as exploitation. A mathematical framework for this problem is described in the book Bandit Algorithms by Tor Lattimore and Csaba Szepesvári [22].

The multi-armed bandit framework is a sequential game with a learner and an environment, in an environment class ε, which is unknown to the learner. The game is repeated T rounds which is called the horizon. In each round $t \in [T]$, the learner first chooses an action A_t from a set of K actions $\mathcal{A} = \{1, 2, ..., K\}$ and then receives a reward $R_t \in \mathbb{R}$ from the environment. The learner chooses his next action based on the past which means that A_t depends only on the past $H_{t-1} = (A_1, R_1, ..., A_{t-1}, R_{t-1})$. The learner chooses a particular policy (decision strategy) to choose actions that lead to the maximum cumulative reward over all T rounds, defined as $\sum_{t=1}^{T} R_t$.

For all bandit problems, two assumptions are made [22]:

1. The learner observes the reward in every round.
2. The learners available choices and rewards in the future are not affected by their current or past decisions.

In the context of threshold tuning, the first assumption means that every time an analyst reviews a transaction, the analyst observes whether the transaction is interesting or not. The second assumption means that the number of possible threshold levels and their suitability remain the same during the threshold tuning process.

Translation to the threshold tuning process
The multi-armed bandit framework can be translated to the threshold tuning process. The environment of the threshold tuning process can be described as an action set from which an analyst can choose to review a transaction from a sample from different threshold levels $\mathcal{A} = \{1, ..., K\}$. The output of a reviewed transaction from the sample can be defined as $Y_t = \{0, 1\}$ with an unknown probability vector $\mu = [0, 1]^k$, such that the probability that $Y_t = 1$ is μ_a given the analyst chose as action $A_t = a$. Assume that the false omission rate $\frac{FN}{FN+TN}$, i.e. the proportion of interesting transactions below the tested threshold level, is chosen as measure for the threshold tuning process. The Risk Tolerance Level (RTL) is chosen as performance criterion, as defined in definition 1. For BTL testing, for each interesting (1) or not interesting (0) transaction that an analyst has reviewed of a certain threshold level, both the proportion of interesting transactions $\hat{p}$ and the upper bound for the proportion p_{upper} can be updated. Given a RTL, the difference, d, between the RTL and p_{upper} for threshold level i at time t can then be determined via:

$$d_i(t) = RTL - p_{upper,i}(t) \qquad \text{for } i = 1, ..., K \tag{4.1}$$

rounded to two decimals. Suppose that 20% is chosen as the maximum allowed percentage of missed interesting transactions below a tested threshold level, with other words $RTL = 0.2$. This has the consequence that the difference $d_i(t)$ between the RTL and the upper bound of the proportion is between -0.8 and 0.2 since the upper bound of the proportion is always between 0 and 1, thus $d_i(t) \in [-0.8, 0.2]$. A small positive difference indicates that the tested threshold level is close and below the most appropriate threshold level. If the difference is larger, this implies that the tested threshold level is considerably lower than the most appropriate threshold level.
On the other hand, a small negative difference indicates that the tested threshold level is close to and above the most appropriate threshold level. If the difference is larger, this implies that the tested threshold level is considerably higher than the most appropriate threshold level.

Definition of reward function
To define an appropriate reward function, the aim is to assign a large reward when there is a small difference between the RTL and p_{upper} for a threshold level. In addition, it is preferable to assign a higher reward to a small positive difference d, where the tested threshold level is just below the optimal threshold level, than to a small negative difference d, where the tested threshold level is just above the optimal threshold level. To achieve this, the following reward function is defined:

$$R_i(t) = \begin{cases} \frac{1}{d_i(t)} * w & \text{if } d_i(t) > 0 \\ \frac{1}{0.005} & \text{if } d_i(t) = 0 \\ |\frac{1}{d_i(t)}| & \text{if } d_i(t) < 0 \end{cases} \qquad \text{for } i = 1, ..., K \tag{4.2}$$

where w is a weight to assign more reward to threshold levels with a positive difference, where p_{upper} is below the RTL instead of above the RTL. The rewards for the rounded difference $d_i(t)$ are between the values 1.25 if $d_i(t)$ is equal to 0.8 and 100 if $d_i(t)$ is equal to 0.01 except when $d_i(t)$ is equal to zero, which would result in an infinite reward. Therefore, it is chosen to assign a reward of $\frac{1}{0.005} = 200$ if the rounded difference $d_i(t)$ is equal to zero, which has as result that $R_i(t) \in [1.25, 200]$. The average reward is then defined as:

$$\hat{\mu}_i(t) = \frac{\sum_{s=1}^{t} R_i(s)\mathbb{I}_{\{I_s=i\}}}{n_{t,i}} \qquad \text{for } i = 1, ..., K \tag{4.3}$$

where $n_{t,i} = \sum_{s=1}^{t} \mathbb{I}_{\{I_s=i\}}$ is the number of transactions an analyst has reviewed for threshold level i until the end of round t.

Definition of rounds
The rounds can be defined as each review of a transaction from a sample of a certain threshold level performed by an analyst. Suppose there are $K = 20$ threshold levels with each a sample of $n = 250$ transactions. In this case, the maximum amount of rounds is 5000 if an analyst would review all transactions.

In the threshold tuning process, an analyst has two options. The analyst can choose for no exploration phase and directly test/exploit a threshold level. Should the tested threshold level appear to not be the most suitable, a new threshold level can then be chosen for testing. In this case, no time is lost in the exploration phase in which the performance of different threshold levels is discovered, but some time is lost if one of the first tested threshold levels appears not to be the most suitable. This has the consequence that several threshold levels have to be tested before the most suitable threshold level is determined.

On the other hand, the analyst can choose to first explore the performance of different threshold levels before the analyst tests/exploits the most promising threshold level after the exploration phase. In this case, time is lost in the exploration phase in which the performance of different threshold levels is discovered, but time is probably gained in the exploitation phase since the most promising threshold level after the exploration phase is possibly close to the most suitable threshold level. For example, an analyst can decide to explore different threshold levels during a time horizon of $T = 1000$ and to exploit the most promising threshold level after that moment.

Objective of analyst
The objective of the analyst is to minimise the rounds needed to determine the most suitable threshold level with a $(1 - \alpha) \times 100\%$ confidence level.

Now that the action set, the reward function and the objective are defined, a threshold tuning strategy can be defined to determine the most appropriate threshold level with as few reviews of transactions as possible. Before delving deeper into various threshold tuning strategies, two possible threshold evaluation approaches are discussed in Section 4.2, describing which transactions above or below a threshold can be sampled since no clear guidelines are given in Comptroller's Handbook about Model Risk Management [23].

4.2 Threshold evaluation approaches

In Section 2.1.1, back-testing of threshold levels via Above-the-line (ATL) and Below-the-line (BTL) testing was introduced. However, to be able to test different threshold levels, suitable threshold values must be determined first.

Explanatory example 1
In the example of smurfing, different threshold levels can be chosen for back-testing. As a simplified example, suppose the data analysis of transaction volumes to foreign accounts shows that the transaction volumes are between 0 and 80,000 euros in a specific period. The transaction volume is defined as the total sum of transactions from one bank account in the measured period. The transaction volumes between 0 and 80,000 euros can be discretised with multiple threshold levels. Suppose that the decision is made to discretise the transaction volumes per 5,000 euro. This implies that transaction volumes between 0 and 5,000 euros fall between threshold levels 0 and 1, transaction volumes between 5,000 and 10,000 fall between threshold levels 1 and 2, etc.

In this thesis, transactions are discretised with fixed threshold levels with a certain fixed width between two threshold levels which is the same between all threshold levels.

4.2.1 Complete population approach

For BTL testing, to get a clear picture of the population proportion of interesting transactions below the threshold level that is evaluated, it is logical to take a sample of all transactions between level 0 and the threshold level that is evaluated. On the other hand, for ATL testing, it makes sense to take a sample of all transactions between the threshold level that is evaluated and all transactions above it. In this thesis, this approach is defined as the complete population approach and is illustrated in Figure 4.1.

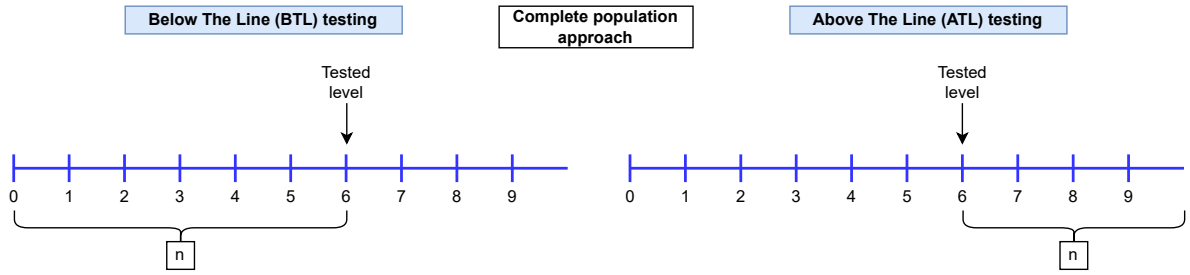


Figure 4.1: Illustration of the complete population approach for BTL and ATL testing.

Subsequently, the proportion of interesting transactions in the sample, the sample proportion, can be determined. Afterwards, the lower and upper bound for the sample proportion can be determined with the Clopper-Pearson confidence interval. The Clopper-Pearson confidence interval will contain the population proportion with a $(1-\alpha)\times 100\%$ confidence level using this approach. This has the advantage that a clear statement can be made about the proportion of interesting transactions for a certain threshold level. Suppose that, with BTL testing, the upper bound for the proportion of interesting transaction volumes is determined to be equal to p_{upper}, then the following statement can be made.

The proportion of interesting transactions in the population below threshold level L is with a $(1-\alpha)\times 100\%$ confidence level less or equal to p_{upper}.

4.2.2 Bucket approach

Another approach is to take a sample of the transactions that are below or above the threshold level but close to the threshold level that is evaluated. A reason to only take a sample of transactions that are close to the threshold level that is evaluated, is that these transactions may be more representative of the proportion of interesting transactions than transactions further below or above the threshold level that is evaluated. In this thesis, this approach is defined as the bucket approach and is illustrated in Figure 4.2.

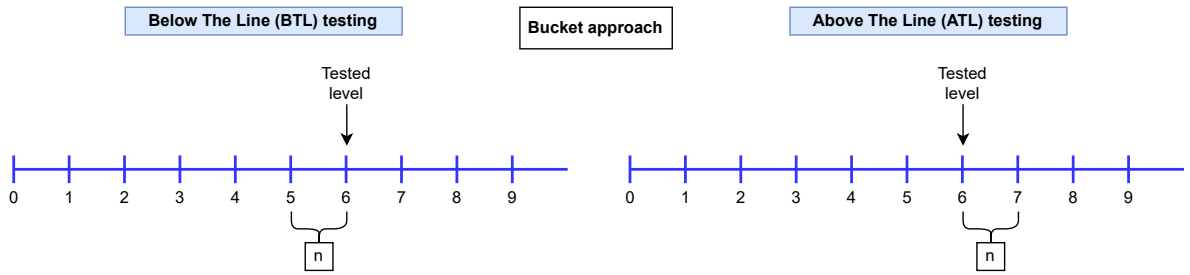
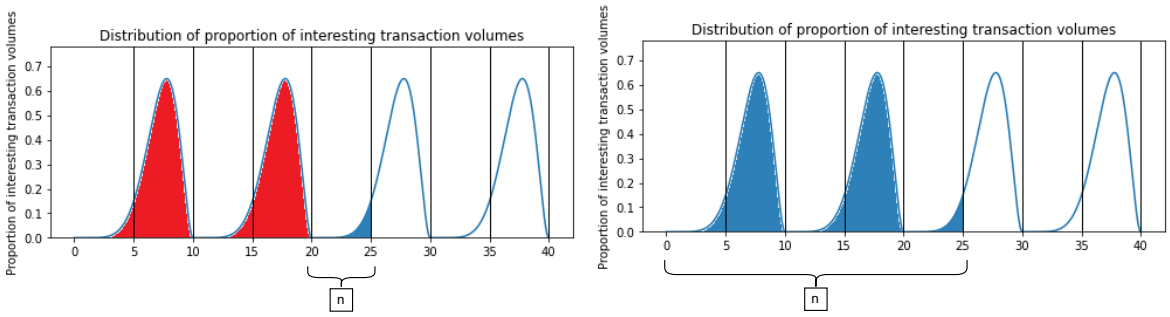


Figure 4.2: Illustration of the bucket approach for BTL and ATL testing.

However, a major disadvantage of this approach is that the number of interesting transactions further below or above the evaluated threshold level is unknown as illustrated in Figure 4.3. Continuing with

exploratory example 1, suppose that the threshold level equal to value 25,000 is evaluated with the bucket approach for BTL testing. A sample of the transactions between values 20,000 and 25,000 is taken to determine the proportion of interesting transactions. This proportion will be an estimate of the blue area in Figure 4.3a divided by the sample size without taking the red areas into account when determining the proportion. The proportion of interesting transactions between values 20,000 and 25,000 is a local minimum, which has as a consequence that a large group of interesting transactions below value 20,000 are missed with the bucket approach. As a result, the proportion of interesting transactions between values 20,000 and 25,000 is considerably lower than the actual proportion of interesting transactions between values 0 and 25,000. In this case, it can be incorrectly concluded that the threshold level which is set equal to the value 25,000 satisfies the RTL while this is not the case. The bucket approach has the risk of not knowing whether a large group of interesting transactions is missed which makes it impossible to make a statement about the proportion of interesting transactions with a $(1 - \alpha) \times 100\%$ confidence level.



(a) With the bucket approach.

(b) With the complete population approach.

Figure 4.3: Illustration of the difference between the bucket and complete population approach.

With the complete population approach, the red areas are included when the threshold level equal to value 25,000 is evaluated, which is illustrated in Figure 4.3b. Therefore, the risk of not knowing how many interesting transactions are missed does not apply to the complete population approach.

4.3 Static threshold tuning strategies

Now that all elements to define a threshold tuning strategy are introduced, various threshold tuning strategies can be explained in more detail. The threshold tuning strategies are divided into static and dynamic strategies and are defined in this thesis as follows:

Definition 3. ***Static strategy*** *A static strategy involves only an exploitation phase and no exploration phase.*

Definition 4. ***Dynamic strategy*** *A dynamic strategy involves both an exploitation and exploration phase.*

This section describes two static strategies using the complete population approach for BTL testing where a sample is taken from transactions between the tested threshold level and level 0. Subsequently, in Section 4.4, two dynamic strategies are described to get an indication of the advantage of the exploration phase. For the bucket approach, the threshold tuning strategies work in the same manner. In addition, the tuning strategies are the same for ATL testing, except that instead of taking a sample from transactions below the threshold level, a sample from transactions above the threshold level is taken.

Before a threshold tuning strategy can be applied, all transactions in the simulated data set should first be labelled. The transactions in the simulated data set are labelled as interesting with ones and are labelled as not interesting with zeros with some underlying distribution which is described in more detail in Chapter 5. Thereafter, the transaction data is partitioned between threshold levels $i = 1, ..., K$

with a fixed width between the different threshold levels. Lastly, for each threshold level i, a sample of the transactions above $S_{i,a}$ and a sample of the transactions below $S_{i,a}$ should be taken. After the simulated transaction data set is set up, a threshold tuning strategy can be applied. The first and perhaps the most obvious strategy is the 'step strategy'.

4.3.1 Step strategy

The purpose of the tuning strategy is to determine the highest threshold level for which the RTL is satisfied. At the start, the tuning strategy needs to determine a first threshold level to test. In principle, any level can be chosen to start with, but experts or data analysts may already have some idea of what a suitable threshold level might be. For BTL testing for the step strategy, the lowest logical threshold level is chosen as the first threshold level to test with the expertise of experts. In this thesis, the lowest logical threshold level is simulated by choosing a random lowest logical level between 10 and 30 per cent of the number of threshold levels in which the transaction volumes are partitioned. In the case of $K = 20$ threshold levels, this means that the lowest logical level is chosen randomly between 2 and 6.

The idea of the step strategy is to test one level lower than the tested threshold level i if $p_{upper,i} \geq RTL$ and to test one level higher than the tested threshold level i if $p_{upper,i} < RTL$ which is presented in equation 4.4.

$$L_{next} = \begin{cases} L_i + 1 & \text{if } p_{upper,i} < RTL \\ L_i - 1 & \text{if } p_{upper,i} \geq RTL \end{cases} \tag{4.4}$$

where L_i is the threshold level that is checked by the analyst and L_{next} is the threshold level that the analyst should test next. This process of testing different threshold levels is repeated until the highest threshold level i is determined for which holds that $p_{upper,i} < RTL$. The idea of the step strategy is described in an algorithmic form in recursive Algorithm 2 in Appendix E.1.

The step strategy is illustrated in Figure 4.4 in 2 situations. In situation A^*, level 3 is the first tested threshold level. A sample of the transactions between levels 0 and 3 is used to determine the lower and upper confidence bound for the proportion of interesting transactions. Suppose that the upper confidence bound for level 3 is lower than the RTL, i.e. $p_{upper,3} < RTL$, then the next step is to test threshold level 4 with the same procedure. This process is repeated up to level 8, but at level 8 the sample of transactions between levels 0 and 8 results in an upper confidence bound for the proportion of interesting transactions higher than the RTL, i.e. $p_{upper,8} > RTL$. This indicates that threshold level 7 is the highest threshold level for which the upper bound of the proportion of interesting transactions is below the RTL. Therefore, the optimal threshold level, in this case, is level 7.

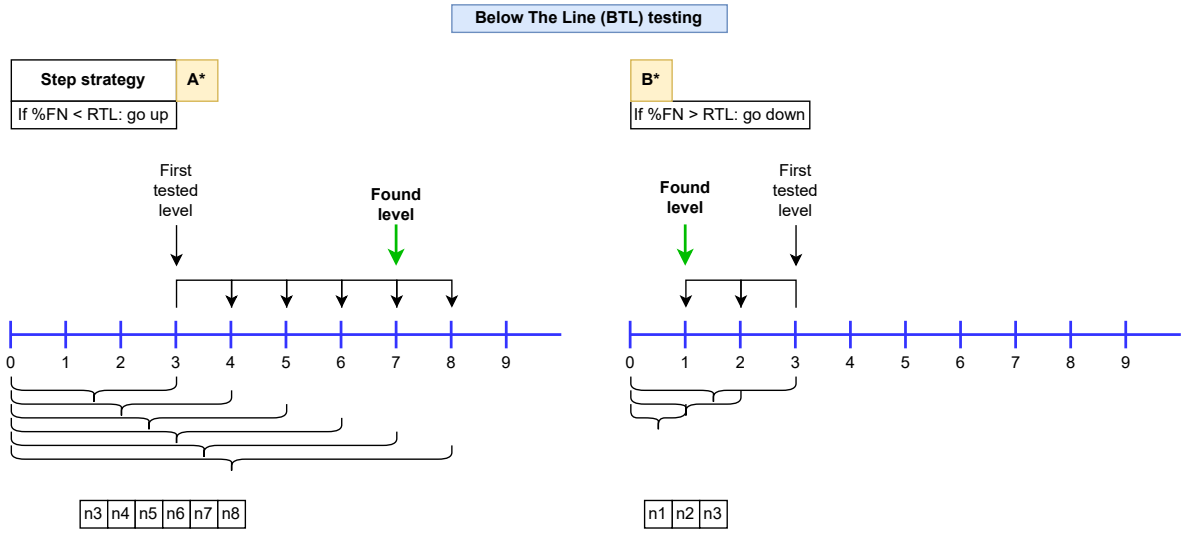


Figure 4.4: Illustration of the step strategy in combination with the complete population approach for BTL tuning.

In situation B^*, a sample of the transactions is used to determine the lower and upper confidence bound for the proportion of interesting transactions. Suppose that the upper confidence bound for level 3 is higher than the RTL, then the next step is to test threshold level 2 with the same procedure. For level 1, the sample of transactions between levels 0 and 1 results in upper confidence bound for the proportion of interesting transactions lower than the RTL. This indicates that threshold level 1 is the highest threshold level for which the upper bound of the proportion of interesting transactions is below the RTL. Therefore, the optimal threshold level, in this case, is level 1.

A possible drawback of the step strategy is that a large number of transactions have to be reviewed and quite some threshold levels have to be tested if, for example, the optimal threshold level would have been threshold level 15. However, this drawback can be diminished with the bisection strategy.

4.3.2 Bisection strategy

The bisection strategy follows mostly the same procedure as the step strategy except for the first threshold level to test and how it chooses the next threshold level to test. The verb bisect means divide into two parts which is the idea for this strategy. Suppose there are $K = 20$ possible threshold levels to test, the first step is to test threshold level 10, which divides the transactions into two parts, transactions below level 10 and transactions above level 10. If the upper confidence bound for the proportion of interesting transactions is lower than the RTL, the next step is to choose the middle level between levels 20 and 10, which is (rounded up) level 15. If the upper bound for the proportion of interesting transactions is higher than the RTL, the middle level between levels 0 and 10 is chosen, which is (rounded up) level 5. This procedure of dividing the transactions in two parts is repeated until the optimal threshold level is determined which is presented in equation 4.5.

$$L_{next} = \begin{cases} \lceil L_i + \frac{L_{max} - L_i}{2} \rceil & \text{if } p_{high,i} < RTL \\ \lceil L_i - \frac{L_i - L_{min}}{2} \rceil & \text{if } p_{high,i} \geq RTL \end{cases} \tag{4.5}$$

where L_{min} and L_{max} are the minimum and maximum checked threshold level by the analyst and equal to $L_{min} = 0, L_{max} = K$ at the beginning. The idea of the bisection strategy is described in an algorithmic form in recursive Algorithm 3 in Appendix E.2.

For the bisection strategy, unlike the step strategy, it is possible to determine in advance the maximum number of threshold levels that have to be tested to determine the most appropriate threshold level. Suppose the transactions are partitioned with K threshold levels, then the maximum number of levels to be tested, m, equals:

$$\left\lceil K * \left(\frac{1}{2}\right)^m \right\rceil = 1 \tag{4.6}$$

Suppose one chooses to partition the transaction volumes with K threshold levels, then the maximum of m threshold levels that should be tested with the bisection strategy as presented in Table 4.1.

k	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**40**
m	3	4	4	5	5	5	6	6

Table 4.1: The maximum of m threshold levels that should be tested with the bisection strategy given one chooses to partition the transactions with K threshold levels.

This is a noteworthy advantage when partitioning the transactions with more threshold levels. Suppose one chooses to partition the transactions with $K = 40$ threshold levels instead of 20 threshold levels. By partitioning the transactions with more threshold levels, it is possible to make a more accurate estimation of the proportion of interesting transactions, since the width between two threshold levels is narrower. In this case, the maximum number of threshold levels that could be tested only changes from 5 to 6, while the width between two threshold levels becomes twice as small, making the proportion estimate more accurate.

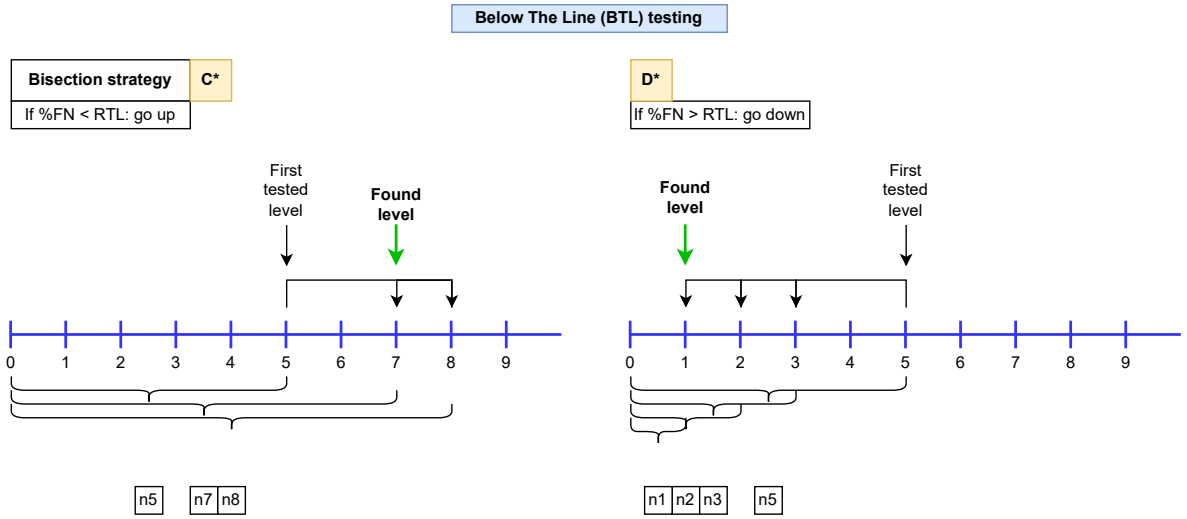


Figure 4.5: Illustration of the bisection tuning strategy in combination with the complete population approach for BTL tuning.

The bisection strategy is illustrated in Figure 4.5 in 2 situations. Comparing situation A^* of Figure 4.4 to situation C^*, it is immediately visible that the bisection method can determine the optimal threshold level faster, in the sense that fewer transactions and threshold levels need to be reviewed, if the most appropriate level is considerably some levels above level 0. In situation C^* the bisection strategy starts at level 5, then continues with level 8 and finally checks level 7. In this case, $6 * n$ transactions need to be reviewed with the step strategy to reach level 7, while only $3 * n$ transactions need to be reviewed with the bisection strategy to reach level 7. However, comparing situation B^* of Figure 4.4 to situation D^*, it becomes clear that the step method can determine the optimal threshold level faster when the most appropriate level is only a few levels above level 0. In situation D^*, the bisection strategy starts at level 5, then continues with level 3, then level 2 and finally tests level 1. Whereas only $3 * n$ transactions need to be reviewed with the step strategy to reach level 1, $4 * n$ transactions need to be reviewed with the bisection strategy to reach level 1.

Situation D^* demonstrates that starting at the middle threshold level by default can sometimes be disadvantageous resulting in the situation that more transactions and threshold levels need to be reviewed to determine the optimal threshold level than with the step strategy. This shortcoming can be enhanced by using an exploration phase to get a first indication of the proportion of interesting transactions for different threshold levels. Thereafter, the analyst can choose to test the most promising threshold level as a starting point. This idea is described by the multi-level strategy in Section 4.4.1.

4.4 Dynamic threshold tuning strategies

In the previous section, it is described that the step and bisection strategy have no exploration phase and that the analyst tests various threshold levels until the most appropriate level is determined. This saves time because there is no exploration phase but could result in a longer exploitation phase than necessary. In this Section, the multi-level and Boltzmann exploration strategies are described in more detail. The multi-level strategy explores a few threshold levels before it continues with the most promising threshold level in the exploitation phase. In the Boltzmann exploration strategy, the exploration phase is further extended by exploring all threshold levels.

4.4.1 Multi-level strategy

The multi-level strategy follows mostly the same procedure as the step and bisection strategy but a short exploration phase is added to this strategy. The idea of the multi-level strategy is to make a first guess of the proportion of interesting transactions by reviewing a part of the transactions in the sample

of a few threshold levels, defined as the exploration phase. Thereafter, this strategy continues with the most promising threshold level in the exploitation phase with the step or bisection strategy. The idea of the multi-level strategy is described in an algorithmic form in Algorithm 4 in Appendix E.3.

Suppose again that $K = 20$ and that the percentage of threshold levels that will be explored equals $x = 0.2$, implying the exploration of $K * x = 20 * 0.2 = 4$ threshold levels. In addition, $\frac{1}{K*x} = \frac{1}{4}$ part of the transactions in the samples, equivalent n^* transactions in the sample, of these threshold levels will be explored. One can choose to review a smaller or larger portion of a sample, for example, 10% or 50% of the transactions in the sample in the exploration phase, making the estimate less or more accurate. However, the main idea of the exploration phase is to get an initial idea of the suitability of certain threshold levels. Therefore, it is chosen to explore $\frac{1}{K*x}$ part of the transactions in the samples. The four threshold levels, L_i, that will be explored, are determined as follows:

$$L_i = \left\lceil K * \frac{j+1}{K*x+1} \right\rceil \tag{4.7}$$

for $j = [0, 1, ..., K * x]$. In this case, this results in threshold levels $i = 4, 8, 12, 16$ to explore. In the situation that the upper confidence bound of the proportion of interesting transactions greater is than the RTL for all four explored threshold levels, the lowest threshold level is chosen to continue within the exploitation phase.

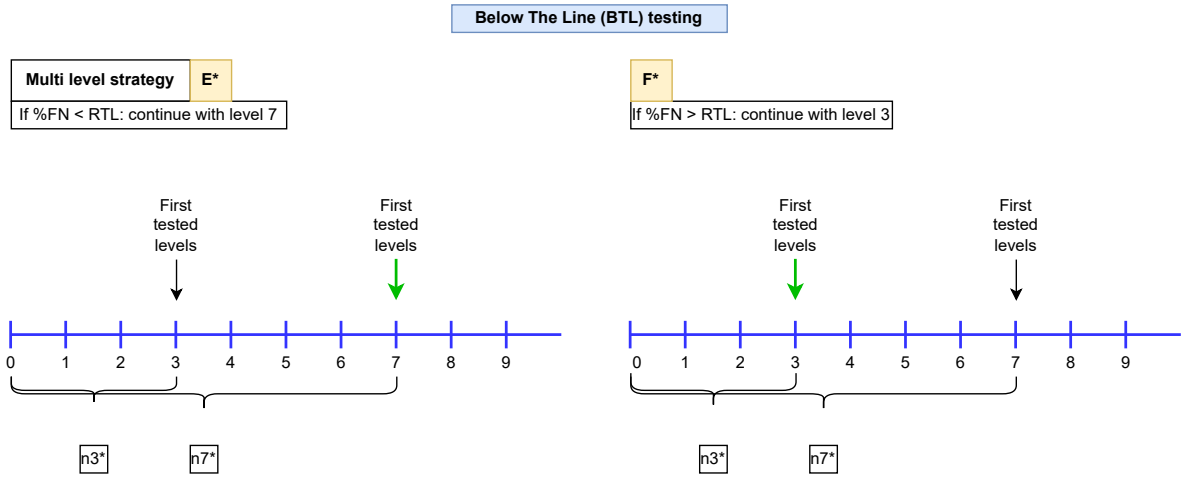


Figure 4.6: Illustration of the multi-level strategy in combination with the complete population approach for BTL tuning.

The multi-level strategy is illustrated in Figure 4.6 in two situations. Suppose that in situation E^*, the first guesses of the proportion of interesting transactions are both below the RTL and it is decided to exploit threshold level 7 as the first threshold level. Then, from level 7 onward, the step or bisection strategy can be used. On the other hand, assume that the first guesses of the proportion of interesting transactions are both above the RTL in situation F^*. In this case, it is chosen to exploit threshold level 3 as the first threshold level. Then, from level 3 onward, the step or bisection strategy can again be used.

To summarise, with the multi-level strategy, unlike the static threshold tuning strategies, a few threshold levels are explored by an analyst who reviews a subset of sampled transactions before continuing with the most promising threshold level to exploit. However, this idea, to explore not just a few but all threshold levels, can be extended even further to minimise the amount of sampled transactions that need to be reviewed by an analyst.

Over time, an extensive number of algorithms have been formulated to address this problem. Methods that imply a binary choice between exploitation (the greedy choice) and exploration (uniform probability over a set of actions) are known as *semi-uniform methods* [40]. The simplest variant is the ϵ-Greedy

strategy. In each round, the ϵ-Greedy algorithm chooses with probability ϵ a random threshold level from which one transaction will be reviewed (exploration) and with probability $1-\epsilon$ the threshold level that has the highest average reward (exploitation) from which one transaction will be reviewed. The main disadvantage of this strategy is that non-optimal threshold levels are still chosen, even after it is identified that these threshold levels are not optimal. Therefore, this strategy is not the most efficient to minimise the required amount of sampled transactions that need to be reviewed to determine the most appropriate threshold level. More details on the ϵ-Greedy algorithm are described in Appendix E.

Another category of strategies is *probability matching methods* that choose actions according to a probability distribution that reflects how likely the actions, in this case the different threshold levels, are to be optimal [40]. A well-known probability matching strategy is Boltzmann exploration, also known as the Softmax strategy.

4.4.2 Boltzmann exploration strategy

Remember the definitions of difference d, reward R and average reward $\hat{\mu}$ which were stated in Section 4.1 and relevant for the Boltzmann exploration strategy:

$$d_i(t) = RTL - p_{upper,i}(t) \qquad \text{for } i = 1, ..., K$$

rounded on two decimals.

$$R_i(t) = \begin{cases} \frac{1}{d_i(t)} * w & \text{if } d_i(t) > 0 \\ \frac{1}{0.005} & \text{if } d_i(t) = 0 \\ |\frac{1}{d_i(t)}| & \text{if } d_i(t) < 0 \end{cases} \qquad \text{for } i = 1, ..., K$$

$$\hat{\mu}_i(t) = \frac{\sum_{s=1}^{t} R_i(s)\mathbb{I}_{\{I_s=i\}}}{n_{t,i}} \qquad \text{for } i = 1, ..., K$$

Boltzmann exploration is a classic strategy for sequential decision-making under uncertainty which, instead of uniformly exploring all threshold levels which is the case with *semi-uniform methods*, chooses each threshold level with a probability that is proportional to its average reward. This strategy selects a threshold level using a Boltzmann distribution, given initial empirical means $\hat{\mu}_0(0)$,...,$\hat{\mu}_K(0)$ [21]:

$$p_{select,i}(t+1) = \frac{e^{\hat{\mu}_i(t)/\tau(t)}}{\sum_{j=1}^{K} e^{\hat{\mu}_j(t)/\tau(t)}}, i = 1, ..., K \tag{4.8}$$

where τ is called the temperature factor, controlling the randomness of the choice to explore a threshold level. In this thesis, it is decided to review $\frac{1}{50}$ of the transactions from the sample from each threshold level to determine the initial empirical means $\hat{\mu}_1(0)$,...,$\hat{\mu}_K(0)$, since it provides an estimate of the average reward for each threshold level but does not require a lot of time from an analyst. It is also possible to choose the initial rewards $\hat{\mu}_1(0) = v$,...,$\hat{\mu}_K(0) = v$ equal to some value v, or to review more transactions from the sample from each threshold level to determine the initial empirical means $\hat{\mu}_1(0)$,...,$\hat{\mu}_K(0)$ with more certainty, but some choice had to be made.

When τ is small, the overall exponential element of each threshold level is exponentially proportional to their current return. As a result, threshold levels with a higher average return will have a higher chance of being chosen in the exploration phase. When $\tau \to \infty$, the overall exponential element of all threshold levels approaches a resultant value of 1, and the algorithm chooses threshold levels uniformly at random.
This strategy could be modified in the same manner as the ϵ-Greedy strategy into decreasing Boltzmann exploration where τ decreases with a particular function with the number of rounds played, for instance $\tau(t) = \tau(0)/t$. Unfortunately, it is noted in literature that determining the correct schedule for $\tau(t)$ is difficult in practice [5]. More precisely, a schedule may choose sub-optimal threshold levels too often even after having estimated all the average rewards correctly, or commit too early to a sub-optimal threshold level and never return to a better threshold level. In Chapter 5, a few schedules for $\tau(t)$ will

be compared to determine which schedule results in the best performance for the Boltzmann exploration strategy for different case studies.

In this thesis, no other bandit algorithms are researched to keep the focus on researching whether dynamic threshold tuning strategies which have an exploration phase are more efficient than static threshold tuning strategies which have no exploration phase. A strategy is more efficient if fewer sampled transactions need to be reviewed by an analyst to determine the most appropriate threshold level. The Boltzmann exploration strategy is described in an algorithmic form in Algorithm 5 in Appendix E.4.

The objective for all threshold tuning strategies is to minimise the required reviews to determine the most suitable threshold level with a $(1-\alpha)\times 100\%$ confidence level. However, there are two methods to reduce the required reviews for all threshold tuning strategies in a simple manner which are described in Section 4.5.

4.5 Methods to reduce the required reviews

This section describes two simple methods to reduce the number of required reviews. The first method is defined as the 'early break-off principle' for BTL testing and is described in more detail in Section 4.5.1. Subsequently, Section 4.5.2 describes how reusing reviewed transactions can further reduce the required reviews. It is interesting to explore methods that can reduce the number of reviews, as this can greatly reduce the workload for an analyst.

4.5.1 Early break-off principle

For BTL testing, it is not always necessary to review all transactions of the sample to determine whether $p_{upper} < RTL$. Suppose that a threshold level is tested with a sample of $n = 250$ transactions and that $RTL = 0.2$. After reviewing 150 transactions, 51 transactions are reviewed as interesting transactions. At this moment the upper bound for proportion interesting transactions is at least $\frac{k}{n} = \frac{51}{250} = 0.204$ which is higher than the RTL although not all transactions in the sample are reviewed. In this case, it is unnecessary to review the remaining 100 transactions, as the proportion $\frac{k}{n}$ can only increase for the tested threshold level. Therefore, the review process for the threshold level can be terminated prematurely which is defined as the 'early break-off principle'. The threshold tuning strategies can continue in the normal manner for the case that $p_{upper} > RTL$. In this case, the time required to review the other 100 transactions in the sample is saved with the 'early break-off principle', resulting in a more efficient procedure to determine the most suitable threshold with a threshold tuning strategy. This principle can also be applied to ATL testing in the situation that the proportion of interesting transactions is below some specified minimum productivity tolerance level.

4.5.2 Reusing information of a previously reviewed sample

Another method to reduce the required reviews is reusing information collected during the threshold tuning process.

Situation that previous tested level is lower than the current tested level
Suppose that the previous tested threshold level is level 5 and the current tested threshold level is level 6 as illustrated in Figure 4.7. In this case, for BTL testing, the transactions between level 0 and 5 in the sample of threshold level 6, S_6, provide no new information compared to the information collected from the transactions in the sample of level 5, S_5, which is illustrated in the left image in Figure 4.7 with the orange line. Only the transactions above level 5 provide new information about the proportion of interesting transactions for threshold level 6 which is illustrated with the green line. Suppose the sample size is $n = 250$ and there are 20 transactions in the sample of threshold level 6, S_6, above the value of threshold level 5, V_{L5}. In this case, the analyst can review the 20 transactions from sample S_6 above V_{L5} and reuse the information of $250 - 20 = 230$ reviewed transactions from sample S_5. A random sample of 230 transactions is taken from sample S_5 and combined with the 20 transactions from sample S_6 to determine the proportion of interesting transactions for threshold level 6. This will

save the analyst the time required to review 230 new transactions.

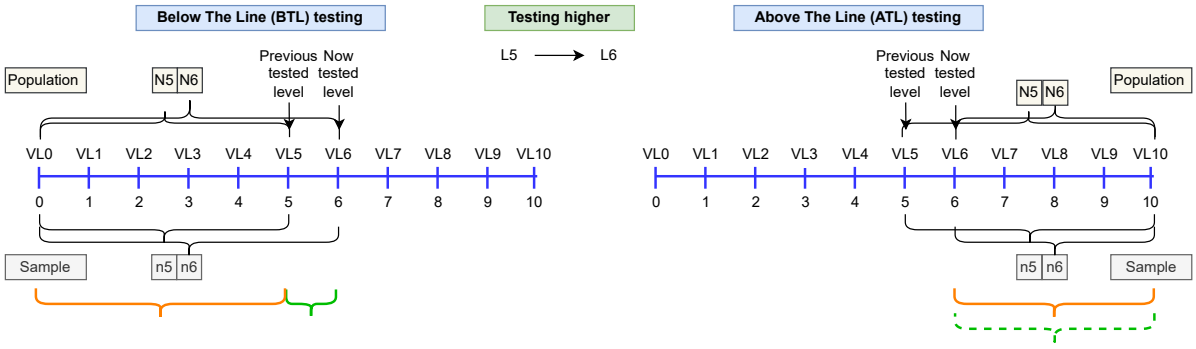


Figure 4.7: Illustration of reusing part of a reviewed sample when the previous tested threshold level was higher.

On the other hand, for ATL testing, the analyst can reuse the information of reviewed transactions from sample S_5 above the value of threshold level 6, V_{L6}. Suppose again that $n = 250$ and there are 210 transactions in the sample of threshold level 5, S_5, above the value of threshold level 6, V_{L6}. In this case, the information of 210 reviewed transactions from sample S_5 can be reused. In addition, the analyst needs to review $250 - 210 = 40$ transactions from sample S_6. A random sample of 40 transactions is taken from S_6 and combined with the 210 transactions from sample S_5 to determine the proportion of interesting transactions for threshold level 6 as illustrated in the right image in Figure 4.7.

Situation that previous tested level is higher than current tested level

Figure 4.8 illustrates the situation where the previous threshold level tested is higher than the current tested threshold level. In this case, reusing the collected information is reversed for BTL and ATL testing. For BTL testing, the analyst can reuse the information of reviewed transactions from sample S_6 below the value of threshold level 5, V_{L5}. The difference between the number of transactions that can be reused and the sample size n can be supplemented with transactions from sample S_5, again via a random sample, to determine the proportion of interesting transactions for threshold level 5 as illustrated in the left image in Figure 4.8.

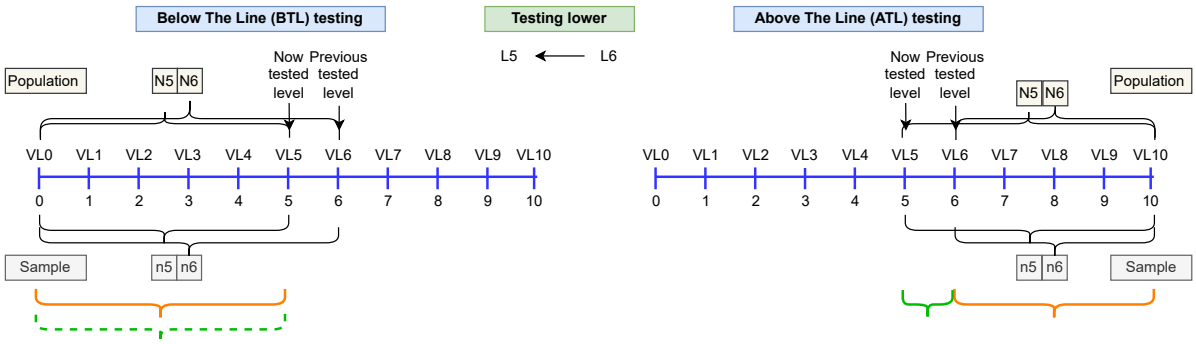


Figure 4.8: Illustration of reusing part of a reviewed sample when the previous tested threshold level was lower.

In contrast, for ATL testing, the analyst can reuse the information of reviewed transactions from sample S_6. Only the transactions below level 6 provide new information about the proportion of interesting transactions for threshold level 5. In this case, the analyst can review the transactions below V_{L6} from sample S_5, for instance 30 transactions, and reuse the information of $250 - 30 = 220$ reviewed transactions from sample S_6, via a random sample, to determine the proportion of interesting transactions for threshold level 5 as illustrated in the right image in Figure 4.8.

In this chapter, various threshold tuning strategies have been described in detail from which naturally the question arises how the performance of the threshold tuning strategies can be compared and quantitatively evaluated, which is described in Section 4.6.

4.6 Evaluation metrics

The performance of the various threshold tuning strategies will be evaluated with three case studies. Two aspects are important in the evaluation of performance. The first aspect is the required amount of transactions that need to be reviewed by an analyst to determine the most appropriate threshold level, in other words, the duration of the threshold tuning process.
In addition, it is important to know whether the most appropriate threshold level determined with a threshold tuning strategy with sampling corresponds to the most suitable true threshold level based on the true transaction population. The difference between the found threshold level $T_{strategy}$, based on a sample of the transactions, and true threshold level T_{true}, based on the population of transactions without sampling can be defined as:

$$D_{threshold} = T_{true} - T_{strategy}. \tag{4.9}$$

In Section 5.2, it will be explained in more detail how the transactions are randomly generated following some distribution and how the transactions are randomly labelled as interesting (1) or not interesting (0) following some distribution. This has the consequence that the number of reviews and the difference can vary a bit each time a threshold tuning strategy is executed due to the random generation of the distributions and random sampling.

In this thesis, the choice is made to repeat the threshold tuning strategies 25 times for a particular case study in which a transaction population is randomly generated following some distribution in each run. In each run, the amount of required reviews and the difference between the found and true threshold level is determined. The choice was made to repeat the threshold tuning strategies 25 times since the maximum, median and minimum of the number of required reviews and the difference between the found and true threshold level stabilise after 25 runs for case study 1 as illustrated in Figures G.1 and G.2 in Appendix G.1. The same conclusion could be drawn for case studies 2 and 3.

4.6.1 Performance evaluation of threshold tuning strategies

To be able to quantify the performance of a threshold tuning strategy and compare it with the other threshold tuning strategies, the following preferences have been defined:

- **The number of reviews**: There is a preference for a threshold tuning strategy that requires as few reviews of transactions as possible to determine the most appropriate threshold level.
- **The difference**: There is a preference for a threshold tuning strategy with a small difference since this implies that the found threshold based on the sample is close to the true threshold based on the population. In addition, there is a preference for a positive difference rather than a negative difference, since a positive difference indicates that the found threshold level is lower than the true threshold level which is a safe choice for BTL testing. On the other hand, if the difference is negative, this indicates that the found threshold level is higher than the true threshold level which implies that the RTL will not be satisfied for the transaction population. Therefore, it is preferred to prevent the situation that the difference is negative.
- **The spread in performance of a tuning strategy**: There is a preference for a threshold tuning strategy for which the spread in the required amount of reviews and difference to determine the most appropriate threshold level is as small as possible over 25 runs. A threshold tuning strategy has a more stable performance if the strategy often requires reviewing a similar number of transactions and the difference between the found and true threshold level is often the same, in comparison to a strategy where the required amount of reviews and the difference fluctuate more. The spread over the 25 runs will be illustrated with box plots showing the maximum, 75% quantile, median, 25% quantile, and minimum values.

5

Experiments

In this chapter, the set up of synthetic data sets based on tree case studies and different experiments conducted in this thesis are described in more detail.
This chapter is divided into several sections. In Section 5.1 three common money laundering scenarios are presented to give more insight into criminal behaviour and how criminals try to circumvent the law. Subsequently, Section 5.2 explains how the money laundering scenarios described in Section 5.1 are employed to construct a labelled synthetic transaction data set to evaluate the various threshold tuning strategies. Lastly, Section 5.3 describes the purpose of the different experiments that will be executed in this thesis.

5.1 Case studies

In this thesis, the threshold tuning strategies presented in Sections 4.3 and 4.4 are applied to three different case studies to evaluate their performance. There are almost no publicly available labelled transaction data sets in which transactions are labelled as unusual or not. However, various money laundering scenarios that financial institutions face are publicly available and can be employed to construct a labelled transaction data set. Some examples of money laundering scenarios that are publicly available are [32]:

- **Money laundering with an exchange market in combination with smurfing**. This method involves large amounts of illegally obtained cash. This case will be discussed in more detail and is used as a case study in this thesis.
- **Money laundering with derivatives**. A derivative contract is a bet placed on the movement of some underlying market factor. This case will be discussed in more detail and is used as a case study in this thesis.
- **Money laundering through charities and non-profit organisations (NPOs)**. The operational model for many NPOs involves cash-intensive fundraising, with numerous combined small cash donations before being deposited into a bank account, and financial transfers to high-risk countries where funds can be deployed for humanitarian work which provides an ideal cover for money laundering. This case will be discussed in more detail and is used as a case study in this thesis.
- **Money laundering with trade finance**. This method is very complicated to detect since the volume of trade flows, the complexities of foreign exchange transactions and the involved long-supply chains indicate that the flows of illicit funds can be hidden from view.

More examples of money laundering scenarios and typologies are publicly available and described by the FIU-NL [36, 38].

5.1.1 Case study 1: detecting smurfing of cash deposits

Consider money laundering with an exchange market in combination with smurfing. This case study is based on the following money laundering scenario 'black market peso exchange' [33]. In this scenario, dollars owned by a drug cartel in the US are exchanged for pesos in Columbia with the help of a broker. In addition, goods are exported from the US to Colombia, since goods can move across borders, but moving money is more complicated. This money laundering scenario between the US and Colombia can be described in 7 steps and is illustrated in Figure 5.1 based on [33].

1. As the first step, a Colombian cartel exports drugs to the US which are sold there. This generates large amounts of cash dollars for which the Colombian cartel arranges a broker to buy the dollars generated by the sale of these drugs.
2. The broker then arranges to sell the dollars to a local importer in Colombia who wants to pay an American supplier for goods, to be exported from the US to Colombia.
3. The broker then arranges for his representative in the US to collect the dollars for the cartel.
4. Subsequently, the broker uses these dollars to pay the American supplier for the goods for export. This can be done, for example, by placing the cash into multiple bank accounts controlled by the broker using structured deposits and then forwarding the funds to the supplier via wire transfers.
5. The American supplier exports the goods to Colombia where they are received by the Colombian importer.
6. The Colombian importer, having received the imported goods, pays the broker in pesos.
7. Finally, the broker now has pesos to pay the Colombian cartel, which he does, after having taken his commission.

In the fourth step, one or more banks are involved and have a responsibility to report unusual transactions. Under the Bank Secrecy Act, banks and other financial institutions have the obligation to report cash transactions exceeding 10,000 dollars [14]. In this case, criminals can try to circumvent the law by structuring large amounts of cash into several small cash deposits which is behaviour that banks will try to detect. In Section 5.2, it will be further explained how a labelled data set can be constructed based on this money laundering scenario.

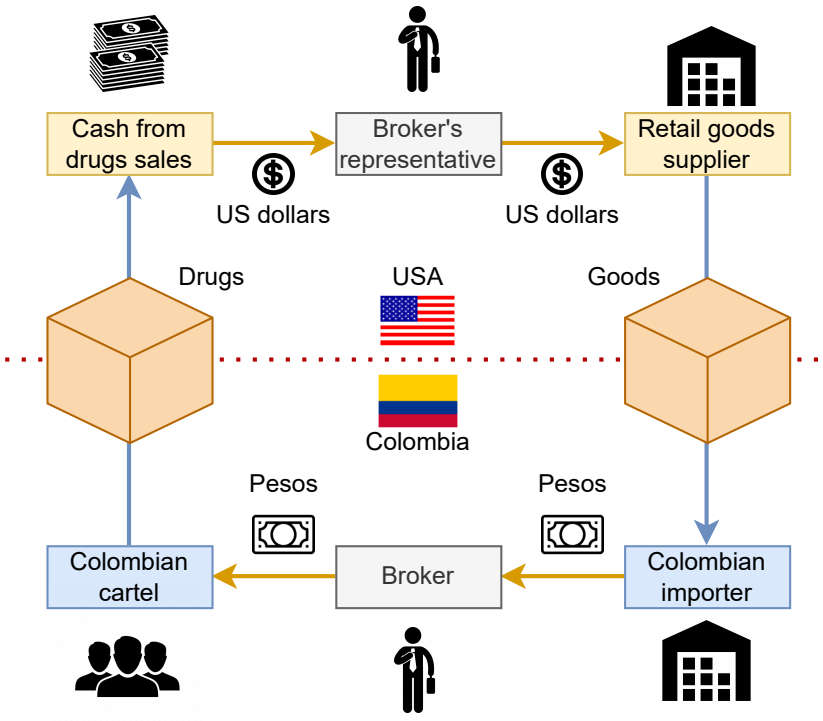


Figure 5.1: Case study 1: detecting smurfing of cash deposits illustrated with an exchange deal [33].

5.1.2 Case study 2: detecting unusual profits from derivatives

Consider money laundering through derivatives. A derivative is a type of financial contract, set between two or more parties, that derives its value from an underlying asset or group of assets. This case study is based on the following money laundering scenario [37]. In this scenario, a Russian criminal organisation wants to launder criminally obtained funds through trading. This money laundering scenario is illustrated in Figure 5.2 based on [37]. Important terminology is to understand the difference between a short and long position on a contract. An investor is said to go 'long' on a contract if the investor is betting that a price will increase in value. Contrarily, if an investor goes 'short' on a contract, the investor bets that the underlying value will decrease [37].
This method for laundering illegal funds is difficult to detect since the complexity of derivatives and the derivatives market provide perfect cover for hiding money laundering activities [37]. This is because it is fairly normal that market participants only know the identity of the broker and not of others. This money laundering scenario can be described in 5 steps.

1. Illicit funds are deposited into a brokerage account, Account A, controlled by a complicit broker.
2. The broker will go both long and short on a particular commodity during a trading day, closing out both positions at the end of the day. Thereafter, the broker will assign the losses to Account A, reducing the balance of the account with the 'dirty' funds, and assign the profits to Account B, thus producing more 'clean' money.
3. As the next step, the broker will go long and purchase derivative contracts for a particular commodity and at the same time, the broker will also go short and sell the same number of derivative contracts for the particular commodity. Later in the trading day, the broker goes back to the market and closes out the two open positions.
4. Now assume the price of the underlying commodity has decreased, meaning that the long position has lost money ad the short position resulted in a profit. The broker then assigns the long (losing position) to Account A and the short (winning position) to Account B.
5. Thereafter, the profit of the winning position will be less than the amount of the loss of the losing position. This difference is the price of laundering the money.

Suppose the winning position has made a profit of 50,000 euros and the losing position has made a loss of 60,000 euros, then it will cost 60,000 euros in 'dirty' money to generate 50,000 euros in 'clean' money. If a particular bank account often results in a profit or loss that is considerably higher or lower than standard gains or losses, this may be an indication of unusual transaction behaviour and require further investigation. In Section 5.2, it will be further explained how a labelled data set can be constructed to detect unusual profits or losses of derivatives.

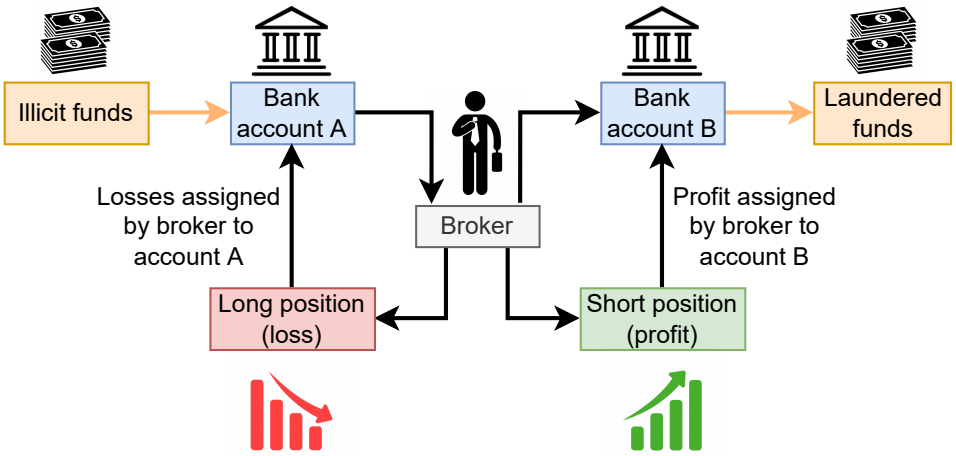


Figure 5.2: Case study 2: detecting unusual profits from derivatives [37].

5.1.3 Case study 3: detecting unusual financial transaction to high-risk countries

Consider money laundering through charities and NPOs with financial transfers to high-risk countries. A list of countries that are designated high-risk countries by the FATF and the European Commission is published three times a year [34]. This case study is based on the following money laundering scenario [35]. In this scenario, a charity or NPO is exploited for money laundering or terrorism funding. This exploitation can occur in the collection phase, transfer phase or point of delivery phase in an NPO operational model. This money laundering scenario is illustrated in Figure 5.3 based on [35]. Money can be laundered in the transfer phase in the following manner.

In the transfer phase, unknown to the majority of the board of an NPO, there is a minority that actively launders money or supports a terrorist group. This minority includes an individual who has access to the NPO's bank account. The funds that are collected in the US, are consolidated in an account in Albania, which is a high-risk country [34], where the NPO has an office, before being sent to the country for which the donations are intended, for example, Somalia. When these various transfers are made, funds are diverted using international wire transfers to the accounts of sham businesses, supposedly for logistical services, and are ultimately used to support the terrorist group.

When money from a charity or NPO is transferred via a complicated or unnecessary financial construction through a high-risk country to the country for which the money is intended, this can be an indication of unusual transaction behaviour. In Section 5.2, it will be further explained how a labelled data set can be constructed to detect unusual financial transactions to high-risk countries.

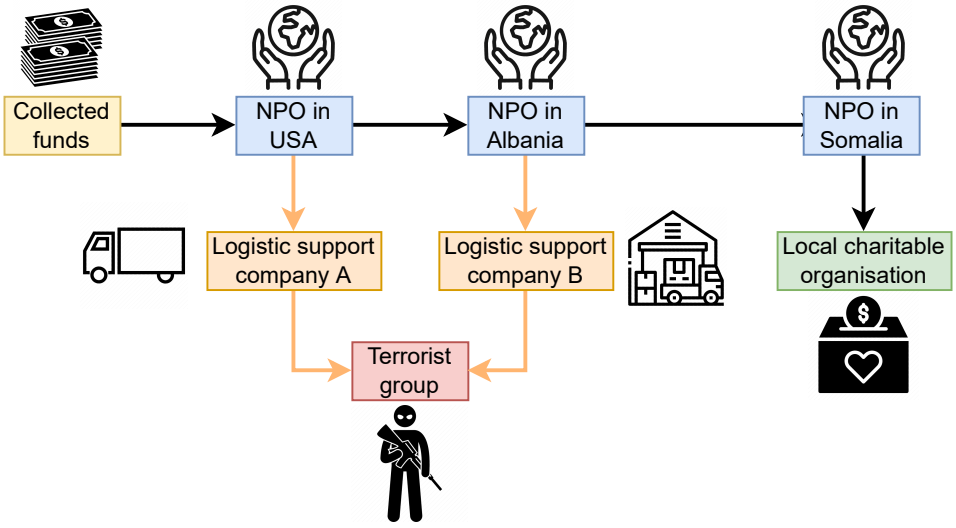


Figure 5.3: Case study 3: detecting unusual financial transactions to high-risk countries [35].

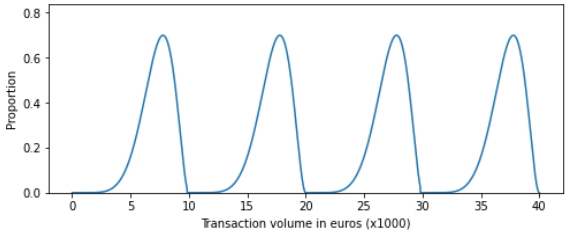
5.2 Translation from scenarios to distributions

This section describes how the synthetic data set can be constructed based on the money laundering scenarios described in the previous section. It is important to note that the assumptions made to construct distributions are used as a starting point to illustrate examples of calculations in which the case studies serve to illustrate possible criminal behaviour. The distributions are constructed to test the performance of the tuning strategies in situations where the proportion of interesting transaction volumes is fluctuating/non-monotonic, monotone non-decreasing or constant since there are almost no publicly available labelled transaction data sets. It is important to emphasise that the plausibility of the distributions and assumptions has not been tested. Research into other distributions may be interesting for future research which will be discussed in more detail in Chapter 7.

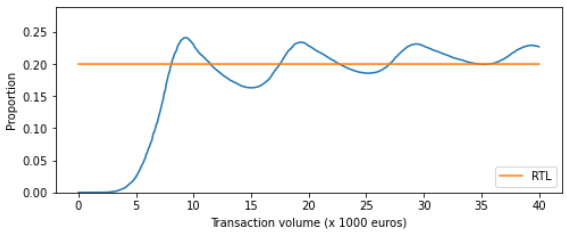
Case study 1: detecting smurfing of cash deposits

To detect smurfing of cash deposits with a rule, a transaction monitoring model should detect whether large sums of cash are structurally transferred that are just below the publicly known threshold. This rule should characterise criminal behaviour that may indicate smurfing.
Assume that criminals will structure the cash deposits into smaller amounts just below the publicly known threshold of 10,000 euros. Based on this reasoning, the following distributions are constructed for the transactions.

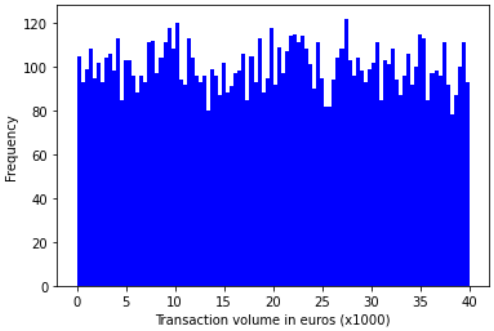
Suppose that in a particular period in which data analysis for threshold tuning is performed, a bank is dealing with a population of 10,000 transaction volumes from different bank accounts. The transaction volume is defined as the sum of the transaction amounts of one bank account in a given period. For simplicity, assume that the transaction volumes lie between 0 and 40,000 euros, whereby the transaction volumes of the 10,000 different bank accounts are uniformly distributed between 0 and 40,000 euros as presented in Figure 5.4c. In addition, assume that there is a higher probability that a transaction volume is interesting just below the public threshold of 10,000 euros and multiples thereof. For this purpose, assume that the interesting transactions are $\beta(8, 3)$ distributed between 0 and 10,000 euros and multiples thereof as presented in Figure 5.4a. The data set is constructed in which the 10,000 transaction volumes from different bank accounts are labelled as interesting (1) with probability p as given on the y-axis and labelled as normal transaction volume (0) with probability $1 - p$. The distribution of transaction volumes and the distribution of interesting transaction volumes are shown in Figure 5.4.



(a) The distribution of unusual transaction volumes of different bank accounts.



(b) The proportion of unusual transaction volumes of different bank accounts as defined by equation 5.1.



(c) The distribution of transaction volumes of different bank accounts.

Figure 5.4: The distributions of case study 1: detecting smurfing of cash deposits.

For BTL testing, the blue line in Figure 5.4b represents the proportion of interesting transaction volumes between threshold level zero and a particular threshold level that is defined as:

$$p_b = \frac{y_b}{N_b} \tag{5.1}$$

where y_b is the amount of interesting transaction volumes from the threshold level equal to value zero up to a threshold level equal to a particular value and N_b the population of transaction volumes from the threshold level equal to value zero up to a threshold level equal to a particular value. Suppose there are 7600 transaction volumes with a value between 0 and 30,000 euros of which 1590 are labelled as interesting transactions (1), then the proportion of interesting transactions below the threshold level of 30,000 euro equals $p_b = y_b/N_b = 1590/7600 = 0.209$

The key reason why this case is interesting is that there is a chance that a threshold tuning strategy determines that the most appropriate threshold level is in one of the local minima. The proportion of interesting transaction volumes, as presented in Figure 5.4b is a non-monotonic function. Therefore, the aim of this case study is to evaluate the performance of different threshold tuning strategies if the proportion of interesting transactions could be described with a non-monotonic function.

Case study 2: detecting unusual profits from derivatives

To detect unusual profits from derivatives, a transaction monitoring model should detect whether profits from a derivative trade are noteworthy higher than the change in market value in a given period. This rule should characterise criminal behaviour that may indicate money laundering using derivatives. The following distributions are constructed for the transaction volumes.

Suppose that in a particular period a bank is dealing with a population of 10,000 trading transaction volumes from different bank accounts which increase or decrease due to a change in the market value. The change in transaction volume of a bank account relative to the change in the market value can be expressed as a ratio:

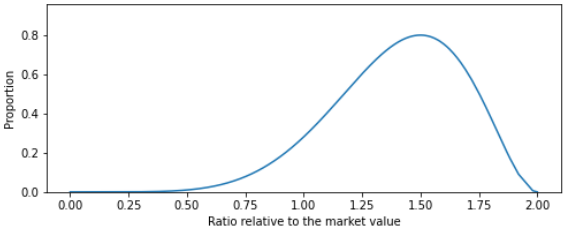
$$ratio = \frac{\Delta_{v_{trade}}}{\Delta_{v_{market}}} = \frac{\frac{v_{t_e}}{v_{t_0}}}{\frac{v_{m_e}}{v_{m_0}}} \tag{5.2}$$

where $\Delta_{v_{trade}}$ is the increase or decrease in transaction volumes of a bank account between the beginning v_{t_0} and end v_{t_e} of a particular period and $\Delta_{v_{market}}$ the increase or decrease in market value between the beginning v_{m_0} and end v_{m_e} of a particular period.

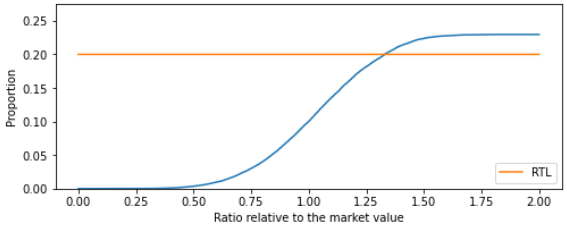
For example, the market value of natural gas per million Btu was 5.712 dollars on 1 July 2022 and 9.33 dollars on 1 September 2022 [31]. During this period, the market value of natural gas changed by $(9.33/5.712) \times 100\% = 163\%$. Now suppose that a transaction volume of a bank account in that period changed by 250%, e.g. $v_{t_e} = 50,000$ and $v_{t_0} = 20,000$, while the change in market value was 163%. In this case, the difference in change between the market value and transaction volume of a bank account is equal to $250/163 = 1.53$. Criminals are likely to minimise large differences between a change in transaction volume on a bank account and the market value because they are aware that it is noticeable.
Therefore, assume that the ratio lies between 0 and 2 whereby the ratio is $N(1, \frac{1}{3})$ distributed as presented in Figure 5.5c. In addition, assume that there is a higher probability of interesting transaction volumes if the ratio is higher than 1 and is relatively large since the clean money is generated in the bank account with unusually high profits. For this purpose, assume that the interesting ratios are $\beta(7, 3)$ distributed between 0 and 2 as presented in Figure 5.5a. The data set is constructed in which the ratio is calculated of 10,000 transaction volumes from different bank accounts and are labelled as interesting (1) with probability p as given on the y-axis and labelled as normal (0) with probability $1 - p$. The distribution of the ratio of 10,000 different bank accounts and the distribution of the interesting ratios are presented in Figure 5.5.

The main reason why this case is interesting is the comparison of multiplicative changes in transaction volumes. The ratio, as presented in Figure 5.5b is a monotonic non-decreasing function. Therefore, the aim of this case study is to evaluate the performance of different threshold tuning strategies if

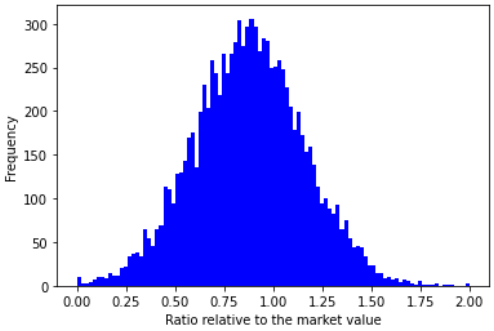
the proportion of interesting transaction volumes could be described with a monotone non-decreasing function.



(a) The distribution of unusual ratios relative to the market value of different bank accounts.



(b) The proportion of unusual ratios relative to the market value of different bank accounts as defined by equation 5.1.



(c) The distribution of ratios relative to the market value of different bank accounts.

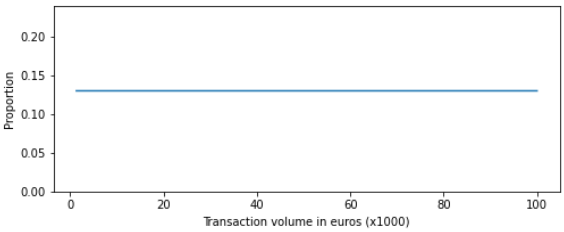
Figure 5.5: The distributions of case study 2: detecting unusual profits from derivatives.

Case study 3: detecting unusual financial transactions to high-risk countries

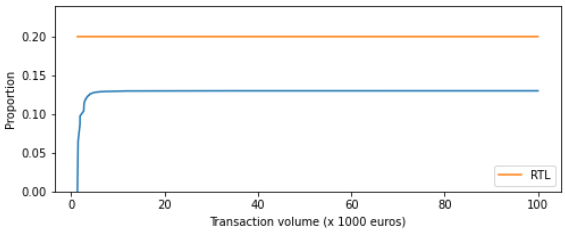
To detect unusual financial transactions to or from bank accounts in high-risk countries, a transaction monitoring model should detect whether financial transactions are transferred via a complicated construction through a high-risk country before the money is transferred to the final bank account in a given period. This rule should characterise criminal behaviour that may indicate money laundering using unusual financial transactions to high-risk countries. The following distributions are constructed for the transaction volumes.

Suppose that in a particular period a bank is dealing with a population of 10,000 transaction volumes from different bank accounts transferred to or from bank accounts in high-risk countries. For simplicity, assume that the transaction volumes lie between 0 and 100,000 euros, whereby the transaction volumes of the 10,000 different bank accounts are $\beta(3, 8)$ distributed between 0 and 100,000 euros as presented in Figure 5.6c and that there will be more low than high transaction volumes. In addition, assume that there is a constant probability of interesting transaction volumes for all financial transactions to or from bank accounts in high-risk countries. This is a strong assumption to make and probably not often the case but money laundering can occur in both larger and smaller companies and with both larger and smaller transaction volumes. For this purpose, assume that the interesting transaction volumes are uniformly distributed between 0 and 100,000 euros as presented in Figure 5.6a. The data set is constructed in which the 10,000 transaction volumes from different bank accounts are labelled as interesting (1) with probability p as given on the y-axis and labelled as normal transaction volume (0) with probability $1 - p$. The distribution of transaction volumes and the distribution of interesting transaction volumes are shown in Figure 5.6.

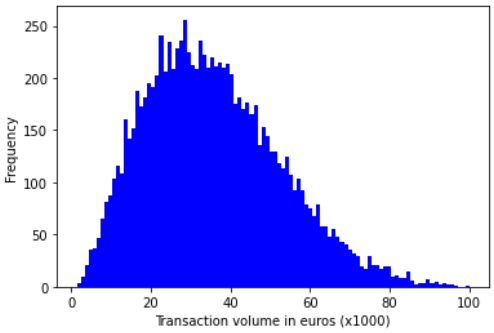
The key reason why this case is interesting is to investigate how tuning strategies handle situations where there is almost no difference between the suitability of each threshold level. The proportion of interesting transaction volumes, as presented in Figure 5.6b, is a constant function. Therefore, the aim of this case study is to evaluate the performance of different threshold tuning strategies if the proportion of interesting transactions has a constant function.



(a) The distribution of unusual transaction volumes of different bank accounts.



(b) The proportion of unusual transaction volumes of different bank accounts as defined by equation 5.1.



(c) The distribution of transaction volumes of different bank accounts.

Figure 5.6: The distributions of case study 3: detecting unusual financial transactions to high-risk countries.

The distributions are generated with *numpy.random.distribution*() and *stats.distribution.pdf*(). In addition, the parameter settings for the different case studies are presented in Table 5.1. In this thesis, the choice was made to partition the transaction volumes with $K = 20$ threshold levels, because the difference in the performance of the various threshold tuning strategies is more visible with a larger number of threshold levels and narrower widths between the threshold levels. The more threshold levels are used to partition the transaction volumes, the more the difference in the performance of the various threshold tuning strategies becomes visible which is illustrated in Section 6.3.3.

Parameter	**Case 1**	**Case 2**	**Case 3**
RTL	0.2	0.2	0.2
α	0.05	0.05	0.05
e (MoE)	0.05	0.05	0.05
p_0	0.2	0.2	0.2
n (resulting from the choices for p_0, e, α)	264	264	264
N	10,000	10,000	10,000
Highest transaction volume (limit x-axis)	40	2	100
Highest proportion value (limit y-axis)	0.7	0.8	0.13
Amount of threshold levels	20	20	20

Table 5.1: The parameter settings for all case studies.

5.3 Purpose of experiments

This section describes the purpose of different experiments conducted in this thesis to answer the research questions 'How can a threshold tuning strategy be incorporated to accelerate the threshold tuning process of transaction monitoring models?' and 'How can the uncertainty in the threshold tuning process for the chosen threshold level be quantified?'. A total of four different types of experiments were conducted in this thesis. The first three experiments were conducted for BTL testing to determine the most appropriate threshold level for which the proportion of interesting transactions satisfied the RTL with a $(1-\alpha)\times 100\%$ confidence level. The fourth experiment was conducted for the combination of BTL and ATL testing. The four experiments and the purpose of each experiment are as follows:

1. **Experiment 1: Threshold evaluation approaches.**
   In the first experiment, the bucket approach and complete population evaluation approaches, which were described in Section 4.2, are compared. The two approaches will be compared based on the difference between the found threshold level, based on a sample of the transactions, and the true threshold level, based on the population of transactions without sampling. The purpose of this experiment is to demonstrate that the complete population approach can guarantee that the found most appropriate threshold level satisfies the RTL with a $(1-\alpha)\times 100\%$ confidence level whereas the bucket approach cannot offer this guarantee. The complete population approach makes it possible to answer the second research question 'How can the uncertainty in the threshold tuning process for the chosen threshold level be quantified?'.
2. **Experiment 2: Performance evaluation of tuning strategies for initial threshold tuning.**
   From experiment 2 onward, only the complete population approach is used for sampling. In the situation when no information is available, as is the case with initial threshold tuning, the purpose of this experiment is to determine which strategy requires reviewing the least transactions to determine the most suitable threshold level. In addition, the purpose of this experiment is to determine which strategy has the smallest (positive) difference between the found threshold level, based on a sample of the population of transactions and the true threshold level, based on the population of transactions. Furthermore, this experiment investigates how performance improves or deteriorates compared to other threshold tuning strategies when transactions are partitioned with more threshold levels. Moreover, it will also be investigated how the performance of the Boltzmann exploration strategy changes due to different schedules for the temperature parameter τ.
   This experiment provides insight into which tuning strategy is most appropriate in a specific situation and which tuning strategy in general results in the best performance which is important for the first research question. The various strategies will be compared according to the required amount of reviews and the difference between the found and true threshold level.
3. **Experiment 3: Reduction of required reviews with reusing information of reviewed samples.**
   Reviewing all transactions from a new sample of a different threshold level is often inefficient since the collected information from a reviewed sample of a previously tested threshold level can partly be reused to save time as described in Section 4.5.2. The purpose of this experiment is to determine how much time can be won if information from a reviewed sample of a previously tested threshold level is partly reused to determine the most suitable threshold level.
   This experiment offers more insight into how, after choosing the most promising threshold tuning strategy, the threshold tuning process can be further accelerated which is the focus of the first research question. The various strategies will again be compared according to the required amount of reviews and the difference between the found and true threshold level.
4. **Experiment 4: Influence of measure choice on most suitable threshold level.**
   The purpose of this experiment is to demonstrate the influence of the chosen measure on the most suitable threshold level as output.
   The reason for interest in this experiment is the limited insight that the false omission rate as a measure, via BTL testing, provides about the quality of a threshold level. The false omission rate $\frac{FN}{TN+FN}$ answers the question; of all transactions that would not generate an alert, how many would the threshold incorrectly predict as not interesting?, but provides no insight into the productivity of alerted transactions above the threshold level. On the other hand, sensitivity as a measure answers the question; of all transactions reviewed as interesting, how many would

the threshold correctly predict as interesting (meaning that the interesting transactions would generate an alert)? This metric is often used in cases where the classification of false negatives is a priority. The sensitivity for a threshold level can be determined by taking a sample of both the transactions below the tested threshold level (BTL testing) and by taking a sample of the transactions above the tested threshold level (ATL testing). The best performing strategies from experiments two and three will be compared according to the required amount of reviews and the difference between the found and true threshold level.

6

Results

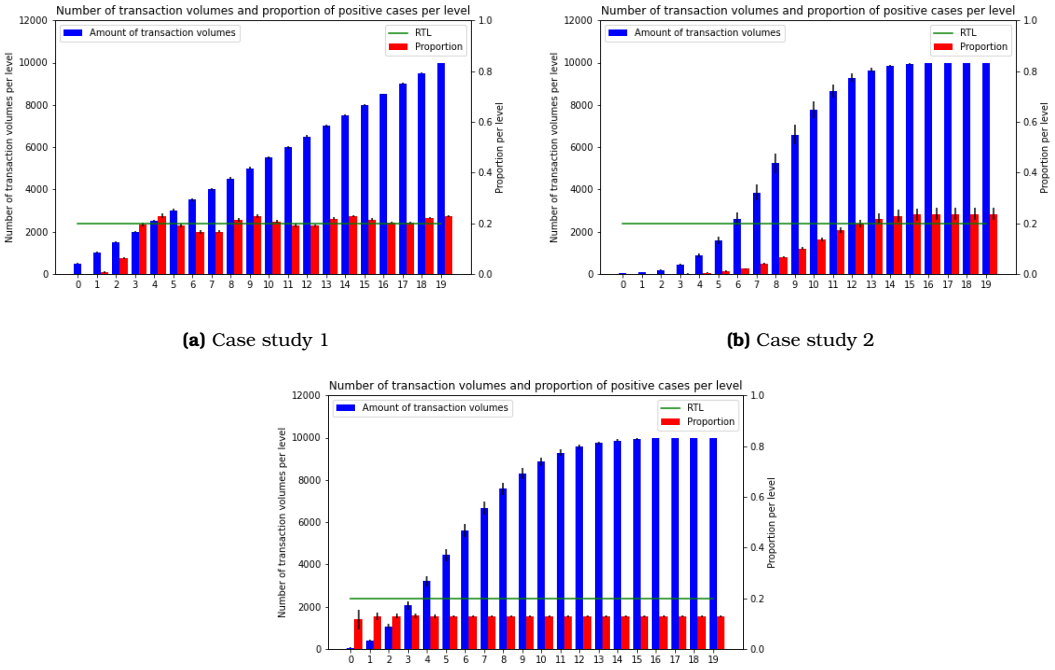
In this chapter, the results of different experiments described in Chapter 5.3 are presented. This chapter is divided into five sections.

Section 6.1 presents the transaction population and the true threshold level for each case study based on the transaction population is described. In addition, Section 6.2 describes the results of the first experiment, in which the bucket and complete population approaches are compared. Subsequently, the performance of the different strategies is evaluated in Section 6.3. This section also investigates the influence of the temperature parameter for the Boltzmann strategy and the effect of the number of threshold levels on the number of required reviews and the difference between the found and true threshold level. Section 6.4 investigates the possible reduction in the required amount of reviews when information from a previously reviewed sample is reused. Lastly, Section 6.5 discusses the influence of the chosen measure on the most appropriate threshold level. The main conclusions of the results are given at the end of each (sub)section.

6.1 Transaction population for each case study

The transaction population for each case study is illustrated in Figure 6.1. The blue bars represent the total number of transactions up to a given threshold level, e.g. the blue bar for threshold level 10 represents the total number of transactions up to threshold level 11. The red bars in Figure 6.1 indicate the proportion of unusual transactions between a certain threshold level and level 0, e.g. the red bar for threshold level 10 represents the number of unusual transactions up to threshold level 11.
The black vertical lines are the standard deviation from the average after 25 runs. The green line is the RTL that is set in advance. The true threshold level based on the population is defined as the highest threshold level for which the red bar is below the green line. The true threshold level per run for each case study can be found in Appendix G.2.

The true threshold level for case study 1 in many runs is equal to level 3 and sometimes level 2, which is also often the point where the red bar first crosses the green line in Figure 6.1a due to the small variation in the proportion each run. For case study 2, the true threshold level is often level 11 or 12 and sporadically level 19, which in Figure 6.1b is also the point where the red bar often crosses the green line for the first time. For case study 3, the true threshold level is equal to level 19 which is shown in Figure 6.1c because the red bar never crosses the green line.



(a) Case study 1

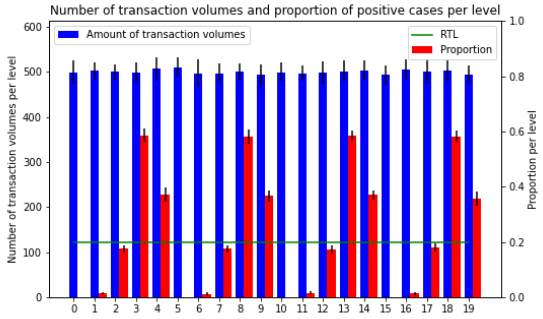
(b) Case study 2

(c) Case study 3

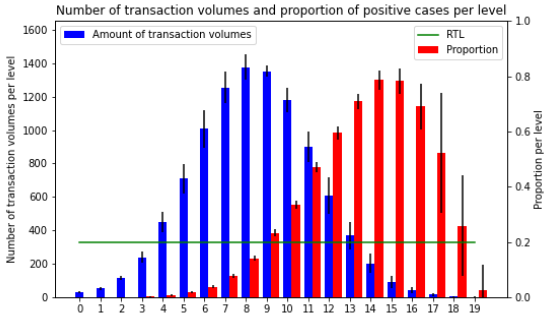
Figure 6.1: The number of transaction volumes and the proportion of unusual transaction volumes of the population below a given threshold level.

6.2 Experiment 1: threshold evaluation approaches

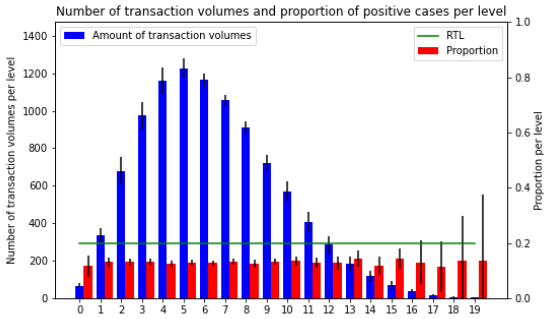
The difference in performance between the bucket and complete population approach is analysed with the step, bisection and multi-level tuning strategies. Section 4.2 briefly explained that the bucket approach for BTL testing has as major disadvantage that the number of interesting transactions further below the evaluated threshold level is unknown, whereby there is a risk of choosing a threshold level in a local minimum for which it is unknown whether a large group of interesting transactions have been missed. This problem is most apparent in case study 1 where the proportion of interesting transactions fluctuates as illustrated in Figure 6.2a. In addition, this problem is the most severe for the strategies that do not start at one of the lowest threshold levels such as the bisection and multi-level strategies. These strategies have a higher risk of testing a threshold level of which the proportion of interesting transactions is a local minimum since these strategies do not start at one of the lowest logical threshold levels which is the case with the step strategy.



(a) Bucket: case study 1



(b) Bucket: case study 1



(c) Bucket: case study 1

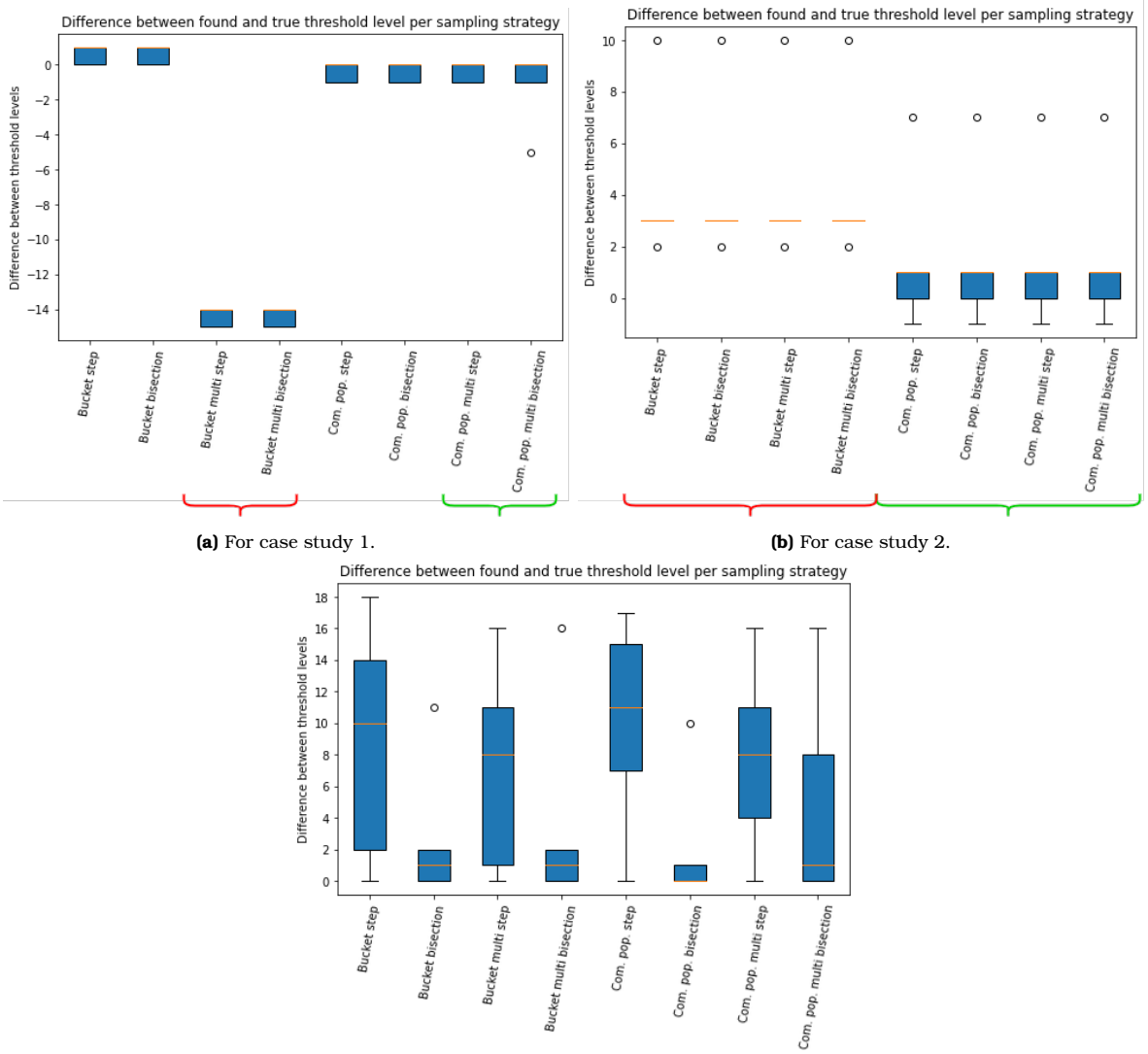
Figure 6.2: The number of transaction volumes and the proportion of unusual transaction volumes of the population in a bucket below a given threshold level.

For BTL testing, in Figures 6.3a, 6.3b and 6.3c, it can be seen that the median difference between the found and true threshold level, $D_{threshold}$, is more often closer to zero using the complete population approach, presented in the four right-hand box plots, in comparison to the bucket approach, presented in the four left-hand box plots.
For case study 1, it can be seen in Figure 6.3a in the red highlighted left-hand box plots that for median difference for the multi-level with the step or bisection strategy using the bucket approach is around 14 levels instead of the median difference of 0, which is highlighted in green, using the complete population approach.
For case study 2, analysing the box plots in Figure 6.3b, it can be seen that the bucket approach results in a median difference of 3 levels, highlighted in red, compared to the median difference of 1 level using the complete population approach, which is highlighted in green. The complete population approach does not necessarily result in a difference of 0 levels. This can be explained by the fact that the most appropriate threshold level is determined based on the upper bound of the proportion p_{upper} in contrast to the true threshold level which is determined based on the population proportion p.
For case study 3, analysing the box plots in Figure 6.3c, it can be seen that the median difference for the bisection and multi-level with the bisection strategy using the bucket or complete population approach is between 0 and 1 level. For the step strategy and the multi-level with the step strategy using the bucket or complete population approach is the median difference between 8 and 11 levels and considerably greater. In addition, the spread in the difference is larger for the step strategy and the multi-level with step strategy which means that the performance is less stable. The step strategy starts by testing the lowest logical threshold level which means that the step strategy must test many threshold levels before it reaches the true threshold level 19. The choice of the constant population proportion $p = 0.13$ could be the cause for the large spread. The difference between the population proportion and the RTL is not large in this case. This has the consequence that a found sample proportion of about $\hat{p} = 0.15$, which results in $p_{upper} = 0.2$ (with a chosen margin of error equal to $e = 0.05$), can already trigger the stop condition for the step strategy. On the other hand, the bisection and multi-level with bisection strategies need to test fewer levels to reach level 19 which reduces the likelihood of triggering the stop

condition, allowing these strategies to end up closer to the true threshold level.

Conclusion

1. Firstly, the complete population approach more often results in a difference close to zero between the found and true threshold level in comparison to the bucket approach.
2. In addition, the bucket approach has the risk of missing interesting transactions that are not in the bucket below the threshold level that is tested.
3. Therefore, the complete population approach as described in Section 4.2.1, in which a sample is taken of all transactions between level 0 and the threshold level that is evaluated with BTL testing, is used in the rest of this thesis.



(a) For case study 1.

(b) For case study 2.

(c) For case study 3.

Figure 6.3: For BTL testing, the difference between the found and true threshold level for the different threshold tuning strategies.

6.3 Experiment 2: performance evaluation

In this section, the performance of the four threshold tuning strategies; step, bisection, multi-level, and Boltzmann exploration is compared. The multi-level strategy combined with the step or the bisection strategy are considered two different tuning strategies resulting in the comparison of five tuning strategies in total. The tuning strategies are compared according to the required reviews of transactions to determine the most appropriate threshold level, the difference between the found and true threshold level and the spread in performance of a threshold tuning strategy over 25 runs as described in Section 4.6.1.

6.3.1 Effect of temperature parameter on Boltzmann Exploration

To compare the performance of the Boltzmann exploration strategy with the other strategies, it is important to first determine the schedule for the temperature parameter for each case study that gives the best performance. Thereafter, the Boltzmann exploration with the best temperature schedule will be compared with the other strategies. The best schedule is defined as the one that results in the fewest required reviews.

Experiments with two different types of temperature schedules were carried out to determine the best schedule for the temperature parameter for each case study.

- **Constant schedule:** the temperature parameter is kept constant over time and is executed for five large values for $\tau(0) = [200, 160, 120, 80, 40]$, focusing on exploration and for five small values for $\tau(0) = [20, 15, 10, 5, 1]$, focusing on exploitation which is described with the formula:

$$\tau(t) = \tau(0) \quad \text{for } t = 1, ..., K * n \tag{6.1}$$

- **Decreasing schedule:** the value for the temperature parameter decreases over time and is executed for five values for $\tau(0) = \left[K * n, \frac{K*n}{2}, \frac{K*n}{3}, \frac{K*n}{4}, \frac{K*n}{5}\right] = [5280, 2640, 1760, 1320, 1056]$ for $K = 20$ threshold levels and a sample size equal to $n = 264$, using $p_0 = 0.2, \alpha = 0.05$ and $e = 0.05$ as described in Table 3.3. The idea of the decreasing temperature is to first focus more on exploration and to shift the focus to exploitation as more time has passed and more information about the different threshold levels is available which is described with the formula:

$$\tau(t) = max\left(\frac{\tau(0)}{t}, 1\right) \quad \text{for } t = 1, ..., K * n \tag{6.2}$$

  It was chosen that $\tau(t)$ should always be greater than or equal to 1, otherwise the exponential component in equation 4.8 explodes.

These two schedules executed for ten and five different values for $\tau(0)$ result in a total of fifteen box plots for each case study for the following values of $\tau(0)$:

$$\tau(0) = [\tau_1, ..., \tau_{15}] = [5280_d, 2640_d, 1760_d, 1320_d, 1056_d, 200_c, 160_c, 120_c, 80_c, 40_c, 20_c, 15_c, 10_c, 5_c, 1_c]$$

where underscore d denotes the decreasing schedule and underscore c denotes the constant schedule.

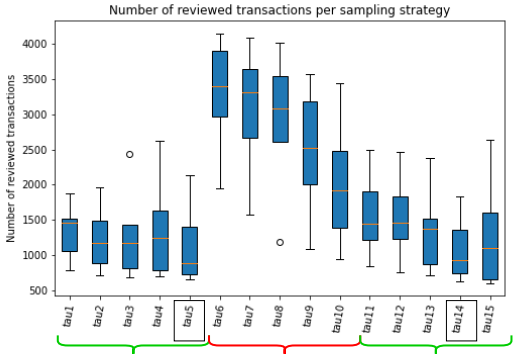
Tuning results of the temperature parameter

The required amount of reviews for all three case studies for $\tau(0) = [\tau_1, ..., \tau_{15}]$ with $K = 20$ threshold levels are presented in box plots in Figure 6.4. The median value is indicated with the orange line, the 25% and 75% quantiles with the blue box and the maximum and minimum values with the black line. A maximum or minimum as an outlier is indicated by a circle. The box plots illustrating the difference between the found and true threshold level for all three case studies for $\tau(0) = [\tau_1, ..., \tau_{15}]$ are presented in Figure F.1 in Appendix F since the difference is only taken into consideration if multiple τ_j result in a similar performance and to restrict the number of figures in the report.
The following conclusions can be drawn about the effect of the temperature schedule and value of $\tau(0)$ for the three case studies with $K = 20$ threshold levels:

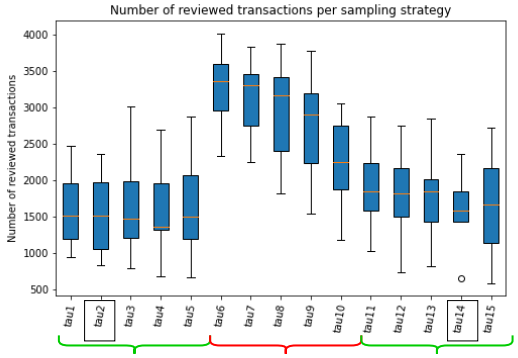
1. For all three case studies, the decreasing schedule, $\tau_1, ..., \tau_5$, generally results in fewer required reviews than the constant schedule with the focus on exploration, $\tau_6, ..., \tau_{10}$, indicated by a green bar for the first five τ_j values and a red bar for the middle five τ_j values in Figure 6.4.
2. For all three case studies, the constant schedule with the focus on exploitation, $\tau_{11}, ..., \tau_{15}$, with low $\tau(0)$ values generally result in fewer required reviews than the constant schedule with the focus on exploration, $\tau_6, ..., \tau_{10}$, with high $\tau(0)$ values, indicated by a green bar for the last five τ_j values and a red bar for the middle five τ_j values in Figure 6.4.
3. For all three case studies, the decreasing schedule, $\tau_1, ..., \tau_5$, generally results in a similar number of required reviews as the constant schedule with the focus on exploitation, $\tau_{11}, ..., \tau_{15}$, indicated by the green bars for the first and last five τ_j values in Figure 6.4.

Conclusion
In this thesis, it is chosen to continue with τ_{14} for case studies 1 and 3 with $K = 20$ threshold levels, since τ_5 and τ_4 resulted in a larger difference between the found and true threshold level and a relatively larger spread as illustrated in Figure F.1 in Appendix F.1. τ_2 is chosen for case study 2 with $K = 20$ threshold levels since it resulted more often in a lower number of required reviews than τ_{14}.

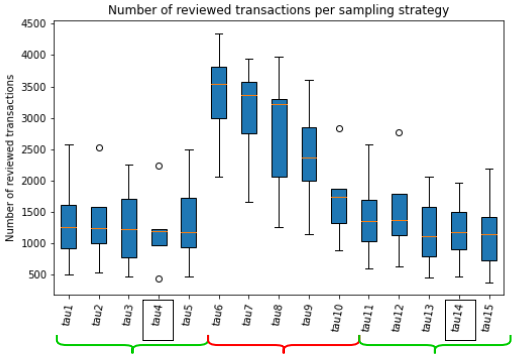
The Boltzmann exploration strategy with these specific τ_j values is compared with the other four tuning strategies.



(a) For case study 1.



(b) For case study 2.

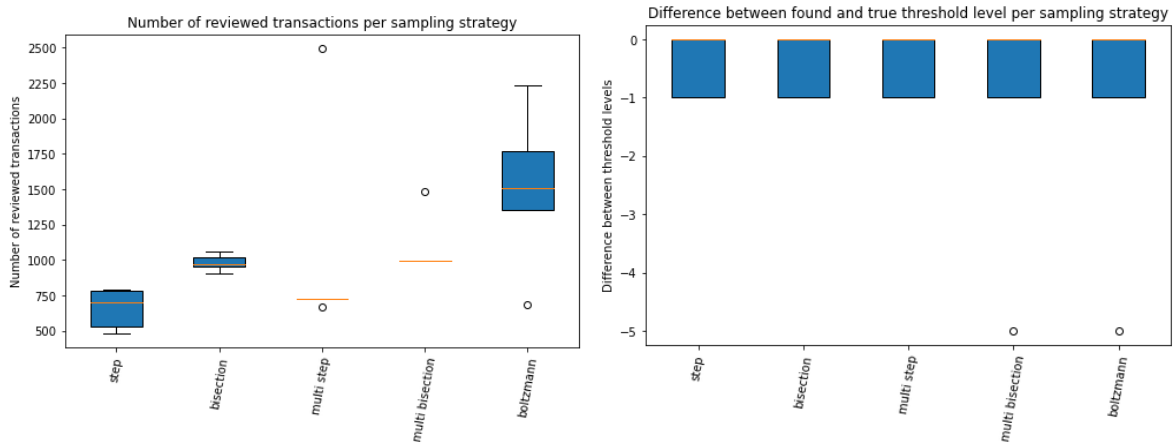


(c) For case study 3.

Figure 6.4: For BTL testing, the required amount of reviewed transactions for a decreasing or constant schedule for different $\tau(0)$ values.

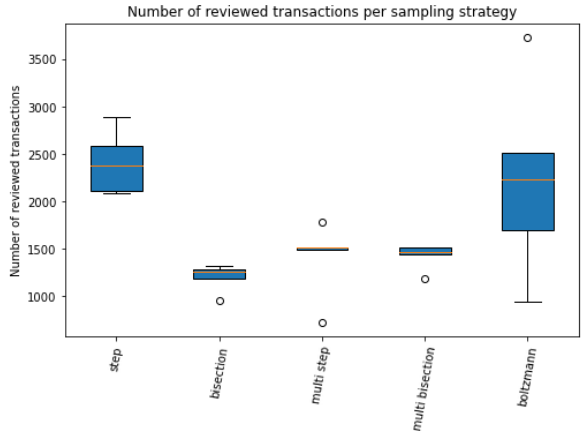
6.3.2 Performance of threshold tuning strategies

In this section, the performance of the five threshold strategies; step, bisection, multi-level with step, multi-level with bisection and Boltzmann exploration are compared. The required amount of reviews and the difference between the found and true threshold level for each case study with $K = 20$ threshold levels and a sample size of $n = 264$ is presented in Figure 6.5.

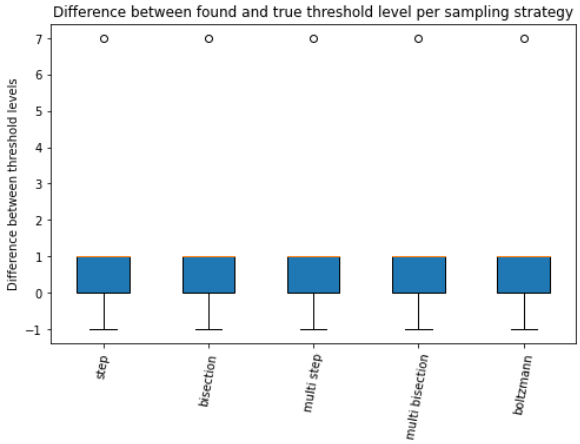


(a) Reviews for case study 1

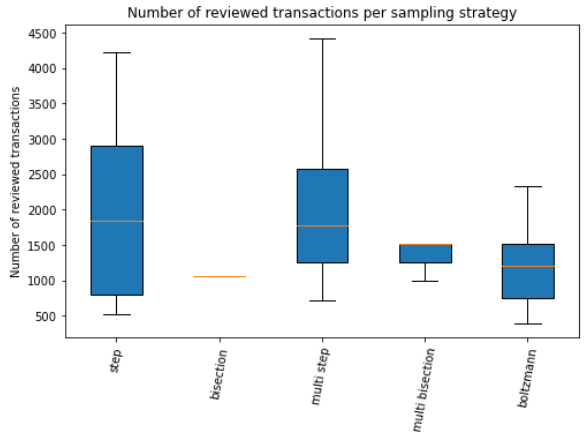
(b) Difference for case study 1



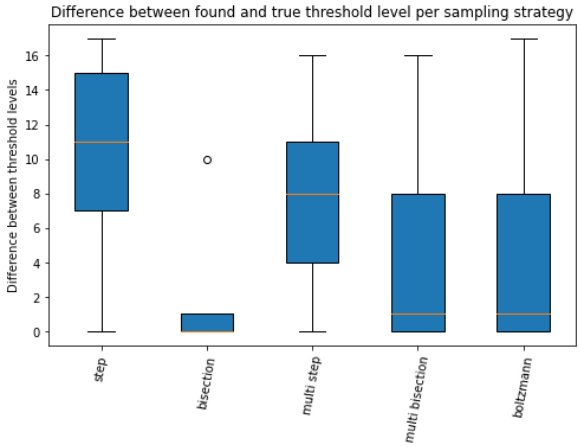
(c) Reviews for case study 2



(d) Difference for case study 2



(e) Reviews for case study 3



(f) Difference for case study 3

Figure 6.5: For BTL testing, the required reviews and difference between found and true threshold level with $K = 20$ threshold levels and a sample size of $n = 264$ for the different tuning strategies.

Evaluation of performance based on the required reviews

The following conclusions can be drawn about the required reviews for the three case studies with $K = 20$ threshold levels:

1. **Median of required reviews:** In general, the median required amount of reviews is the smallest with the bisection strategy, except for case study 1 for which the true threshold is low resulting in

a better performance of the step strategy. On the other hand, the step and Boltzmann strategy generally require the most reviews.

2. **Spread in required reviews:** For all three case studies, the bisection and multi-level with bisection strategies are the most stable strategies and result in a similar or smaller spread of the required reviews compared to the other strategies.
3. Behaviour of the strategies per case study:
   (a) **Case study 1 (non-monotonic function):** Taking into account the median required reviews, the step strategy requires the least reviews of transactions to determine the most suitable threshold level because the true threshold of 2 or 3 is relatively low. After that, the multi-level with step strategy requires the fewest reviews, followed by the bisection and multi-level with bisection strategy. The bisection and multi-level with step or bisection strategies are the most stable with the smallest spread and the Boltzmann exploration strategy is the least stable with the largest spread in performance over 25 runs.
   (b) **Case study 2 (monotonic non-decreasing function):** The same conclusions hold as for case study 1 except that the step strategy, considering the median required reviews, now needs the most reviews because the true threshold, which is often between levels 11 and 19, is now considerably higher than in case study 1.
   (c) **Case study 3 (constant function):** In case study 3, the performance of the different strategies changed slightly compared to case studies 1 and 2. The bisection strategy still requires the least reviews of transactions to determine the most suitable threshold level. However, it is worth noting that the step and multi-level with step strategies have relatively larger spreads compared to the other strategies. The choice of the constant population proportion $p = 0.13$ in case study 3 could be the cause for the large spread, which has the consequence that a found sample proportion of about $\hat{p} = 0.15$, which could result in $p_{upper} = 0.2$ (with a chosen margin of error equal to $e = 0.05$), can already trigger the stop condition for the step strategy. In general, the step and multi-level with step strategies need to test more levels to reach level 19, which makes it more likely that these strategies trigger the stop condition, $p_{upper} > RTL$, on any of the levels between the lowest level and level 19 resulting in the larger spread.

Evaluation of performance based on the difference

In addition, considering the difference between found and true threshold level for the three case studies with $K = 20$ threshold levels, the following conclusions may be drawn:

1. **Median difference:** For all three case studies, the bisection, multi-level with bisection and Boltzmann strategies result in a median difference of 0 or 1 between the found and true threshold level. This is also the case for step and the multi-level with step strategies in case studies 1 and 2 but not for case study 3.
2. **Spread in the difference:** For all three case studies, the bisection strategy is the most stable strategy and results in a similar or smaller spread in the difference compared to the other strategies. The other strategies show a less stable performance which is most evident in case study 3.
3. Behaviour of the strategies per case study:
   (a) **Case study 1 (non-monotonic function):** Considering the median difference between the found and true threshold level, all strategies result in a difference of 0. In addition, sometimes the strategies end up one level too high. A possible explanation for this is that the true threshold level is often level 3 and sometimes level 2. The strategies almost always end up at level 3 since level 3 has a proportion of interesting transactions approximately equal to the RTL, which explains the difference of one level. Furthermore, the bisection, the multi-level with step or bisection and Boltzmann exploration strategies have a chance to determine that the most suitable threshold level is in local minima, illustrated in Figure 6.1a around level 6, 7, 11 or 12 because these strategies do not start at the lowest logical level like the step strategy. The small circles in the box plots for multi-level with bisection and Boltzmann strategies show this behaviour, where the strategy ended in a local minimum 5 levels above the true threshold.

(b) **Case study 2 (monotonic non-decreasing function):** The same conclusions hold as for case study 1, except that the strategies more often end below the true threshold. The true threshold level for case study 2 is often level 11 or 12 and sporadically level 19. The strategies regularly end up one level too low. This happens since the upper confidence bound is used as the stop criterion. This upper confidence bound will often be above the population proportion which has the consequence that all strategies often end up one level below the true threshold. Moreover, threshold level 19 is sporadic the true threshold level while the strategies may end at level 12 which explains the small circles at a difference of 7 levels in Figure 6.5d. Finally, the strategies occasionally end one level too high, which is probably caused by the fact that the proportion of levels 11 to level 13 differ very little.

(c) **Case study 3 (constant function):** The true threshold level for case study 3 is always the highest level which is level 19. In case study 3, the difference between found and true threshold level of the different strategies changes considerably compared to case studies 1 and 2. The bisection strategy still has a median difference of 0 and has a small spread in performance. However, the median difference becomes noteworthy larger for the step and multi-level with step strategies. The step and multi-level with step need to test more levels to reach level 19, which makes it more likely that these strategies trigger the stop condition, $p_{upper} > RTL$, resulting in the larger spread and the large positive median difference. In contrast, the bisection strategy has the smallest spread in the difference because it generally needs to test fewer levels to reach level 19 which reduces the likelihood of triggering the stop condition, allowing this strategy to end up closer to the true threshold level.
In addition, the multi-level with bisection and Boltzmann exploration strategies have a median difference close to 0 but have a large spread. For the multi-level strategy, this is possibly caused by the fact that the strategy variably starts by testing levels 4, 8, 12 or 16, because p_{upper} of a sample proportion is around the RTL. As a result, the strategy has to test a different number of levels which has the consequence that the strategy does not always end at level 19, explaining the spread. In addition, transactions of all threshold levels are reviewed simultaneously with the Boltzmann exploration strategy. The strategy can regularly test a level lower than level 19 for which $p_{upper} > RTL$ applies since for all levels the difference between the upper bound and RTL, $d = RTL - p_{upper}$, is similar and around 0, which explains the spread.

The found threshold level with different tuning strategies and true threshold level per run for each case study can be found in Appendix G.2.

Conclusion

Considering the required amount of reviews, the difference between the found and true threshold level and the spread in performance over 25 runs, the five strategies can be ranked as presented in Tables 6.1 and 6.2. From this ranking, it can be concluded that the bisection strategy generally results in the least required reviews, a median difference close to zero and has the most stable performance expressed in the small spread over 25 runs, which is highlighted in green in the Tables.

Case 1	**Step**	**Bisection**	**Multi step**	**Multi bisection**	**Boltzmann**
Lowest median	1	2	1	2	3
Smallest spread	3	2	1	1	4
Case 2					
Lowest median	3	1	2	2	3
Smallest spread	2	1	1	1	3
Case 3					
Lowest median	3	1	3	2	2
Smallest spread	4	1	4	2	3
Total sum	16	8	12	10	18
Total ranking	4	1	3	2	5

Table 6.1: A ranking of the five strategies based on the required amount of reviews with $K = 20$ threshold levels, 1 indicating the best and 5 the worst performing strategy.

Case 1	**Step**	**Bisection**	**Multi step**	**Multi bisection**	**Boltzmann**
Lowest median	1	1	1	1	1
Smallest spread	1	1	1	2	2
Case 2					
Lowest median	1	1	1	1	1
Smallest spread	1	1	1	1	1
Case 3					
Lowest median	4	1	3	2	2
Smallest spread	2	1	2	2	2
Total sum	10	6	9	9	9
Total ranking	3	1	2	2	2

Table 6.2: A ranking of the five strategies based on the difference between found and true threshold level with $K = 20$ threshold levels, 1 indicating the best and 5 the worst performing strategy.

6.3.3 Effect of amount of threshold levels

This section investigates the effect of the number of threshold levels on the performance of the tuning strategies. To investigate this influence, the population of transactions is doubled from 10,000 to 20,000 transactions and the number of threshold levels from 20 to 40, but all other parameters as α, e and p_0 are kept the same. It is interesting to investigate this effect when an analyst needs to handle a large population of transactions but still wants to accurately determine the most appropriate threshold level, which requires splitting the transactions with more threshold levels.

Both the relative change in the median and spread and the absolute values for the median and spread in required reviews and the difference between the found and true threshold are considered to determine which strategy has the best performance. The change in median $\Delta median$ and spread $\Delta spread$ in the required amount of reviews for all three case studies with $K = 40$ instead of $K = 20$ threshold levels is presented in Table 6.3. The change in the median and spread in the difference between the found and true threshold level for all three case studies with $K = 40$ instead of $K = 20$ threshold levels is presented in Table 6.4. The change in the median and spread is calculated as follows:

$$\Delta median = \frac{X_{median,k=40}}{X_{median,k=20}} \quad \text{for required amount of reviews} \tag{6.3}$$

$$\Delta median = X_{median,k=40} - X_{median,k=20} \quad \text{for the difference} \tag{6.4}$$

$$\Delta spread = \begin{cases} \frac{X_{max,k=40} - X_{min,k=40}}{X_{max,k=20} - X_{min,k=20}} & \text{if } (X_{max,k=20} - X_{min,k=20}) \neq 0 \\ X_{max,k=40} - X_{min,k=40} & \text{if } (X_{max,k=20} - X_{min,k=20}) = 0 \end{cases} \tag{6.5}$$

where X is the number of reviews or the difference between the found and true threshold level.

A comparison of the performance with $K = 40$ instead of $K = 20$ threshold levels is illustrated with box plots as Figure 6.5 in Figure 6.6.

Results: the effect of the amount of threshold levels based on required reviews

From Table 6.3 and Figure 6.6, the following conclusions can be drawn:

1. **Median of required reviews:** When the relative performance of the strategies is compared with the situation with $K = 20$ threshold levels, the median required reviews generally increases. An exception is case study 2, in which the median decreases of the multi-level with step strategy. However, the spread in the required reviews in case study 2 increases significantly as illustrated in Figure 6.6c and presented with the orange cell in Table 6.3. In case study 3, the median decreases for the step strategy, but the median difference between the found and true threshold is large demonstrating that this strategy often ends at a too low threshold level as presented in Figure 6.6f and presented with the orange cell in Table 6.4. Excluding the step and multi-level with

step strategies because they have some undesirable behaviour in case studies 2 and 3, the median required reviews increases the least for the bisection strategy presented with the green marked cell 'average Δ median' in Table 6.3. The median required reviews increases the most for the Boltzmann strategy.

2. **Spread in required reviews:** Not considering the step and multi-level with step strategies again due to undesirable behaviour in case studies 2 and 3, when the relative performance of the strategies is compared to the situation with $K = 20$ threshold levels, the spread in required reviews increases the least for the multi-level with bisection strategy illustrated with the green marked cell 'average Δ spread' in Table 6.3. The spread in required reviews increases the most for the step strategy which is caused by the results in case study 2.
3. Considering the absolute performance, as in the situation with $K = 20$ threshold levels, the bisection strategy results in the best performance with a low median required reviews and a small spread for all cases as illustrated in left box plots of Figure 6.6.

Case 1	**Step**	**Bisection**	**Multi Step**	**Multi Bisection**	**Boltzmann**
Δ median	1.372	1.295	2.136	1.770	1.349
Δ spread	3.083	2.340	0.239	0.785	2.002
Case 2					
Δ median	1.778	1.044	1.883	1.193	1.177
Δ spread	3.455	0.956	2.997	0.988	1.435
Case 3					
Δ median	0.429	1.250	1.019	1.022	1.166
Δ spread	1.429	0.000	0.929	1.500	0.946
Average Δ median	1.193	1.196	1.679	1.328	1.231
Average Δ spread	2.655	1.099	1.388	1.091	1.461

Table 6.3: A comparison of the median and spread in the required amount of reviews with $K = 40$ instead of $K = 20$ threshold levels for all case studies.

Results: the effect of the number of threshold levels based on the difference

From Table 6.4 and Figure 6.6, the following conclusions can be drawn:

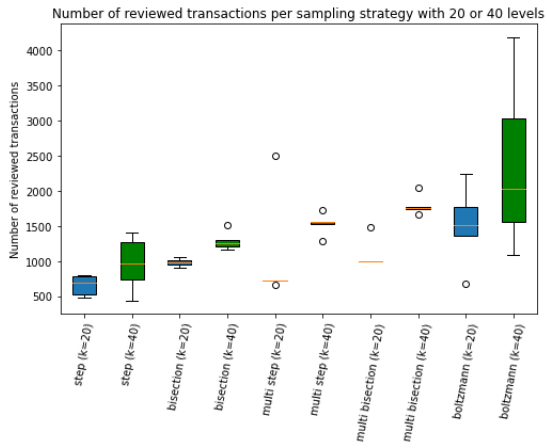
1. **Median difference:** When the relative performance of the strategies is compared with the situation with $K = 20$ threshold levels, the median difference generally increases. The median difference increases the least for the multi-level with bisection strategy presented with the green marked cell 'average Δ median' in Table 6.4, followed by the bisection and Boltzmann strategy . The median required reviews increases the most for the step and multi-level with step strategies which is mainly caused by case study 3 where these strategies often end at a too low a threshold level, presented by the orange marked cells in Table 6.4.
2. **Spread in difference:** Not considering the step and multi-level with step strategies again due to undesirable behaviour in case studies 2 and 3, when the relative performance of the strategies is compared to the situation with $K = 20$ threshold levels, the spread in the difference increases the least for the bisection strategy presented with the green marked cell 'average Δ spread' in Table 6.4. The spread in the difference increases the most for the Boltzmann strategy which is caused by the results in case study 1.
3. Considering the absolute performance, as in the situation with $K = 20$ threshold levels, the bisection strategy results in the best performance with a low median difference and a small spread in the difference between the found and true threshold level for all cases as illustrated in right box plots of Figure 6.6.

Case 1	**Step**	**Bisection**	**Multi Step**	**Multi Bisection**	**Boltzmann**
Δ median	0.000	0.000	0.000	0.000	0.000
Δ spread	2.000	1.000	2.000	1.800	4.000
Case 2					
Δ median	2.000	2.000	2.000	2.000	2.000
Δ spread	1.875	1.000	1.875	1.875	2.250
Case 3					
Δ median	18.000	0.000	13.000	-1.000	0.000
Δ spread	1.353	1.000	2.188	1.938	1.941
Average Δ median	6.667	0.667	5.000	0.333	0.667
Average Δ spread	1.743	1.000	2.021	1.871	2.730

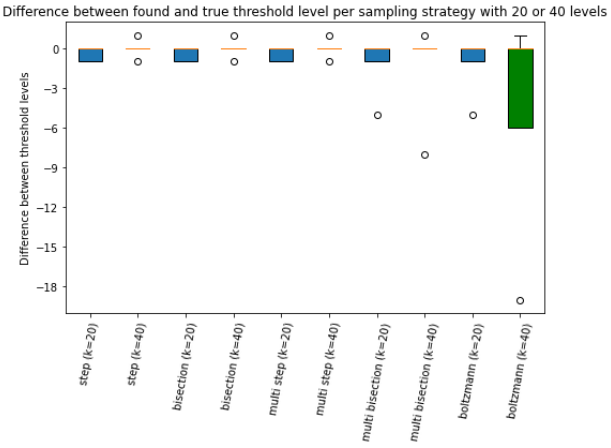
Table 6.4: A comparison of the median and spread in the difference between found and true threshold level with $K = 40$ instead of $K = 20$ threshold levels for all case studies.

Conclusion

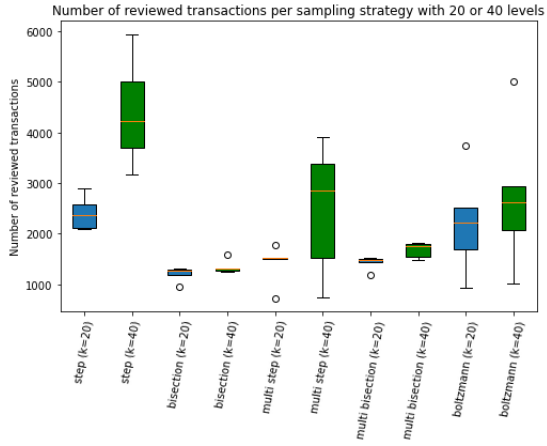
1. Considering the required amount of reviews, the difference between the found and true threshold level and the spread in performance over 25 runs, the number of threshold levels has the effect that the median required reviews and difference in general increases with more threshold levels as well as the spread in the required reviews.
   The bisection strategy generally has the lowest absolute median required reviews and difference and a relatively small increase in the required amount of reviews and difference with twice as many threshold levels. In addition, the spread in the required reviews and difference does not increase considerably, relative to the other strategies, with twice as many threshold levels and the bisection strategy has a stable performance both with $K = 40$ and $K = 20$ threshold levels.
2. The step, multi-level with step and Boltzmann strategies have the least stable performance both looking at the required reviews and the difference between the found and true threshold level where the number of threshold levels clearly affects the performance of these strategies.



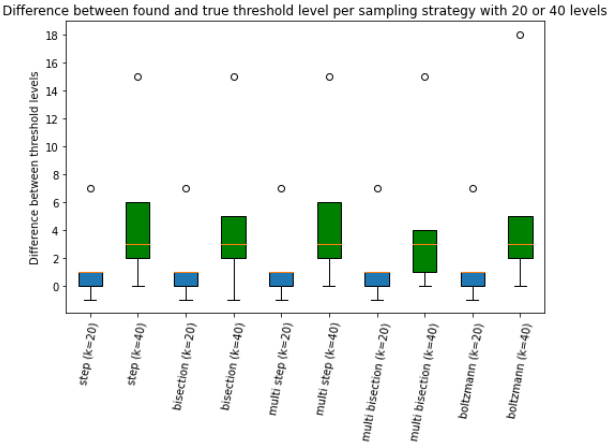
(a) Change in reviews for case study 1



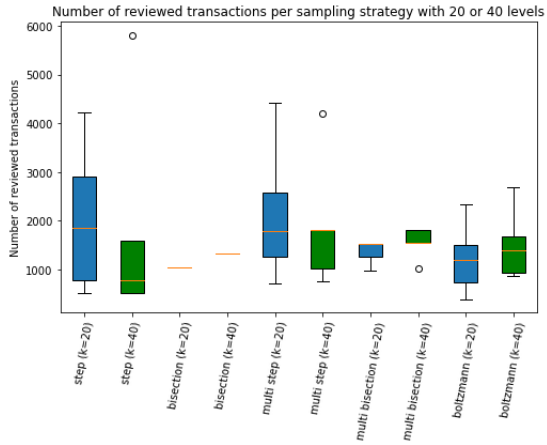
(b) Change in difference for case study 1



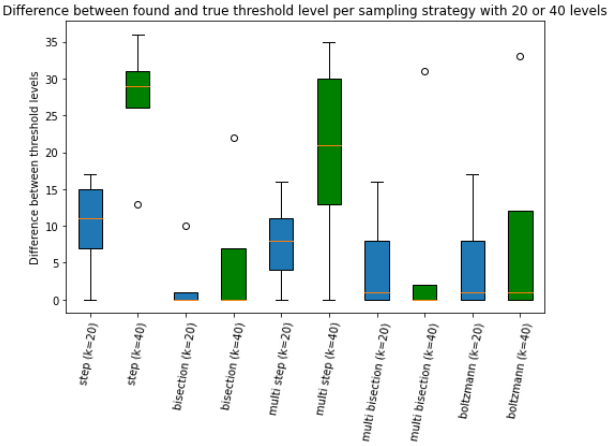
(c) Change in reviews for case study 2



(d) Change in difference for case study 2



(e) Change in reviews for case study 3



(f) Change in difference for case study 3

Figure 6.6: For BTL testing, the effect of the number of threshold levels on the required reviews and the difference between found and true threshold level with $K = 40$ instead of $K = 20$ threshold levels for the different tuning strategies.

6.4 Experiment 3: Reduction of required reviews by reusing information

As briefly explained in Section 4.5.2, reviewing all transactions of a new sample does not always provide new insights. This experiment investigates the reduction in required reviews of transactions if accumulated knowledge of transactions from a previously reviewed sample of a different threshold level is reused as described in Section 4.5.2. The reduction in required reviews is calculated as the required reviews with reusing information divided by the required review without reusing information to determine the most suitable threshold level. The results are presented in Table 6.5 and illustrated in box plots in Figure 6.7.

Results: the effect of reusing information based on the required reviews

The results in Table 6.5 show that, on average, for all strategies, between 37% and 54% of the required reviews is needed when information is reused from a previously reviewed sample of a different threshold level as illustrated in the last row of Table 6.5. This is a large reduction in the median required reviews and workload for analysts. The results show that the bisection strategy has the largest average reduction in required reviews although the difference in average reduction is not large compared to the step strategy. On the other hand, the multi-level with bisection strategy has the least average reduction. A possible reason why the reduction is less for the multi-level and Boltzmann strategies is that these strategies have an exploration phase where 'time' is lost because the reuse of information especially reduces the required reviews in the exploitation phase. In addition, the left box plots in Figure 6.7 show clearly that the spread in required reviews narrows when information from a previously reviewed sample is reused.

However, one can wonder what happens to the difference between the found and true threshold level when information from a previously reviewed sample is reused. For example, does the difference increase because the information from a previously reviewed sample is reused? To answer that question, the median difference between found and true threshold level with and without reusing information from a previously reviewed sample is shown in Table 6.6.

Case 1	**Step**	**Bisection**	**Multi Step**	**Multi Bisection**	**Boltzmann**
With reuse	329	507	529	686	831
Without reuse	698	973	726	990	1504
Reduction	0.471	0.521	0.729	0.693	0.553
Case 2					
With reuse	955	378	653	706	1041
Without reuse	2376	1263	1518	1471	2228
Reduction	0.402	0.299	0.430	0.480	0.467
Case 3					
With reuse	569	315	635	672	507
Without reuse	1848	1056	1782	1518	1207
Reduction	0.308	0.298	0.356	0.443	0.420
Average reduction of reviews	0.394	0.373	0.505	0.539	0.480

Table 6.5: A comparison of the median required reviews with and without using knowledge of a previously reviewed sample of a different threshold level.

Results: the effect of reusing information based on the difference

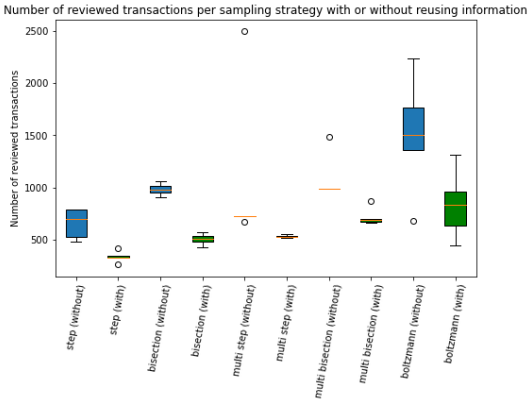
The change in difference is calculated as the difference without reusing information minus the difference with reusing information to determine the most suitable threshold level. The right box plots in Figure 6.7 illustrate that the median difference between the found and the true threshold does not considerably change when information from a previously reviewed sample is reused. The results in Table 6.6 show that the average median difference between the found and the true threshold does not change for the bisection strategy, as presented in the last row of Table 6.6. For the other strategies, the average median difference changed to a median difference between 0 and 1 greater than without reusing information.

Case 1	**Step**	**Bisection**	**Multi Step**	**Multi Bisection**	**Boltzmann**
With reuse	0	0	0	0	0
Without reuse	0	0	0	0	0
Δ in difference	0	0	0	0	0
Case 2					
With reuse	1	1	1	0	-1
Without reuse	1	1	1	1	1
Δ in difference	0	0	0	-1	-2
Case 3					
With reuse	10	0	6	0	0
Without reuse	11	0	8	1	1
Δ in difference	-1	0	-2	-1	-1
Average change in difference	-0.333	0.000	-0.667	-0.667	-1.000

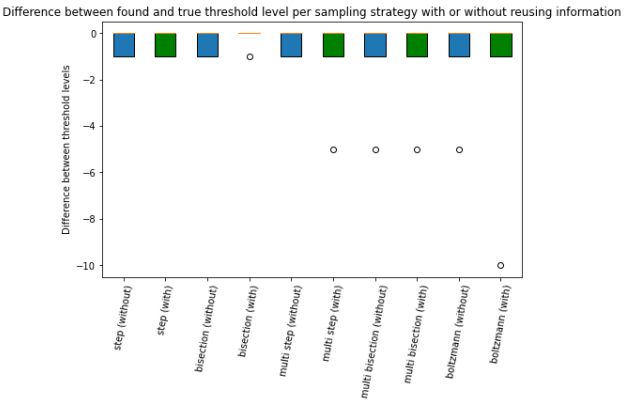
Table 6.6: A comparison of the median difference between found and true threshold level with and without using knowledge of a previously reviewed sample of a different threshold level.

Conclusion

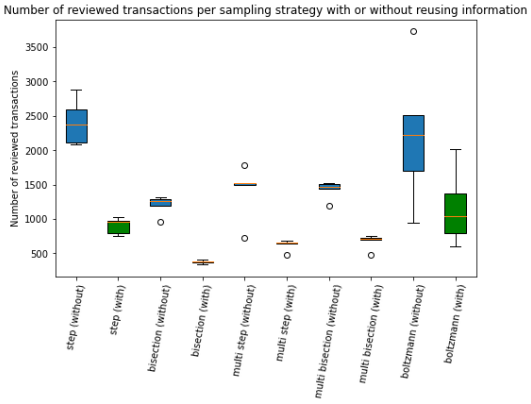
1. Considering the required amount of reviews, the difference between the found and true threshold level and the spread in performance over 25 runs, reusing information of a previously reviewed sample of a different threshold level is beneficial. In general, less than 50% of the required reviews is needed when information is reused from a previously reviewed sample with BTL testing. In addition, the spread in required reviews becomes noteworthy narrower with reusing information from a previously reviewed sample. Overall, the step and bisection strategy have the greatest advantage of reusing information from a previously reviewed sample with the largest reduction.
2. In general, the median difference between found and true threshold level remains similar for the strategies, even though all strategies except the bisection strategy sometimes end up 1 or 2 levels higher compared to the situation when information is not reused. Furthermore, the spread in the difference remains similar compared to the situation without reusing information except for the step and multi-level with step strategies, where the spread widens for case study 3.



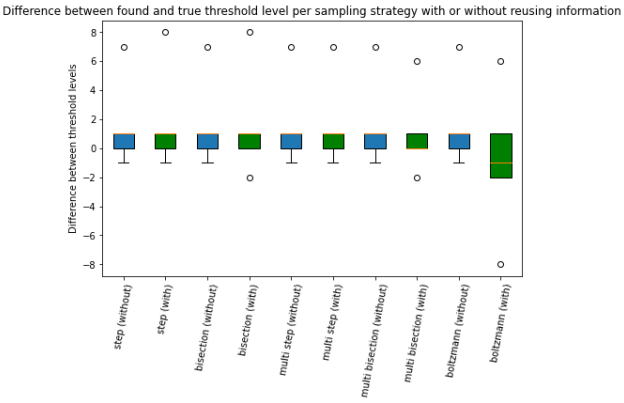
(a) Reduction in reviews for case study 1



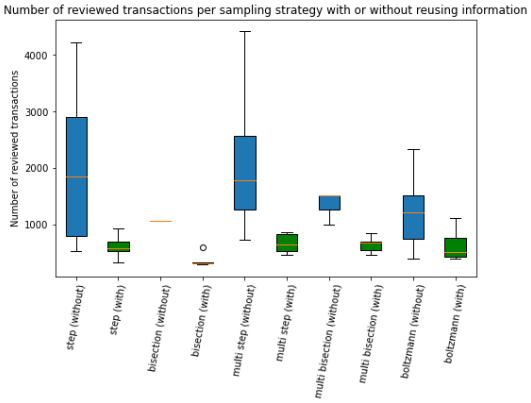
(b) Change in difference for case study 1



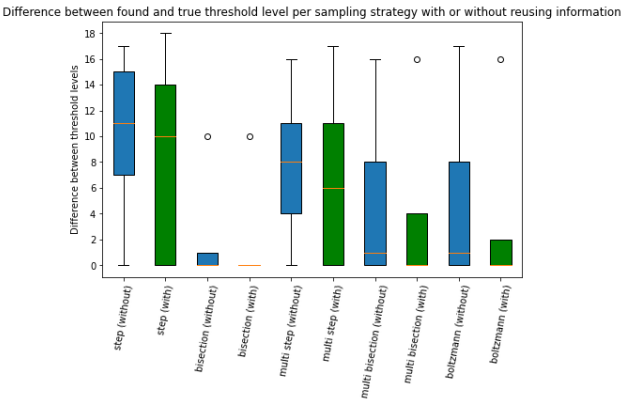
(c) Reduction in reviews for case study 2



(d) Change in difference for case study 2



(e) Reduction in reviews for case study 3



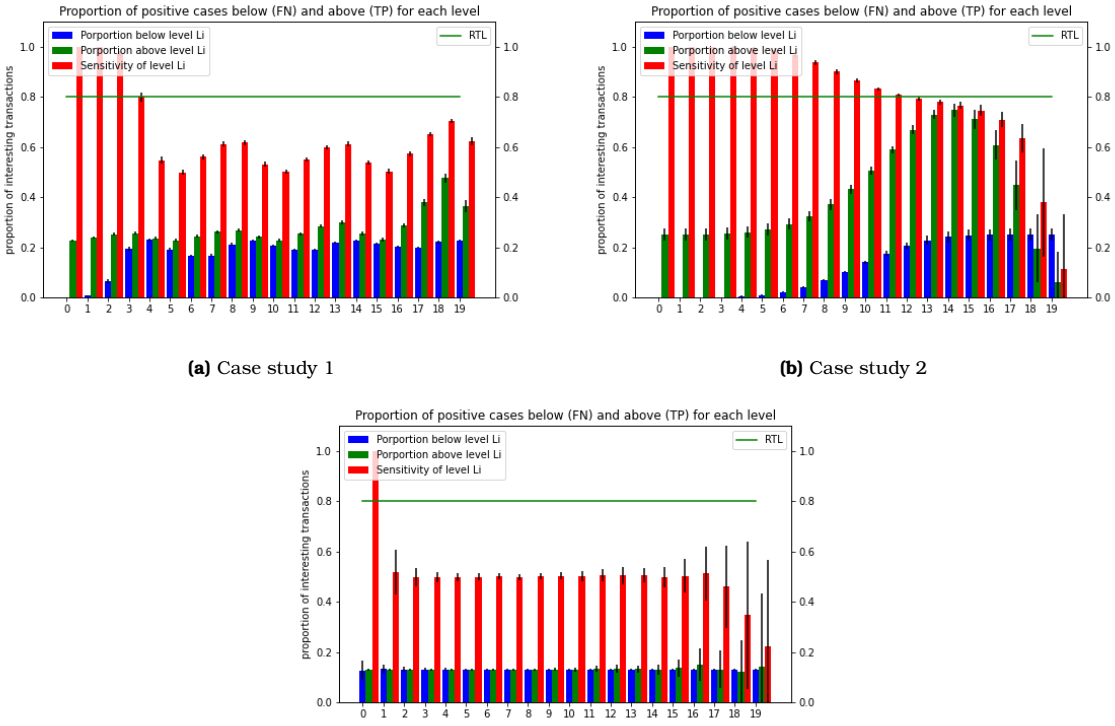
(f) Change in difference for case study 3

Figure 6.7: For BTL testing, the reduction in the required reviews and the change in the difference between found and true threshold level with and without reusing information of a previously reviewed sample for the different tuning strategies.

6.5 Experiment 4: Influence of measure choice

In this thesis and all experiments carried out so far, the main focus was on quantifying the risk of missing interesting transactions below the threshold. The false omission rate $\frac{FN}{TN+FN}$, which is 1 minus the negative predictive value $1 - \frac{TN}{TN+FN}$ was used as a performance criterion in the experiments. However, other measures can also be chosen which can result in a different true threshold level. The false omission rate answers the question: of all transactions that would not generate an alert, how many would the threshold incorrectly predict as not interesting? The sensitivity, on the other hand, answers the question: of all transactions reviewed as interesting, how many would the threshold correctly predict as interesting, meaning that the interesting transactions would generate an alert? This metric is often used in cases where the classification of false negatives is a priority.

Suppose one is interested in the sensitivity of a threshold level in addition to the risk of missing interesting transactions and that there is a preference for four times fewer false negative transactions than true positive transactions. This means that one is looking for a threshold level for which the sensitivity is at least $\frac{4}{4+1} \times 100\% = 80\%$. The sensitivities for the three case studies are illustrated in Figure 6.8. These figures show that the true threshold level can change by choosing a different measure. For case study 3, the true threshold level changes from level 19, when looking for a threshold level for which the maximum false omission rate is lower than the RTL of 20%, while the true threshold level is level 0 when looking for a threshold with a minimum sensitivity of 80%. This example shows the influence of the chosen measure on which threshold is considered the most appropriate threshold level.



(a) Case study 1

(b) Case study 2

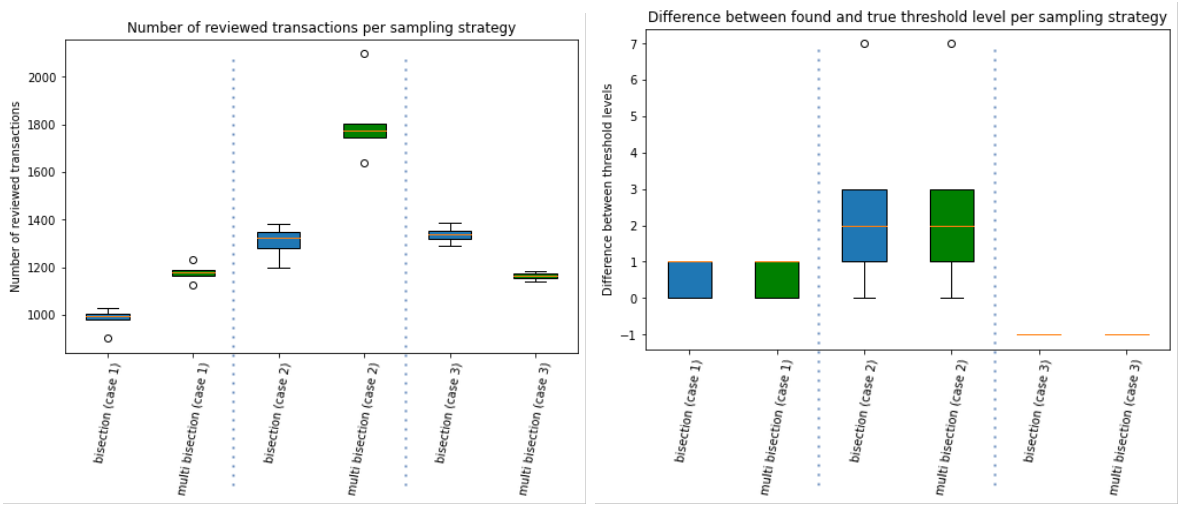
(c) Case study 3

Figure 6.8: The sensitivity and the population proportion of unusual transaction volumes below (FN) and above (TP) a given threshold level.

Results: the influence of measure on the performance of strategies

The influence of the sensitivity as the measure on the performance of bisection and multi-level with bisection strategies was tested on the three case studies with $K = 20$ threshold levels and the results are presented in Figure 6.9. The same sample size is taken for the sample below as above a threshold level that is evaluated. In this case, the early break-off principle described in Section 4.5.1 was not applied as

it was specifically defined for the false omission rate as a measure. However, reusing information from a previously reviewed sample of a different threshold level was applied. It was chosen to investigate only the performance of the bisection and multi-level with bisection strategies as they resulted in the best performance so far.



(a) Reviews for all cases

(b) Difference for all cases

Figure 6.9: The required reviews and the difference between the found and true threshold level for the bisection and multi-level with bisection strategies for all cases using the sensitivity as the measure.

From the results, it can be concluded that the required amount of reviews increases considerably. This could be expected as both transactions above and below a threshold level should be reviewed by an analyst. Using the false omission rate as a measure, case study 1 resulted respectively in around 500 and 600 median required reviews for the two strategies, as presented in Figure 6.7a. However, Figure 6.8a shows that this number roughly doubles for case study 1. For case studies 2 and 3, the required amount of reviews increases even more than a factor of two. A possible cause is that the strategies test different levels to determine the most appropriate threshold level compared to when the false omission rate is used as a measure. For case 3, the path of threshold levels that are tested with the bisection strategy changes from testing levels 10, 15, 18 and then level 19 based on the false omission rate to testing levels 10, 5, 3, 2 and then level 1 based on the sensitivity.
Furthermore, the results in Figure 6.9b illustrate that in general, the strategies ended at a threshold level below the true threshold level with a median difference of 1 or 2 levels. Only for case study 3, where the true threshold level is 0 based on the sensitivity, did the strategies end 1 level too high.
The found threshold level with the bisection and multi-level with bisection strategies and true threshold level per run for each case study can be found in Appendix G.3

Conclusion

1. The most appropriate threshold depends on the chosen measure used to evaluate the suitability of a threshold level. It is therefore important to choose a measure that provides insight into the aspect that is considered important and a requirement that a threshold level should satisfy.
2. The required amount of reviews increases considerably to determine the sensitivity compared to the required amount of reviews to determine the false omission rate. This could be considered a drawback to using sensitivity, specificity or accuracy as a measure.

7

Conclusion and discussion

This chapter provides a summary of the threshold tuning strategies proposed in this thesis, followed by a discussion of the implications and limitations of these strategies. This thesis concludes with a section on topics for further research.

7.1 Conclusions

In this section, the research objectives stated in Chapter 1.1 will be answered and argued using results derived from this thesis. The first research objective to answer is:

How can a threshold tuning strategy be incorporated to accelerate the threshold tuning process of transaction monitoring models?

In this thesis, the performance of five threshold tuning strategies was investigated for initial threshold tuning to accelerate the threshold tuning process of transaction monitoring models. The main focus of this thesis was on BTL testing and the risk of missing interesting transactions below a threshold level. From the results in Chapter 6, it can be concluded that the bisection strategy requires the fewest reviews of transactions to determine the most appropriate threshold level. The multi-level with bisection strategy has a similar performance but requires slightly more reviews of transactions.

Furthermore, the bisection strategy results in a small difference between the true and found threshold level and has a stable performance. Another advantage of the bisection strategy is that the bisection strategy does not require tuning a parameter as is the case for the temperature parameter for the Boltzmann exploration strategy to improve the performance. However, the bisection strategy could end up in a local minimum when the proportion of interesting transactions fluctuates because this strategy does not start by testing the lowest logical level.

Moreover, the threshold tuning process can be accelerated by reusing information from a previously reviewed sample of another threshold level. This can reduce the number of required reviews in the tested case studies by more than 50% resulting in a great reduction in the workload for an analyst.

Lastly, it is important to note that the chosen measure affects the required amount of reviews. The false omission rate as chosen measure results in substantially fewer required reviews compared to the situation that the sensitivity is chosen as the measure.

In conclusion, there is a preference to use the bisection strategy for initial threshold tuning and reuse information from a previously reviewed sample of another threshold level to not waste valuable knowledge in the threshold tuning process.

The second research objective to answer is:

How can the uncertainty in the threshold tuning process for the chosen threshold level be quantified?

In this thesis, five confidence interval methods were evaluated with a simulation study when the main interest is only false and true negative transactions, determined via BTL testing, or only false and true positive transactions, determined via ATL testing. From the results in Section 3.2.4, it could be concluded that the Clopper-Pearson confidence interval method is the most appropriate since the Clopper-Pearson method always guarantees a coverage probability that is at least equal to the nominal coverage probability. In addition, the Clopper-Pearson confidence interval method has the great advantage that it can be applied regardless of the proportion of interesting transactions in the population.

It is possible to quantify the uncertainty in the sample proportion with the Clopper-Pearson confidence interval in combination with the complete population approach and make a statement with a certain confidence level such as; 'the proportion of interesting transactions in the population below threshold level L is with a $(1-\alpha)\times 100\%$ confidence level less or equal to p_{upper}'.

When interested in a variable expressed as a combination of the false and true negative and positive transactions via BTL and ATL testing, it is possible to quantify the uncertainty via bootstrapping in combination with the complete population approach. Using bootstrapping, a sampling distribution for the desired variable can be determined. Subsequently, Wald's confidence interval or the Bias corrected and accelerated (BCa) Percentile confidence interval method can be used to quantify the uncertainty in the variable and to make a statement with a certain confidence level such as; 'the sensitivity of threshold level L is with a $(1-\alpha)*100\%$ confidence level at least or equal to θ_{lower}'. Which method is appropriate for the threshold level that is evaluated, should be based on the determined sampling distribution by bootstrapping.

Lastly, the required sample size in the threshold tuning process depends on the preliminary population proportion and the desired confidence level for the confidence interval. The required sample size increases for a larger preliminary population proportion, with a maximum required sample size for a preliminary population proportion equal to 0.5. In addition, a larger sample size is required for a confidence interval with a higher confidence level or a smaller margin of error.

In conclusion, it is recommended to quantify uncertainty in the sample proportion using the Clopper-Pearson confidence interval method when one is only interested in the false and true negative transactions or only interested in the false and true positive transactions. It is recommended to quantify uncertainty with bootstrapping in combination with the Wald or Bias corrected and accelerated (BCa) Percentile confidence interval method when one is interested in a variable expressed as a combination of the false and true negative and positive transactions.

7.2 Discussion

This section reflects on the contribution of this thesis and discusses the obtained results. The bisection strategy is a suitable strategy for initial threshold tuning in various situations where it is unknown which threshold level is most suitable. The strategy quickly reduces the number of potentially suitable threshold levels by halving the number of options after each investigated threshold level. However, the research conducted in this thesis has its limitations which are discussed below.

Firstly, the performance of the tuning strategies has not been examined for the situation of periodic evaluation where, unlike with initial threshold tuning, some information is already known about the suitability of certain threshold levels. Periodic evaluation of an existing rule could be applicable when new data analysis has shown that the proportion of interesting transactions is higher than desired. In the case of periodic evaluation, the most appropriate threshold might be fairly close to the original threshold level which might enable the step strategy to determine the most appropriate threshold level with fewer reviews of transactions than the bisection strategy. If it is possible to use expert judgement to make a good estimate of the most appropriate threshold level, the step strategy possibly results in a

better performance compared to the bisection strategy. It is therefore always important to consciously consider which strategy is most suitable for a given situation.

In addition, the strategies have been tested on a limited number of distributions that may not be a representative reflection of the transaction data that banks face in reality. To set up the synthetic data sets, strong assumptions were made about possible criminal behaviour, but the plausibility of the assumptions was explored to a limited extent to keep the focus on threshold tuning strategies. It is important to emphasise that other choices of distributions or parameters in the distributions may lead to different results.

More research on the performance of the strategies on synthetic data with other distributions is necessary to investigate the performance of the strategies on a wide variety of possible transaction data. To be more certain of the performance of the tuning strategies, the strategies should be tested on real labelled transaction data and should discussed with experts in the field.

Moreover, the bisection strategy especially results in a good performance when dealing with a reasonable number of threshold levels. However, if it is decided to divide the transaction data only with 10 threshold levels, for example, the performance of the bisection strategy may become relatively worse compared to the other strategies.

Furthermore, in the thesis, a sample size with a margin of error of 0.05 was chosen which was 25% of the RTL in the experiments. However, the relatively large margin of error to the RTL can cause problems if the population proportion is close to the RTL which can result in a found threshold level further below or above the true threshold level. A margin of error that is relatively smaller than the RTL, which requires a larger sample size, mitigates this problem by allowing the sample proportion to be determined more accurately.

Additionally, the impact of the choice of the percentage of threshold levels which are explored with the multi-level strategy on the performance of this strategy is not investigated in this thesis. Since the multi-level also generally resulted in a good performance, it is interesting to investigate whether a higher percentage for threshold levels explored with the multi-level strategy could lead to better performance.

Lastly, relatively limited attention is paid to the confidence interval for bootstrapping in this thesis. However, a sampling distribution does not always result in a distribution that resembles the normal distribution. As an example, consider the sensitivity as a chosen measure which can take values between 0 and 1 and that the sensitivity of a certain threshold level is close to 1. This may result in a sampling distribution that does not resemble the normal distribution as shown in Figure G.3 in Appendix G.4. More research could be performed on the confidence intervals for bootstrapping and its applicability in the threshold tuning process.

7.3 Recommendations for future research

In this section, recommendations for future research are discussed.

Firstly, the threshold tuning framework used throughout this thesis was based on fixed threshold levels. All transactions were split using threshold levels that were a fixed distance apart. However, this could limit the threshold tuning process if, for example, the most appropriate threshold level is between two fixed threshold levels. A more continuous threshold tuning framework that allows flexibility in the distance between threshold levels might make it possible to end up closer to the optimal threshold value. In this case, continuous threshold levels might be considered instead of threshold levels discretised with a fixed distance between them. For the bisection strategy, this would mean that the next threshold level that should be tested is a value in the middle of the two values instead of in the middle of two fixed threshold levels. However, it is important to define the minimum difference between the values of two threshold levels in advance since the threshold tuning process can last a long time if the process is for example continued until the difference between two threshold levels is only 50 euros. In this case, it might

be a waste of time to test an additional threshold level for such a small difference. Further research about a more continuous threshold tuning framework would be an interesting continuation of this thesis.

In addition, the bisection strategy is now formulated to halve the potential threshold levels each time after evaluating a threshold level. However, this is not necessarily the smartest choice. Suppose that transactions are discretised with 20 threshold levels and level 10 is tested first for which an analyst determines that $p_{upper} = 0.22$. In this situation, it might make more sense to test a next threshold level that is about $\frac{RTL}{0.22} = \frac{0.2}{0.22} \approx 0.91$ times the value of the tested threshold level rather than a threshold level at about half the value of the tested threshold level which can be defined as 'adjusted bisection with weights' strategy. It would be interesting to investigate how much more the threshold tuning process can be accelerated using the 'adjusted bisection with weights' strategy.

Lastly, in the case studies, only one variable of interest was considered, or several variables could be summarised in the variable of interest. In practice, however, there are probably multiple variables that influence the suitability of a threshold level and cannot always be summarised in one variable. It would therefore be interesting to investigate and adapt this framework for a situation where multiple variables, which cannot be summarised in the variable of interest, influence the appropriateness of a threshold level.

References

[1] Alan Agresti and Brent A Coull. “Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions”. In: *The American Statistician* 52.2 (1998), pp. 119–126. URL: https://users.stat.ufl.edu/~aa/articles/agresti_coull_1998.pdf.

[2] F J Anscombe. “The Transformation of Poisson, Binomial and Negative-Binomial Data”. In: *Biometrika* 35 (1948), pp. 246–254. URL: https://academic.oup.com/biomet/article/35/3-4/246/280278.

[3] Andrew H. Briggs, David E. Wonderling, and Christopher Z. Mooney. “Pulling cost-effectiveness analysis up by its bootstraps: A non-parametric approach to confidence interval estimation”. In: *Health Economics*. Vol. 6. 4. July 1997, pp. 327–340. DOI: 10.1002/(SICI)1099-1050(199707)6:4<327::AID-HEC282>3.0.CO;2-W.

[4] Lawrence D Brown, T Tony Cai, and Anirban Dasgupta. “Interval Estimation for a Binomial Proportion”. In: *Source: Statistical Science* 16.2 (2001), pp. 101–117. URL: http://www-stat.wharton.upenn.edu/~lbrown/Papers/2001a%20Interval%20estimation%20for%20a%20binomial%20proportion%20(with%20T.%20T.%20Cai%20and%20A.%20DasGupta).pdf.

[5] Nicolò Cesa-Bianchi et al. “Boltzmann Exploration Done Right”. In: *Neural Information Processing Systems* (May 2017). URL: http://arxiv.org/abs/1705.10257.

[6] Azra Pravdic Daniel Mikkelsen and Bryan Richardson. “Flushing out the money launderers with better customer risk- rating models”. In: *McKinsey & Company, Risk Practice* September (2019). URL: https://www.mckinsey.com/business-functions/risk-and-resilience/our-insights/flushing-out-the-money-launderers-with-better-customer-risk-rating-models.

[7] David Zacks. *The Future of Anti-Money Laundering Compliance*. 2018. URL: https://www.acamstoday.org/the-future-of-anti-money-laundering-compliance/.

[8] De Nederlandse Bank (DNB). *De integriteitrisicoanalyse*. Tech. rep. 2015. URL: https://www.dnb.nl/media/nu5pqb1a/integriteitrisicoanalyse.pdf.

[9] De Nederlandse Bank (DNB). *Leidraad Wwft en Sw*. Tech. rep. 2020. URL: https://www.dnb.nl/media/dzicty20/leidraad-wwft-en-sanctiewet.pdf.

[10] De Nederlandse Bank (DNB). *Post-event transactie-monitorings proces bij banken - Guidance*. Tech. rep. 2017. URL: https://www.dnb.nl/media/0eilgixn/guidance-document-transactiemonitoring-banken.pdf.

[11] Dutch government. *Wet ter voorkoming van witwassen en financieren van terrorisme*. 1977. URL: https://wetten.overheid.nl/BWBR0024282/2022-10-01#Hoofdstuk3.

[12] Habib Ahmed Elsayir. “Comparison of Precision of Systematic Sampling with Some other Probability Samplings”. In: *American Journal of Theoretical and Applied Statistics* 3.4 (2014), p. 111. ISSN: 2326-8999. DOI: 10.11648/j.ajtas.20140304.16.

[13] European Commission. *Preventing money laundering and terrorist financing across the EU*. 2018. URL: https://ec.europa.eu/info/sites/default/files/diagram_aml_2018.07_ok.pdf.

[14] Financial Crimes Enforcement Network. *The Bank Secrecy Act*. 1970. URL: https://www.fincen.gov/resources/statutes-and-regulations/bank-secrecy-act.

[15] Gavin Finch. *Worlds biggest banks fined $321 billion since financial crisis*. 2017. URL: https://www.bloomberg.com/professional/blog/worlds-biggest-banks-fined-321-billion-since-financial-crisis/.

[16] Luzia Gonçalves et al. “Sample size for estimating a binomial proportion: Comparison of different methods”. In: *Journal of Applied Statistics* 39.11 (2012), pp. 2453–2473. ISSN: 02664763. DOI: 10.1080/02664763.2012.713919.

[17] Ryan Eichler James Chen Somer Anderson. *Money Laundering: What is is and how to prevent it*. June 2022.

[18] John A. Rice. *Mathematical Statistics and Data Analysis*. Third edition. Belmont: Brooks/Cole, Cengage Learning, 2007, pp. 217–220. ISBN: ISBN-13: 978-0-495-11868-8.

[19] Martin Jullum, Anders Løland, and Ragnar Bang Huseby. "Detecting money laundering transactions with machine learning". In: 23.1 (2020), pp. 173–186. DOI: 10.1108/JMLC-07-2019-0055.

[20] Ramachandran Kandethody M and Chris P. Tsokos. *Mathematical Statistics with Applications in R*. Second Edition. Elsevier Inc., 2015, p. 181. URL: https://www-sciencedirect-com.tudelft.idm.oclc.org/book/9780124171138/mathematical-statistics-with-applications-in-r.

[21] Volodymyr Kuleshov and Doina Precup. "Algorithms for multi-armed bandit problems". In: *Journal of Machine Learning Research* (Feb. 2014). URL: http://arxiv.org/abs/1402.6028.

[22] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge Univiersity Press, 2020, pp. 10–56. URL: https://tor-lattimore.com/downloads/book/book.pdf?msclkid=5444db1ebbf411ec86a8fd015c94529c.

[23] Office of the Comptroller of the Currency. *Safety and Soundness - Model Risk Management*. Tech. rep. 2021. URL: https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/pub-ch-model-risk.pdf.

[24] Ana M Pires and Conceição Amado. "Interval estimators for a binomial proportion: Comparison of twenty methods". In: *Statistical Journal* 6.2 (2008), pp. 165–197. URL: https://www.researchgate.net/publication/237465854.

[25] Friedrich Schneider and Ursula Windischbauer. "Money laundering: Some facts". In: *European Journal of Law and Economics* 26.3 (Dec. 2008), pp. 387–404. ISSN: 09291261. DOI: 10.1007/s10657-008-9070-x. URL: https://link.springer.com/content/pdf/10.1007/s10657-008-9070-x.pdf.

[26] Fritz Scholz. *Confidence Bounds & Intervals for Parameters Relating to the Binomial, Negative Binomial, Poisson and Hypergeometric Distributions With Applications to Rare Events*. Tech. rep. University of Washington, 2019, pp. 4–4. URL: http://faculty.washington.edu/fscholz/Stat498B2008.html..

[27] Gaganpreet Sharma. "Pros and cons of different sampling techniques". In: *International Journal of Applied Research* 3.7 (2017), pp. 749–752. ISSN: 2394-5869. URL: https://www.allresearchjournal.com/archives/2017/vol3issue7/PartK/3-7-69-542.pdf.

[28] Julius Sim and Norma Reid. "Statistical Inference by Confidence Intervals: Issues of Interpretation and Utilization Professional Perspective". In: *Physical Therapy* 79.2 (1999). URL: https://academic.oup.com/ptj/article/79/2/186/2837119.

[29] Mikael Hagstroem Stuart Breslow and Daniel Mikkelsen. "The new frontier in anti-money laundering". In: *McKinsey & Company* (2017). URL: https://www.mckinsey.com/business-functions/risk-and-resilience/our-insights/the-new-frontier-in-anti-money-laundering.

[30] Brigitte Unger Joras Ferwerda Ian Koetsier Bojken Gjoleka and Alexander van Saase Brigitte. *Aard en omvang van criminele bestedingen Opdrachtgever: WODC*. Tech. rep. Wetenschappelijk Onderzoek- en Documentatiecentrum, 2018. URL: https://repository.wodc.nl/bitstream/handle/20.500.12832/2319/2790_Volledige_Tekst_tcm28-355586.pdf?sequence=2&isAllowed=y.

[31] Unknown. *Natural Gas*. 2022. URL: https://tradingeconomics.com/commodity/natural-gas.

[32] unknown. *Case studies*. 2022. URL: https://kyc360.riskscreen.com/case-study/.

[33] unknown. *How does the Black Market Peso Exchange work?* 2016. URL: https://kyc360.riskscreen.com/case-study/the-black-market-peso-exchange/.

[34] unknown. *Jurisdictions under Increased Monitoring - June 2022*. June 2022. URL: https://www.fatf-gafi.org/publications/high-risk-and-other-monitored-jurisdictions/documents/increased-monitoring-june-2022.html.

[35] unknown. *Money laundering through charities and NPOs*. 2016. URL: https://kyc360.riskscreen.com/case-study/abuse-of-charities-and-npos/.

[36] unknown. *Money laundering typologies*. 2022. URL: https://www.fiu-nederland.nl/en/general-legislation/money-laundering-typologies.

[37] unknown. *Money laundering with derivatives*. 2016. URL: https://kyc360.riskscreen.com/case-study/money-laundering-with-derivatives/.

[38] unknown. *Relevant cases*. 2022. URL: https://www.fiu-nederland.nl/en/legislation/relevant-cases.

[39] Adriana Vallejo et al. "New method to estimate the sample size for calculation of a proportion assuming binomial distribution". In: *Research in Veterinary Science* 95.2 (Oct. 2013), pp. 405–409. ISSN: 00345288. DOI: 10.1016/j.rvsc.2013.04.005.

[40] Joannès Vermorel and Mehryar Mohri. *Multi-Armed Bandit Algorithms and Empirical Evaluation*. Tech. rep. 2005, pp. 437–448. URL: https://cs.nyu.edu/~mohri/pub/bandit.pdf.

[41] Sean Wallis. *Accurate confidence intervals on Binomial proportions, functions of proportions, algebraic formulae and effect sizes*. Tech. rep. 2022. URL: https://www.ucl.ac.uk/english-usage//staff/sean/resources/confidence-intervals.pdf.

[42] Timothy Li Will Kenton Somer Anderson. *Anti Money Laundering (AML) Definition: Its history and how it works*. May 22.

A

Terminology and derivations from a confusion matrix

Definitions

- True positive (TP): A test result that correctly indicates the presence of a condition or characteristic.
- True negative (TN): A test result that correctly indicates the absence of a condition or characteristic.
- Talse positive (FP): A test result which wrongly indicates that a particular condition or attribute is present.
- Talse negative (FN): A test result which wrongly indicates that a particular condition or attribute is absent

Terminology

$$TPR = \frac{TP}{TP + FN} = 1 - FNR \quad \text{Sensitivity or True Positive Rate} \quad \text{(A.1)}$$

$$TNR = \frac{TN}{TN + FP} = 1 - FPR \quad \text{Specificity or True Negative Rate} \quad \text{(A.2)}$$

$$PPV = \frac{TP}{TP + FP} \quad \text{Positive Predictive Value} \quad \text{(A.3)}$$

$$FDR = \frac{FP}{TP + FP} = 1 - PPV \quad \text{False Discovery Rate (FDR)} \quad \text{(A.4)}$$

$$NPV = \frac{TN}{TN + FN} \quad \text{Negative Predictive Value} \quad \text{(A.5)}$$

$$FOR = \frac{FN}{TN + FN} = 1 - NPV \quad \text{False Omission rate} \quad \text{(A.6)}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Accuracy} \quad \text{(A.7)}$$

$$FNR = \frac{FN}{TN + TP} = 1 - TPR \quad \text{False Negative Rate} \quad \text{(A.8)}$$

$$FPR = \frac{FP}{FP + TN} = 1 - TNR \quad \text{False Positive Rate} \quad \text{(A.9)}$$

B

Sample sizes

Clopper-Pearson sample size
For the Clopper-Pearson CI, the three cases result in the following systems of equations:
For $x = 0$

$$\begin{cases} p_{upper} \leq 1 - (\alpha/2)^{1/n} \\ p_{lower} \geq 0 \\ p_{upper} - p_{lower} < 2e \end{cases} \tag{B.1}$$

For $0 < x < n$:

$$\begin{cases} p_{upper} \leq Beta_{1-\alpha/2}(x+1, n-x) \\ p_{lower} \geq Beta_{\alpha/2}(x, n-x+1) \\ p_{upper} - p_{lower} < 2e \end{cases} \tag{B.2}$$

For $x = n$:

$$\begin{cases} p_{upper} \leq 1 \\ p_{lower} \geq (\alpha/2)^{1/n} \\ p_{upper} - p_{lower} < 2e \end{cases} \tag{B.3}$$

The sample size that follows from solving the system of equations will be rounded up to the nearest integer.

Arcsine sample size

For the Arcsine CI, the three cases result in the following systems of equations:
For $x = 0$

$$\begin{cases} p_{upper} \leq sin^2\left(min\left(arcsin(\sqrt{p_0}) + \frac{z_{1-\alpha/2}}{2\sqrt{n}}; \frac{\pi}{2}\right)\right) \\ p_{lower} \geq 0 \\ p_{upper} - p_{lower} < 2e \end{cases} \tag{B.4}$$

For $0 < x < n$:

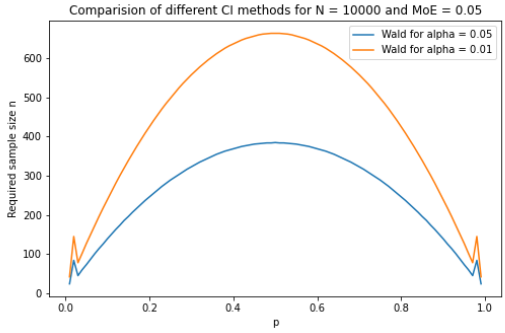
$$\begin{cases} p_{upper} \leq sin^2\left(min\left(arcsin(\sqrt{p_0}) + \frac{z_{1-\alpha/2}}{2\sqrt{n}}; \frac{\pi}{2}\right)\right) \\ p_{lower} \geq sin^2\left(max\left(arcsin(\sqrt{p_0}) - \frac{z_{1-\alpha/2}}{2\sqrt{n}}; 0\right)\right) \\ p_{upper} - p_{lower} < 2e \end{cases} \tag{B.5}$$

For $x = n$:

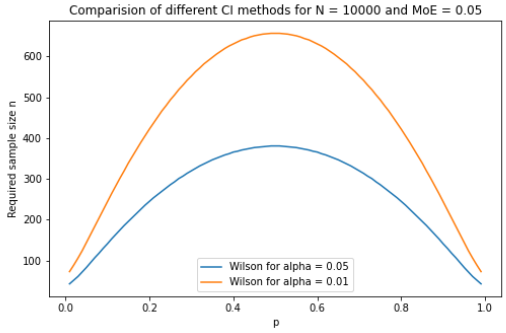
$$\begin{cases} p_{upper} \leq 1 \\ p_{lower} \geq sin^2\left(max\left(arcsin(\sqrt{p_0}) - \frac{z_{1-\alpha/2}}{2\sqrt{n}}; 0\right)\right) \\ p_{upper} - p_{lower} < 2e \end{cases} \tag{B.6}$$

The sample size that follows from solving the system of equations will be rounded up to the nearest integer.

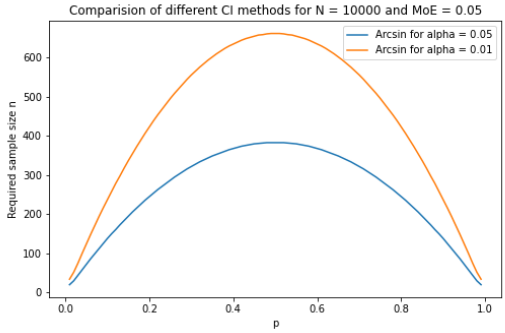
B.1 Sample size figures



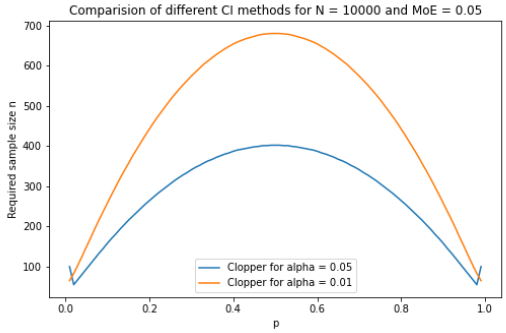
(a) The required sample size using Wald's CI.



(b) The required sample size using Wilson Score CI.

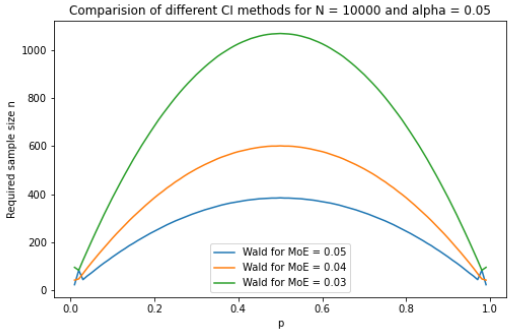


(c) The required sample size using Arcsine CI.

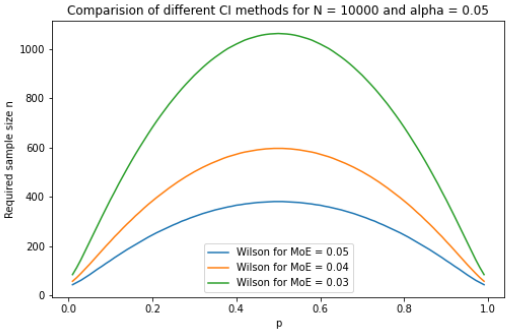


(d) The required sample size using Clopper-Pearson CI.

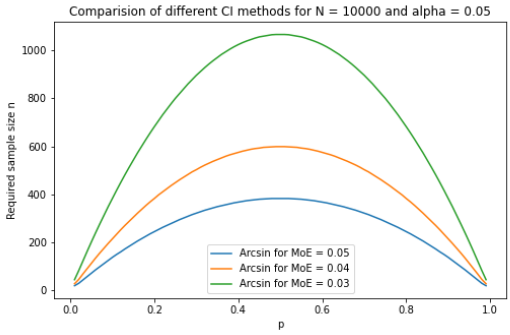
Figure B.1: A comparison of the required sample size with different CI methods for a 95% or 99% confidence interval with a margin of error $e = 0.05$



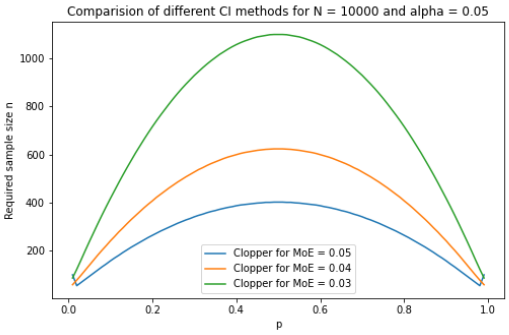
(a) The required sample size using Wald's CI.



(b) The required sample size using Wilson Score CI.



(c) The required sample size using Arcsine CI.



(d) The required sample size using Clopper-Pearson CI.

Figure B.2: A comparison of the required sample size with different CI methods for a margin of error e equal to 0.05, 0.04 and 0.03 with $\alpha = 0.05$

C

Proof of Clopper-Pearson CI limits

A proof that the limits of the Clopper-Pearson confidence interval are equal to the quantiles of the Beta distribution, i.e $P(X \geq k) = I_p(k, n-k+1)$ and $P(X \leq k) = 1 - I_p(k+1, n-k)$, is described in the report of Scholz [26] in which two facts are proven:

1. Let $x(p) = P(X \geq k)$ and $y(p) = I_p(k, n-k+1)$. Firstly, it is proven that:

$$x'(p) = \frac{\partial P(X \geq k)}{\partial p} = \frac{\partial I_p(k, n-k+1)}{\partial p} = y'(p) \; \forall p \geq 0. \tag{C.1}$$

2. In addition, it is proven that:

$$x(p) = P(X \geq k) = I_p(k, n-k+1) = y(p) \text{ for } \; p = 0. \tag{C.2}$$

3. From 1 and 2, it can be concluded that $P(X \geq k) = I_p(k, n-k+1)$ for all values of $p \geq 0$ which proves the relation in equation 3.11.

The conclusion in point 3 follows from the fact that $x'(p) = y'(p) = f(p)$ which results in $x(p) = y(p) + C$. Furthermore, it is proven that $x(0) = y(0)$ from which follows that $C = 0$. Therefore, it can be concluded that $x(p) = y(p) \; \forall p \geq 0$ since the functions are continuous which implies uniqueness of the solution.

Proof
To prove equation C.1, suppose X is a binomial random variable X. Then:

$$P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \tag{C.3}$$

and $P(X \geq k) = 1 - P(X \leq k-1)$. It can be proven that $P(X \leq k)$ is a strictly decreasing in p for $k = 0, 1, ..., n-1$ and that $P(X \geq k)$ is strictly increasing in p for $k = 0, 2, ..., n$. Using the identities $i\binom{n}{i} = n\binom{n-1}{i-1}$ and $(n-i)\binom{n}{i} = n\binom{n-1}{i}$ in equation C.6 and taking the derivative of $P(X \geq k)$ with respect to p gives as result:

$$x'(p) = \frac{\partial P(X \geq k)}{\partial p} = \sum_{i=k}^{n} \binom{n}{i} i p^{i-1}(1-p)^{n-i} - \sum_{i=k}^{n} \binom{n}{i}(n-i)p^{i}(1-p)^{n-i-1} \tag{C.4}$$

$$= \sum_{i=k}^{n} \binom{n}{i} i p^{i-1}(1-p)^{n-i} - \sum_{i=k}^{n-1} \binom{n}{i}(n-i)p^{i}(1-p)^{n-i-1} + 0 \tag{C.5}$$

$$= n\left[\sum_{i=k}^{n} \binom{n-1}{i-1} i p^{i-1}(1-p)^{n-i} - \sum_{i=k}^{n-1} \binom{n-1}{i} p^{i}(1-p)^{n-i-1}\right] \tag{C.6}$$

$$= n\left[\binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} - \binom{n-1}{k} p^{k}(1-p)^{n-k-1}\right. \tag{C.7}$$

$$+ \binom{n-1}{k} p^{k}(1-p)^{n-k-1} - \binom{n-1}{k+1} p^{k+1}(1-p)^{n-k-2} \tag{C.8}$$

$$+ \ldots - \ldots \tag{C.9}$$

$$+ \binom{n-1}{n-2} p^{n-2}(1-p) - \binom{n-1}{n-1} p^{n-1} \tag{C.10}$$

$$\left. + \binom{n-1}{n-1} p^{n-1}\right] \tag{C.11}$$

$$= n\binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} = k\binom{n}{k} p^{k-1}(1-p)^{n-k} > 0 \tag{C.12}$$

where the blue terms cancel each other out. In addition, by the Fundamental Theorem of Calculus, the derivative of:

$$I_p(k, n-k+1) = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} \int_0^p t^{k-1}(1-t)^{n-k} dt = k\binom{n}{k} \int_0^p t^{k-1}(1-t)^{n-k} dt \tag{C.13}$$

with respect to p is equal to:

$$k\binom{n}{k} p^{k-1}(1-p)^{n-k} \tag{C.14}$$

which is equal to the derivative of $P(X \geq k)$ with respect to p. This proves equation C.1. Furthermore, $P(X \geq k) = 0$ and $I_p(k, n-k+1) = 0$ for $p = 0$ which proves equation C.2. Therefore, it can be concluded that the relation $P(X \geq k) = I_p(k, n-k+1)$ holds. By complement, it follows for $k < n$ that [26]:

$$P(X \leq k) = 1 - P(X \geq k+1) = 1 - I_p(k+1, n-k) \tag{C.15}$$

This proves that the upper and lower bound for the proportion p for the Clopper-Pearson CI are equal to the quantiles of the Beta distribution as described in equation 3.12 □.

D

Tuning strategies

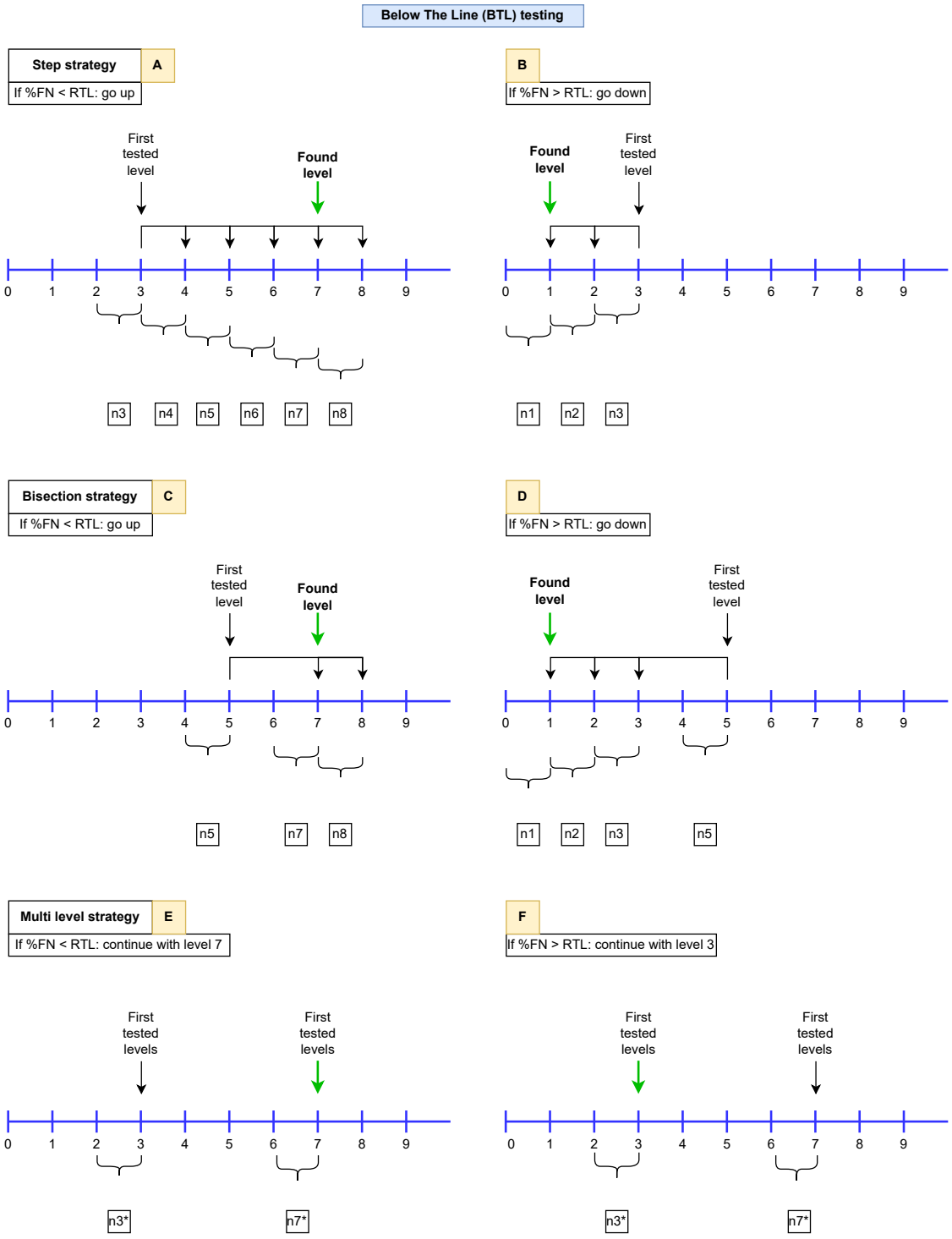


Figure D.1: Illustration of the step, bisection and multi level strategy in combination with the bucket approach for BTL tuning.

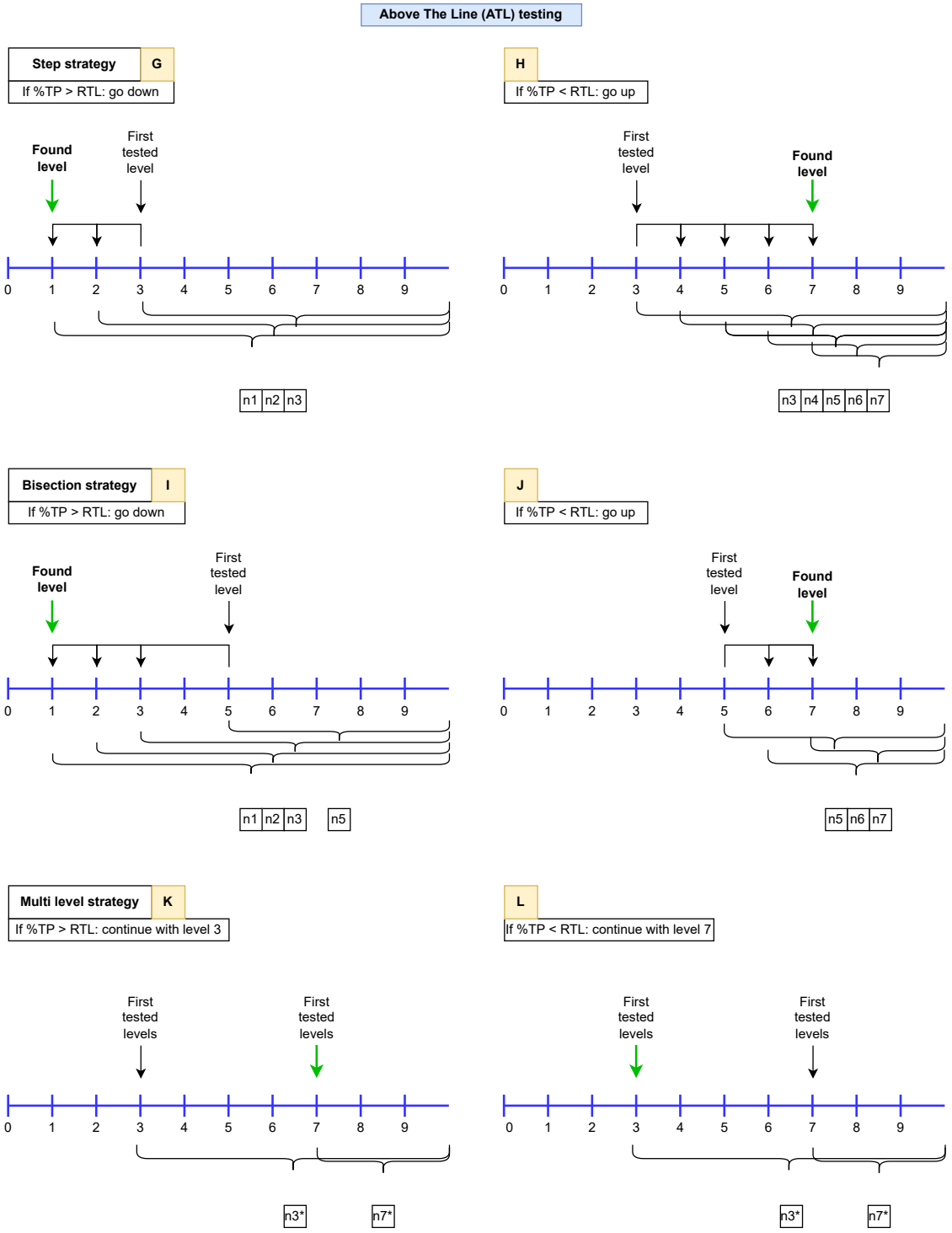


Figure D.2: Illustration of the step, bisection and multi level strategy in combination with the complete population approach for ATL tuning.

E

Strategy algorithms

E.1 Step algorithm

The review process in which the proportion of interesting transactions in the sample is determined, is described in algorithmic form in Algorithm 1. In addition, upper and lower confidence interval bounds for the proportion of interesting transactions are also determined in the review process as well as the amount of reviewed transactions by the analyst.

Algorithm 1 Review Process

Require: $n, \alpha, reviews, S$

$s = 0$ ▷ sum of interesting transactions
for j in range$(0, n)$ **do** ▷ review process by analyst
    $s+ = s + S[j]$ ▷ $S[j]$=0 or 1, if the transaction in sample S is interesting or not
    $reviews+ = 1$ ▷ amount of reviewed transactions
end for
$p = s/n$ ▷ proportion of interesting transactions
$p_{low}, p_{high} = ClopperPearson(...)$ ▷ function returning Clopper Pearson CI limits
return $p_{low}, p_{high}, reviews$

Algorithm 2 Step

Require: $RTL, n, \alpha, reviews, S, p_{high}$

$S_i \leftarrow S[i]$ ▷ sample of level i
$p_{low}, p_{high}, reviews = ReviewProcess(i, ...)$ ▷ simulation of review process by analyst
$p_{high}[i] \leftarrow p_{high}$
$L_{checked} += 1$ ▷ counts number of checked levels
if $p_{high} \geq RTL$ **then** ▷ if $p_{high} \geq RTL$, go lower
    if $p_{high}[i-1] < RTL$ **then** ▷ stop condition
        $L_{final} = i - 1$
    else
        $L_{next} = i - 1$
        $L_{final}, reviews = Step(L_{next}, ...)$
    end if
else ▷ else go higher
    if $p_{high}[i+1] \geq RTL$ **then** ▷ stop condition
        $L_{final} = i$
    else
        $L_{next} = i + 1$
        $L_{final}, reviews = Step(L_{next}, ...)$
    end if
end if
return $L_{final}, L_{checked}, reviews$

E.2 Bisection algorithm

Algorithm 3 Bisection

Require: $RTL, n, \alpha, reviews, S, p_{high}, L_{min}, L_{max}$

$S_i \leftarrow S[i]$ ▷ sample of level i
$p_{low}, p_{high}, reviews = ReviewProcess(i, ...)$ ▷ simulation of review process by analyst
$p_{high}[i] \leftarrow p_{high}$
$L_{checked} += 1$ ▷ counts number of checked levels
if $p_{high} \geq RTL$ **then** ▷ if $p_{high} \geq RTL$, go lower
  if $p_{high}[i-1] < RTL$ **then** ▷ stop condition
    $L_{final} = i - 1$
  else
    $L_{next} = ceil(i - \frac{i - L_{min}}{2})$
    $L_{max} = i$
    $L_{final}, reviews = Bisection(L_{next}, ...)$
  end if
else ▷ else go higher
  if $p_{high}[i+1] \geq RTL$ **then** ▷ stop condition
    $L_{final} = i$
  else
    $L_{next} = ceil(i + \frac{L_{max} - i}{2})$
    $L_{min} = i$
    $L_{final}, reviews = Bisection(L_{next}, ...)$
  end if
end if
return $L_{final}, L_{checked}, reviews$

E.3 Multi-level algorithm

Algorithm 4 Multi-level

Require: $RTL, n, \alpha, reviews, S, p_{high}, L_{min}, L_{max}$

1: $n^* = ceil(n/(K * x))$ ▷ amount of transactions that will be explored
2: **for** j in range$(0, K * x)$ **do** ▷ EXPLORATION PHASE
3: $i = ceil(K * \frac{j+1}{K*x+1})$ ▷ explore level i
4: $L[j] \leftarrow i$
5: $p_{low}, p_{high}, reviews = ReviewProcess(i, n^*, ...)$
6: $p_{high}[i] \leftarrow p_{high}$
7: **end for**
8: $L_{cont} = L[j]$ ▷ continue with highest level for which $p_{high}[i] < RTL$
9: ▷ EXPLOITATION PHASE
10: $p_{low}, p_{high}, reviews = ReviewProcess2(L_{cont}, n^*, ...)$ ▷ Small adjustment to Review Process function
11: **if** $p_{high} \geq RTL$ **then**
12: $L_{next} = ...$ ▷ Determined with chosen strategy to continue with
13: **else**
14: $L_{next} = ...$ ▷ Determined with chosen strategy to continue with
15: **end if**
16: $L_{final}, reviews = Strategy(L_{next}, ...)$ ▷ Continue with chosen strategy Step or Bisection
17: **return** $L_{final}, L_{checked}, reviews$

where x the percentage of threshold levels which will be explored and n^* the amount of transactions in the sample that will be explored. Function 'Review Process2' is the same as the function 'Review Process' except that it continues reviewing transactions of the most promising threshold level, changing line 2 in Algorithm 1 to 'for j in range(n^*, n)' and not starting with $s = 0$.

E.4 Boltzmann Exploration algorithm

Algorithm 5 Boltzmann Exploration

Require: $RTL, n, \alpha, reviews, S, p_{high}, t$

$n^* = ceil(n/50)$ ▷ amount of transactions that will be explored
for i in range$(0, K)$ **do** ▷ GET INITIAL AVG REWARDS, review n^* transactions for each level
    $\hat{\mu}_{0,i} = AverageReward(i, ...)$ ▷ function returns average reward for level i
end for
for i in range$(0, K)$ **do** ▷ INITIAL PICK PROBABILITY
    $p_{select,i} = (e^{\hat{\mu}_{0,i}/\tau_0})/(sum(e^{\hat{\mu}_0/\tau_0}))$
end for
for t in range$(K * n^*, K * n)$ **do** ▷ $K * n$ is max amount of possible required reviews
    $i = int(random(L, p_{select}[1:K], 1))$ ▷ select level i to review one transaction
    $n_{t,i} += 1$
    $\tau_t = taufunction(t, \tau_0)$ ▷ tau function
    $\hat{\mu}_{t,i} = AverageReward(i, ...)$ ▷ update average reward for level i
    ▷ STOP CONDITION
    if $n_{t,i} == n$ **then** ▷ all transactions in sample S_i are reviewed
        $L_{checked} += 1$
        if $p_{high}[i] \geq RTL$ **then**
            $p_{high}[i-1] = ReviewProcess2(i-1, n_{t,i-1}, ...)$ ▷ Determine p_{high} for one level lower
            $L_{checked} += 1$
            if $p_{high}[i-1] < RTL$ **then**
                $L_{final} = i - 1$
                $break$
            else
                for i in range$(0, K)$ **do** ▷ UPDATE PICK PROBABILITY AND CONTINUE
                    $p_{select,i} = (e^{\hat{\mu}_{t,i}/\tau_t})/(sum(e^{\hat{\mu}_t/\tau_t}))$
                end for
            end if
        else
            $p_{high}[i+1] = ReviewProcess2(i+1, n_{t,i+1}, ...)$ ▷ Determine p_{high} for one level higher
            $L_{checked} += 1$
            if $p_{high}[i+1] \geq RTL$ **then**
                $L_{final} = i$
                $break$
            else
                for i in range$(0, K)$ **do** ▷ UPDATE PICK PROBABILITY AND CONTINUE
                    $p_{select,i} = (e^{\hat{\mu}_{t,i}/\tau_t})/(sum(e^{\hat{\mu}_t/\tau_t}))$
                end for
            end if
        end if
    end if
end for
$reviews = t$
return $L_{final}, L_{checked}, reviews$

E.5 Epsilon-Greedy strategy

ϵ-Greedy Strategy (Semi-Uniform method)
The ϵ-Greedy algorithm chooses in each round with probability ϵ a random threshold level (exploration) and with probability $1-\epsilon$ the threshold level that has the highest average reward (exploitation). Therefore, given the initial empirical means $\hat{\mu}_1(0)$,...,$\hat{\mu}_K(0)$ [21]:

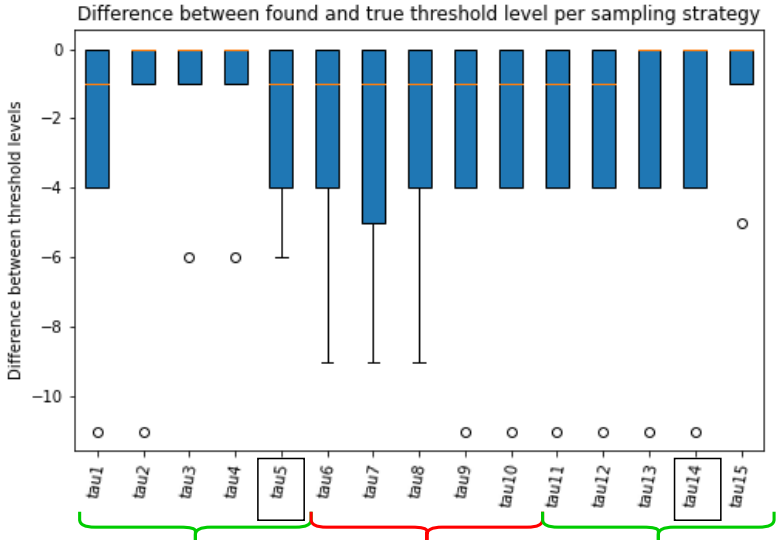
$$p_{select,i}(t+1) = \begin{cases} 1-\epsilon+\frac{\epsilon}{k} & if\ i = argmax_{j=1,...,K}\hat{\mu}_j(t) \\ \frac{\epsilon}{k} & otherwise \end{cases}$$

This strategy purely exploits the threshold level with the highest average reward with $\epsilon = 0$. As epsilon increases, exploration increases resulting in the optimal threshold level being chosen with a higher frequency. As epsilon increases, however, a tipping point is reached where there is more exploration and less exploitation until $\epsilon = 1$. Many variations on the epsilon greedy algorithm have been devised by such as the ϵ-decreasing strategy in which ϵ decreases (less exploration) over time instead of keeping ϵ constant. The main disadvantage of this strategy is that non-optimal threshold levels are still chosen, even after it is identified that these threshold levels are not optimal.

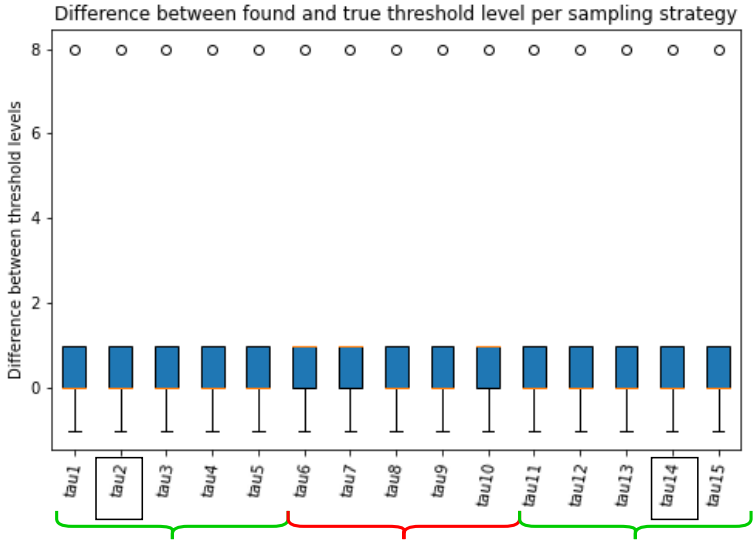
F

Tuning results of the temperature parameter

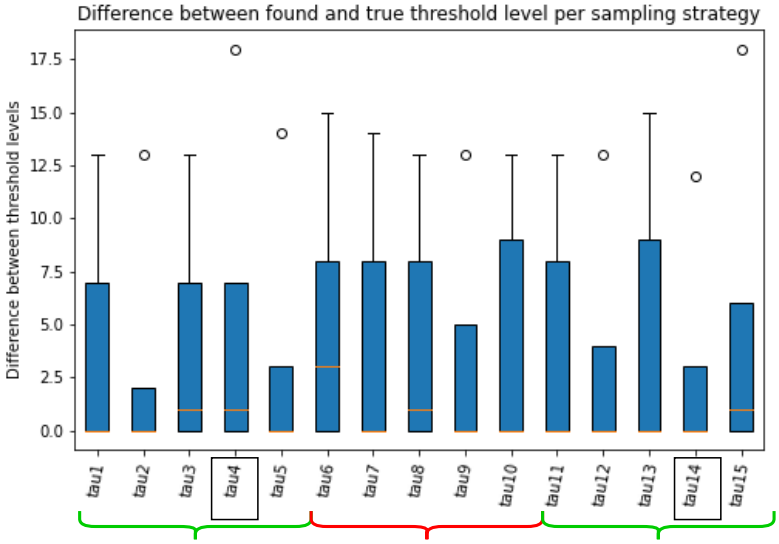
F.1 Difference between found and true threshold level



(a) For case study 1.



(b) For case study 2.



(c) For case study 3.

Figure F.1: For BTL testing, the difference between the found and true threshold level for a decreasing or constant schedule for different $\tau(0)$ values for $K = 20$ threshold levels.

G

Experiment results

G.1 Stability moment

In Figures G.1 and G.2, it can be clearly seen that the maximum, the median and the minimum amount of reviews and difference between found and true threshold level stabilise and do not change considerable after 25 runs for case study 1. The same conclusion could be drawn for case study 2 and 3.

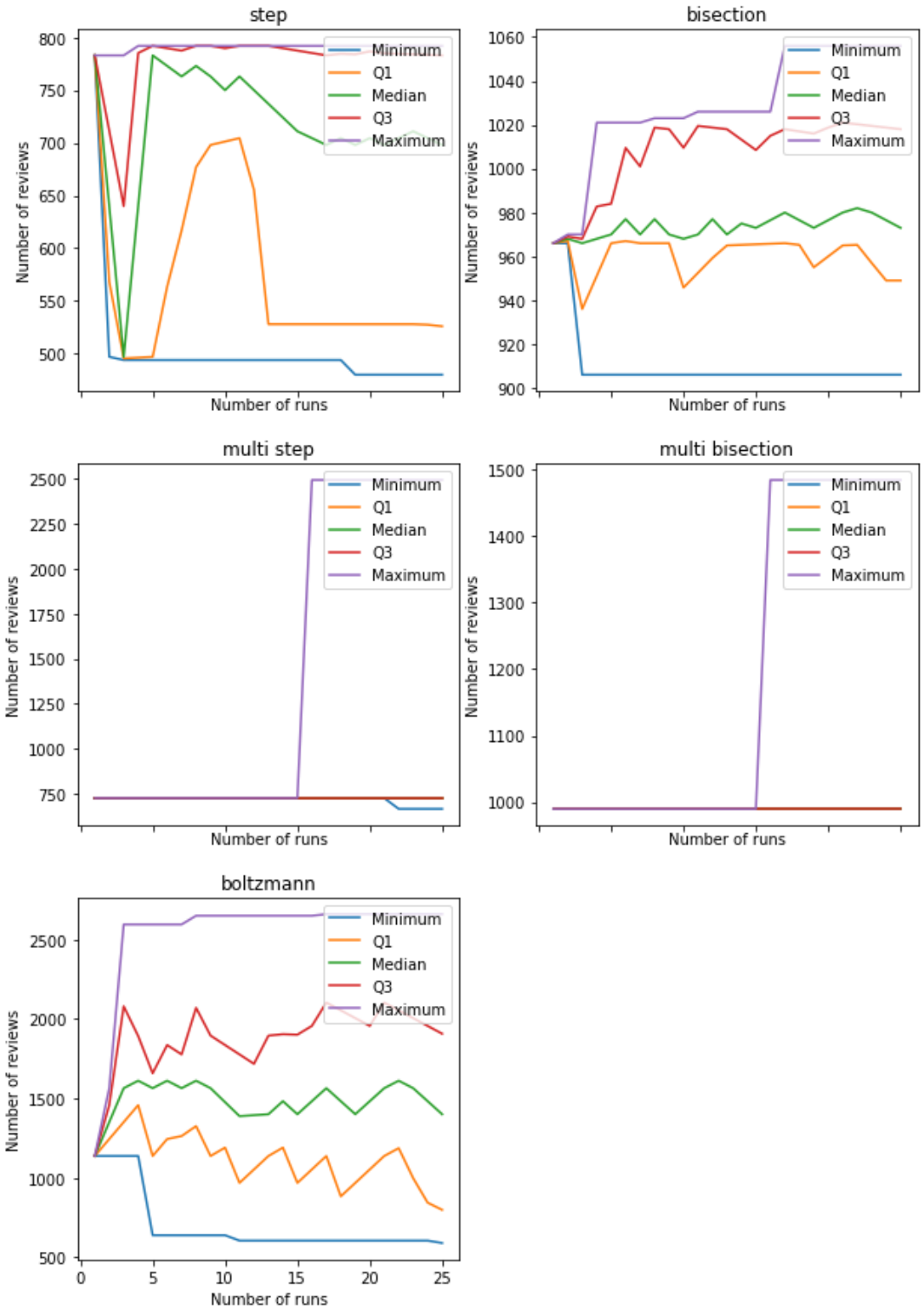


Figure G.1: For case study 1, the maximum, 75% quantile, median, 25% quantile and minimum amount of reviews required for different tuning strategies over 25 runs.

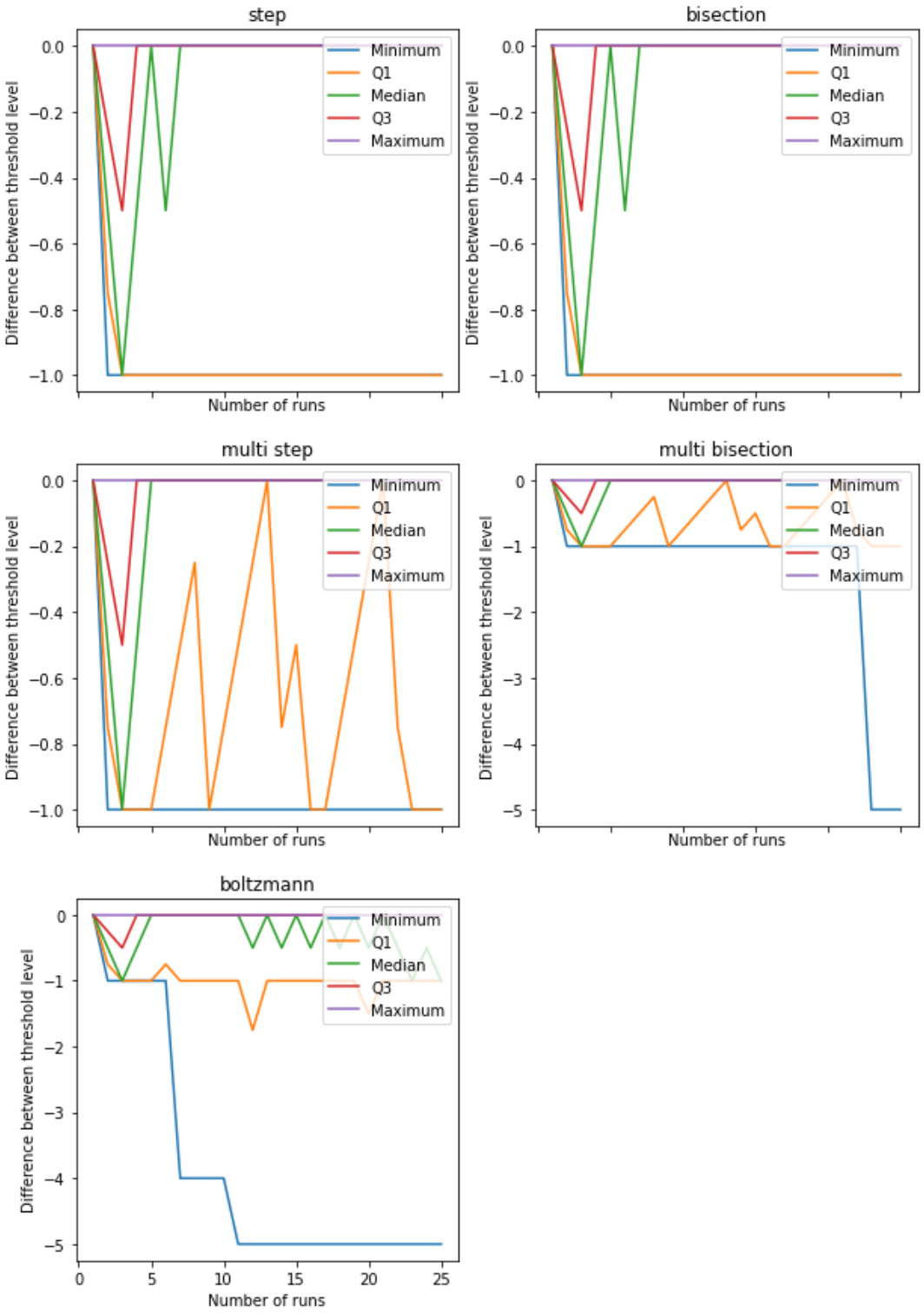


Figure G.2: For case study 1, the maximum, 75% quantile, median, 25% quantile and minimum difference between the found and true threshold level for different tuning strategies over 25 runs.

G.2 Found and true threshold for each case based on the false omission rate

The darker orange the colour of a cell, the further is the found threshold level above the true threshold level. The darker green the colour of a cell, the further is the found threshold level below the true threshold level.

Run	Step	Bisectiion	Multi step	Multi bisection	Boltzmann	TRUE
1	3	3	3	3	7	3
2	3	3	3	3	3	2
3	3	3	3	3	3	2
4	3	3	3	3	3	3
5	3	3	3	3	3	3
6	4	4	3	3	3	3
7	3	3	3	3	7	3
8	3	3	3	3	3	3
9	3	3	3	3	3	2
10	3	3	3	3	3	3
11	3	3	3	3	8	3
12	3	3	3	3	3	3
13	3	3	3	3	3	3
14	3	3	3	3	3	2
15	3	3	3	3	3	3
16	4	4	4	4	4	3
17	3	3	3	3	3	3
18	3	3	3	3	3	3
19	3	3	3	3	3	3
20	3	3	3	3	3	3
21	3	3	3	3	3	3
22	4	4	4	4	4	3
23	4	4	4	8	4	3
24	3	3	3	3	3	3
25	3	3	3	3	3	2

Table G.1: The found threshold with different strategies and the true threshold per run for case study 1.

The darker orange the colour of a cell, the further is the found threshold level above the true threshold level. The darker green the colour of a cell, the further is the found threshold level below the true threshold level.

Run	Step	Bisection	Multi step	Multi bisection	Boltzmann	TRUE
1	11	11	11	11	11	12
2	10	10	10	10	10	12
3	11	11	11	11	11	12
4	11	11	11	11	11	12
5	10	10	10	10	10	11
6	12	12	12	12	12	19
7	11	11	11	11	11	11
8	11	11	11	11	11	11
9	11	11	11	11	11	11
10	11	11	11	11	11	11
11	11	13	11	13	13	12
12	10	10	10	10	10	12
13	11	11	11	11	11	11
14	11	11	11	11	11	11
15	11	11	11	11	11	11
16	11	11	11	11	11	12
17	12	12	12	12	12	11
18	11	11	11	11	11	12
19	11	11	11	11	11	11
20	11	11	11	11	11	11
21	11	11	11	11	11	12
22	10	10	10	10	10	11
23	12	12	12	12	12	19
24	12	12	12	12	12	19
25	11	11	11	11	11	11

Table G.2: The found threshold with different strategies and the true threshold per run for case study 2.

The darker orange the colour of a cell, the further is the found threshold level above the true threshold level. The darker green the colour of a cell, the further is the found threshold level below the true threshold level.

Run	Step	Bisection	Multi step	Multi bisection	Boltzmann	TRUE
1	3	19	15	15	2	19
2	11	19	11	11	11	19
3	8	19	19	19	19	19
4	7	19	7	19	11	19
5	3	18	18	18	16	19
6	15	19	15	15	15	19
7	19	19	19	19	19	19
8	3	9	9	19	19	19
9	19	19	19	19	19	19
10	19	19	3	3	19	19
11	5	9	9	19	5	19
12	2	17	17	17	19	19
13	11	14	11	19	19	19
14	4	19	3	3	19	19
15	9	9	9	9	18	19
16	10	18	10	13	13	19
17	13	19	13	13	19	19
18	12	19	12	19	12	19
19	8	19	8	19	8	19
20	4	19	12	19	10	19
21	19	19	19	19	19	19
22	2	19	6	19	19	19
23	11	19	11	11	19	19
24	5	19	5	5	5	19
25	6	19	6	11	15	19

Table G.3: The found threshold with different strategies and the true threshold per run for case study 3.

G.3 Found and true threshold for each case based on the sensitivity

The right-hand column shows the difference between the true threshold level based on sensitivity as a measure instead of the false omission rate as a measure.

Run	Bisection	Multi bisection	TRUE (based on sensitivity)	TRUE (based on FOR)
1	2	2	2	3
2	2	2	3	2
3	2	2	2	2
4	2	2	3	3
5	2	2	2	3
6	2	2	3	3
7	2	2	2	3
8	3	3	3	3
9	2	2	2	2
10	2	2	3	3
11	2	2	3	3
12	2	2	2	3
13	2	2	3	3
14	2	2	2	2
15	2	2	2	3
16	2	2	2	3
17	2	2	3	3
18	2	2	3	3
19	2	2	2	3
20	2	2	3	3
21	2	2	3	3
22	2	2	3	3
23	2	2	2	3
24	2	2	3	3
25	2	2	3	2

Table G.4: The found threshold with different strategies and the true threshold per run for case study 1, based on the sensitivity as measure.

The darker green the colour of a cell, the further is the found threshold level below the true threshold level. The right-hand column shows the difference between the true threshold level based on sensitivity as a measure instead of the false omission rate as a measure.

Run	Bisection	Multi bisection	TRUE (based on sensitivity)	TRUE (based on FOR)
1	9	9	12	12
2	10	10	13	12
3	9	9	12	12
4	9	9	12	12
5	8	8	11	11
6	8	8	14	19
7	11	11	11	11
8	10	10	11	11
9	10	10	11	11
10	9	9	11	11
11	10	7	12	12
12	7	7	11	12
13	8	8	11	11
14	9	9	11	11
15	9	9	11	11
16	9	9	10	12
17	9	9	11	11
18	11	11	12	12
19	9	9	11	11
20	9	9	11	11
21	10	10	11	12
22	9	9	11	11
23	7	7	14	19
24	9	11	14	19
25	10	10	11	11

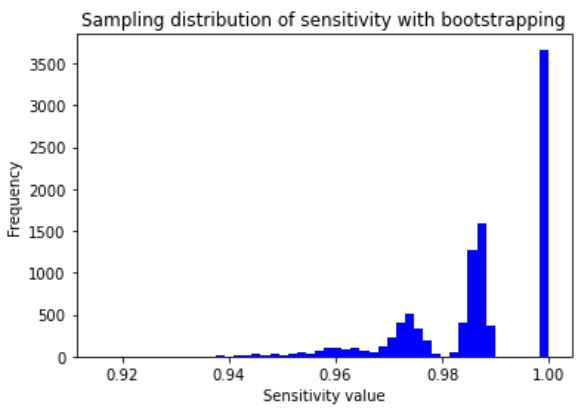
Table G.5: The found threshold with different strategies and the true threshold per run for case study 2, based on the sensitivity as measure.

The darker orange the colour of a cell, the further is the found threshold level above the true threshold level. The right-hand column shows the difference between the true threshold level based on sensitivity as a measure instead of the false omission rate as a measure.

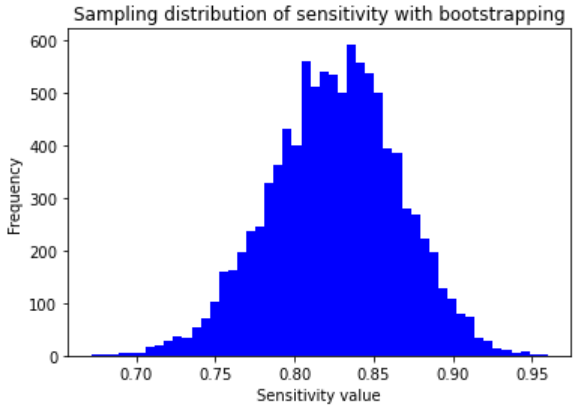
Run	Bisection	Multi bisection	TRUE (based on sensitivity)	TRUE (based on FOR)
1	1	1	0	19
2	1	1	0	19
3	1	1	0	19
4	1	1	0	19
5	1	1	0	19
6	1	1	0	19
7	1	1	0	19
8	1	1	0	19
9	1	1	0	19
10	1	1	0	19
11	1	1	0	19
12	1	1	0	19
13	1	1	0	19
14	1	1	0	19
15	1	1	0	19
16	1	1	0	19
17	1	1	0	19
18	1	1	0	19
19	1	1	0	19
20	1	1	0	19
21	1	1	0	19
22	1	1	0	19
23	1	1	0	19
24	1	1	0	19
25	1	1	0	19

Table G.6: The found threshold with different strategies and the true threshold per run for case study 3, based on the sensitivity as measure.

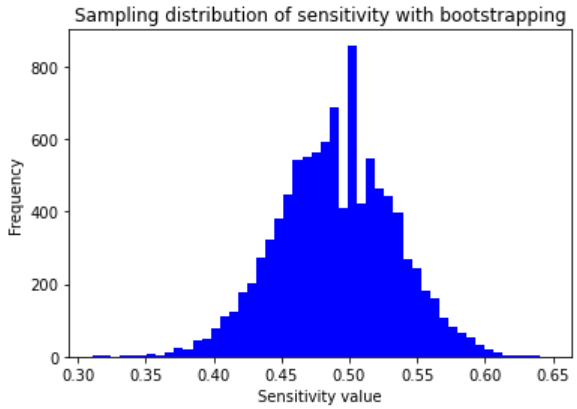
G.4 Bootstrap sampling distribution



(a) Population sensitivity = 0.9734 (level 2)



(b) Population sensitivity = 0.8 (level 3)



(c) Population sensitivity = 0.557 (level 4)

Figure G.3: The sampling distribution of the sensitivity of case study 1 of threshold level 2, 3, and 4 with $B = 10000$ and $n = 264$.