UNIVERSITY OF SOUTHAMPTON

FACULTY OF BUSINESS AND LAW

Southampton Business School

DEVELOPING INSIGHTS RELATED TO PORTFOLIO MANAGEMENT AND INDIVIDUAL INVESTORS BY OVERCOMING PROBLEMS ASSOCIATED WITH ANALYSING LARGE SCALE FINANCIAL DATA.

by

Juan Carlos Moreno-Paredes

Thesis for the degree of Doctor of Philosophy

March 2018

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF BUSINESS AND LAW

Southampton Business School

Thesis for the degree of Doctor of Philosophy

DEVELOPING INSIGHTS RELATED TO PORTFOLIO MANAGEMENT AND INDIVIDUAL INVESTORS BY OVERCOMING PROBLEMS ASSOCIATED WITH ANALYSING LARGE SCALE FINANCIAL DATA.

By Juan Carlos Moreno-Paredes

Despite the evolution of data science during recent years, some problems still persist when studying decision making processes. Issues such as missing data, errors, outliers, imbalance, internal correlations and the lack of unique solutions have to be properly addressed to avoid erroneous inferences. This thesis, addresses these issues in three case studies of decision making problems in the general area of credit risk management, financial investment services and financial trading.

First, in the case of credit risk management, this work overcomes the problem of dealing with several scenarios that financial lenders have to face when trying to re-structure their credit portfolios. A framework is presented that allows the reduction of the solutions' selection and in consequence improve the risk management process within these organisations.

Second, within financial investment services, this thesis overcomes the challenges of profiling individual investors in the spread trading market by using ensemble data mining techniques. The application of such techniques, over this new domain, allows overcoming the complexities of profiling individual investors coming from different backgrounds in a very dynamic environment, and therefore improving the decision making process and risk management in retail brokers.

Finally, within the financial trading context, by applying the appropriate controls and modelling the internal correlations in a high volume of trading data, it is revealed whether new technologies, such as smart mobiles (tablets and smart phones) and their apps, effectively help individual investors make better decisions.

Table of Contents

Dedicado a todos aquellos que han iluminado mi vida en especial mis padres, esposa, hija y hermanos. Que Dios les colme de bendiciones por su gran e invaluable apoyo.

List of Tables

List of Figures

DECLARATION OF AUTHORSHIP

I, JUAN CARLOS MORENO PAREDES declare that this thesis and the work presented in it are my own and it has been generated by me as the result of my own original research.

DEVELOPING INSIGHTS RELATED TO PORTFOLIO MANAGEMENT AND INDIVIDUAL INVESTORS BY OVERCOMING PROBLEMS ASSOCIATED WITH ANALYSING LARGE SCALE FINANCIAL DATA.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

   JUAN C. MORENO-PAREDES et al. (2013). A Multi-Objective Decision Framework for Credit Portfolio Management. *Credit Scoring and Credit Control XIII Conference. Edinburgh 2013.* (Conference proceedings).
   https://www.business-school.ed.ac.uk/crc/wp-content/uploads/sites/55/2017/02/A-Multi-Objective-Decision-Framework-for-Credit-Portfolio-Management-Moreno-Paredes-Mues-and-Thomas.pdf

Signed: ....

Date: ....

Acknowledgements

1. To my beloved wife Raquel and daughter Mariana. Thank you very much for your invaluable support. Without you, it could not have been possible to achieve this goal in my life.
2. To all my family, in a very special way to my mother, father and brothers. Thank you very much for your great support.
3. To my supervisors Prof Lyn Thomas (RIP), Prof Johnnie Johnson, Dr Christophe Mues, Prof Vanessa Sung and Dr Tiejun Ma. Thank you very much for your advice, patience and support during this project. It has been a great honour to work with all of you.
4. Special thanks to Dr Raquel Florez-Lopez for your help and support.
5. Thanks to all the people from Postgraduate Office their great support and patience
6. To my supervisor Nick Lewis from LCG. Thank you for giving me the great opportunity to be part of your team and all the support received from you (personally and financially).
7. Thanks to all the people from LCG for their great support and patience.
8. To my research fellows colleges and good friends Pete Frazer – MacKenzie, Nicholas Kililis and David MacDonald. Thank you for your help.
9. To my good friend, Andrea Eross. Thank you for your friendship and great support.
10. To my good friend in *Mercantil Servicios Financieros* Dr Edgar Osuna. Thank you for your support, advice and friendship.
11. To my colleagues and supervisors in 4-Most Europe LTD. Thank you for your great support (personally and financially).
12. Finally, to my good friend Ingrid Fernández for encouraging to take this important step in my life. *Gracias totales*.

Definitions and Abbreviations

APP: An application, especially as downloaded by a user to a mobile device

AUC: Area Under the received operator characteristic curve

BCBS: Basel Committee on Banking Supervision

BIS: Bank for International Settlements

CDO: Collateral Debt Obligation

CD: Centroid Distance

CVaR: Conditional Value at Risk

CART also CT: Classification and Regression Trees

DE: Disposition Effect

DMC: Data Mining-based Classification

FDIC: Federal Deposit Insurance Corporation

FI: Financial Institutions

FX: Foreign Exchange market also known as FOREX

GCM: Global Criteria Method

GBM: Gradient Boosting Machine

GDPR: General Data Protection Regulation

GIGO: Garbage In Garbage Out principle

IT: Information technology

ITT: Information technology tools

LMM: Linear mixed models

LR: Logistic Regression

MOEA: Multi-Objective Evolutionary Algorithm

MOOP: Multi-Objective Optimisation Problem

NGCM: Normalised Global Criteria Method

NN: Neural Network

NSGA II: Non-Dominated Sorting Genetic Algorithm type II

OTC: Over The Counter markets

PDP: Partial Dependency Plot

ROC: Received Operator Characteristic

SMA: Smart Mobiles and their applications (apps)

SPV: Special purpose vehicles

SVM: Support Vector Machine

UK: United Kingdom

USA: United States of America

VaR: Value at Risk

Chapter 1: Introduction

1.1 Chapter overview

This introductory chapter provides the foundation and background of the topics studied in this thesis. In section 1.2 the decision making process is addressed as a main background of this thesis. Section 1.3 describes problems faced by those who analyse large scale financial data. The general statement of the problem and motivation are presented in section 1.4. The research objectives, research questions and data used in this work are presented and described in section 1.5. In section 1.6 the research methods used in this thesis are outlined. The research contributions and implications of this work are stated under section 1.7. The outline of this thesis is presented in section 1.8.

1.2 The decision making process

According to Tzeng and Huang (2011) the decision making process can be defined as a series of steps, which involve problem identification, establishing preferences and alternatives, and choosing the best alternatives that enable the decision maker(s) to find suitable solutions.

In general, decision making can be classified into two types: Strategic and Operational decisions. Strategic decisions are characterised by broad analysis of the different scenarios where the main objective is to visualise possible impacts in the long term. On the contrary, operational decisions are more tactical; they have a short term scope and their main aims are to trigger immediate responses or actions to solve particular situations.

When decision making processes are investigated, one element that needs to be taken into consideration is the decision maker. There are clear differences between individuals and organisations when they make decisions. See for example Shapira and Venezia (2001) and Brown *et al.* (2006), where behaviour of individual and institutional investors are compared. Moreover, it can be said that decision making in individuals tend to be more tactical, as they normally face several decisions on daily basis with a short term scope. Organisations, on contrary, not only have to deal with the challenges of the day to day business, but at the same time they have to make strategic decisions with much wider and longer term impact.

Perhaps, one of the most relevant elements that influences the decision making process is the degree of uncertainty of outcomes and consequences of decisions. The complexity of decisions is usually related to that degree of uncertainty. Consequently, both individuals and organisations try to minimise uncertainty using different processes.

1.2.1 The decision making process employed by individuals

According to Tversky and Kahneman (1974) individuals tend to rely on heuristics to cope with uncertainty. Heuristics can be defined as mental steps that people follow in order to assign probabilities to scenarios when assessing potential outcomes in the decision making process.

Goodwin and Wright (2014) summarise these heuristics as three types: i) Availability, when individuals assign probabilities to events based on how easy it is for them to recall these events in their minds; for example, if a person is asked to compute the probability of aircraft accidents, she/he will try to remember the most recent news where similar events have been reported. II) Representative, when probabilities are assigned based on how well such events fit into individuals' stereotypes, believes or pre-conceived ideas; for example, when person A is invited to determine the chances of person B being a sales executive, A may undertake the assessment based on his/her ideas and pre-conceptions of how a sales executive looks and behaves. III) Anchoring, in this case, probabilities are framed by individuals' initial thoughts of what these values should be.

Although heuristics can help individuals produce estimates of probabilities and therefore managing uncertainty within the decision making process, this way of thinking is usually subject to bias (Goodwin and Wright, 2014). For instance, the availability heuristic could cause people to underestimate probabilities when is difficult for them to recall previous experiences of the evaluated event. Similarly, the representative heuristic can lead individuals to hold on to their thoughts, ignoring objective data on the assessed event. Consequently, the influence of such biases has to be taken into consideration when studying the decision making process in individuals.

In the study of the decision making processes of individual investors, the Disposition Effect (DE) is the bias widely reported in the literature (see for example: Shefrin and Statman (1985); Odean (1998); Shapira and Venezia (2001); Kaustia (2004); Feng and Seasholes (2005); Barber *et al.* (2007); Choe and Eom (2009)). According to Shefrin and Statman (1985) the DE is the tendency of investors to hold their losing positions and to sell those in profit.

1.2.2 The decision making process in organisations

Organisations may use heuristics and be subject to those biases, after all an organisation is simply a group of individuals that share common visions and objectives. For instance, Shapira and Venezia (2001) and BrownChappelDa Silva Rosa and Walter (2006) expose evidence suggesting that institutional investors are subject to the DE bias, although to a lesser degree than individuals.

Organisations perhaps are less subject to bias because they make more complex decisions, as the potential impact and cost of such decisions are considerably higher in comparison with individuals. Organisations tend to be more systematic in their decision making process and therefore they seek data and reliable information. In consequence, it is not a coincidence that most of them have turned to data analysis science, hoping to get insights from data and mitigate uncertainty in the decision making process. Hill *et al.* (2006) provide a general overview of the development of this field within different industries. Then studying the decision making in organisations implies investigating the limitations of the approaches used by them as well as the challenges that they face when analysing data.

Over the last decades, data science has evolved considerably. Kurgan and Musilek (2006); Kriegel *et al.* (2007); Liao *et al.* (2012) survey the main contributions in this field. Advances in data science have also been propelled by computational power, allowing substantial increments in capacity of data processing as well the ability to perform more complex analyses and consequently get more insights from the phenomena under study. Rohanizadeh and Moghadam (2009) illustrate and describe the evolution of data science over the last five decades.

Despite the evolution of data science during recent years, typical problems still persist in the data such as missing data, treatment of errors and outliers, as well as dealing with data imbalance, i.e. one class or group being overrepresented (García *et al.*, 2009; López *et al.*, 2012), and the lack of unique solutions. Moreover, other relevant data issues began appearing; for instance, now researchers have the capability for collecting a wider spectrum of data, e.g. real-time data from individuals trading in the stock markets. However, collecting several transactions originated from the same individual could introduce internal correlations within the data (Laird and Ware, 1982; Verbeke, 1997; McCulloch and Neuhaus, 2001). Consequently, these issues have to be properly addressed and overcome when data is analysed, particularly in decision making studies, while taking into account that not all methodologies could be suitable for performing analyses over such data sets.

1.3 Challenges when analysing large scale financial data

The explosion of the internet has allowed organisations to collect significantly more data from their clients. An example of this will be Walmart / ASDA (one of the biggest retail shops in the USA and UK), which handles more than a million clients' transaction every hour (Kambatla *et al.*, 2014). The financial sector is not an exception, as the number of transactions that customers make using online platforms and e-commerce has considerably increased during recent decades.

Organisations recognise data as an important asset (Gopalkrishnan *et al.*, 2012), such that with the appropriate mining process it is possible to obtain valuable information from their customers and consequently develop more assertive marketing strategies (Linoff and Berry, 2011). In the financial sector in particular, the use of data has been mainly driven by the Basel accords (BCBS, 2005, 2011), which introduced the necessity of more accurate reporting and quantitative risk management. Consequently, financial firms have started harvesting and maintaining significant number of data repositories.

Despite the benefits of incorporating data analysis into the financial sector, there are still some concerns regarding data management and quality (see for instance GopalkrishnanSteierLewis and Guszcza (2012)), where elements such as accuracy, integrity, relevance, completeness and trustworthiness need to be guaranteed to secure comprehensive quality levels in the data (Cheong and Chang, 2007).

In that sense, the complexity of data integration from aggregated sources could be one of the main problems faced by the financial industry according to GopalkrishnanSteierLewis and Guszcza (2012). This process consists of consolidating data from different sources, both internal and external. Although thorough validation and reconciliation need to be completed to ensure data quality, consolidation processes can often be erroneous due mismatches between sources. For example, including geographical information from external sources can be challenging, as it requires finding common fields to join the datasets, which normally do not match perfectly, thus generating a final dataset with several missing values. In these circumstances, it is almost impossible to apply imputation techniques such as those suggested by Maletic and Marcus (2000).

Perhaps the increase in mergers and acquisitions within the financial sector, after the financial crisis of 2007-2009 (see for instances statistics offered by the FDIC (2013), has contributed to the increased requirement of data integration and aggregation, as merged companies need to consolidate their databases for reporting, risk management and governance.

New regulation may also challenge the financial sector's data management processes. For example, when the General Data Protection Regulation (see GDPR (2018)) comes into effect in

May 2018, data breaches must be reported to the regulators within less than 72 hours. This type of regulation will demand more complex and accurate monitoring systems as well as more robust data management processes.

Data storage and processing are explicitly referenced by Philip Chen and Zhang (2014) as other potential obstacles when manipulating large scale data. In particular, storage could be relevant in the financial sector as sensitive data such as clients' personal information must be kept under the organisation's total control, due security and regulatory reasons. Hence, solutions such as cloud storage may not be totally feasible for the financial industry. Similarly, in the case of data processing, the manipulation of large scale datasets will require considerable investment in software and hardware to be able to cope with enormous volumes.

In summary, financial organisations using large scale data may face several issues such as storage, integrity assurance and data processing, in a more regulated environment. Consequently, this will require not only significant investment in technology and new methodological frameworks, but also more robust data management governance as highlighted by Young and McConkey (2012).

1.4 Statement of the problems and motivations for study

This thesis addresses issues occurring within the data analyses of three different decision making problems: in the general areas of credit risk management, financial investment services and financial trading. First, in the case of credit risk management, this study aims to produce a methodological framework that optimises credit portfolios of commercial and retail lenders, focusing on improving the risk management process within these organisations. In this particular case, the study overcomes the problem of dealing with several scenarios that financial lenders have to face when trying to improve their credit portfolios. Second, in financial investment services, the research focuses on the investigation of the main drivers that influence the individual investors' behaviour who participate in the spread trading market, while aiming to get more specific insights of such drivers by using ensemble data mining techniques. The application of such techniques, over this new domain, overcomes the complexities of profiling investors coming from different backgrounds in a very dynamic environment, and in consequence improving the decision making and risk management in retail brokers. Thirdly, within the financial trading context, by applying the appropriate controls and modelling the internal correlations in a high volume of trading data, this investigation reveals whether new technologies, such as smart mobiles (tablets and smart phones) and their apps (SMA)[1], effectively help individual investors make better decisions.

In the first case study, the motivation for building a methodological approach for improving credit portfolio management comes from the failure of the financial system worldwide during the 2007-2008 crises, where several financial institutions (FIs) in USA and Europe were severely affected due the defaults in the real estate sector. As evidence of the magnitude of such failure, data released by the FDIC (2013) shows that the annual average of merging and acquisition of failed FIs increased from 4.5 FIs (2000 to 2007) to 93 FIs (2008-2012), representing an increment of almost 2000%. As a consequence, FIs are now facing a more regulated environment where decisions over their portfolios become more complex. For instance, a FI may decide to increase its exposure in a specific sector in its credit portfolio, aiming to improve that portfolio's return. This can potentially introduce more risk (i.e. concentration risk or default), which would have to be covered with more capital or reserves, generating additional costs to the detriment of the organisation's shareholders and probably making this lender much weaker. Therefore, it becomes necessary to develop robust quantitative frameworks able to simulate several potential impacts, and in this way help decision makers to assess several scenarios before the final decision is made.

In the second case study, decision makers within FIs have to focus on the short-term impact of their financial conditions (i.e. balance sheets and financial statements). For example, in retail brokers, a FI that allows individual investors to participate in financial markets, a team of monitors (dealing desk) is dedicated to oversee investors' positions and the risk that those investors are taken, particularly in leveraged transactions[2]. The dealing desk has to assess whether an investment of a particular individual could expose the broker into severe losses. As a result the dealing desk has to decide, within a very short period of time, whether this investor's position should be hedged in order to avoid potential losses. Consequently, it is important for retail brokers to profile their investors efficiently in order to decide which investors should be monitored carefully.

Investigating how to improve the risk management system of retail brokers becomes relevant, as they hedge their positions in the underlying markets. Moreover, overreaction in the hedging activity from these institutions could introduce volatility into financial markets and cause

[1] App: an application downloaded by a user to a mobile device

[2] Leveraged transactions are those ones where the losses can exceed investors' account deposits.

even greater losses to institutional investors. Additionally, when these companies fail to manage risk and go bankrupt, they can also harm individual investors' patrimony and undermine confidence in financial markets. For instance, in January 2015, several retail brokers faced high losses and bankruptcy after the volatility of the Swiss franc / Euro exchange rate caused investors' losses and instability in financial markets (FT, 2015).

The third case study examines the degree to which smart mobiles and their apps (SMA) improve financial decisions by scrutinising a large dataset of individual investors' decisions in the spread trading market. The motivation for studying the influence of SMA in the individuals' decision making process is the fact that this new technology has experienced one of the fastest penetrations into the market during the last 10 years (Cisco, 2015). This new technology is now well spread in the field of marketing, social networks, communications, psychology, medicine and finance (Nakasumi, 2012; Buijink *et al.*, 2013; Constantiou *et al.*, 2014). According to Bhömer *et al.* (2011), the impact of such technology in the decision making process has not been completely established due to its rapid deployment. Furthermore, with such a rapid increase to content and information access through smart mobiles (ConstantiouLehrer and Hess, 2014) and with more than a billion mobile apps (MobiThinking, 2013), a fundamental research question has emerged, that is, "*What is the impact of adopting smart mobiles on the nature and quality of people's decisions resulting from the use of this technology".* Given the size of the mobile related economy and the global adoption of SMA by billions of people, understanding such research question is imperative.

The next three subsections discuss the above three case studies in more detail.

1.5 Challenges when optimising credit portfolios

Developments in financial portfolio optimisation were initially presented by Markowitz (1952). This study established the bases for portfolio analysis and optimisation of risk and return measures in financial portfolios. Markowitz (1956) assumed normality of asset prices and used the standard deviation of asset prices as a risk measure. In practice, the normality assumption is not necessarily observed in credit risk portfolios (Cespedes, 2002).

Further research led to the introduction of more robust risk metrics, which do not assume normality in the price movements such as the Value at Risk (VaR) (Morgan, 1996) and Conditional Value at Risk (CVaR) in Artzner *et al.* (1999). Similar metrics were introduced for credit portfolios. In particular, Credit Suisse First Boston (CSFB, 1997) proposed a framework for credit risk management in which it is possible to compute the VaR and CVaR for credit portfolios.

Despite the development of VaR and CVaR for credit portfolios, the optimisation process for these portfolios is usually very complex as there is not an analytical algorithm that solves the optimisation problem over VaR and CVaR. Moreover, other risk metrics such as concentration risk, which arises when there is an uneven distribution of loans in the different sectors, have to be considered when assessing the entire risk of credit portfolios. Therefore, the optimisation of credit portfolios becomes challenging as several risk metrics have to be considered. Some authors sought to develop mathematical models able to integrate concentration metrics into the VaR and CVaR. Such integrated models are difficult to implement in credit portfolios of commercial and retail banks. Orenstein (2011) identifies several drawbacks of these integrated methods proved by the Citigroup failure case. Consequently, researches proposed multi-objective optimisation frameworks, which are able to incorporate several objectives in the credit portfolio optimisation problem (Schlottmann and Seese, 2004; Schlottmann *et al.*, 2005; Branke *et al.*, 2009).

On the other hand, meta-heuristics such as Multi Objective Genetic Algorithms have been proposed to perform optimisation on credit portfolios (Zopounidis and Doumpos, 2002; Schlottmann and Seese, 2004; Moreno-Paredes *et al.*, 2013). However, one of the major drawbacks of these approaches is the production of several potential sub-optimal solutions, where the solution selecting process becomes a problem itself. This thesis addresses these issues by developing a solution selection approach to narrow the search field and allowing the selection process to become less complex for decision makers.

1.6 Profiling individual investors

In the case study of profiling individual investors, the Prospect theory (PT) proposed by Kahneman and Tversky (1979), has to be acknowledged as one of the first attempts to explain how individuals make decisions under uncertainty. This theory is the cornerstone in explaining the most observed investors' behaviour such as the Disposition Effect (DE) (Shefrin and Statman, 1985). The DE is a tendency for investors to hold the losing/winning positions longer/shorter than they should be.

According to Barber and Odean (2011) the DE in investors is one of the main causes of failure of investors in financial markets. Furthermore, this behaviour has been observed in investors worldwide: USA (Shefrin and Statman, 1985; Odean, 1998); Israel (Shapira and Venezia, 2001); Australia (Kaustia, 2004); Taiwan (BarberLeeLiu and Odean, 2007); Korea (Choe and Eom, 2009); China (Yonghong, 2001; Feng and Seasholes, 2005; Chen *et al.*, 2007) and Finland (Kaustia, 2010).

Along with the DE, researchers have found other features that affect the performance of individual investors. For instance, lacking the ability to select diversified portfolios, performing high numbers of transactions, past performance (Barber and Odean, 2011; Nakasumi, 2012); experience (Seru *et al.*, 2010; Barber and Odean, 2011), socio economic level (Dhar and Zhu, 2006); knowledge of local markets (Ivkovic and Weisbenner, 2005) and the influence of peers and neighbours (Brown *et al.*, 2008; Kaustia and Knupfer, 2012).

Despite of the vast contribution in this field, the literature is still reporting the influence of these features in a very general way. For example, some researchers claim that senior investors perform better than younger ones (Greenwood and Nagel, 2009). However, these investigations do not capture whether there are differences within some age groups. In consequence, the precise nature of the relationship between these features and investors' performance remains under-researched.

Perhaps one of the reasons, why the literature continues reporting in a very general context, could be the limitations of approaches used by researchers to study investors' behaviour. In particular, General Linear Models (GLM) have been used by Anderson (2013), Korniotis and Kumar (2010) Korniotis and Kumar (2011) and Barber and Odean (2001) to study demographic characteristics (e.g. gender, age, income, DE, social background and risk attitude of investors). A similar approach is used by Grinblatt and Keloharju (2009) for comparing the performance of local and foreigner investors. Barber and Odean (2000) and Barber *et al.* (2014) used time series analysis to study features such as investors' overconfidence and past performance.

Despite the popularity of GLM and Time Series models among the research community and their capacity for getting insights from data, these methodologies have several shortcomings. One of them is the underlying assumption in the estimation of their parameters. For instance, Lemeshow *et al.* (1995); Mitchell (1997); Rushton (2000) point out that these assumptions can mislead researchers' findings. In consequence, conclusions from previous research could be incomplete. This thesis fills this gap in the literature by proposing the use of more advance data-mining techniques, able to overcome the shortcomings of GLM and Time Series models and, therefore, enabling researchers to obtain more precise insights of how investors' features affect their performance.

1.7 Difficulties to determine whether technological aids improve the decisions making process

The third case study examines the degree to which the SMA technology improves financial decisions by scrutinising a large dataset of individual investors' decisions in the spread trading market.

Molloy and Schwenk (1995) argue that information technology (IT) improves decision making processes and permits individuals and organisations to access more accurate and updated data to enhance their analysis. Such advantages have boosted the development of the mobile internet and particularly SMA, as these provide access to information, allow interactions with the rest of the world and enable decisions almost anytime and anywhere (Sraeel, 2006; Koenig-Lewis *et al.*, 2010; Kourouthanassis and Giaglis, 2012; BuijinkVisser and Marshall, 2013; ConstantiouLehrer and Hess, 2014).

Despite the apparent benefits that these technologies offer to decision makers, it cannot be assumed that them will necessarily lead to an improved decision making process. For example, there is considerable research which suggests that information overload can have an adverse impact on decision quality (Jacoby *et al.*, 1974; Jacoby, 1984; Eppler and Mengis, 2004; Bawden and Robinson, 2009). Furthermore, elements such as individual preferences, context and interaction with other sources of information can exert a significant influence on the results of decisions facilitated by these technologies (Venkatesh *et al.*, 2003; Nakasumi, 2012; Venkatesh *et al.*, 2012; ConstantiouLehrer and Hess, 2014). In consequence, establishing the impact of a particular technology such as the SMA within the decision making process, seems to be less straightforward as there are several elements which can influence this process.

One of the elements that should be taken into account when studying the impact of a particular technology is the digital divide effect. The digital divide is a socio-economic inequality in the population as a consequence of limitation in access and use of information technologies (Norris, 2001; Van Dijk and Hacker, 2003; Forman, 2005; Philip *et al.*, 2017; Tsetsi and Rains, 2017). Furthermore, research studies show the phenomenon of the second level digital divide (Norris, 2001; Graham, 2014), which was sourced from material factors such as the type of Internet connection users have and the frequency of access to the Internet (e.g. Mobile Internet). Therefore, it is expected that SMA users tend to be different, probably "smarter" or they have better decision-making skills.

Previous researches, such as Nakasumi (2012), ConstantiouLehrer and Hess (2014) and BuijinkVisser and Marshall (2013), who investigated the use of SMA to aid decision making in several areas such as marketing, medicine and finance, have not taken into account the effect of external contexts and therefore they do not account for the digital divide effect. Therefore, results from these studies may not be conclusive.

This thesis overcomes the gaps in the literature by studying the decision making behaviour of investors, when using their SMA as a trading channel operating in the Spread Trading Market. In this dynamic environment, investors have to make several decisions in a short period of time (normally within minutes), limiting the interference of external contexts and interaction with other sources of information, as mentioned by VenkateshMorrisDavis and Davis (2003); Nakasumi (2012); VenkateshThong and Xu (2012); ConstantiouLehrer and Hess (2014). The investigation in this thesis also controls for demographic differences securing the inclusion of the digital divide effect, which has been previously neglected in the literature.

1.8 Research objectives, research questions and data used in the analysis

The research objectives of this thesis are:

- To establish a methodological framework which enables decision makers at lending institutions to deal with a variety of scenarios when managing credit portfolios.
- To demonstrate the suitability of ensemble based data mining techniques to overcome the challenge of profiling individual investors, when the modelling involves high dimensionality (i.e. many independent variables), heterogenic population (investors coming from different backgrounds) and data associated to such population may contain errors and omissions.
- To establish profiles of users and non-users of the SMA technology, in order to determine whether there are differences in the type of people who adopt this technology.
- To examine to what extent the SMA as a technology actually improves the quality of decisions.
- To overcome the difficulties of isolating interfering drivers, which may create noise, when studies on large – scale financial data are undertaken.

The following research questions are addressed in this thesis:

> In the first case study, where a framework for improving the decision making process in financial lenders managing credit portfolios is developed, the following research questions are considered:

1. What should be an adequate framework to undertake an optimisation process aiming to improve the performance of credit portfolios?
2. As some optimisation processes produce several potential solutions to be considered, what should be the process to guarantee a degree of consistency and stability in the selected solutions?

In the second case study, when attempting to profile individual investors for improving decision making in retail brokers, the research questions are:

1. Are ensemble methodologies such as Gradient Boosting Machine (GBM) suitable for profiling potential outstanding individual investors (eventually risky for brokers), given the fact that individual investors exhibit a wide range of different characteristics and backgrounds?
2. How do these characteristics interact and influence traders' performance?

Finally, in the last case study, where trying to determine the influence of SMA onto individual's decision–making, the research questions are:

1. To what extent individuals who use SMA for trading are distinguished by their demographic profile and does this technology therefore establish a digital divide between these populations (i.e. users and non-users of the SMA technology)?
2. What are the main differences in the decision-making behaviour and performance of individual investors who do and do not use SMA as a trading tool, whilst controlling for any digital divide effect?
3. Is there any improvement in the decision making of investors when trading via SMA?

This study uses the following dataset:

- Secondary data provided by a US lender with more than US$ 2.8 billion assets corresponding to 2,557 loans.
- Secondary data provided by a UK retail broker with more than 4.5 million trading records over 9 years corresponding to 5,184 individual investors using different trading platforms, including SMA.

1.9 Research methods

To address the credit portfolio optimisation problem the following operational research methods are combined:

1. An actuarial approximation to compute portfolio risk. This method was originally proposed by Credit Suisse First Boston (CSFB, 1997) and adapted by Bürgisser *et al.* (1999) and Haaf *et al.* (2004).
2. For modelling the optimisation problem, the multi-objective approach proposed by Zopounidis and Doumpos (2002) and Schlottmann and Seese (2004) is adapted.
3. For solving the optimisation problem, a Multi-Objective Evolutionary Algorithm is developed based on type II Non-dominated Sorting Genetic Algorithm (NSGA II) proposed by Deb (2008).
4. For selecting solutions and testing stability an approach based on the Global Criteria Method (GCM) proposed by Deb (2008), is developed.

For determining the main drivers that categorise individual investors, the following research methods are used:

1. Four well know classifiers Logistic Regression, Classification Trees, Neural Networks and Support Vector Machines, plus GBMs.
2. GBMs' Partial Dependency Plots.
3. Controlling for factors that may influence traders' behaviour based on the drivers described in the literature.

In the case study of how SMA is impacting the decision-making of individual investors, the following approaches are used:

1. A long-term longitudinal study via linear mixed models and logistic regression to investigate changes in individuals' behaviour caused by particular factors, such as trading with SMA, demographic differences and market conditions.
2. Controlling for factors, which may influence individuals' trading performance.

1.10 Main contributions and implications of this thesis

The main contributions of this work are:

I. The methodological framework for optimising credit portfolios, proposed in this thesis, allows the portfolio planning in a long-term horizon and shifts the discussion from asset to sector planning. Therefore, the framework enables potential improvements in the credit risk management within retail lenders.

II. Additionally, the introduction of a Normalised Global Criteria Method, for selecting preferred portfolios, enables decision makers to overcome the problem of selecting a suitable solution amongst several potential solutions coming from the multi-objective optimisation.

III. In the case of profiling individual investors, the suitability of using ensemble based data mining techniques, for studying and profiling individual investors, is established. The application of this approach, over this type of problem, enables researches to investigate in depth the key drivers that influence the performance of individual investors.

IV. In the study of the use of SMA for trading, the conducted research in this thesis demonstrates that this technology has significant impact on the decision making process of individual investors. Additionally, this study confirms the digital divide theory is also applicable within the financial investment context, as the study finds evidence pointing out clear demographic and behavioural differences between those who do and do not choose to use SMA for trading.

V. Finally, the development and adaptations of several methodologies presented in this thesis permit overcoming the most common problems occurring when using large scale data for undertaking quantitative studies on decision making process in the financial industry.

The implications of this study are the following:

A. In the case of credit portfolios, the methodological proposal can be used as guidance by portfolio managers to improve their credit portfolios.

B. Profiling traders effectively helps retail brokers to improve their hedging strategies. This last element becomes relevant considering these FIs exercise their hedging strategies directly in financial markets[3]. In other words, an effective approach that enables the identification of potential risk investors could avoid overreaction of these companies and consequently reducing their potential negative impact in financial markets.

C. Investigating the impact of smart mobile devices on individual investors' behaviour can help understand how, this new and widely used technology, may influence the social behaviour, under this new emerging mobile-based environment, where elements such as mobile commerce (m-commerce) and social commerce (s-commerce) are the new tendencies in social economical behaviour as argued by Kourouthanassis and Giaglis (2012).

D. With the migration of internet connections towards mobile platforms and the popularity of SMA (Cisco, 2015), it is expected that more retail brokers adopt this new technology as a trading channel for their clients. The research conducted in this thesis suggests that

[3] For example, if a broker's client bets a big quantity in favour of the pound/dollar exchange rate, the broker can buy an option in the forex market to cover the exchange rate risk.

both elements (popularity and migration) can lead individuals' investors to increase their participation in different financial markets, as the interaction via SMA makes trade easier and more accessible. This last fact could introduce more noise and volatility into financial markets.

E. Finally, the use of alternative methodologies allow to overcome issues commonly presented in data associated to decision making process such as solutions' selection (in portfolio's optimisation), overrepresentation, highly dispersed data with complex interactions and relationships (in the case of profiling risk traders), and data with internal correlations and segmentations (in the study of influence of the SMA technology over investors' decision making), which allows researchers to get additional insights as well as improving their results. Therefore, the use of such methodologies in this thesis could provide a new generation of models that will be used for optimisation, classification and studying large scale data associated with complex decision making process.

1.11 Outline of this thesis

The thesis is organised into five chapters. In Chapter 2, namely "A Multi-Objective Decision Framework for Credit Portfolio Management of Non-liquid Assets," the first case study is presented, where a methodological framework for improving the performance of credit portfolios in commercial and retail lenders is produced, while addressing the problem of strategic decisions. Under the title "Predicting investor's success in the spread trading market" in Chapter 3, the drivers and most appropriate methodology which help to profile individual investors, particularly the ones which can produce financial losses to retail brokers, is explored. The third case study under the title "Investors on the move: Do smart mobiles apps enhance financial decisions?" is investigated in Chapter 4, where an empirical analysis is undertaken in order to explore to what extent new technologies, such as the SMA, influence the decision making process of individuals. Finally, in Chapter 5, the general conclusions and limitations of the whole study are presented as well as a potential expansion of the topics investigated in this thesis.

Chapter 2: A Multi-Objective Decision Framework for Credit Portfolio Management of Non-liquid Assets: A Case Study of Commercial and Retail lending

Abstract

A novel framework is proposed that supports financial institutions as they make strategic decisions with regards to their credit portfolios. Specifically, it addresses the question of how retail and commercial banks shall allocate their resources across various sectors (group of loans) in order to meet the competing criteria of maximising profitability and minimising risk. The framework comprises mathematical models, operational research techniques and solution selection approaches which allow portfolio managers to identify and explore alternative strategies. An example case, where this framework is applied to data provided by a US commercial and retail bank, shows how both risk and return can be improved whilst simultaneously diversifying a credit portfolio.

Keywords: Credit Risk Management framework, Multi-Objective Optimisation, Credit VaR, Credit Expected Shortfall.

2.1 Introduction

Since the financial crisis of 2008 - 2009, failures of financial institutions (FIs) around the world have significantly increased as a consequence of the collapse of the Real Estate subprime market in the US and UK, where a substantial number of banks were concentrated. In the US, for instance, figures released by the FDIC (2013) show the annual average of mergers and acquisitions of failed FIs has increased from 4.5 (2000 to 2007) to 93 (2008-2012), which represents an increment of almost 2000%. Consequently, risk management in the financial sector is now facing a tightly regulated environment along with the increased correlation and globalisation of financial markets.

The decision making process becomes more complex and less intuitive in FIs, particularly when the main challenge is to quantify the impact of decisions on their long term financial conditions. In the case of commercial and retail lenders, quantifying such impact is even more challenging, as their credit portfolios are made out of non-liquid assets such as private loans, which unlike corporate loans and bonds, are not publicly traded in exchange markets. Therefore, if a bank decides to either increase or reduce its exposure in its credit portfolio, in order to improve the portfolio's risk/return profile, it is not straight forward in the short term to either increase the number of assets (by acquiring more loans) or dispose of them (by selling loans). In consequence, the management process over this type of portfolios is normally done via long term investment strategies in specific economic sectors.

In addition, the decision making process for investment allocation has to consider multiple risks that could cause losses (e.g. concentration risk and default). In the case of FIs, these risks must be covered with more capital or reserves, generating additional costs affecting profitability and thus shareholder return.

Although the Basel II and III accords (BCBS, 2005, 2011), adopted by regulators worldwide, have promoted improvements in portfolio management, they do not provide explicit guidance on how to perform optimisation in credit portfolios. Hence, there is still considerable scope for developing new quantitative methods for profit/risk optimisation.

Another driver toward the use of more quantitative approaches in credit portfolio analysis is the fact that FIs increasingly have been looking to improve their liquidity by transferring credit risk via securitisation using financial instruments such as Collateralised Debt Obligations (Schlottmann and Seese, 2004). The role of these instruments in the financial crisis of 2008-2009 has come under considerable scrutiny (Bezemer, 2009; Colander *et al.*, 2009; Adrian and Shin, 2010; Ivashina and Scharfstein, 2010). Hence, developing methodologies which consider portfolio behaviour could reduce the possibility of future financial crises.

2.1.1 Perspectives of credit portfolio management

Credit portfolios of retail and commercial banks are a collection of illiquid loans; therefore, they are not easy to sell to a third party[4] unless they are packaged into "special purpose vehicles" (SPV). Each loan is a contract between a FI and an individual or legal entity[5] (called obligor), whereby the FI lends an amount of money to the obligor who agrees to repay it over time. In some cases (e.g. mortgage loans), the obligors have to put up some assets as collateral, so that, in the event the obligor defaults on their payments, the FI could decide to repossess these assets in order to recover the money that has been lent; these are called secured loans, whereas other forms of lending (e.g. personal loans, credit cards) may be unsecured. Another distinction is

[4] Normally, another financial institution.
[5] Individuals, SME and private companies (without credit rating).

between different types of obligors (e.g. individuals or small and medium-sized enterprises (SMEs)). A strategic decision the FI now faces is how to allocate lending over these different loan sectors (or segments), so as to balance two competing goals or perspectives – maximise profitability and/or minimise risk. Therefore, portfolio management can be viewed from two different perspectives: the profitability and risk perspectives.

The *profitability perspective* to portfolio management suggests the main objectives of a portfolio manager are to spot investment opportunities and increase the FI's performance. Therefore, elements such as return, profitability, pricing and planning are the key factors. In contrast, looking at the same problem from a *risk perspective*, portfolio managers have to identify and anticipate potential losses. If they can assess and quantify these losses, then they may be able to mitigate them by setting limits on investments in certain sectors of loans and using capital reserves to cover those potential losses (Thomas, 2009; Van Gestel and Baesens, 2009). The diagram in Figure 2.1 represents these two perspectives as well as the interaction of the different elements within the credit portfolio.

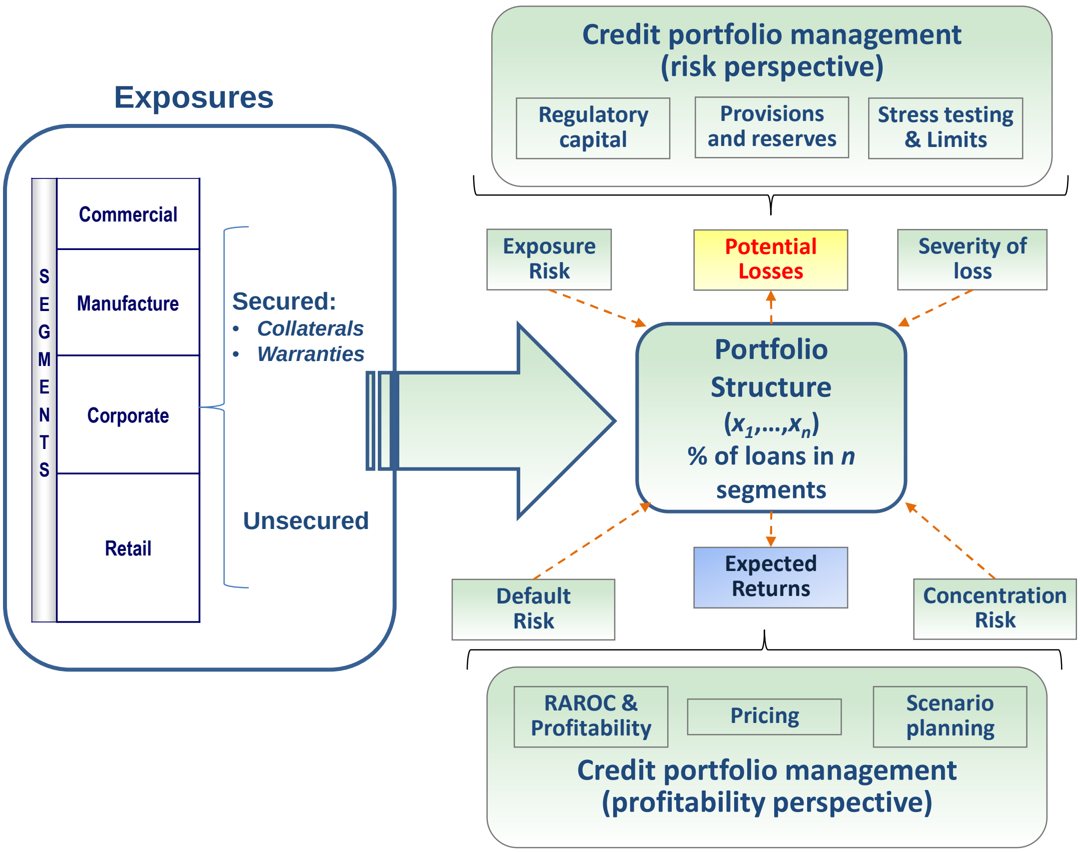


Figure 2.1 Perspective and interactions in a credit portfolio.

2.1.1.1 Risk perspective

Figure 2.1 shows the risks faced by the FIs when holding credit portfolios in their balance sheets. These threats can lead to real losses in the credit portfolio but at the same time present opportunities for profit making.

The risk perspective encourages managers to carry out sensitivity analyses such as stress testing to estimate the losses that can occur in a given credit portfolio. This perspective is enforced by regulatory bodies, such as central banks, to preserve the health of the financial system in each country. They impose a set of rules and regulations that FIs should fulfil. The Basel Accords have become the reference around the globe for such risk assessments. In particular the Basel III Accord (BCBS, 2011) highlights the importance of portfolio management. For example, BCBS (2011 parg. 113) is about assessing individual and portfolio assets, BCBS (2011 parg. 29 - 31) refers to how to monitor and avoid excessive credit growth, and BCBS (2011 parg. 115) addresses concentration risk and portfolio monitoring. Thus, the new Accord outlines the components of a comprehensive risk framework. However, there is not more specific guidance for performing portfolio optimisation, perhaps due to the complexity of modelling the optimisation process itself.

2.1.1.2 Profitability perspective

Secondly, portfolio managers look at improving the profitability or performance of their credit portfolio. One objective is to find investment opportunities, such as sectors where there are possibilities to expand the credit portfolio issuing loans with higher and more stable returns. In order to accomplish this, the analysis of returns over various scenarios of market conditions, transactional costs, and pricing, is needed. More specifically, once the portfolio's measures for one scenario (PD, LGD, etc.) have been established, is possible to run optimisation models for minimising risk and maximising returns, producing solutions that will be the potential strategies to consider by portfolio managers.

2.1.2 Objective, research questions and contributions of this paper

The main objective in this paper is to propose a practical framework for optimising credit portfolios in retail and commercial FIs, considering investment allocation by sector and taking into consideration the profitability and risk perspective of the portfolio. In order to achieve this objective, the following research questions are considered:

a) What would be a suitable mathematical model to represent the credit optimisation problem, taking into account the profitability and risk perspectives?

b) What operational research techniques should be used to solve these models and find strategies (i.e. reducing or increasing exposures in particular sectors) that would improve credit portfolio performance?
c) Are the strategies derived from this framework sufficiently stable to help establish a comprehensive portfolio management decision making process?

A key difference with previous studies is the use of a real credit portfolio of more than 2500 loans from a commercial and retail bank in the US with a value of US$ 2.8 billion. Also, in contrast with previous studies, the framework proposed in this paper focuses on the long term credit portfolio management of retail and commercial lending, taking into consideration market conditions and concentration risk; hence, the main objective is not only reducing the portfolio's associated risk but also increasing its return and therefore considering both the risk and the profitability perspective.

The contributions of this paper are summarised as follows:

1. Extending previous literature by introducing a framework for commercial and retail credit portfolios.
2. Extending the use of Multi-Objective Evolutionary Optimisation by introducing selection methods that guaranties stable solutions.
3. Enabling the integration of the profitability as well as risk perspectives in the optimisation process, i.e. increasing the portfolio's profitability, reducing capital allocations and concentration risk within a credit portfolio.

This paper is organised as follows: section 2.2 visits the relevant literature supporting this study. Section 2.3 describes the proposed process of modelling the credit portfolio. Also, the notation to represent the credit portfolio optimisation problem and its associated measures are established in this section. Next, the methods used for computing and selecting solutions, for the credit portfolio optimisation problem, are presented in section 2.4. Section 2.5 outlines the proposed framework. Section 2.6 presents a practical case to illustrate how the framework operates, followed by an analysis of the stability of the solutions obtained. Finally, section 2.7 draws the conclusions of this research.

2.2 Literature review

2.2.1 Credit risk and credit portfolio optimisation

There is significant literature about techniques and methods to estimate credit risk for retail lending (i.e. lending to individuals and small business) at the individual level (e.g. Thomas *et al.* (2002); Thomas *et al.* (2004); Thomas (2009); Tong *et al.* (2012); Fitzpatrick and Mues (2016)). The literature for corporate and sovereigns bonds is even larger; see for example the books by Saunders and Allen (2010), Van Gestel and Baesens (2009), Lu *et al.* (2013) and Bo and Capponi (2016). Nevertheless, risk assessment and optimisation of credit portfolios are still under investigation, particularly in the case of retail and commercial lenders where there is no public data of the risk that their assets have.

Markowitz (1952) established the key foundations of portfolio analysis and subsequently Markowitz (1956) proposed the "critical line algorithm" to perform optimisation over portfolios. Despite these remarkable contributions[6], the analysis and optimisation process proposed by Markowitz (1956) assume normality of the asset prices and use the standard deviation of these prices as a risk measure. In practice, this normality assumption is not necessarily observed in credit risk portfolios (Cespedes, 2002). Therefore, a different theoretical framework should be considered when analysing credit portfolios.

The first steps for establishing frameworks for credit portfolio assessment were made by CSFB (1997) with CreditRisk+, a model based on actuarial approaches that allows estimating the risk of credit portfolios, whilst overcoming the normality assumption on the assets valuation. CreditRisk+ has been the cornerstone for further research development in credit portfolio management and optimisation. For instance, Gundlach and Lehrbass (2004) compile several papers with applications of CreditRisk+ in the banking industry; Han and Kang (2008) provide an extension of CreditRisk+ for portfolio management; Han (2016) uses CreditRisk+ for modelling risk factor correlation and dependency between default and losses.

Schlottmann and Seese (2001) combines CreditRisk+ and evolutionary algorithms to perform optimisation on credit portfolios. In Schlottmann and Seese (2004), they extended this framework by using hybrid multi-objective evolutionary algorithms allowing improvements in running time. However, their approach modelled assets with a binary representation (loans that

[6] John von Neumann Theory Prize (1989) and Nobel Prize (1990)

are included and excluded in the portfolio). This assumption could be challenging to implement as loans are not liquid assets themselves.

Evolutionary algorithms are also used by Ivorra *et al.* (2009) for optimising Collateral Debt Obligations (a securitisation instrument of mortgages). However, these authors only use Value at Risk (VaR) as a risk metric, not considering other risk measures such as concentration risk. Also, the mathematical limitations, that the VaR has as risk measure, are well established in the literature (ArtznerDelbaenEber and Heath, 1999; Tasche, 2002). Furthermore, IvorraMohammadi and Ramos (2009) do not address the problem of choosing among the vast number of outcomes produced by the evolutionary approaches.

Andersson *et al.* (2001) developed an optimisation framework considering Conditional Value at Risk (CVaR) as risk measure for portfolios of corporate and sovereign bonds. Similarly, Bo and Capponi (2016) use robust optimisation over the same credit portfolio configuration. In those cases, the authors only focused on the optimisation of a single objective, not considering profitability and concentration risk. Furthermore, all of them only consider corporate and sovereign bonds in their studies, which tend to be more liquid and therefore more easily traded than private loans.

2.2.2 Risk metrics for credit portfolios

When assessing the overall risk in credit portfolios it is relevant to consider the interaction between the portfolio's assets. Cespedes (2002) shows that the correlation of defaults can produce bigger losses within a credit portfolio. Additionally, issues such as concentration, correlation and contagion[7] make the measurement of risk in a credit portfolio more complex (Lütkebohmert, 2009; Herbertsson, 2011). Therefore, it is imperative to consider several metrics for assessing the different risks in credit portfolios.

After the "Black Monday" financial crisis in 1987 (Barro, 1989) and the collapse of the Baring Bank in 1995, the JP Morgan group (Morgan, 1996) adopted Value at Risk (VaR) – i.e. a one-factor extension of the Merton model (Merton, 1976) – as a measure for unexpected losses in a portfolio. This is also the measure adopted and endorsed by the Basel II Accord (BCBS, 2005).

[7] Credit contagion arises when a firm files for bankruptcy (normally a bigger one) and this affects the credit performance of the firms in the same sector (Jorion and Zhang, 2009).

Although widely used, VaR is only one of a series of possible risk measures having different properties. ArtznerDelbaenEber and Heath (1999) suggest that for any such risk measure to be "acceptable", it must be mathematically coherent. Even though VaR is in widespread use in the financial sector and supported by BCBS (2005), it is not a coherent risk measure as it fails to satisfy the subadditivity property and the invariance law (ArtznerDelbaenEber and Heath, 1999; Tasche, 2002). Lack of subadditivity complicates the analysis of the risk contribution of a portion of the portfolio (ArtznerDelbaenEber and Heath, 1999; Tasche, 2002). Moreover, due the fact that VaR is law invariant, it can produce the same results for portfolios with thin and fat tails (see e.g. Embrechts *et al.* (1997)), the latter being one of the typical characteristics of the distribution of losses in credit portfolios (Cespedes, 2002). Consequently, Acerbi and Tasche (2002) argued the case against VaR being a suitable measure for estimating the unexpected losses in financial portfolios.

In order to tackle the problems inherent in the VaR measure, Tasche (2002) therefore proposed using Expected Shortfall (also known as Conditional VaR or CVaR) as a risk metric to estimate unexpected losses and be able to decompose the risk contribution of different portions of the portfolio. Additionally, AnderssonMausserRosen and Uryasev (2001) shows that minimising the CVaR also implies minimising the associated VaR of a portfolio.

Another factor that can affect portfolios is concentration in sectors. For instance, in the global financial crisis of 2008, commercial and retail banks whose lending was heavily concentrated in the real estate sector were severely affected when the subprime mortgage market collapsed (Bessis, 2011). Therefore, a concentration measure may be required to further ensure the risk soundness of a FI's strategy.

Concentration risk arises when there is an uneven distribution of the assets in the different sectors. Lütkebohmert (2009) surveys the different methods for calculating concentration risk in credit portfolios, classifying them in two groups: ad hoc methods and model based methods. She proposes several desirable properties that a concentration metric must have to be consider a consistent risk concentration measure such as *transfer principle* — the reduction of the exposure in a loan and the equal increment in another loan should not decrease the measure; *uniform distribution principle* — when all loans have same exposure, the measure should achieve its minimum value; *Lorenz-criterion* — when there are two portfolios (A, B) with the same number of loans M and the first m largest loans are added ($m \leq M$), if $Exposure(A_m) \geq Exposure(B_m)$ then $measure(A_m) \geq measure(B_m)$; *superadditivity* — the measure should not decrease when two or more loans are combined; *independence of loan quantity* — the measure should not decrease (increase) when the number of loans increases (decreases); *irrelevance of small exposures* — the measure should not increase significantly when a relatively small loan is issued.

One of these ad hoc methods is the *Herfindahl-Hirschman* Index (from here on abbreviated as the HHI), a measure that is widely used in the finance industry. Becker *et al.* (2004) prove that the HHI satisfies all the desirable properties of methods used for calculating concentration risk suggested by Lütkebohmert (2009).

The optimisation of credit portfolios thus presents a particular challenge as several risk metrics have to be considered. Some authors have sought to develop mathematical models capable of integrating concentration metrics into the VaR or CVaR. Lütkebohmert (2009 p 72-73) suggests using granularity adjustment, the semi asymptotic approach or saddle-point approximation methods for this purpose. Details of these methodologies are found in Gordy (2004), Emmer and Tasche (2004) and Gordy (2002), respectively.

However, such integrated models are difficult to apply to credit portfolios of commercial and retail lenders as their portfolios are composed of relatively illiquid assets, for which the data required by these models is not available. Also the complexity of these integrated models is a problem in itself. Orenstein (2011) points out several drawbacks of these models in the Citigroup failure case.

Hence, the research conducted here opts for preserving each metric as an objective within the optimisation process. Consequently, in order to allow this restriction, a multi-objective optimisation approach is adopted.

2.2.3 Multi-Objective Optimisation

Zopounidis and Doumpos (2002) suggested using multi-objective programming whereby each metric is preserved as a separate criterion (i.e. several metrics can be maximised or minimised simultaneously). A similar approach is developed by SchlottmannMitschele and Seese (2005) when they attempt combining different risk measures to model credit risk, market risk and operational risk in FIs. The main advantage of the multi-objective framework is the flexibility that it provides for including a selection of different metrics in the modelling process, without sacrificing the intuitiveness of each such metric.

Deb and Goel (2001) propose Multi-Objective Evolutionary Algorithms (MOEAs) for solving multi-objective optimisation problems. Deb (2008) surveys the different types of MOEAs. Schlottmann and Seese (2004) and BrankeScheckenbachSteinDeb and Schmeck (2009) provide two example studies applying MOEAs to credit and financial portfolio optimisation. Ponsich *et al.*

(2013) surveyed the use of MOEA in portfolio optimisation and more recently Mizgier and Pasia (2016) use MOEAs for capital allocation.

Even though MOEAs seem suitable for solving multi-objective optimisation problems, one of their major disadvantages is that they produce several solutions, which tend to be disperse (Deb and Goel, 2001); therefore, choosing a solution becomes a problem in itself for the decision makers. In that sense, there is a practical benefit considering approaches which allow narrowing down this solution space.

The convergence of the MOEA methods has been established by Rudolph and Agapie (2000), but the stability of the solutions they produce is not fully addressed yet in the literature – this, along with solution selection, are issues that will be further addressed in the proposed framework.

2.3 Credit portfolio modelling

A portfolio can be defined as a collection of assets that an investor can purchase using a predefined amount of money (Markowitz, 1952). When planning their portfolios, FIs tend to split the assets into segments or sectors rather than concentrate on individual assets. In line with this practice, a sectorial analysis is adopted, where the assets are grouped in sectors, rather than considering each asset individually. Furthermore, organising the portfolio into sectors enables an analysis of the risk due to common factors. These systemic factors may affect one sector of the portfolio only, or could affect several at the same time.

Some performance metrics of the portfolio can be calculated at sector level but others involve the performance of each asset separately. For that reason, two "levels" are introduced when this type of credit portfolio is studied: the asset level and the sector level. The following notation is used to represent these two levels in credit portfolios:

Let O_{ij} be the portion of money invested in asset j of sector i. The portion invested in sector i is given by:

$$x_i = \sum_{j=1}^{m_i} O_{ij} \qquad (2.1)$$

where m_i is the number of obligors in sector i.

Note that the assets in a credit portfolio are the contractual obligors. O_{ij} can be interpreted as portion of the portfolio for funding the obligor j of in sector i

Definition 3.1: A credit portfolio is defined by the vector $\boldsymbol{X} = [x_1, \dots, x_n]$, where x_i is the proportion of the portfolio invested in sector i; $x_i \in [0,1]$; $\sum_{i=1}^{n} x_i = 1$ and n is the number of sectors.

2.3.1 Credit portfolio performance metrics

As Figure 2.1 shows there are several interactions in a credit portfolio and so there are a variety of performance metrics. The performance of a credit portfolio can be summarised by two main elements: the returns given by the re-payments of the performing assets and the losses caused by the default of the obligors. The "default" event is the main trigger that could produce losses in the credit portfolio.

Definition 3.2: Let the random variable D represent the default of a loan; hence $D_{ij} = 1$ implies loan j in sector i defaults, or $D_{ij} = 0$ otherwise. The probability of default PD_{ij} is the associated probability of this event.

Definition 3.3: The risk of exposure or exposure at default (EAD_{ij}) is the book value of the loan j in sector i when the loan j defaults. The EAD could be a stochastic value, particularly when loans are credit lines or credit cards. The BCBS (2005 parg. 311-315 and parg. 474 - 478) establishes the conditions that FIs should follow to estimate EAD values. In this paper, the EADs are treated as deterministic values.

Definition 3.4: The severity of the loss (SEV_{ij}) is the portion of the value of the loans that is lost after the loan defaults. The severity of the loss is also stochastic and its expected value is the loss given default (LGD_{ij}) (Bluhm *et al.*, 2003). BCBS (2005 parg. 286 - 307) details the conditions to work out the LGDs in a credit portfolio.

Using these previous metrics and following logic proposed by BluhmOverbeck and Wagner (2003) the return for a particular loan can be modelled as:

$$Return(O_{ij}) = EAD_{ij} * R_{ij} * (1 - D_{ij}) - EAD_{ij} * SEV_{ij} * D_{ij} \tag{2.2}$$

where R_{ij} is the annual interest rate associated with a particular obligor, netted cost of funding and expenses.

Given that PD_{ij} is the probability of default of a particular obligor and LGD_{ij} its loss given default, the expected return can be expressed as:

$$EReturn(O_{ij}) = EAD_{ij} * R_{ij} * (1 - PD_{ij}) - EAD_{ij} * LGD_{ij} * PD_{ij} \tag{2.3}$$

Equation (2.3) assumes independence between the LGDs and PDs[8], although this assumption is questionable. The expected return of a credit portfolio is represented by:

$$EReturn(X) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} EAD_{ij} * R_{ij} * \left(1 - PD_{ij}\right) - EAD_{ij} * LGD_{ij} * PD_{ij} \tag{2.4}$$

The associated loss of a particular obligor is given by:

$$L\left(O_{ij}\right) = EAD_{ij} * SEV_{ij} * D_{ij} \tag{2.5}$$

Hence, the expected loss is:

$$EL\left(O_{ij}\right) = EAD_{ij} * LGD_{ij} * PD_{ij} \tag{2.6}$$

In the same way, the expected loss (EL) in a credit portfolio is given by:

$$EL(X) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} EAD_{ij} * LGD_{ij} * PD_{ij} \tag{2.7}$$

Since the EL is the loss likely to occur in a credit portfolio, an enhanced metric which captures the unexpected losses is needed. Comments and examples of the limitations of the EL as a risk measure are given in Thomas (2009).

The VaR, a commonly used risk metric for measuring unexpected losses, as mentioned in section 2.2.2, is the quantile l in the portfolio's distribution of losses, such that the probability of experiencing a larger loss than l is less than $1 - \alpha$, for a specified value of α ($0 < \alpha \leq 1$).

Let $\mathrm{L}(\boldsymbol{X})$ be the loss associated with portfolio $\boldsymbol{X}$. According to ArtznerDelbaenEber and Heath (1999), for a fixed $\alpha \in (0,1]$ the credit portfolio VaR at α level is defined as:

$$VaR(\boldsymbol{X})_{\alpha} = inf\{l \in \mathbb{R}: \Pr(L(\boldsymbol{X}) > l) \leq 1 - \alpha\} \tag{2.8}$$

where inf is the infimum operator.

The conditional VaR (CVaR) is defined as follows (Tasche, 2002):

[8] This assumption has been accepted in the industry. However, the impact of this assumption is still under scrutiny. A review of the implication of this assumption is found in Folpmers (2012).

$$CVaR(\boldsymbol{X})_\alpha = \frac{1}{(1-\alpha)}\int_\alpha^1 VaR(\boldsymbol{X})_u du \tag{2.9}$$

It should be highlighted that $CVaR(X)_\alpha$ is a particular case of spectral risk functions (see Härdle *et al.* (2017)). A spectral risk function can be defined as (Overbeck and Sokolova, 2017):

$$r_w(L(\boldsymbol{X})) = \int_0^1 w(u)VaR(\boldsymbol{X})_u du \tag{2.10}$$

where $w(u)$ is a weight function.

In particular, Overbeck and Sokolova (2017) show that the w for the CVaR is:

$$w(u)_{CVaR} = \frac{1}{(1-\alpha)}\ \mathbf{1}_{u>\alpha} \tag{2.11}$$

where $\mathbf{1}_{u>\alpha} = \begin{cases} 1 & \text{if } u>\alpha \\ 0 & \text{otherwise} \end{cases}$

Taking into consideration the different risk metrics for modelling the unexpecting losses the concept of Economic Capital (EC) is introduced as follows: The EC is the amount of capital that the FI estimates that needs to cover its potential portfolio losses. It is important to contrast this definition with regulatory capital which is the minimum capital requirement set by the regulatory authorities (Thomas, 2009). In this work, the EC is represented by:

$$EC_\alpha(\boldsymbol{X}) = UL_\alpha(\boldsymbol{X}) - EL(\boldsymbol{X}) \tag{2.12}$$

where $UL_\alpha(\boldsymbol{X})$ is the unexpected loss at α level of the portfolio $\boldsymbol{X}$.

Let $UL_\alpha(\boldsymbol{X}) = r_w(L(\boldsymbol{X}))$ depending on the chosen metric. Considering their respective properties discussed earlier in section 2.2.2 and the fact that the CVaR is commonly used in the industry, this study chooses to use $UL_\alpha(\boldsymbol{X}) = CVaR(\boldsymbol{X})_\alpha$ and $\alpha = 99\%$.

In the presented modelling process, the HHI is used to measure concentration risk between the sectors. The HHI is computed as follows:

$$HHI(X) = \sum_{i=1}^{n}(x_i)^2 \tag{2.13}$$

Hence a portfolio concentrated in a few sectors will produce a HHI near to 1 and a more evenly balanced portfolio will produce a HHI near to 1/n.

2.3.2 Credit portfolio optimisation problem

Previous sections have introduced the main goals of the credit portfolio optimisation problem (CPOP), i.e. expected return (representing the profitability perspective) and the economic capital and the concentration risk (representing the risk perspective). Also, the decision variables are given by the vector X which represents the portion invested in each sector. Without losing generality it can be assumed that the whole capital is invested.

Additional restrictions have to be considered to avoid negative values in the decision variables (i.e. not short selling) and the maximum investment that the FI is willing to have in each sector. Specifically, the maximum investment in each particular sector will correspond to the market saturation in that sector[9].

Therefore, the multi-objective formulation of the CPOP can be summarised as follows:

Decision variable:

X: The portion invested in each sector

Objective functions:

Profitability perspective:

$$minimise\ f_1 = -EReturn(X) \tag{2.14}$$

Risk perspective:

$$minimise\ f_2 = EC_\alpha(X) \tag{2.15}$$

$$minimise\ f_3 = HHI(X) \tag{2.16}$$

Restriction: All the budget should be invested

[9] FIs normally have estimations of this saturation given by market share studies. In consequence, it is reasonable to assume that there is a celling in each sector. If this information is not available, we can always set this maximum value as $U_i = 1$.

Subject to:

$$\sum_{i=1}^{n} x_i = 1 \tag{2.17}$$

Restriction: Market conditions in each sector:

$$L_i \leq x_i \leq U_i \tag{2.18}$$

where $L_i \geq 0$ represents the no short sell allowance restriction

and $U_i \leq 1$ is the sector maximum investment.

The CPOP has the following assumptions:

1. The proportion of O_{ij} in a sector remains fixed over the planning horizon.
2. PD_{ij}, LGD_{ij} are the long run average PD and downturn LGD and remain constant during the planning horizon. This assumption is in line with the guidance proposed by the Basel accords for capital calculations.
3. Return rates R_i remain constant during a planning horizon.
4. Each obligor can only belong to one sector.

2.3.3 Methods to compute the solutions

Generating solutions that satisfies the CPOP can be challenging as the computation of the Economical Capital, $EC_\alpha(X)$, is not straightforward. The remainder of this section introduces a numerical approach for computing the economic capital.

2.3.3.1 Methods to compute economic capital

In the case of credit portfolios, the assumption of normality in the distribution of losses can be misleading as the correlation between defaults can produce higher losses. Thus, the distribution of losses is characterised by a large kurtosis (fat tails) (see Cespedes (2002)).

The financial industry has developed different methods to approximate the distribution of the losses in a credit portfolio. Jorion (2009) surveyed the most popular approaches. One such method is CreditRisk+(CSFB, 1997). This approach considers separately the distributions of defaults and the severity of the losses to build the distribution function of losses. The distribution of the number of defaults is modelled using a Poisson distribution, based on the assumption that the individual PDs are uniformly small enough[10] and independent among the obligors. CreditRisk+ splits the portfolio in sectors where the obligors may have some systematic risk factors in common. Initially, in CSFB (1997), the sectors are considered independent. BürgisserKuthWagner and Wolf (1999), Han and Kang (2008) and Fisher and Dietz (2011) present improved versions of CreditRisk+ in which it is possible to use correlated sectors.

CreditRisk+ has several useful features. It produces deterministic solutions; consequently, when the unexpected loses are estimated, volatility is avoided. The method is well documented and explained in the literature and there are several implementations and surveys of it. Gundlach and Lehrbass (2004) present a compilation of the major enhancements of CreditRisk+.

CreditRisk+ is used in this study as a way to illustrate how the framework operates. However, it is important to highlight that the framework is able to implement any method that can estimate unexpected losses.

2.3.3.2 Methods to solve the CPOP

The CPOP is modelled using a Multi-Objective Optimisation Problem approach. In general, solutions from these approaches are characterised by the following properties. First, the objectives are conflicting; for example, better returns usually imply higher risks. Second, solutions can dominate one another. A solution Y is dominated by a solution X, if X performs similarly to Y in almost all the objectives but Y is better in at least one of the objectives. In contrast, X is called non-dominated if there is not a solution Y such that Y dominates X. Third, there is more than one possible non-dominated optimal solution. Finally, all non-dominated solutions are located in a set denominated the Pareto-Optimal Front (also known in the literature as Efficient Front) (Markowitz, 1952). Figure 2.2 illustrates these characteristics.

[10] Usually no more than 10% to fulfil the assumption $\ln(1 + PD) = PD$, see CSFB (1997 p 34).

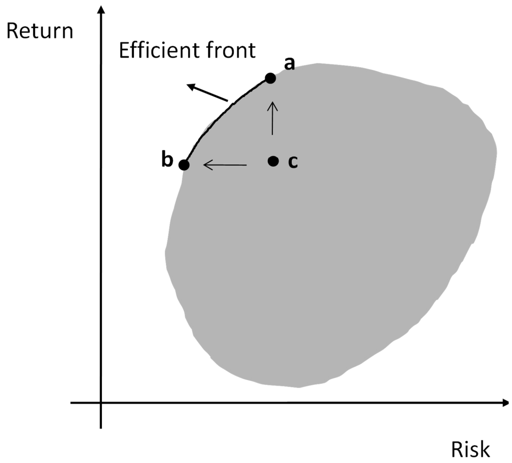


Figure 2.2 Efficient Front in MOOP

Note how in Figure 2.2 solution *c* is dominated by solutions *a* and *b*.

For the CPOP the dominancy criteria can be defined as follows. Let X and Y be solutions of the CPOP. Then X dominates Y if at least one of the following conditions is true:

$$EReturn(X) > EReturn(Y)\ and\ EC_\alpha(X) \le EC_\alpha(Y)\ and\ HHI(X) \le HHI(Y)$$

$$EReturn(X) \ge EReturn(Y)\ and\ EC_\alpha(X) < EC_\alpha(Y)\ and\ HHI(X) \le HHI(Y)$$

$$EReturn(X) \ge EReturn(Y)\ and\ EC_\alpha(X) \le EC_\alpha(Y)\ and\ HHI(X) < HHI(Y)$$

Generic optimisation methods (also called meta-heuristics) are suitable for solving the CPOP, as the complexity involved in the calculations of the objective functions especially in the EC, rule out other, more conventional methods. Moreno-Paredes (2016) showed how conventional methods such as Conjugate Gradient fail to solve such problems.

The Multi-Objective Evolutionary Algorithm (MOEA), one of these meta-heuristics, has been used successfully to solve similar problems (Schlottmann and Seese, 2004; BrankeScheckenbachSteinDeb and Schmeck, 2009; Mizgier and Pasia, 2016).

2.3.3.3 Solutions selection for the CPOP

One of the major drawbacks of MOEAs is the fact that they produce multiple solutions for the optimisation problem. In order to narrow down the set of solutions, a modified version of the ‘global criteria method’ (GCM), described by Deb (2008), is developed in this thesis. This approach is named as *normalised GCM* (NGCM) and it is outlined below:

I. Create an artificial vector $\boldsymbol{F}$ with the objective values by selecting the minimum of each objective from the set of solutions obtained when the MOEA is run ($\boldsymbol{F}$ is referred as objective vector in this paper).

$$\boldsymbol{F}: F_k = \min_{s \in S}(f_s{}^k) \tag{2.19}$$

where S is the set of solutions from MOEA, and

$f_s{}^k$ is the k-th objective value from the solution s. In this case, $k = 1,2,3$.

II. Calculate the normalised distance CD_s between the objective vector, $\boldsymbol{F}_s$, of each MOEA solution and $\boldsymbol{F}$. CD_s is named as Centroid Distance.

$$CD_s = \frac{\|\boldsymbol{F}_s - \boldsymbol{F}\|}{\|\boldsymbol{F}\|} \tag{2.20}$$

where $\| \quad \|$ is the Euclidean norm.

III. Select the solution $s^* \in S$ whose objective vector $\boldsymbol{F}_{s^*}$ has the shortest distance to $\boldsymbol{F}$. In other words, undertaking this process will ensure that the best possible outcome is selected (i.e. the one with objective values that are closest to the ideal $\boldsymbol{F}$). The selection process is represented as follows:

$$\boldsymbol{F}_{s^*}: \text{where } CD_{s^*} \leq CD_s, \forall\, s \in S \tag{2.21}$$

where $s^* \in S$.

When there are two more solutions that fulfil this condition, one of them is randomly selected.

2.3.3.4 Computing the dispersion of the objectives

Dispersion of the objective vectors is computed by averaging the centroid distance between the objective vectors and the ideal objective vector $\boldsymbol{F}$ as follows:

$$\overline{CD} = \frac{1}{|S|} \sum_{s \in S} CD_s \tag{2.22}$$

where:

CD_s is the centroid distance from the solution s,

$|S|$ is the cardinality of S.

The average of the centroid distance is used to determine if the proposed selection method produces more stable solutions, i.e. solutions having less dispersed objective vectors.

2.3.4 Data availability

In order to implement this framework, a FI must have the following set of data for each obligor j: EAD_{ij}, PD_{ij},, LGD_{ij}, R_{ij}, the sector-by-sector Pearson correlation matrix of yearly default rates $\left\{\rho_{i_y i_z}\right\}$ over a certain time horizon[11] and the PD volatilities in each sector σ_i. Those are the inputs required by CreditRisk+ to compute the economic capital. Additionally, the market conditions $L_i \geq 0$ and $U_i \leq 1$, i.e. the minimum and maximum sector investment to be considered.

2.4 Framework outline

The proposed framework, for improving credit portfolios of commercial and retail FIs, can be summarised in these steps:

1) The following inputs, from an initial credit portfolio, are assumed: PD, LGD, correlation matrix of the portfolio's sectors, yearly PD's volatilities and market conditions.
2) The objective values, i.e. Economic Capital, Returns and HHI are computed for the initial portfolio.
3) With the data from steps 1 and 2 the CPOP model in section 2.3.2 is populated.
4) A MOEA is used for finding solutions that solve the CPOP.
5) Solutions are selected using the NGCM.
6) Selected solutions are reported.

And outline of the proposed framework in this paper is presented in Figure 2.3.

[11] The time horizon should be broad enough to capture the different economy cycles according to Thomas, et al. (2002).

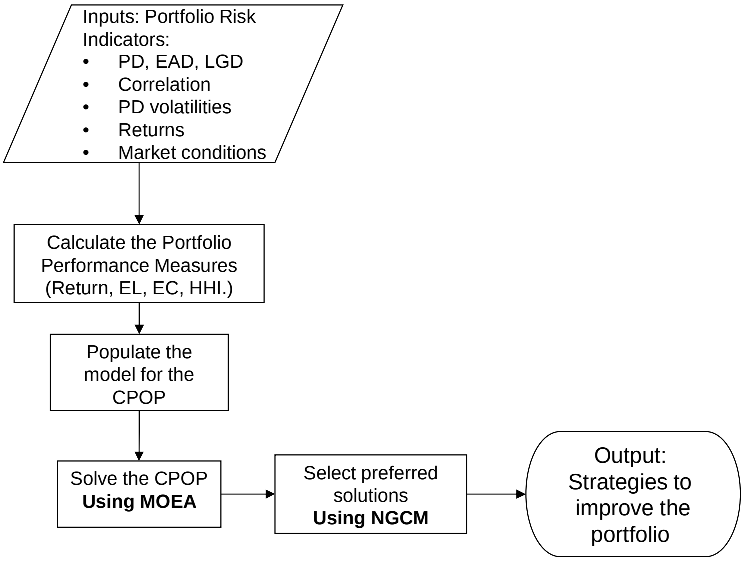


Figure 2.3 Outline of the proposed framework to optimise credit portfolios

2.5 Case Study

In order to illustrate how the proposed framework operates, a case study, based on the data of a US commercial and retail bank, is used. A summary of the information contained in this database is presented in Table 2.1.

Item	**Description**
Total Assets	US$ 2.8 Billion
Total Obligors	2557 exposures
Number of Sectors	16

Table 2.1 Portfolio description

The assets of the credit portfolio are loans to individuals and companies (obligors). For each obligor in the portfolio, values of PD, LGD and EAD are available. The portfolio is categorised into 16 different sectors. Figure 2.4 shows how the portfolio is distributed among the sectors (note that for commercial confidentiality reasons, the sectors are identified by numbers rather than text descriptions).

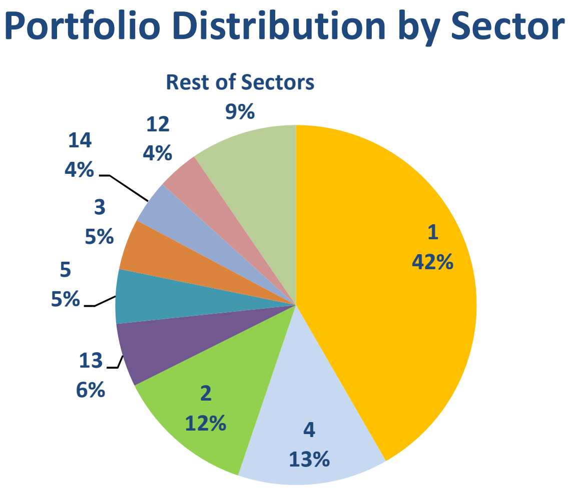


Figure 2.4 Distribution of the portfolio by sector.

The sectors within the portfolio are not independent and the correlation matrix of defaults between the sectors is presented in Table 2.2.

	Sector 1	Sector 2	Sector 3	Sector 4	Sector 5	Sector 6	Sector 7	Sector 8	Sector 9	Sector 10	Sector 11	Sector 12	Sector 13	Sector 14	Sector 15	Sector 16
Sector 1	1.00	0.41	0.54	0.84	-0.05	0.25	0.41	0.96	0.12	0.61	0.19	0.09	0.28	0.27	0.13	0.01
Sector 2	0.41	1.00	0.34	0.41	-0.16	0.10	0.99	0.39	0.82	0.00	0.07	0.09	-0.19	-0.04	0.12	0.30
Sector 3	0.54	0.34	1.00	0.42	0.10	0.53	0.34	0.60	-0.02	0.34	0.33	-0.05	0.34	0.41	0.42	0.16
Sector 4	0.84	0.41	0.42	1.00	-0.01	0.02	0.41	0.83	0.25	0.42	0.33	-0.01	0.45	-0.08	-0.12	-0.15
Sector 5	-0.05	-0.16	0.10	-0.01	1.00	0.40	-0.16	-0.05	-0.08	0.01	0.47	-0.16	0.45	-0.15	-0.25	-0.38
Sector 6	0.25	0.10	0.53	0.02	0.40	1.00	0.10	0.24	-0.06	0.35	0.31	-0.21	0.02	0.15	0.19	-0.04
Sector 7	0.41	0.99	0.34	0.41	-0.16	0.10	1.00	0.39	0.82	0.00	0.07	0.09	-0.19	-0.04	0.12	0.30
Sector 8	0.96	0.39	0.60	0.83	-0.05	0.24	0.39	1.00	0.02	0.51	0.18	-0.07	0.27	0.28	0.19	0.04
Sector 9	0.12	0.82	-0.02	0.25	-0.08	-0.06	0.82	0.02	1.00	-0.10	0.11	0.15	-0.17	-0.33	-0.17	-0.01
Sector 10	0.61	0.00	0.34	0.42	0.01	0.35	0.00	0.51	-0.10	1.00	-0.09	0.11	0.21	-0.04	-0.23	-0.45
Sector 11	0.19	0.07	0.33	0.33	0.47	0.31	0.07	0.18	0.11	-0.09	1.00	-0.19	0.67	-0.03	0.09	-0.10
Sector 12	0.09	0.09	-0.05	-0.01	-0.16	-0.21	0.09	-0.07	0.15	0.11	-0.19	1.00	0.16	0.26	0.03	0.40
Sector 13	0.28	-0.19	0.34	0.45	0.45	0.02	-0.19	0.27	-0.17	0.21	0.67	0.16	1.00	-0.06	-0.20	-0.21
Sector 14	0.27	-0.04	0.41	-0.08	-0.15	0.15	-0.04	0.28	-0.33	-0.04	-0.03	0.26	-0.06	1.00	0.79	0.60
Sector 15	0.13	0.12	0.42	-0.12	-0.25	0.19	0.12	0.19	-0.17	-0.23	0.09	0.03	-0.20	0.79	1.00	0.57
Sector 16	0.01	0.30	0.16	-0.15	-0.38	-0.04	0.30	0.04	-0.01	-0.45	-0.10	0.40	-0.21	0.60	0.57	1.00

Table 2.2 Correlation matrix between sectors.

The volatility of defaults within the sectors is defined as the standard deviation of the annual number of defaults in each sector divided by the average number of annual defaults in each sector. Similarly, the returns of the portfolio are expressed by the annual interest rates less the cost of funding, commissions and fees incurred. In this case study the bank has used five years of data to compute these figures.

The volatilities and the returns of each sector of the portfolio are given in Table 2.3.

Sector	Volatilities	Returns
1	0.22	1.91%
2	0.17	0.41%
3	0.33	-0.72%
4	0.22	0.29%

Sector	Volatilities	Returns
5	0.06	-1.77%
6	0.15	0.15%
7	0.17	0.39%
8	0.35	0.06%
9	0.17	-0.83%
10	0.41	-1.46%
11	0.89	0.11%
12	0.73	0.87%
13	0.32	0.78%
14	0.21	-0.83%
15	0.36	0.12%
16	0.25	-0.34%

Table 2.3 Volatilities and annual returns by sector

The market conditions (i.e. the minimum and maximum investment that the FI is willing to have in each sector) are presented in Table 2.4.

Sector	L	U
1	1.7E-05%	100%
2	4.6E-05%	100%
3	2.1E-06%	100%
4	7.3E-06%	100%
5	1.3E-06%	100%
6	3.8E-06%	100%
7	1.4E-06%	100%
8	3.6E-08%	100%
9	5.8E-07%	100%
10	2.2E-07%	100%
11	7.2E-08%	100%
12	1.1E-05%	100%
13	7.6E-07%	100%
14	4.0E-07%	100%
15	4.0E-07%	100%
16	4.7E-07%	100%

Table 2.4 Market conditions[12]

In order to compute the CVaR, the CreditRisk+ version of HaafRieß and Schoenmakers (2004) is used in combination with the BürgisserKuthWagner and Wolf (1999) approach. Integrating both versions gives a numerically stable version of CreditRisk+ which also allows calculating CVaR and dealing with correlated sectors.

The associated CPOP is solved using the Non-Dominated Sorting Genetic Algorithm type II (NSGA II), a type of MOEA developed by Deb and Goel (2001). The following initial settings and stop

[12] Market conditions are based on internal estimation produced by the strategic department of the bank.

conditions were applied, using the same initialisation suggested by Dinovella and Moreno-Paredes (2005):

Parameter	Value
Pc: Probability of Crossover	0.9
Pm: Probability of mutation	0.9
Max number of generation	500
Initial Population	Random
Number of max generation	500
Number of runs	145

Table 2.5 MOEA initialisation parameters

The numerical representation and genetic procedures in the NSGA II were implemented on Matlab® ver 10.0 using the approaches in Moreno-ParedesMues and Thomas (2013) and Dinovella and Moreno-Paredes (2005).

The optimisation process is run 145 times and at the end of each run a random solution from the final non-dominated front is taken. The solutions are graphically presented in Figure 2.5, where the x and y axis depict CVaR and return, respectively, whilst the size of each dot represents the HHI of each solution portfolio.

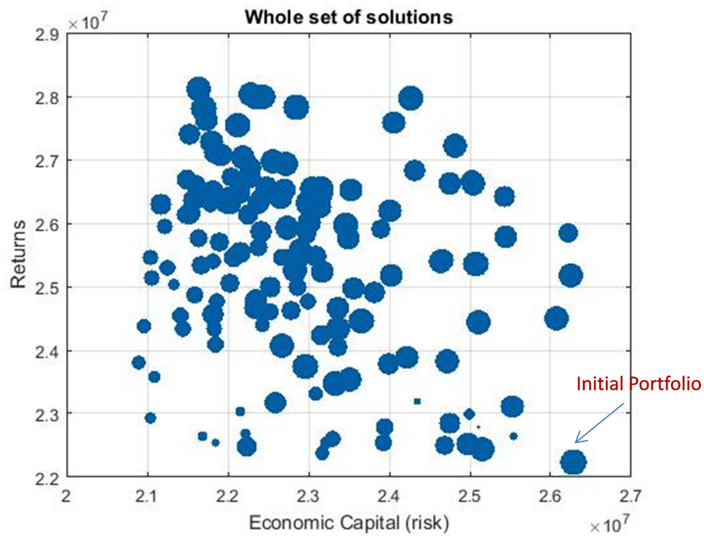


Figure 2.5 Solutions from the MOEA

On average, the solutions obtained from the MOEA improve the performance of the credit portfolio, achieving higher returns with lower economic capital and lower concentration as shown in Table 2.6.

	Initial Portfolio	Average across solutions	Improvement (%)	Std. deviation of solutions
Expected Return (MM US$)	$22.23	$25.27	13.7%	$1.51
Economic Capital via CVaR (MM US$)	$26.28	$22.89	-12.9%	$1.28
HHI	0.2202	0.1993	-9.5%	0.01

Table 2.6 Summary results

For selecting a solution, the NGCM (explained in section 2.3.3.3) is applied. The associated objective vector of the selected solution is presented below.

	Initial Portfolio	Selected solution via NGCM	Improvement (%)
Expected Return (MM US$)	$22.23	$25.03	12.6%
Economic Capital via CVaR (MM US$)	$26.28	$21.33	-18.8%
HHI	0.2202	0.1758	-20.2%

Table 2.7 Objective values of selected solution

Solutions from the optimisation process produce diversification of the assets in the portfolio. For example, Figure 2.6 shows a comparison of the sectorial distribution between the initial portfolio and the selected one via NGCM. It should be noticed that the portfolio associated with the selected solution exhibits a more even sectorial distribution, mainly driven by the significant reduction in sector 1 and increments in sectors 12 and 13.

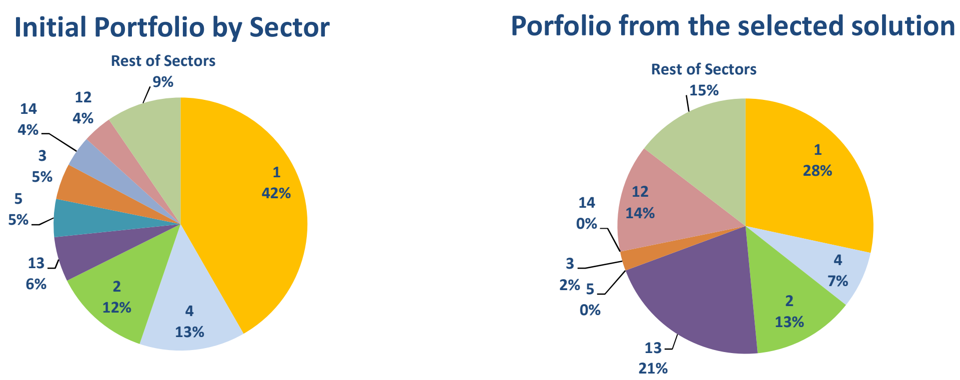


Figure 2.6 Comparison of portfolios

2.5.1 Testing the stability of the framework

In order to test the stability of the solutions generating using this framework, the original set of solutions is divided in H subsets. In this case, $H = 7$, with six subset of 20 solutions and one of 25; such selection ensures enough variability to analyse the dispersion of the solutions' objective vectors. These subsets are randomly configured; this can be achieved without losing generality as each run is independent of one another.

For each subset, a preferred solution is selected via the normalised GCM. Figure 2.7 depicts each of the preferred objective vectors in pink; the blue dots represent the other solutions. The initial portfolio is also included as reference. Similarly than Figure 2.5, the x and y axis depict CVaR and return, respectively, whilst the size of each dot represents the HHI of each solution portfolio.

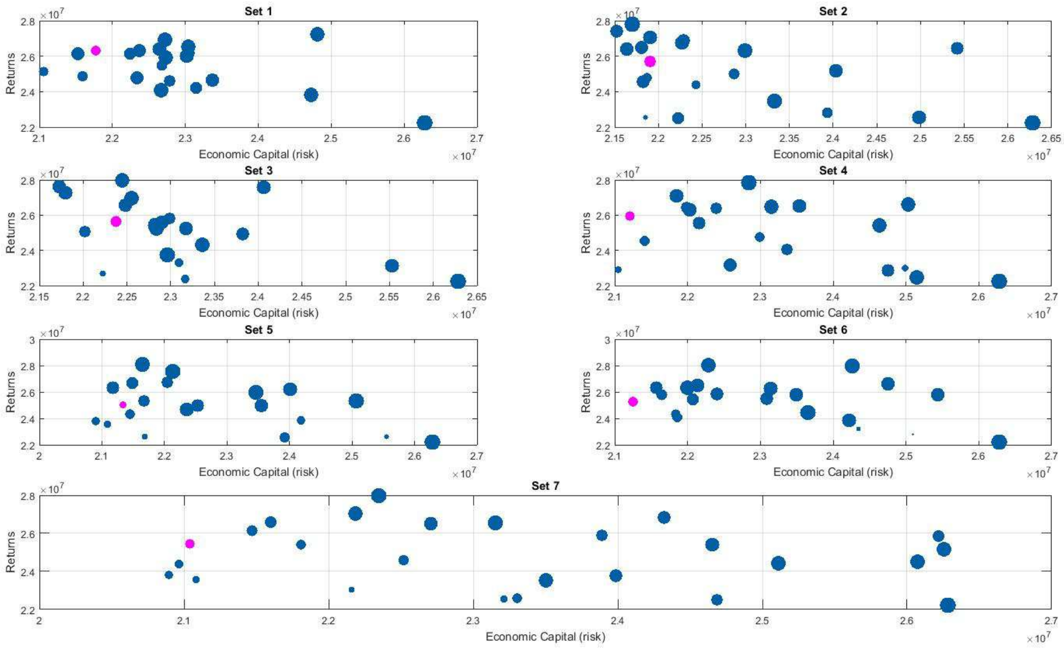


Figure 2.7 Preferred objective vectors in each subset

Figure 2.8 shows all the solutions from all the subsets. The preferred solutions are indicated in pink colour. Similarly than Figure 2.5, the x and y axis depict CVaR and return, respectively, whilst the size of each dot represents the HHI of each solution portfolio.

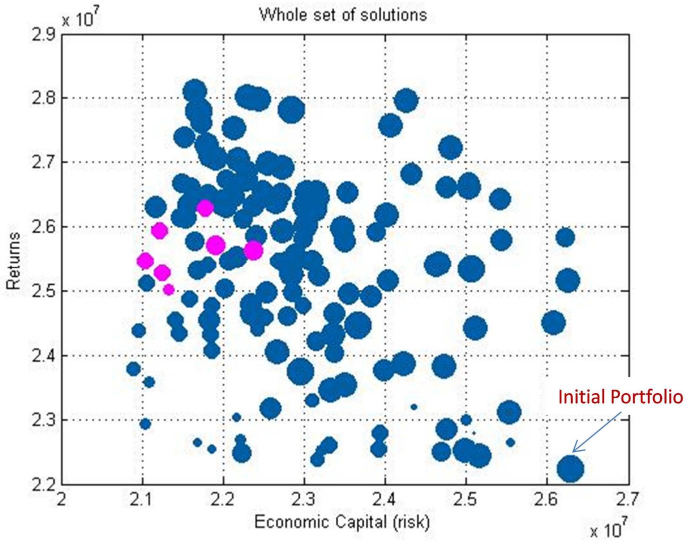


Figure 2.8 Preferred solutions.

Table 2.8 shows the improvement in the portfolio metrics using the objective values of the preferred solution subset (P-subset)

	Initial Portfolio	Average of P-subset	Improvement (%)	Std. deviation of P-subset
Expected Return (MM US$)	$ 22.23	$ 25.62	15.3%	$ 0.39
Economic Capital via CVaR (MM US$)	$ 26.28	$ 21.55	-18.0%	$ 0.44
HHI	0.2202	0.1861	-15.5%	0.005

Table 2.8 Summary results from the preferred front

In order to perform a comparison between the dispersion in the solutions, two subsets are considered: P-subset and the rest of solutions (R-subset).

The Centroid Distance (CD) is computed for each objective vector in the P and R subsets. Subsequently, the average of the CD for P and R is calculated. Figure 2.9 shows that dispersion in P-subset is less than in R-subset; implying that solutions selected via NGCM tend to be less disperse and consequently more stable. Similarly than Figure 2.5, the x and y axis depict CVaR and return, respectively, whilst the size of each dot represents the HHI of each solution portfolio.

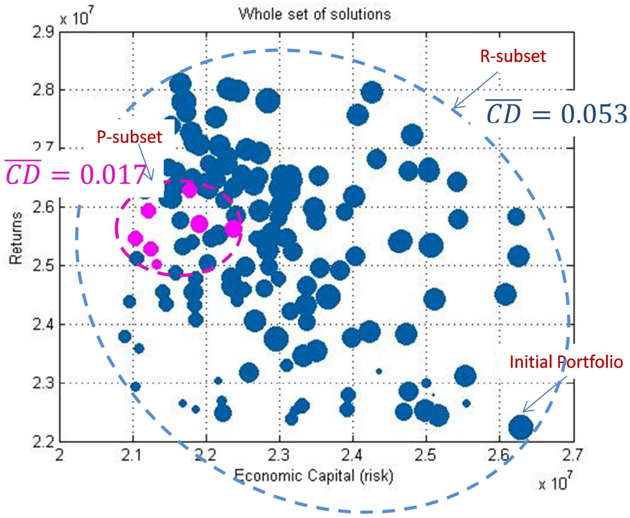


Figure 2.9 Solution's dispersion.

2.6 Discussion

2.6.1 Framework for managing commercial and retail credit portfolios

The proposed framework extends the literature on the optimisation of credit portfolios by considering commercial and retail lending, which is normally made up mostly of non-liquid assets such as private loans. This characteristic makes it difficult to apply previous credit portfolio optimisation frameworks developed in the literature such as AnderssonMausserRosen and Uryasev (2001), LuHuangChing and Siu (2013) and Bo and Capponi (2016), as they assume that the quantity of assets in the portfolio can be easily increased or decreased; in other words, they require that assets can be traded in an exchange market. In consequence, the optimisation is performed assuming a short term planning horizon. Conversely, retail and commercial lenders usually plan their portfolio with a much longer horizon, given the low liquidity of their assets. The framework proposed in this paper demonstrates that it is possible to perform optimisation of this type of portfolios by shifting the decision process from individual assets to sectors, addressing in that way the liquidity issue. Additionally, by using long term risk metrics such as long run average PDs and downturn LGDs, in line with the Basel accords, it is possible to develop longer term plans and strategies for managing this type of portfolios.

Although Schlottmann and Seese (2004) focus on the optimisation of commercial and retail lenders' portfolios, their approach is limited to the assumption that the loans in the portfolio belong to a single sector. By introducing a sector-level analysis and considering the default correlation of the sectors, the proposed approach shows a more realistic view.

2.6.2 Multi-objective optimisation and solution selection

By performing multi-objective optimisation via MOEA, it is shown that the proposed approach produces investment strategies that reshape and improve the portfolio performance in terms of risk and return. This produces a more evenly distributed portfolio and on average improves its return and reduces its economic capital and concentration index by 13.7%, 12.9% and 9.5%, respectively. This result seems to be aligned with the main advantages of MOEA described in the literature (Zopounidis and Doumpos, 2002; Schlottmann and Seese, 2004; SchlottmannMitschele and Seese, 2005). It also suggests that the framework is suitable for solving credit portfolio optimisation problems and it can be a useful methodology for long term scenario planning.

However, as also pointed out in the literature (Deb and Goel, 2001; Deb, 2008; BrankeScheckenbachSteinDeb and Schmeck, 2009; PonsichJaimes and Coello, 2013; Mizgier and Pasia, 2016), MOEAs produce multiple disperse outcomes, making the solution selection a problem in itself. This important drawback of the MOEAs is not fully addressed in previous applications to credit portfolios (Zopounidis and Doumpos, 2002; Schlottmann and Seese, 2004; SchlottmannMitschele and Seese, 2005). In the proposed approach, this obstacle is addressed by introducing the NGCM procedure for choosing preferred solutions. This procedure produces a substantial reduction in the dispersion of the selected solutions, moving the standard deviation of returns from 1.51 to 0.39 (MM US$), economic capital from 1.28 to 0.44 (MM US$) and the concentration index 0.01 to 0.005. Furthermore, the dispersion analysis (see Figure 2.9) suggests that the preferred objective vectors, obtained by the NGCM procedure, are less dispersed than the others, confirming that the NGCM approach produces solutions with more stable objective values.

2.7 Conclusions

This paper develops a framework to assess and improve credit portfolios of commercial and retail lenders. The framework proposes a sectorial investment optimisation strategy rather than asset level optimisation to deal with the specific nature of these portfolios.

The framework uses a multi-objective model which can combine both profitability and risk measures and supports the credit portfolio optimisation by finding effective and efficient investment strategies in sectors, i.e. strategies that not only mitigate possible losses but also convey opportunities to increase the return of a credit portfolio.

The approach that this paper presents is comprised of five major components:

1. Credit portfolio measures: a set of mathematical formulas are proposed to assess the performance of credit portfolios.
2. Computational methods: a methodology is proposed to compute the associated risk of unexpected losses in credit portfolios.
3. Optimisation model: a multi-objective optimisation programme is defined to integrate the different performance and risk measures.
4. Optimisation solver: A Multi-Objective Evolutionary Algorithm (MOEA) is used as an alternative heuristic method for solving the optimisation model.
5. Solution selection: the NGCM approach, a normalised version of the "global criteria method" (Deb, 2008), is used for selecting non-dispersed solutions from the MOEA.

Also in this work, a real life portfolio belonging to a lending institution is used to demonstrate how the framework operates. Particularly, this illustration displays the search and selection process of potential strategies (i.e. increasing and decreasing investment in the different sectors). This process ends up finding solutions that will improve the portfolio's quality in term of return, capital and concentration risk.

Another important characteristic of this framework is its flexibility. The multi-objective approach makes it easy to add more objectives and constraints. Furthermore, the NGCM approach, developed in this research, can also be applicable in other domains where multi-objective optimisation is performed.

One extension of this work is therefore, the incorporation of other risk and portfolio performing measures such as contagion. However, it is not always easy to measure contagion unless interbank data is available.

Another potential extension of this research could be the development of an enhanced framework that considers two levels of optimisation: In the first stage an optimisation is performed over the sectors level and in the second stage the optimisation is performed at exposure level. In other words, if the optimisation, conducted over the sector level, suggests a reduction in a sector, the second level optimisation will indicate which exposure should be reduced.

Finally, the framework shifts the discussion from individual assets to sectors. This new point of view is more suitable for long term strategic decision making, as retail and commercial specialised lenders usually manage their credit portfolios by sectors rather than individual loans, as proposed in previous papers.

Chapter 3: Predicting investor's success in the spread trading market: Case study from UK investors

Abstract

Developing effective data mining-based forecasting approaches for a new dataset often presents significant challenges for researchers and practitioners, such as dealing with outliers, overrepresentation of one of the classes and variable selection, particularly when dealing with large scale data from individual investors in the financial market. Gradient boosting machines (GBM) offer the prospect of overcoming some of these issues. The research conducted in this paper demonstrates the ability of GBMs for producing reliable forecasts, to identify those individual investors in a financial market posing most risk to a retail broker. The research also demonstrates how these forecasts can lead to a better understanding of the individual investor's behaviour.

Key words: Data mining, gradient boosting machines, investor profiling, spread trading

3.1 Introduction

Many forecasting tasks, particularly in the financial sector, involve classification. For example, forecasting if a market will rise or fall, whether an individual is likely to become a rogue trader or whether a trader's trading activity is likely to pose a risk to a broker's profitability. Increasingly, data mining techniques are applied in such circumstances. However, data mining-based classification (DMC) is particularly challenging when dealing with high dimensionality (i.e. several variables) and when the raw data on which the process depends contains many errors or omissions. The challenge arises because the majority of the DMC techniques require significant data pre-processing to avoid misleading results (Azoff, 1994; Fayyad *et al.*, 1996; Berry and Linoff, 1997; Kotsiantis *et al.*, 2006; Tan, 2006; Hall *et al.*, 2009; Han *et al.*, 2011).

In addition, commonly employed DMC techniques, such as General Linear Models (GLM), have several limitations. These limitations include their sensitivity to outliers, the strong assumption of non-correlation between the independent variables, overrepresentation, (i.e. in the data set one class is more frequent than the other) and the requirement of a linear or monotonic relationship between the independent and dependent variables. These limitations may prevent models from adequately assessing the influence of an individual's characteristics over investor's performance, leading to poor or misleading classification and hence inaccurate forecasts.

The research conducted in this paper demonstrates how Gradient Boosting Machines (GBM), can help to overcome many of the limitations of other DMC techniques, when producing accurate forecasts in an environment which is challenging for DMC techniques. Behavioural data from individual investors operating in the Spread Trading Market is used. In particular, behavioural data from individual investors tend to be highly disperse due the different types of investors. For example, they have different demographic background, risk attitude, different skills when operating and some of them are more active than others; posing several trades during the day. Therefore, profiling a particular type of investors could be a true challenge for DMC techniques.

GBM approaches, based on ensembles of basic classifiers, were first developed by Friedman (2001) and Friedman (2002), and has been employed in a variety of domains, including, profiling potential defaulted loans (Fitzpatrick and Mues, 2016), studying properties of wine grapes (Brillante *et al.*, 2015) and in determining the characteristics of soils (Nussbaum *et al.*, 2015). The research conducted in this second paper of this thesis also demonstrates the practical advantages of applying GBMs in a financial market when profiling risk investors. It is explored here how this approach can significantly improve practitioners' understanding about the impact of investors' characteristics and behaviour over their future performance, and how this can lead to greater confidence in the model's forecasts.

To achieve the research objectives of this second paper, a GBM approach is developed to profile individual financial investors. Specifically, this paper attempts to identify the characteristics of traders that enables to forecast effectively those who are likely to be particularly successful and therefore may pose most risk to those who underwrite their trades. The aim is to demonstrate the value of GBMs in dealing with asymmetric populations where the model involves high dimensionality (i.e. many independent variables), and where the raw data contains many errors or omissions, without investing considerable time and resources in data pre-processing. This research benchmarks the performance of GBMs against the most common DMC techniques such as logistic regression (LR), classification trees (CT), neural networks (NN) and support vector machine (SVM). The aim is to demonstrate that GBM provides a more robust forecasting approach, providing greater understanding of those traders' characteristics and behaviours that enable to forecast effectively the risk they pose to those who underwrite their trades.

The remainder of the paper is organised as follows: Section 3.2 provides an overview of the spread trading market and the nature and importance of the classification problem explored in this second paper. Section 3.3 outlines the characteristics of traders that previous research has found can help predict their future performance and the methods that have traditionally been employed to model investors' behaviour. The data is described in the section 3.4. The methodologies employed to classify traders, in terms of the risk they pose, is described in section 3.5. The results are presented in section 3.6 and discussed them and their implications in section 3.7. Finally, section 3.8 draws some conclusions.

3.2 The spread trading market classification problem

Market speculation, involving, for example, the foreign exchange market (FX), and spread trading, has significantly increased during recent years. For instance, the Bank for International Settlements (BIS) disclosed that the FX and over the counter markets (OTC)[13] reached a daily volume of US$ 5.1 trillion in 2016 (BIS, 2016), a 30% increase compared with 2010.

Spread trading, in particular, involves trading on the direction of movement of a financial instrument and can involve buy or sell trades. These, respectively, involve betting a given 'stake per point', that the value of a financial instrument will increase or decrease. Should the financial instrument move *n* points in the direction anticipated by the trader, then they secure a gain of £*n* x stake per point. However, if the financial instrument moves *n* points in the opposite direction, they lose £*n* x stake per point. Further details are given in the Appendix A.

Paton and Williams (2005) argue that the popularity of the spread trading market, derives from its low transactional costs and the favourable tax treatment of gains from spread trading activity. Consequently, it attracts small traders with wealth motivations and larger traders executing risk management strategies (e.g., as a hedge against a fall in the market in the short term).

Investors in spread trading markets place their trades through retail brokers. The success of these firms requires that they develop an appropriate means of managing the risk and this is relevant for the health of financial markets. In particular, these brokers hedge their positions in the underlying markets and any overreaction in their hedging activity could introduce market volatility. Equally, if they fail to effectively manage their risk they may go bankrupt, undermining confidence in financial markets. For example, in January of 2015, several brokers faced bankruptcy after a substantial volatility in the value of the Swiss franc, causing market instability (FT, 2015).

[13] OTC stands for over the counter markets, i.e. instruments that are traded out of the stock exchange.

Retail brokers make their profit from the commission they charge to traders for each transaction (the differential between the buy and sell prices of the instruments that they offer). They manage their risk by hedging the trades of those investors who they believe are more informed (i.e. 'A-book traders' who are likely to secure profits from their trading activity) with other specialised companies (liquidity providers). They aim is to retain the risk of trades from those who are less likely to be successful ('B-book traders). The brokers, therefore, secure further profits from underwriting these trades. The ability to forecast which traders pose the greatest (A-book) and the lowest risk (B-book) at different times is, therefore, vital for effectively managing risk and profit. Importantly, spread traders tend to trade frequently (e.g. several times during a day). Consequently, the ability to forecast within minutes, or even seconds, the risk posed by a trader is vital for effectively deciding which trades should be hedged.

Grouping investors as A- or B-Book is a typical binary classification problem and the literature has identified many characteristics of traders which can be used to help achieve this (e.g. demographic factors, experience, trading frequency, market influence, past performance, trading discipline and risk control) (Odean, 1998; Barber and Odean, 2000; Barber *et al.*, 2009a; Barber *et al.*, 2009b). A number of different classification methods could be used for such a task (see LiaoChu and Hsiao (2012) extensive survey). However, there are specific features of this problem that make it particularly challenging, including the facts that (i) many factors/variables are needed to effectively classify the traders, (ii) there is a very wide dispersion of types of traders (e.g. those with very small and those with very large trading volume) and iii) there are few potential A-Book traders within the population in comparison with B-Book ones (i.e. overrepresentation of the B-Book class).

3.3 Literature review

3.3.1 Characteristics that predict traders' future performance

Researchers have identified a range of factors influencing traders' future performance in different markets around the world. These factors fall into the following major categories: past performance, disposition effect, experience (SeruShumway and Stoffman, 2010; Barber and Odean, 2011), demographics, socio economic level (Dhar and Zhu, 2006), knowledge of local markets (Ivkovic and Weisbenner, 2005), the influence of the environment where he/she normally interacts (i.e. communities where they live or work and social interaction) (BrownIvkovicSmith and Weisbenner, 2008; Kaustia and Knupfer, 2012). A discussion of the contributions and limitations of these studies are presented below.

BarberLeeLiu and Odean (2009a) and Barber and Odean (2011) demonstrated that a relationship exists between traders' past performance, measured by average return, and their future performance. However, they showed a general overview of this relationship, without exploring more in detail whether additional patterns are presented.

The Deposition Effect (DE), a psychological bias, which leads some traders to sell their winning positions too early and to hold their losing positions too long (Shefrin and Statman, 1985), has also been shown to affect traders' performance around the world including in the USA (Shefrin and Statman, 1985; Odean, 1998); Israel (Shapira and Venezia, 2001); Australia (Kaustia, 2004); Taiwan (BarberLeeLiu and Odean, 2007); Korea (Choe and Eom, 2009) China (Yonghong, 2001; Feng and Seasholes, 2005; ChenKimNofsinger and Rui, 2007) and Finland (Kaustia, 2010).

It has been shown that experience and learning reduces the degree to which traders display the DE, and, as a result improve their trading results (Feng and Seasholes, 2005; Dhar and Zhu, 2006; Greenwood and Nagel, 2009; Linnainmaa, 2011). For example, Shapira and Venezia (2001) and BrownChappelDa Silva Rosa and Walter (2006) demonstrated that portfolios managed by professionals (i.e. more experienced and informed) tend to be more diversified, less prone to the DE and more profitable.

However, contradictions are found in the literature examining the impact of experience on performance, particularly when certain proxies for experience are employed. For example, Barber and Odean (2000) found that more active traders underperform (by 6.5% cf. average investors), whereas Dhar and Zhu (2006) and SeruShumway and Stoffman (2010), who used trading frequency and duration of trading in years, found that experience leads to improved performance. Consequently, this suggests that using trade frequency on its own as a proxy could provide an incomplete picture of the impact of traders' experience. On the other hand, Greenwood and Nagel (2009) suggest the trader's age as proxy of experience, as they discovered that older traders tend to be less prompt to the DE.

Demographic factors have also been found to be correlated with traders' performance. In particular, it has been found that found that wealthier, professional and older individuals are less subject to DE and tend to demonstrate better trading performance (Terpstra *et al.* (1993), Dhar and Zhu (2006), Yamaguchi (2006), Anderson (2013) and Korniotis and Kumar (2010)). Research exploring the relationship between gender and performance has produced mixed results: Barber and Odean (2001) and Feng and Seasholes (2008) both reported that men traded more frequently. However, the former found that women achieved lower returns and the latter that there were no differences in the performance of men and women.

Local influence (influence of neighbours and social environment) and local preferences have also been found to affect traders' choices (Ivkovic and Weisbenner (2005), BrownIvkovicSmith and Weisbenner (2008), Kaustia and Knupfer (2012)). For example, Ivkovic and Weisbenner (2005) found evidence of those who trade in local shares achieve better performance, as they tend to have more information of the local market and Kaustia and Knupfer (2012) established that stock market participation increases in neighbourhoods where traders achieve positive returns.

In summary, previous research suggests that traders' characteristics such as experience, age, gender, demographic background, local influence (i.e. influence of neighbours and social environment on traders' investment decisions) or features related to their trading activity, such as past performance and the degree of DE displayed are indicative of their future trading behaviour and performance. However, the precise nature of the relationship between these factors and future trading performance as well as the degree to which these features interact is complex and remains under-researched (Röthig and Chiarella (2007); Lim *et al.* (2009); Bolgorian and Raei (2011)). Consequently, the research conducted in this second paper uses the modelling capabilities of GBMs (e.g. insensitivity to outliers and robust modelling of non-linearities and powerful classifications) to gain greater insight into these relationships, to better predict traders' future behaviour.

3.3.2 Methods used to model individual investor behaviour.

General Linear Models (GLM), in various forms, have been the most popular method for investigating investors' behaviour. For example, they have been used to explore the impact on investors' behaviour of demographic/behavioural characteristics (e.g. gender, age, income, DE, social background and risk attitude) (see for instance Anderson (2013), Korniotis and Kumar (2010), Korniotis and Kumar (2011) and Barber and Odean (2001)). Other researchers have employed time series analysis to examine a range of features, such as traders' overconfidence and past performance (Barber and Odean (2000) and BarberLeeLiu and Odean (2014)).

These approaches used for modelling investors' behaviour suffer from a range of limitations, including the assumptions employed in the estimation of their parameters (LemeshowKlar and Teres (1995); Mitchell (1997); Rushton (2000)). For example, linear regression assumes a linear relationship between the dependent and the independent variables. This assumption has been shown to be violated in a number of trader behaviour studies (Röthig and Chiarella (2007); Lim, Habibullah and Hinich (2009); Bolgorian and Raei (2011)). Consequently, if it is claimed, for example, that younger traders outperform their older counterparts, the result may not capture the fact that it is traders in a particular age group (say 30-40) years old, who are the ones who actually outperform other groups. Furthermore, parametric models are highly sensitive to scale, outliers, erroneous data and correlations between variables and this can lead to erroneous conclusions. In addition, prior to employing GLM and time series analysis, researches must invest time in data preparation to remove correlated variables and to treat outliers missing values and erroneous data.

By contrast, GBMs involve adaptive decision trees' ensembles, which tend to be less sensitive to outliers, erroneous data or missing values and, therefore, do not require intensive data preparation. In particular, decision trees are less sensitive to outliers and the ensemble process progressively corrects the residuals in each iteration, producing a better fit of the data.

However, it is important to highlight that ensembles methods such GBM are not a panacea and a minimum data quality is required. The *Garbage In Garbage Out* (GIGO) principle still applies, consequently results coming from erroneous data cannot be completely trusted.

Another relevant characteristic of the GBMs are the fact they do not require the variables to be independent or to have a linear or monotonic relationship with the target variable, all of which are normally required by GLMs (Friedman, 2001, 2002; Elith *et al.*, 2008). The undertaken study in this thesis demonstrates these valuable properties of GBMs and how they can produce superior classifications to more traditional methods. Furthermore, this study also demonstrates that GBMs enable to compute the contribution of each variable in the discrimination process. This allows developing greater insight into the characteristics of investors (e.g., past performance, demographic factors and levels of disposition effect) and how they may affect their future performance.

3.4 Description of the data

The data employed contains details of 4.5 million trades placed by 5184 UK spread traders, with a trading volume of more than £190 millions, between November 2004 and March 2013. All these trades were associated with the most popular financial instruments in UK spread trading markets: UK FTSE100 index, German Xetra Dax index, Euro - Dollar exchange rate and Pound - Dollar exchange rate. Between them, these instruments account for around 60% of total trading volume in UK spread trading markets. Further details of these trades are presented in Appendix A.

Using data from spread trading market offers two major advantages. First, earnings from spread trading are not taxed in the UK. This ensures that spread traders' decisions are not influenced by seasonal selling for tax reduction purposes, as is common amongst investors in other financial markets (Odean, 1998; Dhar and Zhu, 2006). Consequently, the data can provide a clear picture of the factors which influence traders' behaviour. This is not the case for data from traditional stock markets where tax-related factors influence trading behaviour. Second, because this market has low entry barriers, it attracts a diverse range of investors (Paton and Williams (2005)). This feature leads to a wide range of investors' behaviours being represented, thus leading to considerable variability in the data (e.g., high and low frequency traders exhibit significantly different behavioural patterns). This last characteristic makes this data unique, with the desirable complexity, for challenging DMC techniques attempting to profile individual investors.

Independent Variables

It is clear from the previous research outlined in section 3.3 that in order to predict individual investors' behaviour and therefore profile them, it is relevant to consider several attributes from each of these categories: past performance, disposition effect, experience, demographic background, socio economic level, knowledge of local markets and local influence. The variables employed in this research are outlined, associated with each of these categories in Table 3.1.

The company, who supplied the trading data, discloses that traders often provide inaccurate income and profession data. Consequently, rather than rely on the reported values, proxies were constructed for income information. Specifically, the UK Deprivation Index of the trader's postcode was employed, provided by the ONS (2010), as a proxy for identifying traders with higher incomes and those with access to a better education. In addition, the trader's initial deposit with the spread trading broker was included, as a proxy for income/wealth.

It is also relevant to include, at least as far as possible, the risk preferences of a trader. As a broad indication of this, this study determined the ratio of the average stakes of an individual trader (for a particular number of trades) and their initial deposit. A higher value suggesting that the trader during this period was willing to put at risk a greater proportion of their funds.

Num.	Effect modelled	Variable	Description
1	Experience	Age	Trader's age at time of trade

Num.	Effect modelled	Variable	Description
2	Trading Discipline	$AmendOrderPerTrade_n$	The average number of amendments in the trade's orders[14] (i.e. stop loss or profit limit) in the last n trades. It is expected this variable captures traders' discipline.
3	Income	$AmountDep_n$	Total deposited (£'s) by the trader in trading account over the last n trades.
4	Profitability	$AmountWithd_n$	Total withdrawal (£'s) by the trader from their account in the last n trades.
5	Past Performance	$AvgPL_n$	Average of Profit /Loss(£'s) in the last n trades
6	Past Performance	$AvgPst3_n$	% of trades closed with 3 or more points in profit in the last n trades. 3 points is the average value of those traders who ends in profit after first 100.
7	Past Performance	$AvgReturn_n$	Average rate of return of the trader in the last n trades
8	Trading Discipline	$AvgShortSales_n$	Tendency for the trader to trade short vs long during the last n trades.
9	Psychological bias	$AvgWDE$	Recently weighted average of the DE of the traders over their trade history.
10	Past Performance	$AvgWReturn$	Recently weighted average of return rate of the traders over their trade history.
11	Risk Control	$AvgWSharpeR$	Recently weighted average of the Sharpe ratio of the traders over their trade history.
12	Psychological bias	DE_n	The level of DE for the trader in their last n trades.
13	Proxy for wealth	$DeprivationIndex$	Deprivation index of the trader's notified address. Lower Deprivation index = more deprived areas.
14	Past Performance	$Direction_n$	Dummy variable, indicating if the trader has been in profit/loss during the last n trades
15	Trading Discipline	$DiscIndex = Sum\ (Max\ Papers\ Profit) / Sum\ (Max\ Paper\ loss)_n$	This variable measures how much (£) on average a trader leaves a winning position open vs. a losing position. (e.g. if the variable takes the value of 2, the trader, on average over the last n trades, allowed their trading positions to reach the point where the profit was double that of the losses they allowed to build up in losing positions).
16	Trading Discipline	$DurationRate_n$	The average time that a trader leaves a winning position versus a losing position open during the last n trades
17	Trading Discipline	$DurationRatio_n$	Average Trade Duration (mins) / STD Trade Duration for a trader during their last n trades
18	Demographic	$Gender$	Dummy variable indicating whether the trader was Male / Female.
19	Income	$InitialAmount3W$	The amount of money that a trader deposited in its account during the first 3 week of trading.
20	Technology	$MobileCloseRate_n$	% of trades closed by a trader using mobile apps in their last

[14] When a position is open, the trading system creates an order with a stop lost limit. If the losses hit this limit the trade will be closed automatically. Investors can amend the initial order and also set up profit limits.

Num.	Effect modelled	Variable	Description
			n trades.
21	Dummy	$MobileFlag$	Dummy variable: 1 after 06/10/2010, when the mobile app was launched
22	Income	$NumTopUps_n$	Number of deposits into the trader's account during the last n trades.
23	Experience	$NumTrades_n$	Number of trades closed by the trader until the last n trades
24	Profitability	$NumWithd_n$	Number of withdrawals from the account during the last n trades.
25	Sophistication	$OrderCloseRate_n$	Proportion of trades automatically closed by an order (e.g., stop loss) in the last n trades.
26	Profitability	$ProfitRate_n$	Profit rate of the last n trades.
27	Interaction	$ProfitRateXDur_n$	Interaction between profit rate and the duration rate (variable 26 and 16) during the last n trades.
28	Risk control	$Sharpe\ Ratio_n$	Sharpe ratio of a trader's last n trades (e.g., the mean / standard deviation of the returns during the last 20 trades.
29	Risk Control	$StakeInitDepRatio_n$ $= AvgStakeSize_n$ $/InitialDeposit3w_n$	The ratio between the average stake size of the trader's last n trades and their initial deposit.
30	Frequency	$TradFQ_n$	The average number of trades posed by each trader per day during the last n trades
31	Local Influence	$UKxFTSE_n$	Percentage of trades placed on the FTSE100 index during the last n trades. The FTSE100 index is the traders' home based market as all the traders are from the UK
32	Profitability	$WinTradeRate_n$	This variable is the trader's number of wins/number of losses on trades during the last n trades
33	Profitability	$WithdDepRatio_n$ $= Amt\ Whitdraws$ $/Amt\ Deposits_n$	The ratio of the trader's withdrawals/deposits during the last n trades.

Table 3.1 Independent variables employed in predicting traders' performance

3.4.1 Measuring performance

Average return over stake

A trader's average returns (£) over stake size is commonly used in the spread trading industry to distinguish high and low performing traders. Consequently, this research measured the average returns (£) over stake size for trader i in K_i trades, as follows:

$$AverageReturn_{K_i} = \frac{\sum_{k\in K_i} PL_k}{\sum_{k\in K_i} Stake_k} \tag{3.1}$$

where, K_i is the set of trades placed by trader i, PL_k is the profit/loss in British pounds when trade k is closed and $Stake_k$ is the stake in British pounds for trade k placed by trader i.

Dependent variable

Several risk managers of spread trading platforms were surveyed, and a consensus was found concerning their definition of successful traders (and consequently labelled as A-Book). In particular, these were defined as those with an $AverageReturn_{K_i} > 5\%$, as these traders generally achieve a minimum profit of £1000. This sum is significant, because most risk managers argued that such an amount would fully justify the costs of hedging such an individual's trades. Furthermore, a categorical target variable was chosen (cf. a continuous variable such as average return over stake) because the aim is to distinguish the group of traders with the highest performance. Consequently, this research employed the following dependent variables:

$$Target_{K_i} = \begin{cases} 1 \; if \; AverageReturn_{K_i} > 5\% \\ 0 \; otherwise \end{cases} \tag{3.2}$$

where K_i is the set of trades placed by trader i. K_i is also referred as the horizon trade number. Next section explains how this indicator is determined.

3.4.2 Horizon trade number and buffer

This research aims to establish a relationship between investors' characteristics together with their trading behaviour and their performance. Specifically, this paper aims to predict the probability that a trader will trade in a manner which would lead to them being classified as an A-Book trader (i.e. $AverageReturn_{K_i} > 0.05$) after he/she has performed K_i trades. It was chosen to set $K_i = 100$ trades, as from an initial exploration of the data, this research found that 50% life time of traders is within 100 trades, as it is presented in Figure 3.1.

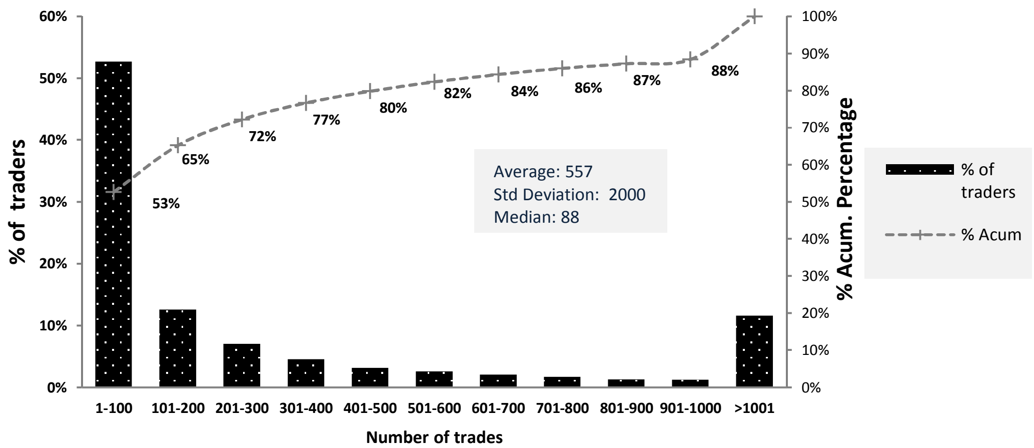


Figure 3.1 Traders' lifetime in number of trades

When collecting information about the traders' behaviour, it is imperative to determine the number of trades that is used as a buffer for computing the behavioural variables. Subsequently, the individuals' trading behaviour is analysed under the following hypothesis: "If an investor is in profit after n ($n < 100$) trades then the investor will be in profit after 100 trades".

To verify this hypothesis, the variable D_n is defined using the following formula:

$$D_n = \begin{cases} 1 \ if \ AvgReturn_n > 0 \\ \\ \quad 0 \ otherwise \end{cases} \tag{3.3}$$

Then a logistic regression model is fit to predict if a trader will be in profit after 100 trades using D_n as unique predictor. This is done for n = 5, 10, …,90. After that, the Area Under the Received Operator Characteristic Curve (AUC_n) is computed for each of these models. The AUC is used to measure the discrimination power of the model (specifically, the discrimination power of D_n as it is the only variable in the model). Figure 3.2 presents the results of this analysis.

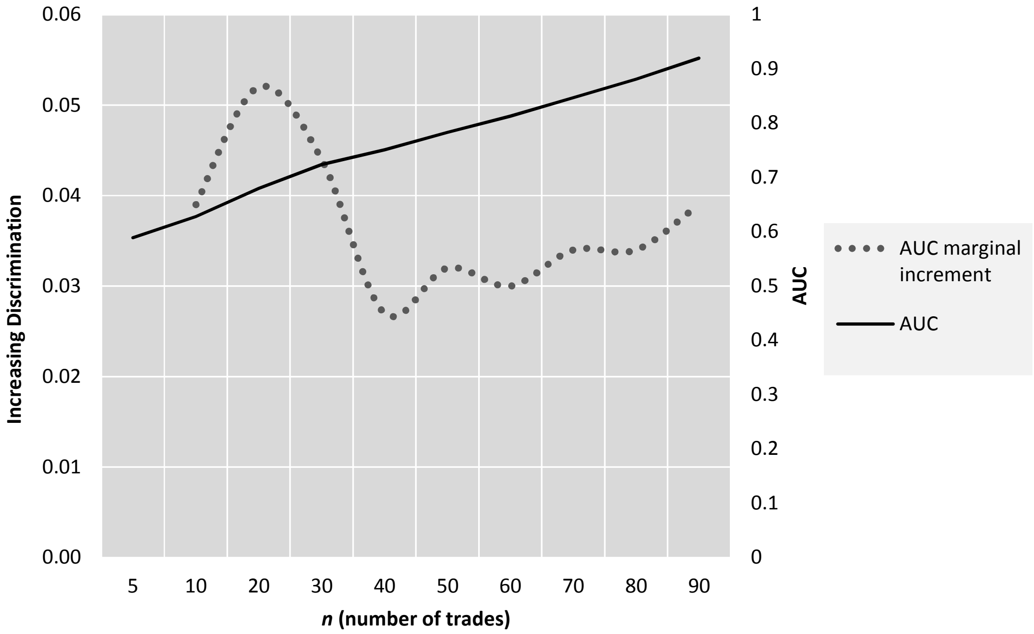


Figure 3.2 AUC for the models using the variable D_n only as independent variable

Figure 3.2 displays the AUC_n of each logistic model, for n = 5, 10, …,90, as well as the marginal increment of the AUC (i.e. $AUC_{n+1} - AUC_n$). Particularly, the figure shows that as n increases, the AUC increases. However, the highest incremental in AUC occurs when $n = 20$ trades. This suggests that 20 trades could be the minimum number of trades needed to observe when predicting the performance of a trader in its subsequent 100 trades.

In summary, the results indicate that traders' information can be used to forecast whether they will have an $AvgReturn_{100} > 5\%$ after 100 trades.

$$Target_{100} = \begin{cases} 1 \ if \ AverageReturn_{100} > 5\% \\ \\ \quad 0 \ otherwise \end{cases} \tag{3.4}$$

Consequently, each trader's behaviour and characteristics (i.e. the independent variables described in Table 3.1) are computed over 20 trades (i.e. $n = 20$) and the performance they achieve over the subsequent 100 trades (i.e. the dependent variable (3.4)), this provides a dataset containing more than 250 thousand observations of trading behaviour and characteristics related to performance of those 5184 spread traders selected over a time span of 9 years.

3.5 Methodology

This section describes the methodologies used for establishing whether a trader is likely to achieve a 5% average return in their next 100 trades and should, therefore be classified as an A-Book trader.

3.5.1 General overview of the data mining process

According to Friedman (2001) and Friedman (2002), the process for data mining can be described as follows: Let y be a response variable of the event to be studied or predicted and a set of variables $\mathbf{x} = \{x_1, \dots, x_n\}$, which represent information associated with y. It is sought to develop a mapping function F such as $y = F(\mathbf{x})$. A training sample $\{y_i, \mathbf{x}_i\}_1^N$ is then used to develop an approximation for the mapping function F^*. Therefore, undertaking an optimisation over the space function of $F's$ would make it possible to establish such an approximation. The function that has to be optimised is presented below:

$$F^* = \arg min_F\, E_{y,\mathbf{x}} L\big(y, F(\mathbf{x})\big) = \arg min_F\, E_{\mathbf{x}} \left[E_y \left(L\big(y, F(\mathbf{x})\big) \right) \mid \mathbf{x} \right] \tag{3.5}$$

where
$L\big(y, F(\mathbf{x})\big)$ is an error function over $\big(y, F(\mathbf{x})\big)$
E is the expected value
$\arg min$ is the minimum argument function

Some of the error functions used in practice are $L(y, F) = (y - F)^2$, the squared error for regressions where $y \in \mathbb{R}$, and $L(y, F) = \log(1 + e^{-2yF})$ the negative binomial log likelihood for binary classification $y \in \{-1,1\}$, used in logistic regression models.

Friedman (2001) suggests the following representation for F:

$$F(\mathbf{x} | \{\beta_m, \mathbf{a}_m\}_1^M) = \sum_{m=1}^{M} \beta_m h(\mathbf{x}, \mathbf{a}_m) \tag{3.6}$$

where $h(\mathbf{x}, \mathbf{a}_m)$ is a base learner (eg. classification trees or logistic transformations) , $\mathbf{a}_m$ are the parameters that characterise the learner itself, β_m is the appropriate weight to minimise the error function $L\big(y, F(\mathbf{x})\big)$ and M is the number of sequences used by the GBM to estimate $F(\mathbf{x})$.

Friedman (2001) suggests that by restricting F to a particular class of parametrised functions, an approximation for F^* can be found.

Several machine learning algorithms follow the form specified in (3.6). Table 3.2 presents different versions of $h(\mathbf{x}, \mathbf{a}_m)$ associated with LR, NN and DT.

Classifier	$\boldsymbol{h}(\mathbf{x}, \mathbf{a}_{\boldsymbol{m}})$
Neural Networks	$\mathbf{a}_m$ are the variables in each perceptron. and $h(\mathbf{x}, \mathbf{a}_m)$ is the activation function in each perceptron.
Classification and Regression Trees (CART)	$\mathbf{a}_m$ are the boundaries of the regions and the splitting variables defined by each terminal node. $h(\mathbf{x}, \mathbf{a}_m)$ is the average value of y in each region.
Support Vector Machine	$\mathbf{a}_m$ are the support vectors. $h(\mathbf{x}, \mathbf{a}_m)$ is the chosen kernel

Table 3.2 Cases of base learners in selected approaches

3.5.2 Gradient Boosting Machines

GBM, developed by Friedman (2001) and Friedman (2002), consist of a sequential approximation of $F(\mathbf{x})$ from a combination of basic learners. This approximation is given by the following expression:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h(\mathbf{x}, \mathbf{a}_m) \tag{3.7}$$

Where $\mathbf{a}_m$ is:

$$\mathbf{a}_m = \arg min_{\mathrm{a},\beta} \sum_{i=1}^{N} [g_m(\mathbf{x}_i) + \beta h(\mathbf{x}_i, \mathbf{a})]^2 \tag{3.8}$$

ρ_m is:

$$\rho_m = \arg min_\rho \sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) + \rho h(\mathbf{x}_i, \mathbf{a}_m)) \tag{3.9}$$

and g_m is the gradient from the optimisation of the error function evaluated in $F_{m-1}(\mathbf{x})$. This is given by:

$$g_m = \left[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x})=F(\mathbf{x}_i)} \tag{3.10}$$

CTs are used as a basic learner and $L(y, F(\mathbf{x})) = (y - F(\mathbf{x}))^2$ as an error function. This is the $\mathrm{L_2\_TreeBoost}$ proposed by Friedman (2001) and Friedman (2002) for solving binary classification problems.

CTs have several advantages over other classification models, in particular because they do not require any particular form of the relationship between the dependent and the independent variables. Furthermore, CTs are not readily biased by outliers (Olshen and Stone, 1984; Steinberg and Cardell, 1998). However, the maximum number of branches ω allowed in the CT has to be defined. If ω is too small, it will lead to very slow convergence. On the other hand, large values of ω will add more complexity to the algorithm and this will consume considerable more processing time.

A part from selecting the base learner, the tuning of the shrinking parameter needs to be considered. The shrinking parameter υ was introduced by Friedman (2001) (also referred as the learning parameter) to control for overfitting[15].

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \upsilon * \rho_m h(\mathbf{x}, \mathbf{a}_m) \tag{3.11}$$

where $0 \leq \upsilon \leq 1$.

Friedman (2002) also added a bagging procedure proposed by Breiman (1996) to improve the fitting process (i.e. help to minimise the error function). This involves using a sample $\mathbf{x}_\pi$ (where π is the sample size) of the training data $\mathbf{x}$, for determining the F_m functions.

Then, in order to optimise the performance of GBMs, the appropriate three meta-parameters have to be determined $(\omega,\ \upsilon, \pi)$. By a process of enumeration across the variable space it was determined that $(\omega = 6,\ \upsilon = 0.1, \pi = 0.5)$ produced a good fit of the data. It is relevant to highlight that the main focus of this paper is simply to demonstrate the value of GBM for predicting high risk traders. Consequently, this research relied on these approximate values rather than determining the optimal values for these parameters. The parameter optimisation process for GBMs is beyond the scope of this paper.

[15] Overfitting is when models start memorising the data, leading to a poor prediction in the out of sample data.

3.5.3 Model comparison

This investigation uses the following models as benchmarks to compare their performance with GBMs for profiling A-Book traders. Also, AUC is employed as a performance metric to undertake the comparison analysis. The main characteristics of the benchmark models are shown in Table 3.3.

Classifier	Characteristic / Implementation
Logistic Regression (LR)	Using the logistic function as a link function.
Classification and Regression Trees (CT)	With max depth of 6, 2 splitting nodes and reduction in the entropy measure and variance.
Neural Networks (NN)	Multilayer perceptron with 5 hidden layers.
Support Vector Machine (SVM)	Using polynomial kernel

Table 3.3 Main characteristics and implementations of selected base learners

To provide such a comparison, the dataset is divided into three subsets: the training set (50% of the data: used for fitting the models); the validation set (30% of the data: used by CT and GBM as stop criteria in order to avoid overfitting) and the test set with (20% of the data: for out of sample assessment of the performance of the models). To ensure that each subset contains the same proportion of A- and B-Book traders, a stratified random sampling procedure is performed when dividing the data, using the target variable as the stratification variable. A similar approach was used by Fitzpatrick and Mues (2016) when using GBMs to forecast potential defaulters in mortgage portfolios.

3.5.4 Main factors predicting A-Book traders

GBM allow identifying which variables (characteristics and behaviour) are most important in determining whether a trader is likely to pose significant risk, and hence should be classified as an A-book trader. This ordering of variables is achieved by determining the proportional contribution that each variable make to reduce the error function selected in 3.5.2 (see Friedman (2001)). This procedure is employed to answer the first research question, namely: *Which characteristics and behaviours are the most important for profiling high risk (A-book) traders*?

As it was mentioned previously, the ultimate goal of the data mining process is to develop a good approximation $\hat{F}(\mathbf{x})$ of the mapping function $F(\mathbf{x})$. GBM produces an explicit approximation of this mapping function given by (3.11). By plotting each independent variable against $\hat{F}(\mathbf{x})$, it is possible to demonstrate the nature of the relationship between a characteristic or behaviour (characterised by each of the independent variables) and the target variable (whether the trader is high risk or not). For example, thus enables to visualise non-linear relationships and potential interactions of these variables in classifying A-Book traders. This approach permits answering the second research question: *In what way are traders' characteristics and behaviours related to whether or not a trader poses a high risk?*

3.6 Results

3.6.1 Model performance comparison

Five different models are used to profile potential A-Booked traders: LR, CT, NN, SVM and GBM. Table 3.4 displays the resulting AUC for each of the models, computed for the three datasets: training, validation and test. The results demonstrate that the values of the AUC, across the three dataset, are very similar; indicating that none of the models is overfitting the training data.

Model	Train	Validate	Test
SVM	0.642	0.648	0.640
CT	0.767	0.772	0.764
NN	0.842	0.843	0.832
LR	0.845	0.857	0.849
GBM	0.922	0.917	0.914

Table 3.4 AUC of the benchmarking models and GBM

3.6.2 Traders' behaviour characterisation

As it was explained in section 3.5.4, the importance of each variable for profiling potential A-Book in the case of GBMs, is given by assessing each variable contribution to the improvement of the error function using the training dataset (see Friedman (2001)). Figure 3.3 displays the results of this analysis and demonstrates the importance of different variables for determining A-Book traders.

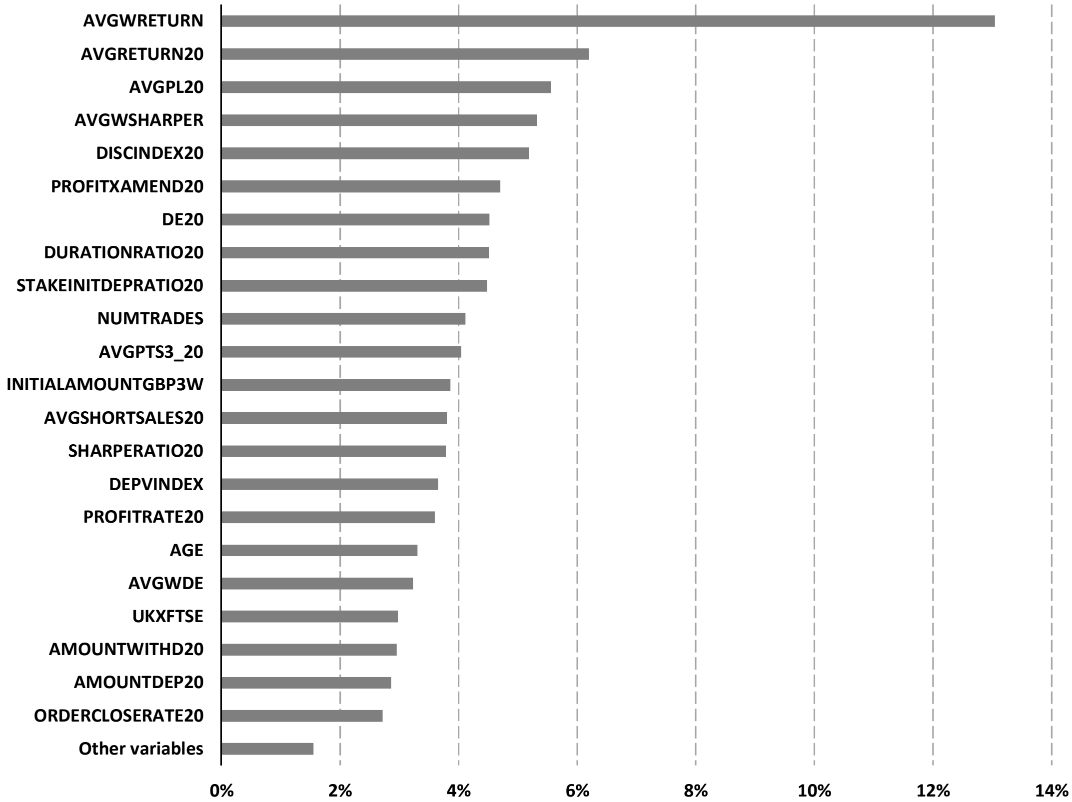


Figure 3.3 GBM variables' importance plot

By using each variable's contribution to the improvement of the error function in the GBM, it is possible to determine that the variables associated with the trader's past return (*AvgWReturn*, *AvgReturn* and *AvgPL20*), risk control (*AvgWSharpeR*), trading discipline (*DiscIndex20 and ProfitXAmend20*) and the disposition effect (*DE20*), are the ones with the highest contribution when profiling A-Book traders.

3.6.3 Cross validation

To verify the stability of the models' performance, the N-fold cross validation is used, with N = 10 (Hosmer Jr and Lemeshow, 2004). This procedure is a standard practice for validating classification models, it consists into split randomly the entire data set, 10 times, in two groups: one with the 90% of the data (train data set) and the other remainder 10% (test data set). Again, a LR, CT, NN, SVM and GBM models are fit using the train data sets and the models are assessed with the test data sets. For the assessment, the AUC is used. To obtain the overall assessment of the models, the average and the standard deviation of the AUC of all test data sets are computed. Ideally, the aim is to look for the model with the highest average and the lowest standard deviation. Table 3.5 presents the results of the cross validation for the models and Figure 3.4 presents a comparison of the performances of the models graphically. The results are consistent with the AUC results shown in Table 3.4, with GBM exhibiting a better and more stable classification performance than the rest of the models (higher average AUC and lower standard deviation of AUC for both the training and validation data sets).

Model	Avg of Train_AUC	StdDev of Train_AUC	Avg of Valid_AUC	StdDev of Valid_AUC
SVM	0.67	0.033	0.67	0.039
CT	0.72	0.028	0.72	0.027
LR	0.82	0.029	0.82	0.031
NN	0.85	0.001	0.84	0.006
GBM	0.92	0.001	0.91	0.007

Table 3.5 Cross validation results

Using the values generated by the 10-fold cross validations, it is possible to compute boxplots of the AUC for the training and validation set. The results are displayed in Figure 3.4. These results indicate GBM exhibits the highest performance and the lowest variability in the AUC in the cross validations, with NN being the second best performing model in terms of both average AUC and variability of AUC.

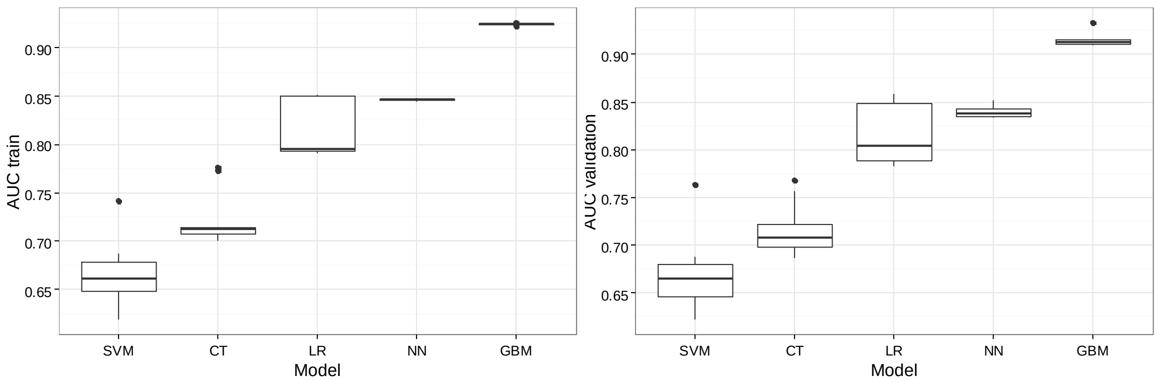


Figure 3.4 Models cross-validation comparison

Therefore, the results demonstrated that GBM significantly outperforms other data mining techniques such as logistic regression, classification trees, neural networks and support vector machines when profiling the risk posed by individual traders.

3.6.4 Exploring the influence of the independent variables over traders' performance

As indicated in 3.5.1 the general data mining process consists of finding a good approximation $\hat{F}(\mathbf{x})$ of the mapping function $F(\mathbf{x})$. Although such approximation can be produced by the other data mining techniques explored in this paper, the fact that GBM displays the highest discrimination power suggests that the approximation $\hat{F}(\mathbf{x})$ generated by GBM is more accurate than the other techniques used in this case study.

Then, the approximation $\hat{F}(\mathbf{x})$ can be used to analyse the influence of each variable over the classification process. Ideally, one way to undertake this analysis is by plotting each variable and $\hat{F}(\mathbf{x})$, and observing their relationship. However, this is not entirely possible when $\mathbf{x}$ has high dimensionality (i.e. more than two variables as in this case). Therefore, a partial approximation of $\hat{F}$ over a particular variable x_i is more desirable for visualising the contribution of the variable x_i over the classification process. Consequently, this research employs the partial approximation of $\hat{F}$ developed by Friedman (2001). A detailed explanation of the partial approximation of mapping functions is presented in Appendix E.

The partial approximation of $\hat{F}$ and the training data (described in section 3.5.3) are used to produce a series of plots. Friedman (2001) refers to them as *Partial Dependency Plots* (PDP). In each PDP the y-axis label as "***Output***" is the partial value of $\hat{F}$ and the x-axis is one of the variable used in the classification process. Higher values of Output in the graphic mean that the evaluated variable has significant influence to profile A-Book traders and lower values otherwise. For example, let's take the variable DE20, the value of Output (i.e. $\hat{F}$) reach its maximum when DE20 is near zero (see Figure 3.5), that means that traders with neutral (zero) DE are likely to be A-Book traders. On contrary when the values of DE20 are away from zero, Output tends to be negative, meaning that traders with this characteristic are less likely to be A-Book traders.

Consequently, the PDPs permit exploring more in depth the influence of each independent variable when attempting to identify A-Book traders. Subsequently, the PDPs allow going beyond the results of the current literature; as previous researches only present a general overview of this influence, because the limitations of the data mining methods employed by them. For instance, Barber and Odean (2011) argue that previous returns will be a good predictor for future performance. Then, it is expected that traders with higher returns in the past will be more likely to be A-Book. Nevertheless, when the PDP for *AvgWReturn* is observed, which measures the weighted average return rate (see Figure 3.6), effectively traders with higher past returns are likely to be A-Book traders, nonetheless, those ones with the lowest returns in the past are also likely to by A-Book traders. Similar results can be found when the variable *AvgWSharperR* associated to risk control is examined (see Figure 3.7).

Interestingly, in terms of traders' experience, the literature suggests the use of number of trades (*NumTrade*) as a proxy for traders' experience (SeruShumway and Stoffman, 2010). It should be expected that experienced traders outperform less experienced traders and are therefore considered as A-Book traders. However, the PDP for *NumTrade* shows that this is not necessarily the case (see Figure 3.8). Traders in their initial stages tend to be in line with A-Book behaviour, but after several trades their behaviour changes and it is more in line with B-Book traders and after several trades more, their behaviour changes again and they start behaving again like A-Book traders. This phenomenon partially support the theory of overconfidence proposed by Barber and Odean (2001) i.e. at the beginning traders seem to be more cautious in their decision. However once they gain confidence they start making mistakes, until finally some of them learn from their past experience.

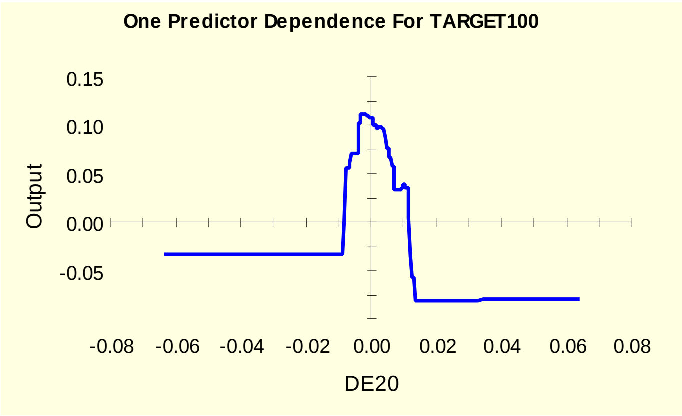


Figure 3.5 Influence of the Disposition Effect when profiling A-Book traders.

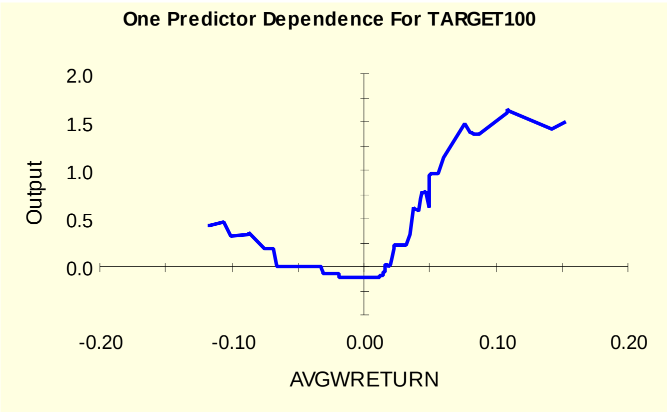


Figure 3.6 Influence of traders' past performance when profiling A-Book traders

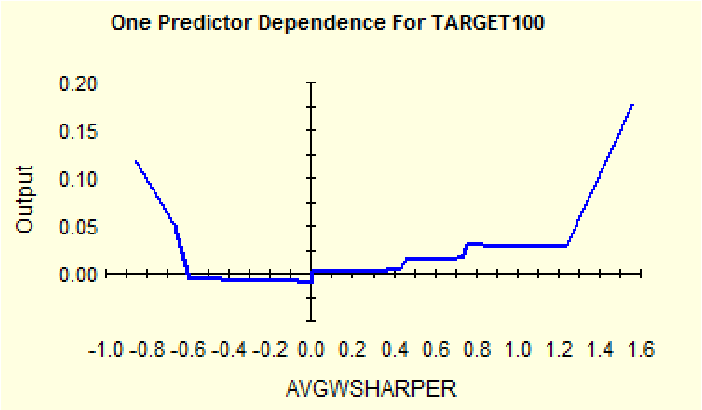


Figure 3.7 Influence of risk control when profiling A-Book traders

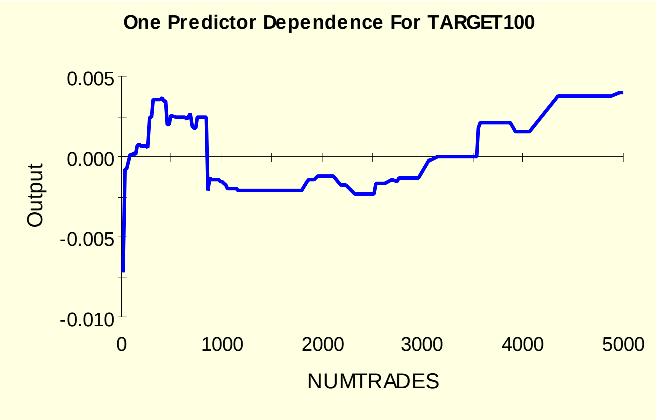


Figure 3.8 Influence of experience when profiling A – Book traders

3.6.5 Exploring the influence of interactions of the independent variables over traders' performance

Two or more characteristics can interact and therefore increase the chances for a better identification of A-Book traders. By using three dimension PDPs, it is possible to visualise the relationship between two independent variables and the partial value of the mapping function $\hat{F}$, which is used to identify potential A-Book traders.

In a similarly manner to the bi-dimensional case, the three dimensional PDPs help to provide further insight into the influence of the trader's characteristics on their performance and consequently expand the current literature. For example, the left hand side of the Figure 3.9 shows that traders investing in local instruments such the UK FTSE100 index tend to be more in line with A-Book traders, supporting the local influence theory of Ivkovic and Weisbenner (2005); traders tend to be more successful in local instruments, in this case the UK FTSE100 index. Also it was found that, younger traders exhibit an A-Book's type behaviour, contradicting Greenwood and Nagel (2009).

Then, the 3D PDP graphics display the influence of the interaction of these two characteristics over the traders' performance. For a better visualisation, colours are used to illustrate the altitude levels of the vertical axe (Output), which is the partial value of the mapping function $\hat{F}$. Like in altitude charts, blue means the lowest level, green medium level and red the highest level of $\hat{F}$. Therefore, Figure 3.9 shows that younger traders investing in the UK FTSE100 are in line with the A-Book traders' behaviour. Consequently, these results complements the conclusion of Ivkovic and Weisbenner (2005), where local influence can be boosted by trader's age.

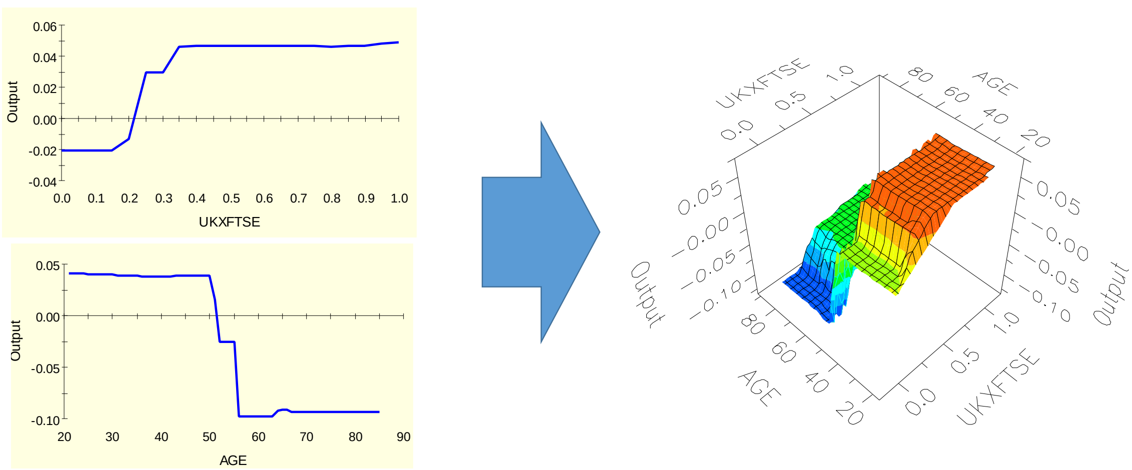


Figure 3.9 Influence of the Interaction between traders' age and investments in local instruments when profiling A-Book traders.

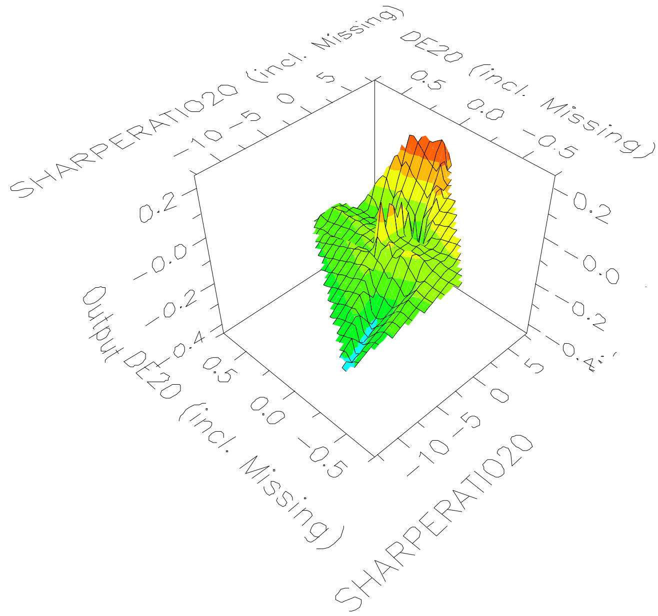


Figure 3.10 Influence of the Interaction between the DE and risk control when profiling A-Book traders.

Figure 3.10, shows the interaction of the Disposition Effect (*DE20*) and the risk control of the traders (*SharpeRatio20*). As it is expected above, the red area shows that traders which exercise risk control (high SharpeRatio20 values) and remain neutral in their DE (i.e. *DE20* = 0) tend to have a behaviour similar to A-Book traders. Interestingly, those traders with neutral DE also tend to exhibit a behaviour in line with A-Book traders (notices the spikes in the middle of the graphic) regardless their risk control levels. These results complement Barber and Odean (2011) regarding to the influence of the DE and past performance over trader's future behaviour.

3.7 Discussion

3.7.1 Suitability of the GMBs for profiling outstanding traders

As it was shown in the PDP graphics in section 3.6.4, most of the important characteristics that describe high risk traders, such as previous returns (*AvgWReturn*), risk control (*AvgWSharperR*) and disposition effect (*DE20*), exhibit strong nonlinear relationships with the mapping function $\hat{F}(\mathbf{x})$ (i.e. a non-clear descending or ascending pattern). This could be one of important issues that cause other data mining techniques to underperform (i.e. showing lower AUC). This view is supported by the fact that the NN model, which is well recognised for being able to deal with nonlinear relationship in the data, obtains the second highest value of the AUC metric.

Fitzpatrick and Mues (2016) presented evidence that GBMs exhibit slightly higher discrimination power (i.e. higher values of the AUC metric) than other data mining techniques when using static data (i.e. non-behavioural). However, the results displayed in section 3.6

suggest that GBMs could be even more effective in terms of classification performance when handling behavioural data, at least in the classification of individual investors (as seen by the results presented in Table 3.4). Additionally, the results demonstrate that GBMs can be suitable for analysing large scale data with high dispersion, (i.e. data coming from a heterogeneous and highly dissimilar population) such as the one used in this thesis.

3.7.2 Distinguishing characteristics of outstanding traders

The PDP capability of the GBMs enables to explore deeply the characteristics of those traders classified as A-Book (outstanding traders). For example, in terms of past performance and risk control, it was found that traders with poor previous returns and lower risk control can also become A-Book traders in the future. This can be seen in Figure 3.6 and Figure 3.7. These results complement Barber and Odean (2011) and BarberLeeLiu and Odean (2014) findings, as they highlight that the trader's previous performance can predict its future performance. However, simply assuming that traders who have achieved high average returns in the past will be those who achieve high average returns in the future is not a good basis for classifying traders. The apparently contradictory behaviour, traders with poor previous returns and lower risk control can become A-Book traders in the future, can be explained by the interaction between risk control and the DE as it can be observed in the PDP exhibited in Figure 3.10. Those traders who manage to keep a neutral DE (i.e. DE = 0) and risk control tend to become A-Book regardless of their past returns. In consequence, these findings are valuable as they show that that not only a trader's success in the past may help predict trader's future success, but also the interaction of their levels of disposition effect and risk control.

Another finding observed by analysing Figure 3.8 is the fact that traders in the early stages of their trading activity with a brokerage can achieve results which would classify them as A-Book traders. This may appear surprising since experience is generally regarded as an important factor affecting performance positively as it is suggested by SeruShumway and Stoffman (2010). This effect may be observed because individuals who have not traded much with a particular broker may, for example, have developed experience trading with other brokers. Consequently, the number of trades with a particular broker, does not necessarily reflect their true degree of experience as it is claimed by SeruShumway and Stoffman (2010).

The results also demonstrate that traders aged between 20 and 50 years old tend to be classified as A-Book traders (see Figure 3.9). Additionally, as shown by Figure 3.9, there are interactions between age and the investment in local instruments (*UKxFTSE*). Figure 3.9 demonstrates that traders aged between 20 and 50 years old investing in local instruments are likely to be A-Book traders. These results support Ivkovic and Weisbenner (2005), as they claim that individual investors tend to be more successful when invest in local instrument. However, at the same time, these findings contradict Greenwood and Nagel (2009), as they claim that experienced traders tend to outperform and age could be a good proxy for trader's experience. Although the results question using age as proxy for experience, they also suggest that traders' exhibit different performance at different ages. Consequently, this suggests that trader's age should not be disregarded when attempting to classify outstanding traders, but that it needs to be incorporated alongside other behavioural variables.

3.8 Conclusions

This paper explores the suitability of gradient boosting machines (GBM) for profiling individual investors in the spread trading market, specifically to classify those who are most likely to achieve significant returns in the long run and this pose the greatest risk to brokers.

Data from individual investors has peculiar characteristics such high dimensionality (i.e. several variables) with errors and omissions and is very heterogeneous, for example, different investors' backgrounds, ages, incomes and risk attitudes, which are intended to be used for profiling investors. Consequently, these characteristics represent a challenge for the classification process for the majority of the data mining-based classification approaches, which require significant pre-processing work to avoid misleading results.

Using a large dataset from the spread trading market, evidence is gathered that demonstrates that GBMs overcome the drawbacks mentioned above. GBM outperformed other data mining techniques, such as logistic regression, classification trees, neural networks and support vector machines when employed to classify individual investors. Furthermore, this research also suggests that GBM could be superior when using behavioural data to profile individual investors.

Consequently, the GBMs can effectively reduce the time invested in data preparation and data modification, proving that the concept of ensembles models could be more appropriate for analysing this type of dataset, reducing preparation times.

By using GMBs capabilities, it was possible to get more insights of the investors' behaviour. For instance, similarities between the behaviour of spread traders and investors in stock markets were found. However, this research reveals that some of the relations between a trader's characteristics and their future performance follow strong nonlinear patterns. For example, it was found that traders who have achieved high returns in the past may be likely to achieve high returns in the future. Similarly however, it was found that traders with very poor returns in the past may also achieve very high returns in the future. It was discovered that one of the reasons for this is that there are important interactions between the levels of disposition effect with returns and risk control. Traders with neutral Disposition Effect are likely to perform well in the future, regardless their previous performance. These interactions lead to complex relationships which are not obviously discovered by using general linear model approaches. Consequently, it is possible that some of the previous conclusions in the literature, derived using such linear approaches, may not be completely accurate.

In conclusion, the study in this research exposes some relevant limitations of the methodologies used by previous researchers in the investigation of the individual investor's behaviour. It was also discovered that by using more robust and adaptive data mining techniques (ensembles models), such as GBM, is possible to achieve a better understanding, in this case study for instance, the features and characteristics that influence individual investors' behaviour.

Even though this study used a large dataset with more than nine years of trading history, this data could be biased due to changes in the trading policies imposed by the broker itself (e.g. fees, minimum deposits, etc.). Unfortunately, there are no records of when these changes happened.

Another limitation of this study is the use of information from UK based investors only (arising from the lack of demographic information of the international investors in the dataset). Consequently, conclusions from this study may not necessarily be applicable to other countries.

An interesting expansion of this work could be testing several ensemble methods using different types of configurations and to perform a comparative analysis in order to determine which of them are more successful when profiling individual investors.

Chapter 4: Do smart mobile apps produce smart financial decisions?

Abstract

Smart mobiles and their 'apps[16]' (SMA), have gained rapid acceptance amongst consumers. However, the impact of this technology on the nature and quality of individual decisions has remained under-researched. This research attempts to fill this gap by examining 4.5 million trades of 5184 investors in the UK spread trading market between November 2004 and March 2013. The results suggest that there are demographic differences between those who do and do not use SMA for trading. After controlling for these factors, it was possible to observe significant performance differentials and differences in the nature of the trading decisions of these two groups. Specifically, this study shows that those who use SMA achieve higher risked-adjusted returns, but exercise less trading discipline (measured by disposition effect). This study also discusses the possible reasons for these findings and the implications for financial market operators, regulators and for the efficiency of markets.

Key words: Decision making, smart mobiles, spread trading, investor behaviour, mobile apps

4.1 Introduction

Growth in the internet and mobile apps' sectors has been exponential, with 7.4 billion connections generated via smart mobiles (e.g., smartphones/tablets) by the end of 2014 and with 20 billion expected in 2019 (Index, 2015). Apple sold over 200 million iPhones in 2016 and in total more than 1.0 billion since 2007, when the iPhone was launch (Statistica, 2016). By 2020 it is estimated that sixty percent of the global population will have a mobile phone subscription (GSMA, 2015). The rise in the availability of mobile apps has also been phenomenal, with more than a billion in circulation in 2013 (MobiThinking, 2013). A leading think-tank reported that the mobile industry generated 3.8% of global GDP in 2014 (i.e. US$3 trillion) and estimated that it will grow rapidly up to 2020 (GSMA, 2015). The number of mobile internet users exceeded traditional web users (e.g. using Desktop/Laptop) in 2014, when mobile internet traffic reached 30 billion

[16] An application, especially as downloaded by a user to a mobile device

gigabytes (an increase of 69% from 2013, and a thirty-fold increase from 2000). These trends have changed social behaviour and business processes.

With such a rapid increase in access to content/information through smart mobiles a fundamental research question has emerged, namely, "*what is the impact of the use of SMA on the nature and quality of individuals' decisions?"* This is of particular interest where SMA users make critical decisions, such as financial decisions, within dynamic environments, whilst often remaining socially connected and regularly receiving mobile-based information on small display screens. Given the size of the mobile-related economy and the global adoption of SMA, answers to this research question are urgently needed.

However, there has been little research examining this question, probably due to the rapid growth in this sector (BhömerHechtSchöningKrger and Bauer, 2011) and the limited availably of suitable data. The work that has been undertaken has yielded inconclusive results. For example, ConstantiouLehrer and Hess (2014) analysed the impact of SMA aids on decisions but their data contained no performance measures.

Research has shown that communication/internet technology has produced a 'digital divide', with social gaps emerging; i.e. a divide between those who do and do not employ the technology appropriately for accessing information and using the acquired information to improve their decisions (e.g., Forman (2005); Forman *et al.* (2005); Arora *et al.* (2010); Hui and Png (2015)). It has proved difficult to determine whether new technology improves individuals' decision making, because research suggests that "smarter" people with better decision-making skills are more likely to choose new technology.

A few studies related to SMA have been undertaken in the areas of marketing, psychology, medicine and finance. However, none of these examine the extent to which SMA technology can help individuals improve their decisions. For example, Nakasumi (2012) studied the impact of the context on the use of SMA to aid decision making, ConstantiouLehrer and Hess (2014) developed a conceptual framework to analyse the cognitive decision making processes of individuals using mobile location-based service apps, BuijinkVisser and Marshall (2013) investigated the impact of clinical decision-making apps on patient care and Koenig-LewisPalmer and Moll (2010) examined the constraints which prevented individuals' adopting mobile banking apps. However, all these studies fail to measure the ultimate impact on the quality of such decisions and none of them account for the digital divide.

The third paper of this thesis fills this gap by examining the degree to which SMA technology improves decisions, having accounted for the digital divide. In particular, this paper examines a large dataset of decisions made by individual traders in a financial market between 2004 and 2013, specifically the spread trading market (4.5 million trades of 5184 traders, including both SMA and non-SMA based trades).

Employing this individual trading data offers three key advantages: first, individuals in these markets generally make several trading decisions in a day (buying and selling instruments) either using SMA or other trading channels, and all these decisions and the resulting profits or losses are recorded. This ensures a fair comparison under similar market conditions in a given day. Second, earnings from spread trading are not taxed in the UK. This ensures that spread traders' decisions are not influenced by seasonal selling for tax reduction purposes, as is common amongst investors in other financial markets (Odean, 1998; Dhar and Zhu, 2006). Third, spread traders have the option of trading via SMA, this technology offering access to instant and relevant information to make financial decisions and an interactive channel to execute the decision in a mobile environment (almost wherever and whenever the user requires). This allows to compare the quality of SMA vs non-SMA based decisions in a naturalistic and dynamic real-world environment.

In assessing the impact of SMA use on the nature and quality of individuals' decisions this paper seeks to:

(i) Establish to what extent individuals who use SMA for trading are distinguished from those who do not by their demographic profile; thus helping to establish the extent of the digital divide;

(ii) Identify differences in the decision-making behaviour and performance of traders who do and do not use SMA, whilst controlling for any digital divide effects;

(iii) Examine to what extent traders improve their trading performance when using SMA.

It was found that the performance of traders overall, in terms of their returns and their Sharpe ratios, improved following the introduction of SMA and that users of SMA achieved higher Sharpe ratios. However, the trading discipline (as measured by DE) of SMA users was worse than those who never used SMA, even when they both used traditional trading channels. This suggests that these results can be explained by the SMA technology enhancing traders' accessibility to accurate and up to date data at time of trading, wherever they are located (Sraeel, 2006; Koenig-LewisPalmer and Moll, 2010; Kourouthanassis and Giaglis, 2012; BuijinkVisser and Marshall, 2013; ConstantiouLehrer and Hess, 2014). Equally, the results suggest that those who chose to use SMA possess superior risk-control skills. In sum, the study performed in this third paper offers important lessons concerning how new technology can be evaluated quantitatively to inform future technology improvement whilst controlling for the digital divide effect.

4.2 Literature Review

Three streams of relevant literature are discussed within this paper. First, the digital divide literature, which suggests that there are differences in the types of people who are early or late adopters of new technology. Second, this paper explores the limited literature examining the impact of the SMA technology on individuals' decision processes and outcomes. Third, this study reviews the literature addressing key aspects of financial traders' behaviour, which could be affected by SMA.

4.2.1 Early adopters of technology and the digital divide phenomena

Much research has suggested that technology produces a digital divide, with an economic and social inequality emerging in terms of access to and use of information technologies (Norris, 2001; Van Dijk and Hacker, 2003; Forman, 2005; PhilipCottrillFarringtonWilliams and Ashmore, 2017; Tsetsi and Rains, 2017). Furthermore, research has suggested a second level digital divide, related to the type of internet connections users employ and their frequency of access to the internet (e.g. mobile internet) (Norris, 2001; Graham, 2014).

Some have suggested that the digital divide may arise initially from the differential appeal of the technology to different groups. For example, Chau and Lung Hui (1998) found that early adopters of technology tend to be young, male, opinion leadership types, who are computer literate and seek novel information. Similarly, Sraeel (2006) found that younger consumers aged between 25 and 34 tend to be more interested in mobile banking and payment apps. VenkateshMorrisDavis and Davis (2003) developed a framework to codify the factors leading to acceptance and use of technology. This includes performance expectancy — the degree of satisfaction that a technology can provide the user when performing a particular activity; effort expectancy — the level of difficulty the user experiences in using the technology; social influence — the importance and relevance of the technology for friends and family, and facilitating conditions — perceptions of having the resources and support to execute a particular task using the technology. VenkateshThong and Xu (2012) suggested additional factors that can drive acceptance and use of new technology, including: Hedonic motivation (pleasure and fun), price value (related to cost), and prior experience with the technology. They also suggest that demographic characteristics may interact with these drivers; e.g., hedonic motivation is stronger in younger males and in those less experienced with the technology.

The popularity of SMA and their high acceptance among individuals worldwide suggests that they have satisfied many of the drivers suggested by VenkateshMorrisDavis and Davis (2003). These drivers are likely to attract different types of users and may well further the digital divide. Indeed, VenkateshMorrisDavis and Davis (2003), VenkateshThong and Xu (2012) and Ghose and Han (2011) suggest that users of a technology such as SMA are likely to be demographically different from those who do not use the technology. If this is the case, then demographic factors need to be controlled when comparing the decision performance of users and non-users of SMA. This is particularly important because early adopters tend to be more sophisticated in terms of their decision quality (Hui and Png, 2015) and it may be the case that demographic factors interact with the degree of an individual's sophistication.

4.2.2 Impact of technology on decision processes and outcomes

Molloy and Schwenk (1995) argued that information technology (IT) aids decision making processes and permits individuals and organisations to access more accurate and updated data to enhance their analysis. The World Wide Web in particular, has allowed people to become more informed, enabling them to make more complex decisions using the minimum level of assistance. Such advantages have boosted the development of the mobile internet and SMA. In particular, SMA provide immediate access to information, enable interactions with the rest of the world and enable decisions almost anytime and anywhere (Sraeel, 2006; Koenig-LewisPalmer and Moll, 2010; Kourouthanassis and Giaglis, 2012; BuijinkVisser and Marshall, 2013; ConstantiouLehrer and Hess, 2014).

Despite the apparent benefits that these technologies offer, it cannot be assumed that they will necessarily lead to improved decision making. For example, there is considerable evidence that information overload can have an adverse impact on decision quality (JacobySpeller and Berning, 1974; Jacoby, 1984; Eppler and Mengis, 2004; Bawden and Robinson, 2009). It has also been shown that there are multiple factors related to mobile technologies that may impact users' decisions, including time constraints (Bang *et al.*, 2013), screen size related to task complexity (Chae and Kim, 2004), information format (Hong *et al.*, 2004), partial display of information (Napoli and Obar, 2014) and even browsing behaviour (Xu *et al.*, 2016). Clearly, measuring the impact of technology on decisions is challenging.

Individual preferences, context and interaction with other sources of information can also exert a significant influence on the results of decisions facilitated by these technologies (VenkateshMorrisDavis and Davis, 2003; Nakasumi, 2012; VenkateshThong and Xu, 2012; ConstantiouLehrer and Hess, 2014). For example, ConstantiouLehrer and Hess (2014) point out that a person who is using a location-based service to find a place to eat can also get, at the same time, information from friends, street ads, etc. Nakasumi (2012) suggests that the final decision in this case may be dependent on one's companion (e.g., a friend or a superior), thus illustrating how context can interact with the decision making process. Previous studies, therefore, suggests that researchers cannot be certain how SMA will impact individuals' decision processes or decision quality since external factors can frame and influence the decision process; quantification of the economic benefits derived from decisions facilitated by SMA is therefore problematic.

In the context of decisions associated with financial markets, mobile phones and more recently SMA enable users to engage in interactions which would not be possible without the technology (Prasopoulou *et al.*, 2006). SMA empower trading while allowing access to relevant information wherever traders are located and whenever they want to trade, thus facilitating more frequent trading (Yamaguchi, 2006). BhömerHechtSchöningKrger and Bauer (2011) examined a sample of 4,125 USA and European SMA users and found that on average they spent nearly an hour per day using their devices. They pointed out that significantly more information is needed concerning how SMAs are used. Xu *et al.* (2011) attempted to fill this gap by investigating patterns of use and preferences of more than 600,000 SMA users of a tier 1 mobile network in USA during one week. They concluded, based on the number of users and traffic volume, that the most popular apps were those with local content e.g. local news, radio and weather apps. Additionally, they discovered that certain apps are likely to be used in tandem (e.g., news and banking apps) and different apps are used at different times of day (e.g., news apps are normally accessed in the morning). Their research suggested that the main motivations for using SMA were seeking information (e.g. news, weather), entertainment (e.g. music, TV) and social networking. Furthermore, they discovered that financial traders tend to use SMA for gathering information and for interacting, particularly to buy or sell financial instruments. Whilst XuErmanGerberMaoPang and Venkataraman (2011) provide an insight into the manner in which SMA are used, the data they employed only covered the period of one week. Consequently, issues such as seasonality were neglected and the study did not explore the impact of SMA on decision processes or decision quality.

In summary, despite the increasing popularity of SMA, there is little literature addressing how this technology influences investors' behaviour in financial markets. The increasing use of SMA for financial trading makes this a pressing question, with potential implications for the efficiency of financial markets.

4.2.3 Traders' behavioural biases

Barber and Odean (2011) argued that individual investors are influenced by past experiences and the media, and, as a result, tend to sell winning and hold losing positions. This phenomenon, first identified by Shefrin and Statman (1985), is known as the Disposition Effect (DE) and is the most widely reported bias in the financial market literature: USA (Shefrin and Statman, 1985; Odean, 1998); Israel (Shapira and Venezia, 2001); Australia (Kaustia, 2004); Taiwan (BarberLeeLiu and Odean, 2007); Korea (Choe and Eom, 2009); China (Yonghong, 2001; Feng and Seasholes, 2005; ChenKimNofsinger and Rui, 2007) and Finland (Kaustia, 2010). Barber and Odean (2000) and Barber *et al.* (2006) demonstrate that this bias leads to poor trading performance in the long-term. Consequently, the DE is often used as an important barometer of trading discipline. Weber and Camerer (1998) found that traders exhibited the DE when choosing when to buy and sell shares. However, when traders were forced to sell their shares by an automatic procedure, such as a stop loss order, the DE was reduced significantly. Studies undertaken by Shapira and Venezia (2001) and BrownChappelDa Silva Rosa and Walter (2006) found that professional and institutional investors, who are generally regarded as exhibiting greater trading discipline, display the DE, but to a significantly lesser extent than individual traders.

The importance of the DE as a measure of trading discipline makes this bias an important element of the enquiry in this paper, concerning the impact of SMA on the decision processes of financial market traders. In particular, it is explored whether SMA can help to reduce this bias, by allowing access to up-to-date information and providing the facility to trade anywhere and at any time.

4.3 Hypotheses

As indicated above, it has been suggested that users of a technology such as SMA are likely to be demographically different from those who do not use the technology. This research sets out to explore the veracity of this proposition, by testing the following, 'Demographic characteristics' hypothesis:

> *H1: Those traders who employ SMA for trading activities differ in terms of their demographic characteristics from those who choose not to use SMA*

As SMA is a recent technological innovation, current users of SMA might be regarded as 'early adopters'. Consequently, they may possess the characteristics described by Chau and Lung Hui (1998) (i.e. younger, males, and computer literates who seek novel information). In addition, as Hui and Png (2015) point out, individuals who embrace new technologies are generally smarter and are more sophisticated in terms of their quality of decisions. Consequently, the following 'performance of SMA traders' hypothesis is tested:

> *H2: SMA users make better quality decisions than non-users, demonstrated by their superior trading performance when not employing SMA technology.*

Nakasumi (2012) and ConstantiouLehrer and Hess (2014) analysed the impact of SMA aids on decisions. However, their results were not conclusive because their data did not permit measuring the performance of the decision makers in study. Access to the individual trading records of traders enables to overcome this limitation. Consequently, this data allow to test the 'impact of SMA on trading performance' hypothesis:

> *H3: SMA has a significant impact on reducing the bias (DE) displayed by traders and has a positive impact on their trading performance.*

4.4 Data and Methodology

4.4.1 Spread trading data

Spread trading, in particular, involves placing bets over the direction of movement of financial instruments on different markets. In particular, the trading process consists into bet whether the price of a financial instrument will either increase or decrease. Details of how this market operates are described in Appendix A.

The data employed to test the hypotheses contains details of 4.5 million trades placed by 5184 UK traders between November 2004 and March 2013. All these trades come from spread trading on the following instruments: the UK FTSE100 index, German Dax index, Euro Dollar exchange rate and Pound Dollar exchange rate. These instruments represent 60% of the spread trading market.

Spread traders can use four different 'trade channels' to open or close trades: web (online platform), SMA, phone (calling the spread trading broker), and automatic orders. The latter can be either "traders' orders" (automatic trades using triggers set by traders) or "system orders" (automatic trades using triggers set by the broker's system: e.g. stop loss orders to help avoid trader insolvency). The study is focus on the trade channel that is used to close positions, as it is only at the time of closing a position that a profit or loss is realised. Consequently, this is when the traders' decisions involve a direct monetary impact.

Traders were divided into two categories: 'Users' — those traders who have used, at least once, the SMA technology to close a trade (995 traders) and 'non-Users' (4189 traders) — those who have never used this technology to close a trade. The SMA trading tool was introduced on 6th October 2010. Consequently, it was possible to identify trades made when SMA was available and when it was not. This allowed to compare the performance of individual traders within the SMA user group before and after the SMA trading tool was introduced, thus facilitating the testing of the 'impact of SMA' hypotheses.

4.4.2 Measuring the quality of an individual's trading decisions

Four metrics of traders' discipline and performance are employed, as these are commonly used in the financial industry to distinguish high and low performing traders: rate of return, long term profitability, DE and Sharpe Ratio.

Rate of return (Return): Measures the performance (profit/loss) of a trader's decisions in monetary terms; the daily average returns for trader i on trading day *d* using trading channel $j = \{0: non - users, 1: SMA\}$ being defined as follows:

$$\mathrm{R}_{ijd} = \frac{\sum_{k \in K_{ijd}} PnL_{ijk}}{\sum_{k \in K_{ijd}} Stake_{ijk}} \quad (4.1)$$

where, K_{ijd} is the set of trades placed by trader i using channel j on day d, PnL_{ijk} is the profit/loss in British pounds when trade k is closed by trader i using channel j and $Stake_{ijk}$ is the stake in British pounds for trade k placed by trader i using channel j on trading day d.

Long-term profitability: The long-term profitability measures the total profit or loss of trader i across their trading history. The long-term profitability is given by:

$$Profitability_i = \sum_{k \in K_i} PnL_{ik} \quad (4.2)$$

Sharpe Ratio: Higher returns can generally be achieved by taking greater risk. Consequently, it is argued that to truly measure a trader's performance one needs to adjust their raw returns by the amount of risk they run to secure these returns. The Sharpe ratio provides such a measure (Sharpe, 1994), defined as follows for trader i using channel j on day d:

$$SharpeRatio_{ijd} = \frac{AvgTrReturn_{ijd}}{STDTrReturn_{ijd}} \quad (4.3)$$

where $AvgTrReturn_{ijd}$ and $STDTrReturn_{ijd}$ are the per trade mean and standard deviation return of trader i using channel j on date *d:*

$$TrReturn_{ijd} = (PnL_{ijk}/Stake_{ijk}) \text{ given } k \in K_{ijd} \tag{4.4}$$

$$AvgTrReturn_{ijd} = \frac{1}{|K_{ijd}|} \sum_{k \in K_{ijd}} TrReturn_{ijd} \tag{4.5}$$

$$STDTrReturn_{ijd} = \sqrt{\frac{1}{|K_{ijd}| - 1} \sum_{k \in K_{ijd}} (TrReturn_{ijd} - AvgTrReturn_{ijd})^2} \tag{4.6}$$

where $|K_{ijd}|$ is the number of trades closed by trader i using channel j on date *d.*

Disposition effect (DE): is a widespread bias amongst financial traders and its incidence is commonly employed as a measure of trading discipline. This research adapts the Dhar and Zhu (2006) metric of the DE for measuring traders' performance. This metric seeks to assess the relative propensity to realise losses and gains. Given that spread traders generally open and close positions frequently, usually several times each day, this propensity was computed minute by minute by examining the average market price during each minute a position was held. It was possible to track every trader's position, because each trade and operation (buy and sell) are linked with unique identifiers. Therefore, the following definition is presented:

$$PaperGains_{ijd} = \sum_{k \in K_{ijd}} Time\ in\ Profit_{ijk} \tag{4.7}$$

$$PaperLosses_{ijd} = \sum_{k \in K_{ijd}} Time\ in\ Loss_{ijk} \tag{4.8}$$

where $Time\ in\ Profit_{ijk}$ is the time in minutes when trade k was in profit between the opening time and the minute before the closing time, $Time\ in\ Loss_{ijk}$ is the time in minutes when trade k was in loss between the opening time and the minute before the closing time and K_{ijd} is the set of trades closed by the trader i using channel j on day d.

Additionally, the following ratios are defined:

$$PG_{ijd} = \frac{Realised\ Gains_{ijd}}{Realised\ Gains_{ijd}\ +\ PaperGains_{ijd}} \tag{4.9}$$

$$PL_{ijd} = \frac{Realised\ Losses_{ijd}}{Realised\ Losses_{ijd}\ +\ PaperLosses_{ijd}} \tag{4.10}$$

where $Realised\ Gains_{ijd}$ and $Realised\ Losses_{ijd}$ count the number of minutes in which trades were closed in profit or loss respectively, by the trader i using channel j on day d. For example, if on day d, trader i closed three trades in profit using channel j, then $Realised\ Gains_{ijd}$= 3, whether or not the trades are closed in different or the same minutes of the day. Consequently, PG_{ijd} and PL_{ijd} measure the propensity to quickly close winning or losing positions, respectively.

The DE displayed by trader i using channel j on day d, is then measured by the difference between the propensity for closing positions in profit and the propensity for closing positions in loss, as follows:

$$DE_{ijd} = PG_{ijd} - PL_{ijd} \qquad (4.11)$$

Therefore, the DE_{ijd} will be positive/negative when a trader tends to realise positions in profit/loss more quickly than positions in loss/profit. In the computation of the DE, the duration time of a trade is measured in minutes. Trades with durations less than one minute were excluded from analysis, as the spread trading firm from whom the data was secured, indicated that these were likely to have been computer generated using trading algorithms.

DE_{ijd} measures the DE displayed by a trader on a given day. An additional measure is introduced, aimed at capturing the degree to which a trader displays the DE over their trading history, across all trading channels. To achieve this, it is measured:

$$PaperGains_i = \sum_{k \in K_i} Time\ in\ Profit_{ik} \qquad (4.12)$$

$$PaperLosses_i = \sum_{k \in K_i} Time\ in\ Loss_{ik} \qquad (4.13)$$

where: $Time\ in\ Profit_{ik}$ is the time in minutes when trade k was in profit and $Time\ in\ Loss_{ik}$ is the time in minutes when trade k was in loss, and K_i is the set of all trades closed by trader i. Also, it is calculated:

$$PG_i = \frac{Realised\ Gains_i}{Realised\ Gains_i\ +\ PaperGains_i} \qquad (4.14)$$

$$PL_i = \frac{Realised\ Losses_i}{Realised\ Losses_i\ +\ PaperLosses_i} \qquad (4.15)$$

where $Realised\ Gains_i$ and $Realised\ Losses_i$ count the number of minutes in which trades were closed in profit or loss respectively, by the trader i across the trading channels and traded days. Consequently, PG_i and PL_i measure the long-term propensity to quickly close winning or losing positions, respectively.

The long run DE exhibited by trader i is then measured by the propensity to realise winning positions more quickly than losing positions across i's trading history, as follows:

$$LR\_DE_i = PG_i - PL_i \tag{4.16}$$

When testing hypotheses 2 and 3 this study identifies those traders who achieve higher returns and higher risk adjusted returns, as those with better performance. When testing hypothesis 3 the study identifies those traders with better trading discipline, as those displaying lower levels of DE.

4.4.2.1 Trading Performance Modelling

The study initially tests the 'Demographic characteristics' hypothesis (H1) by exploring whether there are demographic differences between those who do and do not use SMA. This is achieved by adopting the logistic regression model (Cox, 1958) with demographic variables (e.g., Age and Gender) as independent variables and a *Segment* variable, taking the value 1 for SMA users and 0 for non-users as a dependent variable. If it is observed that the coefficients of any of the independent variables are statistically significant this will imply that users and non-users of SMA differ in terms of that particular characteristic. This will suggest that it is necessary to control for these variables when comparing the performance of users and non-users of SMA, to eliminate demographic factors as a possible cause for any performance differences.

To test the 'performance of SMA traders' hypothesis (H2), the study compares the performance of non-users of SMA with the performance of SMA users, when they are employing no SMA based channels to trade (traditional channels). To test the 'impact of SMA on trading performance' hypothesis (H3), the research compares the trading discipline and performance of users of SMA when they do and do not employ SMA technology.

There are two potential sources of performance and trading discipline variation among traders. The first arise from random variations (in terms of returns, Sharpe ratio and DE), as elements such as stress could impact traders' decision processes. These are known as ***specific effects.*** The second source of variations, known as ***main effects,*** are not random, and may arise from factors that could affect traders' performance, including, for example, age, income and whether or not the individual uses SMA to trade. (Laird and Ware, 1982; Breslow and Clayton, 1993; Verbeke, 1997; Bolker *et al.*, 2009).

Three longitudinal analyses are conducted using linear mixed models (LMM) to test H2 and three to test H3, each with one of the following dependent variables: returns (given by (4.1)), Sharpe ratio (given by (4.3)) and DE (given by (4.11)). LMM are adopted as the data relates to the trading records of individual traders, i.e. each trader placing several trades, commencing and ceasing trading at different times. In such circumstances, it is very likely that the observations for a particular trader are correlated. LMM are designed to cater for these situations, allowing assessing the impact of SMA on traders' performance and trading discipline, after controlling for other potential causes of variation. Following the notation of Laird and Ware (1982), the LMM used for this research are introduced as follows:

Let $\boldsymbol{X}_i$: $(n_i \times p)$ be a matrix with the values of the p control variables, corresponding to the main effects, for each trader i, where n_i is the number of observations of returns, DEs and Sharpe ratios, for trader i. Models of the following form are estimated:

$$\boldsymbol{Resp_i} = \boldsymbol{Z_i} a_i + \boldsymbol{X_i \alpha} + \boldsymbol{\varepsilon_i} \tag{4.17}$$

where $\boldsymbol{Resp_i} = \{Resp_{ijd}\}'$ is one of the dependent variables (Return, Sharpe ratio or DE) for trader i during day d using channel j. $\boldsymbol{Z}_i$ is a $(n_i \times 1)$ vector with specific effects (described above), for trader i; a_i is the coefficient for the specific effects for the trader i, where a_i is assumed random with $a_i \sim N(0, \theta_i^2)$ and independent, θ_i^2 is the variance of the specific effects within trader i. $\boldsymbol{\alpha}$ is a $(p \times 1)$ vector with the main effects arising from the external factors (described above) represented by the control variables (these effects are assumed to be fixed); $\boldsymbol{\varepsilon_i}$ is a $(n_i \times 1)$ vector with the errors, where $\boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \mathbf{R}_{\boldsymbol{i}})$ (multivariate normal with vector of means $\mathbf{0}$ and $\mathbf{R}_{\boldsymbol{i}}$ as $(n_i \times n_i)$ positive definite covariance matrix).

4.4.3 Control variables

The aim is to determine whether the use of SMA affect traders' performance and trading discipline. To achieve this goal, it is important to eliminate or control for the effect of other factors that may influence trading performance. Previous research suggests that there several such factors. For example, it has been shown that traders' experience and learning reduces the degree to which they display the DE, and, as a result improves their trading performance (Feng and Seasholes, 2005; Dhar and Zhu, 2006; Greenwood and Nagel, 2009; Linnainmaa, 2011). Thus, Dhar and Zhu (2006) and SeruShumway and Stoffman (2010) used number of trades and duration in years of trading activity as proxies for a trader's experience. In the same way, BarberLeeLiu and

Odean (2014) presented evidence that a trader's past performance helps predict her/his future performance. However, Barber and Odean (2000) found that more active traders underperform the market index by more than the average trader. Therefore, the study used the number of days traded up to date and the total trades placed up to date, to control the learning effect.

Clearly, trading experience does need to be controlled but proxies based only on trading days may give an incomplete picture of the trader's experience. Consequently, Greenwood and Nagel (2009) suggest that trader's age can also be used as a proxy for experience. Other factors might also affect trader performance. For example, Dhar and Zhu (2006) found that wealthier, professional and older individuals are less subject to DE and tend to demonstrate better trading performance. Similar results were obtained by Yamaguchi (2006), Anderson (2013) and Korniotis and Kumar (2010). Gender has also been shown to influence trading performance, Barber and Odean (2001) reporting that men trade more frequently but have lower returns than women.

The company who supplied the trading data disclosed that traders often provide inaccurate income and profession data. Consequently, proxies for income related factors are constructed. In particular, the study employed the UK Deprivation Index of the trader's postcode, provided by the ONS (2010), as a proxy for identifying traders with higher incomes and those with access to a better education. In addition, the initial deposit that a trader lodged with the spread trading broker is used as a proxy for income/wealth. Also, it is relevant to control, at least to some extent, for the risk attitude of each trader. As a broad indication of this, the study determined for each separate day, an individual traded the ratio of the average stakes of that individual on a given day and his/her initial deposit. A higher value suggests that the trader on that day was willing to risk more of the funds that he/she had made available for trading.

Volatile market conditions can make more difficult to predict the direction of market movement. Consequently, this paper follows Barber and Odean (2000); BarberOdean and Zhu (2009b) and Kelley and Tetlock (2013) by incorporating a market volatility variable in the models to discount market volatility effects on traders' performances (this measure is described in Appendix B).

Based on the results of the previous research discussed above, a set of proxies was developed, which allowed controlling for the variety of factors that could influence a trader's performance, such as wealth, experience, gender and trading frequency. These factors are summarised in Table 4.1.

The dummy variables *Segment, Period* and *Trade Channel* are also employed. *Segment* takes the value 1 if an individual has at some time used SMA and 0 otherwise. Similarly, *Period* takes the value 1 for the period after, and 0 for the period before, SMA was introduced (i.e. after and before 6th October 2010). *Trade Channel* takes the value 1 when the trader closes a position via SMA, 0 otherwise. With the control variables shown in Table 4.1, each individual's trading performance on a daily basis is computed, yielding 566,909 observations of returns, Sharpe ratio and DE.

Variable	**Type**	**Description**	**Controlling**
Age	Demographic	Trader's age at the trade date	Trader experience
Trade Channel	Behavioural	Dummy variable for trades closed using SMA (1 = trading with SMA, 0 otherwise)	Use of SMA to trade
Segment	Behavioural	1 if the trader is a SMA user (i.e. has traded using SMA at least once) and 0 otherwise	Users and non-users
Period	Dummy	0 for "before" and 1 for "after" the SMA technology was introduced	Period when SMA is available
AvgStake/InitDeposit	Behavioural	Sum of daily Stakes / Trader's Initial Deposit	Trader's risk appetite
Initial Deposit (GBP)	Demographic	Trader's initial deposit in GBP (as a proxy for income)	Income
DepvIndex	Demographic	Deprivation Index of the trader post code (as a proxy for wealth and lifestyle). Lower Deprivation index = more deprived areas.	Wealth and lifestyle.
Gender	Demographic	Male / Female	Gender effect
Days Traded UTD	Behaviour	Number of days that a trader has been trading up to date.	Trader experience
Total Trades UTD	Behaviour	Total trades that a trader has closed up to date.	Trader experience
MarketVol	Market	Market volatility of the trader's portfolio during the day the trade is closed.	Market effect

Table 4.1 Control variables employed in the traders' performance modelling

4.4.4 Evaluation Methods

To test hypotheses H1-H3, the study undertook a four-step evaluation process:

Step 1: Statistical summary: Initially, a statistics summary of the data is calculated and it is investigated, via linear regression, the relationship between the levels of a trader's DE and their returns. The aim of the latter analysis is to confirm that the DE has the same impact on spread traders than on individual investors in the stock markets.

Step 2: Determining appropriate control variables: H1 is tested by assessing whether there are demographic differences (see Table 4.1) between SMA users and non-users. Second, it is assessed whether market volatility and the behavioural variables, discussed in 4.4, impact traders' performance. Any variables shown to be significant are then controlled for when testing H2 and H3.

Step 3, Comparing trading performance of SMA users and non-users: First, the study compares SMA users and non-users throughout their trading history. Second, the study compares SMA users and non-users when they employ traditional trading channels. Consequently, these analyses will determine whether there are performance differences between SMA users and non-users, irrespective of their use of SMA (i.e. to test H2).

Step 4, Impact of SMA on trading performance: This step analyses the trading records of SMA users only, comparing their performance when they do and do not use SMA (to test H3).

4.5 Results

4.5.1 Statistical summary

A comparison of the cumulative distribution functions (CDF) of the returns, Sharpe ratio and DE metrics for trades closed using SMA and other trading channels are presented in Figure 4.1. It is clear from a) and b) that the proportion of trades with higher returns and higher Sharpe ratios is greater for those trades closed using SMA (cf. traditional channels). However, Figure 4.1 c) indicates that a greater proportion of trades closed via SMA are associated with a higher DE.

Summary statistics associated with each of the potential control and dependent variables are presented in Table 4.1. It is clear from these figures that there is wide diversity in terms of all these variables. The extremely high variability in the returns and Sharpe ratios, associated with trades executed in a given day, demonstrates the high risk associated with these markets. The wide ranges of values for traders' age, the deprivation index associated with their residence, their initial deposits, the days traded, the ratios of their average stakes to initial deposit and DE displayed in a given day, suggest a wide range of traders with different behavioural characteristics and levels of experience or that these vary significantly throughout an individual's trading history.

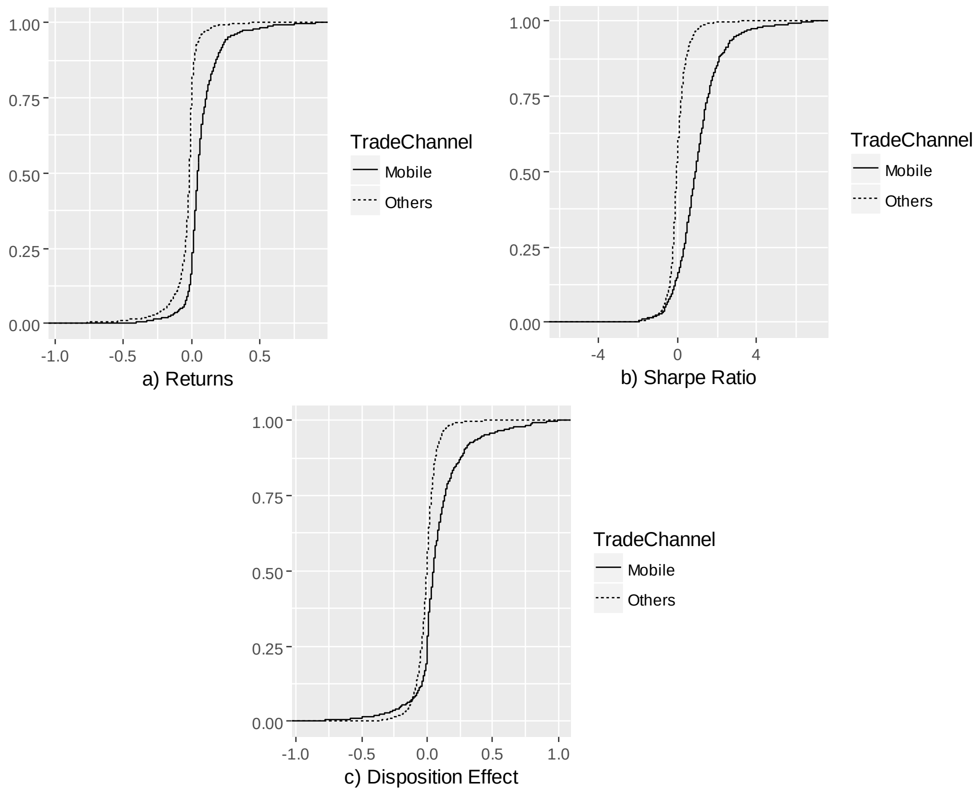


Figure 4.1 Cumulative probability distributions of returns, Sharpe Ratio and DE

Variable	**No. of obs.**	**Minimum**	**Maximum**	**Average**	**Std. dev.**
Returns	566909	-12.608	11.619	-0.007	0.175
Sharpe Ratio	566909	-85,735,612	80,351,481	-329.4	1,822,569
DE	560441	-0.999	1.000	0.010	0.198
AvgStake/InitDep	566909	0	20,000	11.26	224.39
Days Traded UTD	565195	1	1,545	223.37	224.60
Total Trades UTD	565195	2	64,195	1609	2,940
Market VolDaily	563093	0.003	0.792	0.097	0.076

Table 4.2 Summary statistics for potential control variables associated with daily trades of individual traders

Variable	**No. of obs.**	**Minimum**	**Maximum**	**Average**	**Std. dev.**
Initial Deposit	5,184	£500	£160,589	£941.24	£3556.19
Depriv. Index	5,184	0 (high deprived)	80.0 (low deprived)	17.32	13.04
Age	5,184	18	90	45.56	13.11
Gender (Females)	511 (10%)				
Gender (Males)	4673 (90%)				

Table 4.3 Summary statistics of the demographic variables

The Pearson correlation values for the variables in Table 4.2 were computed. There is low correlation between most of the variables (the majority < |0.18|), other than between Total Trades up to date (UTD) and Days Traded UTD (correlation = 0.636). To avoid multi-collinearity, these two correlated variables are combined in the analysis, using the interaction term Total Trades UTD x Days Traded UTD.

The influence of DE on traders' performance is explored using linear regression. In particular, the relationship between the levels of a trader's DE and their returns is examined: first for a given trading day and second over the individual's entire trading history. The results are displayed in Table 4.4. Interestingly, the level of DE is positively correlated with the level of returns for trades placed in a given day. This is perhaps not surprising, as waiting until a trade turns profitable before closing that trade, will often produce positive returns on a given day.

	Target variables			
	Return		Long-term profitability	
Num. of observations	566,909		5,184	
	Variables	**Coefficient**	**Variables**	**Coefficient**
	Intercept	0.0214	Intercept	0.0315
	DE	0.1624***	LR_DE	-0.4069**

* significant at the 0.1 level, ** significant at the 0.05 level, *** significant at the 0.001 level

Table 4.4 Results for linear regression exploring the relationship between the daily DE displayed by a trader and their daily return and long-term profitability.

However, a linear regression exploring the relationship between a trader's long run DE (LR_DE) and the profit they secure across their trading history shows that a high DE is associated with lower profitability (see Table 4.4). This suggests that high DE in the long run leads to losing positions being held too long and eventually forcing the trader to close a trade with large losses.

4.5.2 Demographic characteristics

A logistic regression is fit with *Segment* (1 for users of SMA and 0 otherwise) as the dependent variable and the control variables, designed to capture traders' demographic information, as the independent variables. The results of this regression are displayed in Table 4.5. These results show that SMA users are more likely to be male, younger and to live in a more deprived area (*Gender*, *Age* and *Deprivation index* variables are statistically significant).

Logistic Regression model		***Segment* = 1 (users of SMA)**	
Num. of observations:		**5,184 traders**	
Independent Variables	**Categories**	**Coefficient**	**Standardised Coefficient**
Intercept		0.6039	
Initial Deposit		-0.00001	-0.0216
Deprivation Index		-0.00623	-0.0445**
Gender	Female=1	-0.1338	-0.044**
Average Age		-0.047	-0.3398***

* significant at the 0.1 level, ** significant at the 0.05 level, *** significant at the 0.001 level

Table 4.5 Demographic differences between SMA Users and Non-Users.

An odds ratio analysis derived from Table 4.5 suggests that men are 30% more likely to use SMA than women and a person 10 years younger is 60% more likely to use SMA (further details in Appendix C). Surprisingly, the results also suggest that SMA users are less likely to live in wealthier areas. This can be explained by Tsetsi and Rains (2017), who found that younger, lower income, and less educated SMA users (i.e. expected living in deprived areas) are more likely to be smart mobiles dependent.

4.5.3 Trading performance with control variables

Three Linear Mixed Models (LMM) are fit with returns, Sharpe ratio and DE as dependent variables. The demographic variables, experience, risk preference and market-related control variables (introduced in section 4.4.3) are used, as independent variables. The results, presented in Table 4.6 show that most of the independent variables have a significant impact on traders' performance and trading discipline. Variables related to traders' demographic characteristics, experience and risk appetite and market conditions (volatility) all impact trading performance. This analysis confirms the need to control for these trader's characteristics and market conditions, when testing the hypotheses, concerning the relative trading performance of SMA users and non-users.

Target Variable		**Returns**	**Sharpe Ratio**	**DE**
Num. of observations			**566,909**	
Variables	**Levels**		**Coefficient**	
Intercept		0.0142	-0.0369	-0.0456
Initial Deposit		0.0280***	0.0395***	0.0285***
AvgStake/InitDep		0.0128	0.0703***	0.0347**
Depriv. Index		0.0016	0.0108***	0.0023
Gender	Female	0.0096	0.0520***	-0.0213*
Age		0.0049***	0.0118***	-0.0169
Days Traded UTD x Total Trades UTD		0.0016*	0.0071***	-0.0040***
MarketVol		0.0036***	-0.0036***	0.0025*

* significant at the 0.1 level, ** significant at the 0.05 level, *** significant at the 0.001 level

Table 4.6: Results of estimating linear mixed models with returns, Sharpe Ratio and DE as dependent variables to determine impact of potential control variables on trader performance and discipline.

4.5.4 Comparative trading performance of users and non-users of SMA

Similarly, three LMM are fit with returns, Sharpe ratios and DE as dependent variables to identify differences in the trading performance of users and non-users of SMA. These models control for the demographic, risk appetite, experience and market volatility factors that influence trading performance. The variable *Segment* is included in the LMM as an independent variable, to enable to determine whether users and non-users of SMA differ in their trading performance. The results of this analysis are presented in Table 4.7.

The coefficients for the variable *Period* (which takes the value 1 for the period after the introduction of SMA and 0 otherwise) in the regressions with returns, Sharpe ratios and DE as dependent variables, were positive and significant. Coefficients for SMA users (i.e. *Segment* = 1) took even greater positive values. These results demonstrate that following the introduction of SMA, the returns and Sharpe ratios of traders as a whole increased significantly, as did their DE. These changes were particularly pronounced for SMA users. The greater returns and higher Sharpe ratios achieved by users of SMA may have arisen because they have inherently superior trading skills and/or because of the advantage that SMA affords. Similarly, the higher levels of DE of SMA users, may have arisen because these traders generally hold on to losses relatively longer than gains or because the use of SMA encourages this behaviour.

Target Variable		**Returns**	**Sharpe Ratio**	**DE**
Num. of observations			**566,909**	
Variables	**Levels**		**Coefficient**	
Intercept		0.0074	-0.0613	-0.0564
Initial Deposit		0.0282***	0.0401***	0.0288***
AvgStake/InitDep		0.0130	0.0722***	0.0355**
Depriv. Index		0.0017	0.0103**	0.0027
Gender	Female	0.0104	0.0547***	-0.0197*
Age		0.0057***	0.0096**	-0.0148***
Days Traded UTD x Total Trades UTD		0.0005	0.0037**	-0.0052***
MarketVol		0.0039***	-0.0026**	0.0029*
Period	1: SMA available	0.0073***	0.0241***	0.0075**
Segment	1: SMA user	0.0194***	0.0761***	0.0381***

* significant at the 0.1 level, ** significant at the 0.05 level, *** significant at the 0.001 level

Table 4.7: Results of estimating linear mixed models with returns, Sharpe Ratio and DE as dependent variables in order to test if traders improved their performance after SMA was introduced, and to compare performance and trading discipline of SMA users and non-users

4.5.5 Trading performance comparison with traditional trading channels.

To test the 'performance of SMA traders' hypothesis (H2), the study compares the trading performance of SMA users and non-users when employing traditional trading channels (i.e. non SMA channel). The aim is to compare the skills of these two groups of traders, having eliminated any advantage afforded by the use of SMA and having controlled for demographic, experience and risk appetite differences and market conditions. Once again three LMM are estimated with returns, Sharpe ratio and DE as dependent variables, and the results are presented in Table 4.8.

Target Variable		Returns	Sharpe Ratio	DE
Num. of observations			556,316	
Variables	levels		Coefficient	
Intercept		0.0121	-0.0545	-0.0534
Initial Deposit		0.0299***	0.0421***	0.0294***
AvgStake/InitDep		0.0167*	0.0759***	0.0360**
Depriv. Index		0.0017	0.0099**	0.0030
Gender	Female	0.0106	0.0547***	-0.0199**
Age		0.0063***	0.0077**	-0.0142***
Days Traded UTD x Total Trades UTD		-0.0012	2.5E-05	-0.0063***
MarketVol		0.0027**	-0.0027**	0.0022
Period	1 SMA available	-0.0047***	0.0078***	-0.0004
Segment	1 SMA user	0.0013	0.0452***	0.0269***

* significant at the 0.1 level, ** significant at the 0.05 level, *** significant at the 0.001 level

Table 4.8: Comparison of the trading performance of the users and non-users of SMA when using traditional trading channels only

The most striking finding, based on the coefficients of ‘Segment’ (1 for SMA users and 0 for non-SMA users), is that SMA users achieve significantly higher Sharpe ratios than non-users, even when they use traditional trading channels. However, users of SMA users exhibit greater DE. Interestingly, there is no significant difference in the returns of users and non-users of SMA when they both employ traditional trading channels. Overall, these results provide partial support for H2, namely that *SMA users make better quality decisions than non-users, demonstrated by their superior trading performance when not employing SMA technology.* These results suggest that SMA users manage risk more effectively than non-users when both employ traditional trading channels. However, it is clear that this does not stem from their better management of the time they hold losses compared to the time they hold gains (DE).

4.5.6 The performance advantage stemming from SMA

Having established that SMA users are better at managing risk than those who do not use SMA (i.e. achieve enhanced *Sharpe ratios*), even when they trade using traditional channels, a test is set out to determine if SMA itself affords users any advantage. To achieve this, the study compares the performance of SMA users when they do and do not trade using SMA, following the introduction of this technology. To achieve this, the variable *Trade Channel* is introduced, which takes the value 1 when the trader closes those positions via SMA, 0 otherwise. The results of estimating the LMM are presented in Table 4.9.

The key finding to emerge is that the coefficients of the variable *Trade Channel*, are positive and statistically significant at the 0.001 level in each of the LMM. This suggests that traders achieve significantly higher returns and Sharpe ratios but have significantly higher DE when they trade using SMA compared with when they trade using traditional channels. This result provides support for H3, that SMA technology has a positive impact on their trading performance. However, the results suggest that trading using SMA also significantly reduces their trading discipline.

Target Variable		Returns	Sharpe Ratio	DE
Num. of Observations			79,082	
Variables	levels		Coefficient	
Intercept		-0.0076	-0.0233	-0.0350
Initial Deposit		0.0361***	0.0631***	0.0422**
AvgStake/InitDep		0.0234	0.0835	0.0373
Depriv. Index		-0.0001	0.0189	0.0101
Gender	Female	0.0088	0.0763*	-0.0003
Age		0.0079	0.0209	-0.0163
Days Traded UTD x Total Trades UTD		-0.0082***	0.0018	-0.0174***
MarketVol		0.0073**	-0.0257***	-0.0054
Trade Channel	1: SMA is used	0.3617***	0.5393***	0.2245***

* significant at the 0.1 level, ** significant at the 0.05 level, *** significant at the 0.001 level

Table 4.9 Comparison of the performance of SMA users when the app is and is not used to trade

4.6 Discussion

4.6.1 Similarities between spread traders and investors in traditional financial markets

The results suggest spread traders and traditional financial market investors share some similar characteristics. In particular, it was shown that DE has a negative influence on spread traders' long term performance, and this accords with the results of Barber and Odean (2000) and BarberLeeLiu and Odean (2009a) in relation to individual stock market investors. In addition, more experienced spread traders achieve higher returns, exhibit greater risk control (i.e. higher Sharpe ratios) and are less subject to the DE. Similar results have been obtained in a range of studies examining investors in more traditional financial markets (e.g., Dhar and Zhu (2006); Feng and Seasholes (2005); Linnainmaa (2011); Greenwood and Nagel (2009)). It was also found that female spread traders are generally less subject to the DE and exercise greater risk control than male traders. This result is in line with Barber and Odean (2001) who examined investors in more traditional markets. Interestingly, Feng and Seasholes (2008) found that male investors in stock markets in China are generally less subject to the DE and exercise greater risk control than female investors. The discrepancy may arise from cultural differences since both Barber and Odean (2001) and the case study investigated here analyse largely western investors.

In summary, these results suggest that there are many similarities in the behaviour of spread traders and investors in wider financial markets, suggesting that the analysis conducted in this paper, in relation to the effect of SMA on trading decisions, has wide applications.

4.6.2 Distinguishing characteristics of SMA users

The analysis suggests that there are significant demographic differences between those who do and those who do not choose to use SMA. In particular, it was found that those who use SMA tend to be younger and are more likely to be male. These characteristics are in line with VenkateshThong and Xu (2012)'s theory related to the nature of early adopters.

Additionally, it was found that SMA users are less likely to live in wealthier areas. This can be explained by Tsetsi and Rains (2017), who found that younger, lower income, and less educated SMA users (i.e. expected living in deprived areas) are more likely to be smart mobiles dependent.

4.6.3 The impact of SMA on trading behaviour and performance

Having controlled for factors that were correlated to an individual's trading behaviour and performance (i.e. experience, risk appetite, demographic factors and market conditions), it was found that the advent of SMA resulted in traders' performance improving (in terms of their returns and Sharpe ratios). This result seems to be aligned with the main advantages of this new technology: mobility and accessibility of accurate and up to date data, allowing individuals to execute decisions almost whenever and whatever they are (Sraeel, 2006; Koenig-LewisPalmer and Moll, 2010; Kourouthanassis and Giaglis, 2012; BuijinkVisser and Marshall, 2013; ConstantiouLehrer and Hess, 2014).

The research also found that SMA users achieve higher Sharpe ratios but also display greater DE than those who do not use SMA, even when they both use traditional trading channels. This suggests that those who choose to use SMA possess superior trading skills (at least in terms of risk control), even after controlling for factors which impacts trading behaviour and performance (i.e. demographic characteristics, experience, risk appetite). This could be explained by the digital divide effect, as previous research suggest that technology splits people into groups with different decision making skills (Forman, 2005; FormanGoldfarb and Greenstein, 2005; AroraForman and Yoon, 2010; Hui and Png, 2015). It is interesting to note that the users of SMA did not achieve significantly higher returns than those who did not use SMA, when both use traditional trading channels. This suggests that the superiority of trading skills of SMA users (as displayed by their significantly higher Sharpe ratios) is related to the control of risk. However, this risk control does not extend to their trading discipline, as they display a greater tendency to hold losses and to sell gains than non SMA users.

The greater Sharpe ratios achieved by SMA users may have arisen because they have superior trading skills and/or because SMA affords those who use it some advantage. Consequently, the trading performance of SMA users was specifically compared when they did and did not use SMA. This analysis revealed that they achieved significantly higher returns and a higher Sharpe ratio when they traded using SMA (cf. when using traditional channels). Interestingly, they were also subject to greater levels of DE when they traded using SMA. It might be felt that the improved returns and Sharpe Ratio of traders using SMA could stem from special features of this technology that provided decision support. However, the SMA had no such features and interviews with the developers of the SMA (used by those spread traders) revealed that they had not designed them to provide any additional decision support mechanisms. It appears, therefore, that users of SMA achieve their advantage largely through their capacity to access current information and to be able to react quickly to market conditions. Alternatively, this advantage may come from the more limited information provided by the SMA. Arguably, this could improve decisions as it may prevent information overload (JacobySpeller and Berning, 1974; Jacoby, 1984; Eppler and Mengis, 2004; Bawden and Robinson, 2009). The precise cause requires further investigation.

The results from this research should be robust since they were derived from a large longitudinal study that examined 4.5 million trades of 5184 individuals over a 9-year period. The scale of this investigation helped overcame many of the limitations of the study conducted by ConstantiouLehrer and Hess (2014), where the behaviour of only 56 individuals was investigated. Furthermore, the fact that spread traders generally trade frequently and are likely to only gather information from the SMA prior to trading, means that their behaviour is less likely to be influenced by other sources of information than may occur in other studies. For example, ConstantiouLehrer and Hess (2014) point out that a person who is using a location-based service SMA, to find a place to eat for instance, can also interact with other information sources, such as friends and street ads. It is then difficult to establish the extent to which the decision was influenced by the SMA or other information sources. In the study conducted in this paper, traders can see their potential loss or profit on the SMA screen before they decide to close the position. Additionally, given the fact that markets move very rapidly, SMA users are unlikely to look for information from other sources, when they decide to realise a position.

Nakasumi (2012) argued that one of the main obstacles to appraise the decisions of SMA users more generally is the fact they have to face numerous options. Furthermore, elements such as preferences, desires and colours can affect the users' decision. By contrast, the decision facing a spread trader using an SMA is fairly simple: either to take the profit or loss associated with position or to continue to hold that position. This reduces the influence of the factors identified by Nakasumi (2012) which can make more difficult to discern the influence of the SMA.

4.7 Conclusions

This study explored whether smart mobiles and their apps (SMA) can improve individuals' trading performance. Using a large dataset from the spread trading market, it was revealed that this technology does have a significant positive impact on the returns for those investors who choose to use SMA (users) and their ability to control risk (as evidenced by higher Sharpe ratios), but also increases their tendency to display the disposition effect.

Previous studies examining the impact of new technology on decision-making have not generally explored situations where it is possible to measure, in an objective fashion, the degree of such impact. By contrast, this research focuses on the actions of spread traders, which enables to identify clear changes in the decision bias of individual traders (as measured by the DE) and to measure the monetary implications of their changed behaviour (via their returns and Sharpe ratios).

This research finds evidence that there are clear demographic differences between those who do and do not choose to use SMA, and it is found that SMA users display greater risk control. In addition, the study finds that SMA users improve their performance (in terms of returns) and their risk control (in terms of improved Sharpe ratios), however at the same time exercise less trading discipline (in terms of the DE) on the occasions when they trade via SMA compared to when they trade using traditional channels (i.e. phone calls and web). Consequently, the study concludes that the advantages that traders obtain from SMA stem from the mobility and accessibility that this technology provides; facilitating access to current information and the ability to trade anytime and anywhere. In addition, this technology could be reducing the level of data overload faced by traders at the moment they make their decisions. Nevertheless, the study also concludes that this mobility, accessibility and reduction of data overload may increase biases such as the disposition effect.

Migration from internet connections toward mobile platforms and SMA is becoming increasingly popular (Index, 2015). Consequently, it is likely that more retail brokers will offer this technology as a trading channel to their clients. For example, the company that provided the data for this study is planning to migrate its entire trading platform to SMA channels. The greater accessibility of SMA, as more internet connections are moved to mobile platforms, and its increasing popularity are likely to lead to more investors across a range of financial markets beginning to trade via SMA.

Beside its popularity, SMA has limitations as a trading platform: One of these, for instance, is the size of the device's screen, which restricts access to other sources of information. Thus, in more complex financial markets, where investors have to gather more information before they trade, the impact of using SMA may not be the same as in spread trading market.

Nevertheless, this investigation found similarities between spread traders and investors in other markets, as they share common characteristics. Consequently, the results reported can probably be extrapolated to those markets where high frequency traders participate (i.e. noise traders), which follows the same underlying motivations of spread traders. BarberOdean and Zhu (2006) show evidence that noise traders move the market prices. Therefore, the outcomes of this investigation suggest that the use of SMA in the financial markets are likely have an important impacts on these markets, perhaps with a move towards greater market efficiency.

On contrary, the ease of accessibility of SMA will also have repercussions for the regulation of financial markets. It is likely, for example, that regulators will impose new rules on brokers when offering their services via SMA, as it is shown that this technology is more appealing to younger individuals. Consequently, brokers may need to engage in more exhaustive identity verification procedures. Perhaps more worryingly the growth of SMA will make cybersecurity a growing concern to regulators, as they become concerned with the security of markets accessed from a range of locations and devices.

Furthermore, these results suggest that the SMA technology may lead to greater levels of DE. This could cause destabilisation of markets if many traders continue to hold losses for long periods. This could mean that, following a sudden market crash, many more individual traders would suffer severe losses; producing perhaps a significant decrease in market liquidity and in consequence increasing the market volatility. The clear lesson to financial market operators and regulators is that further research is needed to fully understand the implication for market efficiency and stability of the wider scale use of SMA for financial market investment.

Chapter 5: Conclusions

5.1 Final remarks

This chapter presents the general conclusions and limitations of this thesis as well as potential expansions of the topics investigated.

This chapter contains five additional sections. In sections 5.2, 5.3 and 5.4 the main conclusions on the three cases studied are presented: optimisation of credit portfolios, profiling individual investors and the influence of smart mobiles and their apps in individual investors. These sections answer the research questions of this thesis (section 1.8). Section 5.5 discusses the limitations and potential research paths for future developments of the topics under scrutiny. Finally, section 5.6 summarises the general conclusions of this thesis.

5.2 Credit portfolio management

The first paper of this thesis presents a new framework for optimising credit portfolios based on a multi-objective model. In contrast with approaches that consider a single integrated objective, the multi-objective framework provides a flexible approach for enhancing decision making in credit portfolio management, whilst keeping each objective separately. This framework allows finding effective and efficient strategies, i.e. strategies that not only mitigate possible losses but also convey opportunities to increase the return of a credit portfolio.

In order to apply this framework successfully, the following information has to be available for each exposure within the portfolio: probability of default, loss given default and exposure at default. Additionally, it is relevant that the portfolio is properly segmented into sectors. Each sector should reflect a degree of homogeneity. The default correlation between the sectors is also required in order to compute the portfolio economic capital. Fortunately, this information is now routinely available within financial institutions practice.

One of the main challenges when multi-objective models are used is finding optimal solutions, particularly in cases when conventional methods cannot be applied due to the non-linearity of some of the objective functions. Even though Multi-Objective Evolutionary Algorithms (MOEA) (Deb, 2008) are suitable for solving these types of problems, one of the main drawback of this methodology is the production of several solutions. This issue is not addressed by previous research in this area (Schlottmann and Seese, 2004; SchlottmannMitschele and Seese, 2005; BrankeScheckenbachSteinDeb and Schmeck, 2009). This thesis contributes to close this gap by developing the Normalised Global Criteria Method for systematically choosing solutions in a neighbourhood with lower dispersion of their objective values. This method ultimately increases the portfolio performance in terms of risk and returns, whilst optimising the other objectives as well.

As it is mentioned above, an important characteristic of this multi-objective modelling framework is its flexibility. The multi-objective approach facilitates adding more objectives and constraints. Therefore, one extension of this work is to incorporate other risk measures such as contagion. However, it is not always easy to measure contagion unless interbank data is available.

Additionally, this framework also contributes to shift the discussion from individual loans to sectors expanding Schlottmann and Seese (2004) and Schlottmann and Seese (2005). This point of view is more suitable for decision makers dealing with the optimisation of a portfolio of non-liquid assets, as usually lenders tend to manage this type of credit portfolio by sector rather than on an individual loan basis.

5.3 Profiling individual investors using Gradient Boosting Machines (GBM)

The second paper in this thesis explores the suitability of gradient boosting machines (GBM) for profiling individual investors in the spread trading market, specifically to classify those who are most likely to achieve significant returns in the long run, which poses the greatest risk to brokers.

Datasets associated with individual investors' behaviour tend to be very heterogeneous. For example, they may contain different investors' background, age, income and risk attitude. Additionally, when researchers try to get insights on investor's behaviour, they need to scrutinise several variables associated with different investors' behavioural patterns (e.g., past returns, level of disposition effect and risk control). Therefore, when such datasets are used for classifying investors (i.e. A-Book or B-Book), it could be a significant challenge for data mining-based classification methodologies; in particular those commonly used in previous research (Barber and Odean, 2000; Barber and Odean, 2001; Grinblatt and Keloharju, 2009; Korniotis and Kumar, 2010; Korniotis and Kumar, 2011; Anderson, 2013; BarberLeeLiu and Odean, 2014).

Using a dataset with 4.5 million trades placed by 5184 UK traders between November 2004 and March 2013 participating in the spread trading market, allowed to gather evidence that suggests that GBMs could have a higher classification power compared with other data mining methods (such as logistic regression, classification trees, neural networks and support vector machines), when profiling individual investors. Consequently, using GBMs in this domain could enable researchers to obtain more accurate insights into investors' behaviour.

Additionally, this study exposes some relevant limitations of the methodologies used by previous researchers in the investigation of the individual investor's behaviour. For instance, this thesis found that traders who achieved high returns in the past may be likely to achieve high returns in the future; interestingly, it was also found that traders with very poor returns in the past may also achieve very high returns in the future. These findings contradict Barber and Odean (2011) who claim past performance could predict future performance. Moreover, this research discovered that one of the causes of this apparently contradiction is the interaction between the levels of disposition effect, age, returns and risk control. In consequence, by using GMBs capabilities, it is possible to get more insights into investors' behaviour, expanding the findings of SeruShumway and Stoffman (2010), Barber and Odean (2011) and Barberis and Xiong (2009), as GBMs allow visualising complex interactions between the traders' past performance, risk control, age and disposition effect, which are not totally considered in these previous studies.

5.4 Influence of smart mobile applications in individual decision making.

The final paper of this thesis explores the impact in individuals' decision-making of using the latest and most popular technologies such as the smartphone and their apps (SMA). In particular, the paper focuses on discovering the impact of technological aids, such as trading SMA, on the decision making of individual investors. With such a rapid increase in access to content/information through smart mobiles, a fundamental research question has emerged: "What is the impact of the use of SMA on the nature and quality of individuals' decisions?" This is of particular interest considering that SMA users make critical decisions, such as financial decisions within dynamic environments, whilst often remaining socially connected and regularly receiving mobile-based information.

Studying how external elements can influence the decision making process in individuals represents a true challenge. For instance, previous studies (Nakasumi, 2012; BuijinkVisser and Marshall, 2013; ConstantiouLehrer and Hess, 2014) have not generally explored situations where it is possible to measure, in an objective fashion, the degree of such influence. Furthermore, there are other associated elements that have to be considered and controlled, for instance individuals using SMA could outperform, simply because they may be coming from a wealthier socioeconomic background or they are more skilful. Another relevant challenge when dealing with trading information is the fact that the data usually records several transactions of the same individual, generating in consequence internal correlations; in other words each observation is not totally independent from other observations, thus violating one of the common requirements of the regression models when performing this type of analysis. Therefore, it is imperative to set up the data correctly, establish the appropriate control for factors that may impact investors' decision making and use the appropriate methodology to deal with the internal correlations.

Consequently, after setting up all the appropriate controls and the methodological framework, this research has revealed that SMA has a positive short-term impact by significantly increasing individuals' performance and risk control when participating in financial markets. At the same time, however, this technology can accelerate human biases such as the disposition effect, which subsequently produce negative effects in the long run.

Furthermore, by conducting this investigation, it was possible to obtain additional insights into how this technology can affect the decision making process of individuals. In particular, this investigation overcame obstacles when conducting research on SMA as decision aids. One of these obstacles is the interaction of decision makers with other sources of information. These interactions make it difficult to discern whether the success or failure of a decision was due the use of the technology in study (e.g. SMA), or the additional information obtained from those sources, or the combination of both. This limitation was pointed out by ConstantiouLehrer and Hess (2014) and Nakasumi (2012) as one of the main issues when conducting this type of research.

This thesis also shows how this type of investigation should be conducted, not only by selecting the appropriate methodology to mine the data, but also by controlling for potential causes that could mislead the analysis.

5.5 Limitations and further research

This section discusses the main limitations of the research conducted in this thesis. Additionally, it presents ideas for future developments. As before, this subsection is divided into three parts. The first part addresses the limitations and future developments for optimising credit portfolios. The second part presents the limitations and ideas for future research on profiling individual investors. Finally, in the last subsection, the limitations and potential research paths for investigating the influence of SMA on individual investors, are presented.

5.5.1 Limitations in the investigation of credit portfolios

Data from a commercial bank was used to research the optimisation of credit portfolios. Although, the data has more than 2000 loans, this research did not have access to their performance during previous years. In consequence, this research could not perform a comparison analysis across time while investigating the evolution of this portfolio. Hence, a potential extension would be to perform such analysis and assess the stability of the solutions arising from the optimisation process across time.

On the other hand, due the fact that this research had information from a single financial institution, it could not incorporate contagion metrics in the analysis. A possible extension of this investigation is to combine portfolios from several institutions in order to evaluate the degree of correlation between them, with the aim of producing a contagion metric, which could then be considered in the optimisation process. This type of analysis could be undertaken perhaps with the support of financial regulators and central banks.

5.5.2 Limitations in the profiling of individual investors

For profiling individual investors, this research used data from a UK based broker. Despite this being a large dataset, the data is subject to trading policies established by the broker itself, for example fees, minimum deposits, spread sizes (difference in the selling and buying prices), etc. Some of these policies changed during the data's time period, which may introduce restrictions and boundaries to the investors and, in consequence, it could produce potential bias in the data. Unfortunately, this cannot be completely controlled as the dates, when some of these changes occurred, were neither recorded nor documented by the broker.

Another potential limitation is the use of UK-based investors only, due to the lack of demographic data of investors from abroad. This fact can also introduce a potential bias in the data.

Future research in this area could focus on using different ensembles with several configurations and determining which of these configurations is more successful for profiling individual investors. Similarly, several ensembles could be used to predict traders' profit and losses.

5.5.3 Limitation in the investigation of SMA on decision making

Similarly to the profiling individual investors' case study, the data used in the influence of SMA over individual decision making research, only considers UK-based traders due to the lack of demographic information of foreign investors.

Additionally, due to the limitation of tick data, the four most popular instruments traded in the UK Spread Trading Market (Pound Dollar Exchange Rate, Euro Dollar Exchange Rate, FTSE100 Index and Xetra DAX Index) were used. An interesting expansion of this investigation could be comparing the influence of the SMA considering different types of investors; for instance, investors trading other instruments such as shares or commodities could have different risk attitudes, as these markets tend to be less volatile than index and currencies markets. Therefore, the focus could be to establish whether the impact of the SMA is different, either by the type of investors or by the type of market.

Another potential future research in this area could be to establish whether the SMA improves investors' reaction when significant changes in the market occur. It can be hypothesised that SMA positively improves the reaction of investors to adverse movements of financial markets.

5.6 General conclusions

After conducting this research, where three non-conventional decision making case studies have been investigated, the following conclusions can be drawn: first, applying appropriate selection methods can positively impact decision making processes, as it produces more simplified scenarios for decision makers. Second, although nowadays there is more computational power that enables researchers to process and perform analyses over large scale data in relative short periods of time, it is crucial to be aware of the limitations that methodologies have when undertaking such analyses. Third, as it has been demonstrated in this thesis, it is equally important to establish the necessary controls over the data itself, in order to avoid the analyses would be misled by external factors (e.g. the digital divide effect in SMA users). Hence, by taking into consideration these elements, it is possible to overcome, or at least minimise, the main difficulties when investigating decision making processes using large scale financial data.

Appendices

Appendix A Spread trading

Traders in spread trading markets effectively bet on whether the price of a particular financial instrument will go up or down. The trader decides upon the number of units (*s*) of the instrument to 'buy' or 'sell' at £1 per unit. If the instrument rises in price by *n* points following a buy trade or falls in price by *n* points following a sell trade, then s/he secures *n* x *s* profit. However, if the instrument falls in price by *m* points following a buy trade or rises in price by *m* points following a sell trade then s/he loses *m* x *s*. For example, consider a trader who believes that the FSTE100 index will increase from its current level of 5968.3. They could place a 'buy' trade of say, 100 units (at £1 per unit) and would then have opened what is referred to as a long position[17] (should the FTSE100 index fall to 0 then the trader would have lost 100 x £1 x 5968.3 = £59,683). Despite this being a 'buy' trade, no assets are purchased, as the position is completely leveraged. Should the trader decide to 'sell' 50 units when the FTSE100 is trading at 5973.8, s/he will have secured a profit on these units of £1 x 50 (units) x (5973.8 - 5968.3)/1.0 = £275. Should the trader later decide to sell the remaining 50 units when the FTSE100 is standing at 5962.3 s/he will have made a loss on these units of £1 x 50 (units) x (5968.3 – 5962.3) = £300. Overall, therefore, on the 100 units purchased the trader will have made a net loss of £300- £275 = £25.

When a financial instrument increases or decreases in price, the number of points which rise or fall is determined by its 'percentage in point' or PIP size. For example, the PIP size for the FSTE100 is 1, and the PIP size for the Euro Dollar rate is 0.0001. Consequently, a rise from 1.1574 to 1.1582 in the Euro Dollar rate is regarded as a rise of (1.1582 - 1.1574) / PIP size (i.e. 0.0001) = 8 points.

[17] A position is referred to as 'long' when the trader 'buys' the market and 'short' when the trader 'sells' the market.

Appendix B Market volatility

The volatility within a market is commonly measured as the standard deviation of the returns of that market, during a predetermined period of time (Jorion, 2011). Given the fact that spread traders can bet in instruments in different markets (e.g. FTSE100, Euro Dollar exchange rate, etc.) an algebraic combination of the different markets' prices (market index) has to be created to compute the market volatility for each trader's positions.

Therefore, a variable that combines the daily prices of these markets is constructed. To achieve this, it was created a market index based on the weighted average of the prices of the instruments in which a particular trader invests, as follows:

$$MI_{id}^{t} = \sum_{m=1}^{M_d} W_{idm} * Price_m^t \tag{B.1}$$

where: (B.2)

$$W_{idm} = \frac{TotStake_{idm}}{\sum_{m=1}^{M_d} TotStake_{idm}}$$

and MI_{id}^{t} is the market index for trader i on trading day d during minute t, $TotStake_{idm}$ is the total stakes in British Pounds bet by the trader i on instrument m on trading day d, $Price_{md}^{t}$is the market price of instrument m on day d during minute t and M_d is the number of instruments available to invest in on day.

Then the variable *MarketVol* is given by:

$$MarketVol_{id} = \sqrt{\frac{F}{T-1}\sum_{t=1}^{T}\left(MI_{id}^{t} - \overline{MI_{\iota d}}\right)^2} \tag{B.3}$$

where:

$$\overline{MI_{\iota d}} = \frac{1}{T}\sum_{t=1}^{T} MI_{id}^{t} \tag{B.4}$$

$\boldsymbol{F = 60 * 24 = 1440}$ is the factor used to convert market volatilities from minutes to days and $\boldsymbol{T}$ is the total time in minutes when the market is open to trade.

Appendix C Odds ratio

The odds ratio measures the marginal contribution that each value of a particular variable has in a classification model. For example, if it is desired to compare the relative likelihood of males and females being SMA users, the coefficient from a logistic regression model, which links *Gender* (independent) and *Segment* (dependent), is used. To establish how likely a male (cf. a female) is to become a SMA user, the odds ratio is computed as follows (Hosmer Jr and Lemeshow, 2004):

$$OddsRatio_{Male\ vs\ Female} = \frac{e^{\beta_{Gender}*Gender(male)}}{e^{\beta_{Gender}*Gender(female)}} \quad \text{(C.1)}$$

where β_{Gender} is the coefficient for the variable $Gender$, respectively, in the logistic regression

If the odds ratio from (C.1) is greater than 1 then this suggests that males (cf. females) are more likely to be SMA users. From Table 4.5, which displays the results of a logistic regression comparing the demographic characteristics of SMA users and non-users, the coefficient for $Gender$ is -0.1338. In this logistic regression, the following coding is used $Gender(male) = -1$ and $Gender(female) = 1$. Then the odds ratio is given by:

$$OddsRatio_{male\ vs\ female} = \frac{e^{0.1338*(-1)}}{e^{-0.1338*(1)}} \approx 1.30 \quad \text{(C.2)}$$

This suggests that males are 30% more likely to become SMA users than females.

Similarly, in order to determine how likely it is for a trader, who is for example, 10 years younger than a given trader (A years old), to become a SMA user, the following odds ratio analysis is performed:

$$OddsRatio_{-10\ year} = \frac{e^{\beta_{Age}*(A-10)}}{e^{\beta_{Age}*A}} = \frac{e^{\beta_{Age}*A}e^{-\beta_{Age}*10}}{e^{\beta_{Age}*A}} = e^{-\beta_{Age}*10} \quad \text{(C.3)}$$

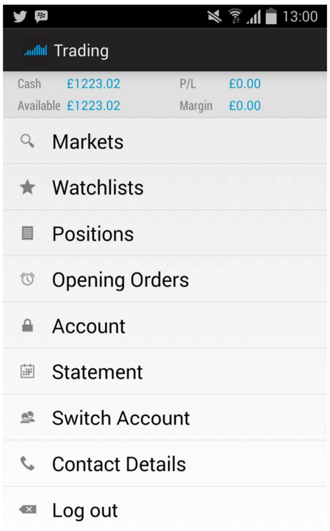
Employing the coefficient for the variable *Age* in Table 4.5, it was found that the odds ratio for a 10 year younger trader is given by.

$$OddRatio_{-10\ year} = e^{0.047*10} \approx 1.60 \quad \text{(C.4)}$$

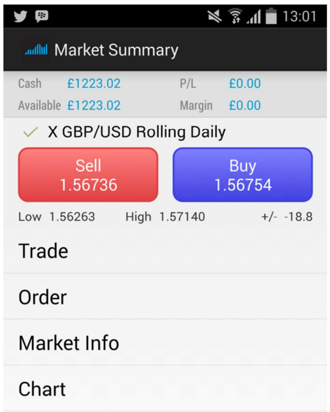
This suggests that on average, when two traders: i_1 and i_2, are compared, where i_1 is 10 years younger than i_2, then i_1 is 60% more likely to be a SMA user.

Appendix D SMA trading screen

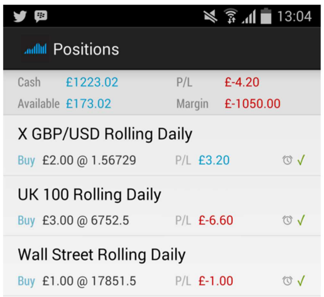
➢ Options



➢ Trade



➢ Watch opened positions



➢ Monitoring the market

Figure D.1 Screen shots from the SDA trading tool examined in this study

Appendix E Partial Approximation of mapping functions

According to Friedman (2001) the partial approximation of mapping functions can be obtaining in the following way: Suppose it is desirable to study the influence of a subset $\mathbf{z}_l$, of size l, of the input variables $\mathbf{x}$, where:

$$\mathbf{z}_l = \{z_1, \dots, z_l\} \subset \{x_1, \dots, x_n\}, \tag{E.1}$$

and $\mathbf{z}_{\setminus l}$ is the complement such as $\mathbf{z}_l \cup \mathbf{z}_{\setminus l} = \mathbf{x}$.

The estimated function $\hat{F}(\mathbf{x})$ can be expressed as $\hat{F}(\mathbf{z}_l, \mathbf{z}_{\setminus l})$. Then the partial approximation of the mapping function is the expected influence of $\mathbf{z}_l$ over $\hat{F}(\mathbf{x})$, which can be expressed as:

$$E(\hat{F}(\mathbf{x}) \big| \mathbf{z}_l) = \int_{\mathbf{z}_{\setminus l}} \hat{F}(\mathbf{z}_l, \mathbf{u}) \omega(\mathbf{u})\, d\mathbf{u} \tag{E.2}$$

Where $\omega(\mathbf{u})$ is the marginal probability density function of $\mathbf{z}_{\setminus l}$ given by:

$$\omega(\mathbf{z}_{\setminus l}) = \int_{\mathbf{z}_l} p(\mathbf{u}, \mathbf{z}_{\setminus l}) d\mathbf{u} \tag{E.3}$$

Where p is the joint density of all the inputs $\mathbf{x}$.

One way to estimate (E.3) is using the entire training dataset as follows:

$$E(\hat{F}(\mathbf{x}) \big| \mathbf{z}_l) \approx \frac{1}{N} \sum_{\mathbf{z}_{\setminus l}} \hat{F}\left(\mathbf{z}_l, \mathbf{z}_{\setminus l}\right) \tag{E.4}$$

However, in the case of GBMs, as they are based on regression trees, $E(\hat{F}(\mathbf{x}) \big| \mathbf{z}_l)$ can be obtained from the trees themselves, without using the data. Friedman (2001) establishes the following procedure to estimate $E(\hat{F}(\mathbf{x}) \big| \mathbf{z}_l)$. Each tree m is visited in a transversal cut, starting from the root until arriving to each terminal node. A weight is assigned to each node, in the case of the root the assigned weight is 1. For each non-terminal node, the splitting variable is inspected; if the splitting variable $x_j \in \mathbf{z}_l$, the right and left sub-sequent node is visited and the weight is not modified. On the other hand, if the splitting variable $x_j \in \mathbf{z}_{\setminus l}$, then both subsequent nodes are visited and the new weight for those nodes is obtained by multiplying the current weight by the fraction of the training observations in each node. When the visit is completed, $E_m(\hat{F}(\mathbf{x}) \big| \mathbf{z}_l)$ is the weighted average of $\hat{F}_m(\mathbf{x})$ using the terminal nodes weights of the tree m. Finally, $E(\hat{F}(\mathbf{x}) \big| \mathbf{z}_l) = \frac{1}{M} \sum_{m=1}^{M} E_m(\hat{F}(\mathbf{x}) \big| \mathbf{z}_l)$.

Appendix B

List of References

Acerbi, C. and Tasche, D. (2002) On the coherence of expected shortfall. *Journal of Banking & Finance,* 26 (7), 1487-1503.

Adrian, T. and Shin, H.S. (2010) The Changing Nature of Financial Intermediation and the Financial Crisis of 2007–2009. *Annual Review of Economics,* 2 (1), 603-618.

Anderson, A. (2013) Trading and Under-Diversification. *Review of Finance,* 17 (5), 1699-1741.

Andersson, F., Mausser, H., Rosen, D. and Uryasev, S. (2001) Credit risk optimization with Conditional Value-at-Risk criterion. *Mathematical Programming,* 89 (2), 273-291.

Arora, A., Forman, C. and Yoon, J.W. (2010) Complementarity and information technology adoption: Local area networks and the Internet. *Information Economics and Policy,* 22 (3), 228-242.

Artzner, P., Delbaen, F., Eber, J.M. and Heath, D. (1999) Coherent measures of risk. *Mathematical Finance,* 9 (3), 203-228.

Azoff, E.M. (1994) *Neural network time series forecasting of financial markets*. John Wiley & Sons, Inc.

Bang, Y., Lee, D.-J., Han, K., Hwang, M. and Ahn, J.-H. (2013) Channel capabilities, product characteristics, and the impacts of mobile channel introduction. *Journal of Management Information Systems,* 30 (2), 101-126.

Barber, B., Lee, Y.-T., Liu, Y.-J. and Odean, T. (2014) The cross-section of speculator skill: Evidence from Taiwan. *Journal of Financial Markets,* 18, 1-24.

Barber, B. and Odean, T. (2011) The behavior of individual investors. *Available at SSRN 1872211*.

Barber, B., Odean, T. and Zhu, N. (2006) Do noise traders move markets? *EFA 2006 Zurich Meetings Paper*.

Barber, B.M., Lee, Y.-T., Liu, Y.-J. and Odean, T. (2007) Is the Aggregate Investor Reluctant to Realise Losses? Evidence from Taiwan. *European Financial Management,* 13 (3), 423-447.

Barber, B.M., Lee, Y.-T., Liu, Y.-J. and Odean, T. (2009a) Just how much do individual investors lose by trading? *Review of Financial Studies,* 22 (2), 609-632.

Barber, B.M. and Odean, T. (2000) Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors. *The Journal of Finance,* 55 (2), 773-806.

Barber, B.M. and Odean, T. (2001) Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics,* 116 (1), 261-292.

Barber, B.M., Odean, T. and Zhu, N. (2009b) Do Retail Trades Move Markets? *Review of Financial Studies,* 22 (1), 151-186.

Barberis, N. and Xiong, W.E.I. (2009) What Drives the Disposition Effect? An Analysis of a Long-Standing Preference-Based Explanation. *The Journal of Finance,* 64 (2), 751-784.

Barro, R.J. (1989) *Black Monday and the future of financial markets*. Richard d Irwin.

Bawden, D. and Robinson, L. (2009) The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of information science,* 35 (2), 180-191.

BCBS (2005) International Convergence of Capital Measurement and Capital Standards: A Revised Framework. *Basel: Basel Committee on Banking Supervision – Bank for International Settlements*. Available from: https://www.bis.org/publ/bcbs107.htm [Accessed 03/05/2012].

BCBS (2011) A global regulatory framework for more resilient banks and banking systems. *Basel: Basel Committee on Banking Supervision – Bank for International Settlements*. Available from: https://www.bis.org/publ/bcbs189.htm [Accessed 03/04/2013].

Becker, S., Düllmann, K. and Pisarek, V. (2004) *Measurement of Concentration Risk - A Theoretical Comparison of Selected Concentration Indices*, Working paper.

Berry, M.J. and Linoff, G. (1997) *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc.

Bessis, J. (2011) *Risk management in banking*. John Wiley & Sons.

Bezemer, D.J. (2009) No One Saw This Coming: Understanding Financial Crisis Through Accounting Models. Available from: http://mpra.ub.uni-muenchen.de/15892/ [Accessed June 2009].

Bhömer, M., Hecht, B., Schöning, J., Krger, A. and Bauer, G. (2011) Falling asleep with Angry Birds, Facebook and Kindle: a large scale study on mobile application usage. Paper presented at Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services, Stockholm, Sweden. Available from: http://delivery.acm.org/10.1145/2040000/2037383/p47-bohmer.pdf?ip=152.78.17.33&id=2037383&acc=ACTIVE%20SERVICE&key=C2716FEBFA981EF1810B722F52EEB3EE17D666AEFB31554D&CFID=371501970&CFTOKEN=87080371&__acm__=1382097760_09324f65bc169da5f3b3eaf8774f247b.

BIS (2016) *Triennial Central Bank Survey of foreign exchange and OTC derivatives markets in 2016*. Available from: http://www.bis.org/publ/rpfx16.htm [Accessed 03/04/2017].

Bluhm, C., Overbeck, L. and Wagner, C. (2003) *An Introduction to Credit Risk Modeling*. London: Chapman & Hall/CRC.

Bo, L. and Capponi, A. (2016) Robust Optimization of Credit Portfolios. *Mathematics of Operations Research,* 0 (0), null. Available from: https://doi.org/10.1287/moor.2016.0790 [Accessed 03/07/2017].

Bolgorian, M. and Raei, R. (2011) A multifractal detrended fluctuation analysis of trading behavior of individual and institutional traders in Tehran stock market. *Physica A: Statistical Mechanics and its Applications,* 390 (21), 3815-3825.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. and White, J.-S.S. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution,* 24 (3), 127-135.

Branke, J., Scheckenbach, B., Stein, M., Deb, K. and Schmeck, H. (2009) Portfolio optimization with an envelope-based multi-objective evolutionary algorithm. *European Journal of Operational Research,* 199 (3), 684-693.

Breiman, L. (1996) Bagging predictors. *Machine learning,* 24 (2), 123-140.

Breslow, N.E. and Clayton, D.G. (1993) Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association,* 88 (421), 9-25.

Brillante, L., Gaiotti, F., Lovat, L., Vincenzi, S., Giacosa, S., Torchio, F., Segade, S.R., Rolle, L. and Tomasi, D. (2015) Investigating the use of gradient boosting machine, random forest and their ensemble to predict skin flavonoid content from berry physical–mechanical characteristics in wine grapes. *Computers and Electronics in Agriculture,* 117, 186-193.

Brown, J.R., Ivkovic, Z., Smith, P.A. and Weisbenner, S. (2008) Neighbors matter: Causal community effects and stock market participation. *Journal of Finance,* 63 (3), 1509-1531.

Brown, P., Chappel, N., Da Silva Rosa, R. and Walter, T. (2006) The Reach of the Disposition Effect: Large Sample Evidence Across Investor Classes. *International Review of Finance,* 6 (1-2), 43-78.

Buijink, A., Visser, B.J. and Marshall, L. (2013) Medical apps for smartphones: lack of evidence undermines quality and safety. *Evidence-Based Medicine,* 18 (3), 90-92.

Bürgisser, P., Kuth, A., Wagner, A. and Wolf, M. (1999) Integrating Correlation. *Risk,* 17 (7), 57-60.

Cespedes, J.C.G. (2002) Credit Risk Modelling and Basel II. *ALGO Research Quarterly,* 5 (1).

Chae, M. and Kim, J. (2004) Do size and structure matter to mobile users? An empirical study of the effects of screen size, information structure, and task complexity on user activities with standard web phones. *Behaviour & information technology,* 23 (3), 165-181.

Chau, P.Y.K. and Lung Hui, K. (1998) Identifying early adopters of new IT products: A case of Windows 95. *Information & Management,* 33 (5), 225-230.

Chen, G.M., Kim, K.A., Nofsinger, J.R. and Rui, O.M. (2007) Trading performance, disposition effect, overconfidence, representativeness bias, and experience of emerging market investors. *Journal of Behavioral Decision Making,* 20 (4), 425-451.

Cheong, L.K. and Chang, V. (2007) The need for data governance: a case study. *ACIS 2007 Proceedings*, 100.

Choe, H. and Eom, Y. (2009) The disposition effect and investment performance in the futures market. *Journal of Futures Markets,* 29 (6), 496-522.

Cisco, T. (2015) Cisco visual networking index: global mobile data traffic forecast update, 2014–2019. *Cisco Public Information*. Available from:

https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html [Accessed 03/12/2016].

Colander, D., Goldberg, M., Haas, A., Juselius, K., Kirman, A., Lux, T. and Sloth, B. (2009) THE FINANCIAL CRISIS AND THE SYSTEMIC FAILURE OF THE ECONOMICS PROFESSION. *Critical Review,* 21 (2-3), 249-267.

Constantiou, I.D., Lehrer, C. and Hess, T. (2014) Changing information retrieval behaviours: an empirical investigation of users' cognitive processes in the choice of location-based services. *European Journal of Information Systems,* 23 (5), 513-528.

Cox, D.R. (1958) The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological),* 20 (2), 215-242.

CSFB (1997) A CREDIT RISK MANAGEMENT FRAMEWORK. Available from: http://www.csfb.com/institutional/research/assets/creditrisk.pdf [Accessed 03/04/2015].

Deb, K. (2008) *Multi-Objective Optimization using Evolutionary Algorithms*. West Sussex: John Wiley & Sons LTD.

Deb, K. and Goel, T. (2001) Controlled elitist non-dominated sorting genetic algorithm for better convergence. Paper presented at First International Conference on Evolutionary Multi-Criterion Optimization.

Dhar, R. and Zhu, N. (2006) Up close and personal: Investor sophistication and the disposition effect. *Management Science,* 52 (5), 726-740.

Dinovella, P. and Moreno-Paredes, J.C. (2005) A computational approach to improve the Benders' Decomposition Method. *Revista de la Facultad de Ingeniería de la U.C.V.,* 20 (2), 5 - 14.

Elith, J., Leathwick, J.R. and Hastie, T. (2008) A Working Guide to Boosted Regression Trees. *Journal of Animal Ecology,* 77 (4), 802-813.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *Modelling Extremal Events*. Berlin: Springer.

Eppler, M.J. and Mengis, J. (2004) The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The information society,* 20 (5), 325-344.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (1996) *Advances in knowledge discovery and data mining*. AAAI press Menlo Park.

FDIC (2013) *Failed Bank List*. Available from: http://www.fdic.gov/bank/individual/failed/banklist.html [Accessed 27/04/2013].

Feng, L. and Seasholes, M.S. (2005) Do Investor Sophistication and Trading Experience Eliminate Behavioral Biases in Financial Markets? *Review of Finance,* 9 (3), 305-351.

Feng, L. and Seasholes, M.S. (2008) Individual investors and gender similarities in an emerging stock market. *Pacific-Basin Finance Journal,* 16 (1–2), 44-60.

Fisher, M. and Dietz, C. (2011) Modelling sector correlations with CreditRisk+. *The Journal of Credit Risk,* 7 (4), 1-20.

Fitzpatrick, T. and Mues, C. (2016) An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market. *European Journal of Operational Research,* 249 (2), 427-439.

Folpmers, M. (2012) The Impact of PD-LGD Correlation on Expected Loss and Economic Capital. *Risk Professional,* 2, 19-25.

Forman, C. (2005) The corporate digital divide: Determinants of Internet adoption. *Management Science,* 51 (4), 641-654.

Forman, C., Goldfarb, A. and Greenstein, S. (2005) How did location affect adoption of the commercial Internet? Global village vs. urban leadership. *Journal of urban Economics,* 58 (3), 389-420.

Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics,* 29 (5), 1189-1232.

Friedman, J.H. (2002) Stochastic gradient boosting. *Computational Statistics & Data Analysis,* 38 (4), 367-378.

FT (2015) Forex brokers suffer escalating losses after Swiss ditch franc cap. *Financial Times*, Jan 16 of 2015. Available from: http://www.ft.com/cms/s/0/a1d0ca62-9d37-11e4-adf3-00144feabdc0.html#slide0.

García, S., Fernández, A., Luengo, J. and Herrera, F. (2009) A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing,* 13 (10), 959.

GDPR, E. (2018) *The EU General Data Protection Regulation*. Available from: https://www.eugdpr.org/ [Accessed 02/04/2018].

Ghose, A. and Han, S.P. (2011) An empirical analysis of user content generation and usage behavior on the mobile Internet. *Management Science,* 57 (9), 1671-1691.

Goodwin, P. and Wright, G. (2014) *Decision Analysis for Management Judgment 5th ed*. John Wiley and sons.

Gopalkrishnan, V., Steier, D., Lewis, H. and Guszcza, J. (2012) Big data, big business: bridging the gap. Paper presented at Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, Beijing, China.

Gordy, M.B. (2002) Saddlepoint Approximation of CreditRisk+. *Journal of Banking and Finance,* 7 (26), 1335- 1353.

Gordy, M.B. (2004) Granularity Adjustment in Portfolio Credit Risk Measurement IN: Szego, G. (ed.) *Risk Measures for the 21st Century*.

Graham, M. (2014) The knowledge based economy and digital divisions of labour. *In Companion to Development Studies,* (3rd edition, eds v. Desai, and R. Potter. Hodder), 189-195.

Greenwood, R. and Nagel, S. (2009) Inexperienced investors and bubbles. *Journal of Financial Economics,* 93 (2), 239-258.

Grinblatt, M. and Keloharju, M. (2009) Sensation Seeking, Overconfidence, and Trading Activity. *The Journal of Finance,* 64 (2), 549-578.

GSMA (2015) The Mobile Economy 2015. (05/10/2016). Available from: http://www.gsma.com/mobileeconomy/archive/GSMA_ME_2015.pdf [Accessed 05/10/2015].

Gundlach, M. and Lehrbass, F. (eds.) (2004) *CreditRisk+ in the banking industry*. Berlin ; London: Springer.

Haaf, H., Rieß, O. and Schoenmakers, J. (2004) Numerical Stable Computation of CreditRisk+ IN: Gundlach, M. and Lehrbass, F. (eds.) *CreditRisk+ in the Banking Sector*. Berlin: Springer.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009) The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter,* 11 (1), 10-18.

Han, C. (2016) Modeling severity risk under PD–LGD correlation. *The European Journal of Finance,* 23 (15), 572-1588.

Han, C. and Kang, J. (2008) An extended CreditRisk+ framework for portfolio credit risk management. *The Journal of Credit Risk,* 4 (4), 63-80.

Han, J., Pei, J. and Kamber, M. (2011) *Data mining: concepts and techniques*. Elsevier.

Härdle, W.K., Hautsch, N. and Overbeck, L. (2017) *Applied quantitative finance*. Springer.

Herbertsson, A. (2011) Modelling default contagion using multivariate phase-type distributions. *Review of Derivatives Research,* 14 (1), 1-36.

Hill, T., Lewicki, P. and Lewicki, P. (2006) *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. StatSoft, Inc.

Hong, W., Thong, J.Y. and Tam, K.Y. (2004) The effects of information format and shopping task on consumers' online shopping behavior: A cognitive fit perspective. *Journal of Management Information Systems,* 21 (3), 149-184.

Hosmer Jr, D.W. and Lemeshow, S. (2004) *Applied logistic regression*. John Wiley & Sons.

Hui, K.-L. and Png, I. (2015) Research Note—Migration of Service to the Internet: Evidence from a Federal Natural Experiment. *Information Systems Research,* 26 (3), 606-618.

Index, C.V.N. (2015) Global Mobile Data Traffic Forecast Update 2014–2019 White Paper.(2015). Available from: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf [Accessed 03/04/2016].

Ivashina, V. and Scharfstein, D. (2010) Bank lending during the financial crisis of 2008. *Journal of Financial economics,* 97 (3), 319-338.

Ivkovic, Z. and Weisbenner, S. (2005) Local does as local is: Information content of the geography of individual investors' common stock investments. *Journal of Finance,* 60 (1), 267-306.

Ivorra, B., Mohammadi, B. and Ramos, A.M. (2009) Optimization strategies in credit portfolio management. *Journal of Global Optimization,* 43 (2), 415-427.

Jacoby, J. (1984) Perspectives on information overload. *Journal of consumer research,* 10 (4), 432-435.

Jacoby, J., Speller, D.E. and Berning, C.K. (1974) Brand Choice Behavior as a Function of Information Load: Replication and Extension. *Journal of Consumer Research,* 1 (1), 33-42.

Jorion, P. (2009) *Financial risk manager handbook*, 5th ed. Hoboken, NJ: Wiley.

Jorion, P. (2011) *Financial risk manager handbook plus test bank FRM Part I/Part II*, 6th ed. Hoboken, N.J. Chichester: Wiley; John Wiley [distributor].

Jorion, P. and Zhang, G. (2009) Credit contagion from counterparty risk. *The Journal of Finance,* 64 (5), 2053-2087.

Kahneman, D. and Tversky, A. (1979) Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society,* 47 (2), 263-291.

Kambatla, K., Kollias, G., Kumar, V. and Grama, A. (2014) Trends in big data analytics. *Journal of Parallel and Distributed Computing,* 74 (7), 2561-2573.

Kaustia, M. (2004) Market-wide impact of the disposition effect: evidence from IPO trading volume. *Journal of Financial Markets,* 7 (2), 207-235.

Kaustia, M. (2010) Prospect Theory and the Disposition Effect. *Journal of Financial and Quantitative Analysis,* 45 (3), 791-812.

Kaustia, M. and Knupfer, S. (2012) Peer performance and stock market entry. *Journal of Financial Economics,* 104 (2), 321-338.

Kelley, E.K. and Tetlock, P.C. (2013) How Wise Are Crowds? Insights from Retail Orders and Stock Returns. *The Journal of Finance,* 68 (3), 1229-1265.

Koenig-Lewis, N., Palmer, A. and Moll, A. (2010) Predicting young consumers' take up of mobile banking services. *International journal of bank marketing,* 28 (5), 410-432.

Korniotis, G. and Kumar, A. (2011) Do Portfolio Distortions Reflect Superior Information or Psychological Biases? *Journal of Financial and Quantitative Analysis,* 48 (1), 1-45.

Korniotis, G.M. and Kumar, A. (2010) Do Older Investors Make Better Investment Decisions? *Review of Economics and Statistics,* 93 (1), 244-265.

Kotsiantis, S., Kanellopoulos, D. and Pintelas, P. (2006) Data preprocessing for supervised leaning. *International Journal of Computer Science,* 1 (2), 111-117.

Kourouthanassis, P.E. and Giaglis, G.M. (2012) Introduction to the special issue mobile commerce: the past, present, and future of mobile commerce research. *International Journal of Electronic Commerce,* 16 (4), 5-18.

Kriegel, H.-P., Borgwardt, K.M., Kröger, P., Pryakhin, A., Schubert, M. and Zimek, A. (2007) Future trends in data mining. *Data Mining and Knowledge Discovery,* 15 (1), 87-97.

Kurgan, L.A. and Musilek, P. (2006) A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review,* 21 (01), 1-24.

Laird, N.M. and Ware, J.H. (1982) Random-Effects Models for Longitudinal Data. *Biometrics,* 38 (4), 963-974.

Lemeshow, S., Klar, J. and Teres, D. (1995) Outcome prediction for individual intensive care patients: useful, misused, or abused? *Intensive care medicine,* 21 (9), 770-776.

Liao, S.-H., Chu, P.-H. and Hsiao, P.-Y. (2012) Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications,* 39 (12), 11303-11311.

Lim, K.-P., Habibullah, M.S. and Hinich, M.J. (2009) The Weak-form Efficiency of Chinese Stock Markets Thin Trading, Nonlinearity and Episodic Serial Dependencies. *Journal of Emerging Market Finance,* 8 (2), 133-163.

Linnainmaa, J.T. (2011) Why do (some) households trade so much? *Review of Financial Studies,* 24 (5), 1630-1666.

Linoff, G.S. and Berry, M.J. (2011) *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

López, V., Fernández, A., Moreno-Torres, J.G. and Herrera, F. (2012) Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. *Expert Systems with Applications,* 39 (7), 6585-6608.

Lu, F.-Q., Huang, M., Ching, W.-K. and Siu, T.K. (2013) Credit portfolio management using two-level particle swarm optimization. *Information Sciences,* 237, 162-175.

Lütkebohmert, E. (2009) *Concentration Risk in Credit Portfolios*. Berlin: Springer-Verlag.

Maletic, J.I. and Marcus, A. (2000) Data Cleansing: Beyond Integrity Analysis *Iq*. Citeseer, 200-209.

Markowitz, H. (1952) Portfolio Selection. *Journal of Finance,* 7 (1), 77-91.

Markowitz, H. (1956) The optimization of a quadratic function subject to linear constraints. *Naval research logistics Quarterly,* 3 (1-2), 111-133.

Mcculloch, C.E. and Neuhaus, J.M. (2001) *Generalized linear mixed models*. Wiley Online Library.

Merton, R.C. (1976) Option pricing when underlying stock returns are discontinuous. *Journal of financial economics,* 3 (1), 125-144.

Mitchell, P. (1997) Misuse of regression for empirical validation of models. *Agricultural Systems,* 54 (3), 313-326.

Mizgier, K.J. and Pasia, J.M. (2016) Multiobjective optimization of credit capital allocation in financial institutions. *Central European Journal of Operations Research,* 24 (4), 801-817.

Mobithinking (2013) *Global mobile statistics 2013 Section E: Mobile apps, app stores, pricing and failure rates*. Available from: http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats/e#lotsofapps [Accessed 09/05/2014].

Molloy, S. and Schwenk, C.R. (1995) The effects of information technology on strategic decision making. *Journal of Management Studies,* 32 (3), 283-311.

Moreno-Paredes, J.C. (2016) *An implementation of a Multi-Objective Evolutionary Approach for Credit Portfolio Optimisation*. Working Paper. University of Southampton.

Moreno-Paredes, J.C., Mues, C. and Thomas, L.C. (2013) A Multi-Objective Decision Framework for Credit Portfolio Management. Paper presented at Credit Scoring and Credit Control XIII Conference, Edinburgh. Available from: https://www.business-school.ed.ac.uk/crc/wp-content/uploads/sites/55/2017/02/A-Multi-Objective-Decision-Framework-for-Credit-Portfolio-Management-Moreno-Paredes-Mues-and-Thomas.pdf.

Morgan, J. (1996) *Riskmetrics: technical document*. Morgan Guaranty Trust Company of New York.

Nakasumi, M. (2012) Decision making aid in mobile environment by behavioral characteristic. Paper presented at Proceedings of the 13th International Conference on Electronic Commerce, Liverpool, United Kingdom. Available from: http://delivery.acm.org/10.1145/2380000/2378118/a14-nakasumi.pdf?ip=152.78.17.14&id=2378118&acc=ACTIVE%20SERVICE&key=BF07A2EE685417C5%2EA13CBF7F1C3C7DF4%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&CFID=481833855&CFTOKEN=90299100&__acm__=1424886847_b877a9e7ee85e7192942502f54e15b21.

Napoli, P.M. and Obar, J.A. (2014) The emerging mobile Internet underclass: A critique of mobile Internet access. *The Information Society,* 30 (5), 323-334.

Norris, P. (2001) *Digital divide: Civic engagement, information poverty, and the Internet worldwide*. Cambridge University Press.

Nussbaum, M., Papritz, A., Fraefel, M., Baltensweiler, A. and Keller, A. (2015) Predictive mapping of soil properties at high resolution by component wise gradient boosting. Paper presented at EGU General Assembly Conference Abstracts.

Odean, T. (1998) Are Investors Reluctant to Realize Their Losses? *The Journal of Finance,* 53 (5).

Olshen, L. and Stone, C.J. (1984) Classification and regression trees. *Wadsworth International Group,* 93 (99), 101.

ONS, O.F.N.S. (2010) *Deprivation Index*. Available from: https://www.gov.uk/government/publications/english-indices-of-deprivation-2010.

Orenstein, A. (2011) A MODIFIED CAREMARK STANDARD TO PROTECT SHAREHOLDERS OF FINANCIAL FIRMS FROM POOR RISK MANAGEMENT. *New York University Law Review,* 86 (3), 766-803.

Overbeck, L. and Sokolova, M. (2017) Risk Measurement with Spectral Capital Allocation IN: Härdle, W.K., Hautsch, N. and Overbeck, L. (eds.) *Applied Quantitative Finance*. Berlin, Heidelberg: Springer Berlin Heidelberg, 139-159.

Paton, D. and Williams, L.V. (2005) Forecasting outcomes in spread betting markets: can bettors use 'quarbs' to beat the book? *Journal of Forecasting,* 24 (2), 139-154.

Philip Chen, C.L. and Zhang, C.-Y. (2014) Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences,* 275, 314-347.

Philip, L., Cottrill, C., Farrington, J., Williams, F. and Ashmore, F. (2017) The digital divide: Patterns, policy and scenarios for connecting the 'final few'in rural communities across Great Britain. *Journal of Rural Studies,* 54, 386-398.

Ponsich, A., Jaimes, A.L. and Coello, C.a.C. (2013) A Survey on Multiobjective Evolutionary Algorithms for the Solution of the Portfolio Optimization Problem and Other Finance and Economics Applications. *IEEE Transactions on Evolutionary Computation,* 17 (3), 321-344.

Prasopoulou, E., Pouloudi, A. and Panteli, N. (2006) Enacting new temporal boundaries: the role of mobile phones. *European Journal of Information Systems,* 15 (3), 277-284.

Rohanizadeh, S.S. and Moghadam, M.B. (2009) A proposed data mining methodology and its application to industrial procedures. *Journal of Industrial Engineering,* 4 (1), 37-50.

Röthig, A. and Chiarella, C. (2007) Investigating nonlinear speculation in cattle, corn, and hog futures markets using logistic smooth transition regression models. *Journal of Futures Markets,* 27 (8), 719-737.

Rudolph, G. and Agapie, A. (2000) *Convergence properties of some multi-objective evolutionary algorithms*. Citeseer.

Rushton, L. (2000) Reporting of occupational and environmental research: use and misuse of statistical and epidemiological methods. *Occupational and Environmental Medicine,* 57 (1), 1-9.

Saunders, A. and Allen, L. (2010) *Credit risk management in and out of the financial crisis: new approaches to value at risk and other paradigms*. John Wiley & Sons.

Schlottmann, F., Mitschele, A. and Seese, D. (2005) A multi-objective approach to integrated risk management IN: Coello, C.a.C., Aguirre, A.H. and Zitzler, E. (eds.) *Evolutionary Multi-Criterion Optimization*. 692-706.

Schlottmann, F. and Seese, D. (2001) A hybrid genetic-quantitative method for risk-return optimisation of credit portfolios. Paper presented at Proceedings of the Conference of Quantitative Methods in Finance, University of Technology, Sydney, Australia.

Schlottmann, F. and Seese, D. (2004) A hybrid heuristic approach to discrete multi-objective optimization of credit portfolios. *Computational Statistics & Data Analysis,* 47 (2), 373-399.

Schlottmann, F. and Seese, D. (2005) Discovery of risk-return efficient structures in middle-market credit portfolios IN: Baier, D. and Wernecke, K.D. (eds.) *Innovations in Classification, Data Science, and Information Systems*. 506-514.

Seru, A., Shumway, T. and Stoffman, N. (2010) Learning by Trading. *Review of Financial Studies,* 23 (2), 705-739.

Shapira, Z. and Venezia, I. (2001) Patterns of behavior of professionally managed and independent investors. *Journal of Banking & Finance,* 25 (8), 1573-1587.

Sharpe, W.F. (1994) The Sharpe Ratio. *The Journal of Portfolio Managemen,* 21 (1), 49-58.

Shefrin, H. and Statman, M. (1985) The disposition to sell winners too early and ride losers too long - theory and evidence. *Journal of Finance,* 40 (3), 777-790.

Sraeel, H. (2006) Hold the Phones: Mobile Commerce Is Here. *American Banker*. Available from: http://www.americanbanker.com/btn/19_12/-297017-1.html [Accessed 03/02/2015].

Statistica (2016) *Apple iPhone sales worldwide 2007-2016*. Available from: https://www.statista.com/statistics/276306/global-apple-iphone-sales-since-fiscal-year-2007/ [Accessed 02/01/2016].

Steinberg, D. and Cardell, N.S. (1998) The hybrid CART-Logit model in classification and data mining. *Salford Systems White Paper*.

Tan, P.-N. (2006) *Introduction to data mining*. Pearson Education India.

Tasche, D. (2002) Expected shortfall and beyond. *Journal of Banking & Finance,* 26 (7), 1519-1533.

Terpstra, D.E., Rozell, E.J. and Robinson, R.K. (1993) THE INFLUENCE OF PERSONALITY AND DEMOGRAPHIC-VARIABLES ON ETHICAL DECISIONS RELATED TO INSIDER TRADING. *Journal of Psychology,* 127 (4), 375-389.

Thomas, L.C. (2009) *Consumer credit models: pricing, profit and portfolios*. Oxford: Oxford University Press.

Thomas, L.C., Edelman, D.B. and Crook, J.N. (2002) *Credit scoring and its applications*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Thomas, L.C., Edelman, D.B. and Crook, J.N. (2004) *Readings in credit scoring : foundations, developments, and aims*. Oxford ; New York: Oxford University Press.

Tong, E.N.C., Mues, C. and Thomas, L.C. (2012) Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research,* 218 (1), 132-139.

Tsetsi, E. and Rains, S.A. (2017) Smartphone Internet access and use: Extending the digital divide and usage gap. *Mobile Media & Communication,* 5 (3), 239-255.

Tversky, A. and Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *science,* 185 (4157), 1124-1131.

Tzeng, G.-H. and Huang, J.-J. (2011) *Multiple attribute decision making: methods and applications*. CRC press.

Van Dijk, J. and Hacker, K. (2003) The digital divide as a complex and dynamic phenomenon. *The information society,* 19 (4), 315-326.

Van Gestel, T. and Baesens, B. (2009) *Credit risk management basic concepts: financial risk components, rating analysis, models, economic and regulatory capital*. Oxford: Oxford University Press. Available from: http://lib.myilibrary.com?id=197871.

Venkatesh, V., Morris, M.G., Davis, G.B. and Davis, F.D. (2003) User acceptance of information technology: Toward a unified view. *MIS quarterly,* 27 (3), 425-478.

Venkatesh, V., Thong, J. and Xu, X. (2012) Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS quarterly,* 36 (1), 157-178.

Verbeke, G. (1997) Linear Mixed Models for Longitudinal Data *Linear Mixed Models in Practice*. Springer New York, 63-153.

Weber, M. and Camerer, C.F. (1998) The disposition effect in securities trading: an experimental analysis. *Journal of Economic Behavior & Organization,* 33 (2), 167-184.

Xu, K., Chan, J., Ghose, A. and Han, S.P. (2016) Battle of the Channels: The Impact of Tablets on Digital Commerce. *Management Science*. Available from: http://pubsonline.informs.org/doi/pdf/10.1287/mnsc.2015.2406 [Accessed 02/11/2016].

Xu, Q., Erman, J., Gerber, A., Mao, Z., Pang, J. and Venkataraman, S. (2011) Identifying diverse usage behaviors of smartphone apps. Paper presented at Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference.

Yamaguchi, T. (2006) Understanding Trading Behavior in 401(k) Trading Plans. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1094993## [Accessed 03/04/2016].

Yonghong, Z.X.W. (2001) The Empirical Study on Disposition Effect in China' s Stock Market. *Journal of Finance,* 7, 011.

Young, A. and Mcconkey, K. (2012) Data Governance and Data Quality: Is It on Your Agenda? *Journal of Institutional Research,* 17 (1), 69-77.

Zopounidis, C. and Doumpos, M. (2002) Multi-criteria Decision Aid in Financial DecisionMaking: Methodologies and Literature Review. *Journal of Multi-Criteria Decision Analysis,* 11, 167-186.