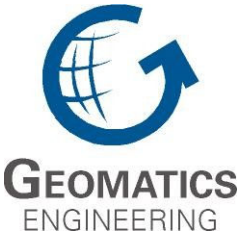


UCGE Reports
Number 20342

Department of Geomatics Engineering

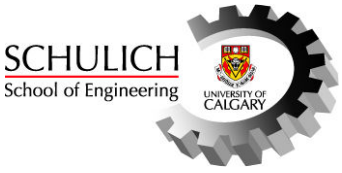
Detecting Fraudulent Activities in Land Record Systems: An Application of Data Mining

(URL: http://www.geomatics.ucalgary.ca/graduatetheses)

by

Thaer Shunnar

September 2011



UNIVERSITY OF CALGARY

Detecting Fraudulent Activities in Land Record Systems: An Application of Data Mining

by

Thaer Shunnar

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF GEOMATICS ENGINEERING

CALGARY, ALBERTA

SEPTEMBER, 2011

Abstract

Tenure security is a principle and critical factor in providing social and political stability. The major goal of this research is to assist the discovery of fraud patterns in land and property transactions using data mining techniques. The methodology starts with analysis of different fraud schemes to discover fraud patterns and their indicators in land records. A data simulator is developed to generate synthetic datasets. Different algorithms are then applied to detect fraud patterns in these datasets. Three major fraud schemes were used to validate the proposed approach (land grabbing in post conflict situations, Oklahoma Flip, and ABC-Construction).

The results have proven that data mining methods can identify fraudulent activities. However, these methods cannot be generalized to be used with all datasets. Finally, data mining can facilitate the building of fraud detection models for land transactions that can then be integrated with registration systems, and act as an alarm system.

Acknowledgments

This thesis would not have been completed without the grace of God. All thanks are due to Him.

I would like to express my gratitude to my supervisor Dr. Michael Barry for giving me the opportunity of joining his group, for all the consistent support, guidance and encouragement, for his patience with my lack of experience, and finally for everything he taught me. I would like also to thank Dr. Andrew Hunter for his help and the insight he gave me, which helped me in tackling the problems of this research.

I am indebted to all the great people who always offered their help and support to me and were there for me whenever I needed them. Thank you to Lani Roux, Mohammed Alshalalfah, Abdel Rahman Muhsen, Derar Alassi, Alaa' Kassab and his wife Nour Kassab, Islam Hegazy, Richard Hall, Elizabeth Hall, Vidya Renganathan, Oday Haddad, Ossama Al-Fanek and Christopher Venus.

I am most thankful and grateful for John Holmlund for his financial support. Without his exceptional generosity, this research would have been impossible.

Lastly, and most importantly, I wish to thank my family. My mother, Sana' Shunnar, my father, Abdel Aziz Shunnar and my two brothers, Ryad and Zyad Shunnar, for they are always there for me, they believe in me, and give me endless emotional support.

Dedication

To whom I owe everything, to my beloved parents Sana' and Abdel Aziz who raised me, taught me, and made me the man who I am. No words can describe my gratefulness to you and my only hope is to always make you happy and proud.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
AC	Average Change in property value
AFP	Average Flip Period of a property
ANOVA	Analysis of Variance
CART	Classification And Regression Trees
CCDM	Core Cadastral Domain Model
CISC	Criminal Intelligence Service Canada
CREB	Calgary Real Estate Board
DA	Discriminant Analysis
DDA	Descriptive Discriminant Analysis
DS	Dataset
DSS	Summarized Dataset
FLDA	Fisher Linear Discriminant Function
GIS	Geographical Information System
ICT	Information Communication Technology
KNN	k^{th} Nearest neighbour
LADM	Land Administration Domain Model
LOO	Leave One Out
LPDS	Labelled Properties Dataset
LRDS	Land Records Dataset
LRDSS1	Summarized Land Records Dataset

LRS	Land Records Simulator
LTV	Loan to Value
LTVR	Loan to Value Ratio
NP	Number of Persons involved in transaction
NT	Number of Transaction on one property
PDA	Predictive Discriminant Analysis
PDS1	Properties Dataset
PTDS1	Property Transactions Dataset
Q-Q	Quantile versus Quantile
RNN	Replicator Neural Networks
SPIN2	Spatial Information System - Government of Alberta
STDM	Social Tenure Domain Model
SVM	Support Vector Machine
TDIDT	Top-Down Induction of Decision Trees
UN-FAO	Food and Agriculture Organization of the United Nations
XML	Extensible Markup Language

Chapter One: Introduction

1.1 Introduction

This thesis identifies some of the illegal activities that occur in the trading of lands and other forms of real estate. It also examines the traceability of these activities in land information management systems. In particular this research project explores the application of data mining methods to identify suspicious transactions in land records which may underlie illegal manipulation of land and property ownership. For data mining to be applied effectively, an understanding of the techniques used to commit illegal activities, such as land grabbing or exploiting the registration systems for personal gain, is vital to enable the tracking of these activities. Tracking land records for fraud and other forms of mischief or negligence is an important facet of tenure security management and this is the prime contribution of this thesis.

Security of tenure is a critical factor underlying social and political stability. Secure tenure is one of mankind's most basic needs, and is a major contributor to the economic and cultural development of civilizations. To enhance land tenure security and property ownership, research has been growing in two main directions. The first stream focuses on creating land administration models to capture the complexity of the different situations which occur in developed and growing economies (e.g. van Oosterom and Lemmen, 2002; Lemmen and van Oosterom, 2006; Lemmen *et al.,* 2007; Barry *et al.*, 2007; Muhsen, 2008). The second stream strives to understand and explain the complexity of land tenure in changing situations in developing economies and employ

new technologies and new trends in software system development to create better applications for land administration in these situations (e.g. Roux and Barry, 2001; Barry *et al.*, 2007; Muhsen, 2008; Hay and Hall, 2009).

This thesis argues that a healthy analysis of land and property records, which form the land information infrastructure, is needed to further enhance tenure security. Enemark (2005) notes that based on the information infrastructure that holds all the land related data in land administration systems, we can support sustainable development through the enhancement of the land administrative functions; i.e., tenure, value, use and development.

Information infrastructures are growing very fast in current integrated land information systems. This data explosion trend can be found in many other fields, such as medical information systems and banking systems. Derived from the availability of data storage technology with low costs, information is being accumulated rapidly from a variety of different resources.

There is a growing realization that all the information being collected and stored may contain important knowledge which has significant potential to help scientists, decision makers, and researchers. This knowledge could make it easier and faster for them to perform their tasks (Bramer, 2007). Niasbit (1982, p. 17) notes "We are drowning in information, but starved for knowledge". As a response to these large volumes of information and the need for knowledge, the concept of data mining has emerged as a way of finding valuable patterns in data, and is now a well established methodology which is applied in many fields (Weiss et al., 2005).

This thesis aims to identify certain forms of knowledge that can be inferred from the information infrastructure that supports land administration systems. It will also assess the integration of data mining with these systems as a means to facilitate the extraction of useful knowledge.

1.2 Problem Definition

Useful knowledge may be hidden in the data stored in land record systems (cadastral survey systems, land registration systems, land information systems). This knowledge, if extracted, may provide good support for planners, decision makers, and legal institutions. This will contribute to the detection of illegal activities, the governance of land, and improved tenure security.

Knowledge discovery from large datasets has been an active field of research for the past two decades. These studies are driven by a desire for automated systems which can search, analyze, and extract knowledge from the massive amount of data collected in many fields. The main goal is to replace the conventional manual examination methods which are expensive, inaccurate, error prone and limited in scope (Bramer 2007).

Land records should also be searched for unusual patterns and undiscovered knowledge. This research demonstrates that different kinds of illegal manipulation of the legal instruments used in property transactions can be discovered by identifying particular patterns and track them in the datasets. Integrated land information systems contain various forms of related datasets such as taxation records, personal details, survey plans, deeds, titles, building plans, property management files, maps, aerial photographs and satellite images. The mining of these various datasets may reveal different kinds of

patterns which could indicate behaviours such as land grabbing and property price manipulation.

Fraud indicators have not been fully explored yet in the literature. Pollakowski and Ray (1997) examine housing price changes by analyzing housing price indices for various U.S. metropolitan areas. They specifically identify price shocks in one area to that of the neighbouring area. Such revelations can highlight unexpected behaviour such as fraud, cheating, and other criminal activities. Other studies that address the problem of fraud in land and property transactions can be found in Kontrimas and Verikas (2011), Auditor General of Alberta (2010), Unger *et al* (2010), and Nelen (2008).

Fraud indicators in land and property transactions are the focus of this thesis. The problem is formulated by 1) recognizing those indicators, the patterns associated with them, and the human behaviour underlying these patterns; 2) a data mining approach to automate the discovery of the illegal activities that generate the patterns.

1.3 Significance

This study will contribute to current efforts in establishing better systems to support tenure security and land governance. To achieve this, two main problems are addressed. 1) Assessing the patterns hidden in land records which can be used to point out useful knowledge. 2) Automating the discovery of some of these patterns from land records by applying data mining techniques.

A major problem is the lack of published work that addresses the automatic extraction of knowledge from land records systems. Therefore, in some aspects, this is a pioneering study. Data mining in particular has been applied sparsely to land tenure

records. To the best of my knowledge, this thesis is one of the first studies in this area. Throughout this research, it has been found that a significant amount of criminal investment is taking place in land and real estate markets. Unfortunately, there are few studies that address this problem. It is also found that the process of searching and analyzing land information to uncover hidden relations and patterns is mainly done in un-automated or semi-automated methods. These conventional methods, which use simple searching techniques and reporting systems, may fail to achieve the desired land administration tasks. There is also a gap in knowledge of the criminal investment in land transactions which should be addressed, but this falls outside the scope of this work.

1.4 Research Objectives

The primary objective of this thesis is:

> *To explore the use of data mining in land record systems and to develop knowledge of where and how data mining can be applied and integrated into these systems, to contribute to the discovery and alleviation of fraud in land and property transactions.*

As stated in section 1.2, the primary objective of this thesis is set to provide a solution to the overarching problem of tenure security and land governance by detecting fraud. To serve this primary objective, four main activities or sub-objectives are set. They are:

a. Identify different fraudulent activities in land record datasets, in a variety of contexts where these activities may take place.
b. Identify suitable data mining techniques that may help in detecting some of the fraud activities found in (a).
c. Design and develop a data simulator to generate land record datasets.
d. Identify existing tools to apply the methods found in (b) above and develop tools where it appears that relevant tools do not exist.

1.5 Research Questions

Section 1.2 mentions that the issues this research addresses should contribute to the solution of a higher problem. The main research question forms part of the answer to the broader question of how to improve tenure security and land governance.

The primary question of this research is *where and how can data mining be used in land record systems in order to improve land governance and tenure security.* To help in answering this question, the following questions are set to underlie the activities of this research:

1. What are land record systems and what is the information infrastructure underlying these systems? Answers to these questions will provide an understanding of the underlying data infrastructures of land records systems. This is important since the application of any form of data mining requires an understanding of the stored information and the relations between the different records. This question is discussed in Section 2.2 of this thesis.

2. How the tasks of information analyses for land information systems are being handled in the current systems? It is vital to understand the current methods followed when analyzing land record systems for two reasons. Firstly, it will provide an assessment of the need for developing automated systems to automate the process. Secondly, it will make it easier to develop such a system by understanding the requirements of the process. This question is addressed in Sections 2.3 and 2.4.
3. What is the knowledge that experts may extract from land records in the different situations to detect fraud and how can this knowledge help in improving the situation? Answering this question constitutes the first step of identifying criminal behaviours in land and real estate transactions. This question is briefly addressed in Section 2.4. More analysis of the types of knowledge is presented Chapter 3Chapter Three:.
4. What are the different fraud indicators, schemes, and patterns that can be found in land and property transactions? Identifying those indicators, schemes and patterns is the first step to developing a mining technique for each of them. It provides the required understanding of the behaviour of the attributes and records inside a land or property transactions dataset. Fraud schemes and indicators are addressed in details in Section 3.2.
5. Which of the schemes and patterns addressed in the previous question could be tracked by analyzing land datasets? The answer to this question scopes the

fraud schemes and patterns that have been tested using data mining methods throughout this research. This scope is provided in Sections 3.3 and 3.4.

6. Can data mining methods be used to discover fraudulent activities or any suspicious behaviour by analyzing land data sets? If yes, how can this be achieved and what are the specific data mining methods that can help? This question is addressed primarily in Chapter 4.
7. And what are the required datasets needed to test the methods developed in this research and how to obtain these datasets? The answer to this question is addressed in details in Chapter 5of this thesis.

1.6 Research Method

This section describes the activities that were carried out by the author in order to address the problem of the research, answer the research questions, and achieve the goals mentioned in the previous sections. Figure 1.1 presents a summary of the primary activities and methods used in this study.

1) **Literature review of land record systems:** This activity serves the goal of understanding the land record systems in general in order to be able to identify and analyze unusual behaviours, which contributes to objective 1.4.a. Many aspects are covered in this review namely.
   a) The development efforts and challenges of land record systems. This will help to better understand the research problem by identifying the functions, goals, and problems of current systems.

b) Search the literature – newspapers, law reports, articles in academic journals - to find any kind of unusual behavioural patterns that may take place in land record. The author looked for is any kind of criminal or illegal activities in trading land property which might leave a noticeable trend in land data.

2) **Interviews with experts in land information systems:** In addition to the literature review; a qualitative survey was conducted using unstructured interviews because literature is sparse. This survey asked experts in the field of land information systems (land agency workers, cadastral system experts and researchers, detectives, planners, lawyers, and consultants), from different countries (Australia, Canada, Great Britain, the Netherlands, and Ireland) and organizations such as the World Bank, UN-Habitat and UN-FAO, about their knowledge in indicating the types of frauds or unusual patterns that may occur in land records. Each of the interviewees was asked about different points that could give more insight into the problem of fraud. Some of the key points asked were:

   a) The types of fraud or unusual behaviour that they have experienced or heard.
   b) Some anecdotes of cases they have come across in the different jurisdictions they have worked.
   c) Any benchmark court cases or other documents that may be related to this study.
   d) The methods followed in land registries for data analysis for the purpose of detecting fraud and errors, planning, or providing decision support.

e) Patterns in land registry data and other land records that could be simulated and then might be identified using various pattern detection algorithms.

f) And finally do they know other people who could have more information to help this study.

3) **Create a set of fraud schemes that reflects the findings from the first two activities above:** In this activity the author studied and analysed the behaviours found in the first two activities to come up with a set of racketeering methods and schemes that are used in the real world. For each scheme, the author tried to find the effect of conducting it on the datasets and how the corresponding records differ from the records of any other usual registration activity. The findings are summarized in lists of fraud patterns and indicators. Based on the patterns and indicators studied for the schemes, some of the schemes are selected for the testing of data mining. This selection process is mostly based on the available data. After completing this activity, Objective 1.4.a would be achieved.

4) **Literature review of data mining**: this phase encompasses two activities:

a) Literature review of data mining concepts and techniques. This will help in understanding the problems data mining techniques might be useful in solving.

b) Literature review to understand the existing fraud detection techniques and examine some case studies of fraud detection in different fields. This is an important activity since the focus of this study is more toward detecting fraudulent activities in land and property transactions. This review helps to

understand how to formulate schemes found in activity 3 above in order to apply detection methods.

5) **Formulation of the problem and algorithms identification:** for each fraud method selected in activity 3 to be part of the experimental work, the problem of detecting fraudulent activities in that scheme is formulated as a data mining problem using a data mining method. Mainly; all schemes are formulated as one of two problems; a classification problem or an outlier detection problem. Algorithms are chosen based on assessments from the literature and applicability to the available data. A search is conducted for available tools such as data mining toolboxes provided in MATLAB or the machine learning/data mining software (WEKA) developed in Bouckaert *et al.* (2010). These tools can provide detection methods that can be applied to extract the targeted pattern from the records. If no available tools were found, the algorithm was implemented by the author.
6) **Data simulation:** In this step, a land records simulation system was developed to provide the required data for the experiments in this research. The need for a simulator stems from not only the lack of uniform data sources but also the lack of access to land or property data. This problem will be discussed in Chapter 5.
7) **Testing:** Perform the tests for each of the developed algorithms on the appropriate datasets and analyze the results. This step includes performance evaluation and critique for the used methods.

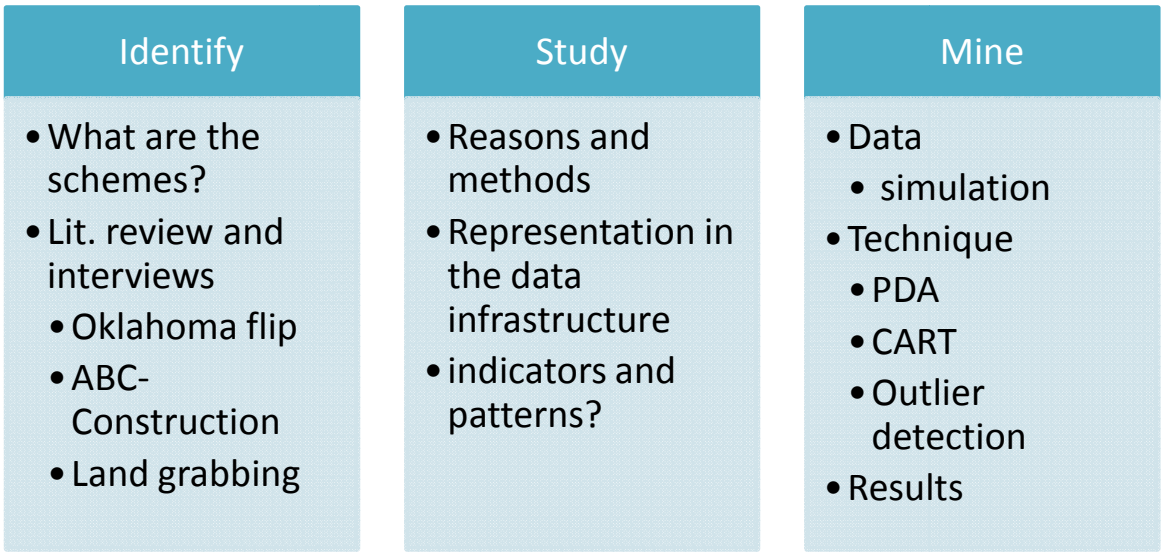


Figure 1.1: Summary of the research methods followed in this study.

1.7 Scope and Limitation

Data mining concepts and techniques can help in solving many problems. This thesis suggests that data mining should be studied and applied in land management; however, it only investigates the application of a set of techniques that are used for fraud detection. It is realized by the author that other data mining techniques may be applied to solve the same problems addressed here. It is also realized that some techniques might be better and provide more accurate results than the techniques used in this research. However, the author tries to apply well-known methods that provide acceptable results as a proof of concept.

It is important to note that there is no intention in this thesis to develop new algorithms for fraud detection. The focus is on trying to find suitable existing algorithms that can be applied and perhaps adapted them to the fraud problems in land records.

Furthermore, it is not the purpose to compare and evaluate performance and efficiency of different algorithms; the main goal is to evaluate the usability and the efficiency of each algorithm in the context of the pattern it is applied to and the compare the results of the used algorithms. The reason of this scoping of the current work is because it is pioneering work and more focused on addressing the problem and the solutions while efficiency can be developed later.

One of the major limitations of this study is the lack of real datasets. As data is a major factor in the success of achieving the objectives, the existence of real datasets would have helped substantially in the progress of the research. A property transactions data simulator was developed to overcome the lack of data availability. While this simulator has advantages, it has some limitations and creates a finer scope for the type of data to be worked with. The advantages and limitations of the data simulator are discussed in Chapter 5.

The fraud patterns in land and property transactions identified in this thesis may be found only in some jurisdictions. That is because of the differences between the jurisdictions in the process of registration and the existence of different policies and regulations. Therefore, there might be different patterns in different jurisdictions which could have similar indications. However, this problem goes beyond the scope of this thesis.

Finally, only the patterns identified from the literature or the surveys conducted in this thesis are handled in the research. This set of patterns does not cover the whole range of patterns that could be found in land records. The author believes there are many other

patterns that need to be identified and could indicate useful knowledge concerning tenure security.

1.8 Thesis Organization

Chapter 2 provides the theoretical background related to the research problem which is the first part of the literature review. It also lists the unusual behaviours that occur in trading in land gathered from literature and interviews. Chapter 3 presents the list of fraud schemes found in this research. In Chapter 3, the author examines in details fraud schemes that would be included in the experimental work and also analyse fraud patterns and indicators of the selected schemes. Chapter 4 provides a technical background about data mining and its usage in fraud detection and then discusses the adopted data mining methods in this study. Chapter 5 presents the developed simulator for the sole purpose of simulating the required data sets for this research. It also examines the simulation process for the datasets used in this research. Chapter 6 includes the experiments and the analysis conducted in this research. Finally, chapter 7 summarizes this research and concludes it then discusses the future work.

1.9 Chapter Summary

This chapter introduced the research problem and the suggested solution. It starts with defining the problem and the contribution of this thesis to the solution. Then, the objectives of the research are listed followed by the research questions. The methods used to answer the research questions and achieve the objectives of this research are then

discussed. Then, the scope and limitation of this study are discussed and lastly the chapters of the thesis are outlined.

Chapter Two: Land Record Systems and the Problem of Fraud

2.1 Introduction

The literature review is divided into two parts: a review of the fraud problem in land record systems and a review of data mining methods that can help solving the problem. This chapter presents the review of the fraud problem in land record systems and it also reports on interviews on the nature of the problem.

Mainly, this chapter expands on the nature of the research problem discussed in Section 1.2 by reviewing fraud literature and examining interviews the author conducted with experts to better understand the problem. The aim in this chapter is to help the reader better understand the fraud problem. It also aims to give the reader an initial understanding of some fraud methods and how they are dealt with in some previous studies.

The chapter contributes to the fulfilment of objective 1.4.a by addressing research Questions 1, 2 and 3. It commences with defining the term *land record system*. Thereafter, it describes the problem of large land datasets with examples of how this data is being analyzed. The chapter concludes by describing major fraud types, methods and causes, and then examines two studies that deal with the fraud problem in land record systems.

2.2 Land Record Systems

Land Record Systems are experiencing a data explosion, as are most other information systems. This study is concerned with the data held inside a system rather than the type of the system. In many parts of the developing world, much of the land data is still in paper format. However, studies are emphasizing the transition from analogue to digital form, and land data is being converted continually into digital form (Hallett *et al*., 2003; Haanen *et al*., 2002; Nyerges, 1989). In this study, I have assumed that all land records are in digital form.

For the purpose of this study, the term *Land Record Systems* refers to the different kinds of systems that are used to manage all aspects of land and property planning, policy making, and ownership. Land Management Systems, Land Administration Systems and Cadastral Systems are all concerned with different aspects of the process of managing the relationship between people and land. All these systems fall under the term Land Record Systems.

Part of the review was to understand the different cadastral models and how relations between people and land are conceptualised. This review was important because for any data mining method to work, the data infrastructure of the targeted system should be reviewed and understood.

Many cadastral data models have been developed to support land record systems. All of these are trying to model the relation between land and people via rights. Some of these models include: Core Cadastral Domain Model (CCDM) (van Oosterom and Lemmen, 2002), Land Administration Domain Model (LADM) (Hespanha *et al*., 2008),

Social Tenure Domain Model (STDM) (Lemmen *et al.*, 2007), and Talking Titler Model (Barry and Khan, 2005; Augustinus and Barry, 2006; Barry *et al.*, 2007 and Muhsen and Barry 2008). Further details of these models can be found in *Appendix A*.

To be aware of these models and other land administration models leads to an understanding of the way data is stored and how relationships are implemented. This is important for future work that builds on this study. However, all the datasets used in this study were simulated using simpler data structures that fit the needs of this study. Simulation is used for many reasons that are discussed in Chapter 5.

2.3 Land Records and the Task of Data Analysis

With all the existing cadastral models to support land record systems, and all the development efforts put into the ICT infrastructure, many of the management tasks are becoming easier. However, vast quantities of data are being collected, which makes decision support more and more complicated. This section describes some examples of data analysis tasks to illustrate two problems. The first is to how large the data infrastructure has become in land management. The second is the problem of a lack of automated tools to explore land records in order to look for errors, negligence and most important, fraud.

Tokyo Sexwale (2009), in the housing budget vote speech in South Africa, mentions that in the first two months of the financial year of 2009, the provincial housing departments in South Africa reported the delivery of more than 22,000 housing units. This delivery raised the number of subsidized homes delivered by the government since

1994 to a total of 2.3 million homes (Tokyo Sexwale 2009). Of course, for each one of the delivered homes, at least one transaction occurs in the registration systems.

This example indicates the volume of information; one can expect some errors due to negligence and also some fraud. According to the Department of Rural Development and Land Reform in South Africa, the deeds controllers in the nine deeds offices of the Chief Directorate in South Africa examined 2,889,867 deeds in 2009. During this examination process, 26% of the examined deeds where rejected, as they were found to be problematic and could not be registered due to conveyancing errors, attachments, interdicts or legal constraints (Department of Rural Development and Land Reform 2009). This is a high ratio of errors; and this examination process is not automated and consumes a lot of resources. However, if it is possible to identify the kind of patterns in the information which define a rejected deed, it would be possible to develop a data mining tool to automate the process of examination and make it faster and cheaper.

More examples of how sizable land records can be are seen in Alberta and in Hong Kong. In Alberta, Registries within Service Alberta processed approximately 1 million document registration requests between April 1, 2008 and March 31, 2009 (Auditor General of Alberta, 2010). In Hong Kong, the land registry received a total of 329,878 deeds for registration during the first seven months of 2009 (Land Registry of Hong Kong 2009).

The examples above show that land information is being collected and stored in massive volumes. Meanwhile there is no shortage of data collection, making it more

important to manage records with innovation and care. Land management incorporates the tasks of strategic planning, decision-making, and policy development based on an infrastructure of cadastral information (Enmark 2005). These tasks are not simple and require deep analysis of the data in order to be able to detect errors, negligence or fraud. However, with the focus on detecting fraud rather than errors, the relevant question is, how is the problem of fraud handled?

In an interview, WG who works as an investigator in the office of the Auditor General in Alberta confirmed that they have a department for fraud detection. WG illustrated that the responsibility of the department is to analyze registration documents in order to detect any suspicious activities. Their analysis process encompasses two steps, In the initial step (first pass), millions of transactions are filtered based on different criteria. The filtration step allows the investigators to decide on the transactions that will go to the second step (second pass). This second pass comprises manual processing and analysis of documents corresponding to transactions filtered in the first pass. It includes files obtained from sources of information other than titles, thus involving more information about each transaction in the analysis. After this pass, the investigators can decide on transactions or properties that look suspicious and would probably indicate fraud cases (WG 2010, pers. comm., 14 June).

The process of fraud detection in Alberta is discussed in detail in Section 2.4.2.2 as reported by Auditor General of Alberta (2010). There is one investigator in the Special Investigation Unit who completes 20 files a year. The reason given for this low throughput is the lack of data-mining capabilities.

This section showed that fraud is a problem in land record systems due to the size of data and the lack of automated analysis tasks. The following section focuses on the types of frauds that have been identified from the literature and personal communications.

2.4 Fraud Types and Methods in Land Record Systems

Fraud can occur in land transactions in different ways. The type of the registration system and the situational attributes surrounding this system are important determinants of the type of fraud. A fraud investigator, NE (2010, pers. comm., 3 March), noted that fraud in land registration can be subdivided into two main categories: fraud by forgery and fraud by impersonation. However, in post-conflict situations, fraud may take place by means of use of power and force (Lewis, 2004; Zevenbergen and van der Molen, 2004; and Future and Commodity Market News, 2009).

Watkins (2007) states that regardless of the differences in the processes and procedures of land conveyancing between the different states, since conveyancing systems are based on title registration, they are inherently susceptible to fraud. Why this fraud happens varies based on the registration process and procedures, the local customs, and the uncertainty of the situation in an area.

Studies show that fraud schemes are increasingly sophisticated due to the use of technology in the registration process (CISC, 2007). For example, one informant from the UK noted that the land registration system in the UK is open to fraud since it is not a very complicated system and can be easily manipulated (AB 2009, pers. comm., 1 October).

On the other hand, sometimes it is the disregard of formal institutions and formal registration processes that facilitate fraudulent activities. People may ignore the formal institutions to avoid discrimination, taxes, bureaucracy, and differences between in-state rules and customary or religious rules (Zevenbergen and van der Molen 2004). Uncertainty provides a rich ground for fraud and manipulation to take place in land transactions. For example, in post-conflict situations, dominant groups, parties, or individuals may manipulate the registration system by destroying or altering old records, or creating new ones for their own gain (Zevenbergen and van der Molen 2004).

In this study, two categories of fraud have been examined: fraud in post-conflict situations, and fraud in real estate transactions in developed communities.

2.4.1 Fraud in Post-Conflict Situations

In post-conflict situations, as described by Daudelin (2003, p. 3), "Tenure is possibly becoming less secure than ever before. It finds itself caught between the common—but not universal— breakdown of customary systems, and attempts by weak national states to replace or do away with them". Also, after a conflict the situation becomes more complex, and existing data records need very careful investigation when trying to fix or re-construct an existing land administration system (Zevenbergen and van der Molen 2004).

Because of the uncertainty created in the post-conflict situation, the registration system is weakened and becomes more vulnerable to fraud and illegal manipulation. Powerful parties, groups and warlords will try to take advantage of the brittle system for

their personal gain. NA (who is an experienced international expert) mentioned that during his experience in dealing with post-conflict situations, almost in every scenario, land is illegally or fraudulently usurped by one party or another. When asked about how these frauds take place, NA noted simply, "Records if they exist are manipulated, destroyed, or altered" (NA 2010, pers. comm., 16 January).

Many countries around the world suffer from conflicts or have suffered from one and are in a post-conflict state. For example, Combodia, Sri Lanka, Brundi, Mozambique, Palestine, Colombia and Guatemala are included in the study of Daudelin (2003). Other studies have looked at situations in Iraq, Afghanistan, Somaliland and Sudan and discuss the challenges with respect to land, housing and property (Barry, 2009; Lewis, 2004; Alden Wily, 2004). Some of these challenges include resolving land, housing and property disputes, removing discrimination from the land, housing and property sector, and re-establishment of the registration system (Lewis 2004).

Since the goal here is to address fraudulent activities in post-conflict situations with respect to land and property transactions, and how data mining can help in detecting these frauds, there will be no discussion of the other issues surrounding post-conflict situations and the structure of land administration systems during these periods. Rather, the discussion is focused on the trends and patterns in manipulating the registration systems and the effects these manipulations have on the records.

In post-conflict situations, fraud patterns identified from the literature are all patterns that indicate land-grabbing cases. Each pattern is produced by a certain way of seizing or grabbing land by individuals or groups.

In general the author has identified five patterns described in Zevenbergen and van der Molen (2004). These patterns are discussed as manipulations in land records that could be practiced by powerful groups during or after a conflict. The five patterns are:

1. "An unusual number of transactions of a certain type in a short time, or even on the same day.
2. Periods without, or with few, transfers, which might indicate that certain parts of the transaction records have been removed.
3. A lack of transfers especially when the overview data showing the situation just prior to the conflict is missing, and/or a new group has come to power.
4. Transfers between members of different groups in the conflict.
5. Transfers from public, common or communal properties to private persons often in the form of privatization."

(Zevenbergen and van der Molen, 2004).

These manipulations can be categorised into two groups. The first group includes manipulations that would affect the total number of transactions within a certain period of time, as in the first three types. This can happen through land-grabbing by warlords and government officials who have access to the registration system. It also can happen through the destruction of records to remove other parties' interests in land.

The second group includes manipulation that would have an effect on the normal distribution of parties involved in the transactions. For example, point number four describes land transfers between members of different groups. This might be a normal case in normal situations. However, during or after a conflict, these kinds of transfers

would appear unusual. Also the direction of the transfers will be different in the sense that land will be transferred most often in one direction to members of the powerful group (Zevenbergen and van der Molen, 2004).

An example of land grabbing in post-conflict situations can be seen in Afghanistan. According to Commodity Market News (2009), the Ministry of Urban Development in Afghanistan says that every day, powerful government officials and warlords in Afghanistan are grabbing between 1,000 and 1,500 jirib (roughly equal to an acre) of land on which families reside. Most of this land grabbing occurred during the early years of the U.S. occupation. Government officials abused their positions during the chaos and used their privileged access to official maps and property ownership records to make counterfeit deeds during the early years of the war. The Ministry of Urban Development says that in total, more than 3.5 million jirib of lands in Afghanistan have been stolen.

Illegal manipulation of land records is not limited to post-conflict situations. According to NA (2010, pers. comm., 16 January), in the so-called developing countries, administration of land rights is a questionable process. This is because the architecture of Land Law and institutions governing land rights are not matched with the capacity either for compliance or enforcement. So, there is a large opportunity for those with insight and/or power to take advantage. Examples can be found in Kenya and Zimbabwe.

In general, the manipulations mentioned above will leave patterns in the records. Zevenbergen and van der Molen (2004) note that after a conflict, records need to be carefully examined in an attempt to find these patterns. If we assume that there is a

computerized information system for land records, data mining would make it easier to identify the patterns and find the manipulations. However, the author could not find any published work that uses data mining for this purpose.

2.4.2 Fraud in Real Estate Transactions in Developed Communities

The real estate sector is large and of high value, which makes it prone to criminal investments more than other sectors such as the bond market. The real estate market includes transactions of private homes, office buildings, public buildings (schools, hospitals, courts, etc.), shops, and company premises (Nelen, 2008). This estate sector is less transparent and hence more attractive for criminals (Unger *et al.*, 2010).

The Dutch Central Bureau of Statistics states that in 2006 the value of real estate transactions in the Netherlands was 35.3 billion euro. The total market value has been increasing quickly, and where it was 764 billion euro in 2000, it was estimated at 2021 billion euro and 2171 billion euro in 2008 and 2009 respectively; and these figures are around three times higher than the value of the bond market in the same country (CBS, 2010). This high value is one of the most important reasons why this market is so attractive to criminal investment.

In Canada, according to CISC (2007), hundreds of millions of dollars are lost annually to mortgage fraud. Organized crime follows the location of strong housing markets across the country and so the problem is concentrated in large urban areas. In one high-profile case in Alberta, over 280 properties were involved in a fraud operation that

amounted to approximately $30 million in alleged fraudulent mortgages financed by 22 lending institutions (Auditor General of Alberta, 2010).

Drawing on Unger *et al.* (2010) and Nelen (2008), There are a number of reasons, excluding size, why real estate fraud is attractive to criminals. The market itself is a high-value market. Also, Property is generally seen as a safe investment, but at the same time there is a long tradition of property speculation. Furthermore, the real estate market lacks the transparency and homogeneity of most financial markets, and so fraudulent transactions may be more difficult to identify. As each property has unique features, the market is heterogeneous. The uniqueness of property as a commodity means that the market itself is not efficient in the same way that financial market prices tend to reflect most of the information available about a particular financial instrument at a given time.

Singularity of properties is a key feature that allows for criminal manipulation and exploitation. Singularity means that each property has a uniqueness which makes it different even from adjacent properties. It makes it difficult to assess the objective value of properties. Geographical location, quality and practical value of the building, as well as supply and demand, are the most influential factors in determining property values (Unger *et al.*, 2010). However, Nelen (2008) notes that getting a precise estimate of a property value is a difficult task, and most lenders have not established a careful assessment for each property they lend over.

In the following section, the author discusses a study done by Unger *et al*. (2010). In that study, researchers are trying to identify different indicators of criminal behaviours

and then use data mining to build a prediction model in order to predict fraudulent activities.

2.4.2.1 Flagging method for fraud detection in the Dutch real estate market

Unger *et al.* (2010) try to find some measurable indicators for criminal behaviours. They use available information from the Dutch real estate sector to be able to systematically analyze criminal investment in it. The study targets the criminal use of real estate in two cities, Maastricht and Utrecht, with the goal of developing a research methodology that enables users to separate real estate transactions into ordinary and conspicuous transactions using outlier mining techniques.

Unger *et al.* (2010) use objective data related to real estate objects such as unusual movements in housing prices and unusual change in ownership to build a prediction model. The objective of the model is to predict objects that might be involved in criminal activities, in order to build a tool that could warn tax and investigation authorities in the Netherlands about conspicuous real estate objects. They use a combination of methods from economics and criminology to establish the model; the role of economists is to identify unusual movements in the prices, while criminologists can point out maleficent behaviour constructs.

Using analysis from economists and criminologists provides good validation for the established fraud indicators. However, real datasets are required. In this study, availability of data is a major setback which is discussed in Chapter 5. The author depends on main indicators taken from Unger *et al*. (2010), which were included in a

land records simulator developed in this study. These indicators are summarized in Chapter 3.

A major question in this research is to identify unusual (exceptional) behaviours inside transactional databases of real estate sectors. In most cases, exceptional behaviours are represented by suspicious data points inside the datasets. These data points can be distinguished from normal data points using abnormalities in the transactions, which translate into abnormalities in a certain attribute or in a combination of attributes to create patterns. The author draws on the Unger *et al.* (2010) and Nelen (2008) studies to identify the main fraud indicators required for this research.

The most common attributes used in predicting unusual behaviours in real estate sector according to Unger *et al*. (2010) and Nelen (2008) are:

1. Owner has unusual number of transactions
2. Unusual fluctuation in housing prices
3. Changes in ownership
4. Foreign ownership
5. Objects that quickly change hands between owners
6. Objects financed without a mortgage
7. Financing methods
8. Unusual purchasing sum

To build the prediction model, Unger *et al.* (2010) identifies 25 measurable indicators. They use a flagging system in which each object is examined against the 25 indicators and receives a red flag for each indicator that applies to it. The more red flags

an object receives, the more suspicious it is. After flagging all objects, a random list of 200 objects is selected, of which 150 with the highest number of red flags were considered as conspicuous (testing group). The remaining 50 objects are considered not unusual (control group). Using the 200 objects, Unger *et al.* (2010) created a model to identify conspicuous objects by measuring the usefulness of each indicator.

In the model created by Unger *et al.* (2010), properties are classified into four classes according to the degree of conspicuousness (non-conspicuous, slightly conspicuous, medium conspicuous and highly conspicuous). This requires a great deal of delicacy, which might be applicable if there is enough information about the classification process which was available in Unger *et al.*'s study. Since the author tries to define a general set of rules to classify properties, the four-class model is replaced with a three-class mode (normal, suspicious and highly suspicious). In this model, it is easier to classify properties, as it is hard to draw the separation limits between the four classes used by Unger *et al.* (2010) using only a subset of the total indicators. It was also recognized that to use all 25 indicators requires access to different data sources which the author did not have.

In the following section, the author discusses another study that addresses the use of data mining to detect fraudulent activities in the property market in Alberta, Canada. This study is reported in Auditor General of Alberta (2010).

2.4.2.2 Data mining to improve detection of fraudulent transactions in Alberta

Service Alberta has established an investigation department to identify and mitigate the risk of fraudulent use of its land titles registration system. Recommendations from the Auditor General of Alberta to the investigation department suggest that the Department can improve its fraud detection by using data mining. The suggested methodology to do so is to first analyse land title data for suspicious transactions using data mining. The results of the analysis can then be used to focus investigations and prosecutions (Auditor General of Alberta, 2010).

Service Alberta identifies in Auditor General of Alberta (2010) two major types of fraud that take place in the registration system in Alberta. The first one is title fraud committed using forged documentation to transfer a title. The fraudster sells the property or obtains a mortgage against it and disappears with the proceeds. The second type is mortgage fraud. The fraudster in this type buys a property, increases its value by flipping it several times between associates, and finally obtains a mortgage against the property. These schemes and others are examined in detail in Chapter 3.

Based on the report of Auditor General of Alberta (2010), the department collected electronic copies of all land titles transactions during 2008-2009. The data was then summarized and grouped by property to obtain a test set. This set is filtered to keep only properties with three to seven transactions over the year. This process is used to test data mining and its ability to detect fraudulent activities.

A similar technique is followed in this thesis. However, the filtration process followed by the author is done by removing properties with only one transaction. Other

reports and studies show that properties with two transactions in a short period of time may prove to be part of a fraud scheme, and that is why they are kept in the fraud detection model developed in this research.

For the mining process followed in the report of Auditor General of Alberta (2010), an initial mining pass is applied to find properties that exhibited indicators of possible fraudulent activities. The criteria used in this pass according to the report are:

1. "Multiple transfers in a short time, or
2. Significant increases in property value with new mortgages taken shortly after the increase, or
3. The presence of individuals or law firms suspected of involvement with fraudulent property transactions."

Auditor General of Alberta (2010)

After the first pass a group of properties is selected, which comprises the top properties that show fraud indicators according to the criteria. This group is then examined using information from SPIN2, the online municipal property tax databases, and the Alberta Corporate Registry. A second pass examination takes place looking for indicators; some of them are new and some are repeated from the first pass. These indicators include:

1. High ratio mortgage which is defined as a loan of 75% or more of the property value.
2. Quick increase in the value followed by a mortgage.
3. Selling the property between the same individuals through corporations.

4. Foreclosure of the mortgage shortly after funding a loan.
5. A person with multiple mortgages on different properties from different lenders.
6. Property value is materially higher than municipal property assessment.

Auditor General of Alberta (2010).

As seen above, the report develops two sets of fraud indicators based on different sourcse of the information. This is because not all the information is available in the initial information source. Also, the second pass is done manually by investigators, so it is necessary to filter the records to reduce the number of documents required for scanning in the second pass, using the available information from the initial set of records.

In the methodology followed in this research, simulation is used to produce the datasets and that is why the author assumes the availability of all the required information in one database. So, indicators from both sets mentioned in Auditor General of Alberta (2010) are used. However, the simulation process itself is not complete and not all of the identified indicators can be taken into consideration.

After auditing the system, it was reported that “the computer system used to process all land titles transactions does not allow easy assembly for a property history. This limits the ability to quickly assess if a transaction is reasonable or falls into a suspicious pattern” Auditor General of Alberta (2010, p. 112).

Finally, the data mining methodology followed in Auditor General of Alberta (2010) was successful in detecting many suspicious cases. In total, 30 properties were

identified with doubtful behaviours. However, the problem is the lack of tools that could make the examination process faster and easier.

2.4.2.3 Discussion

Although the study of Unger *et al.* (2010) specifically targets criminal investments in the Dutch real estate market, it sets the foundation for this problem to be targeted on a larger scale and especially in developed countries. The problem of fraud in land and real estate transactions is widespread, and identifying fraud methods and putting forward effective solutions has drawn increasing attention among researchers, governmental institutions and media (Kontrimas and Verikas, 2011; Auditor General of Alberta, 2010; Unger *et al.,* 2010; Nelen, 2008; CISC, 2007; Watkins, 2007; Kontrimas and Verikas, 2006; Troister *et al*., 2006; CTV News, 2005; The Law Society of Upper Canada, 2004)

In this study, as mentioned above, the author is drawing on previous research in addition to interviews and personal communications with experts in land record systems in order to identify fraud indicators, schemes and patterns in property transactions. Although the expression *property fraud* is the one mostly used throughout this study, it is found that the general and most common criminal behaviour is to use real estate objects as a means to obtain mortgages. So, the term *mortgage fraud* can also be used in this study interchangeably with *real estate fraud*.

For this study, the author adopts the definition of Mortgage as noted by Unger *et al* (2010) which is "a loan where house serves as collateral. If the borrower cannot pay its down payment, the house is sold and the lender can get its money back". The author also

uses the term Mortgage fraud as defined by CISC (2007, p. 2): “a deliberate use of miss-statements, misrepresentations or omissions to purchase or secure a loan. Simply put, mortgage fraud is any scheme designed to obtain mortgage financing under false pretences such as using fraudulent or stolen identification or falsifying income statements”.

Finally, Real estate transactions refer to the buying and selling of properties and not to the rental market. However, the author is aware that the rental market is also used in many fraud schemes. In particular, this study uses statistics from the housing market for the purpose of data simulation. This is because the housing market occupies a high ratio of the total criminal investigation in real estate market (Unger *et al.,* 2010). In addition, the commercial market is very dynamic, so fraud indicators that apply to houses may not apply to commercial property objects.

2.5 Chapter Summary

This chapter introduced and expanded on the problem of fraud in land record systems. It was found that property fraud is a serious problem in many situations. It was also found that few methods are followed in order to track fraud activities. The reviewed methods use different fraud indicators in order to point out fraud cases. However, it was found that the methods examined are not automated and require a lot of manual examination by experts.

Section 2.2 provided a definition of the term *land record systems* as it is used in this thesis. Section 2.3 examined the current processes used to analyse the data stored in

land record systems and provided examples of those processes. Those two sections fulfilled the answers to the first two research questions listed in Section 1.5.

The chapter then addressed the identified fraudulent activities in two scenarios: fraud that takes place in post-conflict situations and fraud that takes place in real estate transactions in developed countries. Some fraud patterns and indicators are identified in this chapter to introduce the fraud problem. In Chapter 3, more focus is set on the fraud schemes, patterns and indicators which have been collected from the literature and personal communications. At the end of this chapter, two studies that have been found to address the problem of fraud in real estate using data mining approaches were discussed. This discussion is provided in the last section, Section 2.6, which contributes to the understanding of 1.5.2 of the research questions and provides an answer to 1.5.3.

This chapter contributed to the achievement of the main objective of this thesis by identifying the problem of fraud in land record systems (sub-objective a, Section 1.4) and providing the answer to many of the research questions. It mainly represents the achievement of the first two activities listed in the research methods in Section 1.6.

Chapter Three: Fraud Schemes and Indicators in Land Record Systems

3.1 Introduction

It was established in Chapter 2 that land transaction fraud is a significant problem in many countries around the world. Technological advances have contributed to significant improvements in operational efficiency in land registration. However, they have also enabled increasingly sophisticated scams (CISC, 2007). In this chapter, addressing sub-objective 1.4.a, the author describes some of the methods fraudsters use in real estate transactions in stable land markets and in post-conflict situations.

This chapter identifies a variety of methods that are available to fraudsters who attempt to profit illegally from land transaction fraud. Favoured methods in a particular jurisdiction depend on the type of registration system and the local land administration environment. Thus, each situation tends to be unique, but certain general patterns are identifiable.

In an initial analysis of the different methods used by fraudsters, key definitions were established that aided both in the analysis of those methods and in the application of detection data mining techniques that discover fraud. Three operational definitions of the important concepts have been developed: a fraud indicator, a fraud scheme, and a fraud pattern.

In this study, a fraud indicator refers to a pointer to a fraud that might have taken place. It reflects a certain action such as obtaining a high-ratio mortgage or the repetition of a person's name on consecutive transactions on the same property. A fraud indicator

may cause an attribute in the corresponding database to cross a certain threshold or to have a certain value that could raise suspicions about the object the attribute describes. However, this is not always the case. Sometimes, the existence of an indicator along with other indicators is what actually raises the suspicions about a property or a piece of land. This means a number of variables together may indicate a fraud, creating a multivariate modelling problem.

A fraud scheme is defined as a series of steps or actions used by fraudsters to commit a fraud. Each of the actions in a scheme is a fraud indicator, and different schemes may share certain indicators. Despite the thought that each action in a scheme may be reflected as an indicator or a group of indicators, not all those indicators can be addressed in the context of this study. The reason is that in most of the schemes, there are steps that affect systems other than land record systems. For example, forged financial or taxation records may be used in income fraud, as discussed later in Section 3.2.3. These records are not part of a land record system but can be used to investigate property fraud.

Finally, a fraud pattern refers to the effects of executing a certain fraud scheme on a record or a group of records in a land records dataset. It is these patterns that the author is looking for to decide if a fraud has occurred. A pattern may be reflected by the effects of a scheme on the values of transaction attributes or by certain correlations between some of the attributes.

The chapter is organized as follows. First in Section 3.2, the author examines a number of fraud schemes with a brief discussion about each scheme. The chapter goes into more detail about the fraud schemes that the author used in the experimental work.

This includes the Oklahoma Flip scheme, and ABC-Construction schemes in Section 3.2.5 and some of the patterns of land grabbing in post-conflict situations in Section 3.4. The patterns and indicators of the three schemes are discussed and analysed to set the bases for the experimental work described in Sections 3.3 and 3.4.

3.2 Fraud Schemes in Real Estate Transactions

Fraud schemes may be classified according to the method or the purpose of the schemes. In a personal communication, NE (2010, pers. comm., 3 March) disclosed that based on the method, fraud can be divided into two categories, fraud by forgery and fraud by impersonation. However, during this study, the author also realized that some frauds may include the abuse of weaknesses and limitations in the registration systems, or the abuse of power in societies that are not open.

Based on the purpose or the motivation of the fraud, two categories are described by Bianco (2008). The first one is fraud for housing, which is derived from the need of the fraudster to find a place to live in. The second one is fraud for profit. Another category, found in CISC (2007), is fraud to further other criminal activities such as marijuana growing operations.

The following sub-sections provide a brief description for some of the common racketeering schemes. Sections 3.2.1 to 3.2.4 introduce some of the schemes. In Section 3.2.5 and its sub-sections, the two mortgage fraud schemes used in the experimental work are examined in detail.

3.2.1 Impersonation Fraud

This is a general scheme that occurs when a fraudster impersonates the true owner (perhaps having stolen their identity documents), sells the home or takes out a mortgage, and then disappears. Our study indicates that in some cases a family member impersonates the owner in the belief that the owner will not prosecute a member of their own family (OC 2010, pers. comm., 25 March). In other cases The Law Society of Upper Canada (2004) reports that fraudsters may appropriate the identity of a strange owner, steal a corporate identity, or appropriate a lawyer's identity.

3.2.2 Occupancy Fraud

This type of fraud involves misrepresentation in a mortgage application to a financial institution. In a mortgage application, the borrower states that the purpose of buying a property is to occupy it as the primary residence or a second home, while the real intention is to purchase the property as an investment. The borrower, if undetected, will often obtain a lower interest rate than allowed for an investment property. Also, lenders may authorize larger loans on owner-occupied homes compared to loans for investment properties. In addition, the owner may also attempt to avoid a capital gains tax on the property (Maggio, 2008, p. 194).

3.2.3 Income Fraud and Employment Fraud

Income fraud also involves misrepresentation to a financial institution. It occurs when a borrower overstates his or her income to qualify for a larger mortgage than the bank

would ordinarily issue to the applicant. These are commonly known as “stated income” mortgage loans or “liar loans”. To accomplish this, the borrower may forge or alter tax returns and bank accounts which show an inflated income (Bourn, 2006).

Employment fraud is a special case of income fraud, where the borrower claims self-employment in a non-existent company or claims a higher position than they actually occupy in a real company (Bourn, 2006).

3.2.4 Air Loans

Air loans involve obtaining a loan on a property that does not exist; for example, by using a non-existent address. The racketeer creates a fictitious property in a realty listing. Then an accomplice acts as a buyer for that property and obtains a mortgage on it. So, the fictitious realty listing can be used to persuade a financial institution to issue a mortgage. The racketeer(s) then disappears with the cash (CISC, 2007).

3.2.5 Appraisal Fraud, Property Flipping and Property Inflation Schemes

Appraisal fraud involves deliberately misstating or inflating the value of a property. In appraisal fraud schemes an appraiser often colludes with a racketeer to overstate or understate the property value. When a property value is overstated, a larger loan can be obtained; or in a case of selling the property, a buyer will pay more than the property is actually worth (CISC, 2007).

Some criminal groups may state that a property is much larger and/or newer than other properties in the same area, or that a property was recently renovated. These claims

will allow them a large mortgage. In the event of a foreclosure the lender may not be able to recover the value of the loan from the sale in execution of debt. Understated values are primarily used to get a lower price on a foreclosed home (CISC, 2007; Law Society of Upper Canada, 2004).

Property flipping and property inflation are special forms of appraisal fraud. Property inflation includes different schemes with the sole purpose of illegally inflating property prices to deceive the lender or a prospective buyer. A widespread method is property flipping. Property flipping involves purchasing a property and then artificially inflating its value by moving it back and forth between a group of people. Sometimes identity theft, straw borrowers and industry insiders are used in these schemes (Financial Crimes Enforcement Network, 2006; CISC, 2007; and Unger *et al.*, 2010). According to the Financial Crimes Enforcement Network (2006) after several flips, the property may be resold at a price that is 50 to 100 percent higher than the original cost to the syndicate, the outcome of which is a significant loss for a financial institution.

Two schemes are most commonly used for unethical or illegal property inflation; these are known as Oklahoma Flip and ABC-Construction. After the execution of these schemes the mortgagee may provide a loan significantly larger than the real property value justifies (WG 2010, pers. comm., 14 June; Unger *et al.*, 2010; Auditor General of Alberta 2010; CTV News, 2005). Further description of these two schemes follows in Sections 3.2.5.1 and 3.2.5.2. Chapter 5 and Chapter 6 deal with simulating datasets which include patterns typical of what was found in these schemes, and use data mining to identify these patterns in a dataset.

3.2.5.1 ABC-Construction

This is a scheme that is widely used for money laundering or to generate a quick profit. It includes the inflation of property price by selling it back and forth between two (or more) colluding persons *A* and *B* before selling it to person *C*. If *C* fails to obtain an independent appraisal he/she will pay an overinflated price (Unger *et al*., 2010).

The basic steps of the scheme are as follows. As shown in Figure 3.1, person *A* inflates the price of his/her property before a final sale takes place, by selling it to a colluder *B*. *A* and *B* may sell the property to each other a number of times using various aliases. The goal is to get the value of the property as high as possible in a short period of time. An unsuspecting buyer *C* will then buy the property at a very high price, as the conveyancing attorney will show *C* the last purchase price. The attorney may be a party to the scheme (Unger *et al.*, 2010).

The scheme relies on *C* not doing a proper inspection of the property. Often *C* is an out-of-town buyer. The scheme works in a buoyant market when prices are rising and real estate agents don't have time to appraise every single property properly (Unger *et al.,* 2010).

It is noted in Ferwerda *et al.* (2007), cited by Unger *et al.* (2010), that ABC-Construction schemes are legal if the transactions are transparent and according to the law. However, it is reported that this scheme is commonly used illegally for profit or money laundering.

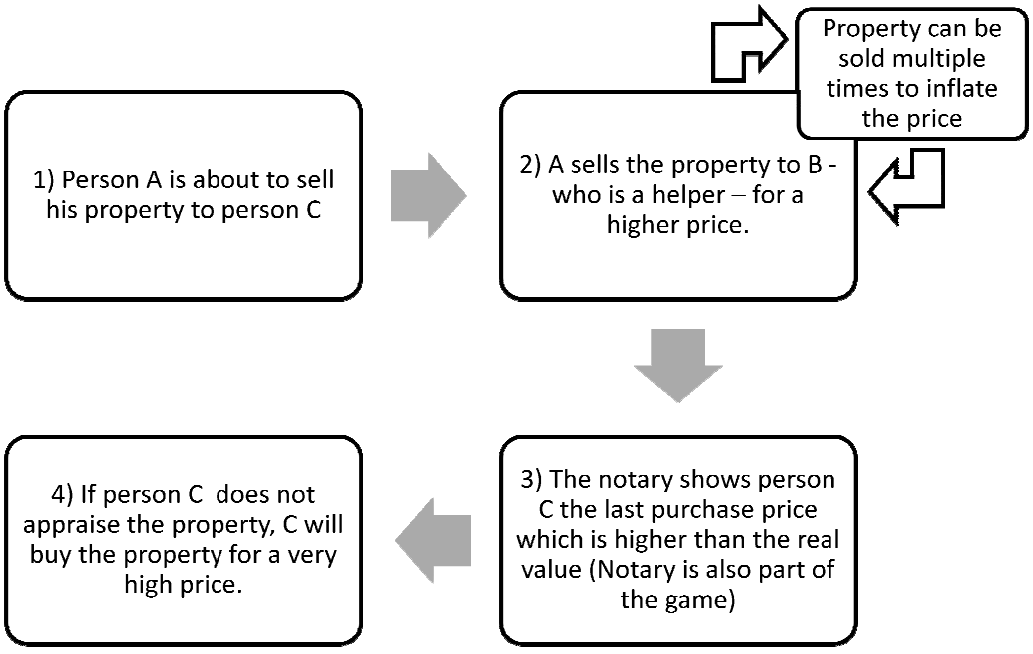


Figure 3.1: Steps for conducting the ABC-Construction fraud scheme

One striking case occurred in the Netherlands. The "Bureau Financieel Toezicht" (a financial watchdog bureau) – which monitors the work of notaries in The Netherlands – discovered that the building in which it is housed was part of an illegal ABC-Construction scheme. It was reported that the former director of "Bouwfonds", who was the prime suspect in the case, made 2.5 million euros on the deal in one day (Kreling and Meeus, 2008).

3.2.5.2 Oklahoma Flip

The Oklahoma Flip, as with the ABC-Construction scheme, inflates a property price. However, the profit in this scheme is gained by obtaining a mortgage on the property from an unsuspecting financial institution rather than selling it to an unsuspecting buyer.

In simple terms, the Oklahoma Flip is about buying a cheap, sometimes rundown property, flipping it several times to inflate its value, and then obtaining a mortgage on the property and running with the proceeds. The inflation happens by selling the property back and forth between the con man and his/her co-conspirators or a company the con man controls (Unger *et al.*, 2010). In general, no money changes hands in the sales, but it allows the con man or syndicate to inflate the value of the house (WG 2010, pers. comm., 14 June).

When the value of the property is inflated, a final transaction takes place wherein the racketeers obtain a mortgage for well over the market value of the property, and then disappears. In some cases, a straw buyer or an unsuspecting intermediary is used in the final transaction. A straw man is "an individual who has no financial or other interest in the property and is recruited and offered a nominal fee solely to allow their name and credit rating to be used to obtain a mortgage from an unsuspecting lender" (Auditor General of Alberta 2010, p. 105). The straw man obtains a mortgage on the property, then everyone shares the proceeds of the mortgage and the straw man defaults on the mortgage, leaving the financial institution with a significant loss (CTV News, 2005; and Unger *et al.*, 2010). Figure 3.2 summarizes the steps common to the Oklahoma Flip.

The Oklahoma Flip is used mainly for quick profit. It is a scheme based on quick inflation of property value through multiple flips in a short period of time, after which a mortgage is taken against the property. According to WG (2010, pers. comm., 14 June), the periods between transactions may be as short as a week. Many players can take part

in executing this scheme, including the seller or the owner of the property, friends of the seller, a straw buyer, and lawyers.

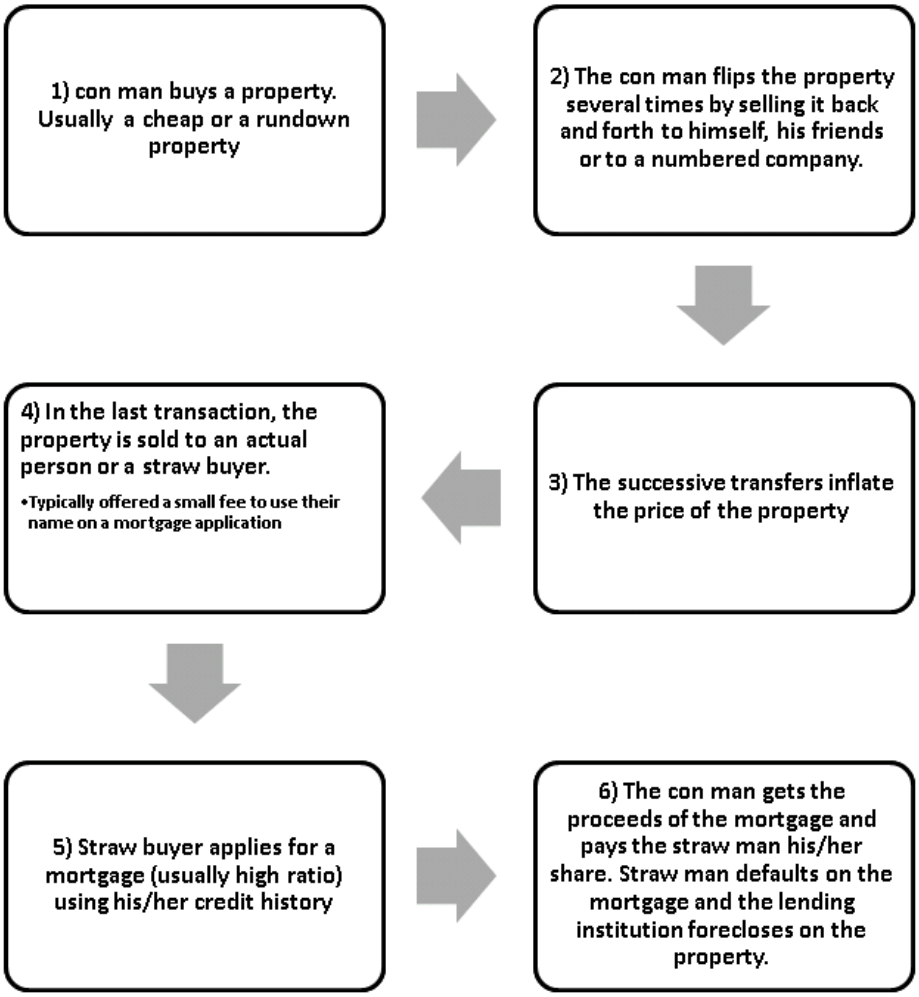


Figure 3.2: Steps for conducting the Oklahoma Flip fraud scheme

3.3 Fraud Indicators and Patterns of the Oklahoma Flip and ABC-Construction

This section describes methods of identifying Oklahoma Flip and ABC-Construction schemes. To do so, the author first developed a list of indicators of the principal activities underlying the two schemes. Then, the indicators were grouped into patterns where each

pattern describes the execution of a complete scheme. These patterns were used to develop the data mining methods described in Chapter 6.

The following five points are general fraud indicators that are common in most of the appraisal frauds:

1. An unusual number of property transactions by one seller, especially in a relatively short period of time. This means that one person has executed an extraordinary number of transactions in a short period of time (Auditor General of Alberta 2010; and WG 2010, pers. comm., 14 June). Any number more than one transaction within a year or even two years can be considered suspicious. However, it is not a definite proof that a fraudulent activity has occurred, since the involved person in those transactions might be a real estate agent or an investor.

   This indicator is a pointer that the person might be suspicious and is not a direct indicator for a suspicious property. However, properties that are part of the transactional activities this person is involved with should be considered suspicious properties (Unger et al 2010, Nelen 2008).

2. Properties that change hands quickly between owners (Unger *et al*., 2010; Nelen, 2008). The indicator here is the period between consecutive transactions on the same properties. In a normal situation, flipping periods for properties are long and a property might stay with the same owner for more than 20 years. It is not specifically mentioned how short a flipping period must be to constitute

fast flipping. However, an average of five years between flips on the same property can still be considered normal (OC 2010, pers. comm., 25 March).

3. The financing method of the purchase can be considered an indicator in different ways. If a property is purchased without a mortgage, this can be considered an indicator, since real estate objects are known to be the most expensive assets for people (Unger *et al*., 2010; and Nelen, 2008), and buying one without any financial assistance is an exceptional behaviour. Also, the existence of high-ratio mortgage can be considered an indicator, especially when other indicators are available (WG 2010, pers. comm., 14 June).
4. Unusual fluctuations in a particular property's price can also be an indicator. In most of the cases, this is an unusual rise in the price which is exceptional relative to the current market and to neighbouring property prices (WG 2010, pers. comm., 14 June, Auditor General of Alberta, 2010). Also, an unusual drop in the price might be considered as a pointer. An example of this case is when a criminal group buys a property, uses it for growing marijuana, and then sells the property (CISC, 2007).
5. Foreign ownership may be an indication of money laundering. On its own, this is not a strong indicator. However, foreign ownership adds weight to a suggestion of racketeering if other indicators are present (Unger *et al*., 2010; and Nelen, 2008).

More fraud indicators can be found in the literature, but for this thesis, the number of indicators used was limited. Unger *et al.* (2010) use a set of 25 indicators in their study

and categorise those indicators into five different groups: indicators related to the financer, indicators related to the financing method, indicators related to the owner, indicators related to the real estate object, and finally, indicators related to the purchase sum. Using such a wide range of indicators would provide higher accuracy in detecting frauds than using a subset. The problem, however, raised in this study is the availability of data. To be able to consider this wide a range of indicators, it is required to have access to different data sources. This was not possible during the development of this thesis, which is an issue discussed in Chapter 5.

Unger *et al.* (2010) use objective data related to real estate objects to build a prediction model using the collected indicators in order to identify objects that might be involved in criminal activities. As discussed in Chapter 2, they use a flagging method to develop their prediction model. In this thesis, the author uses a different approach to build a prediction model for properties based on fewer indicators. Basically, the author uses a set of the most important indicators that can be found in a normal registration system and that can point out exceptional behaviours which are considered fraud.

In chapter 6, a classification model will be developed to identify properties that might have been part of an Oklahoma Flip or an ABC-Construction scheme. This model will be based on 5 fraud indicators:

1. An unusual number of transactions taking place on the same property as fraudsters flip the property back and forth between themselves. The unusual number is anything more than one transaction during a short period of time. The definition of a short period is not clear. In an interview, WG (2010, pers.

comm., 14 June) mentioned that once three or more transactions can be seen per year, this should raise suspicion. He also mentioned that, when looking for Oklahoma Flip or ABC-Construction during the screening process, the investigation department where WG works uses a time frame of two years and looks at transactions that took place during this time frame.

So, in building the classification model, a time frame of two years is used and the model is built based on the transactions that took place during this period. In the evaluation criterion for this indicator, two or more transactions on the same property will result in the property being considered for the analysis and for building the model.

2. A repetition of a name over multiple transactions on the same property. This happens as one person sells a property and then buys it again and sells it a second time, each time using another friend as the second party in the transaction. This may be repeated a number of times and may include two persons or more. In general, the number of different parties truly involved in the transactions on the same property will be less than the number of transactions which took place.
3. An unusual increase in the price of the property. This may happen at two levels. The first is the unusual overall increase of the price in a relatively short period of time. It could be pointed out by comparing the increase of the targeted property with the increase in other neighbouring properties over the same period of time. The second level is a high increment in the price between

any two consecutive transactions, which does not correspond to the appraised value. This can occur because these are not arm's length transactions and no money changes hands between fraudsters. Unger *et al.* (2010) state that this is one of the most visible indicators. One case mentioned in Unger *et al.* (2010) is a case of a building in Ukraine which was purchased for a price that was 10 times higher than the purchase price of three days earlier.

4. A short flip period of a particular property. This indicator is different from the first one. It pertains to the periods between each two successive transfers of the same property, whereas the first indicator is related to the number of transactions within a time frame.

   The shorter the flip period, the more conspicuous the property is. A short period could be as short as a week or two (WG 2010, pers. comm., 14 June); however, as mentioned before, there is no clear separation between what is considered a short period and what is not. WG also mentioned that the period between suspicious flips could go up to 3 or 4 months.

5. The existence of a high-ratio mortgage on the property after the last transaction. The attribute that describes this indicator is defined as Loan to Value Ratio (LTVR), which will be further discussed in Section 5.3.2. Based on information from real estate agents and listing services in Canada, a 90% LTVR is considered high and rare, whereas the average of LTVR is around 80% and is dependent on the country. These values are taken into consideration when working with this indicator as a pointer for fraud.

Those were the five indicators that have been taken into consideration when analyzing fraud patterns of Oklahoma Flip and ABC-Construction schemes in this study. Using those indicators, and building on the two fraud schemes under study, fraud patterns were constructed to help in identifying fraud cases from a dataset of real estate transactions.

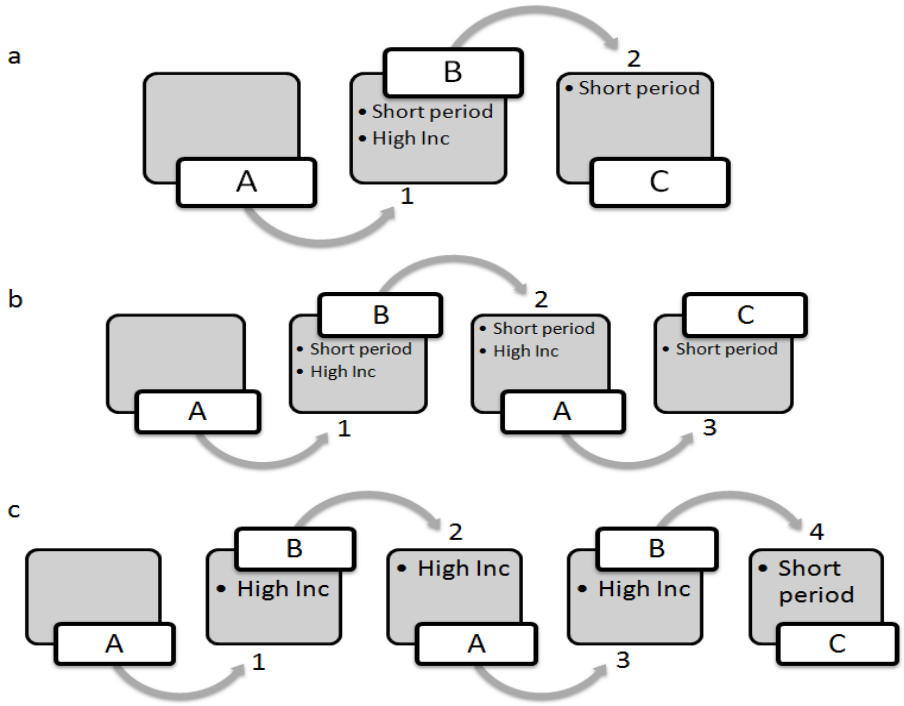


Figure 3.3: Some patterns of the ABC-Construction fraud scheme

Figure 3.3 and Figure 3.4 show some of the fraud patterns that would lead to a suspicious property. Each of the patterns represents a series of transactions on one property. In the figures, an arrow represents the direction of a transaction, the solid white rectangle contains the name of the buyer or the seller, and the gray shaded rectangle contains apparent fraud indicators in each pattern. These indicators might be one or a

combination of three indicators: a short period between a current transaction and the previous one, a high increase of the value of the property, and a high LTVR. Finally, the number of arrows in each diagram represents the number of transactions that took place on a property.

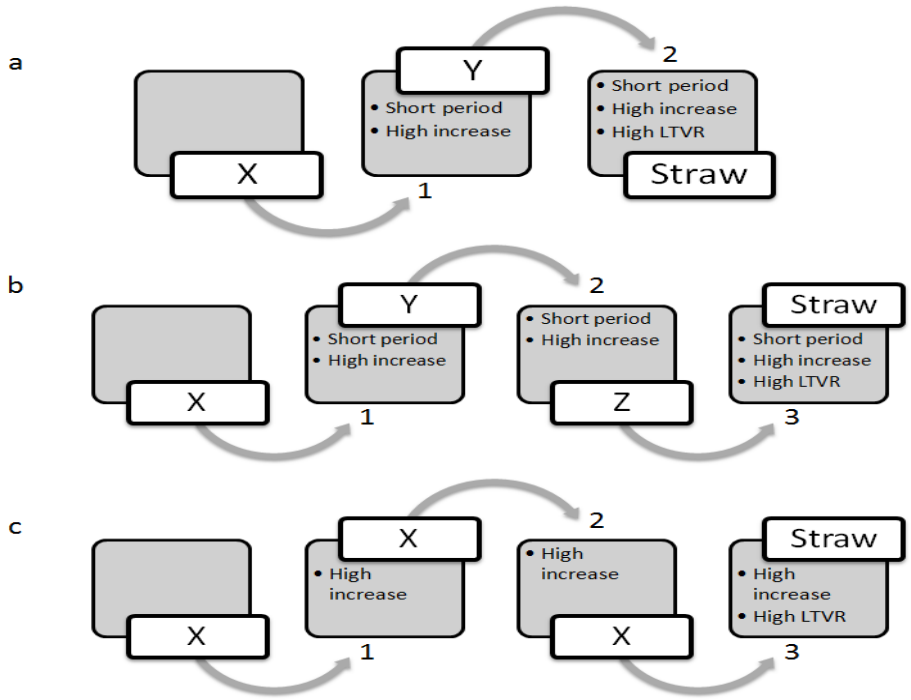


Figure 3.4: Some patterns of Oklahoma Flip fraud scheme

Figure 3.3 illustrates three different patterns which can be considered general cases of the ABC-Construction scheme. For example, in Figure 3.3.c, person *A* sells a property to person *B*. *B* then sells it to *A* who sells it to *B* again. All three transactions share an indicator, which is a high increase in the value of the property between any two successive transactions. The final transaction is a sale from *B* to C who is not part of the

con as *A* and *B*. *C* buys the property for a very high price just shortly after the last transaction between *A* and *B*.

Figure 3.4 illustrates three patterns of the Oklahoma Flip. These patterns are similar to the patterns of ABC-Construction; the main difference is the high-ratio mortgage obtained by the straw buyer in the last purchase.

Patterns of Oklahoma Flip and ABC-Construction are not limited to the patterns portrayed in Figure 3.3and Figure 3.4, and there are many more patterns that can be considered part of these two schemes. These patterns are used in Chapter 5 to construct classification rules that help in building the property classification model.

3.4 Fraud Schemes and Indicators in Post-Conflict Situations

Looking for fraud patterns in post-conflict situations, the only patterns that the author was able to identify are patterns of land grabbing. The problem of land grabbing was discussed previously in Section 2.6.1, with two groups of indicators listed: indicators derived from the number of transactions taking place, and indicators derived from parties involved in transactional activities in addition to the direction of these transactions.

Only the first group is addressed in the experimental work in Chapter 6. More specifically, as mentioned in Zevenbergen and van der Molen (2004), during or after a conflict, an investigator may find periods with an exceptionally high or low number of transactions. The measure of high or low is relative to the number of transactions in a normal case even during or after a conflict. The general behavior is that powerful groups will grab the land from weak ones or other individuals.

Most of the grabbing activities will take place during short periods of the conflict itself, and that is what will cause jumps in the number of transactions during this period. In other cases, powerful individuals may get access to the registration system of a particular jurisdiction and remove transactions, which will cause a high drop in the number of transactions for that period.

The aforementioned two cases represent the patterns addressed in the experimentation of fraud detection in Chapter 6, where outlier detection methods are employed in order to detect identified manipulations.

3.5 Chapter Summary

This chapter concluded the review and the analysis of the different fraud schemes identified in this study. It examined a range of fraud schemes and then put the focus on the schemes used in this study for the experimental work. Namely, the Oklahoma Flip and ABC-Construction schemes were examined in detail. For those schemes, the chapter elaborated on the fraud indicators and patterns that are used to detect the schemes in the experimental work. In addition, land grabbing patterns that will be used in the experimental work were reviewed.

The chapter established five different indicators that are used as a base to build a classification model for suspicious properties. These indicators are used to simulate fraudulent land records activities, as will be discussed in Section 5.3. The building of a classification model to identify suspicious properties is discussed in detail in Section 6.2.

The chapter also established the sole indicator that will be used for detecting fraudulent activities in post-conflict situations. This indicator is the exceptional number of transactions taking place in a certain period of time. This indicator is used for data simulation in Section 5.3 and is what the author is looking for in the outlier detection experiments discussed in Section 6.3.

The identified fraud indicators and patterns contribute to the achievement of sub-objective 1.4.a. This chapter also helps in answering research Questions 4 and 5 stated in Section 1.5.

Chapter Four: Data Mining Methods

4.1 Introduction

This chapter presents the second part of the literature review. Chapter 2 presented the first part – the problem of fraud in land records and some of the methods used to target this problem. This chapter reviews data mining methods used in this study to identify certain fraud schemes that were discussed in Chapter 2 and Chapter 3.

PDA and Classification and Regression Trees (CART) methods are adopted for the problem of detecting suspicious activities in real estate transactions. Entropy-based outlier detection is adopted for the problem of detecting land grabbing patterns in post-conflict situations.

The chapter proceeds as follows. Section 4.2 reviews methods and applications of data mining. Section 4.3 focuses on classification methods and then goes into detail about PDA, decision trees, and methods of classifier evaluation. Section 4.4 reviews outlier detection methods with a focus on entropy-based outlier detection. The adopted data mining methods are justified in Section 4.5. Finally, Section 4.6 summarises this chapter.

4.2 Introduction to Data Mining

The term *data mining* refers to the process that involves the automatic extraction of useful information (knowledge) from large data repositories (Han and Kamber, 2006; and Dunham, 2003). It is a multi-disciplinary field which draws on statistics, databases,

information retrieval, information extraction, machine learning, artificial intelligence, and visualization.

In Bramer (2007), data mining (sometimes referred to as Knowledge Discovery from Databases) is divided into two main branches: supervised and unsupervised learning techniques. Supervised learning includes classification and numerical prediction. Data mining based on clustering and association rules falls in the unsupervised learning category. Witten and Frank's (2000) list of data mining tasks include: classification, numerical prediction, clustering and association rules mining, as well as estimation, and outlier detection. In general, all the tasks share the same primary goal, which is to extract knowledge from the input data.

Supervised learning techniques require previous knowledge of the different classes (groups) that separate the instances, while unsupervised learning techniques do not require this knowledge. An example of supervised techniques is classification, which is considered one of the most common supervised learning techniques (Barmer, 2007). Classification requires the availability of datasets where each instance is assigned to one of the known classes in order to be able to build classification or prediction models (Bramer, 2007). On the other hand, clustering, which is the most common unsupervised learning technique, works by examining the data to find groups of instances that are similar using distance measures in a metric space (Barmer, 2007; and Berkhin, 2006).

Another important technique used in this study is outlier detection. It is most often used in data pre-processing as a step in data cleaning, in order to obtain coherent analysis for the data (Ben-Gal, 2005). An outlier is an observation that is suspicious, as it deviates

significantly from other observations in the dataset (Hawkin, 1980). Consequently, the goal of outlier detection techniques is to discover those points that are dissimilar or exceptional with respect to the rest of the points in a dataset (Li *et al*., 2006 and Angiulli *et al*., 2006). Most outliers are noise or data errors, but there are many cases where they may carry important information, such as fraud cases in credit card transactions (Li *et al*., 2006).

This study uses classification and outlier detection methods because of the nature of the fraud problems in real estate transactions. The fraud indicators and patterns listed in Chapter 3 separates properties into two distinct classes based on their transactional attributes – suspicious and normal. This separation enabled the generation of a training properties dataset, which can be used to build a classification model. One other method that might be also useful for this kind of problem is estimation (multiple regression) which is a similar technique to classification. Multiple regression, however, applies when the required output is continuous; while classification is used when the required output is categorical, such as in our problem.

Outlier detection may identify fraud cases in land in the case of a post-conflict situation. The problem of land grabbing and stealing in a post-conflict situation should show up as outliers from the rest of the data points.

In the following two Sections 4.3 and 4.4, the tasks of classification and outlier detection are further reviewed. Each task is first reviewed in general and the discussion moves to the specific methods used for the experiments.

4.3 Classification

In general, classification involves dividing up objects in a way that each object is assigned to one of a number of categories known as classes or groups. This assignment is done under the condition that an object falls into one class only (Bramer, 2007, p. 23). Specifically, classification is "the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class is unknown" (Han and Kamber, 2006, p. 18). A classification model is developed by analysing a training set. A training set is generally a labelled sample of the main data used to construct a classification model

There are many different classification methods and models. These include decision trees, discriminant analysis, neural networks, Naive Bayes classifiers, k-nearest neighbour classification and support vector machines (SVM) (Tan *et al.*, 2005). Decision trees and discriminant analysis were chosen based on comparative studies for the different methods and as a proof of concept. However, the author is aware that other methods may apply to the same problem addressed in the experimental work. Section 4.5 provides a justification of the methods used.

Dudoit *et al.* (2002) apply different discrimination methods to classify tumours using gene expression profiles. They compare the performance of three methods including nearest-neighbour, linear PDA, and classification trees. In their main conclusion, and based on the datasets they have, Dudoit *et al.* (2002) find that linear PDA and nearest-neighbour classifiers perform much better than classification trees. As will be

illustrated in Chapter 6, PDA performed better than classification trees in the classification of suspicious properties.

Nogueira *et al.* (2005) also perform a comparison between two classification methods. They use PDA and artificial neural networks to classify internet user into 3 groups. The result of their comparison shows that PDA outperforms neural networks. They also conclude that DA is easier to use and its results have simpler interpretation. Nogueira *et al.* (2005) use the Fisher procedure to establish a linear discriminant function (FLDA) that maximizes the ratio of between-group sum of squares and within-group sum of squares *(see appendix C for equations)*. The final results of the application of FLDA in Nogueira *et al.* (2005) show that the model fits the data and is expected to perform well for unseen records.

4.3.1 Predictive Discriminant Analysis (PDA)

In general, discriminant analysis addresses the problem of the extent to which two or more groups of individuals can be separated. This separation is based on available measurements of individuals on several variables (Manly, 2005).

Most of the research done in discriminant analysis in its early years focused on the membership prediction of studied objects, which is referred to as Predictive Discriminant Analysis (PDA). PDA is used when a single set of response variables, which are the attributes that describe an object, is the predictor of one grouping variable. A grouping variable determines the group to which an instance belongs. The question is how well group membership of objects may be predicted. Another branch of discriminant

analysis is Descriptive Discriminant Analysis (DDA), which appeared after PDA to focus on grouping variable effects on response variables (Huberty and Olejnik, 2006).

Prediction itself is used widely in today's applications. The processes of predicting life expectancy, economic growth, academic achievement and sales revenues are examples of those applications. In all cases, the goal is to predict a variable by studying previous behaviours. This concept is applied in this research as a solution for two of the identified fraud problems, namely the Oklahoma Flip and the ABC-Construction scheme. The hypothesis is that PDA may predict if a single property should be classified as suspicious or not based on historical transactional behaviours.

Huberty and Olejnik (2006) divide prediction into two types based on the scale of measuring the outcome variable. For instance, if the outcome variable is quantitative, such as a ratio scale or an ordinal scale, prediction should be conducted using multiple regression analysis. On the other hand, if the outcome variable is categorical and measured with a nominal scale, PDA is the appropriate method.

To be able to predict the group of an instance, a classification model should first be established. To build a model, the goal of PDA is to develop a prediction rule involving as many composites (linear or quadratic) of predictors as the number of groups. Each of the composites is associated with one of the groups and is used to predict the pertinence of an observation to that group. When a new object needs to be assigned to one of the groups, it is evaluated using all the composites and then based on the scores, it is assigned to one of the groups (Huberty and Olejnik, 2006).

It is suggested by Huberty and Olejnik (2006) to do some checks before applying PDA. Some of the recommended checks are correlation measures and the univariate Analysis of Variance (ANOVA) *F*-test. This analysis will identify if the relevant predictors are related to the grouping variable. Composites should be formed using the relevant predictors for a better accuracy of the classification rule.

The *F*-test in ANOVA is used to assess whether the expected values of a predictor variable within several pre-defined groups differ from each other. Basically, it measures the ratio of the between-group variability and the within-group variability. So, if *F* is large, it means that the between-group variability is much larger than the within-group variability (Manly, 2004). This indicates that actual means of the predictor for the different groups are different and this predictor can be used to distinguish between the different groups *(see appendix C for F-statistic equations)*.

If it is found, from the correlation analysis, that two predictors are highly correlated, one of them should be dropped as the predictors should be independent of each other. Also, if it is found through the *F*- test that a variable contributes only noise (*F* value less than 1), it is recommended that the variable be dropped (Huberty and Olejnik, 2006).

It is also recommended by Huberty and Olejnik (2006) that in order to develop a good model, the calibration data should be representative of the population data. This means that proportions of the different groups in the sample reflect the actual proportions in the original population. All these checks and recommendations were applied in the

PDA experiments described in Chapter 6. *Appendix E* provides a brief background of the classification and decision rules of PDA.

4.3.2 Classification Using Decision Trees

Decision trees are used in decision support to map decisions to their possible consequences. In data mining, decision trees are employed for prediction by mapping records to conclusions based on selected features. Matthew and Shiva (2009) state that decision trees are most commonly used for classification since they are easily understood and are based on simple modelling techniques that simplify the classification process.

Two types of decision trees are established in data mining: classification trees, used when predicting a nominal or categorical variable which has no numerical values; and regression trees, used to predict a quantitative or continuous variable (Mirkin, 2011). As discussed in Section 4.2, the required output in this study is categorical and so the focus is on classification trees.

Figure 4.1 shows a simple classification tree structure that classifies instances into two classes based on three features. Each node in a classification tree embodies a subset or a cluster of instances from the total population. The root of a tree embodies the entire population. The children of a node are subsets of the set represented in that node, so the deeper you move in a tree the smaller the subsets become. A node in a decision tree splits over a test condition for one feature to form either new child nodes or reach a class label node (leaf node). In most of algorithms used in building classification trees, only binary

splits are considered both in categorical and quantitative features, in order to make the partitions less arbitrary (Barmer, 2007; and Mirkin, 2011).

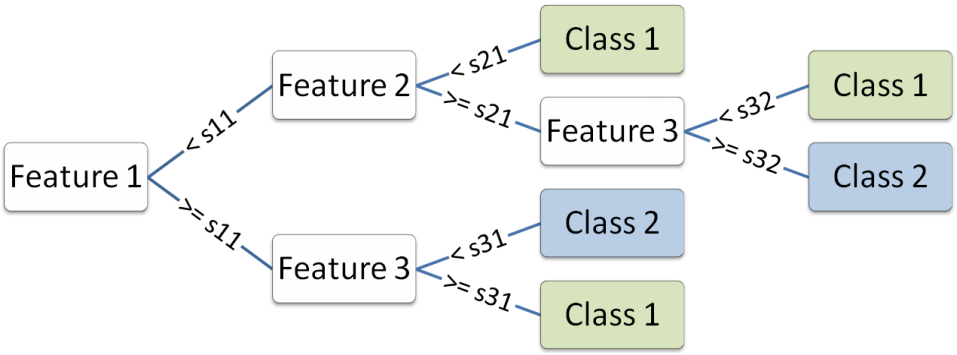


Figure 4.1: A classification tree that classifies instances into two distinct groups based on three features.

Classifying a new instance is straightforward once a decision tree has been constructed. Starting at the root node, the test condition is applied to the instance (e.g., a land parcel). Based on the outcome of the test, this instance follows the appropriate branch. The outcome of a test condition will lead either to another internal node, where the test condition of that node is evaluated against the instance, or to a leaf node. When reaching a leaf node, the class label associated with it is assigned to the observation.

Many decision trees can be constructed from a given set of attributes; some are more accurate than others. The problem of finding the optimal tree is computationally infeasible because of the exponential size of the search space (Tan *et al.*, 2005). These problems are addressed by tree induction.

4.3.2.1 Induction of classification trees

The process of constructing a classification tree from a training set is called induction of the tree. Most of the existing induction algorithms are based on Hunt's TDIDT (Top-Down Induction of Decision Trees) algorithm (Tan *et al.*, 2005). In this algorithm, a decision tree is constructed recursively by splitting the training set into successively purer subsets (Bramer, 2007; Tan *et al.*, 2005). The purity of the subsets is measured by impurity measures. The most common impurity measures are Entropy, Gini Index, and misclassification errors (Tan *et al.*, 2005). These measures are used to evaluate the goodness of a split during the induction process by measuring the homogeneity of the subsets.

Hunt's TDIDT algorithm is a well-established method for tree induction. However, it does not specify the attribute to select for each split. So, many of the algorithms based on Hunt's algorithm try to enhance the attribute selection mechanism (Bramer, 2007). Some of the most famous induction algorithms based on Hunt's algorithm include ID3, CART, C4.5, SLIQ, and SPRINT (Anyanwu and Shiva, 2009).

Anyanwu and Shiva (2009) perform a comparative study for the different decision tree induction algorithms that include ID3, C4.5, CART, SLIQ and SPRINT. The results of their study basically show that there is no significant difference in accuracy when different algorithms are used.

Mingers (1989) also performs an empirical comparison of a number of methods and strategies for the creation of a decision tree. Methods included in the study are mainly used for attribute selection for the split process. The reported results from

Mingers' study show that prediction accuracy is not sensitive to the goodness of split measure. Furthermore, the results of the study show that accuracy is not improved significantly by using a split measure at all. This means that the induction of a tree using randomly selected attributes for the splits is likely to yield a similar accuracy to one using a split measure. Mingers (1989) shows that applying a certain measure affects only the size of the tree and does not have a significant effect on the accuracy.

4.3.3 Evaluation of Classifiers

Assuming that a classification model was constructed, it is important to assess the model and evaluate its performance. To do this, and regardless of the method used to construct the model (PDA, classification trees, SVM, etc.), one can measure two types of errors; training error (also known as resubstitution error), and generalization error. Tan *et al.* (2005) provide a definition for the two types. Training error is a measure of the number of misclassification cases committed on a training set. On the other hand, generalization error is the expected misclassification rate by the model on unseen records, such as a test set.

It is important to have low measures for both types, as it is important for a model to accurately classify records it has never seen. In some cases, models that fit training data too well tend to have a poorer generalization error, which is called model over-fitting. In some cases, over-fitting is caused by a non-representative training sample (Tan *et al.*, 2005).

In order to estimate the generalization error, a validation set approach is generally employed. In its basic concept, the original dataset in this approach is divided into two smaller subsets; one is used to build a classification model, while the other is used to estimate the generalization error (Tan *et al.*, 2005). Such a measure provides an unbiased evaluation of the model. However, in order to apply this approach, the class labels of the test set must be known.

There are several methods to evaluate the performance of a classifier. The most common are the holdout method, the random sub-sampling method, and the cross-validation method.

4.3.3.1 Holdout

In the basic holdout method, the original labelled data set is divided into two sets, a training set and a test set. The training set is used to build a model, while the test set is used to evaluate the performance and the generalization error. There are two common splitting ratios. First, half of the original set is used for training while the second half is used as a test set. The other approach is to use two-thirds of the original set for training and use the remaining third for testing.

The holdout method has three well-known limitations. First, it reduces the number of records available for training. Second, the model may largely depend on the sample selected for the training, and there will be larger variation of the model as the training sample gets smaller. Finally, if a large sample is used for training, fewer records will be available for testing and the calculated accuracy from it is less reliable (Tan *et al.*, 2005).

4.3.3.2 Random sub-sampling

The random sub-sampling method is conducted by repeating the holdout method several times to obtain a more reliable performance estimate. Each time, a random set of observations is selected as the training set and the remaining set is used for testing. The overall accuracy is measured as the average accuracy of the different holdout iterations. A major limitation is that there is no control over the number of times each record is used for training and testing (Tan *et al.*, 2005).

4.3.3.3 Cross-validation

Finally, the cross-validation method is used as an approach to overcome the limitations of previous methods. This approach uses each record exactly the same number of times for training and only once for testing. This is achieved by dividing a dataset of N labelled records into k equal subsets. Each subset is used as a testing set while the rest of the subsets are used for training. This is repeated for each subset for a total of k runs; the method is called k-fold cross-validation.

A special case of the k-fold method is when $k=N$. In which case, each record in the labelled dataset represents a separate subset. This method is called the leave-one-out (LOO) method. LOO utilizes as many records as possible for training, and also provides test sets that cover the whole dataset (Tan *et al.,* 2005).

4.4 Outlier Detection

Outlier detection is the second method used in this study for the detection of fraudulent activities. The previous section reviewed the task of classification, which is the task used for the first part of the experimental work. This section provides a review for the task of outlier detection and examines some of its methods, focusing on entropy-based outlier detection, which is employed in this study.

In many data mining applications outlier detection is a primary step, usually for data cleaning. In general, outliers in most of those applications are treated as noise or errors, and hence unwanted observations, as they might affect the analysis of a dataset. However, outliers in some cases may carry important information (Ben-Gal, 2005).

Hawkins (1980) gives a general definition for an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Other more specific definitions also define outliers as deviating or inconsistent observations when compared to the sample from which they are taken (Barnett and Lewis, 1994; and Johnson, 1992).

In many applications outliers are proven to present useful information, and so, different methods and applications have been developed for their detection. Some applications of outlier detection include fraud detection (especially in credit card transactions), voting irregularity analysis, data cleansing and prediction of severe weather (Hawkins, 1980; Barnett and Lewis, 1994; Fawcett and Provost, 1997).

Outlier detection is used widely in many applications that include fraud detection, as discussed before. In the field of transactional data and particularly in real estate

transactions, Kontrimas and Vertikas (2006) develop a comparative study of many outlier detection methods, which include distance-based methods and robust regression, in order to track doubtful real estate transactions. To the best of my knowledge, no other studies apply outlier detection methods for fraud detection in property transactions.

4.4.1 Outlier Detection Methods

Ben-Gal (2005) provides a categorization of outlier detection according to two criteria. The first criterion is the method used for detection and the second criterion is the dimensionality of the data. According to the method, outlier detection methods are divided into two groups: parametric (statistical) methods, and non-parametric methods. According to the dimensionality of the data, outlier detection methods are classified into univariate methods and multivariate methods.

In parametric methods, a principal assumption is that there exists a known underlying distribution for the observations, or there are estimates for unknown distribution parameters (Ben-Gal, 2005). With these assumptions, the concept of outlier detection is formulated as identifying observations that fall outside the distribution.

Non-parametric methods are more capable of handling large databases than parametric methods, as noted by Ben-Gal (2005). Non-parametric outlier detection methods are divided into five common classes: *distance-based methods* (Hawkins *et al.*, 2002), *clustering methods* (Barbra and Chen, 2000), *classification methods* (Hu and Sung, 2003), *density-based methods* (Breunig *et al.* 2000), and finally, *spatial methods* that are used for spatial outliers (Ben-Gal, 2005).

In *distance-based methods*, different definitions of distance are established in the literature. For example, Knorr and Ng (1998) define an observation as a distance-based outlier if at least a fraction β of observations in the dataset are further than distance r from that observation. In other definitions, outliers are defined based on the distance from the k^{th} nearest neighbour and not from the whole dataset (Ramaswamy *et al.*, 2000).

In *clustering methods*, small clusters are considered outliers (Barbra and Chen, 2000). In *classification methods*, data is partitioned into outliers and non-outliers using classification models, which are constructed using labelled datasets that are already divided into normal observations and outliers (Hu and Sung, 2003). Finally, in *density-based* approaches, a local outlier factor is assigned to each observation based on its local neighbourhood density, and observations with a high outlying factor are considered outliers.

A new method for outlier detection is proposed in He *et al.* (2005); it uses the concept of entropy in information theory to detect deviated observations inside a dataset. This method is expanded upon in Section 4.4.2 and is adopted in this study.

4.4.2 Entropy-Based Outlier Detection

He *et al.* (2005) and Nogueira *et al.* (2010) used a different methodology to approach the outlier problem in categorical data. He *et al.* (2005, p. 400) argue that "conventional techniques do not handle categorical data in a satisfactory manner and most of the existing techniques lack of a solid theoretical foundation or assume underling

distributions". The concept used in their study is built on the fact that outliers in a dataset will cause an "amount of mess" inside it.

The problem of outlier mining in He *et al.* (2005) is defined as an optimization problem, with the assumption that removing outliers from a dataset will create a dataset that is less disordered. To measure the degree of disorder in data, the researchers used entropy, which in information theory represents the amount of uncertainty attached to a random variable (Shannon, 1948) and hence can be used as a measure of information disorder.

Based on the entropy concept, the optimization problem in He *et al.* (2005) is described as removing a subset of k observations from a dataset, which would lead to minimizing the entropy for the rest of the dataset. This information entropy model for outlier detection is considered as a new method and is further exploited by Jiang *et al.* (2010) as an approach for outlier detection in rough sets.

He *et al.* (2005) and Nogueira *et al.* (2010) apply the entropy method on *lymphography* and *cancer* datasets. (2010). Results obtained in both studies suggest the method is superior to other methods applied on the same datasets; these included Replicator Neural Networks (RNN) and the distance-based method (KNN). Building on these results, this concept is adopted in this study for trying to find fraud cases in land transactions during or after a conflict situation using transactional datasets. The application of entropy-based outlier detection to the fraud problem is discussed later in Chapter 6.

4.5 The adopted methods

In this study, as will be discussed in Chapter 5 and Chapter 6, three different methods were used to detect fraudulent activities in property transactions. Two fraud problems were first identified: the problem of property or mortgage fraud in real estate transactions and the problem of land grabbing in post-conflict situations. The following two sections examine the adopted methods to solve the identified problems.

4.5.1 Problem of Property or Mortgage Fraud in Real Estate Transactions

This problem was formulated as a classification problem by seeking a model to classify properties into one of three different classes (normal, suspicious, and highly suspicions). This formulation was based on Unger *et al.* (2010) and uses the concept of classification in the study of mortgage fraud. So, in this study, two different classification methods are applied. The first method is quadratic PDA and the second method is CART.

The choice of PDA was based on two factors. First, based on the results from the studies of Dudoit *et al.* (2002) and Nogueira *et al.* (2005) discussed in Section 4.3.1, PDA is expected to perform better than the other methods. The second factor is based on the use of PDA by Hunter (2007).

Hunter (2007) performs a study to build a prediction model for three locomotion behaviours of a grizzly bear. During this study, a classification rule has been developed using quadratic PDA. This rule can be applied to classify locomotion behaviour of a grizzly bear into one of three groups (stationary, searching, and walking). The prediction model developed in Hunter (2007) achieved a classification accuracy of 0.846. The

property classification problem formulated in this study is treated in the same way that Hunter (2007) formulates the problem of predicting the locomotion behaviour of a grizzly bear.

Another classification method is used for the same purpose in order to validate the results obtained from the quadratic PDA model. The validation method was chosen to be CART, which the author uses to build a classification tree for the same three groups used in PDA.

This choice of CART is made based on results obtained from the studies of Mingers (1989) and Anyanwu and Shiva (2009). The reported classification accuracy of the compared algorithms in Anyanwu and Shiva (2009) and comparison of different split measures compared in Mingers (1989) show that different algorithms perform with relatively equal accuracies and so CART is expected to perform similarly to the other algorithms when applied to the properties dataset.

For the hit rate estimates and performance evaluation of both classification methods, mainly the LOO method is used. This method utilizes all records for training and it also uses each record in the dataset for testing.

4.5.2 Problem of Land Grabbing in Post-Conflict Situations

The land grabbing problem is formulated as an outlier detection problem. Basically, in a post-conflict situation, days with a very high or very low number of transactions will deviate from the rest of normal days. This deviation gives those days outlying behaviour, and so outlier detection methods should be able to detect them.

The entropy-based outlier detection approach presented in He *et al.* (2005) is used for this problem, primarily because the results obtained in He *et al.* (2005) and Jiang *et al.* (2010) show superiority of the method over other methods. In addition, it was found that the entropy method can be easily adapted for the nature of the univariate time series data, which describes the frequency of transaction during fixed periods of time of a conflict.

4.6 Chapter Summary

This chapter presented a review for data mining methods in general and focused more on the methods that are used in developing a solution for the proposed fraud problems in this thesis. Basically, this chapter highlighted activity 4 of the listed research methods in Section 1.6.

Classification is one of the most common methods used in data mining and was discussed here with a focus on PDA, the adopted method in this thesis for identifying suspicious real estate objects. Decision trees were also reviewed, as they are used in this thesis to validate the results obtained from PDA. In addition to the review of classification methods, classifiers evaluation techniques were discussed in this chapter.

Outlier detection methods were also reviewed, focusing on entropy-based outlier detection, which is used in this thesis for fraud detection in land transaction in post-conflict situations. Finally, at the end of this chapter, the adopted methods for the experimental portion of this study were listed and argued.

This chapter helped the author to achieve Sub-objective 1.4.b of the research objectives. It also addressed research question 6 in Section 1.5.

Chapter Five: Dataset Simulation and Development of Land Record Simulator

5.1 Introduction

A major problem in this research has been getting access to land record datasets. This chapter discusses a simulator the author developed to overcome this problem. As mentioned in Section 2.6, this study explores two different fraud scenarios:

1. Land grabbing in post-conflict situations
2. Property or mortgage fraud in the real estate sector

In post-conflict situations, accessing land records is a very difficult and potentially dangerous task. NA (2010, pers. comm., 16 January) described the process of trying to gain access to land records in post-conflict situations as "difficult and somewhat dangerous" especially if the situation has not stabilized yet. People with vested interest in illegal land grabbing in most affected countries (*e.g.,* Colombia, Afghanistan and Kosovo) are the same people that make other people disappear. They usually have control over access to data and want to hide any kind of illegal activities committed during the conflict (NA 2010, pers. comm., 16 January).

In general, even if there is a possibility of gaining access to data, it will still be problematic for two reasons. First, land data in many countries, and specifically developing countries, is not in a form that can be analysed by computers, or it needs pre-processing. To be able to analyse land data using computer systems, it is essential to have the data in digitized form, which often is not the case. Second, even when access is granted, in many cases there are restrictions on access to information (NR 2010, pers.

comm., 10 March). These restrictions may include a limit on the number of records that can be accessed or restrictions on the type of information that can be accessed for each record. These limitations cause a problem in developing analysis methods, as data will be incomplete.

In stable situations, Pollakowski and Ray (1997) state that the lack of a uniform data source is considered one of the biggest problems that researchers in the real estate market have to deal with. They list three reasons that cause this problem: heterogeneity of housing assets, transaction infrequency for individual property, and finally, different sources may provide different dataset structures.

It might be possible to gain access to datasets. However, one has to get permission to access the data; as an independent researcher, negotiating this could take an inordinate amount of time, and it may be refused. For a M.Sc. this is too risky, as permission may never be granted, so simulated data has been used in this exploratory study.

Moreover, in cases where data is available, the data might not be accessible due to restrictions and obstacles that may exist. Some of these restrictions are; (1) limitations on the number of records that can be accessed; (2) limitations on the type of information that can be accessed; (3) limitation of format, usually text; and (4) expense to obtain records, as it is required to pay per record. These restrictions were established during the process of looking for datasets.

For example, in Alberta, Alberta Registries (2002) has developed the SPIN2 system, which provides tools to search and obtain land-related registered documents. However, information is only provided in limited document formats (Acrobat PDF

documents, TIFF images, and plain text). A second problem is that only one document can be obtained at a time. Finally, to obtain a document, the system charges a small fee that ranges from $2 to $5 per document (Alberta Registries 2002).

A simulator makes it possible to control the generated records. It allows for the generation of synthetic land and real estate transactions and at the same time gives the ability to introduce some of the fraud patterns identified in this research. Obtaining synthetic datasets will make it easier to test and validate data mining algorithms, as the researcher has a priori knowledge of the data.

This chapter describes the design and implementation of the Land Record Simulator (LRS). The chapter also presents the simulated datasets used in testing the classification and outlier detection methods used in the experimental work.

The chapter starts by describing the LRS as a software system. The main drivers to develop the simulator are listed, and then, the development environment is discussed. The simulation process flow is then addressed, which comprises three parts: (1) the initial process; (2) the land transactions simulation process; and (3) the property transactions simulation process. After the discussion of the simulation process, the actual generated datasets used in the experiments of this research are examined.

This chapter addresses Question 1.5.7 of the research questions. The chapter also addresses sub-objective 1.4.c.

5.2 Land Records Simulator

This section describes the technical component of the simulator.

LRS is a Windows-based desktop application. It was developed using C# programming language, and Microsoft Visual Studio 2005 as the Integrated Development Environment. The user interface consists of two parts: the controls that are used to set the different parameters for the simulation, and a mapping area that can be used to visualize a map of the land parcels for which the records are generated. A screenshot of the simulator is presented in *Appendix D.*

For mapping functionality, the SharpMap library was used, which is an open source mapping library written in C# and based on Microsoft.NET 2.0. SharpMap was chosen because the library provides access to many types of GIS data, enables spatial querying of that data, and is an open source C# library (SharpMap, 2009).

5.2.1 Simulation Process

The simulator performs three main operations: (1) the general operation, which creates the initial set of parcels to be used by the other two operations; (2) the land transactions simulation operation, which uses the initial set of parcels to simulate land transactions in conflict/post-conflict situations (now labelled as land transaction simulation); and (3) the real estate transactions simulation operation, which uses the initial set of parcels to simulate real estate transactions in stable situations. Although the last two operations are the same in real life, two different modules were developed in the simulator for each. The author uses a different data format for each simulation as required for the detection.

In the general operation of the LRS system, the user creates an initial (*l x w)* parcel where *l* and *w* are the length and width of the initial parcel. This parcel is then

subdivided into a number (defined by the user) of squared equally sized parcels. All the generated parcels are registered as owned by the state. The following two sections provide discussion about the two simulation modules that form the simulator.

5.2.1.1 Land transactions simulation module (Conflict / Post-conflict)

This module is responsible for simulating the land conveyancing process in order to generate land transactional datasets. The output of this model is used for the problem of fraud detection in post-conflict situations. Following the general operation, this module starts generating land transfers between people. The people are randomly generated using a tool that was developed for this study. The tool uses two list of names, first name and last name, and then randomly creates combinations of those names to generate hypothetical individuals. The tool also gives each individual a unique ID. Names were extracted from <http://www.listofbabynames.org/a_boys.html>.

The land transfer process randomly generates land transactions where two types of operations can be performed, parcel transfer and parcel subdivision. A parcel subdivision is always followed by a number of transfers equal to the number of parcels created in the subdivision.

During each transaction, information regarding the transaction is generated and saved using a tree data structure. The root of the tree represents the initial parcel. Figure 5.1 illustrates how the tree structure stores transactions and parcel information. The green (filled) nodes are the leaf nodes of the tree. A leaf node represents the final state of a particular parcel such as current location and current parcel ID. It also contains

information of the last transaction. The final leaf nodes are used to generate a land ownership map at the end of the simulation. The transactions of the nodes in the second level of the tree (the General Operation) are not recorded nor used for the final output dataset, as they did not encounter any transactions and are still owned by the state. They are just used in drawing the final map.

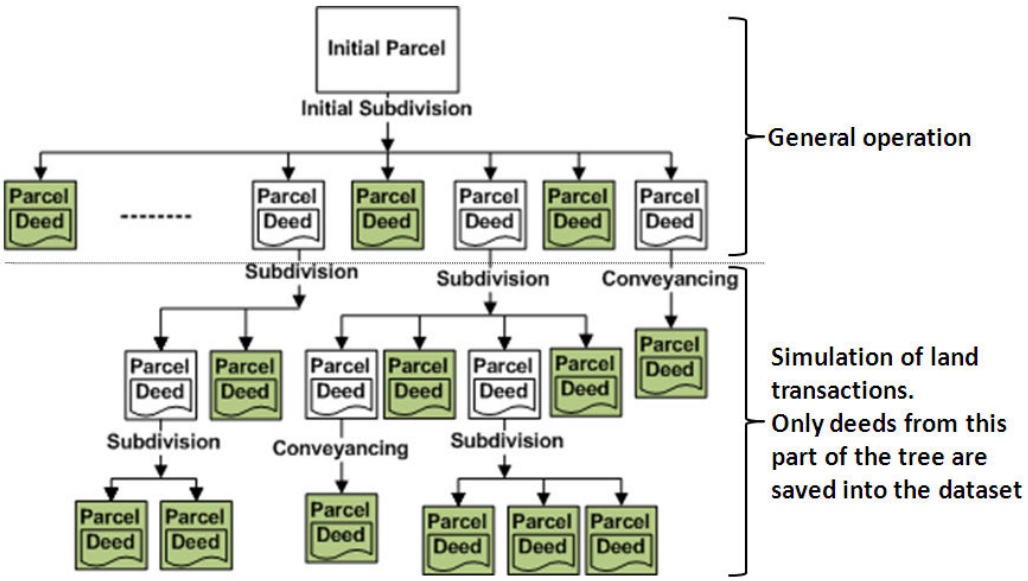


Figure 5.1: The tree data structure used in the simulation of land transactions.

The simulated number of transactions generated per day can be altered as needed. Three different options are available: a fixed value to be used for each day of the simulation, an array of values that determine the number of transactions for a finite number of days, or a function of time to set the number of transactions for each day of the year.

The module allows for the introduction of a small bias in the number of transactions generated per day, regardless of the method used. Generally, if the number of transactions for a certain day is calculated to be x using one of the three options, the

actual number to be used would be $x \pm c$ where c is a randomly generated number between 0 and α, and the user enters α. Thus the number of transactions will differ for the different days of the week. Some days will have higher activity than others, a pattern reflected in the Canadian property market (e.g., see CREB 2010).

The last feature of the module is the ability to introduce outliers when generating datasets. To recap, Hawkins (1980) defines an outlier as "an observation that deviates so much from other observations as to arouse suspicious that it was generated by a different mechanism." Outlier in this case applies to the transactional behaviour in a certain day and not to a single transaction. This means one transaction cannot be identified as an outlier, but rather, the presence or absence of a certain number of transactions in a certain day can raise suspicion that there might be fraudulent behaviour on that day. For example, if the expected number of transactions for a certain day is x but the actual number is y where $y=x\pm\alpha$ and α *is* a large number relative to x, that day can be recognised as having outlying behaviour.

Only one type of outliers can be introduced using the simulator: an exceptional number of transactions in certain periods of time during the simulation period. The way to introduce the outlying number of transactions is by neglecting the generation rule used in the simulation process. The exceptional number could be a very high number or a very low number when compared to adjacent values. It could also be zero, which reflects the removal of all transactions that took place in a certain period of time in a real-world scenario. These patterns reflect some of the patterns that occur in periods of conflicts / post-conflicts, as mentioned in Section 2.6.1.

Finally, generated datasets from this module are outputted in two formats, XML and text. The XML schema is portrayed in Figure 5.2. For the text format, the module generates a colon-separated text file that includes all the simulated transactions. The structure of the text file is set to include all the fields from the XML schema in the following order: *DeedNumber, ParcelOwnedByState, DeedObsolete, DeedRegistered, LotNumber, PartOfLotNumber, ParcelCoordinatesNE_X, ParcelCoordinatesNE_Y, ParcelCoordinatesSW_X, ParcelCoordinatesSW_Y, PrimaryRightHolder, PrimaryRightHolderID, RegistrationDate*.

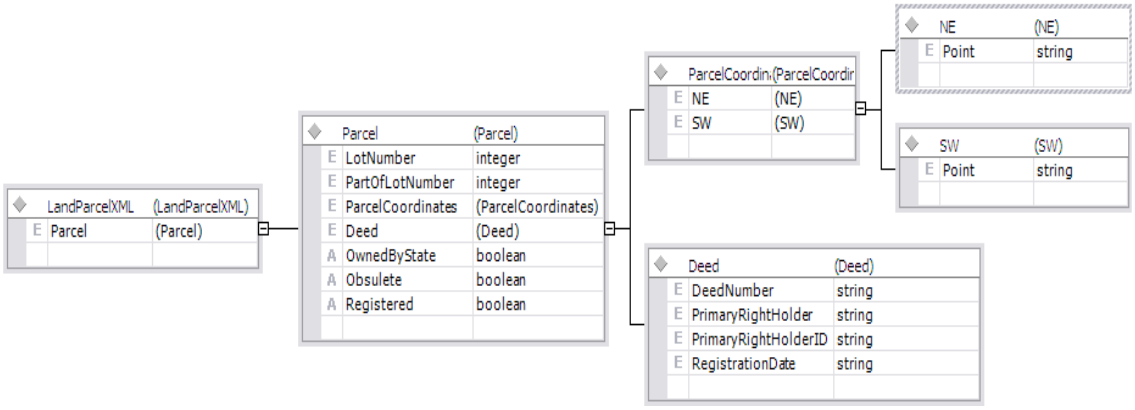


Figure 5.2: XML schema diagram for the output of land transactions simulation module.

5.2.1.2 Property transactions simulation module (Stable Real Estate Markets)

Real estate markets tend to involve improved land parcels (i.e., buildings are constructed on parcels) in secondary markets. For the purposes of this study, the term *property* is defined as an improved land parcel. This module creates properties and performs transactions on them, which is basically a land transaction. The module uses the initial parcels array generated from the general operation (see Figure 5.2) and performs

transactions on them with the assumption that those parcels are improved. The difference between this module and the land transactions simulation module is mainly in the attributes generated with the transactions.

Each property is given an initial value by which it starts the transactions. The initial values for the properties should be provided from an external file. This technique is employed to give the users the flexibility of defining their own values according to their requirements. Initial values can be generated using tools such as MS Excel or MATLAB, or provided directly from available datasets. However, the values should be given to the simulator in a text file with one value in each line of the file. This method is used to allow for the values to be introduced in a way that follows a certain distribution according to the real estate market statistics in the targeted area of the simulation. In the simulations performed for this study, MATLAB was used to generate property values using statistics from the City of Calgary, Alberta. This procedure is discussed in more detail in Section 5.3.2. The author chose the City of Calgary because statistics were found for this city, which helped in determining the attributes of the simulation.

Generation of property transactions starts after each property is assigned an initial value. In a single iteration of the simulation process, a property is selected randomly from the initial list and is sold to a randomly selected person from the community.

After a property is sold to a randomly selected person, a transaction is generated for the sale, and all the attributes (Table 5.1) for the transaction are set. The transaction is then stored into the output dataset. The final output of the simulator is a transactions dataset comprising one table. Each record in that table represents one transaction on one

property. The generated dataset schema can be seen in Table 5.1 with a brief description of each attribute.

Table 5.1: Attributes of the generated table (RealEstateTransactions) from the real estate transactions simulation module.

RealEstateTransactions	
Field	Description
TitleNumber (PK)	A unique number generated in sequence for each transaction; it does not follow the three parts rule followed in Alberta title numbers
PropertyID	A unique number generated for each property
RegDate	The date on which the transaction takes place
Value	The price the property is sold for
BuyerFName	The first name of the property buyer
BuyerLName	The last name of the property buyer
BuyerID	The ID number of the property buyer
Mortgage	A Boolean field describing whether a mortgage is or is not obtained on the property[1]
MortgageRegDate	The registration data of the obtained mortgage
MortgageValue	The value of the obtained mortgage
LTV_ratio	Loan To Value ratio, which is the ratio between the mortgage value and value of the property

The attributes of the generated transactional dataset are chosen based on the Land Title Certificate of the province of Alberta, Canada. Samples were obtained from SPIN2, the online Alberta Land Titles Spatial Information System (Alberta Registries, 2002). A few changes were introduced to the format of the obtained title certificates. First, the original format of the title certificate contains a *legal description* field which is used to identify a property using plan number, block number, and lot number. This field is simplified in the simulator to include only one unique integer as an ID for each property. Plan number, block number and lot number are not used in the methods developed in the

[1] Only one mortgage can be obtained on a property in the simulation. This approach was used because none of the experimented fraud schemes in this research include obtaining multiple mortgages on the same property without transacting the property.

study. The second change was the addition of the *LTV_ratio* (Loan to Value Ratio) attribute. This field is not generated but calculated from the ratio between the mortgage value and the property value.

The module only generates normal transactions. This means that a maximum of two transactions are allowed per property in a period of two years. Also, the increase in the value of the property is constrained. It is set to be between 6% and 12% for every two consecutive transactions on that property. This range is assumed in this study as the normal range of property value increase because there were no real figures showing how much is the normal percentage in real life. So, this range is chosen as a proof of concept.

Finally, the simulator does not allow for a transaction to take place on a property if the last transaction occurred within 100 days of the proposed transaction. These configurations limit the cases of suspicious transactions but do not totally forbid them.

To generate suspicious transactions, a separate process was developed. In this process, a group of properties is selected. This group is assumed to contain properties that would act as targets for suspicious activities. The process iterates through the group. For each property, a number of transactions are executed, and in each transaction some attributes will be set that reflect suspicious activities. The manipulation of attributes to reflect suspicious transactions is derived from the indicators mentioned in Section 3.3. This includes the permission to execute more than two transactions, using relatively short periods of time between successive transactions (can be less than 100 days), using the same individuals from the community to transfer a property back and forth, obtaining a

high-ratio mortgage on the property (it can exceed 90%), and allowing for high-value increases between back-to-back sales that are more than 12%.

In summary, the modules developed for generating synthetic datasets of land and property transactions were examined in this section. The actual datasets that were simulated and used in the experiments are discussed in the following section.

5.3 Datasets Simulation

In this section, the datasets generated for the study are discussed. Primarily, four datasets were generated and used for the experiments. Three of them are Land Record Datasets (LRDS1, LRDS2 and LRDS3). These three datasets were generated with the land transactions simulation module developed in Section 5.2.1.1 for experiments on detecting fraudulent activities in land transactions in post-conflict situations. The remaining dataset is a Property Transactions Dataset (PTDS1). It was generated using the property transactions simulation module developed in Section 5.2.1.2 for the experiment on detecting fraud schemes in real estate transactions.

The following two sub-sections discuss the original datasets generated with the simulator and the "outlier" records in the datasets.

5.3.1 LRDS1, LRDS2, and LRDS3

As mentioned in Section 3.4, in post-conflict situations, a number of indicators in land records could be found that reflect land stealing or grabbing by powerful groups during the conflict or in its aftermath. The indicators show noticeable change in the trend of

conveyancing activities during or after the end of a conflict. Essentially, an exceptional high or low number of transactions in certain periods is the indicator to look for.

To get the datasets required for the experiments, the land transactions simulation module discussed in 5.2.1.1 was used. Three datasets were generated: Land Records Dataset 1 (LRDS1), Land Records Dataset 2 (LRDS2) and Land Records Dataset 3 (LRDS3).

For LRDS1 and LRDS2, the trans-per-day input method was set as a fixed value of 30 with the addition of a random bias of 0-60. So, in LRDS1 and LRDS2, the number of transactions for each day of the simulation process (n) is defined as $n = 30 + c$ and c is a random number from 0 to 60 to give variability to the data. This formula generates the number of transaction for usual days.

For LRDS3, the generation method was set as an array of pre-defined values. The array was obtained from the interpolation of property sales in Calgary, Alberta for the two years 2008 and 2009 (CREB, 2010). Figure 5.3 shows the interpolation results.

The configuration attributes used to generate the three datasets are shown in Table 5.2. For the three datasets, the simulator was set to generate records over the period of two years with an initial number of land parcels of 2500. Other than the trans-per-day input method, there are two other differences in the settings of the simulator used for each dataset: first, in the population used for each simulation; and second, in the number of outliers superimposed into each dataset. The population, however, has no effects on the expected results but was included in the simulation for future use.

Table 5.2: Attributes used to generate the datasets (LRDS1 and LRDS2).

Dataset	*Trans-per-day input method*	*Simulation start date*	*Simulation end date*	*Population*	*Number of generated Records*
LRDS1	Fixed value	Jan 1, 2010	Dec 31, 2012	50,000	59398
LRDS2	Fixed value	Jan 1, 2010	Dec 31, 2012	20,000	52727
LRDS3	Array of values	Jan 1, 2010	Dec 31, 2012	20,000	31577

Table 5.3: Summary for the three datasets simulated for post-conflict situations.

Dataset	*Number of days exhibiting normal number of transactions*	*Number of days exhibiting exceptional number of transactions (outliers)*
LRDS1	748	32
LRDS2	771	9
LRDS3	702	28

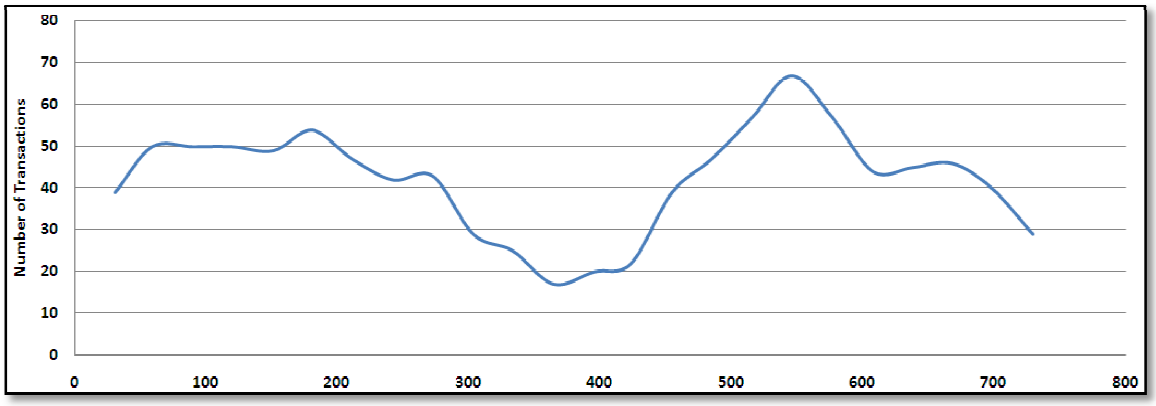


Figure 5.3: transactions-per-day values used for generating LRDS3. Values were interpolated from Calgary property sales statistics for the two years 2008 and 2009 taken from (CREB, 2010).

Table 5.3 shows number of outliers introduced into each dataset. This was achieved by generating a very high or very low number of transactions for the selected days, and superimposing them on the original dataset. Each outlier represents one day in which an exceptional number of transactions occur. For LRDS1 a total of 32 days of suspicious activities are introduced, 9 days for LRDS2 and finally 28 for LRDS3.

5.3.2 PTDS1

In order to test the fraud detection technique to detect property and mortgage fraud schemes, a dataset containing 37,380 records representing transactions over a two year period was simulated. The total number of transactions was determined by the number of transactions set for every day and the chosen period of the simulation. For this simulation, the property transactions simulator module discussed in Section 5.2.1.2 was used. The parameters used in the simulation were determined based on statistics obtained from the real estate market in the City of Calgary, Alberta. In principle, three parameters were vital for the simulation to make it as realistic as possible. These are:

- *Number of transactions per day*
- *Property prices*
- *Loan to Value ratios* (Mortgage Value/ Property Paid Price)

Number of transactions per day: To determine the number of transactions per day used in this simulation, Calgary Real Estate Board's (CREB) monthly statistics for property sales for the months from October 2009 to October 2010 were used (CREB 2010). The statistics include monthly sales of two different kinds of properties, single family houses and condominiums. Sales for each day of the year were interpolated from the monthly sales figures. Figure 5.4 shows the interpolated sales for one full year.

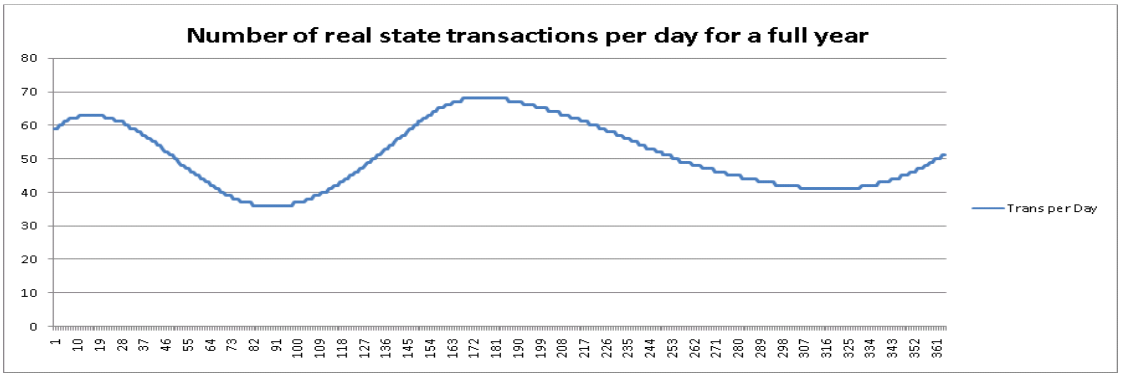


Figure 5.4: Interpolated real estate sales per day for a full year.

Property prices: An initial price is set for every property using sales statistics from CREB (2010) and Teranet (2006). The statistics from CREB do not have prices for all properties. Thus, MATLAB was used to generate prices for 308,315 dwellings based on statistical values obtained from Teranet (2006) for property values in Calgary, Alberta for the year 2006. Prices were generated using a Pearson Distribution random generator based on Ohnishi *et al's.* (2010) investigation of the distribution of house prices in Tokyo, Japan. This was the only published work found relating to property price distribution. Based on this work, values for the *mean*, *standard deviation*, *skewness*, and *kurtosis* used in the random generator were set to *382000, 120000, 0.52,* and *3.1* respectively. The distribution of the generated property prices is shown in the histogram in Figure 5.5.a. These prices were used to set the initial property prices in the simulation process.

Loan-To-Value ratio (LTV): LTV ratio is the value of a mortgage loan as a percentage of the total value of real estate property based on the selling price and not the valuation of the property. No precise statistics were found to be used as a base for the simulation of LTV ratios. However, Montia (2010) mentions that the average LTV ratio

for 2009 in England was 0.75. Also, real estate agents and listing services in Canada note that low LTV ratios are below 80%, and ratios of 90% or more are considered high and rare. Based on those estimates, an array of 400 LTV ratio values was generated. Figure 5.5.b shows the histogram of the generated LTV ratios.

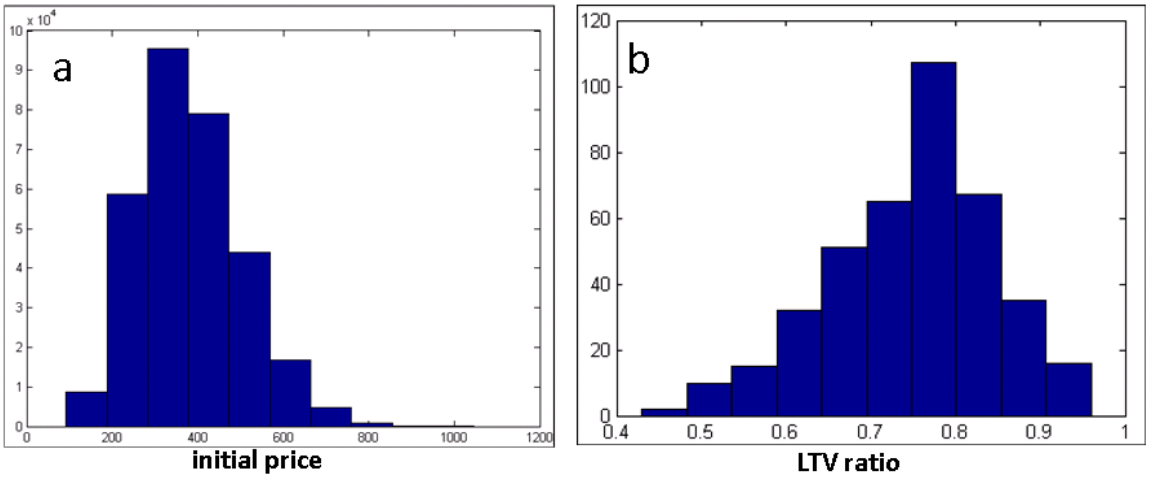


Figure 5.5: a) distribution of 308315 generated dwelling initial prices. b) Distribution of the 400 generated LTV ratios.

The transactions-per-day values list, property prices list, and LTV ratios list were fed to the simulator in order to generate the dataset. The simulation period was set for two years and a total of 37,380 transactions were generated. These transactions took place on 36,917 different properties.

To generate transactions that exhibit fraudulent activities on selected properties, 245 properties were selected from the 36,917 properties. This number of selected properties to be targeted as illegal activities was selected according to the report of Auditor General of Alberta (2010) in which a total of 30 properties were found to exhibit illegal activities out of the 4254 examined properties in that report. This gives a

percentage of around 0.7% of the total properties, found to be targeted by fraudulent activities.

A total of 774 transactions were generated on the selected properties to represent suspicious activities suggesting patterns of fraud, primarily of the ABC-Construction and Oklahoma Flip variety. The 774 transactions were finally added to the original dataset to form a total of 38,150 transactions in the dataset PTDS1 which was used in the experiments.

5.4 Chapter summary

This chapter presented the land and property transactions simulator that was developed as part of this research. The development of this simulator was an essential step in this research because of the lack of real land transactional data.

The chapter discussed the need of this simulator and described the technical realization of it. It also described the modules available in the simulator and the different simulation processes it provides: the general process, the land transactions simulation process, the property transactions simulation process and the processes used to generate fraudulent behaviours.

In addition, the chapter examined the datasets used in this research that were generated using the simulator. It described the configuration used to generate the datasets and addressed some of the statistics used for that purpose. Four different datasets were described: LRDS1, LRDS2, LRDS3 and PTDS1.The first three were simulated to reflect

fraudulent activities in land records during or after a conflict. The last dataset (PTDS1) was simulated to reflect fraudulent activities in the real estate sectors.

This chapter provided the answer for research question 7. In particular, it addressed the execution of Activity 5 in the research methods provided in Section 1.6 to achieve the pre-requisite of sub-objective 1.4.c.

Chapter Six: Experimental Analysis

6.1 Introduction

So far, this thesis examined the problem of fraud in land record systems. Several fraud methods were identified and discussed in Chapter 2 and Chapter 3, and the patterns and indicators for a selection of fraud schemes were examined in detail in Chapter 3; namely, the Oklahoma Flip scheme, the ABC-Construction scheme, and land grabbing in post-conflict situations. Chapter 4 then reviewed the data mining methods that are used in this study to detect the three fraud schemes, and Chapter 5 discussed the Land Records Simulator and the simulation of the data sets used in the experimentation.

This chapter reports on the experimental work the author has done to detect the three fraud schemes. It uses methods reviewed in Chapter 4 and describes the application of those methods to the datasets simulated in Chapter 5, in order to assess the effectiveness of data mining methods in detecting the discussed fraud schemes. This addresses sub-objective 1.4.d.

The chapter is organised in two main sections. First in Section 6.2, the author reports on classification models to detect suspicious properties in property transactions, and develops a model to be applied to the dataset. The model was developed based on patterns and indicators of the Oklahoma Flip and ABC-Construction schemes. Then in Section 6.3, the author reports on the outlier detection method developed to detect fraudulent activities in post-conflict situations. Finally, Section 6.4 provides a chapter summary.

6.2 Detecting Oklahoma Flip and ABC-Construction Schemes

The goal of this section is to assess the possibility of detecting suspicious properties in property transactions datasets using classification methods. Suspicious properties refer to properties that might be involved in fraudulent activities based on fraud indicators and patterns discussed in Section 3.3.

As mentioned in Chapter 4, two classification methods were used; the first one is a classification model based on quadratic PDA. The second model is a classification tree model that was built using the CART algorithm. The models are supposed to classify properties according to how suspicious they are, based on five different attributes.

Section 6.2.1 describes the design of the study and Section 6.2.2 describes the data preparation process. Sections 6.2.3 and 6.2.4 introduce the building of a quadratic PDA classification model as well as the classification results obtained from this model. Finally, the building of a classification model using CART is described in Section 6.2.5, along with the results of this model.

6.2.1 Study Design

In order to detect fraudulent Oklahoma Flips and ABC-Construction schemes from the transactions dataset developed in Section 5.3.2, three groups were defined: Normal (N), Suspicious (S) and Highly Suspicious (H). The classification models have been developed to classify a property into one of these three groups based on its transaction attributes.

At the beginning of the study, only two classification groups were defined, normal and suspicious; however, the separation line between normal and suspicious was not clear enough. This obscurity in the separation between the two groups is due to the method of building the indicators which define a suspicious property. These indicators were built based on subjective analysis of personal communications with experts, and also based on some studies that address the fraud problem. The author could find neither statistics nor empirical studies to provide a clear definition of a suspicious property.

Consequently, three groups were introduced. Using this system, a property that is classified as S or H is suspicious and needs to be investigated. However, S is less suspicious than H and has a high probability of being normal. This middle class was introduced in an attempt to reduce the false positive error, which is the rate of identifying a property as normal when it is actually suspicious. On the other hand, identifying a property as suspicious when it is actually a normal property is considered less harmful.

In order to assess the suggested classification methods, the experimental design was based on the following steps:

1. Transforming the simulated property transactions dataset by grouping the records by properties (rather than by time sequence) and calculating some descriptive attributes for each property. This step is required because each record in the dataset represents one transaction on one property and does not describe the property itself.
2. Filtering the dataset to remove unwanted observations. Unwanted observations in this case are properties that have been through only one transaction.

3. Selecting a representative sample of the dataset in order to use it to build the classification model.
4. Labelling the sample dataset by assigning each property into one of the three groups.
5. Assessing the candidate sets of attributes to be used in a classification model. Originally, each property is described by 10 attributes described in Table 6.2 This assessment should help us to decide on the attributes that should be included in building a classification model.
6. Using quadratic PDA to build and evaluate a property classification model.
7. Using the CART algorithm as described in Section 4.3.2 to build a property classification tree and compare its classification results with the classification results obtained from the quadratic PDA model.

Detailed discussion of these steps is included in sections 6.2.2, 6.2.3, 6.2.4, and 6.2.5.

6.2.2 Data Preparation

The original data set used for the two experiments (quadratic PDA and CART) is PTDS1, which was described in Section 5.3.2. A summary of PTDS1 is shown in Table 6.1.

Table 6.1: Information of simulated PTDS1.

Total transactions	38150
Total properties	36917
Selected properties that exhibit fraud	245
Number of transactions on properties that exhibit fraud	774

The classification models are expected to classify properties and not transactions into one of the three different groups. However, each record in PTDS1 represents one transaction on one property and does not describe the property itself. So, the first step in data preparation was to generate a new dataset, called "Properties Data Set 1" (PDS1), from PTDS1 by grouping the records based on properties. In PDS1, each record represents a summary of all the transactions that took place on one property during the two-year period.

Because the dataset is simulated, it is expected that proportions of the three different groups do not actually represent the actual proportions that might be found in a real dataset. Also, PDS1 is not labelled, and therefore it is impossible to know the size of each class from this dataset. So, to obtain actual proportions, the author used statistics of suspicious properties that were found in the report Auditor General of Alberta (2010) and Unger *et al.* (2010). Table 6.2 shows the attributes of PDS1 and a description of each attribute.

According to the patterns examined in Section 3.3, properties with only one transaction cannot be classified as suspicious and are considered normal. So a filter was applied on PDS1 to remove all properties with only one transaction.

The third step of the data preparation was to select a representative sample out of the filtered PDS1. To achieve representativeness of the sample, two requirements must be met according to Huberty and Olejnik (2006, p. 309 - 310). The first requirement is to select a sample size that satisfies the formula, $\min(n_j) > 5p$, where p is the number of predictors (attributes) used to build the model, and $\min(n_j)$ is the size of the group with

the fewest number of records in the population. This means that the smallest group in our sample dataset should contain at least 5 times the number of predictors. The second requirement is to select a sample with group proportions that reflect the actual proportions of the population.

Table 6.2: Description of PDS1.

Attribute	Description
PropertyID	ID of the property the record describes.
NumberOfTransactions	Number of transactions that took place on the property during the selected epoch.
NumberOfPersonsInvolved	Total number of different persons who were involved in the transactions on the property.
InitialValue	Value of the property in the first transaction in the selected epoch.
LastValue	Value of the property in the last transaction in the selected epoch.
AverageChange	Average change (increase or decrease) in the value of the property between each two consecutive transactions that took place over it. *AverageChange* is calculated as $\sum_{i=2}^{n} \frac{V_i - V_{i-1}}{n-1}$ where V_i is the value of the property in the i^{th} transaction, n is the total number of transactions that took place on the property, and V_1 is the initial price of the property.
PeriodOfTransactions	The period between the registration dates for the first and last transaction on the property. This period is calculated in days.
AverageFlipPeriod	Average flipping period of the property. A flipping period is defined as the number of days between any two consecutive transactions on a certain property. *AverageFlipPeriod* is calculated as *PeriodOfTransactions* divided by *NumberOfTransactions*.
MortgageValue	Value of the mortgage attached to the last transaction on the property.
LTVR	LTV ratio for the loan attached with the last transaction on the property (see Section 5.3.2 for definition of LTV ratio).

Because the dataset is simulated, it is expected that proportions of the three different groups do not actually represent the actual proportions that might be found in a

real dataset. Also, PDS1 is not labelled, and therefore it is impossible to know the size of each class from this dataset. So, to obtain actual proportions, the author used statistics of suspicious properties that were found in the report Auditor General of Alberta (2010) and Unger *et al.* (2010).

As discussed in Section 2.4.2.2, in Auditor General of Alberta (2010), 148 properties exhibited indicators of possible fraudulent activities based on an initial filtration process. After the second scanning process, 30 properties out of the 148 were found to be highly suspicious. This gives a percentage of 21% of highly suspicious properties. In the study by Unger *et al.* (2010, p. 10), the numbers show that 36 out of the 200 selected properties for analysis (i.e., 18%) were identified as suspicious properties.

To simulate known real-world situations, based on the above percentages the author assumed a ratio of 0.20 of the unusual properties in the dataset to be highly suspicious. This gives a proportion of 4:1 between the two groups, N and H. But, the proposed classification model is a three-class model, and so the ratio of group S is still needed. To determine this ratio, the author assumed the same ratio as the H class, which is 0.2. This is based on the fact that all the properties in the dataset after applying the filter are unusual, and so, more properties are expected to exhibit suspicious activates. As a result, the final proportions for the three classes in the sample should be 3(N):1(S):1(H).

Table 6.3: Details of the representative sample (LPDS) selected from the filtered PDS1.

Total number of properties	*properties in each group*		
	N	S	H
315	187	64	64

The final step in the data preparation process was to label the properties in the sample set. To do so, 500 properties were selected from the filtered PDS1. Each of the properties was then assigned to one of the three groups using indicators and patterns discussed in Section 3.3, and based on that assignment it was given a label (N, S or H). Out of the 500 records, the final sample was selected to satisfy the conditions of representativeness as discussed above. The details of the sample can be seen in Table 6.3. This sample contains 315 records of properties labelled using the three defined classes. This set will be referred to as LPDS (Labelled Properties Dataset) in the rest of this thesis.

The author used LPDS to assess the use of quadratic PDA and CART in order to build classification models to detect suspicious properties. The following section, Section 6.2.3, examines the use of Quadratic PDA to build and test a classification model. Then, in Section 6.2.4, the author uses Classification Trees to validate the results acquired from the Quadratic PDA experiments.

6.2.3 Classification of Properties Using Quadratic PDA.

Initially, based on the indicators and patterns explained in Section 3.3, the preliminary candidates for property classification included (1) *NumberoOfTransactions* (NT); (2) *NumberOfPersonsInvolved* (NP); (3) *AverageChange* (AC); (4) *AverageFlipPeriod* (AFP); and (5) *LTVR* (LTVR). These attributes were chosen out of the attributes described in Table 6.2. The remaining attributes were not included because they are not part of the indicators of Oklahoma Flip nor ABC-Construction.

In order to apply PDA, Huberty and Olejnik (2006) recommended that multivariate normality be found in the dataset. If multivariate normality is not tenable, PDA can still be applied, but optimal results cannot be guaranteed. Multivariate normality is a generalization of the one-dimensional normal distribution into higher dimensions (Burdeaski, 2000). According to Burdeaski (2000), in a dataset of two variables or more, one necessary condition to achieve multivariate normality is the univariate normality for each variable; i.e., each variable by itself should follow a normal distribution in order for multivariate normality to be tenable in the dataset. This condition is not sufficient, however, so if this condition applies it does not necessary mean the data is multivariate normal.

To check for multivariate normality, a check was first conducted on the univariate normality of the five predictors as recommended in Burdeaski (2000). Burdeaski suggests that one of the most used tests for univariate normality is the normal probability plot or Q-Q plot (Quantile versus Quantile). In this plot, observations of each variable are ordered in decreasing degree of magnitude and then plotted against expected normal distribution values. Based on Q-Q plots of the five predictors in LPDS, which can be found in *Appendix B*, it can be concluded initially that multivariate normality does not exist in this dataset. In the Q-Q plots, the blue points are the observations and the red line is the expected normal distribution so if observations of one variable fall on the red line it implies the variable is normally distributed. As can be seen in those plots, actual observations do not follow a linear pattern and in general do not fall on the normal

distribution line. The plots prove that the data set is not multivariate normal as each individual variable does not follow a univariate normal distribution.

It is expected that all variables are independent. The multivariate scatter plot shown in Figure 6.1 shows no noticeable correlation between the pairs of variables except in the case of NT and NP. The correlation matrix for the five predictors in Table 6.4, calculated using MATAB, supports this conclusion, as it shows that NT and NP have a high correlation. All other variables may be considered uncorrelated. Because of the high correlation between the two predictors NT and NP, it is suggested not to use both in a PDA, as they may affect the classification results in a negative way (Huberty and Olejnik, 2006, p. 11).

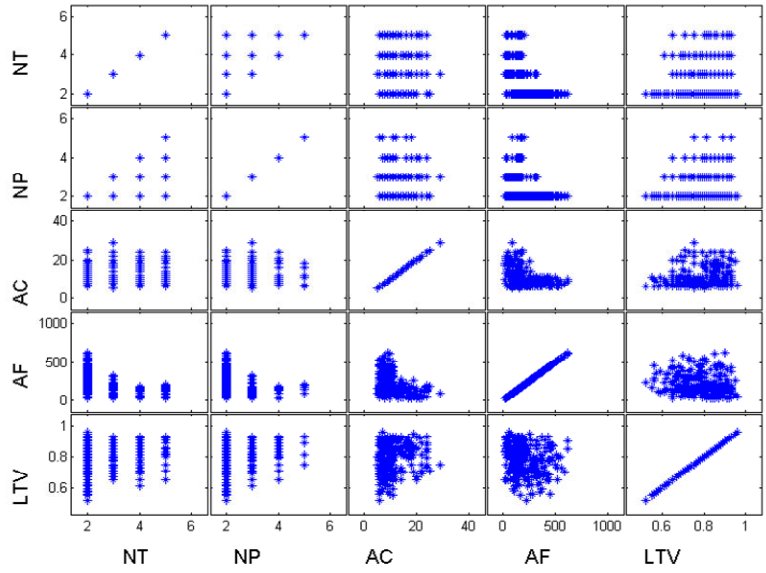


Figure 6.1: Multivariate scatter plot of the five predictor variables.

Logical screening of the initial variable list is a suggested step in deciding on the predictors that should be used in PDA (Huberty and Olejnik, 2006). In the case of our property dataset, the screening was done based on prior research and examination of the

indentified fraud indicators in order to determine influence of each variable on the classification of a property. In spite of the high correlation between NT and NP, logical screening – that is the subjective judgement of the author obtained through the analysis of interviews and fraud schemes – suggests that both NT and NP are important for predicting either of the two fraud schemes under investigation.

Table 6.4: Error correlation matrix for the five predictors in LPDS.

	NP	**AC**	**AFP**	**LTV**
NT	0.763	0.346	-0.436	0.265
NP		0.207	-0.306	0.239
AC			-0.342	0.215
AFP				-0.177

It was found during the examination of the two fraud schemes that both number of transactions and number of people involved in transactions over one property (i.e., NT and NP in LPDS) play an important role in determining highly suspicious properties. These two predictors were always the first to be mentioned in any of the fraud schemes. Accordingly, the author preferred not to drop any of the two attributes but rather examine the impact of both on any classification model individually and together. So, three sets of candidate predictors were formed. The three sets are shown in Table 6.5. Each of these sets was used to build a classification model using quadratic PDA, and then the models were compared to decide on the best set of predictors.

Table 6.5: Candidate sets of predictors for building a classification model for property data.

Set	*Predictors*
Set1	NT, NP, AC, AFP, LTVR
Set2	NT, AC, AFP, LTVR
Set3	NP, AC, AFP, LTVR

To test the variation between the groups and within each group, and test if belonging to a group actually makes a difference, the multivariate null hypothesis is tested. Our null hypothesis H_0 is that belonging to one of the three groups does not make a difference on the predictors. This hypothesis implies that the actual population means for the three groups are assumed to be equal for each of the five predictors.

The results of the univariate hypothesis test (df_1=2, df_2=312) in Table 6.6 show high values of the *F*-statistic. The equations used for this test are included in *Appendix C*. This test measures the ratio of between-groups variation to within-groups variation. df_1 is the degrees of freedom between groups and df_2 is the degree of freedom within groups. High values of *F* imply that variation between the groups is actually higher than the variation within each group. It also means that the three population means for each predictor are most probably not equal, and belonging to one of the three groups creates a difference in the values of the predictors. *P* which was calculated as 0.000 for all *F* values represents the probability of obtaining *F* by chance assuming H_0. In conclusion, the multivariate null hypothesis can be rejected, and thus, all five variables have effect on the grouping variable and are capable of distinguishing between the three groups.

Table 6.6: Descriptive information and univariate test for the property data.

Variable	*N* *Mean(SD)*	*S* *Mean(SD)*	*H* *Mean(SD)*	$F_{2,312}$	*P*
NT	2.336 (0.822)	3.277 (1.096)	3.746 (1.03)	64.5659	0.000
NP	2.208 (0.590)	2.738 (0.888)	2.65 (0.626)	20.37355	0.000
AC	8.475 (1.968)	11.84 (4.305)	16.92 (5.571)	140.0585	0.000
AFP	263.16 (125.27)	142.95 (50.53)	84.31 (49.87)	86.68684	0.000
LTV	0.76 (0.089)	0.826 (0.0811)	0.839 (0.083)	27.01421	0.000

As discussed above, multivariate normality is not expected in LPDS. So, application of PDA would not guarantee the optimality of the classification results; i.e., the rates of correctly classifying observations cannot be assured to be the maximum possible. Furthermore, one more test needs to be done to decide on the classification rule, quadratic or linear.

To decide whether a quadratic or a linear rule should be used for classification, equality of the covariance matrices for the three different classes was tested. If it is established that covariance matrices for the three different classes are not equal, then a quadratic rule is favoured for the classification over a linear model. This condition can be tested using the Box's M test for covariance homogeneity, which tests the log-transformed determinants of the covariance matrices (Hunter, 2007). Equal values mean equal variability within a set of data. The 'MBoxtest' function in MATLAB was used for this test. The test results showed that the covariance matrices are significantly different.

Also, the natural logarithms of the determinants of the three covariance matrices were calculated. For group N the natural logarithm for the group's covariance matrix was 3.00. For group S it was 3.82. Finally, for group H it was 4.98. These values are different, which implies different variability within the dataset. This is further support for the inequality of the covariance matrices (Huberty and Olejnik, 2006, p. 278-279). This provides enough support to use a quadratic classification rule in order to build a classification model. However, optimality is not expected, as mentioned above.

Based on the above discussion, Quadratic PDA was used on LPDS three times to test three possible models. For each model, one of the candidate sets from Table 6.5 was used. The results of the three tests are discussed next in section 6.2.4.

6.2.4 Results of Quadratic PDA

The labelled dataset LPDS was used to estimate the quadratic discriminant functions. However, before establishing a final classification model, an assessment was conducted to choose one of the three candidate sets of attributes (see Table 6.5) to build the final model.

In order to build a classification model from each of the candidate sets, prior probabilities of the three different groups (H, S, N) were taken into consideration while performing classification with PDA. These prior probabilities were calculated based on the different group proportions established during the design of the study, as discussed in Section 6.2.2. The prior probabilities for the three classes were calculated as 0.6, 0.2, and 0.2 for the three classes N, S, and H respectively. These prior probabilities were used in the generation of a quadratic PDA classification model from each set.

To evaluate the models generated from the three candidate sets, first the resubstitution errors (reviewed in Section 4.3.3) for each of the three models is calculated to obtain an initial assessment of the hit rates. Then, a more robust evaluation is conducted using the Leave-One-Out method (reviewed in Section 4.3.3).

First, resubstitution errors and hit rates were calculated for the generated classification models. This was done by using the same dataset (LPDS) both as a training

set and as a test set for each of the three models. This means that for each model, all properties were used to first generate the model and then to test the model. The resubstitution hit rates for the three candidate sets are presented in Table 6.7.

As can be seen from the resubstitution hit rates in Table 6.7, the three classification models appear to perform equally well, with relatively high resubstitution hit rates. However, in contrast to the initial anticipation, Set1 generated a slightly better model than the models generated from Set2 and Set3, which had exactly the same hit rates of 0.841. Also, the results of the separate groups' resubstitution hit rates (presented in Table 6.8) show that the hit rate for group H is significantly higher for Set1 than for Set2 and Set3. This means that using both predictors NT and NP enhances the detection of highly suspicious properties, which supports the early logical screening.

The initial resubstitution results suggest that neither removing NP nor removing NT enhanced the generated model despite the high correlation between NT and NP. However, these results require more review since resubstitution error cannot be generalized. So, a more robust evaluation is conducted using the LOO method.

Table 6.7: Total group hit rates for the three models generated using the three candidate sets.

Predictors set	*Resubstitution hit rate*	*LOO hit rate*
Set1	0.848	0.822
Set2	0.841	0.819
Set3	0.841	0.832

A quadratic LOO was used to further evaluate the performance of the three models generated using Set1, Set2 and Set3. The total group LOO hit rates for the three models are presented in Table 6.7. Results of LOO showed that Set3 has a higher

classification rate than Set1, while Set2 has the lowest hit rate. The results of the LOO method are different from the results obtained using the resubstitution method, so further analysis was conducted by looking at the details of the classification results of LOO.

Table 6.8: Separate groups' resubstitution hit rates for the three models generated using the three candidate sets.

Group	*Set1*	*Set2*	*Set3*
N	0.888	0.909	0.914
S	0.719	0.734	0.687
H	0.859	0.750	0.781

Detailed property classification results of the LOO method for the three models are presented in Table 6.9, Table 6.10, and Table 6.11 in the form of 3 by 3 classification tables. Using these tables, it is possible to better evaluate the classification results for each model. Each row of a classification table shows the actual number of properties in one group. A column shows the number of properties that were predicted by a classification model to belong to a certain group. Finally, the last column shows the obtained hit rate for each of the three groups. This hit rate is calculated from the ratio between the number of properties correctly classified as belonging to a certain group and the total number of properties in that group.

Table 6.9: Quadratic PDA results using Quadratic LOO rule on set1.

		Predicted Group				
		N	S	H	Total	Separate group Hit Rate
Actual Group	N	**165**	22	0	187	0.882
	S	17	**40**	7	64	0.625
	H	4	6	**54**	64	0.844
Total		186	68	61	$N = 315$	

Table 6.10: Quadratic PDA results using Quadratic LOO rule on set2.

		Predicted Group				
		N	S	H	Total	Separate group Hit Rate
Actual Group	N	**169**	18	0	187	0.904
	S	14	**43**	7	64	0.672
	H	5	13	**46**	64	0.719
Total		188	74	53	$N = 315$	

Table 6.11: Quadratic PDA results using Quadratic LOO rule on set3.

		Predicted Group				
		N	S	H	Total	Separate group Hit Rate
Actual Group	N	**170**	16	1	187	0.909
	S	15	**44**	5	64	0.687
	H	6	10	**48**	64	0.750
Total		191	70	54	$N = 315$	

It was established that the purpose of the classification model is mainly to predict suspicious properties, which includes properties that fall into groups S or H. Also, it is of a great importance to obtain a high detection rate for highly suspicious properties even if the detection rate for group N is sacrificed. This is built on the assumption that in practice only properties classified as H or S will be further examined.

Based on this importance of group H, the most important goal is to obtain a high detection rate for group H in the chosen model. Looking at the three classification tables, one can see that Set2 and Set3 generate higher LOO hit rates for groups N, and S than Set1. This means that using these sets, more properties can be correctly identified as normal and suspicious. However, the group H hit rate is much lower when Set2 and Set3 are used than obtained from Set1.

The results also show that using Set1, 4 properties were incorrectly predicted as N when they actually belong to group H, 5 properties for Set2, and 6 properties for Set3. It is very important to get this number as low as possible in any suggested classification model. The reason is that properties described by this number are highly suspicious properties but were actually classified as normal.

As a conclusion of results obtained from the resubstitution method and LOO method, it was evident that Set1 generates a better quadratic classification model which outperformed the models generated from Set2 and Set3. This conclusion was not based only on the total hit rates of the three models, but also on the separate groups' hit rates for the three models, and especially the hit rates of group H. So, the quadratic discriminant functions for the final classification model were established using all five predictors (i.e., Set1).

To assess the effectiveness of the classification rule generated from LPDS using Set1, classification results obtained from the model were compared with those that could be obtained by chance. To prove that the model is effective, the author used a hypothesis test that the number of properties correctly classified by the model (o) does not exceed the number correctly classified by chance (e); i.e., H_0: $o \leq e$. Hence the alternative hypothesis was H_1: $o > e$ where e is the overall chance frequency (i.e., the number of properties that might be correctly classified by chance based on prior probabilities) and o is the observed frequency of hits from the quadratic model.

Assuming the number of groups is J, n_j is the number of instances actually belonging to group j, n_{jj} is the number of hits observed for group j, and q_j is the prior

probability of belonging to group *j*. The following can be calculated (Huberty and Olejnik 2006, p. 316 - 319).

$$e_j = q_j n_j \tag{6.1}$$

$$e = \sum_{j=1}^{J} e_j \tag{6.2}$$

$$o = \sum_{j=1}^{J} n_{jj} \tag{6.3}$$

$$z = \frac{o - e}{\sqrt{e\ (N - e)/N}} \tag{6.4}$$

Where *z* is a standard normal statistic used to test H_0 and indicates whether the actual results are better than results obtained by chance or vice versa. The larger the *z* score, the less probable the experimental result is due to chance. *N* is the total number of observations in the dataset. e_j is the number of instances in group j correctly classified using chance classification based on prior probabilities.

As can be seen in Table 6.12, *o* is significantly greater than *e* and hence the null hypothesis can be rejected. This is also supported by the obtained overall *z* score which indicates better-than-chance results, as it is show in the Table 6.12 that *P* is 0.00 for all groups. This *P* represents the probability of obtaining *z* using chance classification for each of the three different groups.

The results obtained so far are for the total group hit rate, and it might be suspected that not all separate-group hit rates are significantly greater than those to be

expected by chance. So, equation 6.4 above can be also calculated for each separate group as follows.

$$z_j = \frac{n_{jj} - e_j}{\sqrt{e_j\,(n_j - e_j)/n_j}} \tag{6.5}$$

Table 6.12: Comparison of classification results with chance classification.

Group	n_{jj}	n_j	q_j	e_j	z_j	*P*
N	165	187	0.6	112.2	7.88	0.00
S	40	64	0.2	12.8	8.5	0.00
H	54	64	0.2	12.8	12.88	0.00
Overall	***o***=259	***N***=315		***e***=137.8	***z*** = 13.76	0.00

Table 6.12 presents a comparison of the observed PDA classification results obtained from the five-predictors model with the chance classification. As shown, *P* is 0.00 for all separate-group predictions, which means that the probability of getting the z_j score for the j^{th} group using chance classification for all the three groups is almost zero. This provides good evidence that the quadratic LOO classification hit rates for the three groups (N, S, and H), in addition to the overall group hit rate, are all significantly higher than what one might expect to obtain by chance.

6.2.4.1 The final PDA classification model

The final property classification model consists of three quadratic composites to form a quadratic classification rule. Each composite is a composite of the five predictors in Set1 (NT, NP, AC, AFP, and LTVR). The coefficients of the three composites were obtained using the quadratic DA function, *classify*, from the statistical tool box in MATLAB. This

function returns the coefficients of three boundary planes where each plane separates two different groups.

To better illustrate the classification process using the boundary planes, a quadratic PDA classification model was built using only two predictors, AC and AFP, for simplification. This is just an example and is not the actual property classification model. Three scores S1, S2 and S3, were formed from the equations of the three boundary planes for the model. The boundary planes (S_1=0, S_2=0, and S_3=0) in the generated model are shown in Figure 6.2. Based on this figure, if an observation (AC, AFP) falls into the green area, it is classified as normal. If it falls in the blue area it is classified as suspicious. Finally, if it falls in the red area, it is classified as highly suspicious.

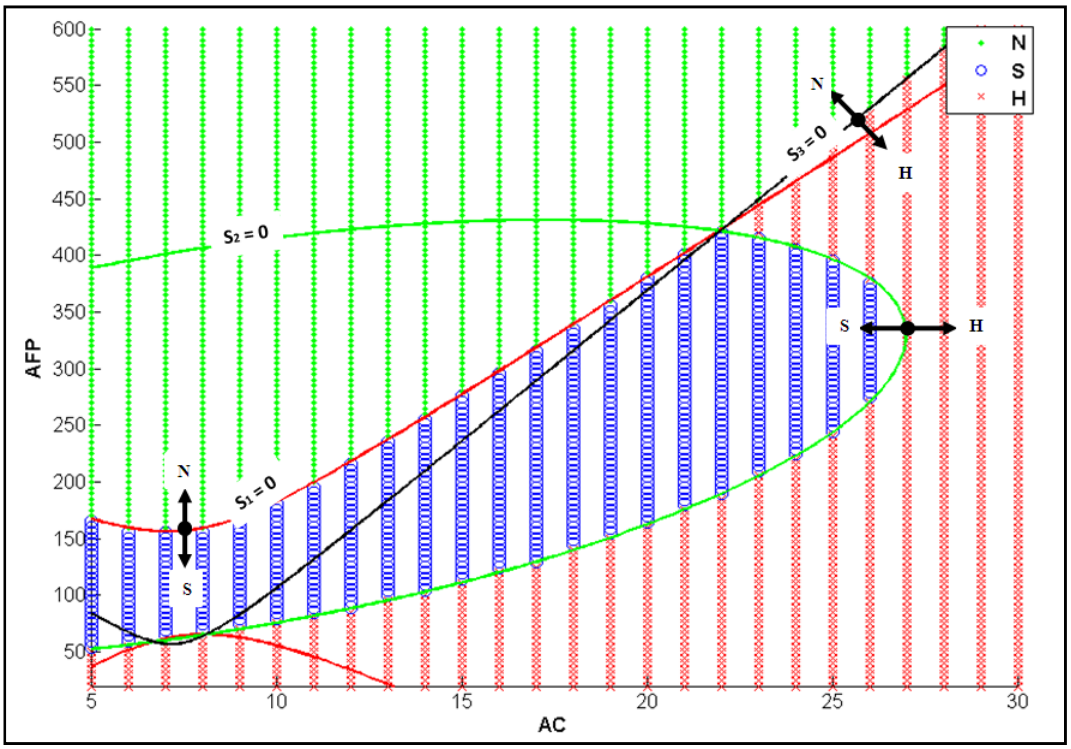


Figure 6.2: Example of three boundary planes obtained from quadratic PDA using only two predictors (AC and AFP).

So, in the final property classification model, which was built using all five predictors, three composite scores were formed (Z_1, Z_2, and Z_3) from the obtained coefficients from MATLAB. Z_1 is the score of the first composite, and $Z_1=0$ is the boundary plane between group S and group N. $Z_2=0$ is the boundary plane separates between groups S and H, and $Z_3=0$ separates between groups N and H. A plot for those planes is not included because each plane's equation is an equation of 5 variables.

Each of the composites (Z_1, Z_2 and Z_3) contains a constant C, a 5 by 1 matrix of linear coefficients **L**, and a 5 by 5 matrix of quadratic coefficients **Q**. So, each of the three scores is written as

$$Z_i = C_i + \mathbf{X}*\mathbf{L}_i + \mathbf{X}*\mathbf{Q}_i*\mathbf{X}' \text{ , } i=1,2,3$$

C_1, C_2, C_3, L_1, L_2, L_3, Q_1, Q_2, and Q_3 were all obtained from MATLAB and following are the obtained values for each coefficient.

C_1 = -24.0156
$C2$ = -6.06087
$C3$ = 17.95469

$$\mathbf{L_1} = \begin{bmatrix} 3.147743 \\ -4.38892 \\ -0.68905 \\ -0.07398 \\ 61.26865 \end{bmatrix}$$

$$\mathbf{L_2} = \begin{bmatrix} 2.468866 \\ -4.14194 \\ 0.710722 \\ -0.02201 \\ 21.91729 \end{bmatrix}$$

$$\mathbf{L_3} = \begin{bmatrix} -0.67888 \\ 0.246982 \\ 1.399769 \\ 0.051977 \\ -39.3514 \end{bmatrix}$$

$$\mathbf{Q_1} = \begin{bmatrix} 2.296996 & -2.80767 & -0.03952 & 0.012757 & -1.66432 \\ -2.80767 & 4.359078 & 0.060493 & -0.01072 & -0.87518 \\ -0.03952 & 0.060493 & 0.093291 & 0.002006 & -0.73416 \\ 0.012757 & -0.01072 & 0.002006 & -0.00044 & 0.069898 \\ -1.66432 & -0.87518 & -0.73416 & 0.069898 & -23.776 \end{bmatrix}$$

$$\mathbf{Q_2} = \begin{bmatrix} -0.74919 & 0.866998 & -0.06499 & 0.00915 & -2.56393 \\ 0.866998 & -0.25672 & 0.034028 & -0.00754 & 1.59609 \\ -0.06499 & 0.034028 & -0.01899 & 0.001385 & -0.40627 \\ 0.00915 & -0.00754 & 0.001385 & -0.00023 & 0.042371 \\ -2.56393 & 1.59609 & -0.40627 & 0.042371 & -13.8526 \end{bmatrix}$$

$$\mathbf{Q_3} = \begin{bmatrix} -3.04619 & 3.674664 & -0.02547 & -0.00361 & -0.89961 \\ 3.674664 & -4.6158 & -0.02647 & 0.003177 & 2.471266 \\ -0.02547 & -0.02647 & -0.11228 & -0.00062 & 0.327891 \\ -0.00361 & 0.003177 & -0.00062 & 0.000206 & -0.02753 \\ -0.89961 & 2.471266 & 0.327891 & -0.02753 & 9.92349 \end{bmatrix}$$

Finally, **X** is a 5 by 1 vector of the five predictors which describe the property in question. **X** should be entered as

$$\mathbf{X} = \begin{bmatrix} NT & NP & AC & AFP & LTVR \end{bmatrix}$$

To use this classification model, all the three scores Z_1, Z_2 and Z_3 should be calculated for an input property of five predictors **X**. Based on the signs of the three

scores which determine the location of **X** relative to each plane, the property described by **X** can then be classified into one of the three groups N, S, or H.

For example, if, for a certain property **P**, the signs of the three scores Z_1, Z_2 and Z_3 were calculated as positive, positive, and negative respectively, then **P** would be classified as suspicious. This classification involves three steps:

1. A positive sign of Z_1 means that **P** is on the suspicious side of the first separation plane and so it is suspicious and not normal.
2. A positive sign of Z_2 means that **P** is on the suspicious side of the second separation plane and so it is suspicious and not highly suspicious.
3. A negative sign of Z_3 means that **P** is on the highly suspicious side of the third separation plane and so it is highly suspicious and not normal.

Although Z_3 suggests **P** should belong to the group N, it was established from Z_1 that **P** cannot be classified as normal. Thus **P** would be classified as suspicious (S).

The complete classification rules based on the signs of the three scores are shown in Table 6.13.

Table 6.13: Predicted group of a property based on the signs of its three scores Z_1, Z_2 and Z_3.

sign of Z_1	sign of Z_2	sign of Z_3	predicted group
-	-	-	H
-	-	+	N
-	+	-	Does not apply
-	+	+	N
+	-	-	H
+	-	+	Does not apply
+	+	-	S
+	+	+	S

The final PDA classification model was implemented in a tool using the three score functions Z_1, Z_2 and Z_3 described above in order to validate the obtained coefficients. The tool was then tested on a set of 35 highly suspicious properties that were not included in the construction of the model. 30 properties out of the 35 were classified as highly suspicious, which gives a hit rate of 0.857. This result validates the hit rate of group H presented in Table 6.9.

In the following section, CART is used to build a classification model for properties in order to compare that model with the quadratic classification model presented in this section.

6.2.5 Classification of Properties using Classification and Regression Trees Method (CART)

In this section, the CART method is used to build a property classification tree that can classify properties into one of the three groups. In order to assess the best set of predictors, the three sets introduced in Table 6.5 were used to build three different classification trees. Then, resubstitution hit rates and LOO hit rates for the total group and the separate groups were calculated for the three trees in order to evaluate the performance and select one tree.

Table 6.14: Total group hit rates for the three CART classification trees generated using the three candidate sets.

Predictors set	*Resubstitution hit rate*	*LOO hit rate*
Set1	0.940	0.857
Set2	0.940	0.867
Set3	0.924	0.831

Table 6.15: Separate groups' resubstitution hit rates for the three CART classification trees generated using the three candidate sets.

Group	*Set1*	*Set2*	*Set3*
N	0.989	0.989	0.984
S	0.859	0.813	0.828
H	0.875	0.922	0.844

From the resubstitution and LOO hit rates for the total dataset (see Table 6.14), it is clear that Set1 and Set2 surpass Set3 in the total group hit rates. This conclusion is also supported by the separate group resubstitution hit rates presented in Table 6.15. As discussed before, the hit rate for group H is of a great importance for any suggested classification model. For the three trees presented here, it was obvious that group H hit rates generated from Set1 and Set2 are significantly higher than those from Set3. Finally, the LOO separate group hit rates show almost the same results as the resubstitution hit rate. It is obvious from Table 6.16, Table 6.17, and Table 6.18 that the separate group hit rate for groups S and H are higher for Set1 and Set2 than for Set3. Therefore, the model generated from Set3 was dropped and the assessment continued to decide between Set1 and Set2.

Based on the total group hit rates for both Set1 and Set2 in Table 6.14, we can see that resubstitution hit rates for both models are equal, while the LOO hit rate for Set2 is slightly better than the one for Set1. A look into the separate group hit rates (Table 6.15, Table 6.16, Table 6.17) show the same pattern repeated, where Set2, in most of the cases, outperforms Set1by a small fraction. In some cases, the hit rates are equal; and Set1 is better than Set2 only in the case of the group S hit rate, which can be seen in Table 6.15.

Looking in these two tables at the group H hit rate, which is the most important for this analysis, we can see that the rates are very close for both sets.

Table 6.16: LOO classification results obtained from CART using Set1 (NT, NP, AC, AFP, LTVR).

		Predicted Group				
		N	S	H	Total	Group Hit Rate
Actual Group	N	**177**	10	0	187	0.947
	S	12	**39**	13	64	0.609
	H	0	10	**54**	64	0.843
Total		189	59	67	315	

Table 6.17: LOO classification results obtained from CART using Set2 (NT, AC, AFP, LTVR).

		Predicted Group				
		N	S	H	Total	Group Hit Rate
Actual Group	N	**178**	9	0	187	0.952
	S	14	**41**	9	64	0.641
	H	0	10	**54**	64	0.844
Total		192	60	63	315	

Table 6.18: LOO classification results obtained from CART using Set3 (NP, AC, AFP, LTVR).

		Predicted Group				
		N	S	H	Total	Group Hit Rate
Actual Group	N	**179**	7	1	187	0.957
	S	12	**33**	19	64	0.516
	H	0	14	**50**	64	0.783
Total					315	

In conclusion, it appears that Set2 has higher hit rates than Set1. But the difference is so small that it is hard to make a general judgement that the model generated from Set2 will have better classification rates than Set1 when applied to unseen

observations. So, the final assessment was to apply both models on the dataset with 35 highly suspicious properties.

Finally, the two models were used to classify the test dataset that contains 35 highly suspicious properties. The results of this test show a hit rate of 0.714 for Set1 while the Set2 hit rate was 0.657. Evidently, the model generated from Set1 outperformed the model generated from Set2 in this test. Also, it is clear that both hit rates were smaller than the hit rates obtained from resubstitution and LOO methods, which might be due to tree over-fitting.

The final classification tree generated using Set1 is presented in Figure 6.3. Using this classification tree, any new property can be classified into one of the three groups. The classification process for a new property starts at the root node of the tree. At each node, the condition of that node is tested against the property, so the property will follow a path that will lead to a leaf node which determines the predicted group for it.

6.2.6 Discussion

As a conclusion from the results of the classification model obtained from quadratic PDA and the classification model obtained from CART, it was established that Set1 generated the best model for both methods, and so the two methods were compared based on the results from Set1.

Looking back at Set1 hit rates for quadratic PDA and CART, we can see that the LOO hit rate for CART (0.857) was clearly higher than the LOO hit rate for quadratic PDA (0.822). However, using the test set of 35 highly suspicious properties, the quadratic

PDA model was able to correctly classify 30 properties, while the CART model correctly classified 25 properties. This suggests that the test set hit rate for the quadratic PDA model (0.857) was significantly higher than the test set hit rate for CART (0.714). It also suggests that quadratic PDA provides a more robust and stable model in this simulated data set.

Based on this result, the final chosen model for the property classification problem was the model generated from quadratic PDA analysis using predictors of Set1 (NT, NP, AC, AFP, and LTVR). However, decision tree data mining methods may well produce superior results in other datasets.

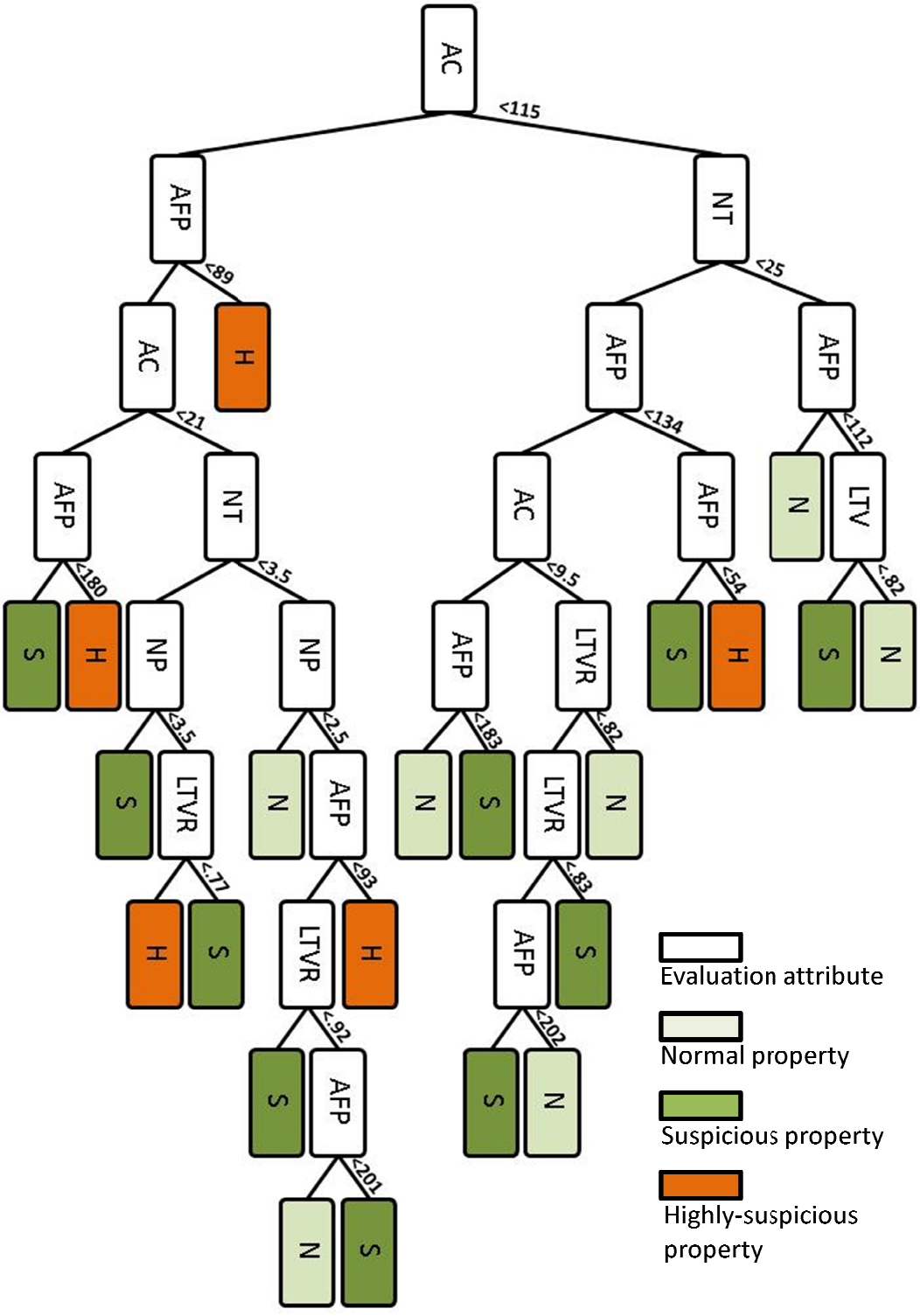


Figure 6.3: Property classification tree generated from CART using predictors of Set1

6.3 Outlier Detection of Land Grabbing in Post-Conflict Situations

This section provides an assessment for the use of outlier detection to highlight periods with suspicious activities in post-conflict situations. As mentioned in Section 3.3, there are a number of indicators that might reflect land grabbing in such situations. However, the only indicator considered in the following experiments is the exceptional high or low number of transactions during a short period of time.

As presented in Section 4.5, the adopted method for this problem is the entropy-based outlier detection proposed by He *et al.* (2005). Following in Sub-sections 6.3.1, 6.3.2, and 6.3.3, the author provides a brief discussion of the method, the formulation of the fraud problem to apply the entropy outlier detection method, and the obtained results.

6.3.1 Problem Formulation

The datasets used in the experiments are transactional datasets that contain land transactions that occurred during a two-year period. The detection problem is formulated to detect periods of fraudulent activities and not the fraudulent transactions themselves. More specifically, the goal is set to detect the days that might have encompassed fraudulent activities (i.e., fraudulent transactions) during the time period covered in the dataset.

The first step was to create a new dataset that summarizes the transactional dataset. The new dataset contains only two attributes, a date and the number of transactions that occurred on that date. The problem then can be formulated as finding the days with an exceptional number of transactions. These days are considered outliers according to the definition of outliers given in Section 4.3.

A dataset that contains many outliers will have a high degree of disorder. So, the target is to identify a subset of the dataset such that removing this subset from the data set will result in a lower degree of information disorder. That is where the concept of entropy can be adopted, as entropy can be used to measure the information disorder. Hence, entropy of the dataset is used as the objective function. The entropy *H(X)* is defined as shown in equation 6.6, where *X* is a random variable which is the number of transactions per day, *S(X)* is the set of values that *X* can take and *p(x)* is the probability function of *X*.

$$H(X) = -\sum_{x \in S(X)} p(x) \log_2 p(x) \tag{6.6}$$

The problem can be formulated as follows. Assuming I have a data set *DS* of *n* records, given an integer *k* that represents the expected number of outliers. The solution would be a subset of outliers *O* contained in *DS* such that the entropy *H(DS – O)* is minimized (He *et al.*, 2005). It is important to mention here that *k* has to be input to the algorithm to represent the desired number of outliers that the algorithm should look for. Thus we should have some idea of the number of outliers. A rule of thumb could be used for *k* if and the method can be applied starting with a small value of *k* and increasing it each time with evaluation of the results in each time.

6.3.2 Algorithm Implementation

The entropy-based outlier detection algorithm was implemented in a tool using C# programming language. A screenshot of the tool is presented in (*Appendix D*). There are two inputs into the algorithm, *k (a prior expected number of outliers)* and a summary

dataset of the land transactions dataset *DSS*. The expected output is a set *O* of *k* outliers that satisfy the objective function of minimizing the entropy of (*DSS*-*O*).

In the algorithm (see Figure 6.4), initially a new data set is selected from *DSS*. This data set is the outliers' data set *O* and contains *k* randomly selected records from *DSS*. Any record that is randomly selected from *DSS* is removed from it and stored in *O*. For processing the datasets, arrays of a simple data structure are used. The data structure contains two variables: RegistrationDate and NumberOfTransactions.

```
Input:      DSS    //the summary of land records dataset
            k      //number of desired outliers

Output:     O      //a subset of DSS with k outliers

Begin:
/* initialization of O */
        For counter equal 0 to k
                record = select a random record from DSS
                add record to O
                remove record from DSS

/* evaluating the initial entropy*/
        MinEntropy = Entropy(DSS)

/*iteration: minimizing the entropy for DSS*/
        foreach record x in O
                foreach record y in DSS
                        swap x and y
                        if Entropy(DSS) less than MinEntropy
                                MinEntropy = Entropy(DSS-y+x)
                        Else
                                Swap x an y

End
```

Figure 6.4: Entropy-based outlier detection algorithm.

The initial entropy of *DSS* is calculated and then an iterative process starts. First, a record is selected from *O*. This record is then swapped with a record from *DSS* and the entropy of *DSS* is measured again. If the new entropy is less than the entropy before the swap, the swap is considered final and the algorithm proceeds to the next record of *DSS*. When all the records of *DSS* are processed, a new iteration is started by selecting the next record of *O*, and this continues until the algorithm reaches the end of *O*.

6.3.3 Experimental Results

The entropy outlier detection algorithm was applied to three different datasets, LRDS1, LRDS2, and LRDS3. The simulation of these three datasets was discussed in Section 5.3.1. Table 6.19 below, which is a copy of Table 5.3, summarizes the three datasets. Each of the three datasets was first transformed into a new dataset that summarizes the original transactional dataset, as discussed in Section 6.3.1. The new datasets are LRDSS1 which summarizes LRDS1, LRDSS2 which summarizes LRDS2, and LRDSS3 which summarizes LRDS3.

Table 6.19: Summary for the three datasets simulated for post-conflict situations (copied from Table 5.3).

Dataset	*Number of days exhibiting normal number of transactions*	*Number of days exhibiting exceptional number of transactions (injected outliers)*
LRDSS1	748	32
LRDSS2	771	9
LRDSS3	702	28

For each dataset, the author did a number of runs of the algorithm using different values for *k*. Table 6.20 presents the results of all the experiments. The values used for *k* were selected to test two things: first, to assess the accuracy of the algorithm when the

value of *k* is less than the actual known outliers in a dataset; second, to assess the possibility of detecting 100% of the outliers at higher *k* values than the number of outliers. Figure 6.5, Figure 6.6 and Figure 6.7 plot some of the obtained results for the three datasets.

In general, outliers were identified in the datasets in most of the tests. However, the success rate of the algorithm in finding the outliers injected into the datasets varied based on the selected value of *k* and based on the pattern of transactions.

For example, in LRDSS1, it can be seen that at *k=20,* the algorithm identified 17 actual outliers and missed in 3 instances, which gives 85% accuracy. At *k=60*, the algorithm was able to identify 100% of the 32 outliers in LRDSS1.Almost similar results were achieved for LRDSS2.

It was noticed in LRDSS1 and LRDSS2 that using high values of *k* led to the detection of 100% of the introduced outliers. However, this conclusion could not be stated for LRDSS3, as can be seen from the results in Table 6.20. It was obvious that the number of outliers detected in LRDSS3 stopped increasing significantly after *k=40*, even with significantly high *k* values. For instance, the number of detected outliers increased only from 20 at *k*=40 to 23 at *k*=120.

Based on these results, it can be concluded that the performance of the algorithm changes according the distribution of the values in the dataset. In the first two datasets, LRDSS1 and LRDSS2, we can see from Figure 6.5 and Figure 6.6 that while the number of transactions for normal days is contained within a limited range of values (~20 - ~100), the number of transactions for the injected outliers falls outside that range. So, the

probability of having a number of transactions that falls within the normal range is much higher than the probability of a value falling outside that range.

From (6.6), it is clear that the higher *p(x)* is, the lower the entropy *E*, as *log(p(x))* approaches 0 when *p(x)* approaches 1. This means that in LRDSS1 and LRDSS2, removing the outliers is expected to cause the entropy to decrease, since we will be removing points with very low probability. This argument, however, cannot be made for LRDSS3.

Table 6.20: Entropy-based outlier detection results for the three datasets LRDSS1, LRDSS2 and LRDSS3.

Dataset	*Total outliers*	*k*	*Detected outliers*	*% outliers detected*
LRDSS1	32	20	17	53%
		40	29	90%
		60	32	100%
LRDSS2	9	7	4	57%
		10	6	67%
		20	8	89%
		40	9	100%
LRDSS3	28	20	18	64%
		40	20	71%
		60	21	75%
		80	21	75%
		120	22	79%

The number of transactions in LRDSS3 does not follow the same trend as in LRDSS1 and LRDSS2. Basically, the transactional trend in LRDSS3 follows the trend of the real estate market in Calgary for the years 2008 and 2009, as was introduced in Section 5.3.1. This trend is represented in the normal data points in Figure 6.7. It can be seen that the normal number of daily transactions is different during the different periods of the two years. So, what might be considered as a normal number of transactions in the

period around day 550 – which looks like a booming period – would be considered as an exceptional value in the period around day 400.

In general, the definition of a normal transaction in the case of LRDSS3 depends on the time of that transaction – unlike LRDSS1 and LRDSS2 where normal data points are expected within a fixed range of values for the whole two-year period. However, the time dimension is not considered in the entropy algorithm as, it deals only with the probability of a certain value to be found in the dataset regardless of the time. This explains the low detection accuracy in LRDSS3 even with high values of *k*.

For example, in Figure 6.7, it can be noticed that a high number of normal data points were identified as outliers around the top of the peak centred at day 550. Even though these were normal points, the probability of finding similar values in the dataset is low, and that is why the entropy algorithm detected these points as outliers.

On the other hand, days 340, 341 and 342 in Figure 6.6 were not identified as outliers even with k =120. However, from Figure 6.7, it is obvious that these three points are exceptional. The problem is that the probability of finding these values in the whole data set is high enough that other normal points in the LRDSS3 show greater outlying behaviour based on the definition of the entropy.

Based on this discussion, in the context of our formulation of the fraud problem in post-conflict situations, the entropy-based outlier detection algorithm cannot be guaranteed to detect fraudulent activities, and its accuracy depends in principle on the distribution of the observations in the dataset. In experiments on datasets that followed a linear trend, the detection accuracy was 100%; however, the algorithm accuracy was low

in datasets with a non-linear trend. One more limitation for this method is the need for a pre-defined number of outliers (k). In the experiments, the values of k used were similar to the number of outliers, since the number of outliers was known. When real data sets are used, a problem will be the establishment of k because the number of outliers is not known.

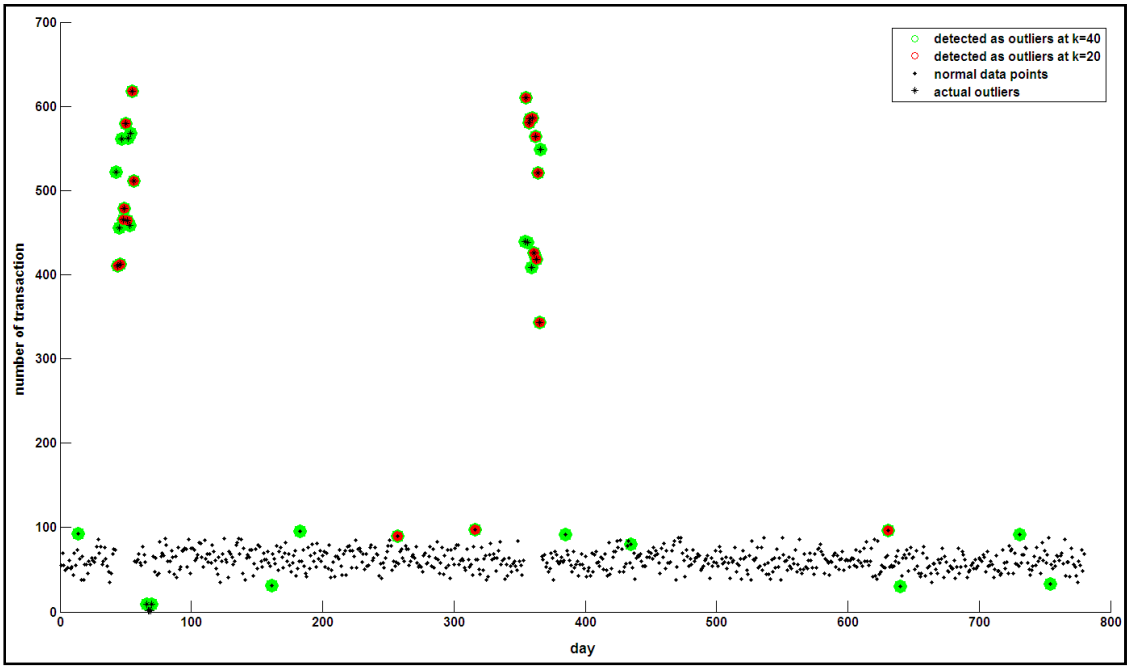


Figure 6.5: Entropy outlier detection results on LRDSS1 for k=20 and k=40.

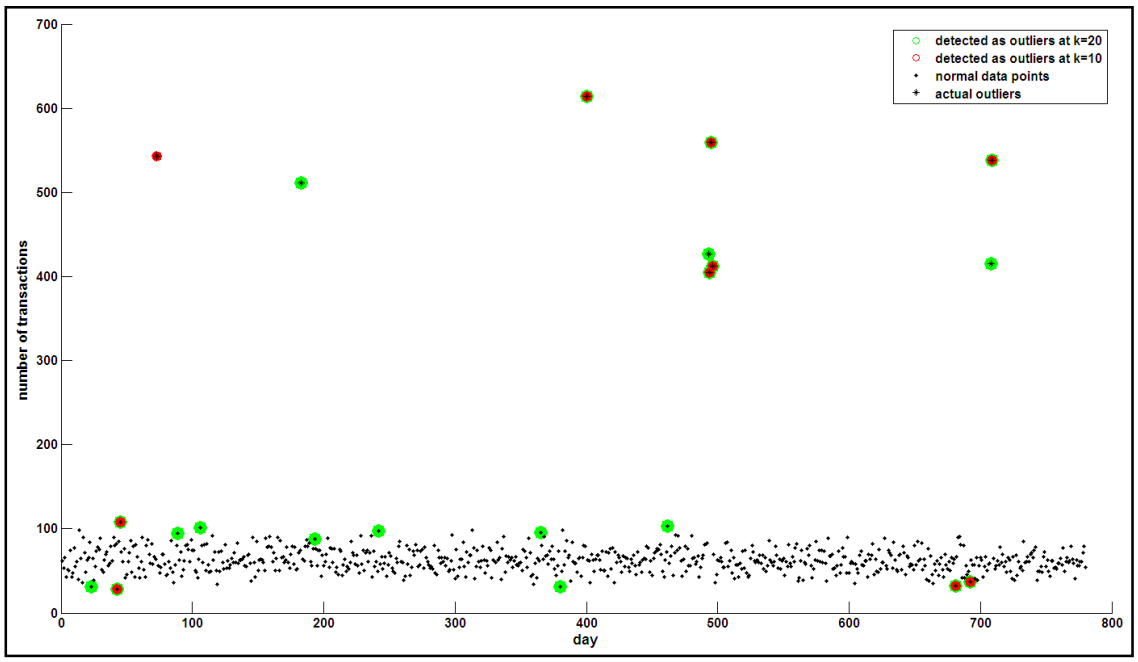


Figure 6.6: Entropy outlier detection results on LRDSS2 for k=10 and k=20.

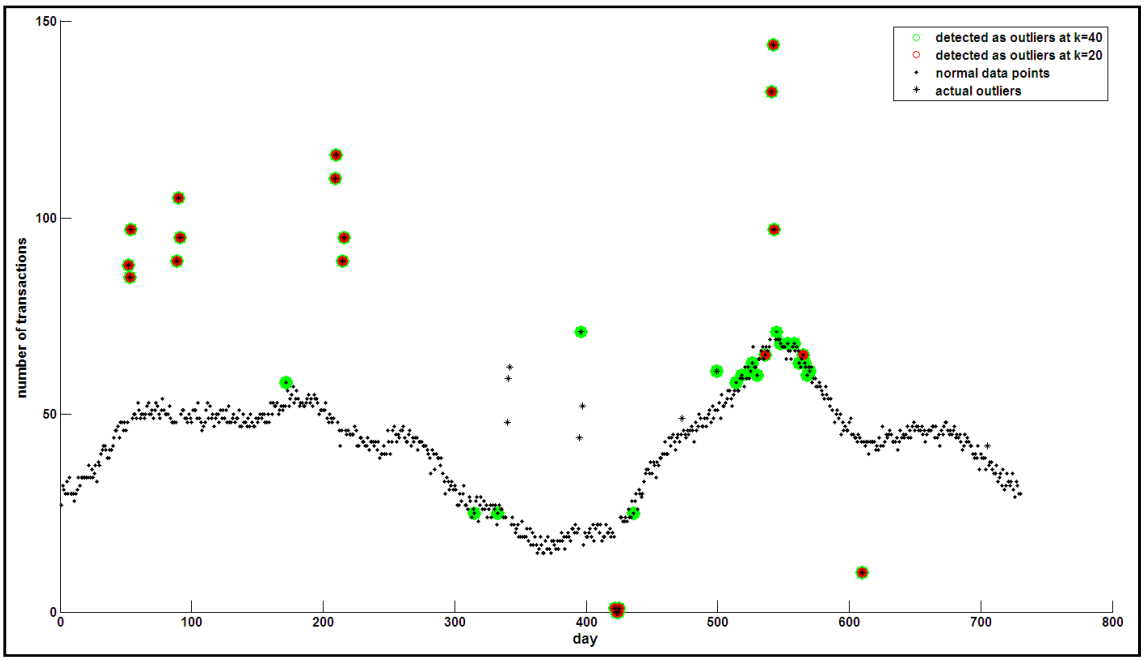


Figure 6.7: Entropy outlier detection results on LRDSS3 for k=20 and k=40.

6.4 Chapter Summary

This chapter reported on the experimental work done in this study. Mainly, the chapter was divided into two parts which were presented in Section 6.2 and 6.3. Section 6.2 reported on the construction of classification models to detect Oklahoma Flip and ABC-Construction fraud schemes in property transaction datasets. The goal of the generated models is to classify properties into three pre-defined groups based on transactional behaviours. The three classes are Normal, Suspicious, and Highly Suspicious. Section 6.3 reported on entropy-based outlier detection as a method to identify land grabbing patterns in post-conflict situations.

In Section 6.2, the results of two different classification models were presented and discussed. However, first the design of classification experiments and dataset preparation tasks were examined. Then results of three quadratic PDA classification models were analysed. Each of the models was built using a different set of predictors. The quadratic PDA rules were then presented for the best model. To validate the results of the quadratic PDA, a property classification tree was built using CART, which is the second adopted classification method. The results of CART were discussed and compared with quadratic PDA results. The results showed that the quadratic PDA model is superior to the classification tree for the simulated dataset. However, further testing should be done to persuasively postulate that one method is superior to the other.

In Section 6.3, results of the entropy-based outlier detection were presented. This method was used for detecting periods of suspicious activities, and more specifically, land grabbing patterns in post-conflict situations. The formulation of land grabbing as an

outlier detection problem was discussed first and then the developed entropy algorithm was presented. Finally, the results of applying the algorithm on three different datasets were examined. It was concluded at the end of the second part that the entropy method is not guaranteed to detect fraudulent activities, and that the accuracy of this method depends largely on the distribution of the dataset.

This chapter has addressed research activities 5 and 7. It has also fulfilled sub-objective 1.4.d.

Chapter Seven: Conclusions and Future Work

7.1 Introduction

This chapter summarizes the research presented in this thesis, discusses main conclusions, links these conclusions to the objectives outlined in Section 1.4 and provides recommendations for future work.

The research project explored the application of data mining methods to detect potential fraudulent activities in property transactions. This was achieved by performing a study on fraud in different land transaction situations, and then selecting and applying data mining methods to land record datasets in an attempt to detect fraud.

The study included gathering data on different fraud schemes through the minimal literature on the topic and interviews and email communication with international experts, investigating the scheme steps, and establishing sets of fraud patterns and indicators associated with the schemes. This comprehensive study of the fraud problem enabled the author to establish what to look for in land record systems to detect the identified fraud schemes.

To summarize, Chapter 1 briefly introduced the research problem and the proposed solution as well as setting out the objectives of this study. Chapter 2 reviewed the problem of fraud in land record systems. The nature of the problem was further explored using data from interviews the author carried out with experts.

Chapter 2 first examined the task of data analysis for land records. Then it described some of the fraud types and methods, followed by a presentation of current

endeavours to detect criminal investments and fraudulent activities in the property market. The examined fraud types and methods were categorized into two groups based on the contextual situations within which they might take place. The two groups are: fraud in post conflict situation and fraud in real estate transactions in developed countries.

In Chapter 3, a variety of methods that are available to fraudsters were examined. The chapter focused on analysis of two fraud schemes in real estate transaction (the Oklahoma Flip and the ABC-Construction), and one fraud pattern in post-conflict situations (land grabbing). These three fraud problems were selected for the experimental work and their fraud patterns and indicators were examined.

Chapter 4 presented a review for data mining in general, focussing on reviewing methods used in this study. Methods of PDA, CART, and entropy based outlier detection were reviewed, argued and adopted as methods in the experimental work.

Chapter 5 described the implementation of a land record simulator developed in this research to overcome the lack of access to land record datasets. The chapter also presented the datasets simulated and used in the study's experiments.

Chapter 6 reported on the experimental work. It described the application of classification methods (PDA, CART) and entropy based outlier detection on land records to detect fraudulent transactions.

Finally, this chapter presents the findings of this study and the key conclusions made throughout the thesis. It also makes recommendations for future work.

7.2 Conclusions

This section presents a synopsis of the findings and links them to the primary objective and the sub-objectives established in the thesis.

> ***Sub-objective (a):*** *Identify different fraudulent activities in land record datasets, in a variety of contexts where these activities may take place:*

This objective was achieved by studying literature and analyzing interviews with land transaction and market experts. A group of fraud methods were identified. It was found that property fraud is a serious problem in many situations and especially mortgage fraud in countries with strong real estate market. Criminal investment in the real estate market is causing financial institutions in countries such as Canada to lose millions of dollars. Yet, few studies are looking into identifying fraudulent activities by using data mining techniques to analyse land records. . Reports and studies focus on investigating fraud methods, and developing strategies to reduce the fraud problem. It was concluded that two principal problems explains the limited number of studies.

1- Only a small percentage of the fraud cases, especially in the case of mortgage fraud, are reported. Financial institutions tend to hide such cases to protect their reputation. Also, in some cases where family members are involved in a fraud case, the victim tends not to report it because they do not want to report on members of the same family.
2- The datasets required to develop detection methods are seldom easily accessible to people outside the institutions which own the data.

Sub-objective (b): *Identify suitable data mining techniques that may help in detecting some of the fraud activities found in (a)*:

This objective was achieved by:

1- Researching fraud detection studies in different fields and the data mining methods that are used in these studies.
2- Analyzing some of the identified fraud schemes and studying their effect on the records inside the land record datasets.

A fraud scheme can be tracked using land records if the effects of that scheme on the records can be extracted. Deriving fraud indicators from the schemes facilitate the formulation of a detection problem from a data mining perspective.

There are many fraud indicators that can lead to the identification of suspicious activities in land transactions. However, those indicators come from a range of systems and cannot be tracked in the registration system only. For example, tax records, credit card histories, and personal information are not available in registration systems. Integration between the different systems or at least interoperability of those systems would enhance any fraud detection process.

Based on the indicators used in the thesis, the problem of fraud detection in land record systems can be formulated in different ways to fit a range of data mining tasks. Three methods from classification and outlier detection were used in the experiments. However, the three methods are not the only ones that can be applied. For instance, the

problem of detecting the fraud schemes examined in this study may be formulated as a clustering problem. Also, only two classification methods were used and there are many other classifiers that might work for the same problem. These were not investigated in the study because of time constraints.

Sub-objective (c): *Design and develop a data simulator to generate land record datasets*:

This objective was reached by the development of a fully working land records simulation system, which was used to generate synthetic datasets for the experiments. Developing the simulator helped in the achievement of the study results. However, it also created limitations on the optimality of the results obtained from the selected detection methods. Essentially, these limitations are:

1- Real datasets are more complex than the simulated datasets. In the simulation processes developed in the land records simulator, the author tried to capture real life factors that influence land transactional activities as best as possible. However, there are limitations to how much simulated data can mirror actual transaction data characteristics.

2- Outliers and fraud cases in the simulator are generated based on the indicators derived from analysis of other studies and interviews. Those outliers are abnormalities in the records. However, it is still expected that those outliers do not reflect actual outliers 100% or might be applied in one situation but not globally. So, it is expected that characteristics of real dataset outliers to be different.

It is concluded that both real datasets and simulator datasets have pros and cons and the best approach would be to use a combination of both in generating and testing any fraud detection system.

> ***Sub-objective (d):*** *Identify existing tools to apply the methods found in (b) above and develop tools where it appears that relevant tools do not exist:*

This objective was achieved by applying three different data mining methods on simulated datasets. MATLAB was used to apply quadratic PDA and CART methods, while a tool was developed for the entropy based outlier detection algorithm.

Entropy based outlier detection was applied on three different datasets to discover some fraud patterns that take place in land records using post conflict as a contextual situation. Results showed that detection accuracy largely depends on the trend of the transactions in a dataset. In experiments on datasets that followed a linear trend, the detection accuracy was a 100%; however the algorithm accuracy was low in datasets with a non-linear trend.

It was also found, that a major limitation of this method is the need for a pre-defined number of outliers (k). In the experiments conducted in this study, the values of k used, were similar to the number of outliers, since the number of outliers injected into the dataset was known. When real data sets are used, a problem will be the establishment of k because the number of outliers is not known. However, this can be solved by iterating through different values of k.

Applying entropy based outlier detection techniques in a real world situation may help in identifying the probable days in which illegal land transactions have taken place. As a result, less effort is needed to examine the suspicious transactions because far fewer documents need to be examined manually instead of conducting and extensive examination on every document.

With regard to the second problem, which is property fraud in stable real estate markets, two classification methods (quadratic PDA and CART) were applied on property transactions data that incorporate fraudulent transactions. The experiments showed that PDA is superior over CART, However, this conclusion cannot be generalized beyond the dataset simulated for this project. Both PDA and CART generated high classification accuracy and are expected to perform well in a real property classification system that follows the same formulation of our problem. However, precise general models cannot be formulated since the fraud indicators' attributes may differ in different locations and situations.

In general, the suggested classification of real estate object methods (Normal, Suspicious, Highly Suspicious) could be applied to increase the security of ownership and help in preventing fraud attempts. In today's fast growing real estate market, different racketeering schemes are used, Oklahoma Flips and ABC-Constructions being the most common.

It appears that PDA is one of a number of methods that may be suitable for flagging properties that are subject to Oklahoma Flip and ABC-Construction schemes. Generating property classification models may help in predicting suspicious properties in

real time once a transaction takes place. As a result, investigators can focus their investigation on fewer cases rather spending lots of time trying to analyze transactions manually.

> ***The primary objective:*** *To explore the use of data mining in land record systems and to develop knowledge of where and how data mining can be applied and integrated into these systems, to contribute to the discovery and alleviation of fraud in land and property transactions.*

With regard to the main objective, this thesis identified different fraud activities in different situations. The effect of those activities on the underlying databases was examined and fraud indicators and patterns derived. Using fraud indicators and patterns, it was possible to use data mining methods to detect the fraudulent activities.

So, using data mining should facilitate the building of fraud detection models for land transactions that can then be integrated in the registration systems and act as alarm systems.

7.3 Future Work

This section proposes some ideas that warrant further investigation to guide future development on this research problem.

First, there are still various fraud schemes that need to be investigated. Many fraud schemes have been identified in this study and some of them were investigated in

detail to find patterns and indicators. However, not all criminal activities could be investigated during this research due to scope and time constraints. The author recommends conducting research to establish a list of fraud schemes and criminal activities in real estate markets. Along with creating such a list, future research should focus on developing lists of fraud indicators and patterns that would allow the tracking of the schemes.

The author suggests including more fraud indicators in the development of any future fraud tracking system, which should improve the detection accuracy. The problem is that those indicators might come from different systems (registration systems, banking systems, and credit card systems). So, to achieve optimal results in detecting frauds, there should be some integration or interoperability between the different systems and a fraud alerting systems should have access to the different systems.

Another area of future development relates to the use of datasets. In this area, the author recommends the using real datasets and synthetic datasets.

With regard to the generation of synthetic datasets, further development on the simulator is required. First, the simulation process can be improved by providing the ability to generate specialized datasets based on the local situation. This can be achieved by allowing more factors to be considered in the simulation. The simulator developed in this study gives the user the option to provide inputs such as market trend, housing prices, population, simulation dates, number of land parcels, and area of land parcels. One important option that may improve the simulation is to allow for the distribution of the prices based on the location where adjacent properties have similar prices. The spatial

attribute of a property is mentioned in different reports as an indicator of fraud if a property is valued much higher or lower than adjacent properties. The location coordinates is provided in the simulator, however it is only used for mapping purposes.

A second aspect of the simulation that need further work is the generation of fraud cases. Currently, the simulator allows only for the three schemes used in the experiment to be generated. With the identification of new fraud methods, the simulator should be updated to generate them.

Finally, in order to apply the results of this study, the methods proposed should be implemented and integrated into registration systems. So, an important area that needs to be investigated is the development of conceptual model how can this integration be achieved.

References

Alberta Registries, 2002, Alberta Land Titles Spatial Information System (SPIN 2), Alberta,
viewed 3 March 2010, < https://alta.registries.gov.ab.ca/spinii/logon.aspx>.

Angiulli, F., Basta, S. & Pizzuti, C., 2006. Distance-based detection and prediction of outliers. Knowledge and Data Engineering, IEEE Transactions on, vol. 18, no. 2, pp.145-160.

Anyanwu, M N, & Shiva, S G, 2009, Comparative Analysis of Serial Decision Tree Classification Algorithms, International Journal of Computer Science and Security (IJCSS), vol. 3, no. 3, pp. 230 – 240.

Auditor General of Alberta 2010. System Audits – Current Year Audits, in Report of the Auditor General of Alberta. April 6, 2010. Edmonton, Alberta, Canada. p 105 – 113

Augustinus, C., and Barry, M., 2006, Land Management Strategy Formulation in Post Conflict

Societies, Survey Review, 38(302), 668-681

Barbra, D, Chen, P 2000, "Using the fractal dimension to cluster datasets", *in Proceedings of ACM KDD 2000,* pp. 260-264.

Barnett, V, and Lewis, T, 1994, Outliers in statistical data, John Wiley.

Barry M., 2008. Multimedia Data In Land Records Systems: Field Trials in Nigeria. In: Canadian Hydrographic Conference and National Surveyors Conference. Victoria, Canada, 5-8 May 2008.

Barry M., 2009. Land Administration Strategy Formulation using GIS in Post Conflict Hargeisa, Somaliland. Surveys and Land Information Science. 69(1), March 2009, 39-52.

Barry M., 2009. Land Administration Strategy Formulation using GIS in Post Conflict Hargeisa, Somaliland. Surveys and Land Information Science. 69(1), March 2009, 39-52.

Barry, M, Hunter A, Muhsen A 2007. Scalable Land Tenure Record Systems. Proceedings of Workshop. Informal Settlements - Real Estate Markets Needs related to Good Land Administration & Planning, 28-31 March 2007, Athens, FIG Com3, UN ECE Working Party on Land Administration, UN ECE Committee on Housing and Land Management

Barry, M., and Khan, K., 2005, Law and Policy Implications of Multimedia Land Records: The Talking Titler Project, FIG Working Week and 8th International Conference of the Global Spatial Data Infrastructure, Cairo, Egypt, 16-21 April 2005

Ben-Gal, I, 2005, 'Outlier Detection', in Maimon, O and Rockach, L, (Eds.) Data mining and knowledge discovery handbook: a complete guid for practitioners and researchers, Kluwer Academic Publishers.

Berkhin, P, 2006, A Survey of Clustering Data Mining Techniques, Accrue software Inc. Viewed 12 April 2011, <http://www.accrue.com/products/rp_cluster_review.pdf>

Bianco M K, 2008, money laundering and mortgage fraud: the growth of merging industry, CCH Mortgage Compliance and Bank Digest.

Bouckaert, R R, Frank, E, Hall, M A, Holmes, G, Pfahringer, B, Reutemann, P, and Witten. I H, 2010, WEKA-experiences with a java open-source project. Journal of Machine Learning Research, vol. 11 (2010) pp. 2533-2541

Bourn, J 2006, International Benchmark of Fraud and Error in Social Security Systems, report by the comptroller auditor general, National Audit Office, London.

Bramer, M., 2007. Principles of data mining. Springer. 2007

Breunig, M M, Kriegel, P, Ng, R T, & Sander, J, 2000. "LOF: identifying density based local outliers". In Proceedings of the 2000 ACM SIGMOD international conference on management of data, pp. 93–104, Dallas, USA.

Burdeaski, T, 2000, "Evaluating univariate, bivariate and multivariate normality using graphical and statistical procedures", *Multiple Linear Regression Viewpoints*, vol. 26, no. 2, pp. 15-28.

CBS 2010, Waarde onroerende zaken: Woningen, niet-woningen, gemiddelde woningwaarde. Viewed on the 25 October 2010, <http://statline.cbs.nl/StatWeb/publication/?VW=T&DM=SLNL&PA=37610&LA=NL>.

CISC 2007. Mortgage fraud and organized crime in Canada. Criminal Intelligence Service Canada – Central Bureau, Intelligence Analysis and Knowledge Development Branch. November 2007.

CREB, 2010, Calgary Real Estate Board, Housing Statistics Archives, viewed 10 November 2010, < http://www.creb.com/public/seller-resources/housing-statistics-archives.php>.

CTV News 2005, 'Stealing Home', *Age* 21 March 2005, Viewed 20 March 2010 <http://www.ctv.ca/CTVNews/WFive/20050321/wfive_stealinghome_050318/ >

Daudelin, J., 2003. Land and Violence in Post-Conflict Situations. Report prepared for the North-South Institute and the World Bank. The Norman Paterson School of International Affairs, Carleton University, Ottawa. May 26, 2003.

Department of Rural Development and Land Reform, 2009, Deeds Registration, Republic of South Africa, viewed 26 August 2009, <http://www.dla.gov.za/land_planning_info/deeds_registration1_copy(1).htm>

Dudoit, S, Fridlyand, J, and Speed, T P, 2002, comparison of discriminant methods for the classification of tumors using gene expression data, Journal of American Statistical Association, vol. 97, no. 457. pp 77 – 87

Dunham, M, H, 2003, Data mining introductory and advanced topics, Prentice Hall/Pearson Education.

Enemark, S. (2005). The Land Management Perspective - Building the Capacity. ITC Lustrum Conference, 14-16 December 2005.

Fawcett, T, & Provost, F 1997, 'Adaptive fraud detection', *Data-mining and Knowledge Discovery*, vol. 1, no. 3, pp. 291-316.

Ferwerda, H., Staring, R., de Vries Robbe´, E. and van de Bunt, J. 2007. Malafide Activiteiten in de Vastgoedsector. Een Exploratief Onderzoek naar Aard, Actoren en Aanpak, SWP, Amsterdam.

Financial Crimes Enforcement Network, 2006, Mortgage loan fraud - An industry assessment based upon suspicious activity report analysis, Office of regulatory analysis, regulatory policy and programs division, Financial Crimes Enforcement Network, United States Of America, November 2006.

Future and Commodity Market News, 2009, Afghanistan: Counterfeit deeds enable sales of public lands, Kabul, 20 August, viewed 29 August 2009, <http://news.tradingcharts.com/futures/8/5/128108558.html>

Haanen, A, Bevin, T & Sutherland, N, 2002, e-Cadastre - Automation of the New Zealand Survey System. Presented at the Joint AURISA and Institution of Surveyors Conference, Adelaide, South Africa.

Hallett, S H, Ozden, D M, Keay, C A, Koral, A, Keskin, S & Bradley, R I, 2003. A land information system for Turkey - a key to the country's sustainable development. Journal of Arid Environments, vol. 54, no. 3, pp. 513-525

Han, J, & Kamber M, 2006, Data mining: concepts and techniques – second edition, Diana Cerra, United States of America.

Hawkins, D, M, 1980, Identification of outliers, Chapman and Hall, New York.

Hawkins, S, He, H, X, Williams, G, J, Baxter, R, A 2002, "Outlier detection using replicator neural networks", *in Proceedings of the 5th conference on Data Warehousing and Knowledge Discovery,* France, Aixen Provence.

Hay, G., and Hall, G. B. 2009. Architecture for an open source land administration systems. FIG working week, Eilat, Israel, 3-8 2009.

He, Z, Deng, S, and Xu, X 2005, "An optimization model for outlier detection in categorical data*", in Proceedings of the ICIC 2005 conference*, August 23-26, Hefei, China, pp. 400-409.

Hespanha, J, van Bennekom-Minnema, J, van Oosterom, P, Lemmen, C, 2008. The Model Driven Architecture Approach Applied to Land Administration Domain Model Version 1.1 - with Focus on Constraints Specified in the Object Constraints Language. FIG working week 2008, Stockholm, Sweden

Hu, T, Sung, S, Y 2003, "Detecting pattern-based outliers", *Pattern Recognition Letters*, vol. 24, no. 16, pp. 3059-3068.

Huberty, C J, Olejnik, S, 2006, Applied MANOVA and discriminant analysis, second edition, New York: Wiley.

Hunter, A 2007, 'Sensor-based Animal Tracking', PhD thesis, University of Calgary, Calgary.

Jiang, F, Sui, Y, & Cao, C 2010, "An information entropy-based approach to outlier detection in rough sets". *Expert Systems with Applications*, vol. 37 no. 9, pp.6338-6344.

Johnson, R, Applied multivariate statistical analysis, Prentice Hall.

Knorr, E, Ng, R, 1998, "Algorithms for mining distance-based outliers in large datasets", In proceedings of the 24th international conference of Very Large Data Bases (VLDB), pp. 27-27.

Kontrimas, V. and Verikas, A. 2006, 'Tracking of Doubtful Real Estate Transactions By Outlier Detection Methods: A Comparative Study', Information Technology and Control, vol. 35, no. 2, pp. 94-104.

Kontrimas, V. and Verikas, A. 2011, 'The mass appraisal of the real estate by computational intelligence', Applied Soft Computing, vol. 11, no. 1, pp.443-448.

Kreling, T, & Meeus, J, 2008, Waakhond zit zelf in 'fout' pand, NRC Handelsblad newspaper, 7 June, viewed 12 January, 2011, <http://www.nrc.nl/binnenland/article1908494.ece/Waakhond_zit_zelf_in_fout_pand>

Lemmen, C., and van Oosterom, P., 2006. Version 1.0 of the core cadastral domain model. Paper presented at XXIII FIG Congress – shaping the change, Munich, Germany.

Lemmen, C., Augustinus C., van Oosterom P., and van der Molen P. 2007. The social tenure domain model – design of a first draft model. Paper presented at Strategic Integration of Surveying Services, FIG Working Week 2007.

Lewis, D 2004, 'Challenges to Sustainable Peace: Land Disputes Following Conflict', Proceedings of Symposium of Land Administration in Post Conflict Areas, FIG Commission 7, Geneva, Switzerland, pp 15-25.

Li, S., Lee, R. & Lang, S.-D., 2006. Detecting outliers in interval data. In Proceedings of the 44th annual southeast regional conference. Melbourne, Florida: ACM, pp. 290-295.

Maggio J, E, 2008, Private Security in the 21st Century: Concepts and Applications, Jones and Bartlett Publishers, United States of America.

Manly, B F J, 2004, Multivariate Statistical Methods: a primer, 3rd edition, Chapman & Hall/CRC

Mingers, J., 1989, An empirical comparison of selection measures for decision-tree induction, Machine Learning, vol. 3, pp. 319-342.

Mirkin, B, 2011, Core Concepts in Data Analysis: Summarization, Correlation and Visualization, ed. 1, Springer.

Montia, G, 2010, Average loan-to-value ratio eases to 70%, Finance Markets, 25 January 2010, viewed 23 December 2010, < http://www.financemarkets.co.uk/2010/01/25/average-loan-to-value-ratio-eases-to-70/>.

Muhsen, A-R, and Barry, M., 2008, Technical Challenges in Developing Flexible Land Records Software, Surveys and Land Information Science, 68(3), 171-181

Muhsen, A-R. (2008). Developing Multimedia Land Record Systems. MSc. Thesis. University of Calgary.

Nelen, H., 2008. Real estate and serious forms of crime. International Journal of Social Economics, Vol. 35 No. 10 pp 751-761.

Niasbit, J., 1982. Megatrends: Ten New Directions Transforming Our Lives, Warner Books 1982.

Nogueira, A, de Oliveira, M R, Salvador, P, Valadas, R, Pacheco, A, 2005, Classification of Internet users using discriminant analysis and neural networks. Next Generation Internet Networks 2005, vol., no., pp. 341- 348

Nyerges, T L, 1989, Information Integration for Multipurpose Land Information Systems, URISA Journal, vol. 1, no. 1, pp. 27-38.

Ohnishi, T, Mizuno, T, Shimizu, C, and Watanabe, T, 2010, “On the evolution of the house price distribution”, Inflation Dynamics - understanding inflation dynamics of the Japanese Economy, working paper 2010, no. 56.

Pollakowski, H O, and Ray, T S, 1997, Housing Price Diffusion Patterns at Different Aggregation Levels: An Examination of Housing Market Efficiency, Journal of Housing Research, vol. 8, no. 1, pp 107-124

Ramaswamy, S, Rastogi, R, Shim, K, 2000, “Efficient algorithms for mining outliers from larg datasets”, in proceedings of ACM SIGMOD international conference on management of data, Dalas, Texas.

Roux, L and Barry, M 2001. Using Video Imagery in Land Tenure Information Systems: A Study of the Algeria Communal Property Association. CONSAS 2001 Conference, Cape Town.

Shannon, C E, 1948, A mathematical theory of communication, Bell System Technical Journal, vol. 27, pp. 379-423.

Tan, P, Steinbach, M, Kumar, V, 2005, Introduction to data mining, Pearson Addison Wesley.

Teranet, 2006, The Teranet National Bank House Price Index, Developed by Teranet in alliance with National Bank of Canada, viewed 12 December 2010, <http://www.housepriceindex.ca/documents/MethodologyEN.pdf>.

The Law Society of Upper Canada 2004, Practice Tips: Recognizing Fraud in Real Estate Transactions, The Law Society of Upper Canada, viewed 29 October 2010 <http://rc.lsuc.on.ca/pdf/fightingRealEstate/july2304_fraud_indicators.pdf >

Troister, S, H, Cohen T, M, LLP A, 2006, Mortgage fraud and discharge issues. Uniform Law Conference of Canada. Edmonton, Alberta. August.

Unger, B., Ferwerda, J., Trouw, J., Nelen, H., and Ritzen, L. 2010. Detecting Criminal Investments in the Dutch Real Estate Sector. Study prepared by the Dutch Ministry of Finance, Justice and Interior Affairs. 19th of January, 2010.

Van Oosterom, P., and Lemmen, C., 2002. Towards a standard for the cadastral domain: Proposal to establish a core cadastral data model. Paper presented at COST Workshop 'Towards a Cadastral Core Domain model', Delft, the Netherlands.

Van Oosterom, P., Lemmen, C. Ingvarsson, T., van der Molen, P., Ploeger, H, Quak, W. Stoter, J., and Zevenbergen, J., 2006. The core cadastral domain model. Computers, Environment and Urban Systems, vol 30 (2006) 627-660

Watkins, P., 2007. Fraud in conveyancing, a paper presented at the Australian Institute of Conveyancers 2007 national conference, March 2007.

Wily L A, 2004, 'Putting Land Registration in Perspective: The Afghanistan Case', Proceedings of Symposium of Land Administration in Post Conflict Areas, FIG Commission 7, Geneva, Switzerland, pp 71-93

Witten, I, H, & Frank, E, 2000, Data mining: practical machine learning tools and techniques with Java implementations, Morgan Kaufmann, United States of America.

Zevenbergen, J. & van der Molen, P., 2004. Legal aspects of land administration in post conflict areas. Symposium on Land Administration in Post Conflict Areas, Geneva, April 29 – 30.

Appendix A: Conceptual Models for Land Management

Land Administration domain Model (Core Cadastral Domain Model (CCDM)

The Core Cadastral Domain Model (CCDM), was suggested as a standard model for the cadastral domain at the FIG Congress in Washington in 2002 (van Oosterom and Lemmen 2002). The first mature version (version 1.0) was presented by Lemmen and van Oosterom (2006) which evolved by Hespanha *et al.* (2008) to version 1.1 - Land Administration Domain Model (LADM). The main incentives behind developing CCDM and LADM are to maintain basis for efficient and effective cadastral system development and to enable the communication of cadastral data between different systems based on a common ontology (van Oosterom et al 2006).

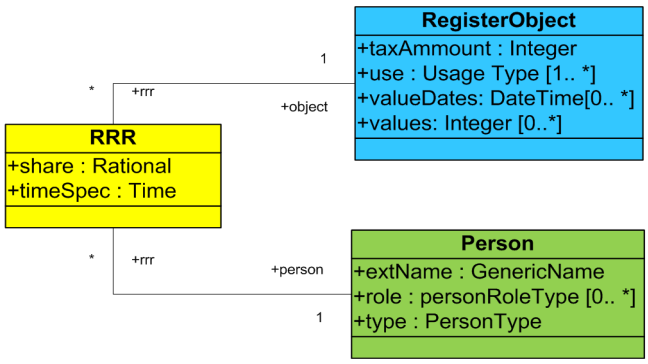


Figure A.1: The three core classes of the CCDM and LADM (after (Lemmen and van Oosterom 2006 and Hespanha et al (2008))).

Three main classes represent the core of the CCDM as shown in Figure A.1, Person, RegisterObject, and RRR (Right, Restriction, and Responsibility). The full design of the model is organized in 5 UML packages each represents an independent aspect.

These packages are Person aspects, RigesterObject and Immovable class specializations, legal/administrative aspects, surveying aspects, and Geometric/topological aspects.

Social Tenure Domain Model (STDM)

This model was presented by Lemmen *et al.* (2007) as a solution for some of the problems that cannot be addressed by CCDM. Basically it is a specialization of CCDM for situation that cannot be handled with CCDM like situations in developing countries, post conflict areas, and informal settlements (Lemmen *et al.* 2007). In other words, STDM aims to model the person-land relationship regardless of its formal/legal status. This model encompasses the main three classes of CCDM and LDM with some changes. The core classes for the STDM are shown in Figure A.2.

Two main amendments are creating the difference between CCDM and STDM. First is that the RRR class in the original CCDM is SocialTenureRelation in STDM and is based on inventories as a source for a large variety of rights, social tenure relations and claims that may exist in the areas where this model could be applied. Some of the social tenure relations that are included in the lookup tables are ownership, apartment right, informal type, customary type, disagreement, conflict, and many others. Especially it is Ramadan time

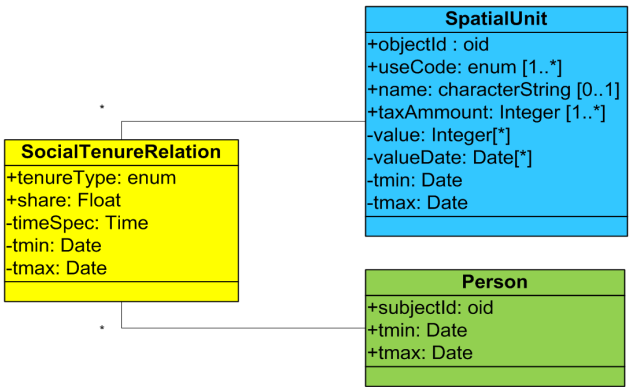


Figure A.2: The core classes of STDM (after (Lemmen *et al*. 2007)).

STDM tries to cover the full spectrum of formal, informal and customary rights. This design enables the model to cover a wider spectrum of situation than it predecessors CCDM and LADM. However, this brings more complications to information representation inside the model and makes it harder to develop supporting tools.

Talking Titler Model

Talking Titler was developed initially as a mean to incorporate multi-media data, such as video and audio clips, in land record system. This idea was then extended to synthesise data from a variety of sources in different contexts of situations (Barry and Khan, 2005; Augustinus and Barry, 2006; Barry *et al.*, 2007 and Muhsen and Barry, 2008).

Muhsen (2008) develops a new methodology for Land Information System (LIS) software development in uncertain situations based on a flexible data model with the Talking Titler conceptual model in the core. The idea behind this model is to start with an initial simple model; the initial model could be used for rapid data collection in uncertain situation, and then to evolve through time to be more suitable for the situation.

The initial design of the developed model consists of three general abstract classes which are Person, Land Object and Media and this model is called the three-class model (Muhsen, 2008). Figure A.3 shows the high-level conceptual view of the three-class model which was developed as part of the Talking Titler project (Muhsen and Barry 2008).

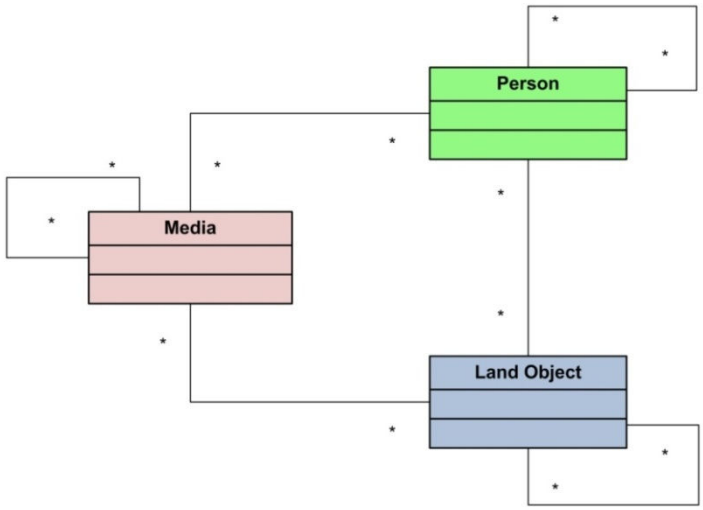


Figure A.3: The Talking Titler Three-class model (after (Muhsen and Barry 2008))

What make this system unique from the other models are the relations among its classes which allow the representation of any relation on the real world into the model. Also, the model represents rights, restrictions and responsibilities as a relation between Person and Land Object. Finally, the model introduces the Media class to capture information that cannot be modelled easily such as modelling all relationships of all persons or relevant land objects that appear in a video clip (Muhsen, 2008).

Talking Titler model enables the collection of more information in various forms. For example, this model provides great flexibility in allowing information to be collected in plain text to collect as much data as possible (Muhsen, 2008). A problem here is the

usability of this information when the size becomes larger and larger. A data mining tool will facilitate the use of the information gathered in the system by providing means to analyze information, find patterns, and extract knowledge. Also, data mining enables the extraction of information from unstructured data such as the unstructured text used in the some parts of the Talking Titler model. This information can then be transformed into a structural format which will make it more useful and easier to use.

Appendix B: Q-Q Plots for the Five Individual Predictor Variables

This appendix shows the Q-Q plots for the 5 predictor variables used for property classification. Each of the plots in the following figure represents the Q-Q plot for one variable to compare its values with a standard normal population on the vertical axis. The quantiles of the standard normal population are represented in the red line (x=y) in each of the figures. The quantiles of the variable being tested against the normal distribution are represented in the blue crosses.

For a variable to be recognized as normally distributed, its quantiles should approximately lie on the standard normal population (x=y) line.

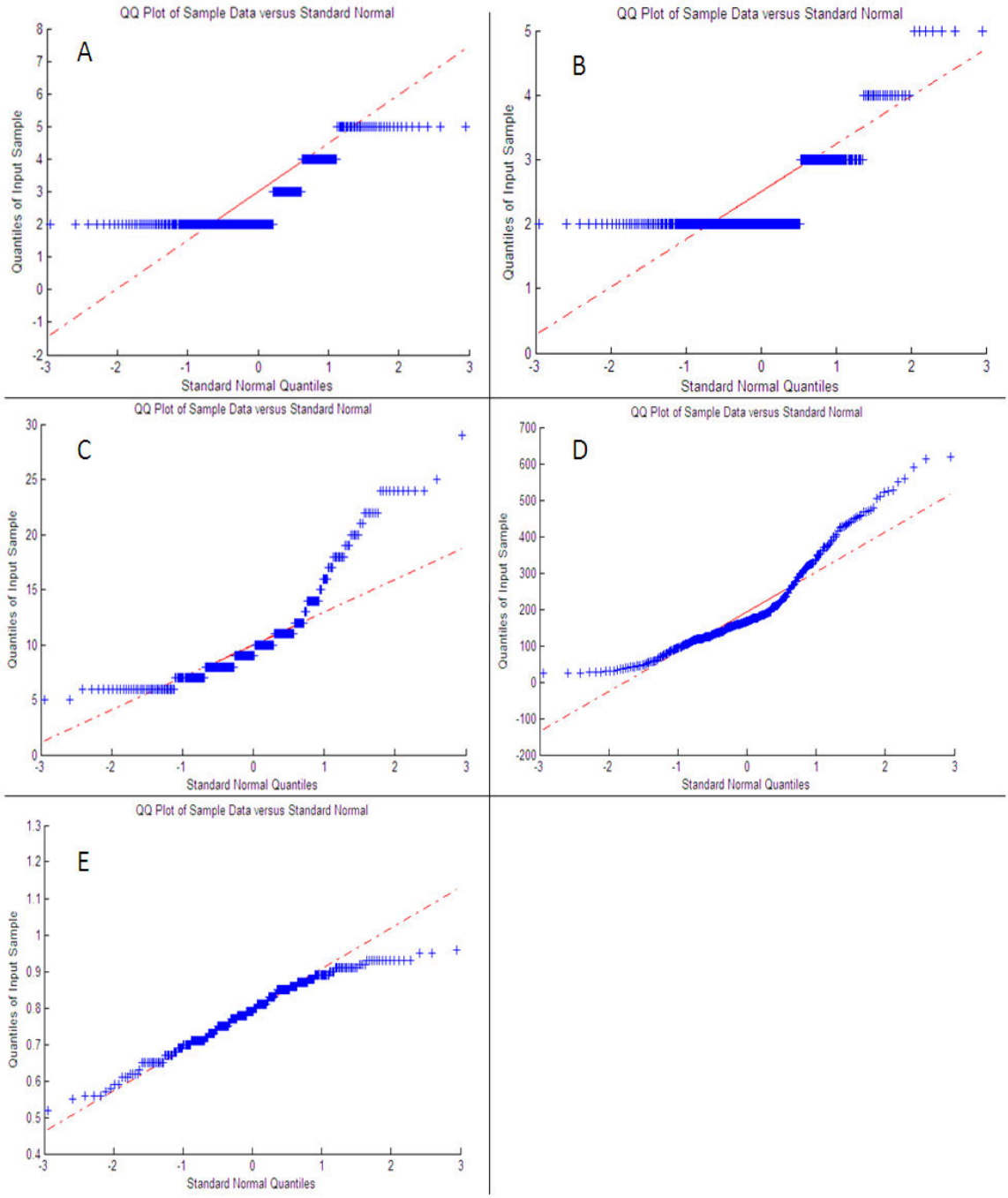


Figure B.1: Q-Q plots for the five individual predictor variables A) Q-Q plot for NT B) Q-Q plot for NP C) Q-Q plot for AC D) Q-Q plot for AFP E) Q-Q plot for LTVR

Appendix C: Groups Variability

This Appendix illustrates the formulas used for the calculation of the F-Statistic.

Within Group Sum of Squares

$$W = \sum_{k=1}^{g} \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2 \tag{C.1}$$

Where g the total number of groups, n_k is is the number of observation in group k, x_{ik} is the i^{th} observation in group k and $\bar{x}_k$ is the mean of group g.

Between-Group Sum of Squares

$$B = \sum_{k=1}^{g} n_k (\bar{x}_k - \bar{x})^2 \tag{C.2}$$

Where $\bar{x}$ is the sample mean.

F-Statistic

$$F = \frac{\frac{B}{df_B}}{\frac{W}{df_W}} \tag{C.3}$$

$$df_B = g - 1 \tag{C.4}$$

$$df_W = N - g \tag{C.5}$$

Where df_B is the between-groups degree of freedom, df_W is the within-group degree of freedom, and N is the total number of observations in the sample.

Appendix D: Screenshot

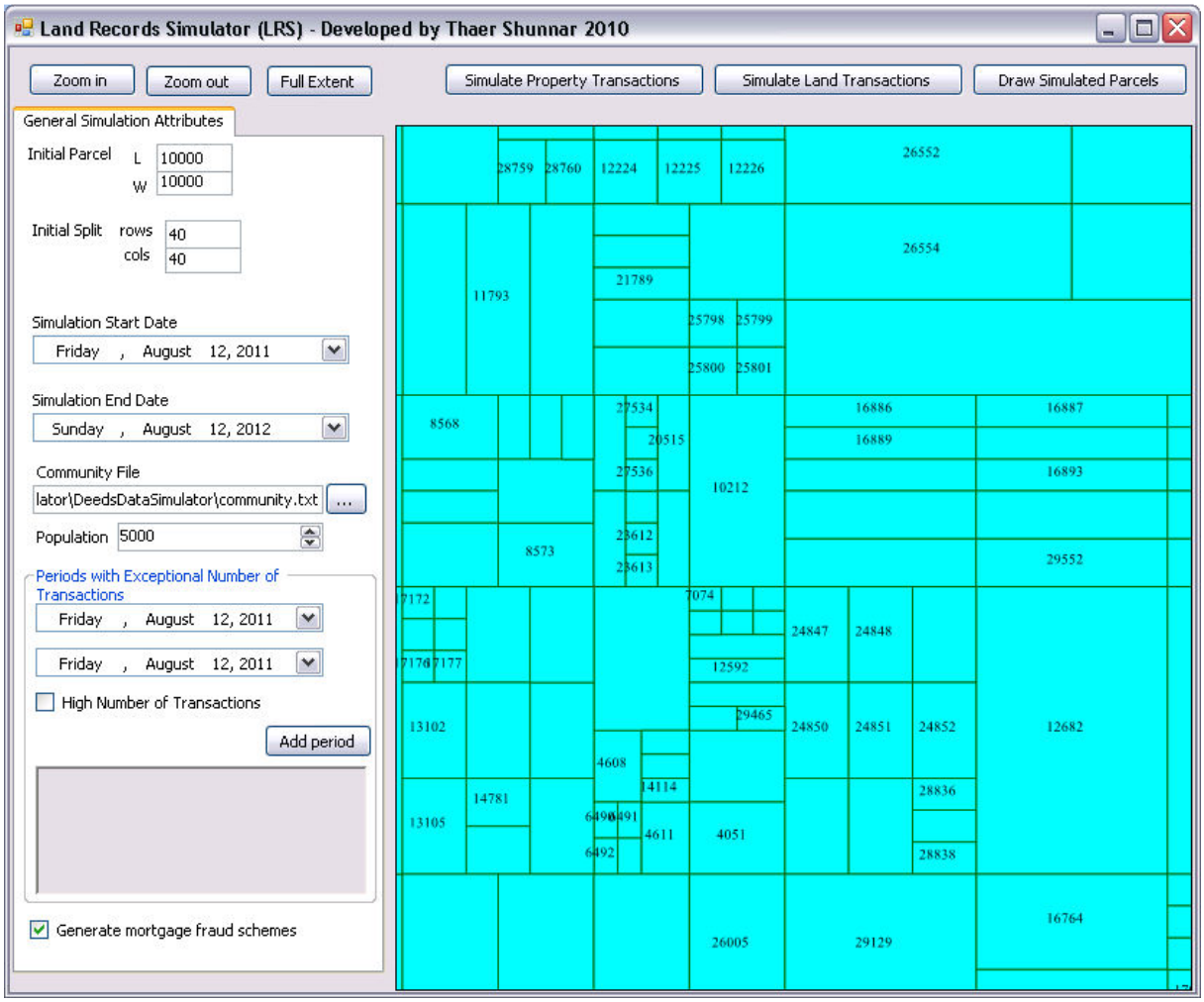
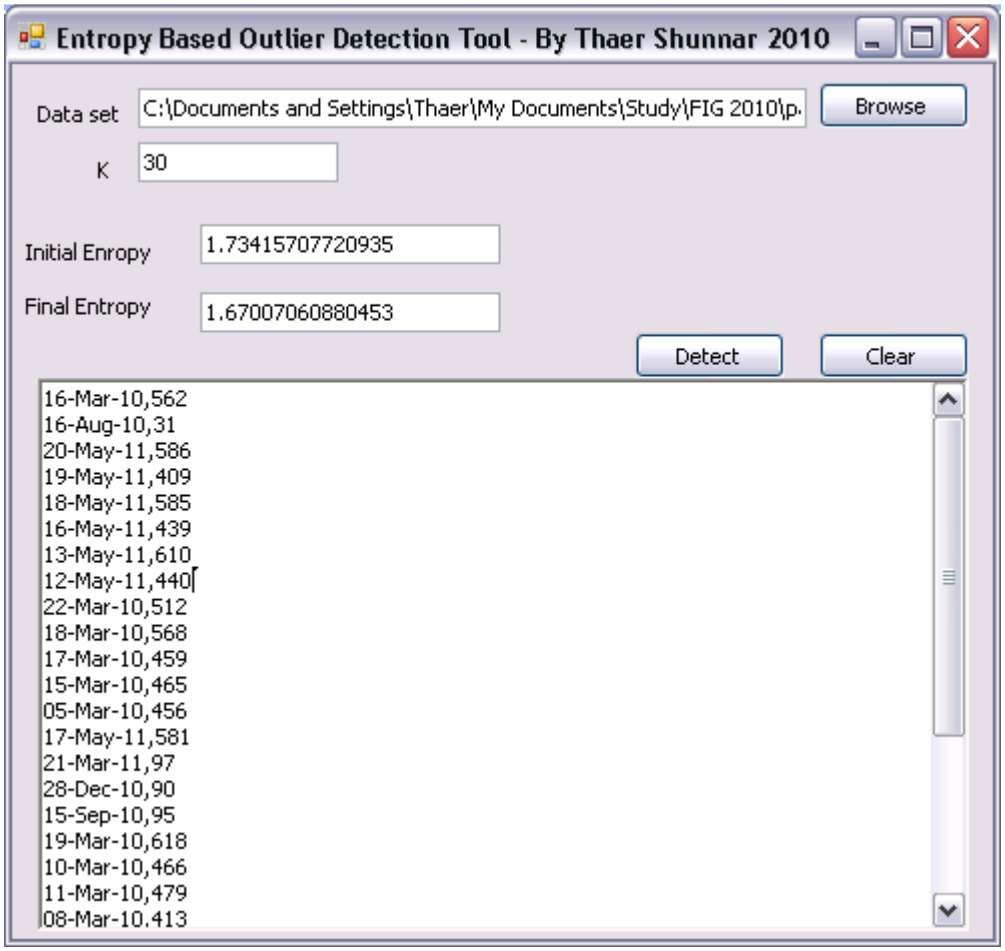


Figure D.1: Screenshot of the Land Records Simulator system.



FigureD.2: Screenshot of the entropy based outlier detection tool.

Appendix E: Predictive Discriminant Analysis (PDA)

The goal of this appendix is provide a brief background of the classification and decision rules of PDA. All the contents of this appendix are summarized from Huberty and Olejnik (2006).

PDA is a form of discriminant analysis used to make empirical predictions. Multiple regression is another approach of making empirical predictions, which can be used when the problem involves a set of p predictors for a single criterion variable Y (outcome variable). The outcome variable in multiple regression is usually a quantitative continuous variable. In PDA however, the outcome variable consists of multiple groups (i.e., it is a categorical variable).

The goal of PDA is to build a rule that will predict the group membership of a unit. The prediction rule may take one of three forms; a composite of predictors it measures, an estimated probability of population membership, or a distance from the estimated population centroid.

The common decision rules of PDA are based on the principle of Maximum Likelihood (ML). This concept states that a unit should be assigned to the population in which the unit's vector has the greatest likelihood of occurrence $P(\mathbf{x}_u|j)$. From the principal of ML, and assuming that we know the prior probability of belonging to a particular population, and also assuming multivariate normality of the population, a posterior probability $\hat{P}(j|\mathbf{x}_u)$ can be calculated from a unit's vector. To use posterior probability for classification the decision rule is:

Assign unit u to population j if

$$\hat{P}(j|\mathbf{x}_u) > \hat{P}(j'|\mathbf{x}_u)$$

For $j \neq j'$, where $\hat{P}(j|\mathbf{x}_u)$ is defined according one of the four rules in Table E.1

Also assignment to a population can be achieved by maximizing the denominator of $P(j|\mathbf{x}_u)$, which can be accomplished by taking the natural logarithm of the denominator. This is achieved in $Q_{uj}$and the decision rule based on $Q_{uj}$is

Assign unit *u* to population *j* if

$$Q_{uj} > Q_{uj'}$$

For $j \neq j'$ where Q_{uj} is defined according one of the four rules in Table E.1

The third rule of assignment is using distance-based classifier where a unit *u* is assigned to the population with the centroid closest to it. This is achieved my minimizing d_{uj}, which is defined in Table E.1.

Table E.1: Alternative forms of classification statistics based on equality of prior probabilities and equality of covariance matrices

Prior Probabilities	Covariance Matrices: Unequal (Quadratic Rule)	Covariance Matrices: Equal (Linear Rule)
Unequal	$\hat{P}(j\|\mathbf{x}_u) = \frac{q_j.\|\mathbf{S}_j\|^{-\frac{1}{2}}.\exp\left(-\frac{1}{2}D_{uj}^2\right)}{\sum_{j'=1}^{J} q_j.\|\mathbf{S}_j\|^{-\frac{1}{2}}.\exp\left(-\frac{1}{2}D_{uj}^2\right)}$ $Q_{uj} = \ln(q_j) - \frac{1}{2}\ln\|\mathbf{S}_j\|\frac{1}{2}D_{uj}^2$ $d_{uj} = \ln\|\mathbf{S}_j\| + D_{uj}^2 - 2\ln(q_j)$	$\hat{P}(j\|\mathbf{x}_u) = \frac{q_j.\exp\left(-\frac{1}{2}D_{uj}^{*2}\right)}{\sum_{j'=1}^{J} q_j.\exp\left(-\frac{1}{2}D_{uj}^{*2}\right)}$ $Q_{uj} = \ln(q_j) - \frac{1}{2}D_{uj}^{*2}$ $d_{uj} = D_{uj}^{*2} - 2\ln(q_j)$
Equal	$\hat{P}(j\|\mathbf{x}_u) = \frac{\|\mathbf{S}_j\|^{-\frac{1}{2}}.\exp\left(-\frac{1}{2}D_{uj}^2\right)}{\sum_{j'=1}^{J}\|\mathbf{S}_j\|^{-\frac{1}{2}}.\exp\left(-\frac{1}{2}D_{uj}^2\right)}$ $Q_{uj} = \frac{1}{2}\ln\|\mathbf{S}_j\| - \frac{1}{2}D_{uj}^2$ $d_{uj} = \ln\|\mathbf{S}_j\| + D_{uj}^2$	$\hat{P}(j\|\mathbf{x}_u) = \frac{\exp\left(-\frac{1}{2}D_{uj}^{*2}\right)}{\sum_{j'=1}^{J}\exp\left(-\frac{1}{2}D_{uj}^{*2}\right)}$ $Q_{uj} = -\frac{1}{2}D_{uj}^{*2}$ $d_{uj} = D_{uj}^{*2}$

u: Measurement unit

J: Total number of groups

$\hat{P}(j|\mathbf{x}_u)$: The posterior of unit *u* belonging to group *j*, given an observed vector $\mathbf{x}_u$

q_j: Estimated prior probability of belonging to group *j*

$\mathbf{S}_j$: Estimated covariance of group *j*

D_{uj}: is the Mahalanobis' generalized distance of a measurement unit $u$from group j centroid and is calculated as

$$D_{uj} = \left(\left[\mathbf{x}_u - \mu_j\right]' \sum_j^{-1} \left[\mathbf{x}_u - \mu_j\right]\right)^{\frac{1}{2}}$$

where $\mathbf{x}_u$ is the observed vector of unit u and μ_j is the centroid of population j.

In the case of the PDA model built for property classification, a quadratic rule was used since covariance matrices were proven not equal. Also, prior probabilities of the three populations (N, S, and H) were not equal and so were considered in the classification rule.