Université de Montréal

Statistical Methods for Insurance Fraud Detection

par

Mathieu Poissant

Département de mathématiques et de statistique

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures

en vue de l'obtention du grade de

Maître ès sciences (M.Sc.)
en statistique

décembre 2008

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Statistical Methods for Insurance Fraud Detection

présenté par

Mathieu Poissant

a été évalué par un jury composé des personnes suivantes :

Louis Doray
(président-rapporteur)

Jean-François Angers
(directeur de recherche)

Alain Desgagné
(Université du Québec à Montréal)
(co-directeur)

Alejandro Murua
(membre du jury)

Mémoire accepté le:
15 décembre 2008

RÉSUMÉ

Une fraude répandue à l'assurance automobile consiste à soumettre de fausses réclamations ou à exagérer les pertes reliées à un sinistre. La performance grandissante des systèmes informatiques et des capacités de stockage crée cependant de nouvelles possibilités pour contrer ce fléau. En effet, les méthodes de forage de données permettent maintenant à une compagnie d'assurance d'analyser une quantité impressionnante d'information afin de déceler les réclamations frauduleuses. Ce mémoire introduit plusieurs méthodes afin de les identifier. Parmi elles, on note l'analyse en composantes principales et l'analyse de classification. Ce mémoire présente aussi les grandes lignes d'une nouvelle méthode statistique de détection des fraudes, appelée PRIDIT. Les résultats obtenus suite à l'application des méthodes à un véritable jeu de données sont aussi présentés.

Mots clés: fraude, assurance automobile, forage de données, classification, composantes principales.

SUMMARY

A common car insurance fraud is to submit false claims or to pad up the severity of an accident. The growing performance of both the information processing systems and the storage capacities however creates new possibilities to deal with this issue. Indeed, the methods of data mining now allow an insurance company to analyze an impressive quantity of information in order to detect fraudulent claims. This thesis introduces several methods to identify potential fraudulent activity. Principal component analysis and cluster analysis are two methods that are discussed. This thesis also introduces an innovative statistical method to detect fraud. It is called the PRIDIT method. The results obtained following the application of those methods to a real data set are also presented.

Keywords: fraud, car insurance, data mining, cluster analysis, principal components.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

REMERCIEMENTS

La rédaction d'un mémoire en milieu industriel procure de multiples avantages. Entre autres, un tel projet permet de nouer des liens avec des gens à l'université, mais aussi au sein d'un environnement de travail. Le nombre de personnes à remercier est, par conséquent, d'autant plus grand.

D'un côté, je souhaite remercier mon directeur de recherche Jean-François pour m'avoir épaulé tout au long de ce parcours... plus long que prévu ! Contrairement à plusieurs qui remercient leur directeur par obligation, je le fais vraiment de bon coeur. D'un autre côté, je tiens à remercier mon co-directeur Alain pour avoir orchestré les premiers milles de ce projet. Sans cette aide, je n'aurais assurément pas vécu cette expérience enrichissante.

De plus, je tiens à remercier Nadine Ouellette pour avoir contribué à mon intégration lors de mes premières semaines chez TD Meloche Monnex. Je souhaite aussi remercier Mathieu Bouvrette d'avoir fait de même lors de mon retour la deuxième année. Je tiens aussi à remercier Catherine Paradis-Therrien pour m'avoir assisté dans le dernier droit de mon parcours chez Meloche Monnex. À un niveau hiérarchique plus élevé, je souhaite remercier Guillaume Gautier, Clément Brunet et Sylvie Mahkzoum de m'avoir témoigné leur confiance.

Je souhaite aussi remercier le Conseil de recherches en sciences naturelles et en génie du Canada ainsi que TD Meloche Monnex pour leur soutien financier.

Finalement, la réalisation d'un projet de cette envergure ne peut se faire sans une bonne dose de passion. Évidemment, il y a la passion de la statistique, mais je parle de l'autre passion, la vraie avec un grand P ! Je remercie donc Marie-Ève pour tout. Simplement !

AVANT-PROPOS

Ce mémoire est le résultat d'une collaboration entre la compagnie d'assurances TD Meloche Monnex, l'Université de Montréal et moi-même. Le projet sous-jacent a été effectué dans le cadre du programme de bourse à incidence industrielle du Conseil de recherches en sciences naturelles et en génie du Canada. Par conséquent, ce mémoire se devait d'avoir une forte composante appliquée afin de respecter l'esprit du programme. Avant sa lecture, deux précisions sont cependant nécessaires. D'une part, le lecteur remarquera que l'ouvrage est rédigé en anglais et non pas en français comme l'exige généralement l'Université de Montréal. Nos interactions avec TD Meloche Monnex exigeaient l'utilisation de la langue anglaise. D'autre part, les données fournies par TD Meloche Monnex avaient au moment de la rédaction un caractère commercial important. Ce mémoire a donc été confidentiel pour un certain nombre d'années avant sa publication officielle.

INTRODUCTION

This Master's thesis is different from most theses since it has a very strong applied nature. Of course, it introduces some theoretical notions from statistics and computer science. We however do not introduce them for their own sake but as a way to handle a precise practical application. The main challenge is therefore to combine business and statistical requirements to produce an academic work. More precisely, we want to identify fraudulent claims within a large data set of car insurance claims. We however do not have any information on previous fraudulent claims.

This thesis has the following structure. The first chapter introduces the reader to main issues about insurance fraud. It also describes the most common car insurance coverages and endorsements. The chapter ends with an overview of the framework of this project, namely, data mining. Statistical notions are first covered in the second chapter. We introduce a useful cross-classification of the types of variables. The second chapter also discusses some useful data transformations that are going to be used in the following analyses. Among them, the RIDIT transformation is of particular interest. We then discuss principal component analysis to reduce the size of our data set. The chapter ends with an introduction to an innovative method to detect fraudulent claims called the PRIDIT algorithm. Unlike the two previous chapters, the third chapter covers a singlewide topic, that is, cluster analysis. This chapter is actually the main one of this thesis. Both hierarchical and nonhierarchical clustering techniques are introduced. Finally, the last chapter presents the results obtained by performing the previous methods. We also give some recommendations to improve the quality of further analyses and results.

Chapter 1

INSURANCE FRAUD IN A DATA MINING FRAMEWORK

This first chapter introduces the reader to the insurance fraud framework of the project. It does not include any statistical contents. Nevertheless, the reader shall read this chapter to understand the following chapters. This chapter begins with a discussion of insurance fraud. Prevalence of insurance fraud, major areas of potential fraud, and anti-fraud activities are among the covered topics. The chapter then introduces the province of Ontario laws and regulations about automobile insurance while this project deals with a data set of claims from policyholders living in this province. The chapter next discusses the selection of statistical methods for identification of potential claims. This initial chapter finally ends with an introduction to the main data mining concepts and techniques by explaining what is data mining and reasons why this is relevant to the current project.

1.1. SPONSORING COMPANY

TD Meloche Monnex Group (TDMMG) is a Canadian company that offers personal home and automobile insurance products in Canada. TDMMG also provides, to a lesser extent, travel and small business coverage. TDGMM exploits two brands, namely, TD Meloche Monnex and TD Insurance. TDMMG insures

professionals, alumni, and affinity group members while TD Insurance insures individuals as traditional insurers do. The reader should note that a confidentiality issue prevents disclosure in this Master's thesis of some information.

1.2. INSURANCE FRAUD

Fraud is an important topic in most businesses. At any moment, some people try to exploit the failures of the system. Telecommunication fraud, credit card fraud, money laundering are all common problems in their respective fields. However, insurance fraud is different on one important aspect from other types of fraud. An insurance fraud in not discovered unless an investigation team found that a fraud occurred. A financial institution, for instance, usually knows credit card fraud, quite rapidly. This difference makes difficult the development of a statistical model for insurance fraud since claims do not have fraudulent or non-fraudulent labels, unlike credit card transactions. The current project is about insurance fraud, which is quite a vast field of study. Two major areas of potential insurance fraud are home and automobile insurance. Home insurance provides coverage against perils that may occur to the residence. Automobile insurance provides protection against loss to covered vehicles. Some possible perils are collision, theft of vehicle, and corporal damages due to an accident. The chapter later describes automobile insurance in details. This section introduces one common definition of fraud. This section is based on Viaene and Dedene (2004).

The legal system defines a fraud as a combination of three components, that is, it involves material misrepresentations, there is intent to deceive, and the main objective is to gain an unjust advantage. Those three components are now discussed.

(1) A fraudulent behavior involves material misrepresentations. First, fraudsters may use concealment, which is the intentional dissimulation of important facts. He can also use falsification, which is the modification of information for his own profit. Finally, a material misrepresentation may

simply consist of a lie. For example, a policyholder who says during underwriting that his 16-year-old child never drives his car knowing that it is wrong is a kind of material misrepresentation.

(2) A fraudulent behavior involves intent to deceive. This premeditation contributes to make of an abuse of insurance a criminal activity. This Master's thesis does not however suppose fraud to be a criminal behavior. The next section on functional classifications of insurance fraud discusses this distinction.

(3) A fraudulent behavior involves an aim of gaining an unauthorized benefit. For instance, a policyholder who says he lost five hundred compact discs during a robbery knowing that he had lost only one hundred of them is actually an unauthorized benefit. The possible unauthorized benefits are infinite and range from a few dollars to many millions dollars undue profits.

The previous definition of fraud is used in most legal systems in the world. Although it is widely used, the insurance industry usually employs the word fraud to designate an abuse of insurance performed by a policyholder. The insurance literature makes a distinction between soft and hard fraud. This chapter later covers more precisely this distinction.

This section defined the concept of fraud. The next section justifies resources devoted to controlling insurance fraud with a particular emphasis on automobile insurance fraud.

1.3. Prevalence of insurance fraud

Two major organizations concerned with insurance fraud in the United States are the Coalition Against Insurance Fraud (CAIF) and the National Insurance Crime Bureau (NICB). Those two organizations provide their own estimates of the prevalence of insurance fraud. The CAIF estimates at $80 billion each year the amount of insurance fraud in the United States. The NICB provides a quite different estimation of $20 billion each year in the United States. Although estimates are very different, they show that fraud is an important problem in the

insurance industry. The Insurance Bureau of Canada (IBC) is the Canadian counterpart of the previous two organizations. The information on the prevalence of insurance fraud in Canada is however not accessible. The prevalence of insurance fraud varies not only among regions but also varies among lines of business (Viaene and Dedene, 2004).

The extent of fraud is also related to the image of the insurance industry. A survey conducted on a random sample of 1000 U.S. adults concluded that 25% of them consider acceptable to overstate the value of an insurance claim (Coalition Against Insurance Fraud, 2003). Moreover, this same survey concluded that 10% of U.S. adults found acceptable to submit a claim for items not lost or damaged or for treatments not provided. This study also concludes that the extent of fraud depends on the economic climate and the context in which the fraud happens. Thus, during an economic recession, fraud costs are more important. On that issue, this survey finds that 66% of U.S. adults say they are more willing to commit an insurance fraud in an economic downturn than in an otherwise good economic vitality. Literature also shows that when a claim is part of a major catastrophe, fraud is more likely to happen (Viaene and Dedene, 2004).

The main point is that different studies lead to very different estimates of the prevalence of fraud. The main reason is the absence of consensus on a precise definition of fraud. This makes the evaluation of this concept quite difficult, say, impossible. Previous estimates may therefore be largely underestimated or overestimated. Although estimates of the prevalence of fraud are questionable, insurance industry admits the importance of such a phenomenon. In addition, although large losses are incurred by insurers, they are usually reluctant to spend money and time for dealing with this problem. This chapter later discusses this issue.

This section showed that insurance fraud is not a marginal phenomenon. The quantification of insurance fraud is however largely different between studies. The next section introduces the concept of insurance fraud by describing common functional classifications found in the literature.

1.4. FUNCTIONAL CLASSIFICATIONS OF INSURANCE FRAUD

Insurance fraud is a general expression by itself. In fact, there are many types of insurance fraud making inappropriate the creation of a single category including all of them. This section presents common classifications of insurance fraud. Such a discussion is useful to better understand the scope of the project and to get a clear and precise situation of the problem. This section introduces the three classifications defined by Viaene and Dedene (2004).

First, there is a distinction between an internal and an external fraud. Those two terms represent whether insiders or outsiders of the insurance industry commit a fraud. Insiders include insurers, agents, brokers, managers, insurer employees or representatives. Outsiders include applicants, policyholders, and claimants. There is also a possibility that outsiders collaborate with insiders to commit a fraudulent activity. Although TDMMG has an interest in both internal and external fraud, this project focuses on the latter.

The second classification differentiates between an underwriting and a claim fraud. An underwriting fraud is committed at the underwriting stage of the insurance process. A policyholder who gives false information to the insurer about the uses of a covered vehicle to obtain a reduced premium commits an underwriting fraud. More precisely, a policyholder may say he drives ten kilometers to reach his working place while he actually drives fifty kilometers. The second type of fraud refers to claim occurring after a peril happens. For example, a policyholder may exaggerate the severity of a loss to gain undue advantage. The term fraud used alone usually indicates the latter type of fraud.

The third classification differentiates between a hard and a soft fraud. This classification refers to the difference between the law connotation of fraud and a broad notion of fraud. Derrig and Krauss (1994) gives the following definition of an insurance claim hard fraud.

Definition 1.1. *An hard fraud is reserved for criminal acts, provable beyond a reasonable doubt, that violate statutes making the willful act of obtaining money or value from an insurer under false pretenses or material misrepresentations.*

According to Derrig (2002), possible hard insurance fraud in personal automobile insurance are:

- staged accident,
- claimant not involved in accident,
- duplicate claims for same injury,
- bills submitted for treatment not given,
- injury not related to accident,
- fictitious injury,
- misrepresentation of wage loss.

Those fraudulent activities are criminal offenses and are therefore quite rare. Abuses of insurance are, by far, more common. Those abuses are what the insurance industry calls soft fraud, which is the exaggeration of a legitimate claim. A hard fraud is usually a planned activity while a soft fraud is an opportunistic activity. The words planned and opportunistic are often used instead of hard and soft fraud. For the purpose of this Master's thesis, the term fraud will be used to designate an abuse of insurance (soft fraud) and not the legal meaning (hard fraud).

The previous three classifications help to refine the scope of the project. Therefore, it can be reformulated as the study of external insurance fraud occurring at the claiming step of the insurance process such that behaviors of interest range from the exaggeration to organized fraud namely, soft fraudulent behaviors.

1.5. Reasons for committing fraudulent behavior

Fraudulent behaviors are not committed for their own sake but for some reasons which may be quite numerous. However, some reasons are surely more common than others. This section introduces reasons why a policyholder might want to commit a fraudulent behavior.

The analysis of behaviors is the nature of behavioral sciences like psychology. It would therefore be inappropriate to pretend being able to analyze such a broad and difficult issue in a single section. Nevertheless, Viaene and Dedene (2004) provide a simple but useful framework to fraudulent behaviors that will be used in

this project. They consider such behaviors as a product of two elements: motivation and opportunity. While those two concepts may seem obvious, they are very far from being simple to understand and explain. This Master's thesis assumes the intuitive definition of those terms. A precise study of fraud motivations from a behavioral perspective is presented by Duffield and Grabosky (2001).

The motivation of a policyholder to commit a fraudulent behavior may take several forms. According to Viaene and Dedene (2004), an economic motivation is present in most cases of insurance fraud, that is, fraudsters want to make undue profits. However, many other motivations might explain a fraudulent behavior as illustrated by the vast quantity of papers on the subject of motivation to defraud an insurance company. As an example, Dionne *et al.* (1993) justify insurance fraud by causes like changes in morality, increased poverty, modifications in the behavior of the intermediaries (medical doctors or mechanics for instance) and attitudes of insurers. Empirical evidence shows that changes in morality are quite important in explaining fraud.

A fraudulent behavior against the interests of an insurer is committed when there is a difference in the possession of information in favor of the policyholder. This principle is called *information asymmetry* in the literature. A claim fraud would certainly not be committed if the insurer knows everything about the underlying peril. However, if the policyholder has important information on the peril and the insurer does not have it, a fraudulent behavior is more likely to occur. For example, there is an information asymmetry when a claimant exaggerates the severity of an injury because he hides some information to the insurer.

1.6. Problems facing an insurance company

Up to this point, this chapter considered insurance fraud from the policyholder perspective. The main objective of the project is however to find ways TDMMG may deal with insurance fraud. This section, based on Viaene and Dedene (2004), identifies six problems an insurer may encounter when dealing with fraud.

(1) A major problem with insurance fraud is that it is not *self-revealing.* In fact, there is no way to know if a claim is actually fraudulent unless it is

discovered. Other types of fraud like the cloning of a credit card are more than likely to be known by the financial institution since the victim shall eventually be aware of such a situation and will communicate with the institution. An insurance company has to consider a claim non-fraudulent unless a special investigation has proved otherwise. Furthermore, everyone would agree that it is easier to control something known and noticed over something unknown and unnoticed.

(2) An insurer may also have difficulties to prove that a claim is legally fraudulent. Moreover, a legal action requires many financial and human resources, that is, money invested in suing potential fraudsters is usually more important than its potential benefits.

(3) An insurer must also consider the dynamic nature of fraud, that is, fraud evolves with time and economy. Current fraudulent behaviors are quite different from those performed 30 years ago when, for instance, Internet was inexistent. Thus, fraud control is a continuous process. This means that an insurer must be aware of economic evolution to update frequently its process.

(4) An insurer faces the problem of the selection of an efficient method. This is a problem because there are a large number of methods and the choice of a method depends on many factors such as available data and computing resources. Many efforts have to be devoted to the selection of the method.

(5) The expression used by Viaene and Dedene (2004) best describes and explains the fifth problem: *News on fraud is always bad news.* In most if not all businesses, management is interested in concepts like investments, budgeting, and profits, not in a pessimistic concept like fraud. Work compensation of high-level managers is closely related to financial results outperforming a given threshold. No compensation is associated with the reduction of fraud by a given amount, that is, this is not a factor in assessing the performance of managers.

(6) A sixth problem facing insurers addresses the quantification of the fraud. As explained earlier, the task of defining the fraud concept makes it difficult to measure its extent. It also makes difficult the assessment of financial impacts of fraud control measures because the validity of this measurement is more or less interesting.

This section introduced the main problems facing an insurer on the issue of fraud control. Fortunately, the literature contains efforts to fight this threat. The next section summarizes actual developments on this subject.

1.7. FIGHTING FRAUD

This section provides a summary of common activities that could improve the fight against fraud by insurers as highlighted by Viaene and Dedene (2004). Anti-fraud activities are performed at both the community and the insurer-level. At the community-level, one main activity consists in the foundation of fraud bureaus and creation of anti-fraud alliances. At the insurer-level, fraud control consists in two approaches, that is, prevention and detection. This section introduces those activities.

Community-level anti-fraud activities are actions undertaken by the insurance industry to fight fraud. One activity is to found fraud bureaus and create alliances among insurance stakeholders. This includes governments, insurance companies, and some customers associations. For instance, the Coalition Against Insurance Fraud and the Insurance Bureau of Canada are two important associations of insurers aimed at fighting insurance fraud in Canada. This project does however exclude this type of activities. The reader interested in this large topic is referred to the discussion papers of the National Insurance Fraud Forum held in 2000.

Fraud control at the firm-level is possible using two approaches, namely, *fraud prevention* and *fraud detection*. They are complementary approaches because each has its specific goals. This section introduces both approaches but puts emphasis on fraud detection.

First, *fraud prevention* refers to actions made by the insurance industry before the fraud is made. Viaene and Dedene (2004) consider prevention as the best

way to fight fraud. Insurers devote large amount of money to prevent fraud. Advertisements of potential drawbacks of insurance fraud are an example of fraud prevention.

Fraud detection identifies activities undertaken by an insurance company to identify policyholders committing fraud. According to Viaene and Dedene (2004), fraud detection is a three-step process. Identification of potentially fraudulent claims is the first step. Actually, TDMMG relies on a Special Investigation Unit (hereafter named SIU) to identify fraudulent claims. The SIU itself relies on precise fraud indicators to achieve this step. Automobile fraud insurance indicators include aggressive claimant, absence of witnesses at the time of loss, and a large number of days between two events, usually the loss and the claim. The second step consists of an investigation of identified claims. The SIU may perform everything for a small phone call to a complete investigation by a claim adjuster. The third step is about decisions taken by the insurer on fraudulent claims. On one hand, the insurer may act unilaterally according to any of the following activities. At this step of the fraud detection, the insurer can do nothing against the fraudulent policyholder. The insurer may also decide to dismiss compensation to the policyholder. To a lesser extent, insurer may simply reduce allowed compensation. Finally, the insurer may press charges against the policyholder. On the other hand, the insurer may negotiate with the fraudulent policyholder. This way, insurer may want the claimant to drop his claim. The claimant may also decide to reduce his claim.

The current project is about increasing the efficiency of the identification step of the detection fraud process. The actual SIU performs essentially manual work in identifying potential fraud. Despite the good work performed by the SIU, manual work is influenced by things such as the analyst's mood and workload. Furthermore, it takes a lot of time and many resources. The implementation of an automatic process would be preferable. More precisely, an algorithm able to identify a potential fraudulent claim as it is received by the claims department allows a rapid dispatch of those threats to the SIU.

Community and insurer level activities are complementary. An insurer should devote resources at both levels to fight fraudulent behaviors. The literature however considers insurer-level activities as more efficient in fighting fraud (Viaene and Dedene, 2004). Proximity of fraud explains this recommendation. This is the objective of the current project.

By now, this chapter has introduced main aspects on insurance fraud especially related to the automobile fraud insurance. The next section gives a precise description of the automobile insurance system in Ontario.

1.8. Automobile insurance system in Ontario

This project considers personal automobile insurance fraud and policyholders whose principal residence is located in Ontario. In fact, great differences in legislation between provinces require us to study only one province. However, our statistical methods would certainly be applicable to other provinces after some adjustments. Two reasons explain the choice of Ontarian policyholders as the framework for this project. The nature of the automobile regime in Ontario is a first reason, where private insurance companies are responsible for both corporal and material damages. In some other provinces, private insurance companies are responsible for the latter only. For instance, in Québec, a public organization called Société de l'assurance automobile du Québec is responsible for corporal damages. It is therefore possible to perform statistical analyses on a data set with both types of claims (corporal and material). The large volume of data available for Ontario policyholders also explains this choice, that is, a larger volume of data means potentially more useful information.

In Ontario, laws require any driver to purchase four coverages while some others are optional. This section, provided by courtesy of the Insurance Bureau of Canada (2006), introduces those compulsory coverages and main optional coverages.

1.8.1. Mandatory coverages

First, every driver must have a *Liability* coverage of at least $200,000 while an higher amount is largely recommended. Liability coverage protects the policyholder for corporal damages (injury or death) caused by his negligence to someone else. It covers the damage amount, costs for processing the claim and defense costs.

Every driver must purchase an *Accident Benefits* coverage. It covers injuries and death that might happen to the policyholder caused by a vehicle accident. This coverage does not depend on who caused the accident. Minimum coverage requirements depend on the type of injury. The reader is referred to the Insurance Act of Ontario for a description of all requirements.

By law, each driver must have an *Uninsured Motorists* coverage. It covers costs related to corporal damages incurred by the policyholder (injury or death) in case of an accident for which an uninsured motorist or a hit-and-run driver is held responsible. The regulation requires a minimum coverage of $200,000. In addition, each driver must have such coverage for damages caused to the policyholder's automobile. In this case, regulation requires a minimum coverage of $25,000.

Finally, any driver must purchase a *Direct Compensation - Property Damage* coverage. This last mandatory coverage provides benefits in case of damages to the policyholder's vehicle due to someone else's fault. It covers also vehicle contents. There is no precise minimum coverage requirement.

1.8.2. Optional coverages

Whereas the previous four coverages are mandatory, the following coverages are optional, but are usually included in an automobile insurance contract. Due to their optional nature, they are not subject to any minimum or limit requirements. In addition, a policyholder has to choose only one of them. This section introduces those four optional coverages.

First, insurers commonly offer *Collision Coverage*. It covers damages to the policyholder's vehicle in two situations, namely, when the policyholder is at-fault or when his vehicle is damaged by an unidentified object or vehicle.

A second common coverage, called *comprehensive coverage*, covers the policyholder against damages to his vehicle caused by perils that do not fall under the previous coverage. In other words, it covers the policyholder against fire, theft, and vandalism.

A policyholder might want to select the previous two coverages. If so, the *all perils coverage* is available. It is a package deal including both collision and comprehensive coverages. Although this coverage provides benefits to the policyholder against most perils, some are still not included. The insurance contract specifies those exclusions.

A policyholder may want coverage for selected perils. For that reason, insurers usually offer a *specified perils* coverage.

1.8.3. Common endorsements

An insurance contract usually includes some additions (endorsements) to the standard policy. Insurers commonly offer the following two endorsements.

First, a policyholder may add a *loss of use* endorsement to his insurance contract. This endorsement allows the policyholder to rent a vehicle free of charge following a covered peril for the time the policyholder's vehicle is out of use. This endorsement is subject to a limit amount that depends on the insurance contract.

Second, insurers usually offer a *waiver of depreciation* endorsement. In case of damages to a vehicle purchased for less than a predetermined amount of time when the loss occurs, the policyholder is entitled to a compensation equivalent to the value of the vehicle at the time of the purchase. Fraud related to this specific endorsement has already been studied by Dionne and Gagné (2002). In their article, they show that policyholders having this endorsement in their insurance contract have a higher probability of theft near the end of the coverage period. Unfortunately, this information is not available in the current project.

Insurers offer other endorsements to their policyholders. The websites of insurers usually provide plenty of information on their products.

This section introduced the reader to the automobile insurance regime in Ontario. The next section discusses business and statistical constraints to the current project.

1.9. Selection of a statistical method

One challenge of the project is to conciliate both business and statistical perspectives. This section discusses the selection of an appropriate statistical method in the actual business framework.

From a *business perspective*, the selected method must

- be applicable while requiring minimal resources,
- be able to predict future incoming claims on their fraud likelihood.

From a *statistical perspective*, the selected method must

- take into account the lack of available information about the outcome variable (fraud),
- be efficient with a large volume of data.

There are two statistical approaches to identify potentially fraudulent automobile claims. On one hand, there are methods designed in an insurance fraud framework. They are common statistical methods with some modifications for a better use in insurance. They are found in insurance and risk journals. An example is the Kohonen's self-organizing feature map as Brockett *et al.* (1998) applied to insurance. Those methods are however not appropriate in all situations. Kohonen's self-organizing feature map is not appropriate because it requires many resources mostly computational. Most economical and statistical models are usually not appropriate because they require much information about previous fraudulent claims. Brockett *et al.* (2002) propose a method, called *principal component analysis of RIDITs*. This last method seems to respect the four criteria previously exposed. The second chapter covers this method. On the other hand, there are general statistical methods. Among them, the cluster analysis respects the previous four criteria. The third chapter introduces this statistical

analysis. Before discussing any statistical analysis, let us introduce the reader to the main framework of the current project, that is, data mining.

1.10. DATA MINING

The increasing power of today's computers allows businesses to gather a large volume of data. Businesses store most of their operational transactions in a system. TDMMG manages three such systems called AS-400, the staging, and the datawarehouse. Such a large volume of data opens the way to the use of *data mining* techniques. There are many subfields with different objectives inside the large data mining field. Among them, the subfield of data clustering encompasses both clustering and data mining techniques to explore large data sets. This is actually the framework of the project. This section introduces the main concepts of data mining. The third chapter covers the cluster analysis issue.

There are different definitions of data mining in the literature. Some authors consider data mining as a step toward a general process called *Knowledge discovery in databases (KDD)* while others do not make that distinction and consider both concepts as equivalent (Han and Kamber, 2006). This Master's thesis uses the latter interpretation as expressed by the following definition of data mining (Berry and Linoff, 2004)[p.4].

Definition 1.2. *Data mining is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules.*

This definition highlights three aspects of data mining. First, data mining is a process performed on a large amount of data. Moreover, large volume of data also means large amount of noisy data. A data cleaning must therefore precede all data mining tasks. Second, it is an exploration tool. In other words, data mining is not an end by itself but a way to obtain new information for further analyses. A predictive result is usually possible with appropriate subsequent analyses. This is completely different from common statistical tools such as hypothesis testing. Thirdly, data mining implies the automatic detection of patterns using computational resources.

The data mining process consists of seven steps:

(1) data cleaning,

(2) data integration (combination of multiple sources),

(3) data selection,

(4) data transformation,

(5) data modeling,

(6) evaluation of patterns,

(7) knowledge presentation.

The previous process is adapted from Han and Kamber (2006). It is important to note the dynamic nature of this process. For instance, data cleaning may be performed at different times during the data mining process since typing errors may be found later in the process. This project is concerned with all steps of the process.

Hastie *et al.* (2001) categorizes data mining techniques as either *supervised* or *unsupervised* methods. On one hand, supervised methods require the company to possess information about the outcome variable for all claims or at least a sample of it. In the current project, the company would have to know if claims are either fraudulent or non-fraudulent. Due to unavailability of that outcome variable, this project does not cover those types of methods. On the other hand, unsupervised methods do not require the company to have that information about the outcome variable. However, it is difficult to draw strong conclusions from those methods.

This section introduced the reader to the main data mining concepts. It first gave a formal definition of data mining. The typical data mining process was then introduced. Finally, the section made a distinction between supervised and unsupervised methods.

1.11. USES OF DATA MINING

This section introduces six main uses of data mining concepts and techniques: classification, estimation, prediction, association, clustering, and profiling.

- Classification: Classification is one of the most important objectives in data mining and in all sciences in general. Classification requires the outcome variable to be qualitative with categories already determined. For

instance, an insurer may want to classify its policyholders as good or bad drivers. In that last situation, the categories good or bad are determined prior to classification. Classification also requires availability of information about the outcome variable. Therefore, a supervised method is necessary to achieve classification.

- Estimation: Whereas the previous task assumed qualitative variables, the present task is about quantitative variables. The determination of this year's number of claims for a population of policyholders given this information for a sample of policyholders is an example of an estimation task. There are two reasons why one would like to estimate a value. First, it allows observations to be ranked based on its estimated value. For example, classes may be ranked based on the estimation. Second, estimation allows the classification of observations into different groups (range of values) based on its estimated value. For instance, policyholders may be classified as having many or no claim. Note the emphasis on obtaining actual information and not future information. Like the previous task, a supervised method is necessary for estimation purposes.
- Prediction: Prediction is similar to estimation. The emphasis is however on obtaining information relating to a future period. The determination of next year's number of claims for a given policyholder is an example of prediction. As classification and estimation, prediction requires a supervised method due to its need of information about an outcome variable.
- Association: The association task is also called *affinity analysis* or *market basket analysis* by some authors (Larose, 2005). Data mining may allow the generation of rules, which is defined as a relation between two or more variables. Determining the proportion of male policyholders that have a sport car is an example of an association task. In that case, the gender of a policyholder and his car's model are associated. The generation of rules may also be appropriate, in some situations, to predict future observations. This predictive ability mainly depends on the context. It would not make sense to predict the car's model of a policyholder based on its

gender. It would however make sense to predict the driving ability based on the age of the policyholder. Association rules are often formulated as if-then clause. Unlike previous tasks, the creation of rules does not however require information on the outcome variable. Association is thus performed by an unsupervised method.

- Clustering: Clustering is defined as the creation of groups (or clusters) in minimizing the similarity between groups while maximizing the similarity inside those groups. Unlike classification, clustering does not require predetermined classes. The analyst obtains groups that are statistically different but not necessarily according to the outcome variable. It is therefore important that a knowledgeable person observes the results. Clustering is usually performed at the beginning of the data mining process. Results obtained are then used in subsequent statistical analysis to provide new and more precise knowledge of a data set. Clustering is an unsupervised method because it does not require information on the outcome variable.
- Profiling: Profiling is also called description by some authors (Larose, 2005). An insurer may want to know more about data and discover new things that may be appropriate for further analysis. Profiling is thus seen as giving to the company's executives new interesting issues to consider by unveiling otherwise unknown situations. In some ways, profiling is similar to fishing. The analyst is not sure if something will come out of his exploration of the data. Furthermore, if something is observed, it is not sure if it would be interesting or useful to the company. However, it may reveal important facts that would require further investigation. Both supervised and unsupervised methods can perform profiling.

The available data set suggests to use an unsupervised statistical method because the data set contains no information on whether or not a claim is fraudulent. Thus, three tasks may be performed in this project, namely, affinity grouping, clustering, and profiling. However, clustering is chosen for business reasons as the main objective of the project in a data mining framework. We will create groups

(or clusters) of similar claims in order to highlight the characteristics shared by most fraudulent claims. Clustering techniques are covered in Chapter 3.

Data mining is an appropriate framework to deal with the current project about fraud detection. Other interesting points about data mining include current applications of data mining, reasons for the development of that field of study and methodologies associated with it. However, it is beyond the scope of this project. The interested reader is referred to Berry and Linoff (2004) for an overview of data mining applied to business context. Many other references are also available for other fields of study.

This chapter first provided the reader with a knowledge of insurance and fraud concepts. Some covered topics were the prevalence of insurance fraud and functional classifications of reasons for committing fraudulent behaviors. This chapter then discussed the laws and regulations of the automobile insurance system in Ontario. It then discussed the selection of an appropriate statistical method for the detection of fraud. This chapter ended with an introduction to the main data mining concepts and techniques.

Chapter 2

STATISTICAL NOTIONS, PRINCIPAL COMPONENT ANALYSIS, AND THE PRIDIT METHOD

This chapter begins with a discussion on the statistical aspects necessary to subsequent reading. The chapter is broadly divided in three parts. First, it introduces the reader to some statistical notions. Among other things, it proposes a potential classification of variables, discusses data standardization and RIDIT transformation. This chapter then introduces principal component analysis (PCA). The PCA is used here to obtain a smaller data set given a very large number of variables. The third part introduces the reader to the principal component analysis of RIDITs method (PRIDIT). Broadly, it is a method designed by Brockett *et al.* (2002) to classify incoming claims according to their fraud likelihood.

2.1. CLASSIFICATION OF VARIABLES

Statistical analyses generally involve a relatively small number of variables and few different types of variables compared with data mining. From a data mining perspective, a data set however contains many variables of different types. This would not be a problem if one could use the same statistical method for all variables. Unfortunately, this is not the case. Any data mining task requires a thorough understanding of the available variables. Therefore, there is a necessity to introduce the reader to the type of variables classification that will be used for the remainder of this thesis. This is the topic of this section.

There are various schemes to classify variables according to their type. It would be great to have a single classification appropriate for all data sets. Unfortunately, the issue is by far more complex. Generally, in statistic, a single classification is sufficient. For example, one may describe all variables in terms of nominal, ordinal, interval, and ratio variables. Due to the large number of different types of variables in a data mining task, this classification is not sufficient. Anderberg (1973) proposes a very interesting cross-classification of variables. This book actually provides one of the best discussions on this issue. It classifies a variable according to both its scale of measurement and its range.

On one hand, one may classify a variable according to the size of its range set. In other words, a variable is either a continuous, a discrete, or a binary variable. Those categories are assumed mutually exclusive.

Anderberg (1973) gives the following definition of a continuous variable.

Definition 2.1. *A continuous variable is a variable having an uncountably infinite range set.*

We find the expression uncountably infinite redundant because an uncountably range set is, by definition, infinite. In an insurance framework, a common continuous variable is the time interval between two events like an accident and a claim. Note that the observed time interval between two events is no more a continuous but a discrete variable. In fact, the limited number of decimals of a continuous variable makes it a discrete variable. A common mistake is to define a discrete variable as a variable that may take only integer values. This is often but not always the case.

Anderberg (1973) also gives this definition of a discrete variable.

Definition 2.2. *A discrete variable is a variable having a finite, or at most countably infinite range set.*

An example of a discrete variable is the indemnity given to a policyholder. This variable is discrete because an amount of money takes a limited number of decimals (two). However, such a variable is often considered as continuous in order to use a particular statistical method.

Finally, Anderberg (1973) considers a binary variable as a special case of a discrete variable.

Definition 2.3. *A binary variable is a discrete variable which may take on only two values.*

The driver's gender is an example of a binary variable. A binary variable is also called a dichotomous variable. Some authors do not make a distinction between a discrete and a binary variable. We however do make this distinction in this thesis because we later introduce some similarity measures that are specific to binary variables. Note that a binary variable is usually coded as either zero or one.

Some authors consider two sub-types of binary variables, that is, a binary variable may be symmetric or asymmetric (Han and Kamber, 2006). Symmetric means that the two possible values on a variable are given the same importance. For example, the driver's gender is a symmetric binary variable because it does not matter whether the male or the female value is coded as zero or one. Obviously, asymmetric means that two values on a variable are given different importance. The more important outcome is now coded as one while the other is coded as zero. For instance, assume a variable that indicates if a car accident caused death. A car accident with death is less probable but more severe than a car accident without death. Therefore, a death value should be coded as one while a no-death value should be coded as zero. This distinction is subtle and it is not necessary for this project. This distinction is however used by many authors when discussing similarity measures such as the Jaccard coefficient in cluster analysis.

On the other hand, we may classify a variable according to its scale of measurement. A variable is either a nominal, an ordinal, an interval, or a ratio variable. Those categories are assumed mutually exclusive.

Suppose x_A and x_B are values for objects A and B on variable X.

Definition 2.4. *A nominal variable satisfies one of the following relations:*

$$(i) \quad x_A = x_B,$$

$$(ii) \quad x_A \neq x_B.$$

The driver's gender is an example of a nominal variable. The only thing we know is whether an observation is different from another on this variable.

Definition 2.5. *An ordinal variable satisfies one of the following relations:*

$$(i) \quad x_A = x_B,$$

$$(ii) \quad x_A < x_B,$$

$$(iii) \quad x_A > x_B.$$

In an insurance framework, an ordinal scale variable may be the risk category of a policyholder. For instance, an insured may be classified as a low-risk or a high-risk policyholder. We know that an observation is identical or different from another on this variable but also that a high-risk policyholder is riskier to insure than a low-risk policyholder.

Definition 2.6. *An interval variable satisfies one of the three relations of an ordinal variable. However, $x_A - x_B$ is now defined.*

In that case, it is possible to say that an observation is greater or lower than another observation on that variable while the difference between them is also meaningful. The accident year of a claim is an example of an interval variable. An accident that occurred in 2006 is more recent than an accident that occurred in 2005. Furthermore, we can say that there is a one-year difference between both events.

Definition 2.7. *A ratio variable satisfies one of the three relations of an ordinal variable. However, x_A/x_B is now defined.*

A ratio variable may be the annual premium for a policyholder. In that case, we know that an observation is greater or lower than another, while their ratio has a useful meaning. For instance, we may say that the annual premium of policyholder A is twice the annual premium of policyholder B.

The difference between an interval and a ratio scale in this thesis has mainly an educative purpose. In fact, no clustering techniques benefit from additional information given by a ratio scale over an interval scale variable (Anderberg, 1973).

TAB. 2.1. Cross-classification of variables in our data set.

	Continuous	**Discrete**	**Binary**
Ratio	N/A	Available	N/A
Interval	Rare	Available	N/A
Ordinal	Rare	N/A	N/A
Nominal	Aberrant	Available	Available

Anderberg (1973) combines those two classifications to obtain a cross-classification that may be illustrated by a contingency table. Note that it is impossible to have a nominal-continuous variable. Moreover, ordinal-continuous and interval-continuous variables are rare in practice. Table 2.1 indicates the types of variables available in our data set immediately after the data selection stage.

This table shows that the data set contains four types of variables but no continuous variables. Moreover, the data set does not contain any ordinal variables and binary variables are all nominal. The next section discusses some ways to transform ratio-discrete, interval-discrete, nominal-discrete, and nominal-binary to solve some issues.

2.2. DATA TRANSFORMATION

In Chapter 1, we introduced the reader to a typical data mining process. The current project supposes the completion of the first three steps (data cleaning, data integration, and data selection) of this process. The fourth step is about transforming the data in order to perform subsequent statistical analyses. Anderberg (1973) overviews a wide range of methods to transform variables. For instance, he identifies several methods to obtain a nominal variable from an ordinal one, an ordinal from a nominal, and so on. We give here some transformations relevant to our project.

The previous cross-classification shows a hierarchy between each type of variables. For instance, a ratio variable gives more information than a nominal variable. It is usually preferable to have an high-information variable than a low-information variable, but it is not always true. For example, it is sometimes necessary to transform a continuous variable into a discrete variable because a particular statistical method works only with the latter type. This transformation

process is called *discretization.* However, this section only considers transformations used to upgrade the level of information. Explanations on discretization are provided only when required.

2.2.1. Ratio-discrete and interval-discrete variables

Except for ratio variables, interval variables give the highest level of information. One important issue to consider with those variables is the possible difference between the units of the variables. For instance, a data set that contains two variables such that one variable is expressed in US dollars and the other in Yen is difficult to manage. A way to solve this problem is to transform those two variables on a same scale. This type of transformation is called *standardization.* There are two common ways to standardize variables. The first common method is to use the *Z-score standardization.* This transformation is given by the following equation:

$$Z_i = \frac{X_i - \mu}{\sigma},$$

where X_i is the value of observation i on variable X, μ is the mean of X, and σ is the standard deviation of X. This method is analogous to the transformation of an arbitrary normal distribution with mean μ and variance σ^2 into a standard normal distribution with mean 0 and variance 1. We usually expect $|Z_i| \leq 3$ for most observations. Furthermore, values outside this range are usually considered outliers.

A second common method, called *min-max normalization*, is given by the following equation:

$$X_i^* = \frac{X_i - \min\{X\}}{\delta_x},$$

where X_i is the value of observation i on variable X, $\min\{X\}$ is the minimal value on variable X, and δ_x is the range of values on variable X. Unlike the previous transformation, this method maps all values on a [0,1] scale. Note that this method works if all X_i are finite.

2.2.2. Nominal variables

The *Z-score standardization* and the *min-max normalization* are inappropriate transformations for nominal variables. For example, it does not make sense to standardize the policyholder's gender variable. This subsection on nominal variables discusses two approaches to handle nominal variables.

A first common way to deal with a nominal variable with two categories is to transform it on a binary 0-1 scale. For example, we can code the policyholder's gender variable such that a *female* value is coded as zero and a *male* value is coded as one. However, this coding is very arbitrary. One may code without any difficulties a *male* value as 0.5. When prior information on the variable is available, it is possible to find an appropriate value. Most of the time, however, one uses data mining to explore a data set with no prior information or knowledge. This method is not a problem for statistical analyses like regression because they are invariant to any binary transformations. Cluster analysis does not have this property. Therefore, another method is usually preferred.

Another approach is to transform a nominal variable into an ordinal variable. A common way to do this is to use a reference variable. Suppose again the policyholder's gender variable. Clearly, all we can say is that a policyholder is similar or different from another, that is, a male or a female. Now, suppose we know that a male policyholder has usually a higher premium than a female policyholder does. We may now rank the policyholders using this reference variable. Only the perspective is changing.

This section introduced some common transformations to deal with interval and nominal variables. From now on, we consider the *min-max normalization* for the interval variables and the reference variable method for the nominal variables. We use this last method because it allows us to use the RIDIT transformation, which is the subject of the next section.

2.3. RIDIT

RIDIT, a concept introduced by Bross (1958), stands for *Relative to an Identified Distribution unIT*. It is a linear transformation that gives empirical values to

the categories of an ordinal variable. In other words, it may be used to obtain an interval variable from an ordinal variable using its empirical distribution. Bross (1958) gives the following definition of a RIDIT.

Definition 2.8. *Suppose X is a discrete-ordinal variable with k possible values. Let $\hat{\mathbf{p}}_X = (\hat{p}_{X1}, \hat{p}_{X2}, \ldots, \hat{p}_{Xk})$ denote the vector of observed proportions for the k possible values of variable X. The RIDIT score for the category i of variable X is given by the following transformation:*

$$R_{Xi} = \sum_{j<i} \hat{p}_{Xj} + \frac{1}{2}\hat{p}_{Xi}.$$

Brockett *et al.* (2002) use a slightly different definition of a RIDIT.

Definition 2.9. *Suppose X is a discrete-ordinal variable with k possible values. Let $\hat{\mathbf{p}}_X = (\hat{p}_{X1}, \hat{p}_{X2}, \ldots, \hat{p}_{Xk})$ denote the vector of observed proportions for the k possible values of variable X. The modified RIDIT score for the category i of variable X is given by the following transformation:*

$$\begin{aligned}
B_{Xi} &= \sum_{j<i} \hat{p}_{Xj} - \sum_{j>i} \hat{p}_{Xj} \\
&= \sum_{j<i} \hat{p}_{Xj} - \left(\sum_{j} \hat{p}_{X_j} - \hat{p}_{X_i} - \sum_{j<i} \hat{p}_{X_j} \right) \\
&= 2\sum_{j<i} \hat{p}_{Xj} - 1 + \hat{p}_{X_i} \\
&= 2\left(\sum_{j<i} \hat{p}_{Xj} + \frac{1}{2}\hat{p}_{X_i} \right) - 1 \\
&= 2R_{Xi} - 1.
\end{aligned}$$

The modified RIDIT is used throughout this thesis.

Example 2.1

Suppose X is the degree of satisfaction of a policyholder toward its insurance company. Table 2.2 shows the empirical distribution of a random sample of 100 observations.

For instance, the RIDIT value for the Satisfied ($i = 2$) is computed as follows:

TAB. 2.2. Data set for Example 2.1

i	**Category**	n_i	$\hat{p}_{Xi}$
1	Very satisfied	25	0.25
2	Satisfied	59	0.59
3	Unsatisfied	15	0.15
4	Very unsatisfied	1	0.01
	Total	100	1.00

$$
\begin{aligned}
B_{X2} &= \sum_{j<2} \hat{p}_{Xj} - \sum_{j>2} \hat{p}_{Xj}, \quad j = 1,2,3,4. \\
&= \hat{p}_{X1} - \hat{p}_{X3} - \hat{p}_{X4} \\
&= 0.25 - 0.15 - 0.01 \\
&= 0.09
\end{aligned}
$$

Table 2.3 gives the RIDIT values of the four categories after the transformation.

TAB. 2.3. RIDIT values for Example 2.1

i	**Category**	B_{Xi}
1	Very satisfied	-0.75
2	Satisfied	0.09
3	Unsatisfied	0.83
4	Very unsatisfied	0.99

The RIDIT transformation has three properties. First, a RIDIT transforms all variables into a $[-1, 1]$ scale. It allows variables with different number of categories to be included into the same analysis. It also allows the comparaison of variables with different scales and units. Second, a RIDIT is relative to an empirical distribution. The analyst does not have to suppose a theoretical distribution. This is particularly useful when the identification of a theoretical distribution is difficult or impossible. Third, a property of the RIDIT transformation is that the mean of all B_{Xi} equals 0, that is $\sum_{i=1}^{k} B_{Xi}\hat{p}_{Xi} = 0$. This property ensures that

all RIDIT values are included in the $[-1, 1]$ interval. Using the data of Example 2.1, we have

$$\begin{aligned}\sum_{i=1}^{k} B_{Xi}\hat{p}_{Xi} &= B_{X1}\hat{p}_{X1} + B_{X2}\hat{p}_{X2} + B_{X3}\hat{p}_{X3} + B_{X4}\hat{p}_{X4} \\ &= -0.75 \times 0.25 + 0.09 \times 0.59 + 0.83 \times 0.15 + 0.99 \times 0.01 \\ &= 0.\end{aligned}$$

This section on the RIDIT transformation ends our discussion on the types of variables and their related issues. We return to the concept of RIDIT in Section 2.5.

2.4. Principal component analysis

A large number of variables is usually a problem when using data mining because it makes difficult the interpretation of the results. Several methods exist to reduce the number of dimensions in a data set. Among all the reduction techniques, principal component analysis (PCA) is one of the most common in the literature. This is the topic of this section. First, it introduces the concept of principal components. It then introduces two of the main issue when performing PCA, namely, the selection of the input matrix, and the determination of the number of principal components.

A PCA may be used for various purposes but two of them are more common (Jolliffe, 2002). As mentioned earlier, PCA is often used to reduce the number of variables in a data set. This is why we use PCA in this project. Another common use of PCA is to transform a set of correlated variables into a new set of uncorrelated variables. Other applications of the PCA exist but are less common.

The data set used for this project has initially more than sixty variables. Performing an analysis using all those variables could be quite difficult or even impossible. An interesting solution would be to have fewer variables without

losing too much information. Before continuing, we define the concept of total univariate variance.

Definition 2.10. *Let $\mathbf{\Sigma}$ denote a covariance matrix of order p associated with a $n \times p$ data set, denoted by $\mathbf{X}$, where n is the number of observations and p is the number of variables. The total univariate variance of $\mathbf{X}$, denoted by $T[\mathbf{X}]$, is given by the trace of $\mathbf{\Sigma}$, which is the sum of the p diagonal elements of $\mathbf{\Sigma}$.*

In the previous definition, we can replace $\mathbf{\Sigma}$ by $\mathbf{P}$ in case of a correlation matrix. The reader should note that the notation $\mathrm{T}[\mathbf{X}]$ is used instead of $\mathrm{V}[\mathbf{X}]$ in order to make a clear distinction between the variance of $\mathbf{X}$ and the univariate variance of $\mathbf{X}$.

The idea behind PCA is that some variables are less important than others in explaining the total univariate variance of a data set. For instance, a variable with only one possible value does not explain any proportion of this total variance. Therefore, this variable may be deleted without changing the total variance. That is the extreme case. Some variables may contribute to a lower degree to the total univariate variance of a data set and therefore deleting them would result in losing little information and gaining effectiveness. A PCA reduces this dimensionality by determining new variables that are called principal components. We give a definition of the kth principal component.

Definition 2.11. *For each $k = 1, 2, \ldots, n$, let $V[Z_k]$ denote the variance of the kth principal component Z_k. Let $\alpha_k' = (\alpha_{kX_1}, \alpha_{kX_2}, \ldots, \alpha_{kX_p})$ denote the vector of coefficients of the original variable X_i where $i = 1, 2, \ldots, p$ for Z_k. The kth principal component Z_k is a linear combination of the p original variables:*

$$Z_k = \alpha_k' \mathbf{X} = \sum_{i=1}^{p} \alpha_{kX_i} X_i \tag{2.4.1}$$

such that $V[Z_k]$ is maximal, $Cov[Z_k, Z_{k-1}] = 0$ for $k = 2, \ldots, p$ and $\alpha_k' \alpha_k = 1$. The last condition ensures that the coefficients of α are finite.

It can be shown that the solution to equation (2.4.1) is $\alpha_k = \mathbf{e}_k$ where $\mathbf{e}_k$ is the kth eigenvector of $\mathbf{\Sigma}$ such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$ and $\mathbf{e}_k' \mathbf{e}_k = 1$ (or $||\mathbf{e}_k|| = 1$) for all $k = 1, 2, \ldots, p$ (see Jolliffe (2002)). Therefore, the first principal component is $Z_1 = \mathbf{e}_1' \mathbf{X}$, the second principal component is $Z_2 = \mathbf{e}_2' \mathbf{X}$, and so on.

Note that the kth principal component is not necessarily unique because there may be multiple solutions of the variance of Z_k.

The most common measure to assess the importance of the kth principal component is to compute the proportion of the total univariate variance $\mathrm{T}[\mathbf{X}]$ that is explained by $\mathrm{V}[Z_k]$. From the previous section, we know that the trace of a covariance matrix is equal to the sum of all its eigenvalues. Therefore, this measure may be given by:

$$\frac{\mathrm{V}[Z_k]}{\mathrm{T}[\mathbf{X}]} = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda i}.$$

This measure is often used to choose the appropriate number of principal components. This issue is covered later in this section.

It is easy to show that $\mathrm{V}[Z_k] = \lambda_k$ for $k = 1, 2, \ldots, p$ and that $\mathrm{Cov}[Z_k, Z_{k-1}] = 0$ for $k = 2, \ldots, p$. Before doing this, the reader should know that $\mathbf{e}_k'\mathbf{e}_k = 1$ for $k = 1, 2, \ldots, p$ and that $\mathbf{e}_k'\mathbf{e}_{k-1} = 0$ for $k = 2, \ldots, p$. Furthermore, the following proposition is needed.

Proposition 2.1. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *denote two random matrix. Let* $\mathbf{a}$ *and* $\mathbf{b}$ *denote two scalar vectors. Then,*

- $V[\mathbf{aX}] = \mathbf{a}'\,V[\mathbf{X}]\,\mathbf{a}$,
- $Cov[\mathbf{a}'\mathbf{X}, \mathbf{b}'\mathbf{Y}] = \mathbf{a}'\,Cov[\mathbf{X}, \mathbf{Y}]\,\mathbf{b}$,
- $Cov[\mathbf{X}, \mathbf{X}] = V[\mathbf{X}]$.

We now prove that $\mathrm{V}[Z_k] = \lambda_k$ for $k = 1, 2, \ldots, p$.

PROOF.

$$\begin{aligned}
\mathrm{V}[Z_k] &= \mathrm{V}[\mathbf{e}_k'\mathbf{X}] \\
&= \mathbf{e}_k'\mathrm{V}[\mathbf{X}]\,\mathbf{e}_k \\
&= \mathbf{e}_k'\Sigma\mathbf{e}_k \\
&= \mathbf{e}_k'\lambda_k\mathbf{e}_k \\
&= \lambda_k\mathbf{e}_k'\mathbf{e}_k \\
&= \lambda_k.
\end{aligned}$$

□

Then, we prove the identity $\text{Cov}\,[Z_k, Z_{k-1}] = 0$ for $k = 2, \ldots, p$.

PROOF.

$$\begin{aligned}
\text{Cov}\,[Z_k, Z_{k-1}] &= \text{Cov}\,\left[\mathbf{e}_k'\mathbf{X}, \mathbf{e}_{k-1}'\mathbf{X}\right] \\
&= \mathbf{e}_k'\text{Cov}\,[\mathbf{X}, \mathbf{X}]\,\mathbf{e}_{k-1} \\
&= \mathbf{e}_k'\text{V}\,[\mathbf{X}]\,\mathbf{e}_{k-1} \\
&= \mathbf{e}_k'\Sigma\mathbf{e}_{k-1} \\
&= \mathbf{e}_k'\lambda_k\mathbf{e}_{k-1} \\
&= \lambda_k\mathbf{e}_k'\mathbf{e}_{k-1} \\
&= 0
\end{aligned}$$

□

In practice, the covariance matrix $\mathbf{\Sigma}$ is usually unknown. Therefore, the covariance matrix $\mathbf{\Sigma}$ has to be estimated. A common estimator of $\mathbf{\Sigma}$ is given by the sample covariance $\mathbf{S}$. Here is its definition.

Definition 2.12. *Let* $\mathbf{x}' = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ *denote a random sample of* n *observations. Let* $\bar{\mathbf{x}}$ *denote the sample mean of* $\mathbf{x}$. *The sample covariance matrix* $\mathbf{S}$ *is defined by:*

$$\mathbf{S} = \frac{\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'}{n-1}.$$

If the correlation matrix $\mathbf{P}$ is used instead of the covariance matrix, a common estimator of $\mathbf{P}$ is given by the sample correlation matrix $\mathbf{R}$. This estimator is usually computed using the following identity:

$$\mathbf{P} = (\text{diag}\,\mathbf{S})^{-1/2}\mathbf{S}(\text{diag}\,\mathbf{S})^{-1/2}.$$

Here is an example of a principal component analysis using a random sample and an unknown correlation matrix.

Example 2.2

Consider three variables X_1, X_2, and X_3 generated from a normal distribution

with mean 0 and variance 1 where $\rho_{X_1X_2} = -0.9$, $\rho_{X_1X_3} = -0.5$ and $\rho_{X_2X_3} = 0.5$. We assume that the correlation matrix $\mathbf{P}$ is unknown. We obtain the following sample correlation matrix:

$$\mathbf{R} = \begin{pmatrix} 1.0000 & -0.9013 & -0.5141 \\ -0.9013 & 1.0000 & 0.4764 \\ -0.5141 & 0.4764 & 1.0000 \end{pmatrix}.$$

The three eigenvalues of $\mathbf{R}$ are the solutions to the characteristic equation. Most statistical softwares have the capability to solve this matrix equation. A possible solution is $\lambda_1 = 2.2836$, $\lambda_2 = 0.6189$, and $\lambda_3 = 0.0976$. Those eigenvalues give the following eigenvectors, $\begin{pmatrix} 0.6248 & -0.6165 & -0.4791 \end{pmatrix}'$, $\begin{pmatrix} 0.3069 & -0.3703 & 0.8768 \end{pmatrix}'$ and $\begin{pmatrix} 0.7180 & 0.6948 & 0.0422 \end{pmatrix}'$. Therefore, the three principal components are:

$$Z_1 = 0.6248X_1 - 0.6165X_2 - 0.4791X_3, \tag{2.4.2}$$

$$Z_2 = 0.3069X_1 - 0.3703X_2 + 0.8768X_3, \tag{2.4.3}$$

$$Z_3 = 0.7180X_1 + 0.6948X_2 + 0.0422X_3. \tag{2.4.4}$$

Note that some softwares (like R) give the standard deviations of the principal components instead of their eigenvalues. To obtain those values, we square the standard deviations. Also, the proportion of the total univariate variance that is explained by Z_1 is given by:

$$\begin{aligned} \frac{\lambda_1}{\sum_{i=1}^{3} \lambda i} &= \frac{2.2836}{2.2836 + 0.6189 + 0.0976} \\ &= 0.7612 \\ &= 76.12\%. \end{aligned}$$

Similarly, we find that the second and third principal components explain 20.63% and 3.25%, respectively. Equation 2.4.2 shows that the first principal component Z_1 explains most of the variation of both X_2 and X_3 while equation 2.4.3 shows

that the variation of X_3 is largely explained by the second component Z_2. Such conclusions are drawn by comparing the coefficients of the two equations.

An important issue when using PCA to reduce the number of variables is to determine an appropriate number of principal components.

2.4.1. Number of principal components

A PCA performed on a data set of 20 variables gives 20 principal components. The objective of PCA is to reduce the number of variables, that is, it makes no sense to replace 20 variables with 20 principal components. Unfortunately, there are no clear rules to determine an appropriate number of principal components. The literature proposes three methods to do this (Timm, 2002). A first method is to select the principal components with eigenvalues higher than one. Using this rule of thumb, we would select one principal component in Example 2.2. Secondly, one may select the principal components that explain at least 70% of the total variance of the data set. Using this criterion, we would also select one principal component in the example. A third method is to use a scree plot. A common rule of thumb is to select the first principal components until a large decrease in variances (or eigenvalues) occurs. For instance, Figure 2.1 suggests one principal component. In this case, all three methods suggest to choose one principal component, that is, Z_1.

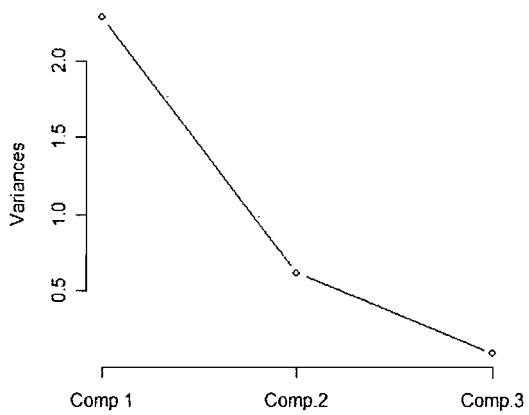

FIG. 2.1. Scree plot for Example 2.2.

All principal components share some properties. A first property is that all principal components are uncorrelated, namely, $\mathrm{Cov}\,[Z_k, Z_{k-1}] = 0$ for $k = 2, \ldots, p$. This property was shown earlier in this chapter. A second property is that all principal components are orthogonal. From linear algebra, we know that all eigenvectors are orthogonal. Since the coefficients of a principal component form an eigenvector, all their coefficients are also orthogonal and their scalar product is zero. Using the principal components Z_1 and Z_2 of Example 2.2, we have

$$\begin{aligned}
\alpha_1' \alpha_2 &= \sum_{i=1}^{3} \alpha_{1i}\alpha_{2i} \\
&= 0.6248 \times 0.3069 - 0.6165 \times (-0.3703) \times -0.4791 \times (0.8768) \\
&= 0.
\end{aligned}$$

In the same manner, it is possible to show that Z_1 and Z_3 are orthogonal and also that Z_2 and Z_3 are orthogonal. A second important issue to consider with PCA is the selection of an input matrix. This issue is now covered.

2.4.2. Input matrix

PCA first requires the selection of an input matrix. This section covers the two most common input matrices, that is, the covariance matrix and the correlation matrix.

A covariance matrix may be used as the input matrix for PCA. This input matrix gives different weights to variables with different variances. More precisely, a covariance matrix gives to a variable with a large variance a greater importance than a variable with a small variance. This result is not surprising because it uses directly the definition of a covariance between two variables. A correlation matrix is also commonly used as the input matrix for PCA. However, the use of a correlation matrix gives equal weights to all variables. We obtain this result directly from the definition of a correlation, that is, all values are transformed

on a same [-1,1] scale. The choice of an input matrix depends on the context of study and *a priori* information.

This section introduced the principal component analysis as a way to reduce the number of variables in a data set. This statistical method may also be used as an iterative process to detect fraudulent claims. This is the subject of the next section.

2.5. Principal component analysis of RIDIT

Brockett *et al.* (2002) propose an innovative method to detect fraudulent claims and identify variables that are the best indicators of fraud. They call it the Principal Component Analysis of RIDITs method or simply the PRIDIT method. The name clearly states the nature of the method. It uses both concepts of principal component analysis and RIDIT. This section overviews the original article of Brockett *et al.* (2002). More precisely, it discusses its goals, its assumptions, and the corresponding algorithm. We also give a complete example of its use.

The PRIDIT method has two main uses. First, the PRIDIT method gives an overall suspicion score for each claim. Those scores may then be used to rank all claims based on their likelihood of fraud. The claim department may investigate only the claims above a given suspicion score. The PRIDIT method may therefore help reduce the costs of investigating the incoming claims. Second, the PRIDIT may be used to assess the discriminatory power of each variable.

The PRIDIT method makes one important assumption, that is, it assumes that all variables are of a ranked-order categorical nature. Using the terminology introduced at the beginning of this chapter, the variables are assumed to be discrete-ordinal. Moreover, all categories of a variable must be ordered in a decreasing likelihood of fraud suspicion. Consider, for example, the driver's gender variable after we transformed it into an ordinal variable using fraud suspicion as a reference variable. The reader is referred to Section 2.2 for a definition of a reference variable. There are evidences that a man is more likely to commit a fraudulent claim than a woman is. Therefore, a man must have a lower value

than a woman must on the variable. An example would be to give a value of zero to a man and a value of one to a woman.

The PRIDIT method is an easy-to-understand procedure composed of two preliminary steps and an algorithm. The first preliminary step is to compute a RIDIT score for each category of all variables. For instance, three binary variables require the computation of six RIDITs. The second step is to replace each value of the data set by its corresponding RIDIT value. This "new" data set, denoted by $\mathbf{F}$, is a $n \times m$ matrix where n is the number of claims and m is the number of variables. Once those two steps are completed, the PRIDIT algorithm begins.

Let f_{it} denote the RIDIT score of claim i on variable t ($i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, m$). In addition, let $\mathbf{W}^{(0)}$ denote a $m \times 1$ vector consisting only of one. A first $n \times 1$ vector of suspicion scores, denoted by $\mathbf{S}^{(0)}$ is given by the product of the matrix $\mathbf{F}$ and the vector $\mathbf{W}^{(0)}$, that is:

$$\mathbf{S}^{(0)} = \mathbf{F}\mathbf{W}^{(0)}.$$

$\mathbf{S}^{(0)}$ is the vector of suspicion scores when all variables have the same weights. However, some variables are better indicators of fraud. A vector of n ones may therefore be inappropriate. According to Brockett *et al.* (2002), a better vector of weights may be given by:

$$\mathbf{W}^{(1)} = \frac{\mathbf{F}'\mathbf{S}^{(0)}}{||\mathbf{F}'\mathbf{S}^{(0)}||}.$$

where $||\mathbf{F}'\mathbf{S}^{(0)}||$ is the Euclidean norm of $\mathbf{F}'\mathbf{S}^{(0)}$. Then, this new vector may be used to obtain a more precise vector of suspicion scores by the following elementary operation:

$$\mathbf{S}^{(1)} = \mathbf{F}\mathbf{W}^{(1)}.$$

By repeating those steps, we obtain $\mathbf{W}^{(2)}$, $\mathbf{S}^{(2)}$, $\mathbf{W}^{(3)}$, and so on. The process may be repeated until convergence is reached. In practice, we repeat the process q times. When q is reached, we obtain the two vectors $\mathbf{S}^{(q)}$ and $\mathbf{W}^{(q)}$. The component $S_i^{(q)}$ of the $n \times 1$ vector $\mathbf{S}^{(q)}$ is the suspicion score for claim i. In other

words, a suspicion score for claim i is given by $S_i = \sum_{t=1}^{m} f_{it} w_t$. The component $W_i^{(q)}$ of the $m \times 1$ vector $\mathbf{W}^{(q)}$ is the weight to give to variable t. As any other consistency measure, a high value indicates a high consistency while a low value indicates a low consistency of variable t with respect to the suspicion score. Here is an example of how the PRIDIT method works with a random data set. Appendix provides our R version of the PRIDIT algorithm.

Example 2.3

Consider the data set of Table 2.4. The categories of the three variables are ranked such that a higher category is related with a lesser degree of fraud. For instance, a value of 0 on variable X_3 means a high likelihood of fraud while a value of 3 means a low likelihood of fraud. Table 2.5 gives the RIDIT values for the categories of those three variables. Table 2.6 shows how we transform the values of claim $i = 1$ into RIDIT values.

TAB. 2.4. Data set for Example 2.3

i	X_1	X_2	X_3
1	1	1	3
2	1	2	3
3	1	2	3
4	0	0	1
5	1	0	2
6	1	1	3
7	1	1	0
8	1	2	3
9	1	1	1
10	1	0	2

TAB. 2.5. RIDIT values for each category of Example 2.3

Category	X_1	X_2	X_3
0	-0.9	-0.7	-0.9
1	0.1	0.0	-0.6
2	-	0.7	-0.2
3	-	-	0.5

TAB. 2.6. RIDIT transformation of claim $i = 1$ for Example 2.3.

	X_1	X_2	X_3
Old	1	1	3
New	0.1	0.0	0.5

We now replace all values of the data set with their corresponding RIDIT values to obtain the RIDIT matrix:

$$\mathbf{F} = \begin{pmatrix} 0.1 & 0.0 & 0.5 \\ 0.1 & 0.7 & 0.5 \\ 0.1 & 0.7 & 0.5 \\ -0.9 & -0.7 & -0.6 \\ 0.1 & -0.7 & -0.2 \\ 0.1 & 0.0 & 0.5 \\ 0.1 & 0.0 & -0.9 \\ 0.1 & 0.7 & 0.5 \\ 0.1 & 0.0 & -0.6 \\ 0.1 & -0.7 & -0.2 \end{pmatrix}.$$

Once $\mathbf{F}$ is computed, we begin the PRIDIT algorithm. First, we compute $\mathbf{S}^{(0)} = \mathbf{F}\mathbf{W}^{(0)}$.

$$\mathbf{S}^{(0)} = \mathbf{F}\mathbf{W}^{(0)} = \begin{pmatrix} 0.6 \\ 1.3 \\ 1.3 \\ -2.2 \\ -0.8 \\ 0.6 \\ -0.8 \\ 1.3 \\ -0.5 \\ -0.8 \end{pmatrix}.$$

The next step is to compute the vector $\mathbf{W}^{(1)}$.

$$\mathbf{W}^{(1)} = \frac{\mathbf{F}'\mathbf{S}^{(0)}}{||\mathbf{F}'\mathbf{S}^{(0)}||} = \begin{pmatrix} 0.2816 \\ 0.6899 \\ 0.6669 \end{pmatrix}.$$

After, we compute $\mathbf{S}^{(1)}, \mathbf{W}^{(2)}, \mathbf{S}^{(2)}$ and so on until convergence is reached. In practice, we usually set a fixed number of iteration of the process. For example, we obtain after 10 iterations:

$$\mathbf{W}^{(10)} = \frac{\mathbf{F}'\mathbf{S}^{(9)}}{||\mathbf{F}'\mathbf{S}^{(9)}||} = \begin{pmatrix} 0.2261 \\ 0.6992 \\ 0.6783 \end{pmatrix}$$

and

$$\mathbf{S}^{(10)} = \mathbf{F}\mathbf{W}^{(10)} = \begin{pmatrix} 0.3617 \\ 0.8512 \\ 0.8512 \\ -1.0999 \\ -0.6025 \\ 0.3617 \\ -0.5878 \\ 0.8512 \\ -0.3843 \\ -0.6025 \end{pmatrix}.$$

After 10 iterations, this algorithm gives us the discriminatory ability of the three variables which is given by $\mathbf{W}^{(10)}$. In this example, variables $t = 2$ and $t = 3$ are good indicators of fraud with values of 0.6992 and 0.6783 while variable $t = 1$ is a poor indicator with a value of 0.2261. This algorithm gives also the suspicion scores for the 10 claims. For instance, the suspicion score for claim $i = 1$ is 0.3617. In other words, this claim is more suspicious than a claim with a 0.8512 value but less suspicious than a claim with a value of -1.0999.

As mentioned earlier, the PRIDIT algorithm may be used to classify all claims based on their suspicious scores. Table 2.7 shows the suspicion scores for the 10

claims in a decreasing order of fraud likelihood. Brockett *et al.* (2002) propose to select the negative values as the fraudulent claims and the positive values as the non-fraudulent claims. Therefore, five claims may be considered as fraudulent and the same number as non-fraudulent. The insurance company should therefore investigate the five claims with negative values ($i = 4, 5, 7, 9, 10$).

TAB. 2.7: Claims of Example 2.3 in a decreasing order of fraud likelihood.

Claim	**Suspicion score**
4	-1.0999
5	-0.6025
10	-0.6025
7	-0.5878
9	-0.3843
1	0.3617
6	0.3617
2	0.8512
3	0.8512
8	0.8512

Brockett *et al.* (2002) show that the PRIDIT algorithm converges, that is, the two sequences $\{\mathbf{W}^{(i)}\}_{i=1}^{\infty}$ and $\{\mathbf{S}^{(i)}\}_{i=1}^{\infty}$ converge. We however do not discuss this convergence property in this thesis. In fact, this property uses the concept of unique variance which is covered in factor analysis. Although factor analysis is similar to PCA, a discussion of this method is beyond the scope of this thesis.

This method has many interesting properties. We now discuss four of them. First, PRIDIT is an unsupervised method as discussed at the end of Chapter 1. Therefore, it may be used when there are no outcome variables. Second, it works with discrete-ordinal variables, which are usually the most common variables in a data set of claims. Third, one may include many variables with varying number of categories within the same analysis because all variables are transformed onto a $[-1, 1]$ scale. Finally, the PRIDIT method is simple and easy to implement in a business context, which is a more than important point to any company. A mathematically complex technique would not be applicable.

This chapter covered three main topics. It began with a discussion on the types of variables usually found in a data set. We exposed the cross-classification proposed by Anderberg (1973), which combines two common classifications into a single classification. The chapter then discussed some ways to transform a given variable into another type. An emphasis was given to the RIDIT transformation. Third, we highlighted main issues on principal component analysis. The chapter ends with an overview of the PRIDIT method to detect fraudulent claims. This method is interesting because it helps the company to allocate its limited resources (financial and human) to the investigation of the most suspicious claims. Chapter 4 shows that this method works with real data.

Chapter 3

CLUSTER ANALYSIS

This chapter covers the broad field of cluster analysis. Unlike the principal component analysis of RIDITs method (PRIDIT) introduced in the second chapter, a cluster analysis is not specifically designed to detect insurance fraud. They are general methods to obtain new knowledge when no previous information on a dependent variable is available.

The idea behind cluster analysis is simple. Given a data set, it creates clusters of observations in a way to both maximize the similarity of observations of a cluster or group and maximize the dissimilarity of observations between different clusters or groups. For instance, an insurer may group all its policyholders in a way that similar policyholders are included in the same cluster and dissimilar policyholders are in distinct clusters. In that case, there are no predefined classes of policyholders such as high-risk and low-risk policyholders. In other words, obtained clusters have no label attached to them. This last task pertains to classification methods and not to clustering techniques. This issue was exposed in the first chapter.

The use of clustering techniques in this project on fraud detection is subject to the following warning. A clustering technique identifies patterns when no information on a dependent variable (fraud status) is available. However, those patterns are not necessarily caused by differences between fraudulent and non-fraudulent claims. Therefore, a claim adjuster (or any other knowledgeable person in insurance fraud) must interpret the results. Among other things, he has to assess the relevance to insurance fraud of the clusters.

This chapter has the following structure. The first section gives an overview of the main concepts on cluster analysis. Then, it discusses the quantification of the proximity between two observations. It also briefly introduces clustering of variables and common classifications of clustering techniques. Next, it introduces the two main types of clustering techniques, that is, agglomerative hierarchical clustering techniques and nonhierarchical techniques through the k-means algorithm. The chapter ends with a discussion on the determination of the number of clusters and the relation between principal component and cluster analyses.

3.1. Definitions and uses of cluster analysis

This section overviews the main concepts underlying cluster analysis. First, it gives a common definition of cluster. It then discusses some possible uses of a cluster analysis. Finally, it makes a distinction between crisp and fuzzy clustering techniques.

Everitt (1980) gives the following definition of a cluster.

Definition 3.1. *A cluster is defined as a group of contiguous elements of a statistical population; for example, a group of people living in a single house, a consecutive run of observations in an ordered series, or a set of adjacent plots in one part of a field.*

A cluster analysis is therefore performed to obtain clusters. More precisely, a cluster analysis is a tool to explore a large data set. Its main objective is the discovery of new knowledge by the generation of new hypotheses for a more precise study or by the identification of patterns in the data. A cluster analysis is also useful to create classification schemes like those commonly found in zoology and other biological sciences. Although the main function of a cluster analysis is not to predict but to discover new patterns, it is however possible to perform predictive analyses. For example, if a claim adjuster finds the clusters relevant to the prediction of the fraudulent status of a claim, he may classify future claims based on those last clusters. Finally, Jobson (1992) uses the cluster analysis as a data reduction technique.

There is a wide variety of clustering techniques. Broadly, the literature on computational pattern recognition identifies two types of cluster analyses. On one hand, a crisp cluster analysis creates mutually exclusive clusters. In other words, a given observation is included into a unique cluster. On the other hand, a fuzzy cluster analysis allows the observations to be in two, three, or more clusters. We performed both types of methods. However, this Master's thesis presents only the results obtained by common crisp clustering techniques.

This section introduced some fundamentals of cluster analysis. A definition of both cluster and clustering concepts and the possible uses of a cluster analysis were the covered concepts.

3.2. PROXIMITY MEASURES

Any cluster analysis needs to quantify the proximity between two objects in a p-dimensional space. The quantification of the proximity between two objects is made through two different types of measures. On one hand, there are dissimilarity measures. This type of measures quantifies the difference (or dissimilarity) between two objects. Therefore, two dissimilar objects are given a high value of dissimilarity while two similar objects are given a low value of dissimilarity. On the other hand, there are similarity measures. This type of measures quantifies the similarity instead of the dissimilarity between two objects. When using this type of measures, two similar objects are given a high value of similarity while two dissimilar objects are given a low value of similarity. Those two types of measures are both part of a more general type of measures referred to as proximity measures. According to this classification, this section is divided in two parts. The first one covers dissimilarity measures while the second part covers similarity measures. Note that most clustering techniques use a dissimilarity measure. However, a similarity measure is often needed when some or all variables are binary. This section also gives some approaches to handle a data set with different types of variables.

3.2.1. Dissimilarity measures

Let p denote the number of variables. Let r, s, and q denote three objects in the p-dimensional space. Let d_{rs} denote the distance between r and s. Here is the definition of a *dissimilarity measure.*

Definition 3.2. *A dissimilarity measure satisfies the following three axioms:*

$$d_{rs} \geq 0, \quad \forall r, s,$$

$$d_{rs} = 0 \Leftrightarrow r = s,$$

$$d_{rs} = d_{sr}.$$

Those three axioms are called positivity, reflexivity, and symmetry, respectively. A measure satisfying those last three axioms is also called a *semi-metric* by some authors (Timm, 2002).

It is possible to add a fourth axiom to the previous definition to obtain a *metric.* Here's its definition.

Definition 3.3. *A metric satisfies the following four axioms:*

$$d_{rs} \geq 0, \quad \forall r, s,$$

$$d_{rs} = 0 \Leftrightarrow r = s,$$

$$d_{rs} = d_{sr},$$

$$d_{rs} \leq d_{rq} + d_{qs}, \quad \forall r, s, q.$$

This fourth axiom is called the triangle inequality. Furthermore, it is possible to replace this last axiom by a new one to obtain an *ultrametric.* A formal definition of this term is now given.

Definition 3.4. *An ultrametric satisfies the following four axioms:*

$$d_{rs} \geq 0, \quad \forall r, s,$$

$$d_{rs} = 0 \Leftrightarrow r = s$$

$$d_{rs} = d_{sr},$$

$$d_{rs} \leq \max(d_{rq}, d_{qs}), \quad \forall r, s, q.$$

The following proposition is useful to identify the type of measure.

Proposition 3.1. *A measure that satisfies all four axioms of an ultrametric is also a dissimilarity measure (semi-metric) and a metric.*

The literature proposes many dissimilarity measures for continuous variables. The *Minkowski distance* (or $L_p - norm$) is however used to derive most popular measures. Its definition is now given.

Definition 3.5. *Let y_r and y_s denote two objects of the p-dimensional space. The Minkowski distance is given by:*

$$d_{rs} = \left(\sum_{j=1}^{p} |y_{rj} - y_{sj}|^{\lambda} \right)^{\frac{1}{\lambda}}, \quad \lambda > 0.$$

Note that $p \geq 1$. It is easy to show that the *Minkowski distance* satisfies the four axioms of an ultrametric. Two common special cases of the Minkowski distance are the *Euclidean* and the *Manhattan* distances where $\lambda = 2$ and $\lambda = 1$, respectively. In the literature, the former measure is, by far, the most frequently used. Its intuitive simplicity and relation to classical geometry might be one of the reasons. As it will be seen later, the most popular does not mean being the most efficient in all circumstances. To state things properly, the definition of the *Euclidean distance* is now given. Note that it is obtained by setting $\lambda = 2$ in the previous definition.

Definition 3.6. *Let y_r and y_s denote two objects of the p-dimensional space. The Euclidean distance is given by:*

$$d_{rs} = \sqrt{\sum_{j=1}^{p} (y_{rj} - y_{sj})^2}.$$

Since the Euclidean distance is a special case of the more general Minkowski measure, it is an ultrametric. By using Proposition 3.1, the Euclidean distance is also a dissimilarity measure and a metric. It is often convenient to express the Euclidean distance in a matrix form.

Definition 3.7. *Let* $\mathbf{y}_r$ *and* $\mathbf{y}_s$ *denote two objects in the p-dimensional space. The Euclidean distance is given by:*

$$d_{rs} = \sqrt{(\mathbf{y}_r - \mathbf{y}_s)'(\mathbf{y}_r - \mathbf{y}_s)}.$$

The Euclidean distance is not always appropriate because of an important scale-unit problem. Suppose the following two variables. The first variable is the indemnity given to a claimant and the other is the age of this claimant. The former variable may range between \$0 and \$200,000, while the latter variable may range between 15 and 85. Obviously, more weight is given to the former than the latter variable when computing the distance between the data for two claimants. We say that the Euclidean distance is not invariant in scale unit. Fortunately, there are many solutions to solve this scaling problem. The main solution is to standardize variables, as explained in the second chapter. This scaling problem affects all Minkowski-based distances at some degree. In fact, a high value for λ increases the importance of the problem while a lower value decreases it.

A second special case of the Minkowski distance, when $\lambda = 1$, is called the *Manhattan distance.* Some authors prefer the use of *city-block* or *taxi-cab* distance (Timm, 2002). Here is the formal definition.

Definition 3.8. *Let* y_r *and* y_s *denote two objects of the p-dimensional space. The Manhattan distance is given by:*

$$d_{rs} = \sum_{j=1}^{p} |y_{rj} - y_{sj}|.$$

Since the Manhattan distance is a special case of the more general Minkowski distance, the Manhattan distance is an ultrametric. By Proposition 3.1, we find that the Manhattan distance is also a metric and a semi-metric. The Manhattan distance is more robust to outliers than the Euclidean distance. Intuitively, the exponent $\lambda = 2$ associated with the Euclidean distance gives outliers more weight than the $\lambda = 1$ associated to the Manhattan distance.

The previous measures are all derived from a more general measure, that is, the Minkowski distance. They therefore share some characteristics. As explained before, all Minkowski-based measures share the scale unit problem. Second, the

distance between two observations on a variable does not affect its counterpart on another variable. Therefore, all those measures assume independence of variables.

Other dissimilarity measures are found in the literature. While most of them are metric measures, there are also nonmetric measures. They differ by not satisfying the fourth axiom of a metric. The attention is however restricted in this Master's thesis to metric measures. Those interested in nonmetric dissimilarity measures for interval variables may read Anderberg (1973).

Dissimilarity measures are one type of proximity measures. They are usually designed to deal with continuous variables and most discrete variables. Binary variables and discrete variables with few categories are usually best handled with similarity measures. However, a particular emphasis is given here to binary variables.

3.2.2. Similarity measures

As mentioned earlier, there are two types of proximity measures: dissimilarity and similarity measures. This second part of the section covers similarity measures. They are sometimes called *correlation-type* measures because they are smaller or equal, in absolute value, to one (Jobson, 1992). Here is the definition of a similarity measure.

Definition 3.9. *Let r and s denote two objects in the p-dimensional space. Let s_{rs} denote the similarity measure between r and s. A similarity measure satisfies the following three axioms:*

$$|s_{rs}| \leq 1, \quad \forall r, s,$$

$$s_{rs} = 1 \Leftrightarrow r = s,$$

$$s_{rs} = s_{sr}.$$

The following proposition is the link between dissimilarity and similarity measures.

Proposition 3.2. *The relation between a similarity measure and a dissimilarity measure is given by:*

$$s_{rs} = \frac{1}{1 + d_{rs}}.$$

The most common similarity measure is the Pearson correlation coefficient, which is now defined.

Definition 3.10. *Let y_r and y_s denote two objects of the p-dimensional space. The Pearson correlation coefficient is given by:*

$$q_{rs} = \frac{\sum_{j=1}^{p}(y_{rj} - \bar{y}_{r.})(y_{sj} - \bar{y}_{s.})}{\sqrt{\sum_{j=1}^{p}(y_{rj} - \bar{y}_{r.})^2 \sum_{j=1}^{p}(y_{sj} - \bar{y}_{s.})^2}},$$

where $\bar{y}_{r.} = \frac{\sum_{i=1}^{p} y_{ri}}{p}$ and $\bar{y}_{s.} = \frac{\sum_{i=1}^{p} y_{si}}{p}$.

A similarity measure is most of the time transformed into a dissimilarity measure since most clustering techniques use this last type of measure. Moreover, most statistical softwares, like the SAS®CLUSTER procedure, use dissimilarity measures. Therefore, similarity measures are usually less important than dissimilarity measures. The situation is however different when a data set contains at least one binary variable. This issue is now considered.

3.2.3. Discrete variables

In an insurance context, available variables are usually discrete or even binary variables. This section first introduces common similarity measures used when the data set contains binary variables. It then discusses similarity measures for nominal and ordinal variables. This section assumes that a binary variable is coded as one if a given characteristic is present and zero if it is absent. The reader may refer to Section 2.1 for an explanation of the differences between the types of discrete variables. There is a wide range of similarity measures for binary variables. However, they are all based on Table 3.1.

TAB. 3.1. Contingency table for the similarity coefficients

	1	0	Total
1	a	b	$a+b$
0	c	d	$c+d$
Total	$a+c$	$b+d$	$q=a+b+c+d$

Using this contingency table, a similarity measure may be obtained for each pair of variables. Timm (2002) introduces nine similarity measures to handle

binary variables. They differ by the weights associated to each variable of a pair. In this Master's thesis, we consider only the simple matching measure. Here is the definition.

Definition 3.11. *Let a, b, c, d, and q be frequencies as represented in Table 3.1. The simple matching measure is given by*

$$\frac{a+d}{q}.$$

This similarity measure gives equal weights to pairs with (0 - 0) matches. Using the relation between a similarity and a dissimilarity measure and the definition of a metric, it is easy to show that the dissimilarity measure corresponding to the simple matching coefficient is actually a metric. An example of the use of clustering with binary variables and the simple matching coefficient is now exposed.

Example 3.1

Consider the two claims of Table 3.2 such that X_1, X_2, X_3, and X_4 are four binary variables.

TAB. 3.2. Data set for Example 3.1

Claim	X_1	X_2	X_3	X_4
1	0	1	0	1
2	0	1	1	0

Those data may be arranged using the contingency table shown in Table 3.1. Table 3.3 shows a similar contingency table for this example.

TAB. 3.3. Contingency table for Example 3.1

	1	0	Total
1	1	1	2
0	1	1	2
Total	2	2	$q = 4$

In this simple example, $a = b = c = d = 1$ and $q = 4$. Therefore, the similarity coefficient between both claims is $\frac{1}{2}$.

A data set more than likely contains some nominal variables. Xu and Wunsch II (2005) propose two strategies to handle such variables. A first strategy is to transform all nominal variables as binary (or dummy) variables and to apply the previous procedure. For example, one may transform the type of insurance (car, moto, or home) variable into three new dummy variables. A second strategy is to use the matching criterion:

$$S_{rs} = \frac{1}{p}\sum_{i=1}^{p} S_{rsi},$$

where $S_{rsi} = 1$, if r and s match and $S_{rsi} = 0$ if r and s do not match. Here is an example that illustrates the use of this second strategy.

Example 3.2

Consider the two claims of Table 3.4 where $i = 1$ denote the policyholder's language, $i = 2$ denote the model of the insured vehicle, and $i = 3$ denote the policyholder's gender. This information is typically found in an insurance company data set of claims.

TAB. 3.4. Data set for Example 3.2

Claim	Language	Model	Gender
1	English	Toyota	Male
2	English	Honda	Male

The similarity measure between claims 1 and 2 is

$$\begin{aligned} S_{12} &= \frac{1}{3}(S_{121} + S_{122} + S_{123}), \\ &= \frac{1}{3}(1 + 0 + 1), \\ &= \frac{2}{3}. \end{aligned}$$

Finally, a data set may contain ordinal variables. Xu and Wunsch II (2005) propose to use a dissimilarity measure, like the Euclidean distance, to handle

this type of variables. Also, Sokal and Sneath (1963) propose a particular coding of ordinal variables in binary variables. Finally, the RIDIT transformation introduced in Chapter 2 is an alternative.

3.2.4. Data set with different types of variables

From a data mining perspective, a data set usually contains many different types of variables. Many approaches are found in the literature to deal with this issue. This section discusses three of them. A first approach is to perform a separate cluster analysis for each type of variable available in the data set (Anderberg, 1973). A second approach is to transform all variables into a same type of variables (Anderberg, 1973). Since conversion of types should be restricted to a minimum, variables should be transformed into the dominant variable type. Finally, Gower (1971) proposes to use the following Gower's similarity coefficient:

$$S_{rs} = \frac{\sum_{i=1}^{p} \delta_{rsi} S_{rsi}}{\sum_{i=1}^{p} \delta_{rsi}},$$

where S_{rsi} is the similarity measure between r and s in the ith dimensional space, where $i = 1, 2, \ldots, p$, and $\delta_{rsi} = 0$ if i is missing for either r or s and, $\delta = 1$ otherwise.

This section introduced common proximity measures. A dissimilarity measure is mainly used with continuous variables while a similarity measure is usually used with binary variables. This section finally discussed approaches to deal with a data set that contains many different types of variables.

3.3. CLUSTERING OF VARIABLES

There are two types of objects that may be grouped in clusters, namely, variables and observations. It is the reason why we referred to objects and not to observations in the previous sections. We wanted to emphasize the fact that variables can also be clustered. The main objective of a cluster analysis on variables is to reduce its number. We prefer the use of a principal component analysis. The reader interested in the clustering of variables is referred to Anderberg (1973). This chapter now discusses some clustering classifications.

3.4. CLUSTERING CLASSIFICATIONS

The field of cluster analysis involves algorithms developed in various fields of study including biology, psychology, and sociology. There is an impressive number of them so that making a review of all clustering techniques is fastidious. In fact, comparisons of all existing algorithms may be a research subject by itself. Fortunately, many classifications of cluster analysis techniques exist depending on the technique used to construct the clusters. However, some classifications and categories overlap because the distinction between them may be unclear. Cormack (1971) proposes a five-category classification: *hierarchical*, *optimization*, *density or mode-seeking*, *clumping*, and *other* techniques. According to this author, hierarchical clustering techniques may be further divided between *agglomerative* and *divisive* techniques. Some authors, like Timm (2002), propose a two-category classification that makes a distinction between *agglomerative hierarchical clustering* and *nonhierarchical* methods. The reader must also be aware that different techniques are not mutually exclusive but rather complementary. For instance, Timm (2002) suggests following hierarchical clustering with nonhierarchical clustering to refine or validate the results.

3.5. HIERARCHICAL CLUSTERING TECHNIQUES

This section introduces hierarchical clustering techniques. They are first considered because of their importance in the literature on data mining and multivariate analysis. All hierarchical clustering techniques follow the same idea. Given N objects, one merges or splits them in clusters that are themselves further merged or split, and so on. This series of successive mergers or splits stops when a *stopping criterion* is satisfied. This criterion usually depends on the number of clusters denoted by k. Therefore, depending on the direction (merges or splits) of the process, there are two classes of hierarchical clustering techniques: *agglomerative* and *divisive* hierarchical clustering techniques. Note that Gordon (1987) considers two other classes of hierarchical clustering techniques: constructive and direct optimization. They are however not discussed in this Master's thesis.

An agglomerative hierarchical clustering technique typically starts with N objects and N clusters $C_1, \ldots, C_N$ such that $C_k \neq \emptyset, k = 1, \ldots, N$. The two most similar clusters are merged to obtain $N-1$ non-empty clusters. This step is then repeated until either the initial N objects are grouped in a single cluster or a stopping criterion is met.

A divisive hierarchical clustering technique is similar but uses splits instead of merges. One starts with one cluster containing N objects. This cluster is then divided into two clusters according to one of the $2^{N-1} - 1$ possible splits. One cluster therefore contains t objects while the other contains $N - t$ objects. A similar split is performed again on those two clusters obtaining this time three clusters. One repeats this process until the initial N objects are divided in N clusters $C_1, \ldots, C_N$ such that $C_k \neq \emptyset, k = 1, \ldots, N$ or a stopping criterion is met.

This Master's thesis covers only the former type of methods, that is, agglomerative hierarchical clustering techniques. The reader interested in divisive hierarchical clustering techniques may read Seber (1984).

While not necessary to determine clusters, interpretation of hierarchical clustering results is greatly simplified by using a *dendrogram.* It is also called a *tree diagram.* Figure 3.1 shows a typical dendrogram. Its use is quite simple. The tree illustrates successive splits or merges at each step of the selected process whether agglomerative or divisive. The distance between clusters is indicated on the y-axis. Note that such a tool may not be appropriate when dealing with a large amount of objects. Gordon (1987) provides a good discussion on that topic by comparing many types of dendrograms and giving mathematical foundations of them.

According to Anderberg (1973), agglomerative hierarchical clustering techniques may be themselves further divided. For instance, those methods may be classified as *linkage*, *centroids*, and *error sum of squares* or *variance* methods. All methods may be used for clustering observations. However, only a *linkage* method clusters variables. This section first introduces the three common linkage methods, namely, *single linkage*, *complete linkage*, and *average linkage* methods. Note that the three linkage methods differ only by the way the distance between

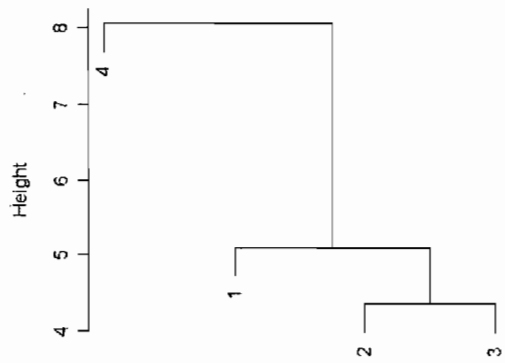

FIG. 3.1. A dendrogram

two clusters is calculated. It then introduces the reader to the *centroid method.* It finally ends with the *Ward's method*, which is the more common *error sum of squares* or *variance* method.

3.5.1. Single linkage

The *single linkage* method has many other appellations. A common but different name is *nearest neighbor* method (Everitt, 1980). Let $d_{(R)(S)}$ denotes the distance between the clusters R and S. The single linkage distance is given by the following equation:

$$d_{(R)(S)} = \min_{r \in R, s \in S} d_{rs},$$

where d_{rs} denotes the distance between the objects r and s. An example will help to understand how it works.

Example 3.3

Consider the following lower triangular distance matrix **D**.

$$\mathbf{D}_{5\times5} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} - & - & - & - & - \\ 3.16 & - & - & - & - \\ 3.61 & 4.12 & - & - & - \\ 7.07 & 8.94 & 5.00 & - & - \\ 2.23 & 1.00 & 4.00 & 8.54 & - \end{pmatrix} \end{array}.$$

First, we merge the two closest items (2 and 5). We denote this cluster (2;5). We then compute the distances between (2;5) and the three remaining items:

$$d_{(2;5)1} = \min(d_{21}, d_{51}) = \min(3.16, 2.23) = 2.23,$$
$$d_{(2;5)3} = \min(d_{23}, d_{53}) = \min(4.12, 4.00) = 4.00,$$
$$d_{(2;5)4} = \min(d_{24}, d_{54}) = \min(8.94, 8.54) = 8.54.$$

The lower triangular distance matrix is now reduced to:

$$\mathbf{D}_{4\times4} = \begin{array}{c} \\ (2;5) \\ 1 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (2;5) & 1 & 3 & 4 \end{array} \\ \begin{pmatrix} - & - & - & - \\ 2.23 & - & - & - \\ 4.00 & 3.61 & - & - \\ 8.54 & 7.07 & 5.00 & - \end{pmatrix} \end{array}.$$

We again select the two most similar items, (2;5) and 1, merge them, and so on. We repeat this process until all objects are in the same cluster. Figure 3.2 illustrates the final structure on a dendrogram. The method first merges the items 2 and 5. It then merges successively the items 1 and 3 to this new cluster. The method finally merges the item 4.

The single linkage method has two main problems (Anderberg, 1973). First, this method is usually unable to separate close clusters. Second, this method tends to create ellipsoidal clusters. Therefore, two observations located at both

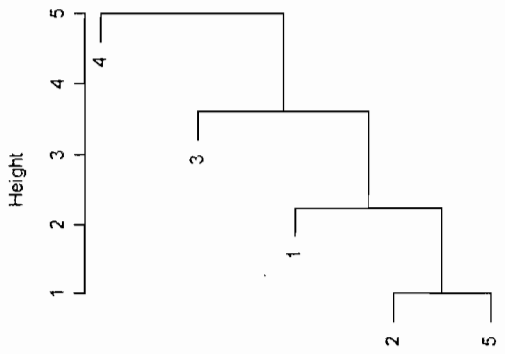

FIG. 3.2. Dendrogram for the single linkage example.

extremities of an ellipsoidal cluster tends to be in the same cluster even if they are distant. This drawback of the single linkage method is called the *chaining* property. This property is sometimes an advantage instead of a disadvantage. For instance, Johnson and Wichern (2007) considers this method as one of the few clustering methods that can delineate non-ellipsoidal clusters.

3.5.2. Complete linkage

The *complete linkage* method, often called *furthest neighbor* method, is very similar to the *single linkage* method. As stated earlier, the two methods differ only by the way they define the distance between two objects. The distance between clusters R and S is now defined by:

$$d_{(R)(S)} = \max_{r \in R, s \in S} d_{rs},$$

where d_{rs} denotes the distance between the objects r and s. Here is an example.

Example 3.4

Consider the lower triangular distance matrix $\mathbf{D}$ of the previous example. By merging the items 2 and 5, we obtain the same (2;5) cluster. The distance between this cluster and the three remaining items are:

$$d_{(2;5)1} = \max(d_{21}, d_{51}) = \max(3.16, 2.23) = 3.16,$$
$$d_{(2;5)3} = \max(d_{23}, d_{53}) = \max(4.12, 4.00) = 4.12,$$
$$d_{(2;5)4} = \max(d_{24}, d_{54}) = \max(8.94, 8.54) = 8.94.$$

Therefore, the following 4×4 lower triangular distance matrix is obtained.

$$\mathbf{D}_{4\times4} = \begin{array}{c} \\ (2;5) \\ 1 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (2;5) & 1 & 3 & 4 \end{array} \\ \begin{pmatrix} - & - & - & - \\ 3.16 & - & - & - \\ 4.12 & 3.61 & - & - \\ 8.94 & 7.07 & 5.00 & - \end{pmatrix} \end{array}.$$

We repeat this process until all items are in the same cluster. Figure 3.3 illustrates the clustering process on a dendrogram. The merging process is exactly the same as the previous example but the distances between successive clusters are different.

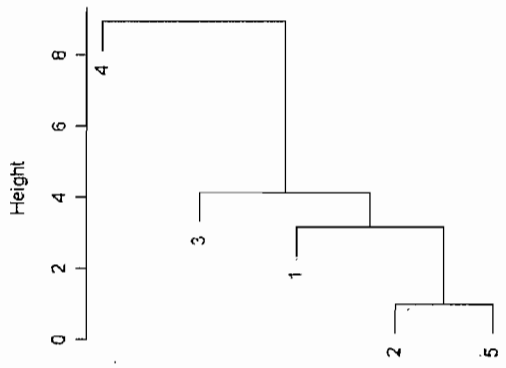

FIG. 3.3. Dendrogram for the complete linkage example.

This result is however not surprising because both single and complete linkage methods are invariant to monotonic transformations of the proximity measures

(Anderberg, 1973). This property describes the fact that the structure of a dendrogram does not change when one uses either the former or the latter method. Johnson (1967) discusses this property.

3.5.3. Average linkage

Anderberg (1973) identifies two average linkage methods. The first method computes the distance of the items within the new cluster. The second method computes the distance of items between merged clusters. This section introduces the latter method only. Everitt (1980) calls it the *group average* method. The reader interested by the first method can see Anderberg (1973). The distance between R and S is now given by the following expression:

$$d_{(R)(S)} = \frac{\sum_r \sum_s d_{rs}}{n_R n_S},$$

where d_{rs} denotes the distance between the objects r and s. An example can make things clearer.

Example 3.5

Reconsider the lower triangular distance matrix **D** of the previous examples. We first merge the two most similar items, namely, 2 and 5. We again denote this new cluster (2;5). The distances between this cluster and the items 1, 3, and 5 are:

$$d_{(2;5)1} = \frac{\sum_{i\in\{2,5\}} \sum_{j\in\{1\}} d_{ij}}{2 \times 1} = \frac{5.39}{2} = 2.695,$$
$$d_{(2;5)3} = \frac{\sum_{i\in\{2,5\}} \sum_{j\in\{3\}} d_{ij}}{2 \times 1} = \frac{8.12}{2} = 4.06,$$
$$d_{(2;5)4} = \frac{\sum_{i\in\{2,5\}} \sum_{j\in\{4\}} d_{ij}}{2 \times 1} = \frac{17.48}{2} = 8.74,$$

Those new distances allow the computation of a 4×4 lower triangular distance matrix.

$$\mathbf{D}_{4\times4} = \begin{array}{c} \\ (2;5) \\ 1 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (2;5) & 1 & 3 & 4 \end{array} \\ \begin{pmatrix} - & - & - & - \\ 2.695 & - & - & - \\ 4.06 & 3.61 & - & - \\ 8.74 & 7.07 & 5.00 & - \end{pmatrix} \end{array}.$$

Again, we repeat this process until all items are in the same cluster. Figure 3.4 illustrates the clustering process on a dendrogram. The merging process is exactly the same as the previous examples.

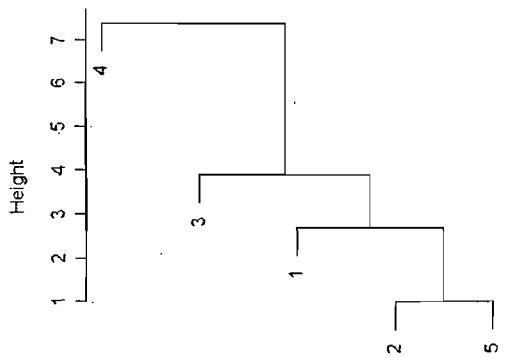

FIG. 3.4. Dendrogram for the average linkage example.

Jobson (1992) proposes to use the average linkage method when the data set contains extremes and outliers. He describes this method as less sensitive (more robust) to those observations than the single and complete linkage methods.

3.5.4. Centroid method

The centroid method is different from the previous three methods. Unlike the linkage methods, the centroid method computes the distance between two clusters using their centroids (or means). This process is best explained by an example.

Example 3.6

Consider the data set of Table 3.5. The three variables X_1, X_2, X_3 are randomly generated from three independent normal distributions with mean 0 and variance 1.

TAB. 3.5. Data set for Example 3.6

Observation	X_1	X_2	X_3
1	0.19	1.11	0.22
2	-0.43	-0.28	-1.05
3	0.91	1.02	-0.29
4	1.79	0.05	0.48
5	1.00	1.58	-1.22

The first step is to compute the distance matrix from the data set. It gives the following lower triangular matrix:

$$\mathbf{D}_{5\times5} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} - & - & - & - & - \\ 1.98 & - & - & - & - \\ 0.89 & 2.02 & - & - & - \\ 1.94 & 2.72 & 1.52 & - & - \\ 1.72 & 2.35 & 1.09 & 2.42 & - \end{pmatrix} \end{array}.$$

Again, we select the two most similar items. Therefore, we merge the items 2 and 3 to obtain the cluster (2;3). Unlike the linkage methods, we now consider a reduced data set as shown on Table 3.6.

TAB. 3.6. Reduced data set for Example 3.6

Observation	X_1	X_2	X_3
(2;3)	0.24	0.37	-0.67
1	0.19	1.11	0.22
4	1.79	0.05	0.48
5	1.00	1.58	-1.22

Using this reduced data set, we obtain the following lower triangular distance matrix:

$$\mathbf{D}_{4\times4} = \begin{array}{c} \\ (2;3) \\ 1 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (2;3) & 1 & 4 & 5 \end{array} \\ \begin{pmatrix} - & - & - & - \\ 1.16 & - & - & - \\ 1.96 & 1.94 & - & - \\ 1.53 & 1.72 & 2.42 & - \end{pmatrix} \end{array}.$$

We repeat this process until all items are in the same cluster. This process gives the dendrogram of Figure 3.5.

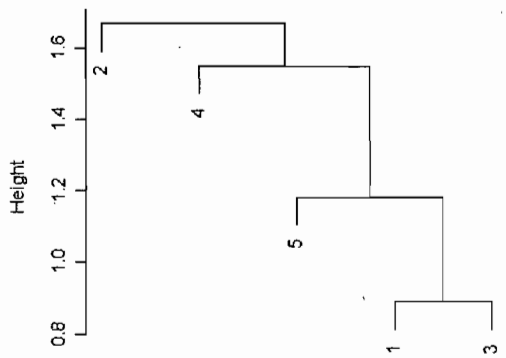

FIG. 3.5. Dendrogram for the centroid method example.

3.5.5. Ward's method

Ward's method is neither a linkage method nor a centroid method. This method differs from previous methods because it does not use a dissimilarity matrix. Moreover, this well-known method uses the concept of sum of squares, not used previously. The first step of the Ward's algorithm is to compute the within sum of squares for all possible merges. Then, the two items that result in the smallest increase in the within sum of squares are merged. Then, the centroid of this new cluster is computed. This process is repeated until all observations

are in a single cluster. Note that an agglomerative method begins with N clusters of only one observation. Therefore, the initial within sum of squares is 0 for all clusters. Let us give an example.

Example 3.7

Consider the initial data set introduced in the previous example on the centroid method. First, we compute the within sum of squares associated with all possible merges. Since the data set contains five observations, we must compute 10 sums of squares. Table 3.7 gives those sums of squares. Note that the incremental within sum of squares equals the within sum of squares at the first iteration.

TAB. 3.7. Incremental within sums of squares for Example 3.7

Merge	Incremental sum of squares
(1 ; 2)	1.9647
(1 ; 3)	0.3933
(1 ; 4)	1.8756
(1 ; 5)	1.4753
(2 ; 3)	2.0316
(2 ; 4)	3.6891
(2 ; 5)	2.7667
(3 ; 4)	1.1541
(3 ; 5)	0.5933
(4 ; 5)	2.9275

Among all sums of squares, merging the items 1 and 3 gives the smallest increase in the within sum of squares (0.3933). Again, we compute all possible within sums of squares and merge the two items with the smallest increase. This procedure is repeated until all items are in a single cluster. The dendrogram of Figure 3.6 illustrates the complete clustering structure.

3.5.6. Computational issues

The previous examples use only a small volume of data. The computation of a dissimilarity matrix requires large resources when the data set contains thousands of observations and variables. To solve this problem, softwares usually compute the distances and sums of squares using the Lance-Williams recursive

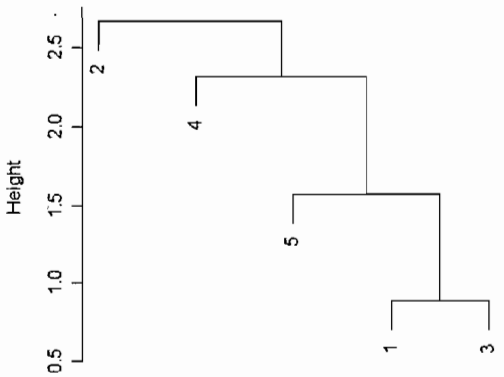

FIG. 3.6. Dendrogram for the Ward's method example.

formula (Lance and Williams, 1967). Note that Jambu and Lebeaux (1978) give a more general formulation a decade later. The Lance-Williams formula is however sufficient to generate the five hierarchical methods introduced in this chapter.

Proposition 3.3. *Let $d(C_i, C_j)$ denote the distance between clusters C_i and C_j. Let α_i, β, and γ denote the parameters specific to a method. The Lance-Williams formula is given by:*

$$d(C_i \cup C_j, C_k) = \alpha_i d(C_i, C_k) + \alpha_j d(C_j, C_k) + \beta d(C_i, C_j) + \gamma |d(C_i, C_k) - d(C_j, C_k)|.$$

This formula allows the construction of most agglomerative hierarchical clustering techniques by replacing each parameter by an appropriate value. For a complete list of them, the reader may refer to Gordon (1987). Table 3.8 shows the parameters needed to obtain the five methods introduced in this chapter (Everitt, 1980). In this table, n_i denote the number of objects in the cluster C_i.

There are other common agglomerative hierarchical clustering techniques like the Ward's method and those obtained by the general Lance-Williams formula. However, they will not be introduced in this thesis. The interested reader is referred to Everitt (1980) or to any textbooks on multivariate analysis.

Hierarchical clustering techniques (agglomerative and divisive) are designed for the analysis of relatively small data sets because of their high requirements of computational resources. In insurance fraud detection, a cluster analysis has to be performed on a large volume of data. Therefore, hierarchical techniques are

TAB. 3.8. Parameters needed to compute all five methods from the Lance-Williams formula.

Method	α_i	α_j	β	γ
Single Link	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{-1}{2}$
Complete Link	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Average Link	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0
Centroid Method	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$-\alpha_i\alpha_j$	0
Ward's Method	$\frac{n_k+n_i}{n_k+n_i+n_j}$	$\frac{n_k+n_j}{n_k+n_i+n_j}$	$\frac{-n_k}{n_k+n_i+n_j}$	0

not appropriate for the purpose of the current project. Description of hierarchical techniques was nevertheless necessary since it is fundamental in any discussion on cluster analysis.

There are three main advantages from using general hierarchical clustering techniques. First, it is a flexible tool about the level of granularity. In other words, clustering allows analysis on a chosen level of precision represented by the stopping criterion. Second, linkage methods provide an easy way to handle different types of proximity measures. Thirdly, it is applicable to any type of data and/or variables. In the same clustering process, it is possible to consider both continuous and discrete data using appropriate proximity measures. Four shortcomings are however worth noting. Difficult issues arise in choosing the number of clusters, that is, some effort must be made in determining a stopping criterion. In addition, hierarchical clustering techniques do not allow reassignment of elements. Once an object is allocated to a cluster, it is impossible to include it in a different cluster. Furthermore, hierarchical clustering algorithms are not robust in the sense that they do not handle outliers well. Finally, they are not appropriate when there is a large volume of data.

As a solution to some of the previous drawbacks, in particular to deal with large data sets, many algorithms have been designed. Among them, the algorithms BIRCH (Zhang *et al.*, 1996), CURE (Guha *et al.*, 2001), ROCK (Guha *et al.*, 2000), and Chameleon (Karypis *et al.*, 1999) are widely used. Due to length restrictions, those algorithms are not covered here.

3.6. NONHIERARCHICAL CLUSTERING TECHNIQUES

This section covers the second main type of clustering techniques. They are called nonhierarchical clustering techniques They are also called *partitioning* or *optimization clustering techniques.* The idea behind these techniques is the division of a data set into several subsets.

All nonhierarchical clustering techniques share at least four properties. First, nonhierarchical clustering techniques are appropriate for clustering observations and not variables (Timm, 2002). Second, they may work with a fixed number k of clusters or may allow k to be determined during the clustering process (Anderberg, 1973). Third, nonhierarchical methods do not require the computation of a dissimilarity matrix making it possible to deal with a larger volume of data than hierarchical clustering techniques. In fact, the clustering is performed directly on the data matrix $\mathbf{Y}_{n\times p}$. Finally, nonhierarchical clustering techniques allow the reassignment to a new cluster of an observation previously included into a different cluster.

Nonhierarchical clustering techniques typically follow a five-step process (Timm, 2002).

(1) Selection of the initial k seed points or an initial partition of items in k groups.

(2) Assignment of each observation to the nearest seed point.

(3) Computation of the centroid of the cluster where the observation is added.

(4) Reassignment of each observation to one of the k clusters.

(5) Return to step 2 if there is no possible reassignment or a convergence criterion is attained.

There is an impressive number of variants of this process. For example, some authors prefer to compute the centroids of the clusters after the assignment of each observation to a cluster.

This section has the following structure. First, it introduces the k-means algorithm. Then, it gives possible alternatives to the original method as proposed by MacQueen (1967) and an example to show how this method may be applied to a data set.

3.6.1. k-Means algorithm

The most common nonhierarchical clustering technique is called k-means algorithm (MacQueen, 1967). However, this is more a generic name for a wide range of algorithms than the name of a unique method. Moreover, different statistical softwares more than likely use different k-means algorithms. One must therefore give great care to this issue before using any statistical software for a cluster analysis. This section introduces a modified instance of the original algorithm. On one hand, MacQueen (1967) proposes to calculate the centroids of a cluster each time an observation is added to it. On the other hand, the algorithm used in this Master's thesis computes the centroids once all assignments are made. Moreover, MacQueen (1967) proposes a single pass through the data set. In other words, the original algorithm excludes the fifth step of the process. Before giving an example, this section discusses one possible way to adapt the original method, that is, through the selection of the initial seeds. Note that the Euclidean distance is used in this section.

MacQueen (1967) proposes to select the first k observations of the data set to be the seeds. There are however many possible alternatives (Anderberg, 1973). The choice of the initial seeds is known to give very different results. A second approach is to randomly select k complete observations to be the initial seeds. A third approach is to select the first observation of the data set to be the initial seed. Then, one selects the second seed based on its distance with the first seed. If the distance between them is greater than a specified radius, then the former observation is now the second seed. The distances between the following observation and the previous two seeds are then computed. If both distances are greater than the radius, then this observation is the third seed. Other seeds are selected the same way. This process is repeated until the specified k seeds are chosen. It is also possible to perform seed replacement based on tests like hypotheses testing.

To make things clear, let us summarize the steps of our modified k-means algorithm. First, we select k data points to be the initial seeds. The last chapter compares the three ways to select the initial seeds with a data set. We then assign

each observation to the closest seed based on the Euclidean distance. The new centroids are computed once all observations are assigned to a seed. Then, those centroids become themselves seeds for the next iteration of the algorithm. The algorithm is repeated until a convergence criterion is satisfied. The more common criterion is to compare the distances between each observation and its centroid to a fixed value of the within sum of squares (Timm, 2002). If the distances are all smaller than the fixed value, then we assume that convergence is reached. Anderberg (1973) gives main steps to prove that this k-means algorithm satisfies a convergence property. Fortunately, SAS®CLUSTER procedure has the ability to compute all previous variants of the k-means algorithm. An example of the application of this algorithm is introduced.

Example 3.8

Consider the data of Example 3.6. We assume $k = 2$ and that the observations 1 and 3 are the initial seeds for clusters R and S, respectively. Then, we compute the distances between all remaining observations and the two centroids. Table 3.9 gives those distances where d_{ij} is the difference between the observation i and the seed j.

TAB. 3.9. Distances for Example 3.8

Observation	d_{iR}	d_{iS}
2	3.93	4.06
4	3.75	2.31
5	2.95	1.19

Next, we look for the smallest distances to assign the observations to either cluster R or S. Therefore, the observation 2 is assigned to cluster R while the observations 4 and 5 are assigned to cluster S. We then compute the centroids for both clusters. Table 3.10 gives those new centroids.

To see if this reassignment occurs, distances have to be computed again. The process continues until there are no possible reassignments.

TAB. 3.10. New centroids for Example 3.8

Cluster	X_1	X_2	X_3
R	-0.12	0.42	-0.42
S	1.23	0.88	-0.34

This section introduced the reader to nonhierarchical clustering techniques. It discussed the common k-means algorithm and potential alternatives. For instance, one may use different methods to select the initial seeds. One may also change when the reallocation is made in the process.

3.7. CHOOSING THE NUMBER k OF CLUSTERS

One of the most important issues in both hierarchical and nonhierarchical clustering techniques is the choice of the number k of clusters. This section introduces some methods to determine it. As expressed by Everitt (1980), it is important to note that no completely satisfactory solution is available. This review is restricted to common criteria generated by the SAS®CLUSTER procedure.

A first criterion is defined by the following expression:

$$R_k^2 = \frac{SST - \sum_k SSW_k}{SST},$$

where k is the number of clusters, SST is the total sum of squares, and SSW_k is the within sum of squares for cluster C_k. This criterion is similar to the coefficient of determination found in regression analysis. For this reason, it is also denoted R_k^2. Broadly, this criterion indicates the proportion of variance explained by the difference between the clusters. A high value of R_k^2 suggests better clusters than a low value. Some authors propose to choose the clusters that explain at least 70% of the total variance (SAS Institute Inc., 2004). Other authors propose to observe if there is a large decrease in R_k^2. The reader should note the similarity between this last rule of thumb and the scree plot method presented in the section on principal component analysis. A similar criterion is to take the expected value of the previous criterion under the null hypothesis of a single uniform cluster (SAS Institute Inc., 2004). Again, some authors propose the 70% rule of thumb and the scree plot method.

Another approach to determine k uses hypotheses testing. The principle is exactly the same as any t-test used to compare the means of two samples. However, it now compares the difference of two centroids to a critical value. Unlike the previous criteria, those methods need some assumptions, that is, the n p-vectors are assumed independent and normally distributed. When a data set does not satisfy the two assumptions, we call them pseudo tests.

A third criterion is given by the following statistic:

$$\text{pseudo } t^2 = \frac{[SSW_T - (SSW_R + SSW_S)](n_R + n_S - 2)}{SSW_R + SSW_S},$$

where SSW_T is the within sum of squares of the new cluster C_T, SSW_R is the within sum of squares of cluster C_R, and n_R is the number of observations in cluster C_R. One then compares this value to a Fisher distribution with p and $p(n_R + n_S - 2)$ degrees of freedom (SAS Institute Inc., 2004). However, it is widely accepted that performing many t-tests to compare more than two means is inappropriate unless the α level is adjusted. One usually prefers to use the principles of the analysis of variable and the associated F-test. Therefore, a fourth criterion is given by the expression:

$$\text{pseudo } F_k = \frac{(SST - \sum_{l=1}^{n} SSW_l)/(k-1)}{(\sum_{l=1}^{n} SSW_l)/(n-k)},$$

where k is the number of clusters, n is the total number of observations, SST is the total sum of squares, and SSW_l is the within sum of squares for cluster C_k. This observed value is then compared to a Fisher distribution with $p(k-1)$ and $p(n-k)$ degrees of freedom (SAS Institute Inc., 2004).

We introduced in this section four criteria to choose a number of clusters. More precisely, we discussed the R_k^2 criterion along with its expectation value. We also discussed the common t-test and F-test statistics as alternatives to them. We finally explained why they are called pseudo criteria.

3.8. Principal Component Analysis and Cluster Analysis

The current project involves a large number of variables. It would surely be a good idea to reduce this number of variables before applying any clustering

analysis. The principal component analysis is selected for this purpose. There are however two issues to care about when using a principal component analysis before a cluster analysis, that is, scaling and weighting.

First, Jobson (1992) indicates that a correlation matrix is usually more appropriate than a covariance matrix to a principal component analysis because of the scaling problem. It was also argued in a previous section that standardization needs to be performed prior to any cluster analysis when there are various scales of measurements. We gave an example between U.S. dollars and Yen. It would be redundant to do the same standardization twice. Jobson (1992) proposes to use the correlation matrix along with principal component analysis before any cluster analysis in most situations. However, if the proximity measure is of a correlation type, then the covariance matrix should be used instead, which also prevents redundancy.

Furthermore, some variables may be highly correlated. This is a problem for a cluster analysis because a dimension (or a factor) may be given too much importance. The use of a principal component analysis helps to reduce the dimension of a data set while preserving important information. However, as Jobson (1992) points out, using principal component analysis prior to a cluster analysis makes outliers less detectable by a cluster analysis. Therefore, special care has to be given to outliers when using both techniques.

Most clustering techniques require the computation of a proximity matrix which can be either a similarity or a dissimilarity matrix. To obtain such a matrix, one needs to select a measure to quantify the distance between two objects. Among those discussed in this chapter, the Euclidean and the Manhattan distances are commonly found in the literature. Each has its own advantages and disadvantages. There are two main types of clustering techniques. Hierarchical techniques were first introduced with an emphasis on five agglomerative techniques. Then, nonhierarchical techniques were introduced while a particular emphasis was given to the k-means algorithm. Two important issues were then discussed at the end of the chapter, that is, the determination of the number of clusters and the relation between principal component analysis and cluster

analysis. By now, all theoretical notions have been described. The next chapter provides the reader with the analysis of the TDMMG fraud data.

Chapter 4

STATISTICAL RESULTS

This chapter presents the results obtained using the various methods discussed previously. It first introduces the data set and some other basic issues. It then discusses the reduction of the data set through a principal component analysis. The PRIDIT results are also described at that point. The chapter next introduces the clustering techniques results. Both agglomerative hierarchical and k-means techniques are extensively discussed. It ends with the presentation of some recommendations for subsequent analyses.

4.1. DATA

The data set contains 38,043 observations on 63 variables. Each observation represents a claim received by TD Meloche Monnex Group (hereafter named TDMMG) between June 1st, 2006 and May 10th, 2007. The data set therefore spans approximately a one-year period. More than one observation can identify the same claim. For instance, a claim with both corporal and material damages is entered in two observations. One observation represents the corporal damage part of the claim while the other represents the material damage part of the claim. The reader is also aware that one policy usually insures more than one driver. For example, a policy can insure all members of a family or both members of a couple. In those cases, we consider the driver appearing first on the policy contract as the main policyholder. All claims are closed by the time of the extraction. All data are extracted from the three information systems (AS-400, datawarehouse, and

datastagging). Note that TDMMG does not use the usual meaning of the word datawarehouse (Berry and Linoff, 2004). We use SAS®8.02 for most analyses.

An important step in the data mining process is the selection of the variables as introduced in the first chapter. The original data set contains more than sixty variables. Many variables are however irrelevant to data mining. The identification key variables are not useful in any data mining task. Furthermore, some variables in the data set are redundant because they give the same information as other variables. All those variables are therefore excluded from the analyses. Some variables have values entered in a free format text. Due to time restriction, those variables are excluded from the analyses. Some variables are excluded from analyses because they are not valid or are simply irrelevant to the detection of fraud. Other variables are also excluded because they require time-consuming transformations. According to the literature and to a claim adjuster, the model of the vehicle is relevant to the detection of fraud (Brockett *et al.*, 2002). A code uniquely identifies each model in the data set. The large number of arbitrary codes makes however impossible the consideration of this variable in the project. The affinity group of a policyholder is also potentially relevant to the detection of fraud. The situation is similar to the previous variable with more than two thousands different unordered codes.

The selection of variables reduces the data set from more than sixty variables to eighteen variables. Those selected variables may be grouped in four categories. First, the claim variables give information on the claim. The loss variables give information on the loss incurred by the policyholder. The policyholder variables give information on the main policyholder and its interaction with TDMMG. Finally, the vehicle variables give information on the vehicle associated with the claim. Table 4.1 lists the selected variables along with their definition. A sharp sign indicates a binary variable.

A major advantage of this project over other similar academic projects is the availability of a claim adjuster throughout the project. His expertise provides clues to obtain better results. He identifies nine risk factors to car insurance fraud. According to his knowledge, a claim is more likely to be fraudulent when:

TAB. 4.1. Description of the selected variables

Name	Description
Claim variables	
Expenses	Indicates the expenses incurred by TDMMG to manage the claim.
Indemnity	Indicates the indemnity given to the policyholder.
Opening time	Indicates the number of days between the first and the last financial transaction associated with the claim.
Inforce policy	Indicates the number of days between the effective date of the policy and the date of the loss.
Time interval before report	Indicates the number of days between the date of the loss and the date of the claim.
Recovery	Indicates the amount recovered by TDMMG.
Loss variables	
Catastrophe (#)	Indicates if the loss occurred at the time of a catastrophe.
Policyholder's liability	Indicates the policyholder's liability (in %). The possible values are 0, 25, 50, 75, and 100%.
Season of the loss	Indicates the season when the loss occurred. The possible values are spring, summer, winter, and autumn.
Time of the loss (#)	Indicates if the loss occurred during either the night or the day .
Policyholder variables	
Claim free	Indicates the number of consecutive years without any claim from the main policyholder.
Client since	Indicates the number of consecutive months the main policyholder is insured by TDMMG.
Policyholder's age	Indicates the main policyholder's age.
Policyholder's gender (#)	Indicates the main policyholder's gender.
Policyholder's residence	Indicates the socio-economic level of the main policyholder's residence. The possible values range between 100 and 900. A value of 100 indicates a vicinity with a high socio-economic level. A value of 900 indicates a vicinity with a low socio-economic level.
Vehicle variables	
Age of the vehicle	Indicates the age of the vehicle.
Ownership (#)	Indicates if the vehicle is either rented or bought.
Price of the vehicle	Indicates the price of the vehicle.

(1) the associated loss occurs at night,

(2) the associated loss occurs at the time of a catastrophic event,

(3) the car is rented (in opposition to bought),

(4) the policyholder's last claim was filed lately,

(5) the associated loss occurs in the autumn,

(6) there is a large number of days between the date of the loss and the date it is reported to TDMMG,

(7) the claim file is opened for a long time,

(8) the claimant is a recently insured policyholder,

(9) the car is expensive.

This *a priori* information helps to interpret the results of the cluster analyses and to identify potential fraudulent claims. That information is however based on his intuition and not on scientific facts. Therefore, we cannot blindly rely on that expertise. A good approach is to use this useful knowledge along with statistical methods. A challenge of this project would be to combine both sources of information.

4.2. COVERAGES

From Chapter 1, it is clear that the type of coverage has to be used as the grouping variable, that is, we have to divide the data set in smaller data sets based on that variable. In fact, there is great evidence that risk factors are different across coverages. For example, it is inappropriate to include collision claims and third-party liability claims in the same analysis because intuitively the risk factors are different. Table 4.2 presents the distribution of the claims by type of coverage. The reader is referred to the first chapter for a complete description of coverages offered to Ontarian policyholders.

TAB. 4.2. Distribution of the claims by type of coverage

Coverage	**Frequency**	**Percent**
Collision	16,153	42.46
TPL - Property damage	13,593	35.73
Comprehensive	4,897	12.87
Other	3,399	8.94
Total	38,042	100.00

Table 4.2 indicates three major types of claims. Collision claims are the most frequent type with 16,153 claims (42%). Third-party liability for property damages are the second most frequent type of claim. The data set contains 13,593 (38%) claims of this type. With 4,897 claims (13%), comprehensive coverages are the third type of claim in importance. The *Other* category contains the four remaining types of claims, that is, third-party liability for corporal damages, medical and disability coverages, and direct compensation. The reader is referred to Chapter 1 for a description of those coverages. We analyzed all types of claim but only collision claims are presented in this thesis. Collision claims also have the highest average severity. The average indemnity for a collision claim is \$3,938 while the average indemnity for all other claims is \$3,568. The selection of this type of coverage is therefore justified.

Table 4.3 presents main descriptive statistics for the collision claims (mean, standard deviation, first and third quartiles). From now on, we assume that those variables are continuous. Note that this table excludes the four binary variables and the *Season of the loss* variable. A good approach to understand the data is to compare the mean on a variable with its standard deviation. On one hand, some variables have a standard deviation smaller than its mean. For instance, a vehicle is, in average, 5.73 years old with a standard deviation of 3.57 years old. On the other hand, some variables have a standard deviation much larger than its mean. For example, the standard deviation of the time interval before report variable is more than three times larger than its mean. On other variables, the standard deviation may be similar to the mean. The important point here is to note the large differences between the standard deviations of the various variables. This observation will later be important when discussing standardization. Tables 4.4, 4.5, 4.6, 4.7, and 4.8 give the distribution of collisions claims based on the five discrete variables.

4.3. Principal component analysis

A principal component analysis (PCA) sometimes helps to reduce the number of variables in a data set. The second chapter covered the theory of this data

TAB. 4.3. Descriptive statistics for the 16,153 collision claims

Variable	Mean	St. Dev.	Q1	Q3
Age of the vehicle (years)	5.73	3.57	3	8
Claim free (years)	2.33	2.65	0	3
Client since (years)	7.15	6.83	2	10
Expenses	$358	$324	$240	$277
Indemnity	$3,938	$5,019	$882	$5,247
Opening time (days)	68.86	61.35	25	95
Policyholder's age (years)	43.47	11.94	34	51
Policyholder's liability	39.9%	48.2%	0%	100%
Policyholder's residence	565	259	344	786
Price of the vehicle	$22,956	$14,606	$13,000	$30,700
Recovery	$102	$619	$0	$0
Inforce policy (days)	175.4	105.5	83	266
Time interval before report (days)	3.93	10.62	0	3

TAB. 4.4. Distribution of the collision claims on Catastrophe variable

Catastrophe	Frequency	Percent
Not part of a catastrophe	16,136	99.89
Part of a catastrophe	17	0.11
Total	16,153	100.00

TAB. 4.5. Distribution of the collision claims on Season of the loss variable

Season of the loss	Frequency	Percent
Spring	1,507	9.33
Summer	4,973	30.79
Autumn	4,909	30.39
Winter	4,764	29.49
Total	16,153	100.00

TAB. 4.6. Distribution of the collision claims on Time of the loss variable

Time of the loss	Frequency	Percent
Day	15,572	96.40
Night	581	3.60
Total	16,153	100.00

reduction technique. Briefly, a principal component is a linear combination of the original variables such that its variance is maximized. For some business reasons,

TAB. 4.7. Distribution of the collision claims on Gender variable

Policyholder's gender	Frequency	Percent
Female	5,712	35.47
Male	10,390	64.53
Total	16,102	100.00

TAB. 4.8. Distribution of the collision claims on Ownership variable

Ownership	Frequency	Percent
Bought car	12,717	78.73
Rented car	3,436	21.27
Total	16,153	100.00

only nine of the 13 continuous variables and no discrete variables are included in the analysis. Table 4.9 lists the nine variables that are included in the PCA.

TAB. 4.9. List of the variables included in the PCA analysis

Inforce policy
Time interval before report
Policyholder's liability
Claim free
Client since
Policyholder's age
Policyholder's residence
Age of the vehicle
Price of the vehicle

The choice of an appropriate input matrix is a major consideration in a PCA. The previous section, which provided a description of the variables, indicates that there are large differences between the variances of the variables. The correlation matrix is therefore an appropriate choice. We use the SAS®FACTOR procedure to perform the analysis. Table 4.10 presents the eigenvalues of the correlation matrix.

An important point in a PCA is the selection of an appropriate number of principal components. As exposed in the second chapter, there are many rules of thumb to determine this number. A first rule of thumb is to select the principal components with an eigenvalue larger than 1. This rule suggests the retention of

TAB. 4.10. Eigenvalues of the correlation matrix

PC	Eigenvalue	Proportion	Cumulative
1	2.0439	0.2271	0.2271
2	1.6132	0.1792	0.4063
3	1.0489	0.1165	0.5229
4	1.0014	0.1113	0.6342
5	0.9599	0.1067	0.7408
6	0.7811	0.0868	0.8276
7	0.7019	0.0780	0.9056
8	0.4402	0.0489	0.9545
9	0.4094	0.0455	1.0000

four principal components. A second rule uses a scree plot which is a graph of the eigenvalues versus the number of principal components. It is common to select the number of principal components that corresponds to an elbow in the plot. Figure 4.1 presents the scree plot for the collision claims. This scree plot suggests the retention of three principal components because an elbow occurs between the second and the third principal components. A third rule of thumb is to select the principal components that explain 70% of variance. This last rule suggests the retention of five principal components. Three rules give three different solutions. The business context however suggests to use no more than two or three principal components. The retention of three principal components however explains only 53% of the total variance. It is therefore inappropriate to use this technique with the data given by TDMMG.

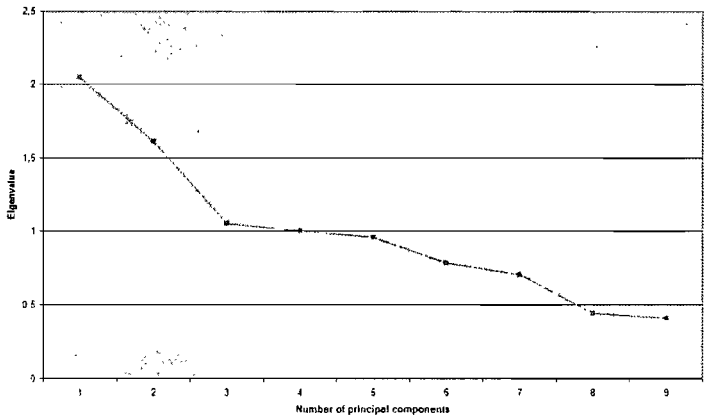

FIG. 4.1. Scree plot of the eigenvalues of the correlation matrix

The previous results indicate that a PCA is not useful to reduce the number of variables in our data set. While a PCA does not give satisfying results, the PRIDIT method actually gives interesting results. They are presented in the following section.

4.4. Principal component analysis of RIDIT

Brockett *et al.* (2002) propose an innovative unsupervised method to detect potential fraudulent claims at the screening level of the claiming process. The second chapter provided an overview of this method while the current section presents the results. This method requires simple matrix operations. However, the Base SAS system does not have such capabilities. The SAS/IML module allows such mathematical operations but was not available by the time of this project. This section therefore presents results conducted with the open software R.

The second chapter introduced the concept of RIDIT, which is a linear transformation performed on an ordinal variable in order to obtain an interval variable. More precisely, a RIDIT maps an ordinal variable into a [-1,1] scale. However, our data set mostly contains interval variables. To handle this issue, we discretize the interval variables into ordinal variables. We choose to discretize all interval variables into five-category variables such that all categories are of equal lengths. This method however results in a loss of information. Nevertheless, it allows us to use the RIDIT transformation in order to perform a PRIDIT analysis.

In Chapter 2, we mentioned that the PRIDIT method requires input variables that have categories ranked in a decreasing likelihood of fraud. To have such a data set, we use the *a priori* information given by the claim adjuster. For example, the claim adjuster told us that fraudulent claims are more likely to occur in autumn but gave no precise information on the other seasons. Hence, we give the lowest value to the "Autumn" category. In addition, we group the three remaining seasons into a same category and give it the highest value. In this section, the *Season of the loss* variable is therefore binary. Unfortunately, we do not have such *a priori* information for all variables. The claim adjuster gave

us information on nine of them. Only those variables may therefore be included in a PRIDIT analysis. Table 4.11 gives the RIDIT values for the nine selected variables.

TAB. 4.11. RIDIT values for the nine variables included in the PRIDIT analysis

t	**Variable**	B_{t1}	B_{t2}	B_{t3}	B_{t4}	B_{t5}
1	Catastrophe	-0.9989	0.0011	-	-	-
2	Season of the loss	-0.6961	0.3039	-	-	-
3	Time of the loss	-0.9640	0.0360	-	-	-
4	Ownership	-0.7873	0.2127	-	-	-
5	Claim free	-0.3603	0.4620	0.7251	0.8558	0.9530
6	Time interval before report	-0.9998	-0.9989	-0.9962	-0.9875	0.0096
7	Opening time	-0.9883	-0.9396	-0.8244	-0.5028	0.3703
8	Client since	-0.2686	0.6338	0.8969	0.9944	0.9999
9	Price of the vehicle	-0.9998	-0.9976	-0.9737	-0.6587	0.3172

Using the RIDIT values, the PRIDIT algorithm gives the following vector after 10 iterations:

$$\mathbf{W}^{(10)} = \frac{\mathbf{F}'\mathbf{S}^{(9)}}{||\mathbf{F}'\mathbf{S}^{(9)}||} = \begin{pmatrix} -0.0005 \\ 0.0722 \\ 0.0046 \\ 0.1175 \\ 0.7804 \\ -0.0029 \\ 0.1899 \\ 0.5705 \\ -0.1018 \end{pmatrix}.$$

This vector is very interesting. The *Claim free* and *Client since* variables both have large positive values (0.7804 and 0.5705). Those two variables are therefore good indicators of fraud. In other words, a policyholder who have submitted its last claim a long time ago is not likely to submit a fraudulent claim. Moreover, a loyal policyholder is not likely to submit a fraudulent claim. Those results may seem obvious but they give credibility to the PRIDIT method to detect insurance

fraud. Other variables are not significant. For obvious reasons, the vector of suspicion scores $\mathbf{S}^{(10)}$ is not presented here.

4.5. HIERARCHICAL CLUSTER ANALYSIS ON RELEVANT VARIABLES

This section presents a comparative study of the various hierarchical clustering methods introduced in the third chapter. The large number of claims is however a major problem because all hierarchical methods require huge computational resources. All attempts to perform a single link method on most available resources was still running more than two hours after it started. A solution to this problem is to select a random sample of k claims. This option is appropriate because the data set shows no meaningful order of observations. We therefore randomly select a sample of 100 claims without replacement. The analyses include the nine continuous variables. However, they exclude the five discrete variables because their inclusion gives poor results. In fact, our analyses show that using various types of variables is inefficient in the current context. We apply the transformation $\frac{X-\min\{X\}}{\max\{X\}-\min\{X\}}$ on the included variables. The traditional transformation $\frac{X-\mu}{\sigma}$ gives poor results. We perform the analyses using the SAS®CLUSTER procedure. Note that the analyses use a dissimilarity matrix of Euclidean distances. Although the analyses consider only 100 claims, a dendrogram is inappropriate due to the still large number of claims.

There are many criteria to determine the number of clusters in a cluster analysis. The third chapter introduced four of them. Table 4.12 reports the number of clusters obtained by using those criteria along with the five hierarchical methods also introduced in Chapter 3.

TAB. 4.12. Number of clusters chosen using the four criteria along with the five hierarchical methods

Method	R^2	Expected R^2	F	t^2
Single link	31	15	14	13
Complete link	14	15	2	2
Average link	17	15	3	2
Centroid	21	15	5	4
Ward	13	15	2	3

The second and third columns report the number of clusters chosen with the R^2 statistic and the expected R^2 statistic, respectively. The common way to use those statistics is to select the smaller number of clusters that explains at least 70% of the total variance. This rule gives a too large number of clusters for both statistics with values that range between 13 and 31. This rule is therefore inappropriate for our data set. The fourth and fifth columns report the number of clusters chosen with the pseudo-F statistic and the pseudo-t^2 statistic. A general rule for those statistics is to select the number of clusters that corresponds to a peak in the F and t^2 values. This last rule suggests the existence of an average of three clusters. The existence of two and four clusters are also good assumptions. From now on, we assume the existence of four clusters. Note that this rule is inappropriate when using the single link method since this algorithm tends to truncate the tails of the distribution (see Timm (2002)). This is clearly the case with results of 13 and 14 clusters. Those results suggest the use of the pseudo-F statistic and the pseudo-t^2 statistic to identify the number of clusters among collision claims. It also suggests the use of one of the three last methods (average link, centroid, and Ward) for clustering those claims.

We still have not identified the best method to cluster the collision claims. To do so, we compare the solutions given by the clustering algorithms. We exclude the single link algorithm because it gives poor results. Tables 4.13, 4.14, and 4.15 report the number of claims by cluster for the four methods under the assumptions of two, three, and four clusters, respectively.

TAB. 4.13. Number of claims for the four methods under the assumption of two clusters

Cluster	Complete	Average	Centroid	Ward
1	43	98	99	43
2	57	2	1	57

The complete link and Ward methods give identical results under the assumption of two clusters. They however provide different results under the other

TAB. 4.14. Number of claims for the four methods under the assumption of three clusters

Cluster	Complete	Average	Centroid	Ward
1	20	43	98	43
2	57	55	1	45
3	23	2	1	12

TAB. 4.15. Number of claims for the four methods under the assumption of four clusters

Cluster	Complete	Average	Centroid	Ward
1	20	41	95	18
2	49	55	3	45
3	23	2	1	25
4	8	2	1	12

assumptions. Both methods have a high discriminative power because they create relatively large clusters. Unlike the previous methods, the average link algorithm gives conservative results because it discriminates only the two most distant claims under the first assumption (two clusters). This method provides conservative results under the other assumptions with small clusters having as low as two claims. The centroid method also provides conservative results because it discriminates only the most different claims. This method discriminates only one claim under the assumption of two clusters, two claims under the assumption of three clusters and five claims under the last assumption.

No method outperforms the other. The four methods can be ordered on a scale based on their discriminative power with the complete link being the most discriminative method and the centroid method being the most conservative method. The ideal method depends on the resources allocated to investigate the potential fraudulent claims. The centroid method should be used with small resources and the Ward's method when large resources are available. This business decision is up to TDMMG.

It is interesting to note the hierarchical structure of the results. For example, the centroid method gives one cluster containing 99 claims and another one containing one claim under the assumption of two clusters. Therefore, this last

claim remains into a different cluster under the two other assumptions. In other words, when one observation is clustered into a different cluster, it will remains clustered whenever a larger number of clusters is assumed. This observation does not apply to the Ward's method.

From now on, we assume the choice of the Ward's method and the existence of four clusters. A useful step is to compute descriptive statistics for each cluster. They are presented in Table 4.16 for the continuous variables. Tables 4.17, 4.18, 4.19, and 4.20 presents the frequencies for four out of the five discrete variables. The random sample does not contain any claim part of a catastrophic event. Although the discrete variables were excluded from the analyses, it is interesting to see their distributions.

The most distant cluster is the cluster D. It is however not clear that this cluster actually contains the fraudulent claims. We can compare the descriptive statistics with the risk factors provided by the claim adjuster. The claim adjuster identified a high-valued car as a potential risk factor. The average price of the vehicles included into cluster D is clearly higher than the average price in the other clusters. The claim adjuster also identified a large number of days between the date of the loss and the date it is reported as a risk factor. Cluster D contains claims reported in average five days after the loss, which is higher than the time interval for the other clusters. The results are unfortunately not all in concordance with the *a priori* information given by the claim adjuster. For instance, the claim adjuster identified a recently insured policyholder as a potential risk factor for fraud. The claims in the cluster D are submitted by a policyholder insured for an average of 20 years while the other clusters have lower values on this variable. The same observation is also valid for the number of years without any claim for the policyholder.

This section introduced the results obtained from the most common hierarchical clustering methods performed on a sample of 100 collision claims. The five covered methods were the single, complete, and average link methods, the centroid method, and the Ward's method. We first compared four common rules to determine the number of clusters in the sample. On one hand, the results

TAB. 4.16. Average values for the four clusters using the Ward's method

Variable	A	B	C	D
Age of the vehicle	5.11	5.96	7.20	4.42
Claim free	0.44	1.73	1.96	3.92
Client since	2.94	5.27	7.52	20.0
Expenses	371	339	473	432
Indemnity	6,233	2,924	3,098	1,615
Opening time	84.2	69.7	67.8	62.5
Policyholder's age	35.3	41.9	48.9	56.2
Policyholder's liability	100.0	1.1	100.0	4.2
Policyholder's residence	820	601	384	364
Price of the vehicle	19,192	22,795	20,091	38,485
Recovery	122	0	76	0
Inforce policy	188	161	142	187
Time interval before report	2.22	2.78	4.00	5.00
Number of observations	18	45	25	12

TAB. 4.17. Distribution of the collision claims on Season of the loss variable by clusters

Season of the loss	A	B	C	D	Total
Spring	3	3	3	1	10
Summer	4	15	4	1	24
Autumn	6	17	9	4	36
Winter	5	10	9	6	30
Total	18	45	25	12	100

TAB. 4.18. Distribution of the collision claims on Time of the loss variable by clusters

Time of the loss	A	B	C	D	Total
Day	17	42	24	11	94
Night	1	3	1	1	6
Total	18	45	25	12	100

suggest the existence of two, three or four clusters. On the other hand, they suggest the use of a specific method based on the available resources at TDMMG. If we assume that the most distant cluster contains the fraudulent claims and the existence of four clusters, about 12 claims are potentially fraudulent. However, the results do not clearly indicate that those claims are fraudulent. Nevertheless, a particular attention has to be given to those claims.

TAB. 4.19. Distribution of the collision claims on Gender variable by clusters

Policyholder's gender	A	B	C	D	Total
Female	13	20	9	2	44
Male	5	24	16	10	55
Total	18	44	25	12	99

TAB. 4.20. Distribution of the collision claims on Ownership variable by clusters

Ownership	A	B	C	D	Total
Bought car	10	34	19	9	72
Rented car	8	11	6	3	28
Total	18	45	25	12	100

4.6. NONHIERARCHICAL CLUSTER ANALYSIS

The second main type of clustering methods is the nonhierarchical cluster analysis. Unlike hierarchical cluster analysis, those methods require a known number of clusters. In addition, they require less computational resources and reallocation of objects is now possible. The reader is referred to the previous chapter of this thesis for a discussion on the popular k-means algorithms. Both SAS®Enterprise Miner™(EM) and SAS®FASTCLUS procedure were used to generate nonhierarchical clusters. Since EM uses the FASTCLUS procedure to generate results, we have expected the same results with both softwares. Surprisingly, this is not the case. By the time of the project, documentation of what EM is actually doing was inaccessible and largely incomplete. In this section, we therefore briefly introduce EM. We prefer to emphasize on the FASTCLUS procedure.

EM is a software designed to perform most data mining tasks. The reader is referred to the first chapter for a discussion on data mining. While the FASTCLUS procedure may perform most nonhierarchical cluster analyses, the CLUSTER node of EM provides additional capabilities. One interesting addition to the basic procedure is the option of selecting the initial seeds using the principal components. In addition, the CLUSTER node may automatically determine the

number of clusters. EM also gives the relative importance of the variables entered in CLUSTER node. However, the official documentation of EM gives no clues on this ambiguous concept of relative importance.

As mentioned in Chapter 3, an important issue to consider with k-means algorithm is the selection of the initial seeds. A good approach is to compare results obtained from different methods of selecting those seeds. We compare the three methods introduced the previous chapter. We however do not consider seed replacement in this thesis. Table 4.21 gives the number of claims for each cluster with the first two methods and the third method with different radius, that is, 0.75, 1, and 1.25. The table shows very different results. The best we can say in this project is that there is no meaningful order of observations in the data set. This makes therefore all methods acceptable. We consider the simple random sampling method to be consistent with the previous section on hierarchical cluster analysis where a sample of 100 claims was selected. Using the results on the hierarchical methods and the results obtained with EM, we assume the existence of four clusters. We standardize the values with the min-max normalization. As the previous cluster analyses, we exclude the five discrete variables.

TAB. 4.21. Number of claims for the four clusters using three methods of selecting the initial seeds

Methods	**A**	**B**	**C**	**D**
First k complete claims	6,255	3,392	3,279	3,227
Random sampling	2,967	3,473	5,123	4,590
Radius (0.75)	3,035	4,581	5,000	3,537
Radius (1)	6,255	3,615	4,614	1,669
Radius (1.25)	3,473	5,124	4,589	2,967

Table 4.22 gives the average values for the 13 continuous variables. It is interesting to compare this table with Table 4.16. We can see that the Ward's method and the k-means algorithm give largely different results. Those differences tell us that we actually need more information of a claim adjuster in order to clarify the characteristics of a fraudulent claim. The effectiveness of any statistical

method depends on the quality of its interpretation. Unfortunately, we do not have sufficient information for an adequate interpretation. For Tables 4.23, 4.24, 4.25, 4.26, and 4.27 give the distribution of the five discrete variables by clusters.

TAB. 4.22. Average values for the four clusters using the k-means algorithm

Variable	**A**	**B**	**C**	**D**
Age of the vehicle	6.00	5.90	5.61	5.55
Claim free	2.35	2.38	1.49	3.23
Client since	7.97	6.43	4.62	9.99
Expenses	387	370	341	347
Indemnity	4,715	4,619	3,457	3,458
Opening time	70.8	70.7	68.7	66.3
Policyholder's age	44.4	43.3	39.2	47.8
Policyholder's liability	97.9	98.4	1.3	1.2
Policyholder's residence	548	579	773	338
Price of the vehicle	22,388	21,967	21,048	26,201
Recovery	147	144	59	87
Inforce policy	275	88	178	175
Time interval before report	4.15	4.68	3.29	3.93

TAB. 4.23. Distribution of the collision claims on Catastrophe variable by clusters

Catastrophe	**A**	**B**	**C**	**D**	**Total**
Not part of a catastrophe	2,965	3,469	5,119	4,583	16,136
Part of a catastrophe	2	4	4	7	17
Total	2,967	3,473	5,123	4,590	16,153

TAB. 4.24. Distribution of the collision claims on Season of the loss variable by clusters

Season of the loss	**A**	**B**	**C**	**D**	**Total**
Spring	301	279	486	441	1,507
Summer	884	1,011	1,658	1,420	4,973
Autumn	846	1,054	1,615	1,394	4,909
Winter	936	1,129	1,364	1,335	4,764
Total	2,967	3,473	5,123	4,590	16,153

TAB. 4.25. Distribution of the collision claims on Time of the loss variable by clusters

Time of the loss	A	B	C	D	Total
Day	2,840	3,321	4,974	4,437	15,572
Night	127	152	149	153	581
Total	2,967	3,473	5,123	4,590	16,153

TAB. 4.26. Distribution of the collision claims on Gender variable by clusters

Policyholder's gender	A	B	C	D	Total
Female	1,040	1,273	1,987	1,412	5,712
Male	1,918	2,194	3,119	3,159	10,390
Total	2,958	3,467	5,106	4,571	16,102

TAB. 4.27. Distribution of the collision claims on Ownership variable by clusters

Ownership	A	B	C	D	Total
Bought car	2,367	2,738	4,014	3,598	12,717
Rented car	600	735	1,109	992	3,436
Total	2,967	3,473	5,123	4,590	16,153

Cluster D is the most distant cluster using the Euclidean distance. However, there is no clear evidence that the claims belonging to the cluster D are actually fraudulent. For instance, claims of the cluster D are submitted by policyholders insured for an average of 10 years by TDMMG. Nevertheless, those results give an implicit segmentation of collision claims. Therefore, we expect them to be useful in some following statistical analyses. For example, it will be possible to select better samples for more precise statistical analyses of insurance fraud.

4.7. NEXT STEPS...

In this project, we considered two broad categories of methods to obtain new knowledge on insurance fraud, that is, methods based on the PCA and those based on the concept of cluster. The traditional PCA does not help to reduce the number of variables of the data set provided by TDMMG. The PRIDIT method is however promising. Since the objective of this thesis was to explore and to

describe various methods to detect insurance fraud, a small portion of it was voluntarily devoted to the PRIDIT method. Our analysis show that the two variables *Claim free* and *Client since* are good indicators of a fraudulent claim. Further analyses should be performed on more potentially important indicators.

The assumption that the most distant cluster contains the fraudulent claims is maybe too restrictive. Since we had no information on previous fraudulent claims, unsupervised methods were the only possible methods to use and they require this assumption. TDMMG should therefore take the problem from another perspective. An interesting approach is to consider supervised methods instead of unsupervised methods. For example, TDMMG might select a sample of recent claims, depending on their resources, in order to create a data set that contains this variable. Logistic regression seems particularly interesting.

This thesis covered four methods to detect fraudulent claims and most of the results were presented in this chapter. Main PCA results were first introduced. We find PCA not efficient to reduce the data set. In fact, it seems afterward that PCA is not an appropriate method when mining data. This conclusion is drawn from our experience with many of the TDMMG data sets. Second, we discussed the innovative PRIDIT method as proposed by Brockett *et al.* (2002). This method seems to be the best one to detect insurance fraud. The chapter then covered agglomerative hierarchical clustering results. More precisely, we studied the usefulness of those techniques to detect fraudulent claims. Those methods require large computational resources and may therefore prevent TDMMG from using them. Fortunately, some algorithms are developed to deal with this problem. The BIRCH and the CURE algorithms are two algorithms that use the ideas of the hierarchical methods but that reduce the required computational time. TDMMG should consider using those algorithms. This chapter ended with the k-means algorithm and its corresponding results. We however do not have the actual expertise to assess the potential value of those clusters. TDMMG should consult different claim adjusters and other claim staff members for interpretation.

CONCLUSION

Let us recall what we have seen in this Master's thesis. The first chapter introduced the reader to the main issues of insurance fraud. We also discussed the car insurance regime in Ontario. We concluded this first chapter with an overview of data mining. The presentation of statistical concepts began with the second chapter. We first described the types of data framework used in this Master's thesis. We next introduced the concept of RIDIT, which is simply a linear transformation of ordinal variables. Before introducing the innovative PRIDIT method, we discussed principal component analysis in order to understand this method. The third chapter covered the vast field of clustering techniques. This chapter began with an explanation of how to measure the proximity between two objects. We then considered two major types of clustering techniques. On one hand, we explained main hierarchical clustering techniques. On the other hand, we discussed non-hierarchical clustering techniques. We concluded our chapter on cluster analysis with a discussion on the choice of an appropriate number of clusters. The last chapter presented the data set given by Meloche Monnex and the results obtained when performing the methods introduced in the second and third chapters.

The main idea of this project was simple. Given a large data set of car insurance claims, is it possible to identify the fraudulent claims? However, we were facing an important problem since we had no information on previous fraudulent claims. We therefore choose an unsupervised method, which restrict largely the range of appropriate methods. Given a data set of claims, we first performed a principal component analysis to reduce the dimensionality of our data set. We then performed a PRIDIT analysis to identify the suspicious claims and the most

relevant variables. We finally used various clustering techniques to create clusters of fraudulent claims.

This thesis provides quantitative results obtained by various statistical methods. At first glance, the PRIDIT method seems to be a good option to detect fraudulent claims while the various clusters seem irrelevant to insurance fraud. It is maybe true but it may also be false. We tried to give to them the best interpretation we could. However, an in-depth interpretation of the results and clusters is beyond the scope of this thesis. In fact, we do not have the experience to do this. A claim adjuster should interpret the results. We frequently mentioned that in this thesis, but we feel it is not a sufficient but a necessary option to consider.

The main point is that any statistical method depends on the data set and there are unfortunately no magical algorithms. All statistical methods covered in this thesis showed to be efficient on examples and literature gives many examples of appropriate applications. By now, there is therefore no reason to discard those methods for subsequent analyses.

From now on, there are two options to get better results. They are not mutually exclusive but complementary to each other. The best is therefore to perform both options. First, TDMMG may continue with the unsupervised methods introduced in this thesis. The in-depth interpretation is necessary to precise the direction to take. Second, TDMMG may consider using a supervised method on a sample of claims already investigated. A logistic regression seems to be particularly interesting to us.

Finally, we would like to point out that this project has been completed during a major restructuring of the activities of TDMMG. In other words, the priority given to this project has dramatically decreased near the end of the project. It largely explains why a reader may find that the results are incomplete. Nevertheless, the results have, in practice, a great business value to TDMMG.

Bibliography

ANDERBERG, M. R. (1973). *Cluster Analysis for Applications.* Academic Press.

BERRY, M. J. A. and LINOFF, G. S. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management (Second Edition).* Wiley Computer Publishing.

BROCKETT, P. L., DERRIG, R. A., GOLDEN, L. L., LEVENE, A. and ALPERT, M. (2002). Fraud classification using principal component analysis of RIDITs. *The Journal of Risk and Insurance*, **69** 341–371.

BROCKETT, P. L., XIA, X. and DERRIG, R. A. (1998). Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. *The Journal of Risk and Insurance*, **65** 245–274.

BROSS, I. D. J. (1958). How to use ridit analysis. *Biometrics*, **14** 18–38.

COALITION AGAINST INSURANCE FRAUD (2003). The fourth of Americans say it's acceptable to defraud insurance. website: www.insurancefraud.org.

CORMACK, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society*, **A(134)** 321–367.

DERRIG, R. A. (2002). Insurance fraud. *The Journal of Risk and Insurance*, **69** 271–287.

DERRIG, R. A. and KRAUSS, L. (1994). First steps to fight workers compensation fraud. *Journal of Insurance Regulation*, **12** 390–415.

DIONNE, G. and GAGNÉ, R. (2002). Replacement cost endorsement and opportunistic fraud in automobile insurance. *Journal of Risk and Uncertainty*, **24** 213–230.

DIONNE, G., GIBBENS, A. and ST-MICHEL, P. (1993). *An Economic Analysis of Insurance Fraud.* Les Presses de l'Université de Montréal.

DUFFIELD, G. and GRABOSKY, P. (2001). *The Psychology of Fraud.* No. 199 in Trends and Issues in Crime and Criminal Justice, Australian Institute of Criminology.

EVERITT, B. (1980). *Cluster Analysis (Second Edition).* Halsted Press.

GORDON, A. D. (1987). A review of hierarchical classification. *Journal of the Royal Statistical Society,* **A(150)** 119–137.

GOWER, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics,* **27** 857–871.

GUHA, S., RASTOGI, R. and SHIM, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. *Information Systems,* **25** 345–366.

GUHA, S., RASTOGI, R. and SHIM, K. (2001). CURE: An efficient clustering algorithm for large databases. *Information Systems,* **26** 35–58.

HAN, J. and KAMBER, M. (2006). *Data Mining: Concepts and Techniques (Second Edition).* Academic Press.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag.

INSURANCE BUREAU OF CANADA (2006). Ontario auto insurance: Frequently asked questions. website: www.ibc.ca.

JAMBU, M. and LEBEAUX, M. O. (1978). *Classification automatique pour l'analyse des données.* Dunod.

JOBSON, J. D. (1992). *Applied Multivariate Data Analysis.* Springer-Verlag.

JOHNSON, R. A. and WICHERN, D. W. (2007). *Applied Multivariate Statistical Analysis (Sixth Edition).* Pearson Prentice Hall.

JOHNSON, S. C. (1967). Hierarchical clustering schemes. *Psychometrika,* **32** 241–254.

JOLLIFFE, I. T. (2002). *Principal Component Analysis (Second Edition).* Springer-Verlag.

KARYPIS, G., HAN, E.-H. and KUMAR, V. (1999). Chameleon: A hierarchical clustering algorithm using dynamic modeling. *Computer,* **32** 68–75.

LANCE, G. N. and WILLIAMS, W. T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. *The Computer Journal,* **9** 373–380.

LAROSE, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining.* John Wiley & Sons.

MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium*, vol. 1. 281–297.

SAS INSTITUTE INC. (2004). *SAS/STAT®9.1 User's Guide.* SAS Institute Inc.

SEBER, G. A. (1984). *Multivariate Observations.* John Wiley & Sons.

SOKAL, R. R. and SNEATH, P. (1963). *Principles of Numerical Taxonomy.* W. H. Freeman and Company.

TIMM, N. H. (2002). *Applied Multivariate Analysis.* Springer-Verlag.

VIAENE, S. and DEDENE, G. (2004). Insurance fraud: Issues and challenges. *Geneva Papers on Risk and Insurance*, **29** 313–333.

XU, R. and WUNSCH II, D. (2005). Survey of clustering algorithms. *Institute of Electrical and Electronics Engineers Transactions on Neural Networks*, **16** 645–678.

ZHANG, T., RAMAKRISHNAN, R. and LIVNY, M. (1996). Birch: An efficient data clustering method for very large databases. In *Proceedings of Association For Computing Machinery Special Interest Group on Management of Data.* 103–114.

APPENDIX

```
# By: Mathieu Poissant
# Date: August 12th, 2008
# This program computes the PRIDIT algorithm.

# The data set is entered here.
# V1 = c(1,1,1,0,1,1,1,1,1,1)
# V2 = c(1,2,2,0,0,1,1,2,1,0)
# V3 = c(3,3,3,1,2,3,0,3,1,2)
# Data = cbind(V1,V2,V3)

# Use this part if the data set is a delimited file.
Data = read.table("PRIDIT.data", header=T)

Data = as.data.frame(Data)
attach(Data)

# Sets the number of iterations of the PRIDIT algorithm.
nb.iterations = 10

# Determine the number of categories for each variable.
k = matrix(, nrow = ncol(Data), ncol = 1)
for(t in 1:ncol(Data))
{
    k[t] = length(unique(Data[,t]))
}

# Compute the RIDIT value for each category.
B = matrix(, nrow = max(k[t]), ncol = ncol(Data))
for (t in 1:ncol(Data))
{
    for (i in 0:k[t]-1)
    {
        B[i+1,t] = sum(Data[,t]<i)/length(Data[,t])-sum(Data
           [,t]>i)/length(Data[,t])
    }
}

# Compute the F matrix.
F = matrix(, nrow = nrow(Data), ncol = ncol(Data))
```

```
for (t in 1:ncol(Data))
{
    for (i in 1:nrow(Data))
    {
        Value = Data[i,t]
        F[i,t] = B[Value+1,t]
    }
}

# The PRIDIT algorithm.
S = matrix(,nrow = nrow(Data), ncol = 1)
W = matrix(1, nrow = ncol(Data), ncol = 1)
S = F %*% W    # S0


for (j in (1:nb.iterations))
{
    W = (t(F)%*%S)  / sqrt(sum((t(F)%*%S)^2))
    S = F %*% W
}

detach(Data)
```