ACTIVE ALLOCATION IN PRIVATE DEBT USING NATURAL LANGUAGE PROCESSING WITH ALTERNATIVE DATA SOURCES

submitted in accordance with the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in the subject

OPERATIONS RESEARCH

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: PROFESSOR PETRUS H. POTGIETER

Date submitted: June 2023

by

STUART ROYDEN-TURNER

Saturday 6th January, 2024 at 10:45

Contents

List of Tables

List of Figures

Abbreviations

ACF Autocorrelation Function

APT Arbitrage pricing theory

AI Artificial intelligence

ALM Asset liability modelling

AR Autoregressive

ARIMA Autoregressive integrated moving average

AUM Assets under management

BERT Bidirectional encoder representations from transformers

BVAR Bayesian value at risk

BOW Bag of Words

CAGR Compound annual growth rate

CAPM Capital asset pricing model

CBOW Continuous bag of words

CCAPM Consumption capital Asset Pricing Model

CCI Credit cycle indicator

CDAR Conditional drawdown at risk

CDS Credit default swap

CIR Cox-Ingersoll-Ross

CNN Convolutional neural network

CRRA Constant relative risk aversion

CSS Composite sentiment score

CVAR Conditional value at risk

DB Defined benefit

DCC-VAR Dynamic conditional control Vector Autoregression

DE Diagnostic expectations

DSGE Dynamic stochastic general equilibrium

EDF Expected default frequency

EL Expected loss

ESG Economic scenario generator

ETF Exchange-traded funds

ETL Extract, transform and load

FED Federal Reserve bank

GARCH Generalized autoRegressive conditional heteroskedasticity

GBM Geometric Brownian motion

GBR Great Britain

GDELT Global database of events, language and tone

GDP Gross domestic product

GFC Global financial crisis

GloVe Global vectors

GVAR Global vector autoregressive

HWV Hull-White-Vasicek

HY High yield

IFRS9 International financial reporting standard no. 9 for credit risk

IG Investment grade

IRB Internal ratings based

JPM JP Morgan bank

KNN K Nearest neighbour

LDA Latent Dirichlet allocation

LGD Loss given default

LIWC Linguistic inquiry and word count

LSI Latent semantic analysis

LSTM Long short-term model

MAD Mean absolute deviation

MAE Mean absolute error

MDP Markov decision process

ME Maximum entropy

ML Machine-learning

MDP Markov decision process

MVO Mean variance optimisation

NB Naïve Bayes

NFCI National financial conditions index

NLP Natural language processing

NLTK Python natural language toolkit

NYT The New York Times

OECD Organisation for economic co-operation and development

P-ACF Partial autocorrelation function

PCA Principal component analysis

PD Potential for default

PE Price-to-equity ratio

PLC PD LGD correlation

POS Part-of-speech

PSP Performance seeking portfolio

RBC Real business cycles

ReCNN Recursive convolution neural network

RMSE Root mean square error

RNN Recurrent neural network

RSA Republic of South Africa

SAA Strategic asset allocation

SDF Stochastic discount factor

SIG Sub-investment grade

SVAR Structural vector auto-regression (VAR)

SVM Support vector machine

TAA Tactical asset allocation

TF-IDF Frequency-inverse document frequency

TFP Total factor productivity

UK United Kingdom

US United States

USA United States of America

VADER Valence aware Dictionary and sentiment reasoner

VaR Value at risk

VAR Vector autoregressive

VIX The volatility index

YOY Year on year

Declaration

Name: STUART J ROYDEN-TURNER

Student number: 50797891

Degree: DOCTORATE OF PHILOSOPHY IN OPERATIONS RESEARCH

ACTIVE ALLOCATION IN PRIVATE DEBT USING NATURAL LANGUAGE PROCESSING WITH ALTERNATIVE DATA SOURCES

I declare that the above thesis is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

I further declare that I submitted the thesis to originality checking software and that it falls within the accepted requirements for originality.

I further declare that I have not previously submitted this work, or part of it, for examination at UNISA for another qualification or at any other higher education institution.

SIGNATURE DATE

Acknowledgements

Writing a PhD is an exciting but strenuous challenge and from appearances it is a solitary one. Many hours spent researching hundreds of sources of information, the consuming focus required to master and code mathematical problems, understanding complex theory or phrasing arguments to compel your reader; for years. It is also an extraordinary journey of discovery, with the exception of editing my own writing, I thoroughly enjoyed writing it and I am so grateful for the opportunity to have completed it. To me it was far from solitary and was not possible to do alone. I involved everyone in some way, but I would like to acknowledge and thank the following people for their contribution.

Firstly to Petrus, Professor Potgieter. How do I even begin to thank you enough. This journey started with my masters and I again thank you for your high standards, your extraordinary experience and knowledge during my PhD study. I recognise the care you have taken in steering me, I am deeply grateful. I have always enjoyed our meetings, your friendship and amusing dry sense of humour. Thank you!

It is so important to acknowledge the role that my beautiful wife, Tamryn, played. I cannot thank you enough for your patience and incredible support. For years there have been so many hours you contributed to organise kids and family affairs, manage our lives, proofread mathematical documents and patiently wait till the last page was written. Without your support this would not have been possible. I am forever grateful. Thank you!

To my kids, Ethan and Megan, I am so appreciative for your support and patience. I know it has been frustrating for your Dad to have been occupied during the weekends, I will be forever grateful for your sacrifice and giving me the time to complete this study.

To my family, Mum, Jess, Shan and Muz. How lucky am I to be part of such a crazy gang. You have always been supportive in many ways. I am very thankful for all your love and support. Dimitri Tsopanakos, your flexibility allowing me to continue a quest whilst trying to balance a demanding job. Craig Sherman and Justin De Kock, your contributions are meaningful to me and I thank you!

Abstract

Portfolio analysis is benefiting from the surge of alternative sources of data coupled with new modelling frameworks, introduced by machine-learning. I collected alternative data and applied new frameworks (machine-learning techniques and technology) to the domain of private debt. This is an interesting and complex asset class, which has a significant shortage of data from which to model. To counter this issue, I incorporated advanced macro-finance and asset-pricing models. Such will provide the correct context in which to model this asset class as part of a sophisticated multi-asset portfolio construction framework. To ensure that the credit-risk models are fully understood, I selected a modelling technique from a broad array of options in a mature environment of credit modelling largely performed in banking (whilst ensuring the technique is suitable for asset management). The modelling framework is geared to account for the dynamics of business cycles, this being an important results driver in unlisted credit and other asset classes alike. My thorough macro-finance research allows me to design non-trivial processes to incorporate alternative signals as part of an asset-pricing framework on which to generate information for use in portfolio construction via the data simulated. My economic scenario generator considers the relative changes to asset classes at various points in the business cycle as part of a long-term investment program suitable for a defined-benefit pension funds portfolio. The final portfolio models are put together using a reinforcement learning framework. This framework connects the macro-finance theory dealing with the business cycle dynamics, together with the credit-risk techniques, to portfolio modelling techniques which I believe compounds to a sophisticated modelling framework for strategic asset allocation in a data-sparse environment.

Key terms

Macro-finance, asset-pricing, private debt, business cycles, alternative data, natural language processing, sentiment analysis, neural networks, economic-scenario-generation and reinforcement learning.

Opsomming

Portefeulje-analise trek voordeel uit die opwelling aan alternatiewe databronne, gekoppel met nuwe modelleringsraamwerke wat uit masjienleer spruit. Ek het alternatiewe data versamel en nuwe raamwerke (masjienleertegnieke en -tegnologieë) toegepas in die gebied van private skuld. Dít is 'n interessante en komplekse bateklas wat 'n beduidende gebrek aan modelleringsdata het. Om dié kwessie die hoof te bied, het ek gevorderde makrofinansiële en batebeprysingsmodelle geïnkorporeer.

Daardeur word die korrekte konteks verskaf vir die modellering van hierdie bateklas as deel van die raamwerk vir die 'n gesofistikeerde meerbate-portefeuljekonstruksie. Ten einde te verseker dat die kredietrisiko-modelle volledig begryp word, het ek 'n modelleringstegniek gekies uit 'n uitgebreide reeks opsies binne 'n volgroeide kredietmodelleringsomgewing wat hoofsaaklik in die bankwese verwend word (onderwyl daar verseker word dat die tegniek geskik is vir batebestuur).

Die modelleringsraamwerk is gerat om die dinamiek van die konjunktuur (sakesiklus) in berekening te neem aangesien dít 'n beduidende dryfkrag is vir ongenoteerde krediet en eweneens vir ander bateklasse. My deeglike navorsing in makrofinansies laat my toe om nie-triviale prosesse te ontwerp om alternatiewe seine as deel van 'n batebeprysingsraamwerk waarop inligting vir gebruik in die portefeuljekonstruksie gegenereer word, te inkorporeer.

My generator vir ekonomiese scenarios neem die relatiewe veranderings in bateklasse op verskillende tydstippe in die konjunktuur in ag as deel van 'n langtermyn-beleggingsprogram wat geskik is vir 'n gedefinieerde voordeel-pensioenfondsportefeulje. Die uiteindelike portefeuljemodelle word saamgestel deur 'n versterkingsleer-raamwerk te gebruik. Dié raamwerk verbind die makrofinansies-teorie vir die konjunktuurdinamiek asook die kredietrisiko-tegnieke met portefeuljemodelleringstegnieke wat (volgens my) tesame 'n gesofistikeerde modelleringsraamwerk vir strategiese batetoewysing in 'n data-arm omgewing daarstel.

Sleutelbegrippe

Makrofinansies, batebeprysing, privaat skuld, konjunktuur, alternatiewe data, natuurlike taalverwerking, sentimentanalise, neurale netwerke, generering van ekonomiese scenarios, versterkingsleer,

Isifinyezo sokuqukethwe

Ukuhlaziywa kwephothifolio kuyinzuzo ekuqhubukeni kominye imithombo yedatha kanye nezihlaka eziyisibonelo eyethulwe yinqubo yokufunda yomshini. Ngiqoqe olunye uhlobo lwedatha mtase ngifaka uhlaka lusha (inqubo yokufunda yomshini enobuchwepheshe) esizindeni sesikweletu sangasese. Loki kuyinto ethokozisayo futhi nesigab sempahla esiyinkimbinkimbi, okunoku shoda kakhulu kwedatha ekuzomodelwa kuyo. Ukulwa naloludaba , ngihlanganise imali enkulu ethuthukiswe nokuthuthukisa imodeli yentengo yempahla. Lokhu kuzohlinzeka umongo olungile lokumodela isigaba sempahla nje ngenxenye yohlaka oluyinkimbinkimbi lokwakhiwa kwephothifoliyo yezimpahla eziningi. Ukuqinisekisa ukuthi kamamodeli engciphe yekhredithi aqondwa ngokugcwele, ngikhethe imodeli elinobuchwepheshe kusuka kuhlu olubanzi uma izinketho endaweni evuthiwe yokumodela izikweletu eyenziwa kakhulu emabhange (ngenkathi eqinisekisa ubuchwepheshe okufanele ukuphatha kwempahla). Loluhlaka lokumodela kwenzelwe ukulandelisa ngokuguquguquka kwemijikelezo yebhizinisi lokhu kuyimiphumela ebalulekile ekushayeleni kwekhredithi engafakwanga nezinye isigaba zempahla ngokufanayo. Ugcwaningo lwemali enkulu olujulile lungivumele ukuklama inqubo engeyona into encane ukuhlanganisa ezinye izimpawu njengenxenye yohlaka lwentengo yempahla lokhu kuzoletha imininingwane yokusebenzisa ukwakhiwa kwephothifoliyo ngedatha elingisiwe. Ijeneretha yami yesimo somnotho ibheka izinguquko ezihambisanayo ukuthola ikilasi lempahla ezindaweni ezahlukahlukene kwezamabhizinisi nje ngenxenye yohlelo lotshalozimali lwesikhathi eside ezifanele iphothifoliyo yezikwama zempesheni ezichaziwe. Amamodeli nephothifoliyo okugcina ahlanganiswe ngokusebenzisa uhlaka lokufunda oluqinisayo. Loluhlaka luhlanganisa imibono yezimali ezinkulu ekubhekene nemijikelezo okushintshashinrsha kwebhizinisi, kanye nequbo yobungozi besikweletu, kumasu nokumodela iphothifoliyo engikholwa engikholwa ihlangisa uhlaka lokumodela oluyindida ngokwabiwa kwempahla yeah endaweni eyingcosana yedatha.

Imigomo ebalulekile

Imali enkulu, Intengo yempahla, Imijikelezo yebhizinisi, Enye idatha, Ukucutshungulwa kolwimi lwemvelo, Ukuhlaziywa kwemizwelo, Amanethiwekhi angathathi hlangothi, Isizukulwane sesimo sezomnotho, Ukuqinisa ukufunda

Part I

Introduction

In this research, I am primarily interested in the use of quantitative techniques and research to provide the necessary tooling for a sophisticated quantitative approach to a multi-asset portfolio-choice problem. This task is set against a current theme in portfolio management of modelling and risk management of complex assets, such as private debt or credit, as part of a strategic asset allocation program for long-term financial returns. This is fashioned for use in pension funds investment programs.

My strategy is based on the tenets of macro-economic research and is aimed at an advanced portfolio theory for specific assets. I will also be discussing the methods and theories developed to tackle difficult problems in this area. Part of this complexity is to configure an appropriate framework for dealing with business cycles. This is because business cycles relate to returns achievable over and above the risk free benchmark (referred to as risk premia), especially the challenges presented by illiquid credit. I research the macro-finance theory for dealing with business cycles, connecting the reader with the theory underpinning the portfolio construction approach taken later in this study. The use of non-traditional or alternative data sources then becomes a major focus of my investigation. My data is both readily available and raw, which is important as it is not, essentially, procured signal information. How I deal with much of this data, advancing the research point for techniques that are used for processing unstructured data, using modern machine-learning techniques, is a pertinent study theme. I dedicated much time to consider whether more complex modelling techniques are beneficial. The modern techniques and technology that are freely available are relatively easy to use and are so powerful and flexible. In the case of market signals from NLP, I found powerful and flexible contemporary methods, however the shortage of time-series data and the typical nuances found in financial markets need to be accounted for.

In part II, I research macro-finance methods to understand the key asset-pricing and business cycle models in this complex and interesting research area. Important foundational concepts such as stochastic deflator and utility theories are discussed. I provide an overview of advanced asset-pricing techniques as they relate to my central problem. In part III, I forecast a risk appetite proxy using machine-learning analysis. My experiment uses live data from alternative data sources and machine-learning for a time-series analysis for providing forward-looking sentiment signals. In part IV, I provide some quant modelling techniques and theory which support my approach to asset-pricing and private debt credit-risk measurement. In part V, I detail an economic scenario-generator modelling section explaining components of my integrated asset-pricing model. I also detail the portfolio construction techniques, a portfolio choice model and planning model using reinforcement learning methods. This section is analytical and illustrative in demonstrating to the reader not only the underlying assumptions, but also the rich information made possible only by sophisticated modelling and simulation techniques.

As context around the need for a study of this nature, the importance of risk management of private debt in the investment management environment should not be underestimated as the extent of the private asset investing is more extensive and private lending is more significant. Kokoszka (2023) reports that International Organisation of Securities Commissions (IOSCO) suggests that private credit funds may finance portfolio companies that take on levels of leverage that exceed the risk appetite of banks; and this additional debt, alongside leverage incurred at the fund level, could have a cumulative negative impact on the financial system. The UK Financial Conduct Authority (FCA) is gearing up to initiate a examination of valuations in the private market, driven by the combination of increasing interest rates and an economic downturn. This situation is placing significant strain on private equity firms and the companies within their portfolios. The FCA is concerned about the robustness of current market practices for valuing privately held assets (Kokoszka, 2023). This study is focussed on methods to advance risk management practice for private debt exposures as part of a larger investment solution. My approach is a robust framework that is useable in the construction of a portfolio with comprising private assets.

1 Overview of this study

I will begin by providing an overarching explanation of the document, the areas of focus in this study and how these are interrelated. This study is an exercise in thorough academic research of key components of a market practitioner's portfolio construction process. In this study I do not consider the full portfolio construction process, for example I wont detail aspects such as environmental, social, and governance (ESG) factors, or differences in cost due to legal wrappers. This focus here is on balancing long-term investment returns against the costs associated with market and credit risk, in the context of an investment portfolio for a private asset portfolio. The alignment in the approach to quantifying the risks between liquid and illiquid assets is an important advancement.

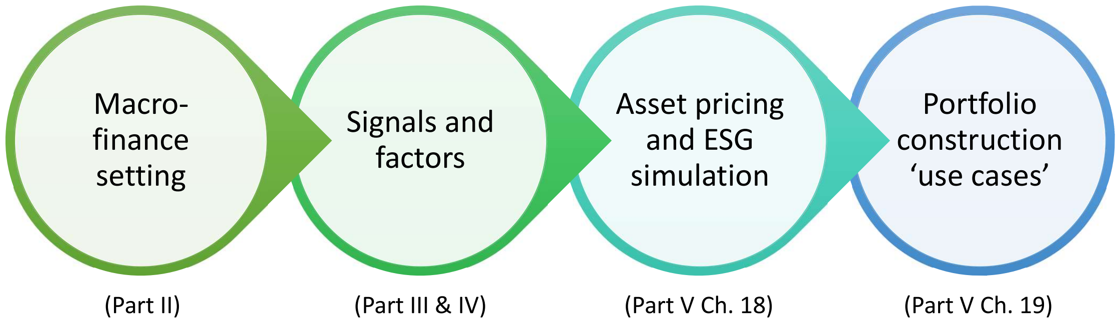


Figure 1: High level schema of the interrelated component parts in this study, as part of a broader framework for constructing a portfolio. This document is structured to represent key stages of the portfolio construction for long-term liability based investments.

The document starts with a macro-finance section where I investigate the underlying theory and available models, so as to correctly frame the investment portfolio problem. With the macro-finance context set, I then focus on specific market signals in Part III and asset class risk drivers and modelling in Part IV. The next section, Part V Chapter 18, I design the simulation of asset pricing pathways to more fully represent financial risks. Simulation of asset prices is an exercise in mathematical formulation and analytical review. The asset pathways are useful in next phase which is an exercise in portfolio optimisation, back-testing and stress-testing. This final stage, Part V Chapter 19, is an asset allocation optimisation framework to complete the portfolio construction process. In this particular 'use-case' we focus on solving for a long-term planning problem that requires asset liability modelling (ALM), I have also selected to make use of more sophisticated machine learning techniques. The following subsections go into bit more of the detail you will read in each section.

1.1 Long term financial returns

Over the past 40 years both stock and bond returns have seen sustained periods of higher returns than their respective long-term averages (calculated from 1900). Dobbs et al. (2016) explain that lower inflation and interest rates are largely responsible for the increased returns on financial asset classes, with three decades of low inflation driving most of the interest rate changes in the European Union and the United States. Paul Volker, chairman of the Federal

Reserve Board from 1979 to 1987, was the first to implement inflation targeting measures, which set the ball in motion. This action, and the actions of other central banks, quickly stabilised inflation rates to acceptable levels (e.g. inflation at 25% in 1979 in the U.K. was reduced to 0.54% by 1982). An early drop in interest rates, followed by successive rate cuts as a result of a decline in the perceived risk of unexpected inflation and the accompanying inflation risk premium, explains the decelerating pace of inflation. (Dobbs et al., 2016).

To understand the dynamics of financial asset class returns, we must review the decline in investment returns over the thirty years from 1985 to 2014; Dobbs et al. (2016) explore these dynamics well. Interest rates are influenced by a number of variables, including unanticipated inflationary shifts, central bank actions and communication, rising credit supply and demand, and country risk (referring to investment uncertainty introduced from a particular country operating profits or value of assets). Of the last thirty years in which the amount bond returns have increased, the most significant driver relates to higher nominal capital gains as a result of lowering yields. The second-largest contribution is from lower inflation rates as a consequence of inflation targeting policies, and lastly due to lower nominal yields on bonds (Dobbs et al., 2016). For equity returns, the reduction in inflation transmits increases in equity returns through the effect on the price-to-equity ratio (PE) and the dividend pay out ratio (Dobbs et al., 2016). The PE ratio has increased over the previous 30 years, while the equity price has increased by 3.6% above its long-term average. This gain is due to adjustments in PE by 2.5% and increases in the net income margin by 1.1% during a protracted period of falling inflation (Dobbs et al., 2016). It is evident from their research that the decline in inflation and interest rates over the past four decades has impacted both bond and stock prices (Dobbs et al., 2016). The near-zero-bound interest rate and low inflation environment have been constraining factors for investment managers of fixed-income asset classes. If the reduction in yields (now close to zero bound) and low inflation are driving core asset-class returns, this does not bode well for future asset-class returns.

Historically, equities and fixed-income investments have been insurers' and pension funds' primary asset classes due to their income-generating potential and risk-diversification qualities. Concerned about decreased returns, investors with a long-term horizon have examined alternative asset classes that offer equivalent yields at comparable levels of risk. This circumstance necessitates that asset managers and asset owners re-evaluate existing methodologies for balancing equity and bond allocations to meet return and risk requirements.

1.2 Alternative assets

Alternative investments offer a range of options for investors seeking returns that are not correlated with traditional assets. Possible investment strategies include investments in real estate, infrastructure, hedge funds, structured goods, and alternative financing sources such as private debt (including investments in or direct exposure to loans to small-to-medium and corporate entities). Baker and Filbeck (2013) declare that the rising demand in alternative asset markets is a result of the prolonged bear market and low interest rate environment. Classic investments are often seen as 'long-only' holdings in equities, fixed-income, and cash. Chambers et al. (2015) stress that other assets, such as stamps and wine, can be categorised as 'alternative'. The defining characteristic of an alternative investment is that it fulfils the standards for institutional grade in order to be included in the investment plans of insurers and pension funds. Benrud (2011) explain that institutional investors and fund managers are focusing on alternative assets having distinctive and compelling characteristics, such as longer-term illiquid investments.

Alternative assets have the potential to increase the return on a portfolio by providing additional return (or alpha, which is defined as the return in excess of a selected benchmark) at a lower volatility risk profile. In stressed markets, alternative assets offer a compelling alternative to the low rates on the fixed-income market, which are not considered correlating with growth assets. This enticing notion is accompanied by a health warning: alternative asset risks are more difficult (the risks are more complex and mostly hard to measure, or have a complete lack of data to quantify) and generally more complex than traditional asset risks. However, this barrier to entry enables a professional investor to flourish in markets, implementing investing methods that are normally too complex for retail investors.

1.3 Private debt and credit-risk

Investors have traditionally regarded credit as the purchase of investment grade corporate and sovereign debt. Reviewing liquid assets, Whyte (2018) and Nimisha (2010) underline that alternative credit encompasses securities such as high-yield bonds, structured credit, emerging market debt, and bank loans. The illiquid side of investments includes distressed debt purchases, unlisted direct lending, and speciality finance. Whyte (2018) notes that private debt provides a solution for institutional investors grappling with low interest rates, higher-than-average volatility, and an environment of rising debt levels, making it an increasingly prevalent option for strategic asset allocation. Credit loans offer a smooth return profile in good times, as a result of their ability to generate income and risk-diversification attributes, highly favourable asset class features.

Private credit loans (private debt) are one type of alternative asset that investors find particularly appealing. We gain insight into the annual alternative asset-management survey by Preqin in 2017; 2018; 2019; 2022. These researchers announce that large financial institutions intend on extending their private debt holdings in the future. Low-yielding fixed-income is generally thought to correlate with growth assets in times of economic stress, making alternative assets an attractive option. Giuzio et al. (2018) comment that portfolios invested in private debt, such as direct lending, result in portfolios with higher returns. Such portfolios are more efficient, reporting improvements in investment risk metrics, such as drawdown and the Sharpe ratio.

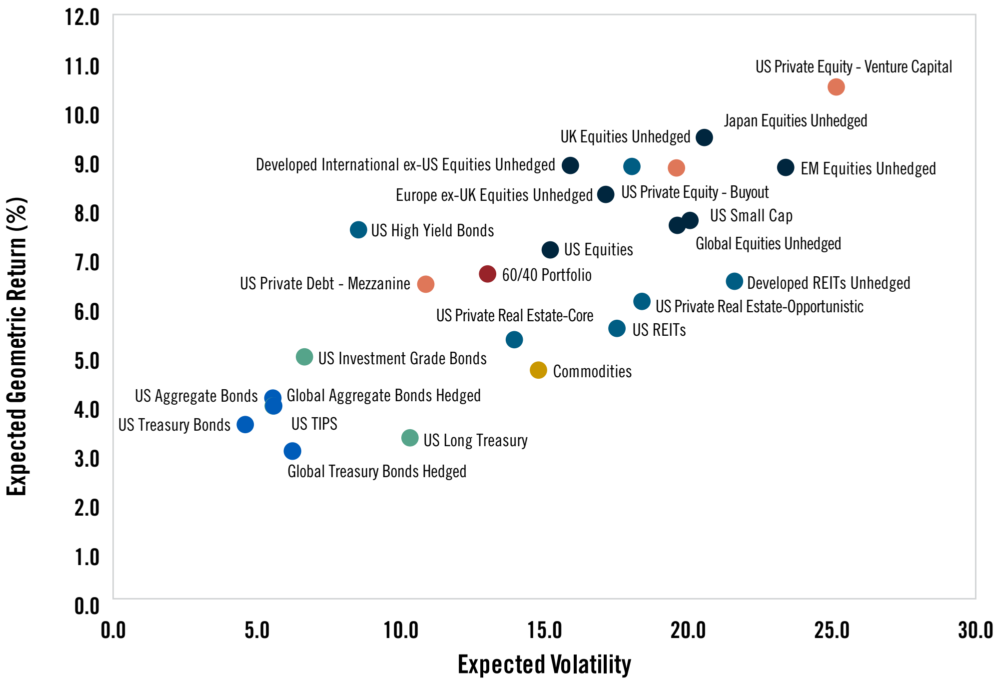


Figure 2: Ten year forecast of returns and volatility plotted in the familiar risk return scatterplot, with data sourced from PGIM (2022), page 23. This comparable XY chart shows the comparable expected volatility and expected return spectrum credit strategies. This the risk/return assumptions for one asset manager, where later in this study I summarise a broad range of asset manager's capital market assumptions.

Over the past five decades, private credit has expanded quickly relative to gross domestic product (GDP), mainly in industrialised economies but also in emerging markets. Real estate, construction, and other non-trading industries have seen a rise in lending as access to credit has decreased in tradable industries (Müller and Verner, 2021). This is also explained by historically low mortgage rates and the increase in demand for real estate. Durdu and Zhong (2021) discuss, in their paper on banking intermediation between banks and non-banks, that there has been a dramatic rise in borrowing over the previous sixty years, with much of this expansion coming from non-bank sources. In this context, non-bank lending encompasses any form of credit extended by any financial institution, this is compared as an asset class in chart 2. Financial institutions include pension funds, traditional banking, foreign entities, mutual funds, insurance companies, and money market funds. Whyte (2018) remark that this rise in private debt is driven by three primary factors that I outline below:

- Pension fund trustees and asset managers are unable to obtain sufficient returns from traditional income assets in a low yield environment. Solvency II's regulatory capital requirements similarly limit the capacity of insurance businesses to increase risk in pursuit of higher returns. Solvency II is a risk-based capital framework for insurers comparable to Basel II in the banking industry;
- Investors are willing to trade yield for the liquidity premium on offer in assets such as private debt with strong default protective clauses as a response to liquidity risk; and
- Banks have curtailed lending to the lower end of the middle market, in part due to industry consolidation and increased regulatory capital requirements imposed by banks.

Preqin (2022) and Nimisha (2010) both report that private debt continues to be a favoured asset class with increases in assets under management (AUM), reaching record levels. Preqin (2022) forecasts that private debt will continue to grow, with AUM more than doubling to $2.69 trillion by 2026, to become the second-largest private capital asset class. To confirm, private debt is comprised of loans, which differ from corporate bonds in that they are more cyclical in nature; both the potential for default (PD) and the loss given default (LGD) are related to and positively correlated with the business cycle. Keijsers et al. (2018) enlighten that, fundamentally, loans differ for the following reasons:

- Bonds do not attain the same level of on-going monitoring achieved by loans.
- Most loans are more senior and secured by collateral that the bank can actively manage.

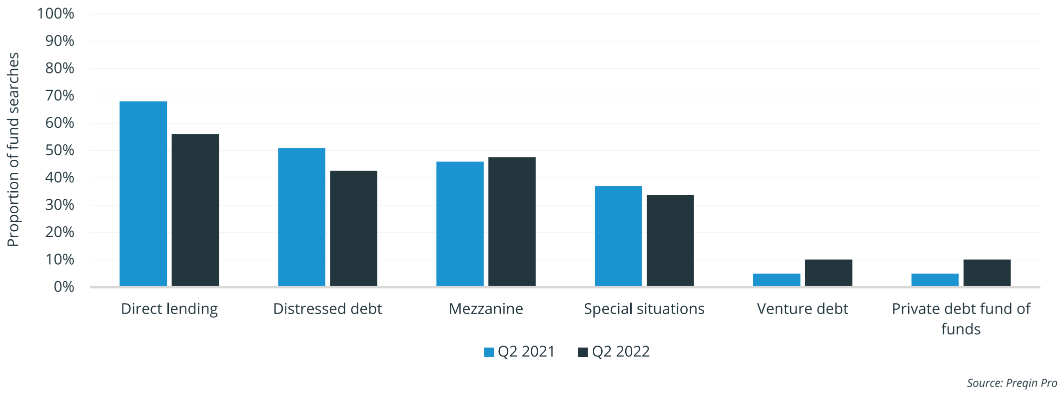


Figure 3: The graph depicts fund types targeted by investors in private debt. Within the private debt space direct lending is already the most significant, and it is expected to increase. If we consider a broader view, direct lending will surpass other asset classes and by 2026 to become the second-largest class of private capital assets, according to Preqin (2022)).

- Another feature relates to the lender being able to hold onto collateral for longer, avoiding the need to sell under fire-sale conditions, where recovery values are stressed.
- LGDs are based on actual workouts, whereas bond studies are implied, often by structural models linked to equity prices.
- The most important difference is that loan losses can exceed 100% and be lower than 0%. Corporate bonds, however, lie strictly within the interval from 0% to 100%. Such is determined in contractual terms, whereas banking loans are subject to a recovery process that can extend beyond that range. This feature means that losses are not normally distributed and will often have the majority of losses with a small loss percentage and a tiny proportion with significant loss levels.
- Bank losses are bimodal in shape, with losses from the loan-recovery process often low (in the range 0% to 20%); and a second cluster of losses closer to a full write-off, ranging from 80% to 100% of the value of the loan amount granted to the bank customer.

Private debt is not without risk. For loan portfolios there is an inherent risk of asymmetric losses due to defaulting credit assets, which can increase in stressed credit markets (Çelik and Isaksson, 2020). Long-term investors, many of whom are more likely to feel the repercussions of the full credit cycle, will find the risk especially strong as they are subject to the full effect of the credit dynamics and credit quality trends. Smith and Balint (2019) interviewed private-debt asset-managers, specifically professionals in the private debt-asset-management industry, asking them to name the risks they were most concerned about. The impact of the business and credit cycle on private debt assets emerged as their highest ranked concern.

Private debt, despite its attractive benefits, is experiencing 'material deterioration in credit underwriting', therefore private debt managers will seek to specialise (Towers Watson, 2018). According to the OECD 2020 report, the quality of non-financial issued credit has deteriorated between 2009 and 2019. Recent evidence suggests that the current loan stock is of lower quality, has larger repayment requirements, and provides inferior investor protection compared to that of past cycles Çelik and Isaksson (2020). Credit sanctioning entities will need to improve their credit scoring and underwriting capabilities for this growing market space. The market has seen an increase in interest-rates over the recent past which has increased the debt-service coverage ratio, making it more difficult to repay debts.

1.4 Risk premium framework

Public finance, modern finance and business cycle theory have all made use of the neo-classical growth model and its various expansions, such as stochastic extensions (Mehra, 2007). Much of our economic intuition revolves around the key concepts of current and future consumption. Traditional models of this kind are often built on the assumption of a trade-off; where an individual considers holding off on consumption due to their perception of future improved pricing of the goods. The desire of a person to substitute between these items is proportional to the relative pricing of the goods, if the firm is capable of offering the choices. Investors are rewarded in the form of a risk premium when they purchase an asset. This is the compensation that investors receive for taking on more risk than they would with a risk-free asset (Schularick et al., 2019). According to Martellini and Milhau (2017), risk premiums are interpreted as an investor's utility of intertemporal consumption and are described in terms of stochastic discount factors linked to variables. Stochastic discount factors (SDFs) in consumption-based models are adjusted to reflect changes in marginal utility (Martellini and Milhau, 2017). This is a crucial link, investors seeking compensation for retaining assets not compensated timeously. In this instance, the investor's wealth is low, but the investor's marginal utility is exceptionally high (Martellini and Milhau, 2017). This is known as a risk premium, which is passed on to the investor for bearing risk. Schularick et al. (2019) posit that log-normal returns with power utility for investors' risk appetite are the most accessible and arguably the clearest explanation. For the introduction to this study I will keep this simplifying assumption: use the risk aversion parameter γ, to determine how steeply the utility curve slopes; this is illustrated below using the constant relative risk aversion (CRRA) utility curve, utility U from per capita consumption c_t, reflected in Mehra (2007).

$$U(c, \gamma) = \frac{c^{1-\gamma}}{1-\gamma} \text{ for } 0 < \gamma < \infty, \ \gamma \neq 1 \tag{1}$$

$$U(c, \gamma) = \log(c) \text{ for } \gamma = 1 \tag{2}$$

Setting γ to 1 defines the function as a logarithmic utility curve[1], which has been widely used in growth models and business cycle models (Mehra, 2007). An intertemporal framework should span several time periods. This framework is an expression in which the marginal utility of one unit now equals $R_{i,t+1}$ units in the future, the foundational connection for applying the Euler equation to value assets (this will be covered in Section 26). This SDF sets up the risk premium framework. Assuming that the growth rate and dividend are independent and identically distributed (i.i.d.) and log-normally distributed (log-normal), the risk-premium as Mehra (2007) outline the SDF as follows:

$$\underbrace{\log E(R_i) - \log E(R_f)}_{\text{Risk premium}} = \underbrace{\gamma}_{\text{Risk aversion}} * \underbrace{\sigma(\log R_i, \log g)}_{\text{Covariance of return \& growth}}$$

The expected return of asset i, the risk free rate, and consumption growth are denote as R_i, R_f and g respectively. In reviewing the CRRA utility characteristics, one can express the rate of increase in real gross consumption per person as follows: $g_t = \frac{c_{t+1}}{c_t}$. If the correlation holds, a risk-averse investor will be less likely to put money into assets that track or 'covary' with consumer spending. For the equilibrium to persist, there must be a reward for taking on this level of risk.

According to Tenreyro (2018), business cycle and macro-economic causation analysis requires sifting through a mountain of data. At times, models must be resorted to make sense of the complex web of relationships between various aggregate-level economic parameters. Expecting to comprehend this intricacy using merely linear relationships would be unproductive. Tenreyro (2018) opine that models are used, even knowing that the underlying data can be sparse, to make sense of the macroeconomy that is highly interrelated, many market-level attributes depending on one another. This process is not the same as reviewing simple correlations to explain causality.

[1] This can be shown by using L'Hospital's rule applied to the CRRA utility model, in equation 1.

1.5 Alternative data signals

Exciting alternative data streams have emerged in financial services as a result of the growth of new forms of data. Improved processing power at a fraction of the cost and hardware that is a fraction of the size greatly facilitate this. In addition, advancements in data storage and a significant decline in the price of data hosting have made data more accessible. Sentiment analysis has indeed become a fascinating field of study with applications in various areas, including business, marketing, politics, and finance. As mentioned, recent research has shown that mathematical techniques can be used to extract emotion cues from large amounts of voice data in order to better understand sentiment and its impact on decision-making. Early indications from Azar and Lo (2019) demonstrate application in financial markets where the Federal Open Market Committee (FOMC) sentiment signals may be extracted using mathematical techniques applied to a huge database of equity-news events rated for sentiment, relevance, and novelty. Azar and Lo (2019) developed the Composite Sentiment Score (CSS), which measures news sentiment. Another study by Feuerriegel and Gordon (2018) shows that signals taken from news language can more accurately predict market performance.

I investigate the use of natural language processing and machine-learning techniques in this study. I also provided a review and comparison of the incremental benefits from the use of alternative data, machine-learning techniques and traditional methods such as econometric techniques. Finally, I directly modelled the use of alternative data in investigating econometric signals based on natural language sentiment modelling. I applied these techniques not only to model against well-known sentiment indices; I also used these indices as modelling variables for three separate regions (South Africa, United Kingdom and United States) over a period from 2000 to 2020, to understand the interaction of business cycles and asset class returns, notwithstanding private debt values.

1.6 Economic scenario generation

A computer-based model known as an economic scenario generator (ESG) will predict the likely range of values for a wide range of economic and financial variables in the future. These scenarios, together with an analysis of the stochastic distribution of scenario outcomes, shed light on the nature of risk factors inside the economy that underpin financial unpredictability. An ESG can, therefore, shed light on the advantages and disadvantages of various operational and strategic options.

Although most ESGs share a common structure, there is some variation in the number of modules included and how they work together. Input, output, and the intermediate components are the three main components of every model, and ESG models are no exception. An ESG simulation can provide a more accurate picture of scenario probabilities, a larger variety of scenario outcomes, and a higher level of scenario complexity than deterministic economic scenarios, econometric models, and macro-finance models only. Risk-neutral (or market-consistent) frameworks are mandated by certain regulatory bodies for the assessment of insurance liabilities. Real-world modelling is more suited to forecasting the future values of economic and financial variables often used in portfolio modelling – this is the focus of this study.

I designed and built a novel ESG that generates pathways to represent the evolution of private debt assets of the investment returns. This is all built as part of an integrated multi-asset and liability modelling program that takes as input the current states of the business cycle effect and extrapolates for the full term of an investment. There is special focus on the underlying fundamentals that determine a credit asset return and cost of credit using sophisticated modelling techniques to cater for market anomalies, such as the credit puzzle and correlation breakdown. I created a framework for the relative comparison of asset-class returns in a non-trivial way. This is useful in designing investment strategies in private debt for 'use-cases' such as pension fund investing requirements.

1.7 Portfolio construction

The prior sections come together in the portfolio construction section. Portfolio construction is an exercise in understanding how asset classes, funds, and weightings affect performance, risk, and investor goals. One must examine all assets, investments, and debts before designing the portfolio investment and setting short- and long-term financial goals. Given that this exercise only deals with asset classes in aggregate at an index level, the purpose of this analysis is not to select specific securities. After establishing a risk-return profile, I construct a diversified, high-return asset allocation strategy using optimisation programs. In this study, I split out the portfolio construction phase into two parts. Firstly, I create two different building-block portfolios constructed with the aim of reducing downside risk in the form of drawdowns. The first building block is based on core fixed-income, I compare the various results in the introduction of credit loans (investment grade and sub-investment grade asset sub-asset classes). The final section of portfolio construction is on investigating the optimal planning for long-term investment needs. In this study I probe the use of reinforcement learning techniques and this is an important link in this study, as the core of these methods is directly related to the underlying theory we established in the asset-pricing section (Part II) that I explore in the next section.

1.8 Contributions of this study

A summary of the work documented in Part III of this document has been accepted for publication in the ORiON journal[2], which is the official journal of the Operations Research Society of South Africa (ORSSA) and is published biannually.

[2] https://orion.journals.ac.za/

Part II

Macro-finance

A simple question can be posed, that reveals the natural human aversion to certain risks, elucidates Cochrane (2017): if equity shares are so much more rewarding on a cumulative basis over long horizons, then why wouldn't an investor only hold equity shares? The answer seems to lie somewhere in how we, as humans, perceive risk. Cochrane (2017) points out the key challenge is not one of seeking to explain facts or understanding what is happening now but rather to find verifiable, quantified and theoretically sound measures that can be used to express risky outcomes and account for asset price puzzles and fluctuations.

Macro-finance is the study of macro-economic fluctuations and their relationship with asset-prices (Cochrane, 2017). Practically, this refers to stocks and bonds, where price growth is improves in good times and depreciates in stressful periods, or where real and nominal rates fluctuate over the business cycle. No model can adequately explain the fundamental sources of risk that account for returns and as Campbell and Cochrane (1999) explain, standard economic models cannot replicate these cyclical features perfectly. Current economic conditions may have a profound influence on future predicted pricing, and this section outlines some of the fundamental theories and models in this area. It would be an oversight not to speculate on the historical evolution of economic cycles and the underlying ideas in such a complex and fascinating field. In this section I review business cycle theory and risk premia models that attempt to explain business cycles.

2 Historical introduction to business cycle theory

Early business cycle studies were all concerned with the instability of inventories, fixed capital, and real asset purchases financed by a supply of institutional credit, according to Zarnowitz (1992). Borio (2012) explains that financial cycles also include attitudes towards risk, financing constraints, and perceptions of value and risk. All these factors result in cycles of expansion and contraction over time.

Business cycle theory has two broad schools of thought: firstly the classical (neo-classical or supply-side) school, which is anchored in the view that macro-economic changes are driven by exogenous causes such as the impact of monopolies, natural causes, unions, or the impact of government (Zarnowitz, 1992). The second school is the Keynesian (or demand side) school that holds the view of a change in the macroeconomy is explained by known features, or endogenous causes. One such theory, the financial instability hypothesis in the work by Minsky (1978), looks at the dynamics of firms and bank lending. In this model during a period of growth, low interest rates encourage increased borrowing and excessive debt, which ultimately cause the economy to move into a recession. Since the great depression in the 1930s, Keynesian economics has been the primary view from which business cycles have been understood or explained. Since the recent great financial crisis (GFC) starting in 2007, there has been a resurgence of neo-classical approaches, like the real business cycle (RBC) theory.

Dobrescu et al. (2012), another literature review on business cycles, starts with Austrian economists Ludwig von Mises and Friedrich von Hayek (thus the name of the class of business cycle models they developed), proceeding to today. The model assumed that low interest rates would allow for credit expansion, but that new business owners would take advantage of this opportunity to make more, riskier, investments thanks to the reduced cost of credit. This would set up a cycle in which risky investments would be financed further, leading to an unsustainable situation. A recession would correct such a situation, and a subsequent credit crunch would be the result.

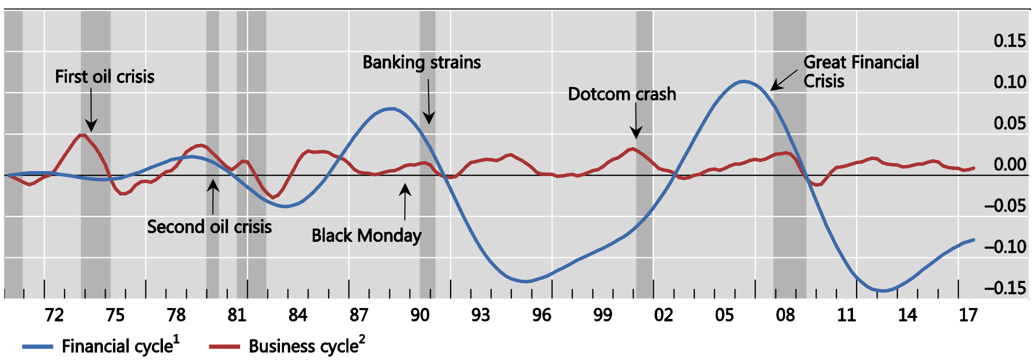


Figure 4: Graph 1 from Filardo et al. (2019) showing the different cycles; included are the business cycle and financial cycle in the USA. The terms 'business' and 'financial' cycle are clearly different (please refer to Filardo et al. (2019) for detailed definitions of how these indices are compiled).

From the 1960s onward, Keynes's famous writings emphasise the importance of money and policy, but they only care about large-scale shifts in demand, which are tied to inflexible shifts in output and wages (Zarnowitz, 1992). Real-world data does not align with this view, leading to the more recent advances of the 'neo-classical' or 'new-Keynesian' models driven by the need to match basic empirical data using stochastic dynamic models. The next part is to accept the micro-foundations are consistent with the utility-maximising behaviour of firms and individuals. Zarnowitz (1992) notes that this ramped up the level of complexity, and looking to solve both is a more than ambitious task.

In a review of business cycles one should understand fully how the particular business cycle index is compiled; there are differences between the global and local business cycles, according to Borio et al. (2019). The domestic cycle covers a particular country, with the global cycle covering them all, including cross-border co-movement. Stiglitz (2017) explain that open economies, with their short-term expectations on capital and volatile flows of capital across borders, have arguably led to higher levels of macro-economic volatility. Borio et al. (2019) go on to explain that the domestic cycle covers credit and housing prices, often an aggregate of the two. The global cycle, however, also includes debt and equity, as well as cross-border flows and financial market prices. Herwartz et al. (2020) explains that co-movement of the business and financial cycles, in particular contractions, can lead to more severe changes in the business cycle.

2.1 Theories of friction and exogenous shocks

In this section we review two recognised classes of macro-economic model: neo-classical and Austrian models. The key point of difference between the two classes is the interpretation of the mechanism that is said to cause key variables (productivity and consumption) to fluctuate in a business cycle. Neo-classicalists believe that the cyclical evolution of macro-economic variables is driven by the general equilibrium framework, whereas the Austrians believe that the business cycle is a systematic inter-temporal imbalance (Dobrescu et al., 2012). López-Salido et al. (2017) explain that many of the mainstream models (Bernanke and Gertler (1998) and Kiyotaki and Moore (1997)) follow this class of framework, where financial market frictions play a central role in the propagation and amplification of shocks in the economy. Also the central assumption is that role players are all seen to be rational. This assumption is often questioned in macro-economics, especially since the GFC.

López-Salido et al. (2017) go into greater detail on how these models are focused on balance-sheet metrics, credit growth, and leverage. Central focus in this theory is the control of credit limits (as a function of exogenous frictions or endogenous borrower limits), with debt contracts being the primary mode of finance. This relates to the fact that the key mechanism and propagator of amplification or a credit cycle are exogenous shocks, however there is not much detail around the driver of a credit cycle or how it is expected to be triggered.

Bernanke and Gertler (1998) and Kiyotaki and Moore (1997) incorporated agency issues between borrowers and lenders to represent financial market frictions as endogenous variables. A direct result of this is a rise in the price of both external financing and credit. Bernanke and Gertler's model of the economy and finance contributed greatly to our comprehension of the credit market. The model was a neo-classical model using a balance-sheet approach, reviewing the impact of macro-economic fluctuations on real variables. Economic expansion improves the net worth of borrowers, causing a reduction in demand for funding and therefore the cost of borrowing, and amplifies investments, this has become known as the 'financial accelerator' (Bernanke and Gertler, 1998). It establishes a connection between the cost of external financing and the level of actual output for fixed-income assets. Since the external financial premium is inversely proportional to the financial strength of the borrower, widening credit spreads is associated with the widening of the external financial premium.

The real business cycle (RBC) belongs to the category of new-classical[3] models that consider economic fluctuations to be real rather than mere nominal shocks. Kiyotaki and Moore (1997) explains that RBC is a coherent dynamic stochastic general equilibrium (DSGE) that is efficient due to the use of exogenous shocks and that can be calibrated to reflect a very wide range of economic indicators. The RBC models may also incorporate aggregate utility functions, where Kiyotaki and Moore (1997) use a log-utility function of wealth levels. RBC, on the other hand, disregards credit and money issues and does not account for pricing variations, large unexpected changes in production technology with no evidence and that unemployment is a reflection of the fact that individuals would prefer to work less. This is simply not plausible in an economic crisis (the very thing it seeks to explain). According to Kiyotaki and Moore (1997), what is difficult to detect in the data is a major external shock to aggregate productivity as the underlying mechanism for triggering the RBC. This challenge remains true for many macro-economic models today.

2.2 Theories of behaviour and endogeneity

As soon as the theory of rational expectations was adopted in the 1970s, mainstream economics largely overlooked the impacts of psychological effects once, explains Sahin (2021). Prior to Keynes, there was a focus on business uncertainty and the associated volatility of

[3] Not to be mistaken here with Neo-classical models.

expectations. Life-cycle and income theories of consumption are linked to consumer sentiment, which is a better indication of the future of income than an extrapolation based on the recent past. Zarnowitz (1992) explains the notable divergence between accelerator multiplier models that focus on endogenous factors featuring non-linearities and lags, and the exogenous models that are driven by shocks. This divergence was primarily driven by the adoption of mathematical techniques in the 1930s and 1940s. The early macro-economic models then adopted exogenous stochastic approaches. This is perhaps explained by the absence of readily available technology and processing power to deal with the non-linearities in sophisticated analytical and econometric techniques. Borio et al. (2019) argue that the discovery of endogenous linkages between policy measures and sensitivities or weaknesses in an economy that allow for a more complete understanding of the dynamics of boom and bust cycles is crucial for long-run macro-financial stability. Nowzohour and Stracca (2017) cite the 2008 financial crisis as proof that extreme shifts in public opinion can have major effects. There are three competing frameworks to explain the transition mechanism for shocks and to explain how sentiment is related to the business cycle, following Nowzohour and Stracca (2017):

- **Irrational animal spirits.** This term was coined by Keynes (1936), where business cycles, which are linked to pessimism and optimism in waves that cause macro-economic fluctuations. The implication is that animal spirits lead to booms and busts even when the fundamentals are not reflecting the same measure of change.
- **Self fulfilling animal spirits** also based on the belief that sentiment drives macro-economic fluctuations. This is caused by sentiment waves (pessimism and optimism), which then cause expectation changes and the change to materialise.
- **News signals** show that agents have access to details of the economy and future developments, implying that there are recurrent booms and infrequent busts linked to correct signals and incorrect signals, respectively.

2.3 Modelling of business cycles

One would be pleased to have a scenario where the models that have been developed for macro-economics both fitted the data well and had a tight elegant theoretical structure (Blanchard, 2018). The reality is that the theory is often complex, abstract and at times simply does not fit the data well. From working in this field and in other quantitative fields, I always find it surprising how long one needs to consider and work with data (challenging to find, hard work to organise and estimated parameters are often unstable overtime). When working with economic data, or any data that is the aggregate of individuals' behaviour into one measure or attribute, Blanchard (2018) explains, the process of aggregation needs to be carefully examined as the underlying correlations can be more complex than an aggregate proxy reports. On this, there are many modelling techniques in macro-economics, and one form of modelling is not better than another. Models can be seen as having a part to play in the process of understanding the inner workings of a large and intricate system, that is the macro-economy. According to Blanchard (2018) five broad classes of models are found in macro-economics (and, applicable to macro-finance).

- **Foundational.** This model class is intended to establish deep theoretical learnings that are conveyed. This may include such models as the equity premium model and the overlapping generational model (Blanchard, 2018).
- **Dynamic stochastic general equilibrium (DSGE).** This group of models is designed to dig into the consequences of economic distortions. These models are useful to policy makers and researchers when discussing macro-economic functions. Blanchard (2018) explains that the model design should be based on a central model that should the agreed-upon base model, this is currently the RBC model and this seems to be at odds with reality. Another important consideration is the use of forward-looking models based on Euler equations and rational expectations (which may be too focused on a forward-looking approach, can be an agent-based approach that relies heavily as network effects). These models really aim to align with reality, as Blanchard (2018) explains, DSGE models require more accurate fitting but are useful as policy models.
- **Policy.** As one can infer, policy models are needed to support policymakers with decisions as they pertain to the economy at large. The models will provide answers to hard questions, such as 'if production in China slows down, what is the effect in South Africa?'. These models need to capture dynamics by using shocks and tracing the effects of the insertion of policies into the same model.
- **Toy.** This class of models allows for a quick first pass question and answer approach and this pedagogical approach that is often featured in undergraduate text books, examples of this include the original RBC, IS-LM models and New Keynesian model (Blanchard,

2018).

- **Forecasting.** This class needs no introduction. With the primary objective of providing the most accurate forecast, this is the territory of statistical methods that can be over-parameterised (Blanchard, 2018).

I have chosen a few of sections that I would like to explore further, as these relate to topics in this study.

2.3.1 Dynamic stochastic general economic (DSGE) models

DSGE models first appeared with the introduction of the RBC model by Edward C. Prescott (1986). In her speech Tenreyro (2018) starts by detailing macro-economic models, starting with John Maynard Keynes in his famed paper ‘General Theory’. Tenreyro (2018) explains that this theory, with its clear focus on demand management, set a new course and most macro-economic models now follow this philosophy. From this, the well-known IS-LM model stems. Tenreyro (2018) goes further to explain that large-scale models in the 1970s failed, partially due to stagflation and oil shocks that bore witness to the Philips curve breakdown. The general models were further reassessed with the introduction of the ‘Lucas critique’[4].Tenreyro (2018) and Blanchard (2018) separately explain the first prototype DSGE model, developed by Finn Kydland and Edward Prescott, called the RBC. The model only had a few variables (consumption, labour output, investment input, and labour productivity) and had no role for central banks. The real benefit of the RBC at the time was the way it connected the theory to the data by using exogenous shocks into the model that are linked to known functions in the economy. Shocks are introduced to preferences, technology, or market structure. The model’s outputs gradually failed to match the data. Since then, DSGE models have converged and are often referred to as ‘post-Keynesian’. With the introduction of an assumption that firms would take time to adjust to policy changes or shocks, then model frictions where introduced to the model. These new model features began to match data and have now elevated themselves and avoided criticisms from the 1970s. Blanchard (2018) explains that a DSGE model built with clear micro-economic assumptions makes the models easier to extend and relate to market dynamics. In terms of frictions being endogenous, the

[4]The Lucas critique, named after Robert Lucas’s economic policy work. Policy acts will impact relationships, so one can assume that stable relationships in historical data are not independent of policy (Campante et al., 2021).

RBC approach is aligned in this assumption.

Kiyotaki and Moore (2019) use a dynamic stochastic general equilibrium (DSGE) approach to model the economy. They use a model that looks at how the capacity to sell assets affects market liquidity overall. The model incorporates a premium between cash and stock returns and it is assumed that this difference in risk premiums is explained by variances in the assets' liquidity and resale value. It is assumed in the DSGE model (where the shocks are calibrated with long-run US data in an structural vector auto-regression (SVAR) model) that the bond-market channel cushions an economy alongside bank lending, reducing amplification by 10%-70% relative to a world without this channel (Drechsler et al., 2018).

I also reviewed how accelerator multiplier models focusing on endogenous factors with non-linearities and lags diverged from classic exogenous models driven by shocks. Many of these models are developed using highly specified vector autoregression (VAR) and econometric approaches. I will summarise a few studies in this areas for context below.

2.3.2 Econometric and VAR modelling of business cycles

The introduction of econometric and VAR modelling of business cycles can be directly linked to the progression of mathematical and statistical techniques that emerged in the 1930s and 1940s. Until more recently, the available technology and processing power was either too expensive or not powerful enough to deal with the non-linearities in sophisticated analytical and econometric techniques. This section will summarise some recent econometric studies that are related to business cycles. There is a significant body of available research, which is briefly focussed on here to provide context.

Using a Bayesian VAR model, Evgenidis and Malliaris (2022) examine and analyse financial stability from the standpoint of a central bank. As part of their research, Evgenidis and Malliaris (2022) evaluate how unexpected events affect consumer sentiment and overall spending patterns in the economy. The BVAR model utilises data segments including credit, asset prices and spread, altogether, the model totals fifteen parameters. To start, the structural VAR models are specified, per Evgenidis and Malliaris (2022) as follows:

$$y_t = c + \sum_{j=1}^{p} y_{t-j} B_j + \varepsilon_t \tag{3}$$

Where c is a vector of constants, y_t is an endogenous variable matrix, that is a function of p-lags on itself and other endogenous variables in the system. Thus B_j is the coefficient matrix, and finally the residual covariance matrix is captured in $\varepsilon_t \sim N(0, \sum)$. Evgenidis and Malliaris (2022) point out that since the above specification is based on the ordinary least squares method and suffers from the curse of dimensionality, they use Bayesian shrinkage methods to solve it, hence BVAR. Evgenidis and Malliaris (2022) perform an extensive econometric analysis of business cycle tracking, finding that a rise in the non-financial sector shock leads to a temporary drop in industrial production, but a more sustained drop in economic activity over the longer term. Interestingly, their deeper investigation reveals that this decline is attributable to financial distress and consumer mood, as opposed to a systemic monetary contraction caused by a credit shock. Finally, Evgenidis and Malliaris (2022) discover that, across the board, unique shocks to credit and asset-price variables have a large effect on manufacturing output. It is shown that the BVAR model can be used as an early warning system for economic decline. Evgenidis and Malliaris (2022) conclude that the model demonstrates that decreases in economic growth associated with a credit shock is due to an increase in financial distress and a fall in consumer confidence rather than a systemic monetary contraction caused by the credit shock.

Financial and macro-economic indicators provide valuable information both before and after the GFC. Kucera (2017) models economic shocks using economic and financial variables in a vector auto-regression (VAR) model to explain this phenomena. The sentiment of the financial markets has been a major contributor to understanding the movement of term premiums in recent years. This is clearly demonstrated by the fact that bond yields have steadily dropped from the highs in the 1980s to historical lows since the GFC despite improving economic conditions and optimistic forecasts for the future of interest rates. Sentiment may be a suitable measure in relation to financial variables. A good example is the time of the Brexit vote, as this was a financial period where economic and financial metrics were favourable, whereas the sentiment variables at the time would have not signalled the same conditions.

Afanasyeva et al. (2020) I built a model in a similar way, structured as a BVAR on macro-factors. This model factors in an empirical review of the Minsky hypothesis, which describes the build-up of a buoyant economic environment where risk tolerance changes, expectations increase and constraints loosen. Afanasyeva et al. (2020) seek to develop an approach to quantify this relationship. The model is based on macro-economic fundamentals including the output gap, GDP growth rate, and unemployment gap, and the Chicago FED's National Financial Conditions Index (NFCI) measures financial stability. Their findings show that buoyant financial conditions are early signals of economic stress and that business cycle shocks are not insignificant contributors to non-financial leverage.

Pesaran et al. (2003) study lenders and discuss that they face commercial fluctuations due to the business cycle. Pesaran et al. (2003) build a model or modelling framework that links credit assets and measurable macro-economic inputs. This model shows a difference in risk between firm-specific (or idiosyncratic) and systemic risk. Their model is based on a set of publicly traded loans in twenty countries globally, including Western Europe, South East Asia, the Middle East, and Latin America. Pesaran et al. (2003) used a GVAR framework as part of an arbitrage pricing theory (APT) model to isolate firm-specific risk. That approach has the added advantage of retaining the correlation structure of the macro-economics variables under investigation. Pesaran et al. (2003) make use of a Merton approach to recognise the value of a firm, using equity returns and volatilities as proxies. The firm-specific default probability model is denoted as, following Pesaran et al. (2003):

$$r_{ji,t+1} = \mu_{ji,t} + \xi_{ji,t+1} \tag{4}$$

Where $r_{ji,t+1}$ is the return of a firm j in region i over period t to $t+1$, the conditional mean μ is forecastable, but ξ is not, for all regions i over time t. The μ part is composed of firm-specific factors that can be modelled, whereas ξ is the unexpected part that is influenced by the effects of shocks, whether macro-economic, firm-level, or global shocks that are exogenous, or outside of the model.

Pesaran et al. (2003) find that symmetric shocks do not result in symmetrical loss. They also find that losses are not proportional either. The Pesaran et al. (2003) model is capable of producing multi-period forecasts that cover the entire range of losses or distributions that are conditionally linked to the macro-economic environment. Another approach was taken by Prabheesh et al. (2020) who use the SVAR model and a dynamic conditional control (DCC) generalized auto-regressive conditional heteroskedasticity (GARCH), (DCC-GARCH) model to analyse the relationship between the global financial cycle and the domestic cycle of Indonesia and India. In the case of India, the global credit cycle was strongly synchronised with the domestic credit cycle. Dees (2016) reviews international spillovers with macro and financial variables related to credit supply shocks. He uses a GVAR model applied to the global markets. Dees (2016) investigates the link between business and financial cycles using a GVAR model. The GVAR model can be expressed as a vector of country VAR models, following Dees (2016):

$$\Phi_i(L, p_i)\boldsymbol{x}_{i,t} = \boldsymbol{a}_{i0} + \boldsymbol{a}_{i1}t + \Upsilon_i(L, q_i)\boldsymbol{d}_t + \Lambda(L, q_i)\boldsymbol{x}^*_{i,t} + \boldsymbol{u}_{i,t} \tag{5}$$

Where $\boldsymbol{x}_{i,t}$ is a vector of modelled variables that are constructed as country-specific trade-weighted averages over the values of the other countries, $\boldsymbol{d}_t$ is a vector of international macro-variables, $\boldsymbol{x}^*_{i,t}$. $\Phi_i(L, p_i)$ are the domestic and country specific foreign variables and the $\Lambda(L, q_i)$ is the lag operator L of $\Phi_i(L, p_i)$. Finally, the $\Upsilon_i(L, q_i)$ vector is the set of country stresses, as part of the idiosyncratic shocks, d_t and $u_{i,t}$ is a vector country shocks. Using a modelled effect of credit and asset prices, Dees (2016) evaluates the true economic activity in 38 nations. The model illustrates the connection between credit and asset-price swings and their significance in explaining observable shifts in the economy. International shock propagation can be substantial and long-lasting. Dees (2016) confirms that global financial cycles are heavily impacted by the US economy and have strong impacts on real variables.

2.4 Credit, lending and business cycles

According to Azariadis (2018) a business cycle is characterised by substantial cross-correlation between credit, output and the persistence of credit over the course of a business cycle. The research of Azariadis (2018) looks into the possibility that shocks in people's confidence in the credit market can trigger cycles. Credit is highly pro-cyclical, as demonstrated by the research on unsecured debt in a business cycle (Azariadis, 2018). Gilchrist and Zakrajšek (2012) examine the inner workings of the GFC, when credit spreads transmit vital signals about the state of the actual economy, economic activity, and asset prices. Substantial evidence now suggests that risk premiums change over time and are countercyclical, as explained by Gourio (2013). Recession-related optimism and pessimism can be part of the time-varying views discussed in Gourio (2013), which is very different from an objectively measured probability of disaster models.

Credit spreads are an excellent early warning signal and can be used to anticipate business cycles, according to Berndt et al. (2018). They find that credit-risk premia are subject to timing and pro-cyclicality and are heightened by time-varying market (il)liquidity. Berndt et al. (2018) estimate corporate credit premia by considering the difference between the true cost of credit and the (priced) default losses, which provide a healthy additional premium, also known as the "Credit Puzzle". Following Berndt et al. (2018), credit premiums can be calculated using the credit default swap rate and compared to the actual losses from sold reference assets.

Herwartz et al. (2020) advance the understanding of the source a credit shock and its link to the real economy and business cycles in their study. They see demand and supply shocks in credit working slightly differently, but both play a role. The credit demand shock relates to the exogenous factors affecting the borrower's preferences for taking credit, and the supply shock is when the lender's appetite for lending increases. Both shocks would increase the volume of credit in the market, but the effects are different depending on whether they were driven by a supply shock or a demand shock. Herwartz et al. (2020) point out that it is understood that supply shocks driven by intermediaries' increasing volumes affect business cycle fluctuations, and spreads are used to signal changes in stressed markets. They observe that second area of understanding is where the composition of credit matters, be it in private or public markets. Public credit is funnelled to government activities, whereas private credit is directed to household lending activities, which differ in productivity. Herwartz et al. (2020) note that shocks to private lending have had negative consequences for household consumption in the medium term (such was the case for the GFC). The impacts of shocks on public credit volumes are not as clear. Herwartz et al. (2020) build a factor model to link global liquidity to the proposed underlying factors or drivers. Modelling by Herwartz et al. (2020) can explain significant proportions of the fluctuations of financial and monetary indicators. In stressed market conditions, public entities are known to require credit, whereas private individuals and businesses tend to borrow more in favourable market conditions or in the run-up to crises.

Müller and Verner (2021) explore the impact of different sector borrowing effects on financial fragility and business cycles. Their explanation of the theory of the economy provides three links, in an open economy, to boom and bust cycles driven by sectors. The paper reviews the relative impacts of sectoral allocation as it pertains to cycles. First, household credit and non-tradable firm credit can set up unsustainable demand booms that end in busts. Secondly, the relatively non-tradable sectors contribute to fragility, and lastly, allocation to non-tradable sectors may cause inefficient allocation of resources across sectors, given the tradable sectors' higher productivity. Müller and Verner (2021) find that with aggregate credit growth, the relative lending activities in the non-tradable sectors increase. This is also associated with a lower growth rate than tradable lending. When I review the impact that non-tradable sector lending has during boom times, with increases in consumption relative to GDP, non-tradable activity, and exchange appreciation, the enabling factor can be attributed to credit expansions in the non-tradable sectors. Naturally, if the lending goes to financially sensitive firms, financial fragility is likely to be the result. According to Müller and Verner

(2021), the non-tradable loan non-performing rate was double that of the tradable sector at the height of crisis periods, and most of the losses were from the non-tradable sector. Although an increase in firm debt is typically associated with economic underperformance in the long-run, this is not always the case because the sector one lends into is a factor, as pointed out by Müller and Verner (2021). This is because of the relatively higher growth rate of the non-tradable firms, financing moves into these sectors where medium-term labour and total factor productivity are lower. Increasing household and non-tradable debt can lead to a real appreciation of non-tradable products, increased labour costs, and a decline in competitiveness, according to Müller and Verner (2021). Collateralised debt with real-estate collateral is also far more widespread in the non-tradable sector, as evidenced by statistics from 56 nations (Müller and Verner, 2021); this may imply riskier debt due to the reliance on collateral to cover the debt exposure.

Durdu and Zhong (2021) review the fragility in the lending systems, as part of a DSGE model design. Their interests include understanding the links between the business cycle and interest rates on non-financial lending. One of the key areas of analysis is how the lending activities of entrepreneurs are linked to shocks. Durdu and Zhong (2021) show that the entrepreneurial shock accounts for more than 50% of the changes to lending levels. They also argue that as much as 70% of all lending is linked to inter-sectoral stresses and shocks. They also find that entrepreneurial shocks are specifically important to the non-financial lending sector, even more so than sectoral shocks as a factor.

Gulen et al. (2019)'s study accurately depicts the influence of credit sentiment on corporate finance for both publicly traded and privately held companies in the United States. They suggest using analysts' revision and forecast restatements to estimate disappointment (a measure of expectation). Credit market sentiment is connected with earnings estimates by analysts (Gulen et al., 2019). Gulen et al. (2019) have shown that an increase in analysts' expectations is linked to both the growth in credit and the decline in credit quality the following year. Systematic downward revisions in analysts' expectations follow credit upswings.

2.5 Sentiment and credit cycles

Consumer sentiment is often treated as an independent variable where the law of large numbers would have the impact of cancelling out the individual effects (Katona, 1947). Then, Katona (1947) goes on to summarise, at times aggregate behaviour measures fail because we know that correlations exist and the influence of media or other outside influences can drive aggregate behaviour. In an age of mass social media, I can think of an example, the enabling force that a free platform such as Facebook had during the Arab Spring (Abdulrahman Naef Farhan and Varghese, 2018). Returning to an investor context, Sahin (2021) contends that as consumer sentiment decreases, so does the perception of financial distress, with investors moving toward safer asset classes.

Sahin (2021) concludes from the literary evidence that consumer confidence is fundamentally linked and affects savings, borrowing, consumption, and therefore income. Sahin (2021) builds a stock and flow model that captures this dynamic and tests the effects on various economic factors. The consumer confidence indicator (CCI) is the variable of interest among many others. All readings above 0 are considered to be positive, with a normal reading sitting at zero on the CCI's scale of -1 to 1. In the Sahin model, CCI is used as an exogenous variable in the logistic function, as follows:

$$\rho = \frac{1}{1 + e^{\gamma(CCI-1)}} \tag{6}$$

γ is a sensitivity parameter, ρ is the household target wealth durables proportion. Sahin (2021) shows consumer sentiment shocks and scenarios in the model to explain changes to aggregate household income, consumption, and leverage, highlighting that this model considers portfolio choice, consumption, and borrowing behaviour. Macro-economic variables in a particular country can be materially impacted by consumer sentiment, with negative feedback loop delays coupled with strengthened positive feedback loops. This is reminiscent of aggregate credit extrapolation and in this study I adopt CCI to denote a credit cycle variable.

The work of Correa et al. (2020) is a review of sentiment using the text from the IMF and ECB financial stability reports. These reports present key fundamentals and commentary from 63 countries with great relevance to central banks. The text analysis was performed using natural language processing (NLP) methods, using simple dictionary methods for mapping words to sentiment valence. In their case, they created their own financial stability dictionary using Harvard's general inquirer and Tim Loughran and Bill McDonald financials dictionary, adding 30% more words using this dictionary method. This was used in the modelling to create a set of financial stability sentiment (FSS) indexes. Valence is a numerical ranking construct (which is simply applying low ranking buckets to low FSS scores and visa versa) for the sentiment values. The FSS index is defined as:

$$\text{FSS}_{t,j} = \frac{-\text{ve words} - +\text{ve words}}{\text{Total words}} \tag{7}$$

$-ve\ words$ ($+ve\ words$) refers to words that have been mapped to a negative (positive) sentiment. For each time period t and country j, the index increases when the proportion of words that are negative grows, showing higher levels of negative sentiment during those times. Correa et al. (2020) investigate what is driving the FSS index using a topic analysis of the underlying text that produces the FSS index. This is done with the use of sub-FSS indices based on topics. Separate indices are created in the same way as the Economist r-index (indicator the Economist created for use to measure the severity of a recession; uses the proportion of words in a corpus containing recession as a weighted measure of overall word usage). The topics indexed are external, banking, real estate, valuation, household, corporate, and sovereign (refer to Correa et al. (2020), page 100 for the full definition). Their study covers a few areas, but the question of interest in this study is how the FSS index created from NLP methods is related to the financial cycle. Correa et al. (2020) modelled using a panel VAR model, with the dependent variables being financial variables and their FSS index as a time varying function of the topic indexes for each topic j and country i. The topic attribution model is broken down by country and time period by the following model.

$$\text{FSS} = u + \sum_{j=1}^{S} B_j \text{FSS}^j + \sum_{j=1}^{S} C_j \text{Freq}^j + e \tag{8}$$

B_j is the sensitivity coefficient per topic j. Freq is the frequency of each topic found in the report, and C then becomes the sensitivity of FSS to this variable. At the 1% level of significance, the following topics are shown to be statistically significant.

Banking	External	Valuation
Household	Corporate	Real estate

From that confirmed set of data, it was used to develop a panel VAR model, denoted as:

$$X_{i,t} = c_i + \sum_{z=1}^{p} X_{i,t-p} A_l + e_{i,t} \tag{9}$$

c is the vector of country parameters, $e_{i,t}$ is the vector of idiosyncratic variables, and $X_{i,t}$ is a vector of the dependent variables that includes variables related to the financial cycle and a FSS index. p is the number of lags in the autocorrelation exercise. A_l is a matrix that was estimated using a generalised method of moments (GMM) procedure. Correa et al. (2020) find that a protracted decrease in the financial cycle follows the sentiment decreases. Similarly the debt service cover ratio, asset prices follow the increases in FSS index. Correa et al. (2020) also finds, using the probit model, that the sentiment of the central bank communication is a significant predictor of the GFC. When viewed through a different lens, one that uses the credit-to-GDP ratio as a proxy for the financial cycle (therefore offering more data points than the recessions data in a typical financial cycle variable), their model shows that FSS has considerable predictive capacity or model power. Correa et al. (2020) conclude that central banks communicate crisis effectively, this is evidenced by the increase in their FSS index signals a sentiment risk and the real risk of financial cycle reductions.

Chen et al. (2021) investigate the connection between a macroeconomy with particular focus on sentiment, which they define as the way people are inclined to be negative (positive) in low (high) sentiment values. Chen et al. (2021) observe that sentiment is influenced by macro-variables and, in turn, influences macro-variables. For this reason sentiment is seen as the bridge between financial markets and macro-variables. Chen et al. (2021) point out that the relationship between financial markets and sentiment is well-known (such as certain sectors rallying during high-sentiment regimes). More recently, the focus has turned to how sentiment and the macro-economy are linked (please refer to Chen et al. (2021)'s introduction for further examples in finance literature). Often, the approach to sentiment modelling in financial literature is to treat variables as exogenous, which means that there is no inquiry into what determines market sentiment. In the Minsky hypothesis, during an economic upswing, payments for firm debt are made earlier and the resulting euphoria and positive sentiment get firms borrowing further (Chen et al., 2021). In general, when debt levels rise, interest rates fall, they go further to explain that at a certain threshold, the debt burden becomes a binding constraint, with leverage levels too high; resulting in a financial crisis. Chen et al. (2021) construct a model by using a the DSGE design. Sentiment is incorporated using the expected premium (rate differential) and the connection to the state of the economy. Sentiment is seen as the expected profit rate differential, ρ and modelled as follows.

$$\rho_t = \gamma_0 + \gamma_1 \rho_{t-1} + \gamma_2 (i_{t-1} - \iota_{t-1}) + \gamma_3 r^s_{t-1} \tag{10}$$

Where ι and i are nominal interest rates, long and short term rates respectively. r^s is the stock market return over one period ($\log P^s_t - \log P^s_{t-1}$). The measure of lagged sentiment is γ_1 and γ_2 is the interest rate. Lastly, the effect on the stock market return on sentiment is

γ_3. Chen et al. (2021) explain that when investors plan for the long term, they use estimates of growth rates that includes an expectation of future dividends, $E[\text{Div}]$ and $\Delta\rho_t$ denotes the change in ρ_t. The modelled sentiment is then incorporated into the expected future stock price, as Chen et al. (2021) describe in the following model:

$$E_t[\text{Div}_{t+1}] = (1 + c_p\Delta\rho_t)\text{Div}_t \tag{11}$$

This model structure then allows for sentiment changes to be directly connected to the valuation of a security, through the sensitivity of sentiment to lagged sentiment, interest rates, and stock price increases. Gupta et al. (2021) explain that there was widespread acceptance that sentiment played a role over the course of the GFC. Gupta et al. (2021) argue that policy shocks are known to impact markets due to their effect on prices due to changes in expectations. This mechanism can influence the cost of capital or preferred discount rate that is used to calculate a present value of future cash flows. These impacts will have a bearing on market risk premiums and valuations.

Their model is set to review regime changes when considering the impact of monetary policy shocks. The model is an econometric VAR on panel data with sentiment and macro-variables. Gupta et al. (2021) reason that the high-sentiment states are subject to mispricing mechanisms. It is at this point that noise traders enter and the role of animal spirits, herd behaviour, and over extrapolation occur, driving a period of overvaluation where the macro-fundamentals don't adequately explain the returns. Greenwood et al. (2016) explain that agents sustain the incorrect perceptions with a view of 'this time will be different' perception. One of the main claims of the Gupta et al. study is that it details a new transmission mechanism by which policy shocks might affect stock prices: the behaviour channel. Given the rational investor, both high and low sentiment regimes should garner a similar response, and this was tested. The VAR model in Gupta et al. (2021) specified as follows:

$$X_{i,t} = \alpha_i + P(Z)_i X_{i,t-1} + R_i\varepsilon_{i,t} \tag{12}$$

Where $X_{i,t}$ is a vector of endogenous variables like GDP and consumer confidence and α is a vector of country constants i. $P(Z)_i$ is a back shift operator for the matrix of polynomials and finally the contemporaneous disturbances denoted as $R_i\varepsilon_{i,t}$. Gupta et al. (2021) test their hypothesis that the policy impacts on prices are more effective during times of high sentiment, using the VAR model. This evidence supports the hypothesis for the entire panel of data and at the country level, which is, on the face of it, compelling.

2.6 Expectations and diagnostic extrapolation

Greenwood et al. (2016) discuss the growing body of research in support of people's expectations of financial markets and that the results are extrapolative. Their work reviews extrapolated investors' beliefs linked to recent and past credit market outcomes. Greenwood et al. (2016) explain how fundamentals can cause markets to swing from boom to bust, with fundamentals shifting temporarily during periods of strong credit-market optimism and extrapolated beliefs based on low defaults. This mechanism can create scenarios that increase market fragility, where there is a mass increase in leverage during times when economic fundamentals are deteriorating.

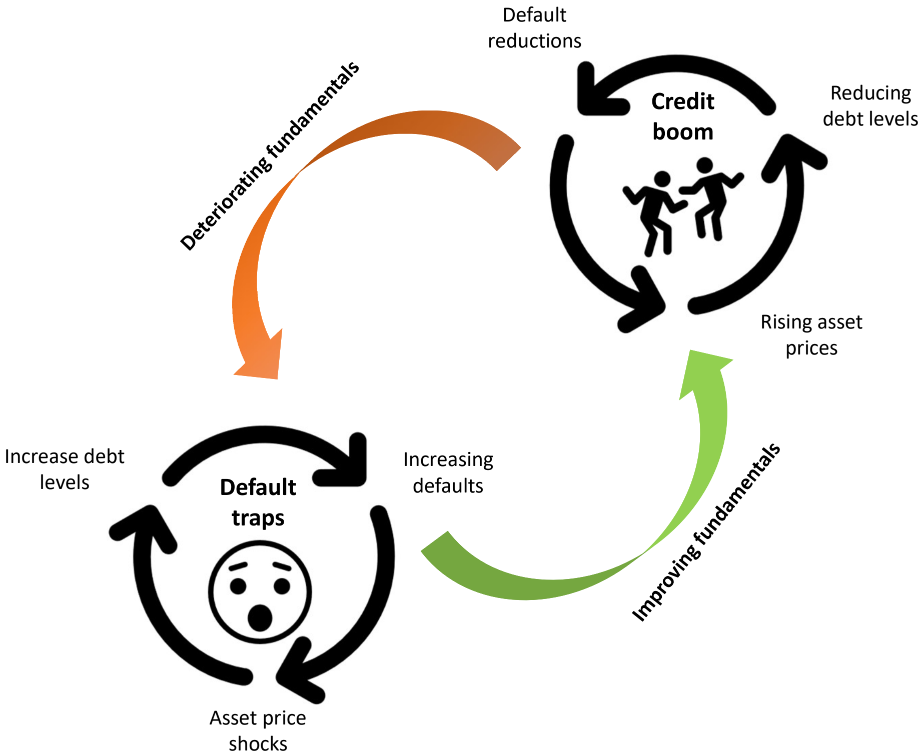


Figure 5: This illustration is adapted from Greenwood et al. (2016), page 6. The study describes the evolution of credit markets and the role that feedback plays in a theoretical framework. In the case of a low default environment, access to finance and lending is easier, bond yields decrease, the sentiment and forward expectation beliefs of the market sustaining this virtuous cycle. Higher default environments make refinancing more harder and this can lead to stressed market conditions and fire-sales of asset in illiquid markets.

Significant rises in outstanding levels of credit is a driver in the run-up to the GFC and López-Salido et al. (2017) study this with the use of a group of models and empirical studies focused on balance sheet measurements, credit growth, and financial leverage. Credit spread forecasts based on lagging valuation indicators are less informative about future economic activity than actual credit spread movements, as explained by López-Salido et al. (2017). Diagnostic extrapolation is inherently forward looking and therefore, Bordalo et al. (2018) argue, it is not subject to the Lucas critique. Diagnostic extrapolations is applied to a macro-economic model of investment, where Bordalo et al. (2018) indicated that the model could explain the empirical aspects of credit cycles without resorting to financial frictions. Diagnostic extrapolation and financial frictions can lead to models that can account for both optimistic forecasts and sharp downturns in the economy, as stated by Bordalo et al. (2018). López-Salido et al. (2017) uses a set of proxies for expected return on credit assets rather than the frictions approach, which focuses on balance sheet measures. They forecast credit spread deltas (the change in credit spread over a time period) by using lagged credit spread values. López-Salido et al. (2017) hypothesise that a lagged value of credit market sentiment (at $t-2$) at a peak is a signal for a negative change in credit spreads (widening), when combined with economic recessionary measures at t. López-Salido et al. (2017) on the other hand, maintain that elevated sentiment changes the debt composition of external finance, equity issuance increases, and net debt issuance declines. López-Salido et al. create credit indicators showing that the pricing of credit-risk becomes increasingly aggressive as it is approaching a cyclical high. After this point, the credit spread widens, and the economy undergoes a severe deceleration that we call a recession. López-Salido et al. also found that an increase in sentiment proxies occurs around the time a shift occurs in the debt composition. This is the stage in which additional funding could be secured through stock sales. As a result loans supply decreases, especially for businesses with poor credit ratings, or limited avenues for additional funding.

Greenwood et al. (2016) observed that present perceptions on loan defaults greatly influenced future views. This extrapolation is a forecast, in the case of good news, the expectation is that the ‘good times’ would continue. This is captured by Bordalo et al. (2018) in their model where news that falls short of peoples expectations serves as the mechanism for contraction. Bordalo et al. (2018) draw the conclusion that credit spreads are often very unstable, with predictable reversals and disproportionate reactions to news. Greenwood et al. (2016) evidence this by showing that current credit spreads are low enough, investors routinely underestimate future credit spreads, as we observe from surveys. López-Salido et al. (2017) contend that belief based models whilst reviewing segmentation in sentiment between equity and credit linked assets, show that there is a belief based expectation issue in both expected cash-flows and lower tail probability. López-Salido et al. (2017) note that the predictive power of sentiment for credit markets is more powerful when modelled alongside debt levels, where high debt levels provide a further signal. They conclude that credit market sentiment has good forecasting potential for credit supply approximately two years in advance, which in a stress scenario can account for a good portion of the economic deterioration.

According to Pedro Bordalo and Terry (2021), investment growth, GDP declines, and credit spreads rise in developed markets during economic and financial boom-bust cycles. Too much optimism drives excessive expansion, which in turn cools when prices are overreached with credit markets tighten, real business activity drops off, and defaults increase. Therefore, the role of expectations, in particular overinflated expectations, plays an important role. Pedro Bordalo and Terry (2021) indicate there is evidence to support this in expectations data. Pedro Bordalo and Terry (2021) model diagnostic extrapolations into a DSGE framework, using a model anchored in overreactions of people to news into a mainstream business cycle model. They explain some of the insights from the model results, namely that a negative shock to total factor productivity (TFP) is far more damaging when it occurs just after good times, whereas positive shocks have left firms, in aggregate, being further leveraged as a consequence of overoptimism. This leaves firms in a riskier position going into a crisis, such as the GFC. The economy is then said to be fragile and not able to withstand shocks. According to Pedro Bordalo and Terry (2021), economies are more sensitive to shocks during good times than during bad times. They showed that overreaction to news without the need for exogenous shocks can generate data resembling realistic credit cycles. Pedro Bordalo and Terry (2021) remind us that this result is drawn from a standard economic model of RBC with one parameter (overreaction) that is measured using company manager expectations about their profitability, based on firm level data. Pedro Bordalo and Terry (2021) conclude that micro-level overreactions to news by managers, when exaggerated, create conditions for credit cycles and fragility in boom times, therefore setting the stage for sharp reversals in bust times. Bordalo et al. (2018) introduce the human biases that can be explained as part of a behavioural theory, such as 'conjunction and disjunction fallacies' and 'base rate neglect.' An example is when one's future beliefs are overweighted in light of the information they currently have. These explanations have significant implications. Based on the same fundamentals, the path of improving news can lead to optimism, and the path of deteriorating news sentiment can in turn lead to pessimism. Bordalo et al. (2018)'s models are classified as behavioural models because they use expectation as an endogenous factor and then include additional information as factors.

Bordalo et al. (2018) formalise the model for representativeness that deals with judgmental biases and sets up the model for diagnostic extrapolation. If the true joint distribution is $h(T = t|S)$, an individual may have judged the representativeness of trait T for group S, so the representative likelihood ratio is calculated as follows:

$$\frac{h(T = t|S)}{h(T = t|\mathbb{S})} \tag{13}$$

where $\mathbb{S}$ is the comparison group. In this case, the ratio would represent the example quoted in Bordalo et al. (2018) who used Irish people with red hair as an example to review. The true distribution of people with red hair in Ireland is 10%, whereas the world average is 1%. The representative likelihood ratio is therefore 10 ($\frac{10\%}{1\%}$), and this model therefore predicts the exaggerated frequency of Irish hair colour. The model has a variable to capture the limits of working memory, θ. The likelihood variable then captures the relative focus that people put on the recent past, where larger values cause extrapolation. When taking into account the representative likelihood ratio, θ represents the agent's memory; if $\theta > 0$, the agent will not fully correct in the representative sample, and will, by implication, overweight the most recent information.

$$E_t^{\theta}(w_{t+1}) = E_t(w_{t+1}) + \theta \cdot [E_t(w_{t+1}) - E_{t-1}(w_{t+1})] \tag{14}$$

The diagnostic extrapolation is represented linearly to that of rational expectations at time t and $t-1$, where the extent of the oversampling for the future state is added to the current state in this function. $E_t(\cdot)$ represents the rational expectation, w_{t+1} the value of total factor productivity (TFP) in the future, and θ the reaction amplitude controlling parameter. If $\theta > 0$, this will drive overreactions to news about TFP. The base model uses the same construct, and has the $\theta = 0$, which implies this model is strictly the rational expectations model. Pedro Bordalo and Terry (2021) report that their rational expectations model matches macro-economic data outcomes well. The overreaction model created scenarios where the volatility of beliefs disproportionately affected the supply of credit. In the model credit demand fluctuates, with decreases furthered by narrowing spreads and increased bankruptcy costs. This implies that overoptimistic investors supply capital at a reduced spread. The rational expectation model generates results with pro-cyclical spreads, which is due to shifts in demand being stable. Interestingly, Pedro Bordalo and Terry (2021) review the size of the TFP shock that is required to create a credit spread increase such as was observed in the GFC, the introduction of a TFP decrease of 1.5% generates a scenario that matches the credit spreads observed in the GFC. Other economic variables include a drop in investment, earnings, and credit demand, these are all consistent with the GFC data. If I review the rational expectations model, it cannot generate the same increase in credit spreads.

In another approach, Beaudry and Portier (2014) make use of information from the news, which is then processed as a signal. News can thus be simply modelled as an ARIMA process that ε_t are innovations of the process, following Beaudry and Portier (2014):

$$S_t = \varepsilon_{t+q} + \nu_t \tag{15}$$

Signal, S_t at time t, using a linear model with ν_t is the variance or noise in the signal. Falato and Xiao (2020) observe that, despite the growing volume of literature, there is still very little written about the links between actual economic outcomes in real markets and credit markets. There is nothing specifically written about the underlying transmission mechanism. Falato and Xiao (2020) argue that if investors see a signal of profit deteriorating, this can be the cause of pessimism in credit-risk of corporates. In fact, Falato and Xiao (2020) establish that a key feature of credit cycles is a contribution from both micro-economic and macro-economic information in the expectations of corporate profits. The interplay between financial and information frictions is a new amplification channel through which changes in expectations are developed, not simply or directly a change in key fundamentals. The Falato and Xiao (2020) paper follows the field of research linking changes in expectations as the source of business cycle fluctuations and specifically news shocks in credit markets are captured by public sentiment indices. These shocks have forecasting power for determining credit spreads and real activity. The information affecting investors is based on public data and relates to incomplete markets with asymmetric information. The signals are noisy, but as investors' beliefs about future prices change, they incorporate more information in a dynamic fashion, which Falato and Xiao (2020) explain this is not a departure from rational expectations model. Falato and Xiao (2020) use data from the professional forecasters corporate earnings survey that provides a quarterly update on earnings expectations. After a revision of a company's corporate profits, they measure the change in investor expectations for the next quarter corporate profits, Rev_t:

$$\text{Rev}_t = E_t[\Phi_{t+1}] - E_{t-1}[\Phi_{t+1}] \tag{16}$$

The second variable of choice is the volatility measure σ_t, which is incorporated into the credit premium model, which is calibrated using a multi-variate regression to forecast changes in investor expectations:

$$R_{t\to t+k} = \alpha + \beta X_t + \sigma C_t + u_{t+k} \tag{17}$$

$R_{t\to t+k}$ is the the excess return per quarter, k. C is the vector of macro-fundament control variables, including aggregate consumption, business investment, GDP, return on assets, excess stock returns, treasuries, and fed fund rate). The explanatory variable, X_t, was tested for both the Rev and σ, the dispersion of Rev. β is the factor loading of X. The simulations reveal that a 0.14 percentage point decrease in bond returns can be transmitted for every one standard deviation change in market expectations, this reveals the strong predictive power of expectation in forecasting bond returns.

3 The origins of utility theory

Utility is a key input in portfolio and asset-pricing theory, where mean-variance optimisation (MVO) is just one example of a useful model that incorporates the crucial concept of utility into its construction. Probability theory also plays an essential role in this framework. In the following section I provide a historical perspective on the development of key models in portfolio and asset-pricing theory.

3.1 Historical background

Christiaan Huygens, Blaise Pascal, and Pierre de Fermat laid the groundwork for probability theory in the seventeenth century, and it was widely employed for decision making in uncertain scenarios and is still in use today. For a finite random variable A, its expected value can be expressed as $E(A) = \sum x_i p_i$, where the probabilities are p and values x. In 1738, Daniel Bernoulli questioned the idea that rational choices could be made based on nothing but projected return. He said that expected value might not be the most useful guiding concept for decisions because the value of a financial gain or loss might vary from person to person depending on factors like socioeconomic status. His main point was that projected return was not the only thing that could be used to make logical judgements. He focused on his argument that something's worth to a person is based on how much it helps that person. Bernoulli's studies laid the groundwork for what is now called Expected Utility Theory (EUT). Based on this theory, one can think of a utility function that ties rising income to at least a corresponding rise in satisfaction. To make a selection, one must first ascertain the highest value that may be reasonably anticipated from the utility (Hens and Rieger, 2010).

3.2 Risk aversion coefficient

A useful and related notion is the Arrow-Pratt risk-aversion measure after the work done by Arrow (1965) and Pratt (1964). It is a standard measure of risk aversion is widely used in portfolio management. This measure is flexible yet stable even after an affine transformation, where the second derivative is divided by the first derivative of the agents utility function $U(w)$, for absolute levels of wealth $\alpha(w)$:

$$\alpha(w) = -\frac{U''(w)}{U'(w)} \tag{18}$$

The equivalent relative measure, ρ, is a follows:

$$\rho(w) = -\frac{wU''(w)}{U'(w)} \tag{19}$$

A risk-averse investor is one whose utility function is concave, meaning it rises more slowly as wealth rises (Back, 2010). When investing, a risk-averse person will go for the safer, more certain option (Back, 2010). A risk-averse investor's Arrow-Pratt score decreases as wealth increases. That is why it is crucial for a risk-averse investor's utility function to have a positive first derivative, $U' > 0$. The concavity of the investor's utility curve is important and is governed by its second derivative in utility case this is most often expected to be negative, $U'' \leqslant 0$. Linear utility functions, quadratic utility functions, power utility functions, and logarithmic utility functions are some of the most used. The utility function exists as long as von Neumann-Morgenstern's axioms are satisfied. If we consider expected utility theory, then a rational person is one who makes decisions that are complete, transitive, independent of irrelevant alternatives, and continuous. For more technical review please refer to the work of Varian (1999). Back (2010) explains that the aggregate risk tolerance for an economy with the hypothetical aggregation of H investors' risk aversion coefficient, which is the reciprocal of the coefficient of risk tolerance.

$$\alpha = \frac{1}{\sum_{h=1}^{H} 1/\alpha_h} \tag{20}$$

The harmonic mean is essentially the mean function in this formula. This aggregate function does not specifically deal with the impact of correlation. This assumption could be a challenge, when we consider these models apply to all investors with time-varying risk tolerance.

3.3 Stochastic deflater function

This section is seemingly simple, however there are a few short equations that require much thought, as their meanings are deep. These simple equations are central to many finance and economic theories, most especially asset-pricing models. It is natural to think of asset prices represented the by the value of a stream of discounted expected payments or asset payoffs (Cochrane, 2005). He explains the discount function is the investors marginal utility and so the asset prices change in a recession or boom time as they are explained by marginal utility and its covariance with consumption. This means that an asset of little value may be of higher value in a boom state (where consumption is higher and the investor feel wealthy) by comparison to that asset in a recessionary state where the investor feels poor (Cochrane, 2005).

The stochastic discount factor (SDF) is a stochastic process that is used when discounting a set of expected future cash flows to determine its current value. The SDF is also known in literature as a state price deflator, state price kernel, state price density and marginal rate of substitution. Whereas a pricing kernel is a commonly used mathematical term, a stochastic discount factor (SDF) broadens asset pricing concepts to adjust for risk, sometimes referred to as 'risk adjusted'. The SDF is also central to the 'law of one price', which is a fundamental law of pricing. For the law of one price to hold, we require that one SDF is applied to all cashflows, pricing all assets simultaneously (I explain the law of one price more in Section 4.3). By using SDFs, investors are able to evaluate a choice of investments and make their choices based on their risk-adjusted return profiles for the proposed investments.

Back (2017) and Martellini and Milhau (2017) explain that the SDF determines risk premium; this is also expressed in a way that relates an investor's utility of inter-temporal consumption to measurable variables, which is a real challenge. Martellini and Milhau (2017) adds that linkages between marginal utility of consumption and the SDF are at the heart of consumption-based models. So this link is key, it explains the value of a risk premium and the rewards required for holding assets using an investor's wealth (Martellini and Milhau, 2017). We know that a reward is passed to an investor for bearing risk and the reward needs to be compelling enough for the investment to take place. In factor models, this is represented with a beta weighted combination of factor risk premia. I formalise this process by considering the intertemporal portfolio choice problem for an investor who can trade in asset i. This is a maximisation problem that that incorporates a time separable utility function, where β_j is a discount factor over time t, following Campbell et al. (1997):

$$\text{Max } E_t \left[\sum_{j=0}^{\infty} \beta_j U(C_{t+j}) \right] \tag{21}$$

Where C_{i+j} is the investor's consumption at time $i + j$ and $U(\cdot)$ is the utility of consumption over time $i + j$. Cochrane (2005) explains this is the basic pricing equation which is an investors first-order conditions that help us understand the value of an uncertain array of income (without the uncertainty it would be a simple case of discounting of cash-flows). What we need now is a way of capturing a typical investor's valuation of the cash flows with a mathematical model, that is achieved by way of a utility function, as Cochrane (2005) explains and I will detail. Campbell et al. (1997) explain the utility function for the optimal consumption plan where $U'(C_t)$ is the marginal utility of consuming one unit at time t, as follows:

$$U'(C_t) = \beta E_t \left[(1 + R_{i,t+1}) U'(C_{t+1}) \right] \tag{22}$$

Asset i generates returns of a rate of $(1 + R_{i,t+1})$. This equation explains the expected marginal utility that an individual expects from asset returns and the consumption of proceeds at time $t+1$. Back (2010) explains that this is the first-order condition for portfolio choice in a dynamic setting, also called the 'Euler equation'. This equation captures both the investors risk appetite or aversion to risk and their desire to consume now, or impatience. As Cochrane (2005) explains that the impact is captured in the β factor, it is known as the subjective discount factor. The curvature of the utility function introduces the aversion of risk and the link to intertemporal substitution[5]. You'll notice at this point, that asset pricing it is not simply a discussion around the relative ratio of asset returns to volatility.

So, how do we capture the decisions of the investor such as what to buy and when to buy it? Would this be done by using a mathematical model? To help me, I follow Cochrane (2005) to develop a model that describes the investor choice and solving it to form the first-order condition, this function is known as the stochastic deflater. Let us assume e represents

[5] For me this term requires a helpful reminder of the intuition: intertemporal substitution refers to the saving and spending decisions that individuals make, as well as the effects of those decisions on the future.

the opening consumption level and ξ denotes the amount the investor chooses to buy. The investor choice problem can stated as, following Cochrane (2005).

$$\begin{aligned} \max_{\xi} U(C_i) + \ & \mathrm{E}_t[\beta U(C_{t+1})] \quad s.t. \\ C_t &= e_t - p_t\xi \\ C_{t+1} &= e_{t+1} + R_{t+1}\xi \end{aligned} \tag{23}$$

Price p_t at time t and the constraints are then substituted into the objective function and solving the equation obtained by calculating the derivative with respect to ξ and setting the result equal to zero, we get the optimum consumption and portfolio choice function.

$$\begin{aligned} p_t U'(C_t) &= \mathrm{E_t}\left[\beta U'(C_{t+1})R_{t+1}\right] \\ p_t &= \mathrm{E_t}\left[\beta \frac{U'(C_{t+1})}{U'(C_t)} R_{t+1}\right] \end{aligned} \tag{24}$$

This is a consumption pricing equation and Cochrane (2005) explains that we can now compartmentalise into two parts that are more intuitive. The first part is $p = \mathrm{E}(MR)$. The second part is $M = \left[\beta \frac{U'(C_{t+1})}{U'(C_t)} R_{t+1}\right]$ and then M_{t+1} is the is the *stochastic deflator function (SDF)*. To complete this equation, we equate the marginal cost and marginal benefit, and if we divide both sides of equation(22) by $U'(C_t)$ the resultant equation is:

$$1 = E_t\left[(1 + R_{i,t+1})M_{t+1}\right] \tag{25}$$

$M_{t+1} = \beta U'(C_{t+1})/U'(C_t)$ and we know that M_{t+1} is the pricing kernel and is the SDF. This is comparable to the discounted marginal utility ratio, often known as the intertemporal marginal substitution rate. This can be restated to suit an asset-pricing problem, Campbell et al. (1997) explains this is based on the one period unconditional form, $1 = E[(1+R_{i,t})M_t]$.

$$E[1 + R_{it}] = \frac{1}{E[M_t]}\left(1 - \mathrm{Cov}[R_{i,t}, M_t]\right) \tag{26}$$

can then be addressed by assuming a zero beta asset for SDF and this implies that the assets expected gross return ($E[1 + R_{0t}] = 1/E[M_t]$) can be substituted into equation(26). This is the excess return of asset i over a zero-beta return,

$$E[Z_{it}] \equiv E[R_{i,t} - R_{0,t}] = -E[1 + R_{0,t}]\mathrm{Cov}[R_{i,t}, M_t] \tag{27}$$

which shows that higher returns imply a smaller covariance with the pricing kernel. The intuition is that an investor requires a higher risk premium if the asset does not have enough return when it is most needed by the investor and this drives the requirement for higher risk premiums (Campbell et al., 1997). Back (2017) explains that the SDF and risk premia are in fact consistent, where asset-pricing theory is based on an SDF that determines a specific form of for M_t, then risk premia are essentially a set of covariances between variables.

To consider this for a moment, as Cochrane (2005) explains, a single SDF to encapsulates all risk corrections. The key is that the utility function can change, and from formula(24) would still enable one to connect the SDF to the data. This is powerful and it prevents the need to redesign the modelling framework when researching new approaches to model asset-pricing anomalies (Cochrane, 2005). When I review the more advanced asset-pricing models, in Section 5, we can see how the framework has has been utilised to express many different asset-pricing models of an uncertain space that is market asset pricing.

4 Basic asset pricing, factor models

The detail in prior sections are a good foundation to now explore asset-pricing models and advanced asset-pricing models, starting with factor premium models. As a reminder, factors are used to help explain returns within asset classes and are often expressed as an asset class return relative to a risk free asset such as cash.

4.1 Factor and risk premium recognition

Cochrane (2017) highlights that well established factors do exist such as momentum, value and earnings premium. According to Roncalli (2017), the goal of factor investing is to deepen our understanding of asset classes in what they offer in terms of systematic risk factors and diversification. As an introduction, investment returns can be explained by the standard price factors from asset-pricing theory. When factor returns are higher than or equal to predicted returns based on the systematic risk borne by the asset, then it is said the factor is ‘priced’, and that the factor earns a risk premium (asset returns in excess of the risk-free rate) over longer time horizons. So, this is fundamentally important in asset management.

Cochrane (2017) reports that many factor models exist (in one count there are more than three hundred), many of which are spurious or simply correlated with other factors. Munk (2013) explains that a lot of work is required in identifying factors, it is not only about parsing lots of data using machine-learning techniques and powerful hardware to extract patterns that explain a ‘data-only’ factor. This approach of simply using data can be inappropriate and

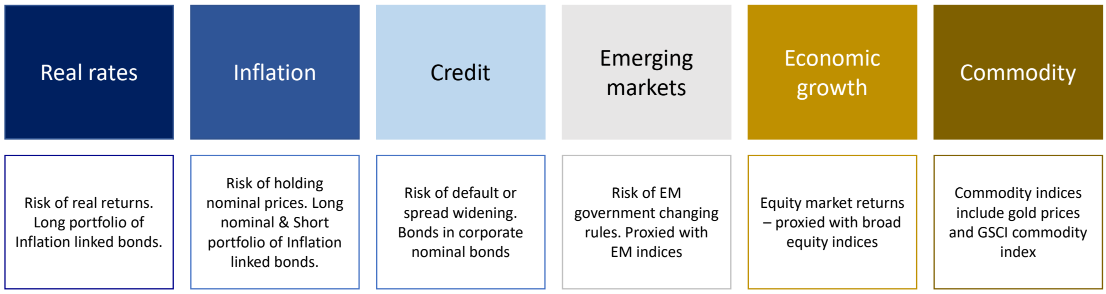


Figure 6: This is a representation of Bass et al. (2017), page 3. It is a nice way to capture the core market risk factors and definitions.

leads to data snooping. The best known data-only model is that of Fama (1970). Munk (2013) explains that the real challenge is to make use of fundamental reasoning in factor design that is supported by formal economic arguments or a fundamentally sound asset-pricing model (this is arguably the overarching goal in this study and my belief is this is the correct way to define asset-pricing for asset classes that are not data rich). Without adding to the increased set of factors, there are a few factors that already explain the largest proportion of risk; this includes rates (real and nominal), inflation, economic growth (equity exposure). Commodity, emerging market, and credit factors also play a significant role. Foreign exchange (FX) however is not compensated with a risk premium over the long-run, but it should still be taken into account when building a portfolio because of its profound impact on portfolio volatility (Bass et al., 2017).

Unlike the traditional long-only fundamental factors, such as size, value, and momentum identified in Fama (1970), alternative risk premia can be exposed to not only bonds and equities but also credit, rates, commodities, and currency. Alternative factors and risk premia provide a schedule of returns outside of what is generally offered in equities and bonds, which is a desirable characteristic in asset management. Roncalli (2017) highlights that in low yielding rate markets, alternative risk premia will provide an interesting option for generating returns. Sheikh (2018) explains that factor models are often used in industry as multi-variate models used to generate random paths for asset-pricing. He explains that commercial software such as RiskMetrics and MSCI Barra use asset-pricing models such as arbitrage pricing theory (APT) and the capital asset-pricing model (CAPM) with large variance-covariance matrices. Due to the demand of this commercial software, we know there is a high penetration of these models in the market.

There are two distinct classes of factor models, APT and the CAPM (Martellini and

Milhau, 2017). The APT is based on the seminal research of Professor Ross (1976) and incorporates multi-factors in the model. CAPM is a well-known model and it relies on a single market component to determine the investments value.

4.2 Single-factor models

There are a few versions of the CAPM, but really the foundation for these models was laid by Markowitz. Two notable versions include Black (1972) and those by Sharpe (1964) and Lintner (1961). Black's version is more general and is based off a zero-beta portfolio in the absence of a risk-free asset and a market portfolio, whereas Lintner and Sharpe's version is based off a risk free rate R_f. There is no specific measure of market portfolio, so core equity indices such as S&P500 or FTSE100 are used to proxy the market portfolio m. Here, I follow Campbell et al. (1997) who defines the more general Black version:

$$\mathrm{E}[R_i] = R_f + \beta_{i,m}(\mathrm{E}[R_m] - R_f) \tag{28}$$

Where $\beta_{i,m}$ is equal to

$$\beta_{i,m} = \frac{\mathrm{Cov}[R_i, R_m]}{\mathrm{Var}[R_m]}. \tag{29}$$

4.3 Multi-factor models

The 'law of one price', Fabozzi et al. (2007) explain, is a foundational axiom in finance. The value of a financial asset must have one price in the market, regardless of how the asset is implemented (security type, contract type), and then this asset can be synthetically created for trading in the market. Ross (1976) introduced the law of one price, it is a key assumption of factor models, therefore foundational for the APT model. Fabozzi et al. (2007) explain that equilibrium prices are maintained because of market liquidity and traders actively closing down the pricing differential on single stocks, including arbitrage opportunities. This is one major factor in why only systematic risk will be rewarded and not unsystematic risk (that is related to the underlying covariance with common factors). In the following section, we follow Back (2017) in his detailing of risk premia and asset-price modelling. Factor pricing models are configured as linear versions of consumption-based marginal utility models. β is a sensitivity parameter that is calibrated, often by regression. In the case of risk factor modelling, this is set up as:

$$\mathrm{E}[R_{t+1}] = \alpha + \beta' f_{t+1} \tag{30}$$

For factor modelling the dependent variable is excess returns, for some constant α and f factors. Back (2017) explains that there is an equivalence between factor models and SDFs. An important assumption in asset-pricing and macro-economics alike is how the measures of aggregate marginal utility are determined and how the measure accounts for bad states of the world. Back (2017) warns that if this the measure is not appropriately defined, the average return premium in the model will not be high enough to warrant an individual's investment as the perceived risks would be too high:

$$\gamma \frac{U'(c_{t+1})}{U'(c_t)} \approx \alpha + \beta' f_{t+1} \tag{31}$$

The left side of the approximation is the same as the SDF that we developed in equation 26 and is the marginal rate of substitution. This is a theme that you will see in this study, I explore the relative risks in many sections, but in the ESG modelling, part V, I model this for private credit. If the APT is set out in a linear factor model with factor f, and with a risk free rate R_z, then following Back (2017), the model is denoted as follows:

$$\mathrm{E}[R] - R_z = \gamma \frac{\mathrm{Cov}(f, R_m)}{\mathrm{Var}(f)} \tag{32}$$

The γ is called the factor risk premium, also known as the price of risk, defined as the extra return premium required per unit beta. In the CAPM, the factor model is set up using a linear model and a risk-free rate of return (per the Sharpe and Lintner configuration) and it is termed the market risk premium. The result is the price of risk is the factor mean. For more detailed readings on the APT model, please refer to Ross (1976) and Roll and Ross (1984). For an asset the return R, can be broken down into constituents and expressed in a factor model as follows:

$$E[R] = \alpha + \beta f + \varepsilon \tag{33}$$

$E[R]$ is the expected return of the security or asset and α is the asset return constant. The beta loading of an asset to movements in the systematic factor is denoted as β. The systematic returns are denoted by f and the final term, ε, the 'noise term' (Roll and Ross, 1980). This is the residual of the model's ability to predict the asset value, and this is attributed to idiosyncratic factors. The error term represents idiosyncratic risk, and as diversification increases so the error term is expected to decrease. Roll and Ross (1984) explain that idiosyncratic risks can be reduced to zero through diversification when dealing with a large enough portfolio. In their opinion, three to four indicators are sufficient to characterise the risk. The vector form of the APT model can be written as:

$$E[R] = R_f + \sum_{k=1}^{K} \beta_k (E(F_k) - R_f) + \varepsilon \tag{34}$$

$E(F_k) - R_f$ is the excess return of the k^{th} systematic factor over the risk free rate, R_f. This relationship is referred to as the APT, unlike the simplified CAPM, which is based on only one factor, APT takes into account a wider variety of variables. APT can be extended to include multiple factors. The model designed by Fama and French is arguably the most famous of the multi-factor market models. A great example of a commercially successful factor model is MSCI's Barra active risk model that Alexander (2008a) defines as

$$E[R] = \alpha + \beta X + \beta^F \mathbf{R}^F + \beta^I \mathbf{R}^I + \varepsilon. \tag{35}$$

α and β are relative portfolio alpha and portfolio beta values respectively (these are found by off-setting portfolio returns with the defined market benchmark), X denotes the market returns. $E[R]$ is the expected active portfolio return, $\mathbf{R}^f$ and $\mathbf{R}^I$ are returns on risk indices and industrial indices respectively, which are represented in bold to denote a vector. α is the benchmark specific return and ε is the portfolio specific return, not explained by market and risk factors and industry factors. There are however four classes of multi-factor models (Fabozzi and Pachamanova, 2016):

- **Statistical approach** making use of the available historical data, by applying methods such as principal component analysis (PCA), panel regression, and econometric analyses such as autoregression, vector-error correction, and cointegration. In general, data methods are used in the search for factors, because the data typically contains a large number of stocks for a short time period, data reduction techniques were required.
- **Macro-economic approach** models are based on macro-economic factors and data that can fundamentally explain stock returns. Typical factors can include interest rates (real or nominal), debt levels, gross domestic product, inflation indices, sentiment indices, and money supply.
- **Fundamental approach.** These models are based on fundamentally recognised factors, starting with the work of Fama and French, whose model includes factors such as small-minus-big and high-minus-low. Typical factors include credit ratings, price-to-earnings ratios, momentum, and equity dividends.
- **Hybrid models approach** Fabozzi and Pachamanova (2016) explain that hybrid models use traditional sets of factors to explain as much of the variation of returns as possible,

then make use of statistical methods to further explain what remains in the residual. According to Fabozzi and Pachamanova (2016), hybrid models provide the best of both worlds, where financial theory and economic context impose a structure and then relatively sophisticated techniques are used to mine the data for what the theoretical model does not predict.

4.4 Factor and risk premium modelling

Factor investing is a specific approach that targets the true source of excess return by taking a comprehensive review of what factors are driving that return. In this section I provide an overview of these core factors.

4.4.1 Rate

Core to a factor framework is the rate of return from risk-free assets, which is the primary building block for establishing other assets' rate of return above risk-free assets. Understanding risk-free rates entails understanding elements of short-end rate securities models, which are workhorses in finance and have been extensively researched. Refer to Munk (2011) for a detailed modelling review. The short rate, or level factor, incorporates macro-economic data to describe monetary authorities' decisions, one of which is current rate and expectations of rates. Rates are often described by the driving factors using the Taylor rule, and I follow the notation of Ang (2014),

$$SR_t = r^* + \pi_t + 0.5(\pi_t - \pi^*) + 0.5I_t \tag{36}$$

Here SR_t is the short rate, r^* is the long-run real interest rate, π_t is the current inflation rate, π^* is the target inflation rate, and I is the output gap, which is the difference between the potential and real GDP growth rate. The first term, which is made up of real rates and inflation, is also described by the well-known Fisher hypothesis. What is critical in this model is the rate term. Often termed the 'level' of rate, it is material to the degree that it accounts for the majority of the total variance of fixed-income factors (Alexander, 2008b). It is in this category that one should expect to see models such as the Vasicek, Cox-Ingersoll-Ross (CIR), Hull-White-Vasicek (HWV), Ho-Lee, and Black-Karansinki as prevalent. The multi-factor models that incorporate further information in describing bond prices are more accurate at explaining bond prices, an example of this is the multi-factor model build by Ang and Piazzesi (2003).

4.4.2 Term

The next fixed-income factor has to do with the premium earned from bearing the risk that interest rates for bonds may change over the life of the bond. The difference between rates when comparing a bond or treasury at different maturities is known as the 'term premium.' For the full range in tenor, the term premium increases with the term to maturity. What explains this increase in premia is a question of the shape of the term structure of interest rates. The theories that have traditionally been used to explain the 'term premium' include:

- **Expectations theory** explains the term structure of interest rates as this reflects the markets expectations of future spot rates.
- **Liquidity preference theory** is a macro-economic theory that was developed by Keynes in his book 'The General Theory of Employment, Interest and Money' written in 1936 to explain interest by supply and demand for money. The theory explains that investors are risk averse and prefer shorter maturities, so longer term securities will have to offer additional returns to compel investors for their investment money.
- **Preferred habitat theory** is a variation of the liquidity preference theory. This theory states that aside from the liquidity preference investors have a preference of maturity, and for investors to purchase bonds at a different maturity they require a premium.
- **Market segmentation** is slightly different where long and short interest rates are not assumed to be related. Yield curves are simply a result of differing demand levels between markets with different segments along the term structure.

In the next section I go into a bit more detail around how the underlying fixed-income factors are dealt with.

4.4.3 Slope, level and curve

Modelling fixed-income factors is closely tied to well-known term structure models, where an affine approach is often used by industry practitioners and academia. The affine modelling framework is flexible, tractable and often more robust (Munk, 2011). An important part of the modelling is to decompose the term premium, this often done, quite accurately, using a principal component analysis (PCA). Three underlying factors that explain the term premium are well-known in the industry. The factors are 'slope', 'level', and 'curve' and refer to a yield curve's influence, not the economy's shocks. Fixed-income models can explain asset-pricing movements without arbitrage, unlike general equilibrium asset-pricing models. Ang and Piazzesi (2003) create a model that adds macro factors to term structure model latent factors ('slope', 'level', and 'curve'). Inflation is linked to slope and level factors.

Ang (2014) explains that long-dated bonds are more sensitive to unexpected changes in macro-economics. The term premium is the driving factor and changes in the interest rates of risk-free bonds and can be harvested by investing in government bonds. Thus, longer-term bonds are sensitive to risk due to certain factors, which therefore expect adequate risk premiums so as to reward the customer for bearing risk. Alexander (2008b) confirms in a PCA analysis that two other factors, the 'twist' factor and the 'butterfly' factor, explain up to 20% of the variance. Macro-economic information, mean reversion, and more complex volatility models are specified in multi-factor short rate models.

Even with additional considerations, term spreads' predictive potential remains strong, according to Engstrom and Sharpe (2018). The one- to two-year spread indicates the market's monetary policy expectations. When negative, the market expects a recession, and actual data suggests that it is usually right (Engstrom and Sharpe, 2018). Interestingly, many market participants use the near-term spread to predict recessions. Economic factors affect mid-to-long-term spread more than short-term spread, according to Connolly et al. (2018). The market knows that the FED and BOE effect term spread, market expectations and inflation specifically affect the short-term spread. The short-term curve is normally upward-sloping, but the ten-year term spread is usually downward-sloping, unless rates are expected to fall soon. Yield volatility and higher convexity drive term-structure (Engstrom and Sharpe, 2018).

4.4.4 Credit default premium

Credit and duration are the two traditional factors in fixed-income (Martellini and Milhau, 2017). It has been observed that bonds subject to default are higher yielding than safe treasury bonds. This is especially true given that credit defaults are likely to occur during economically difficult or stressed markets. Dor et al. (2012) explain the credit spread premium is harvested by investing in credit risky bonds, where the investor can expect a return premium over and above the loss attributed to credit in the bond. This is an interesting feature as corporate bonds are not seen to provide impressive investment returns in recent history. In fact, the corporate bonds from 1990 to 2009 as measured by the U.S. Corporate Index returned 0.27% above duration matched treasury bonds, which is low even in today's low yield environment.

Gourio (2012) explains that there is a puzzling market feature in holding credit assets to maturity. According to Houweling and van Zundert (2017), the benefit of investing in corporate bonds is the ability to harvest both the term premium and the default premium. Dor et al. (2012) conclude that investors who are able to hold the bonds through the stressed periods without having to liquidate the portfolio will recoup the spread net of default losses and benefit from the wide credit spread (approximately 90% of the credit spread exceeds the amount required to cover credit losses). This is because investments in private debt or unlisted credit carry a large additional liquidity premium. Anson et al. (2011) define credit-risk as financial instruments issued by organisations or entities that have the potential to default, where outstanding money owing to the investor is at risk of default. There is reward for holding assets that are exposed to credit risk and it can be shown empirically by comparing credit spreads to the expected cost of credit, expressed as the potential for default multiplied by the loss owing to that default (assumed to be 50%). In a traded market setting, corporate bonds default with a 0.4 percent probability and have experienced recoveries of around 50%, resulting in a credit cost of 0.2 percent, whereas the BAA-AAA credit spread is 1 percent. The corporate bonds are therefore being priced at a discount, and the investor can expect increased excess returns, at a premium (Anson et al., 2011).

Berndt et al. (2018) highlight that the movement in risk premia for high-yield (HY) firms is less correlated with macro-fundamentals than investment grade (IG) firms. Berndt et al. (2018) conclude that the relationship between risk premia and macro-economic fundamentals, once controlled for firm and market based variables, is clear and contributes to a powerful model. Berndt et al. (2018) discuss how in stressed periods (including 2002 and 2009) will have to show marked decreases in risk-bearing capacity relative to the actual risk levels. This will likely explain the larger increases in the risk premia relative to the expected losses. Finally, Berndt et al. (2018) conclude that the risk premia to expected losses, when reviewed over the full time horizon, is actually less pronounced for HY firms than it is for IG firms.

4.4.5 Final considerations

Campbell et al. (1997) warn that using a factor approach without limits has two major flaws. First, the statistics-only models will over-fit the data because of data-snooping bias, where data is misused or manipulated to artificially generate meaningful results. This means that the models can't be used to predict returns outside of the sample or out-of-time, so they are not very useful. The second major reason, which ties neatly to large parts of this thesis, is that these models do not account for market inefficiency or investor irrationality, both of which will imply Sharpe ratios that are modelled above what can be expected when one considers the equilibrium of the underlying market. This, I suspect, is a matter of truly understanding the risk levels.

5 Advanced consumption-based asset models

In the next section I summarise the more advanced asset-pricing models. Part of this thesis is a review of the literature of puzzles that the asset-pricing pricing models look to solve and will include parts from macro-economics and finance.

5.1 Inter-temporal substitution models

The standard power utility consumption-based model, following Cochrane (2017):

$$M_{t+1} = e^{-\delta} \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} \quad \text{or} \quad \mathrm{E}(R^e_{t+1}) = \gamma \operatorname{cov}(\Delta C_{t+1}, R^e_{t+1}) \tag{37}$$

C is consumption, ΔC represents the consumption growth and γ is the coefficient of risk aversion. Essentially, as the risk of a recession increases, so consumption falls. Excess returns are denoted as R^e. This is a quantifiable measure-based approach, and in the work from Mehra and Prescott (1985) and Hansen and Jagannathan (1991) shows from the analysis that a risk coefficient would need to be very large to induce an equity risk premium that is as high one can capture in the market:

$$\frac{\mathrm{E}(R^e)}{\sigma(R^e)} \leqslant \gamma \sigma(\Delta C_{t+1}) \tag{38}$$

If a representative agent can accept the high risk aversion, the model would require that risk-free rate levels are inconsistent, explains Cochrane (2017). Lastly, the risk premium is known to display time variation, with a clear relationship to the business cycle, and this can be calibrated by regression of the following form $R^e_{t+1} = a + by_t + \varepsilon_{t+1}$. Regression variables may include prices, dividends, price-dividend ratios, and bond spreads. The predictive power of these models is not a useful approach with low R^2 values. Macro-finance models are non-vacuous and when processed in a formal testing framework would be rejected. Cochrane (2017) points out that at the top (bottom) of a cycle, expected returns are low (high), risk premiums are low (high) and prices are high (low). Another feature is the excess volatility in equity, due in part to the dividend expectation for stocks that have relatively high prices so we consider the model: either the dividend is expected to increase or the prices will decrease.

Cochrane (2017) explains that the economic state we are in when equity shares fall sharply is just around the same time one may be at real risk of losing a job or their business. So what is this fear really about? How can it be measured? And how is this connected to the business cycle? These central questions point to a subtle difference: that people fear recession more than can rationally be explained, and when people are in a recession, their risk appetite is affected, implying a high risk premium required to bear future risk (Cochrane, 2017). Recessions seem to capture both the risk of loss in the here-and-now, coupled with the additional risk aversion for future prospects. These may in fact be interrelated; the fear of loss may cause people to sell stocks, and the well-known effect of inelastic supply could in fact drive deeper recessions.

5.1.1 Power utility

The power utility model is often seen in academic articles, it makes uses of a time-separable power utility function for a representative function as follows (Campbell et al., 1997):

$$U(C) = \frac{C^{1-\gamma} - 1}{1 - \gamma} \tag{39}$$

The coefficient of risk aversion is denoted as γ . A key, interesting, mathematical feature here is that as γ approaches one, the function becomes $U(C_t) = \log C_t$. The features of this model, power utility, is that it is scale invariant, which implies that even individuals that are at differing wealth levels, than be aggregated to a representative investor, which is a strong justification of the use of this model in the Consumption CAPM (CCAPM) (Campbell et al., 1997). The model is a power form and this means that it implies a restrictive elasticity of intertemporal substitution assumption.

5.1.2 Recursive utility

Cochrane (2017) explains that although the many approaches adopted by different specifications, including technology, markets, preferences, which can be altogether different arguments, the models end up looking quite similar. The recursive utility model is one in which the future and current utility are connected, using a non-linear function, again following

Cochrane (2017):

$$U_t(C_t) = \left[(1-\beta)C_t^{1-\rho} + \beta\left[\mathrm{E}_t\left(U_{t+1}^{1-\gamma}(C_{t+1})\right)\right]^{\frac{1-e}{1-\gamma}}\right]^{\frac{1}{1-\rho}} \tag{40}$$

Elasticity of inter-temporal substitution is $\frac{1}{\rho}$ (that is, ρ is the relative risk aversion) and risk aversion is set to γ. Cochrane (2017) explains that with $\rho = \gamma$ this gives a time-separable power utility. The discount factor then arrives with the following formulation, following Cochrane (2017):

$$M_{t+1} = \beta\left(\frac{C_{t+1}}{C_t}\right)^{-\rho}\left\{\frac{U_{t+1}C_{t+1}}{\left[\mathrm{E}_t\left(U_{t+1}^{1-\gamma}C_{t+1}\right)\right]^{\frac{1}{1-\gamma}}}\right\}^{\rho-\gamma} \tag{41}$$

The tricky issue with the models is that U, the utility index, is not observable, and this needs to be swapped for something that is (Cochrane, 2017). For wealth portfolio outcomes, Cochrane (2017) explains that Epstein et al. (2014) used market returns as a proxy. More recent advances utilise data on consumption to generate a utility.

5.1.3 Habit model

The habit model explains more than one puzzle, such as a low short rate with low volatility, high equity risk premium and cyclical variation of volatility stock prices (something akin volatility clustering that increases during economic decline). The habit model is based on a standard power utility that introduces a benchmark or subsistence point called 'external habit'. First, I will define the standard model here, following Hansen and Jagannathan (1991):

$$\mathrm{E}\left[\sum_{t=0}^{\infty}\delta^t\frac{(C_t+\theta C_{t-1})^{(\gamma+1)}-1}{\gamma+1}\right] \tag{42}$$

C is consumption and when $\theta = 0$ this model becomes a contingent claims market where all consumers have the same preferences. When $\theta > 0$, which is the assumption in economics models, there is an intertemporal substitution from consumption goods. This means the consumer would rather forgo consuming now to consume at a later date (Hall, 1988). He explains that this is a central question in macro-economics and where customers are sensitive to changes in interest rate. In another approach in Campbell and Cochrane's habit model is that it specifies that an individual will have spending habits that are linked to the history of aggregate consumption, rather than an individuals personal history on consumption levels after one period. The model is specified in a way that changes to habit are slow moving relative to changes in current consumption. As Campbell and Cochrane (1999) explain, this replicates mean-reversion in price, dividends and stock returns for longer horizons. This model is defined as, following Campbell and Cochrane (1999):

$$\mathrm{E}\left[\sum_{t=0}^{\infty}\delta^t\frac{(C_t - X_t)^{(1-\gamma)-1}}{1-\gamma}\right] \tag{43}$$

The formula expresses the maximised utility of an individual or agent. This model does not consider how current consumption decisions will affect future risk aversion (such as in non-separable goods models). The time-varying benchmark (or subsistence level) X is determined externally to the model by observing others' consumption, so surplus consumption is $S_t \equiv (C_t - X_t)/C_t$. This ratio is configured so that when $S_t = 0$, this is a bad state of consumption that equals the habit level X_t. The curvature of the utility function then becomes

$$-\frac{CU''(C))}{U'(C)} = \gamma\left(\frac{C}{C-X}\right) = \frac{\gamma}{S} \tag{44}$$

S is the surplus consumption ratio, X subsistence level and consumption C. As I previously explained, the habit is determined by historical aggregate consumption, following Campbell and Cochrane (1999):

$$S_t^a \equiv \frac{C_t^a - X_t}{C_t^a} \tag{45}$$

The reaction of an individual to aggregate consumption can be modelled as an autoregression (AR(1), accounting for heteroskedasticity) where dependent variable $s_t^a \equiv \log S_t^a$. Campbell and Cochrane (1999) go on to specify the slow moving habit formation model where $X_t \approx \phi X_{t-1} + kC_t$. This formulation allows for a changing appreciation for risk aversion as consumption grows and likewise as consumption drops (e.g., wealth levels drop) the benchmark changes too, and thus the risk aversion rises. This model is a regression and can be specified to match local characteristics and therefore be directly linked to a market. The effect is that the returns required to account for the changes in risk aversion increase. This would result in lower dividend asset prices.

Cochrane (2017) explains that if the dividends were constant, then changes in the dividend ratio would be entirely driven by varying risk premiums. This model can then explain expected dividends by comparison to excess price volatility. The model naturally creates time-varying recession-linked risk premia and returns related to dividend yields, especially at long horizons. For habit models, risk free rates are usually a challenge. So to define a slowly fluctuating pro-cyclical interest rate function, with marginal utility $(C-X)^{-\gamma}$, following Cochrane (2017):

$$r dt = \delta dt + \gamma\left(\frac{C}{C-X}\right)\mathrm{E}\left(\frac{dC}{C}\right) - \frac{1}{2}\gamma(\gamma+1)\left(\frac{C}{C-X}\right)^2\sigma^2\left(\frac{dC}{C}\right) \tag{46}$$

The subjective discount factor δ denotes the real interest rate that is added to the inverse elasticity of inter-temporal substitution multiplied by the expected consumption growth, plus risk aversion squared times the variance of consumption growth. This habit model features a recession indicator ($S \equiv (C - X)/C)$) for the power utility model.

$$M_{t+1} = e^{-\delta} \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} \left(\frac{S_{t+1}}{S_t} \right)^{-\gamma} \tag{47}$$

The model will reflect a small effect when stock prices decrease; this is due to the impact of $\gamma = 2$ relative to the recession indicator. Cochrane (2017) summarises the model capability, stating that the model does feature an equity risk premium, a reasonable Sharpe ratio and low constant time varying risk free rate. The model is successful in generating the observed returns, volatility and even generates heteroscedastic returns. It still features a relatively high risk aversion and, as such, it does not fully solve the equity risk-premium puzzle.

5.1.4 Long-run risks

Long-run, idiosyncratic, and rare disaster models feature time-varying risk variables (Bansal and Zhou, 2002) and can explain many features of financial markets. These models, as Bansal (2007) explains, are focussed on economic uncertainty and changing long-run growth potential, and this drives equity risk premia and price volatility. The models for non-separable goods work in a way that affects a person's appetite for risk depending on the size of their consumption (i.e., the size of your house may affect your risk aversion). For models of the behavioural or risk ambiguity type, a person's probability assessments vary over time. Preferences are not inherent to finance models, but the leverage value of intermediaries can affect the market's risk capacity. The approaches listed above look different but are often discount factors and/or marginal utility generalisations.

$$M_{t+1} = \beta \left(\frac{C_{t+1}}{C_t} \right)^{\gamma} Y_{t+1} \tag{48}$$

Cochrane (2017) explains that Y_{t+1} is the more significant contributor in the model. Even the probability distortion or behavioural models follow this basic form, where the expectation is a sum over states s:

$$p_t U'(C_t) = \beta \sum_s \pi_s U'(C_{t+1,s}) x_{t+1,s} \tag{49}$$

Payoff and price are p and x respectively. β is the market beta, π is the probability assessments, that are can effectively distorted to account for the impacts of sentiment. The state variable Y is there to set distortion in probability or marginal utility (given these always enter the equation together, so it is the same thing), Cochrane (2017). In this model, time-varying risk aversion for individuals is endogenous, where an individual will be less willing to take on risk during stressed market scenarios. Bansal (2007) concludes that the strength of the model is in how it describes the features of financial markets and can be used in an economic policy setting for markets.

5.1.5 Long-run equity premium

The long-run equity premium is a surprising result (Cochrane, 2017). From research performed by Hansen and Jagannathan (1991), we know that the need for greater levels of volatility are required to explain the equity premium, which is a central theme in this class of model. It is not only consumption that drives the habit discount factor, if we define the model following Cochrane (2017):

$$M_{t,t+k} = e^{-k\delta}\left(\frac{C_{t+k}}{C_t}\right)^{-\gamma}\left(\frac{S_{t+k}}{S_t}\right)^{-\gamma} \tag{50}$$

The long term horizon is k. The S term provides the additional habit model volatility, as Cochrane (2017) explain it is a contributing factor in asset prices and expectations of return rates in the long term. The second term is stationary, whereas C is a random walk that increase with term, so asset-price variability is less of a feature over the long term horizons (Cochrane, 2017).

5.1.6 Heterogeneous preferences

Heterogeneous preferences is an interesting and dynamic setting where people relate to differing levels of risk aversion, where more risk averse people hold less stock. A short introductory note on Pareto optimality is needed here. For an economy the goal of a social planner is to maximise individuals' wealth over the planning horizon, in a way that the expected utility for one investor will not reduce or sacrifice the expected utility of another investor, if this is achieved, then the property is termed Pareto optimality (Back, 2010). Munk (2013) explains that Pareto optimality allows for aggregate consumption risk in a that is shared amongst the participants in the economy, thus allocation is seen to be is solved when the allocation of resources in the most efficient way possible. This being the case, larger stock brokers will lose a greater proportion of their money during a shock resulting in a smaller share of the market, thus the overall market risk aversion goes up. In the heterogeneous preferences model, Gârleanu and Panageas (2015) review the scenario where differing risk aversions in a time-varying wealth setting, the less risk averse people then to hold more stock. They explain the dynamics, where sharp drops in the market will mean that the people with the lower risk aversion will account for a smaller proportion of the market, and hence the risk aversion will increase as a result. Cochrane (2017) defines this relationship in the following model, where two consumers are denoted as A and B respectively:

$$\Lambda_t = \lambda_A e^{-\delta t} C_{A,t}^{-\gamma_A} = \lambda_B e^{-\delta t} C_{B,t}^{-\gamma \lambda_B} \tag{51}$$

Λ_t is a discount factor and λ_i are time invariant Pareto weights. Phrased using risk-aversion terminology, I get $C_t = C_{A,t} + C_{B,t}$ for total aggregate consumption across time and discounting (aggregate of all individuals), denoted as:

$$\frac{1}{\gamma_{m,t}} = \frac{1}{\gamma_B}\frac{C_{B,t}}{C_t} + \frac{1}{\gamma_A}\frac{C_{A,t}}{C_t} \tag{52}$$

From the model, I recognise the same working mechanics as the habit model: after the aggregate individuals' sharp decrease in consumption reduces their risk capacity, this is due to a relative increase in risk-averse people during stressed periods. This poses a good question: why do markets show changes in risk aversion in stressed markets, however the individuals do not? Cochrane (2017) elaborates on how risk aversion of the representative agent may change over time, whereas the risk aversion of an individual does not necessarily do so.

5.1.7 Rare disasters

Another reason that could go to explaining the equity premium is the substantial effect and perceived risk from a very low probability event with significant impact, otherwise known as the 'Peso problem' (Back, 2017). Barro (2006) notes that macro-finance models have made the connection of asset-pricing puzzles and rare disasters. These models are controlled where the equity premium and other asset-pricing features are triggered by the fear of extremely unlikely events. The foundational assumption describes the relationship of real per capita consumption C, described in the following equation:

$$\log(C_{t+1}) = \log(C_t) + g + u_{t+1} + v_{t+1} \tag{53}$$

When $g \geqslant 0$ the productivity increases that is an exogenous variable. The rare shock variable is covered by u_{t+1}. The last term is set as a stochastic variable (i.i.d.) representing the expected levels of macro-economic variability. The arrival of these rare events are seen to arrive at a low frequency. Barro (2006) cite the use of a power-law density for the arrival distribution. To discuss, I make use of similar functionality later in this study, although the distribution for the rate of arrival is different.

Nakamura et al. (2013) provide a different approach in a model seeking to explain the consumption disasters with crises and recoveries. They conclude that the stock premium is low for a given level of risk aversion and that disaster risk dramatically magnifies the equity premium, which persists for years. This was a study using an empirical model of 24 nations spanning 100 years.

$$c_{i,t+1} = \tilde{c}_{i,t+1} + v_{i,t+1} + I^d_{i,t+1}\psi^d_{i,t+1} \tag{54}$$

$\tilde{c}$ represents the error in the measure of consumption c. The volatility of c is denoted at v and is normally distributed (i.i.d.). The second shock is denoted as $I^d_{i,t+1}\psi^d_{i,t+1}$ and this produces the disaster induced shock in consumption. This function has two parts, where I is a dummy variable and ψ the extent of addition consumption and has a distribution of $N(\mu_d, 1)$. Nakamura et al. (2013) has introduced a break in volatility with a second shock they refer to as a transitional variation in consumption due to rare stressful events. This features a binary variable I and ψ is the extent of the disaster due to the shock. Nakamura et al. (2013) explains that this model extends the work done in Bansal (2007).

5.1.8 Other considerations

Finding an explanation for why, or what drives, a factor risk premium is the real work of macro-finance (Cochrane, 2017). Factors can be explained as features that describe portfolio returns over time. Importantly, finance cannot by itself explain premiums. Using the CAPM model as an example, it will explain average stock returns but says nothing about equity risk premiums; it sits as a free parameter that needs macro-finance to explain it.

Cochrane (2017) argues that macro-finance is defined by risk premiums that fluctuate over time and the accompanying changes in risk aversion, that it is linked to precautionary savings, and that it is central to economic cycles. Business cycles are not caused by changes in the risk-free government interest rate or by consumers seemingly and collectively delaying their purchases in order to save money for a rainy day. Cochrane (2017) argues that recessions are definitely not times when individuals are just driven to be thrifty and, hence, make savings for a better tomorrow, and these are not times of high interest rates. The macro-finance view is that individuals are in a state of real fear and are concerned about job losses and business failures in their personal capacity. This is not simply a case of consuming less and having the wrong level of overnight rates. This fear also drives insurance premiums higher.

Cochrane (2017) concludes that no model is superior; they differ in many ways, but the habit and recession will always be a function of risk-bearing capacity, which influences investment decisions and the level of risk taken while investing. He points out that business cycles deal with risk premiums and movements from riskier aggregate positions to the safer options. What it is not is a detailed interpretation of subtle movements of the risk-free rate or driven by the view that during stressful periods an individuals need to substitute their consumption of today versus some future state changes, or is even a consideration.

5.2 Individual optimality

As Cochrane (2001) highlights, consumption-based asset prices provide the theoretical answer to most asset-pricing questions, but the link to actual markets is not as clearly demonstrated. The most popular use of models is the linear factor-based pricing models, which are most often calibrated in discrete time. The optimal consumption of utility-maximising individuals in a market can be studied with the use of an SDF, which will also take care of the state contingent dividends and development at various periods of time (Munk, 2013). In a consumption model an individual's consumption c at time $t = 0$ and again at time $t = 1$, is denoted as c_0 and c_1 respectively. The individual's initial wealth is $e_0 \geqslant 0$, which will be greater than zero at time $t = 0$. I assume this is based on the law of one price for cash invested from the client portfolio, denoted as θ, then the price of $P^\theta = \theta^T P = \sum_{i=1}^{I} \theta_i P_i$, and similarly, the portfolio will deliver dividends $D^\theta = \theta^T D = \sum_{i=1}^{I} \theta_i D_i$. The budget constraints are therefore, following Munk (2013):

$$
\begin{aligned}
c_\omega &\leqslant e_\omega + D_\omega^\theta = e_\omega + \sum_{i=1}^{I} D_{i\omega}\theta_i \quad \text{for all } \omega \in \Omega \\
c_0 &\leqslant e_0 - P^\theta = e_0 - \sum_{i=1}^{I} \theta_i P_i
\end{aligned}
\tag{55}
$$

The individual considers the portfolio θ and consumption at (c_0, c_1), which is also set to be non-negative, with a marginal utility at zero consumption levels following a CRRA. The consumption at c will be informed, just as the consumption at $t = 1$ will be informed based on the portfolio performance θ for all states ω, which become very large. In this model of solving optimal choice, by assuming a concave shaped utility curve, this allows for solving the problem by first-order maximisation, as the second-order maxima are already satisfied (Munk, 2013).

5.2.1 Time-additive expected utility

If we consider a model based on a concave utility, using a single date, set as a time-additive expected utility problem for individuals that would express their optimisation problem, following Munk (2013):

$$\max_{\theta} U(c_o) + \mathrm{E}[e^{-\delta} U(c_1)] \tag{56}$$

The subjective time preference rate is denoted as δ. When the budget constraints are added, then:

$$\max_{\theta} U\left(e_0 - \sum_{i=1}^{I} \theta_i P_i\right) + \mathrm{E}[e^{-\delta} U(e - \sum_{i=1}^{I} \theta_i P_i)] \tag{57}$$

The portfolio's first order criterion is:

$$-P_i U'\left(e_0 - \sum_{i=1}^{I} \theta_i P_i\right) + \mathrm{E}\left[e^{-\delta} D_i U'\left(e + \sum_{i=1}^{I} \theta_i D_i)\right)\right] \tag{58}$$

This formula then shows, for consumption c_1 and c_0 that:

$$\begin{aligned} P_i U'(c_0) &= \mathrm{E}[e^{-\delta} D_i U'(c_1)] \\ P_i &= \mathrm{E}[e^{-\delta} \frac{U'(c_1)}{U'(c_o)} D_i] \end{aligned} \tag{59}$$

This forms the consumption plan that is generated by the θ optimal portfolio at $t = 1$, which can then be restated as the price. This can then be restated as the price deflator:

$$\xi = e^{-\delta} \frac{U'(c_1)}{U'(c_o)} \tag{60}$$

This is a good time to explain that the state price deflator ξ represents the willingness of the individual consumption at time $t = 0$ for the individual to hold off and consume at time $t = 1$.

5.2.2 Habit formulation in a discrete time framework

This framework extends so that individuals that have a investing strategy $\theta = (\theta_t)_{t=0,1,\ldots,T}$, choose consumption at various states or time periods, $c = (c_t)_{t \in T}$ and where time is a defined set, time t. The individual also has a starting endowment then the endowment is denoted by $e = (e_t)_{t \in T}$, with income, e_t, at state times t. If I now consider a time non-additive preference configuration, the objective function is as follows, per Munk (2013):

$$\max_{\theta=(\theta_t)_{t=0,1,\ldots,T-1}} \mathrm{E}\left[\sum_{t=0}^{T} e^{-\delta t} U(c_t, h_t)\right] \tag{61}$$

The habit level, h_t, is governed by $h_t = \beta c_{t-1}$, and from this we know that prior habits drive current habits, time preference rate is denoted by δ, this is a one period framework. β will lie between 0 and 1, and if $\beta = 0$ then we are back to a time additive utility. The variational argument is applied, $c = (c_t)_{t\in T}$ and $h = (h_t)_{t\in T}$ that denote consumption and the habit levels respectively. ε represents the perturbation in consumption by purchases. The habit levels will be effected from the perturbation as follows $h_{t+1} - \beta\varepsilon P_{i,t}$ and $h_{t+1} - \beta\varepsilon(D_{i,t} + P_{i,t})$ for time $t = 1$ and $t = 2$. This function then features quite a few variables, following Munk (2013):

$$\begin{aligned}
&U(c_t - \varepsilon P_{i,t}, h_t) - U(c_t, h_t) \\
&+ e^{-\delta}\mathrm{E}_t\left[U(c_{t+1} + \varepsilon(D_{i,t+1} + P_{i,t+1}), h_{t+1} - \beta\varepsilon P_{i,t}) - U(c_{t+1}, h_{t+1})\right] \\
&+ e^{-2\delta}\mathrm{E}_t\left[U(c_{t+2}, h_{t+2} + \beta\varepsilon(D_{i,t+1} + P_{i,t+1})) - U(c_{t+2}h_{t+2})\right] \leqslant 0
\end{aligned} \tag{62}$$

This function is then divided by ε, with allowing $\varepsilon \longrightarrow 0$ and rearranging terms and as Munk (2013) explains the last step involves the law of iterated expectations, arriving at:

$$P_{i,t} = \mathrm{E}_t\left[e^{-\delta}\frac{U_c(c_{t+1}, h_{t+1}) + \beta e^{-\delta}\mathrm{E}_{t+1}\left[U_h(c_{t+2}, h_{t+2})\right]}{U_c(c_t, h_t) + \beta e^{-\delta}\mathrm{E}_t\left[U_h(c_{t+1}, h_{t+1})\right]}(D_{i,t+1} + P_{i,t+1})\right] \tag{63}$$

and finally the state price deflater is represented by the following formula:

$$\xi_t = e^{-\delta t}\frac{U_c(c_t, h_t) + \beta e^{-\delta}\mathrm{E}_t\left[U_h(c_{t+1}, h_{t+1})\right]}{U_c(c_0, h_0) + \beta e^{-\delta}\mathrm{E}\left[U_h(c_1, h_1)\right]} \tag{64}$$

As discussed, the formula reverts to time additive utility when $\beta = 0$. If the habit levels are considered to be related to all prior values, possible in a way that aggregates, then I could use $h_t = \sum_{s=0}^{t-1}\beta^{t-s}c_s$. The state price deflater will then be, following Munk (2013):

$$\xi_t = e^{-\delta t}\frac{U_c(c_t, h_t) + \sum_{s=1}^{T-t}\beta^s e^{-\delta s}\mathrm{E}_t\left[U_h(c_{t+s}, h_{t+s})\right]}{U_c(c_0, h_0) + \sum_{s=0}^{T}\beta e^{-\delta s}\mathrm{E}\left[U_h(c_s, h_s)\right]} \tag{65}$$

As Munk (2013) notes that the formulation is complicated with many terms. This formulation has a simplifying assumption of the utility function, $U(c, h)$, that assumes the individual utility is dealt with via exogenous factors, which then brings it back to a time-additive utility with marginal utility linked to exogenous factors.

$$\xi_t = e^{-\delta t}\frac{U_c(c_t, X_t)}{U_c(c_0, X_0)} \tag{66}$$

This is the state price deflator for a discrete time preference rate δ with the time additive preferences in the utility function $U(c, X)$ which feature exogenous stochastic process $X = (X_t)_{t\in T}$, where the individual's optimal consumption plan $c = (c_t)_{t\in T}$ and the deflator process is $\xi = (\xi_t)_{t\in T}$.

5.2.3 Dynamic programming

In this section, I introduce optimisation techniques for solving complex problems that span multiple time periods. These solutions are termed "dynamic programming". I'll share an amusing anecdote from Bellman (1984) about how he came up with the name:

> What title, what name, could I choose? In the first place I was interested in planning, in decision making, in thinking. But planning, is not a good word for various reasons. I decided therefore to use the word, 'programming'. I wanted to get across the idea that this was dynamic, this was multistage, this was time-varying—I thought, let's kill two birds with one stone. Let's take a word that has an absolutely precise meaning, namely dynamic, in the classical physical sense. It also has a very interesting property as an adjective, and that is it's impossible to use the word, dynamic, in a pejorative sense. Try thinking of some combination that will possibly give it a pejorative meaning. It's impossible. Thus, I thought dynamic programming was a good name. It was something not even a congressman could object to. So I used it as an umbrella for my activities.

5.2.4 Optimality in discrete time

Firstly, we need to define maximum expected utility as the key equation for indirect utility, but this will need to be expressed over the current and future utility of consumption (Munk, 2013). Keeping in mind that wealth levels W_t will determine the levels of consumption, as will larger investments that can affect consumption in the long-run, all these variables that can affect consumption will be bound by one parameter, x_t, where indirect utility is denoted as J. Following Munk (2013):

$$J_t = \sup_{(c_s, \theta_s)_{s=t}^{T}} \mathrm{E}_t \left[\sum_{s=t}^{T} e^{-\delta(s-t)} U(c_s) \right] \tag{67}$$

E_t denotes the start of a continuous-time Euler equation. Munk (2013) explains that x_t is concerned with multiple variables and when one review (W_t, x_t) is a Markov process, in which case the indirect utility takes the following form $J_t = J(W_t, x_t, t)$. This problem requires no portfolio choice for the final step, and this means we set θ_T to zero. I display a very important assumption for this approach is the envelope condition, which is the partial derivative of J with respect to Wealth, W_t:

$$U'(c_t) = J_W(W_t, x_t, t) \tag{68}$$

The logic inherent in this model is that one would need to attain equal value from consuming an extra unit at a time as would be returned from the markets from investing in an optimal way. From this envelope condition, the following can be determined, the state price deflator:

$$\xi_t = e^{-\delta t}\frac{U'(c_t)}{U'(c_0)} = e^{-\delta t}\frac{J_W(W_t, x_t, t)}{J_W(W_0, x_0, 0)} \tag{69}$$

If I now turn to the portfolio dynamics, where risk free assets and market exposed risky assets are invested into a portfolio θ. Following Munk (2013), I see the wealth of an individual from t to $t = 1$ is captured in the following program.

$$\begin{aligned}
W_{t+1} - W_t &= \sum_{i=0}^{d} \theta_{i,t}(P_{i,t+1} - P_{i,t}) + \gamma_t - c_t \\
&= W_t + \gamma_t - C_t + \sum_{i=0}^{d} \theta_{i,t}P_{i,t}r_{i,t} \\
&= (W_t + \gamma_t - C_t)\left(1 + \sum_{i=0}^{d} \pi_{i,t}r_{i,t}\right) \\
&= (W_t + \gamma_t - C_t)R^W_{t+1}
\end{aligned} \tag{70}$$

θ represents a fraction of investment in different assets $\pi_{i,t} = \frac{\theta_{i,t}P_{i,t}}{W_t+\gamma_t-c_t}$. The risky assets θ_t, where it is assumed for now that there are no dividends paid. γ_t represents the individual's income at time t. $R^W_{t+1} = \pi_t^T[R_{t+1} - R^f_t 1]$. Maximising over the entire solution, for both current and future expected consumption is a complicated task, hence breaking it down, as follows, into smaller maximisation problems, following Munk (2013):

$$\begin{aligned}
J_t &= \sup_{(c_s,\pi_s)_{s=t}^T} \mathrm{E}\left[\sum_{s=t}^{T} e^{-\delta(s-t)}U(c_s)\right] \\
&= \sup_{(c_s,\pi_s)_{s=t}^T} \mathrm{E}\left[U(c_t) + \sum_{s=t+1}^{T} e^{-\delta(s-t)}U(c_s)\right] \\
&= \sup_{(c_s,\pi_s)_{s=t}^T} \mathrm{E}\left[U(c_t) + \mathrm{E}_{t+1}\left[\sum_{s=t+1}^{T} e^{-\delta(s-t)}U(c_s)\right]\right] \\
&= \sup_{(c_s,\pi_s)_{s=t}^T} \mathrm{E}\left[U(c_t) + e^{-\delta}\mathrm{E}_{t+1}\left[\sum_{s=t+1}^{T} e^{-\delta(s-(t+1))}U(c_s)\right]\right] \\
&= \sup_{(c_t,\pi_t)} \mathrm{E}_t\left[U(c_t) + e^{-\delta}\sup_{(c_s,\pi_s)_{s=t+1}^T} \mathrm{E}_{t+1}\left[\sum_{s=t+1}^{T} e^{-\delta(s-(t+1))}U(c_s)\right]\right]
\end{aligned}$$

This then is a set of equalities, built by separating the current time period from that of the future. The inner supremum is the same as indirect utility; however, for $t = t + 1$, I then arrive at:

$$J_t = \sup_{c_t,\pi_t} E_t\left[U(c_t) + e^{-\delta}J_{t+1}\right] = \sup_{c_t,\pi_t}\{U(c_t) + e^{-\delta}\mathrm{E}_t[J_{t+1}]\} \tag{71}$$

This is also known as the Bellman equation, with J, the indirect utility, having dynamic programming properties. Munk (2013) explains that one can consider the optimal decisions for the current time period and then consider the optimal decisions for future periods, which also need to be considered together as they are inherently linked due to the c_t and π_t effect on future investments, portfolios, and consumption. The dynamic programming is suited to the backward iterative procedure that begins with an optimal view based on terminal values using first-order conditions, or an Euler equation subject to constraints per equation 70 above. This requires numerous of scenarios to work with. Ultimately, the maximisation will depend on the choices for π_t and c_t. To solve for optimal decisions, the modeller will make use of the indirect utility function, as this problem cannot be solved directly, which is where the envelope condition is used, and solved using the partial derivatives. This process will eventually lead back to the SDF.

6 Summary of analysis

Asset pricing, like economics has a both positive and a normative conflict. Cochrane (2005) observe that this is to do with how one seeks to answer a question, from the perspective of 'how the world works' versus 'how the world should work'. Asset pricing can be understood by posing questions about asset prices, as Cochrane (2005) notes, can be posed in two ways. Firstly, the absolute prices must be reviewed with the goal of explaining what makes up the full price of the asset. This will require an understanding of asset exposure to macroeconomic risks. Secondly, the relative pricing route that seeks to answer, given the prices of other assets in the market, what the price of the asset is, with very little explanation of where the price comes from or what fundament factors the asset is exposed to (Cochrane, 2005). The researcher cites the Black-Scholes model as an example of this relative approach. Cochrane argues that, and I agree, the most appropriate way to approach an asset-pricing problem is to consider both techniques. This allows one not to be caught in either extreme therein becoming idealistic, but to rather review and assess the problem at hand, then being purposeful in one's calculation methods (Cochrane, 2005). In this study, I endeavour to walk this path, seeking to proxy with data from areas where data is available and at all times to provide an adequate absolute perspective from the research and theory. This provides a solid departure point, because I know, regardless of my data-access constraints here, there is no appropriate end-to-end data available at this point in time. So, I will thus compile a pricing framework which I describe as an ensemble of many methods and techniques, to my mind is structured, and always underpinned by dedicated research. I posit that, relative to investing models in use for data-rich assets, this framework be applied to private debt as part of an SAA. Such is my most significant innovation in this study – it brings closer the alignment of modelling frameworks with more sophisticated data rich asset-pricing environments. From the perspective of introducing and alternative asset in an SAA process, this is, I argue, crucial.

In this part I provided the broad set of foundational concepts. I also provided an up to date review, in both business cycles and asset-pricing, assisted with a macro-finance framework. I started by investigating business cycles, in which we review the main classes of models, classical and neo-classical models. The recent use of post Keynesian models that focus heavily on monetary policy are under scrutiny. Another school of thought brings the behaviour and beliefs in as they are now known to have a bearing on and helping us to understand the nature of the driving factors of business cycles.

I detailed the key modelling techniques and approaches taken in literature. Recent evidence points to the link between credit and the business cycle, in which tradable versus non-tradable lending can have a direct bearing on the characteristics of the credit cycle. Interesting insights included are the direct contribution that entrepreneurs make to on cyclical variability. Looking further into the research on the behavioural conditions and credit lending during business cycles, I found evidence citing the powerful effects of diagnostic extrapolation on large changes in business cycles.

I also detailed the key asset-pricing models and broken down the factor risk premium framework and the most prominent models that explain the asset-pricing models. The challenge with all asset-pricing models is their ability to replicate what we observe in the market, these not being able to account for all the known market 'puzzles'. Lastly I detail the key models that one would use when considering the intertemporal portfolio choice. These models rely on an important, classic, technique for optimisation by dynamic programming using the Bellman equations.

Part III

Forecasting risk appetite using machine-learning sentiment analysis

An exciting consequence of the advent of alternative data coupled with machine-learning techniques is the topic for this part. As a part of my research into the causes and effects of business cycles by making use of machine learning techniques coupled with alternative sources of information. This investigation required that I codify a natural language programming interface to multiple sources of unstructured data, with data processing components and data analytics. The analytics outputs were then used to investigate usefulness in a time-series modelling analysis. This was a very interesting journey, that required significant effort to research the available data, then to source the data (freely available) that adequately covered each region (including South Africa, United Kingdom and United States) in this scope of study. I also provided a detailed review of the modelling techniques for use in the field of NLP, as it pertains to modelling in time-series analysis. This study was interesting, challenging at times (as the effort required to process data is significant), but the insights from the analysis are revealing, especially to do with the relative merits of modelling techniques and alternative data.

7 Introduction

In addition to its practical uses in business, sentiment analysis is a rapidly expanding field of study in the academic world. For instance, a recent work by Azar and Lo (2019) demonstrated that mathematical techniques may be successfully applied to massive amounts of voice data in order to identify emotion cues from the Federal Open Market Committee (FOMC) speeches. Dang et al. (2020) discuss areas of application, including modelling product review data for user acceptance, voice recording processing to understand products and processes, product and client recommendation systems. Sources of input for sentiment analysis include web-scraped data from internet pages, online social media, blog sites, and user forums, all of which require a big data approach for processing unstructured data (Dang et al., 2020). This new era of alternative data in the financial sector is promising. My interest is in how to process signals for macro-economic and asset class returns based new data types and analysis of this kind.

Questions of interest and focal points The following questions have been asked prior to the design and build of the models and I will end this part with a conclusion that includes a review of these questions.

- Are social media data appropriate or value adding in macro-econometric and investment signal research?
- Does the inclusion of additional sources of information, such as social media or news data, improve modelling accuracy of sentiment indices?
- Does the geographic meta data increase predictive power of modelling sentiment indices?
- Do more independent factors improve the predictive power when modelling sentiment indices?
- Do lagged factors improve the predictive power when modelling sentiment indices, for both single-factor and multi-factor models?
- Will traditional modelling techniques improve machine-learning only models?
- How effective are freely available modelling frameworks and trading?

Alternative data in finance Pozzi et al. (2017) point out that social media has quickly become a part of peoples everyday lives. Popular platforms include blogs, online chats forums, and platforms such as Facebook, LinkedIn, Twitter, Instagram, YouTube, WeChat and WhatsApp. We now have a wealth of data to describe individuals, markets, and networks. The initial use of analytics to comprehend this data begins with so-called vanity metrics such as 'number of clicks,' 'number of followers,' and 'number of page visits' (Pozzi et al., 2017). More recently the types of analysis that can further improve our understanding of the data and reveal new information, this includes how emotion and opinion are conveyed in natural persons writing and this is achieved with natural language processing techniques. Pozzi et al. (2017) explain that an online digital media platform is media hosted on digital technologies that are different from those classic media sources, including radio, press, television, etc. A subset of this digital media is social media and social networks, which include those platforms (not only the world wide web), where people connect and share information with each other.

Base data sources include data feeds, news articles on blogs, and social media data such as Twitter feeds. The data may then be examined for information and powerful signals, such as demand-side sentiment and economic signals using machine-learning on signal processing. Preqin (2019) identifies a force of change in the application of advanced data analytics to the creation of more customised investments products. The most probable effects will be on the incorporation of alternative data into credit evaluation, and specifically on the quality of credit. Fundamental data, although it is differentiated by source (like individuals trends, business process and sensor data), it is characterised because it is primary information which has not been used (Prado, 2018). He goes on to explain that alternative data can often be difficult to process and even harder to use (transform, filter, store), it does offer the allure of new information which could translate to alpha (or more efficient risk-adjusted returns).

The three core types of financial data		
Fundamental data	**Market data**	**Alternative data**
Balance sheet information	Prices	Satellite/CCTV
Income statement information	Returns and dividends	Internet search data
Macro-economic variable information	Risk numbers (e.g. volatility)	Social media sentiment
Financial ratios of companies	Fixed-income data (yield, coupons, duration)	News Sentiment
Earnings data	Rates data	Twitter chats
Analyst recommendations	Commodities / property specific data	Card expenditure data
	Credit ratings	Private debt scoring history

Table 1: The breakdown of broad data types in the table explains that fundamental data is well-known and a normal input into the market pricing process, thus regularised with low value to add in terms of alpha contributions. Market data encompasses all publicly available pricing and trading information from exchanges. As I move to the right, we start to see more diversified sources of information, ranging from satellite data showing infrastructure project progress to retail credit card spend trends to internet browser searches (Prado, 2018)

.

8 Basics of NLP and sentiment analysis

In terms of introduction, linguistic programming is concerned with language at increasingly complex levels. In the first chapter of his book, Mitkov (2005) explains that the first level is the morphological level, which examines word structure and word creation by examining each word in isolation. In the process of deconstructing words (i.e., prefix, stem, and suffix), which are the smallest units of meaning, morphemes are defined. Prefixes and suffixes can alter the meaning of stemmed words; for instance, the prefix 'un' modifies the meaning of 'untidy'.

The next level, lexical, is concerned with the meaning of words within the context of speech, explains Mitkov (2005). This technique extracts more meaning than the individual words, producing a lexeme as the fundamental unit of lexical meaning. An example in finance is the difference between 'this show may collect a lot of *interest*' and 'this loan may collect a lot of *interest*. Next, a syntactic level analysis is performed by tagging the output of the lexical analysis' part-of-speech tags. Complexity in this field increases when one ventures into semantic, discourse, and pragmatic processing, which is not covered as part of the scope in this study. Kurdi (2016) indicates that NLP is an interdisciplinary field incorporating computer science, cognitive psychology, and artificial intelligence (AI). The classic approach to natural language processing and sentiment analysis consists of a series of discrete tasks (Kurdi, 2016). Liu (2015) contributes that the prerequisite for sentiment modelling is to determine which scoring level applies. The first level is at the message level, which is a review of a whole passage of text or paragraph. At the sentence level, sentiment scoring is applied to each sentence, followed by fine-grained analysis, which is based on the idea that an opinion has a sentiment component and a target.

8.1 Component parts of the NLP process

The following general steps are a high-level introduction to an NLP process and pipeline:

1. **Tokenisation.** This involves processing substantial amounts of text (paragraphs or sentences) into words, or chunks of data, which can be seen as the unit of grammar in text processing languages (Goldberg and Hirst, 2017). This is slightly more challenging than splitting by spaces, as words like 'high-end' need to be either split or collected as a term, which makes the task complicated and involves establishing the set of single words, terms, and n-grams (of words like 'ice cream', which can be seen as a bi-gram).

In this first step of the NLP pipeline, care must be taken the rest of the analysis being impacted by the decision made here.

2. **Lexical analysis.** The next step will be to find the morphemes and grouping the words into lexemes. Then, the next step involves grouping together tokenised words to form single meanings in a process known as stemming or lemmatization. For example, 'sickness' and 'sickest' would be mapped to the word 'sick'. This can then be applied to term frequency distributions in documents.

3. **Syntactical analysis.** The analysis focuses on capturing the meaning of text, builds up a structure for a part of the text in a sentence and allows for the relationships between words and the dependency between words to be established. This can then be illustrated using a 'parse tree', a sort of structural and hierarchical diagram. This phase entails checking the grammar rules to determine the meaning of the text and to detect deviations in sentence structure from the grammar rules. This includes part-of-speech mapping, which is a key grammatical task and involves mapping each word to one of the eight parts of speech (noun, pronoun, adjective, verb, adverb, preposition, conjunction, and interjection). The output of the task is to apply a Part-of-speech (POS) tag to each token, which can then be linked to a syntactical category. Syntax is the grammatical structure of the text, whereas semantics is the meaning being conveyed. This allows a program to determine whether a sentence, such as 'hot on sauce any-more' or 'man corn-dog his this', makes any sense according to grammar rules.

4. **Semantic Analysis.** The semantic analysis operation articulates the proper meaning of a sentence. This is an analysis that is broader than sentiment. This analysis will also identify the meaning held in the text, the relationships between words, expressing how the meaning of text changes with various combinations of words. This area that can become very complex and nuanced.

5. **Corpus.** This is referring to a collection, or body, of texts (written or spoken), used in language research, that allow a machine to annotate text or specific words. There are numerous corpora available, which can be domain-specific (medical, financial, or military) or language-specific. As language is constantly changing over time, to be updated constantly. The ease with English is that because of the features of language,

or the significant use of the language rather than the size of the corpus or demand (Boiy and Moens, 2009).

8.2 Framework for dealing with unstructured text data

The framework is a structured process that is put in place to extract the key parts of the text that can provide information; and to map words to remove features that serve no purpose. The following frameworks are in place for this study:

- **Pre-processing**
  - ◇ Removal of punctuation
  - ◇ Setting text to lower case
  - ◇ Remove stop words which are uninformative, such as 'the' and 'and'
- **Case-folding** is the process by which decisions are made to normalise words consolidating words with different spellings but the same meaning. Dealing with *all-caps* words is another aspect of case folding Goldberg and Hirst (2017).
- **N-gram**: I touched on this earlier, N-gram refers to combining words into term-based vectors, which may be composed of one word (uni-grams) or two words (bi-grams) or longer (length dependent N-gram). An example of a bi-gram is 'machine-learning'.
- **Stemming and lemmatization** of words removes redundancy and assists in frequency-based measures (so words such as 'frequent' and 'frequently' are set to the same stem). The final root form of the word is mapped; and this process is called lemmatization (Rao and McMahan, 2019).
- **Part of speech mapping**: This is made simpler by the lemmatization step, however this process involves marking up each word's functional role in the sentence and tagging it with the part of speech (POS) tag. For example, a POS is mapped in brackets: 'David (*propn*) ate (*verb*) the (*det*) cold (*adj*) meat (*noun*)'.
- **Negation**: This process of changing the valence of a word to negative (or positive) depends on whether there are other words immediately before it that may change the meaning, for example, 'not an amazing experience'.

In Figure 7, I provided an illustration of natural language processing methods at a high level. One may read how a sentence is influenced at each phase, beginning with text replacement, filtering, tokenising paragraphs into smaller word-size pieces, stemming to prepare for mapping, and lastly constructing the syntactical tree. In this study, I perform these codification steps using Python and the Python Natural Language Toolkit (NLTK) package for natural language processing of unstructured text data.

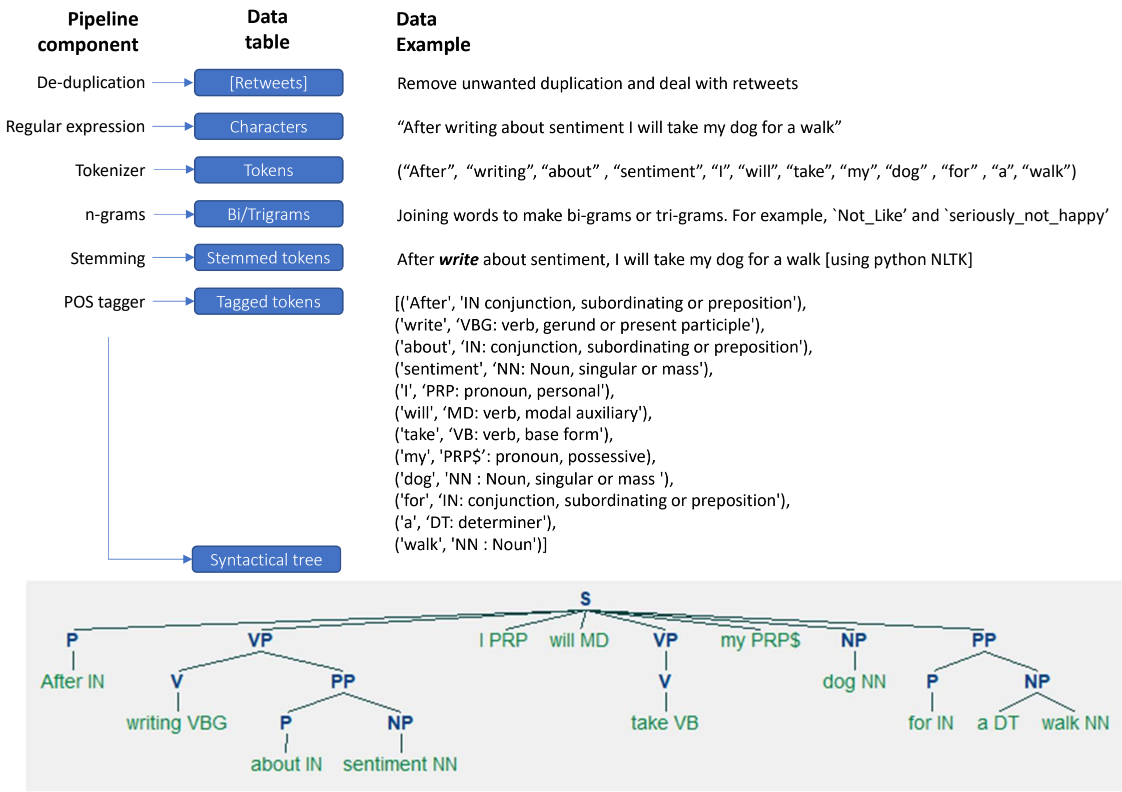


Figure 7: Example layers for an NLP pipeline. At the POS tagging step, the information is mapped to words so as to enable further filtering and information addition. The steps are carried out using the Python NLTK package.

8.2.1 Nuances of Twitter data ('un-wanted tweets').

Modelling with NLP can be a challenge for one who is used to structured data. I invested a significant amount of time in learning methods of extraction and processing, for instance, by collecting from websites, making use of API's and data storage techniques in Python using formats such as xml or json. This was not time wasted, as the challenges of NLP were also heightened once the Twitter data was collected in a database. However, I need to highlight that the process of collecting Twitter data was by no measure simple or easy, it involved the near constant supervision to stream data; lasting more than a year. This is because Twitter does not supply large amounts of free historical data; therefore streaming was the only option. In contrast, the collection of newspaper data has been made easy as the APIs are relatively simple to connect to, the data is well organised, and one can work with the data immediately.

Twitter data, however, requires a fair amount of cleaning prior to use. The data would not even parse in the Python NLP package NLTK or other packages fit for NLP modelling purposes. Cleaning included the removal of white spaces, punctuation marks, non-characters, retweets (RT), 'RT @', '@links' and emoticons, foreign language, hidden characters that are obviously hard to detect. A note on retweets; I needed to work on multiple filtering methods to ensure that the retweets are either removed, or de-duplicated. This was frustrating as I had already set the data feed up as a filtered search in which retweet indicator is already set to exclude retweets. Table 2 below shows that the largest proportion of tweets still came through as being manually retweeted. One must also consider whether to remove the retweets in the first place, depending on the modeller's goals. As a retweet could be seen as a positive momentum, as a signal. Basically, from the representative sample statistics in Table 2, most of the data was not useful.

Tweet sample	Filtered sample of tweets	% of total sample
Regular tweets (usable data)	3,180	18.52%
Reply	2,516	14.65%
Retweet (RT @)	7,158	41.68%
Foreign language	4,321	25.16%
Grand Total	17,175	100.00%

Table 2: Sample of the Twitter data collected (streamed on 19^{th} Dec'20). Most of the data was not useful and not possible to filter out in the collection process, one needs to model a filtering process.

8.3 Descriptive analysis of data sources

A descriptive analysis is used to quantitatively describe the contents of the data. The purpose is to provide estimates as they relate to the study objectives by using simple summaries, tables, and graphical analysis. We may be interested in presenting a cursory review of patterns and unexpected results that may require further inquiry in the study.

8.3.1 Social media: Twitter data

Twitter is a freely available micro-blogging website and mobile phone application that allows free publication of 140 character posts called 'tweets'. Twitter provides streaming data via the Twitter API version 2, as well as historical data at a significant cost.

Figure 8: The Twitter word cloud is a representation of the most frequently used words from the Twitter database. The sizes of the words are determined by their frequency. Importantly, this analysis is based on the remaining text after the stop words have been removed, emojis, special characters have been removed, and lastly, internet address strings such as 'https'.

8.3.2 News data: The Guardian

The Guardian newspaper is a daily broadsheet that has been published since 1821. The Guardian provides complete access to news articles that are freely available via API. The UK news regulator, OFCOM [6], provides readership reports on news sources, including TV, newspapers, and social media. With a readership of about 22%, The Guardian newspaper is unquestionably a mainstream news source in the UK. Relative to other news sources, the

[6] Ofcom is the regulator for the communications services, this include news.

readership is reasonably well spread across gender, age, and the UK National Readership Survey social grading of ABC1 (middle class) and C2DE (working class), with a slight tilt toward ethnic versus white readers (OFCOM, 2020).

8.3.3 News data: The New York Times

The New York Times is a daily morning newspaper from New York City. It has a long history, having been founded in 1851. It is an important newspaper that is regarded as the 'newspaper of record'. The New York Times makes data freely available via an API.

8.3.4 News data: GDELT

Google built a free sentiment scoring solution and this been available since January 2017 that is based on a massive database of the world's broadcast, print, and web news sites. The project provides free outputs and is supported by Google Jigsaw and GDELT.

9 Methods for modelling of NLP and sentiment analysis

Liu (2015) states that sentiment analysis (SA) is a technique that originated in the field of computer science and is mostly used to classify text into categories. It has entered the realms of economics and finance as well as management science. In the field of natural language processing, sentiment analysis has seen a great deal of attention. *Opinion mining* is another name for sentiment analysis. Pozzi et al. (2017) enlighten that businesses and governments both place a premium on citizens' thoughts and opinions. In sentiment analysis, the bulk of the effort goes toward labelling utterances and natural language text as either positive or negative. Other methods of analysis, such as subjectivity analysis, opinion and sentiment extraction, and emotion analysis, are listed in Pozzi et al. (2017).

There are two types of models used in text-based analytics and linguistics: those that use predefined rules and/or algorithms to convert text from unstructured data into quantitative measures (referred to as 'rules-based' methods). The rules-based approach includes methods that incorporate detailed mapping and complexity-based corpus rules, known as lexical methods. The second category includes methods that employ statistical approaches, which can range from simple frequency type measures to more sophisticated learning approaches known as machine-learning. Kalamara et al. (2020) notes that the most common technique used for text analysis are rule-based approaches, as these techniques are freely available and their use does not require calibration (which is an intensive data exercise that consumes large volumes of data with specific hardware processing requirements to deal with the load).

Shapiro et al. (2020) also explain that sentiment analysis is a key method in NLP and is primarily a classification problem, in which sentiment texts (passages, paragraphs, sentences, and words) are classified as either positive, negative, or neutral; this is often then added to a scoring/rating model that generates valence (rating from 1–5). This allows the user to specify more grades between positive and negative. Shapiro et al. (2020) posit that NLP sentiment analysis can be classified using the following broad approach to standardising the scores into categories for analysis; such will be adopted in this study. The Valence Aware Dictionary and sEntiment Reasoner (VADER) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiment expressed in social media. VADER is fully open-sourced under an MIT License. VADER calculated a compound score is the main output of the VADER tool; this is a valence score calculated by adding the scores of each scored word in the document, which is then normalised between -1 and 1, with -1 (1) being the most extreme negative (positive) sentiment score; and standardised by establishing bands, that I illustrate below.

$$\text{Sentiment band} = \begin{cases} \text{compoundscore} <= -0.05 & \longrightarrow \text{negative sentiment} \\ -0.05 < \text{compoundscore} < 0.05 & \longrightarrow \text{neutral sentiment} \\ \text{compoundscore} >= 0.05 & \longrightarrow \text{positive sentiment} \end{cases} \tag{72}$$

Shapiro et al. (2020) constructed a series of sentiment indices using a simple 'bag of words' approach. This included the Harvard General Inquirer (GI) and the updated version of Loughran's indices (Loughran, 2014), which have the benefit of being finance-domain-specific; and are a useful contextualising example of the various sentiment analysis techniques deployed. Language is complex; simply mapping words to meaning in a general manner can produce inaccurate results. For example, words such as *interest*, *liability*, and *tax* carry different meanings when used in the context of finance and family relationships. According to Shapiro et al. (2020), the general 'off-the-shelf' models are widely available and free; however these models are frequently tailored to a specific domain, such as movie reviews. In the context of financial and macro-economic outcomes this may produce an unacceptable level of accuracy.

The Hu and Liu lexicon (L&M) is a lexicon approach based on movie reviews featuring more words mapped to negative and positive sentiment. Hu and Liu (2004) find that none of the lexicons are outright winners in this task of scoring finance and economics news based text. Tim Loughran and Bill McDonald (2014) appears frequently in economics and finance articles but not in general text. Given this, Shapiro et al. (2020) decided to combine the outputs of lexicons to increase the power of the models. They also note that a large number of words were simply not scored. According to Shapiro et al. (2020), VADER's performance is equivalent to the combined L&M lexicon. A simple but powerful rule added to the 'bag-of-words' approach is the negation rule (multiplying the valence score by -1 for each word found with a negative word within three words preceding it).

This increment improved the scores slightly. Overall, the highest scoring model was the L&M and Liu's lexicon model with the negation rule applied. Shapiro et al. (2020) consider two other lexical based NLP models. One is the GloVe, developed by the Stanford NLP group (Pennington et al., 2014), the other is BERT, developed by Google (Devlin et al., 2018). Both models use pre-trained word vectors that embed semantic information about each word. BERT also has bidirectional, context-aware word and document embeddings that review text both forward and backward, which makes for a powerful NLP model. Shapiro et al. (2020) tested using the 700 articles in the training and development data set and testing using 100 articles, their results show the GloVe model performs worse then and the BERT model performs better than the lexical models.

9.1 Preliminary definitions, notations and terminology

In preparation for the modelling and literature section, I formalise the following foundational terms, following Blei et al. (2003):

- a single unit vector $(1, \ldots, V)$ is used to represent words. This structure is also fit for purpose for working with a corpus, a dictionary or vocabulary. Using the superscripts to represent component parts, the v^{th} word is vector w, such that $w^v = 1$ and $w^u = 0$ for $u \neq v$.
- a document, M, is sequence of words, N, denoted by $\mathbf{w} = (w_1, w_2, \ldots w_N)$, where w_n is the n^{th} word in a sequence.
- a corpus can then be defined as a collection of documents, D where $D = (\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_M)$.

9.2 Lexical, rule-based, models

Pozzi et al. (2017) document that the first sentiment resource was started in 1966 and was configured so that sentiment was scored as a categorical binary output, called General Inquirer. The first lexicon, that looks at nuances across and within parts-of-speech, is the extension to WordNet (from Princeton University) called SentiWordNet. In this tool, sentiment is seen as a gradient and not as a categorical feature. This lexicon is seen to have a positive bias, though, and does not detect negative opinions well.

9.2.1 Simple lexical approaches

Lexical approaches are based on a database or corpus of predefined words that are tagged for emotion, which in this case is sentiment, and sentiment valence is added. This simple approach uses sets of words to match against and then tag emotional characteristics, the structural hierarchy, and parts of speech. Lane et al. (2019) explain that this approach will not consider the context of a word or order within the text; this is often termed the 'bag of words' approach. Pozzi et al. (2017) also designed a popular lexicon, the opinion lexicon, based on 6800 English words manually classified as positive or negative. The lexicon also uses a part-of-speech approach to assess polarity and thus can recognise the ambiguous parts of a sentence that would not be understood at the word level.

9.2.2 Simple dictionary-mapped models

The next of the simple models is based on a mapping to dictionary terms, tags, definitions. The user would be able to easily provide sentiment (or other scores), based on specific words. Again the use of stemming can help reduce the corpus load. Simple mappings may include high sentiment scores for 'awesome' and low scores for 'horrific'. Passages of text with these words appearing frequently, or at all would presuppose a simple sentiment score by mapping to words found in the text. A significant benefit of the approach would be the ease of being able to manage specific domain related terms, such as the word 'growth' in economics may arguably carry a higher sentiment mapping than in cancer research. Simple dictionary-based methods would not be context-aware and would potentially score the sentiment in a phrase incorrectly due to the word being part of a sentence with meaning (e.g., 'less than expected growth').

9.2.3 Orientation-based sentiment lexicons models

Lexicon models use dictionaries to mark words as negative or positive, which is why they are called binary orientation lexical classifications. Labelling words according to their sentiment is not a new technique; it is a central goal in linguistic programming. Other objectives include opinion and sarcasm detection. Hutto and Gilbert (2014) summarise the different notable lexical-sensational analysis models. The first was a high speed, accurate model which is widely used in the in social media application is called the linguistic inquiry and word count (LIWC), by Pennebaker et al. (2001). LIWC uses a dictionary of 4500 word and sentiment is scored using a set of only 905 words, which are split into 406 positive words and 499 negative words. There are many other categories that cover other topics than sentiment (in total, there are 76 categories). Hutto and Gilbert (2014) mention that no emoticons or slang are listed in the dictionary, which may be an issue in this study given the application to social media.

General Inquirer. This lexicon model was designed by Stone et al. in 1966 for use by social scientists, politicians, and psychologists as a content analysis tool to identify characteristics of written messages. This lexicon was dictionary-based, with a set of 11,000 words, 1,915 of which were labelled positive and 2,291 of which were labelled as negative. The merits of the lexicon, much like others in the orientation-based category, are that valance or sentiment intensity is not resolved, context is not taken into account, and social text such as slang is not included (Hutto and Gilbert, 2014). Hutto and Gilbert (2014) also discuss a more recent addition which was initially constructed by a bootstrapping method, this was created by Hu and Liu (2004) and Pennebaker et al. (2007). Both of these models are word based, where the 'Hu-Liu04' model is a lexicon which maintains 6,800 words (2,006 positive and 4,783 negative words). WordNet, introduced in 1998 by Fellbaum (1998), is another popular English based method that also does not cater for emoticons, slang, or valence.

9.2.4 Valence based sentiment lexicon models

Hutto and Gilbert (2014) note that Affective-Norms-for-English-Words (Anew) was introduced by Bradley and Lang (1999). This was the first in a list of intensity-based lexicons to be introduced. This lexical method is also a list of dictionary words. Whereas the binary-based methods do not have valence, the words in this dictionary are ranked by valence. The valence score is based on a lower score (sentiment) of 1 to the highest sentiment (positive) at 9, where a neutral word is set at 5. This dictionary also ranks in terms of dominance, pleasure, and arousal. Much like other binary methods (such as GI and LIWC), this lexicon does not cater for social text, such as slang and emoticons. SenticWordNet is an extension to WordNet, but Hutto and Gilbert (2014) report that this Lexicon is noisy and fails to score many synsets that do not get mapped. Cambria et al. (2017) explains that SenticNet introduced by in 2012 and was constructed using AI and Semantic Web techniques, including network graphs. Sentiment uses scoring, and the valence is reported as a number between -1 and 1, with zero being neutral. SenticNet also applies common-sense concepts like adoration and wrath to sentiment scores and word mappings, this specialises in concept-level sentiment analysis and maps to sixteen core emotions. SentiWordNet is based on the dictionary from SentiWord and provides a measure of valence, consider the POS mapping in the sentence.

9.2.5 Context aware lexicon models

The challenge for simple lexical techniques, is that a word could be positive, but if the preceding word is negative, this changes the meaning completely, which is **not** a simple fix. Context awareness in lexicons starts with a more significant understanding of a sentence; a way to do this is by understanding the parts of speech, which is a deeper lexical feature in linguistics (Hutto and Gilbert, 2014). Word sense disambiguation provides better context for a single word and how it is used in a sentence. This context-aware lexicon was introduced by (Akkaya et al., 2009). A prominent lexical approach that deals with the issues is an open-source Python package named VADER (Valence Aware Dictionary and sEntiment Reasoner), developed by Hutto and Gilbert (2014). VADER is a hybrid approach that uses a lexical approach based on a unigram of several thousand words that are scored (from -4 to 4) and a second set of guided heuristic rules that examine the sentence containing the word in question and establish the context around the word's meaning or sentiment. The initial lexical score is then scaled and multiplied by a factor that is based on context rules for the work within the sentence. An example is a word that is preceded by a degree modifier such as **not**, **sadly** and other examples include capitalisation and exclamation.

Shapiro et al. (2020) explains that sentiment analysis is a rich field of study. A core part of the modelling is dealing with the meaning of words and this can vary depending on the domain one is working in. For instance, the language used in finance may be very different from that used in rating movies. Words such as 'liability', 'taxing', and 'mean outcomes' could imply different totally different things from a sentiment perspective when comparing finance and common language. Shapiro and Shapiro (2019) argue that context-specific words in finance are important since dictionaries such as the Harvard psycho-sociological dictionary can mis-classify neutral financial terms such as liability, tax, excess, and capital. One would prefer to create models or rules-based dictionaries based on a domain specific collection of words, or corpus. The second critical factor is the language's complexity, which is greater than can be determined by word counts. There are multiple levels of complexity in language, starting with simple negation ('not' in front of 'happy' will convey the opposite sentiment or valence). There are many more examples of how the use of words is structured in any language with hidden and/or subtle meanings that impact the sentiment score.

Hutto and Gilbert (2014) find in their study, which includes comparisons to popular machine-learning techniques, that VADER was the best-performing technique for scoring sentiment on tweets, Amazon reviews, and New York Times articles. The only time the technique was beaten was for movie reviews from rottentomato.com, where Naive Bayes and maximum entropy placed first. Hutto (2014) concludes that VADER is a simple and effective approach with gold-standard NLP characteristics because it is domain-agnostic and requires no setup or calibration.

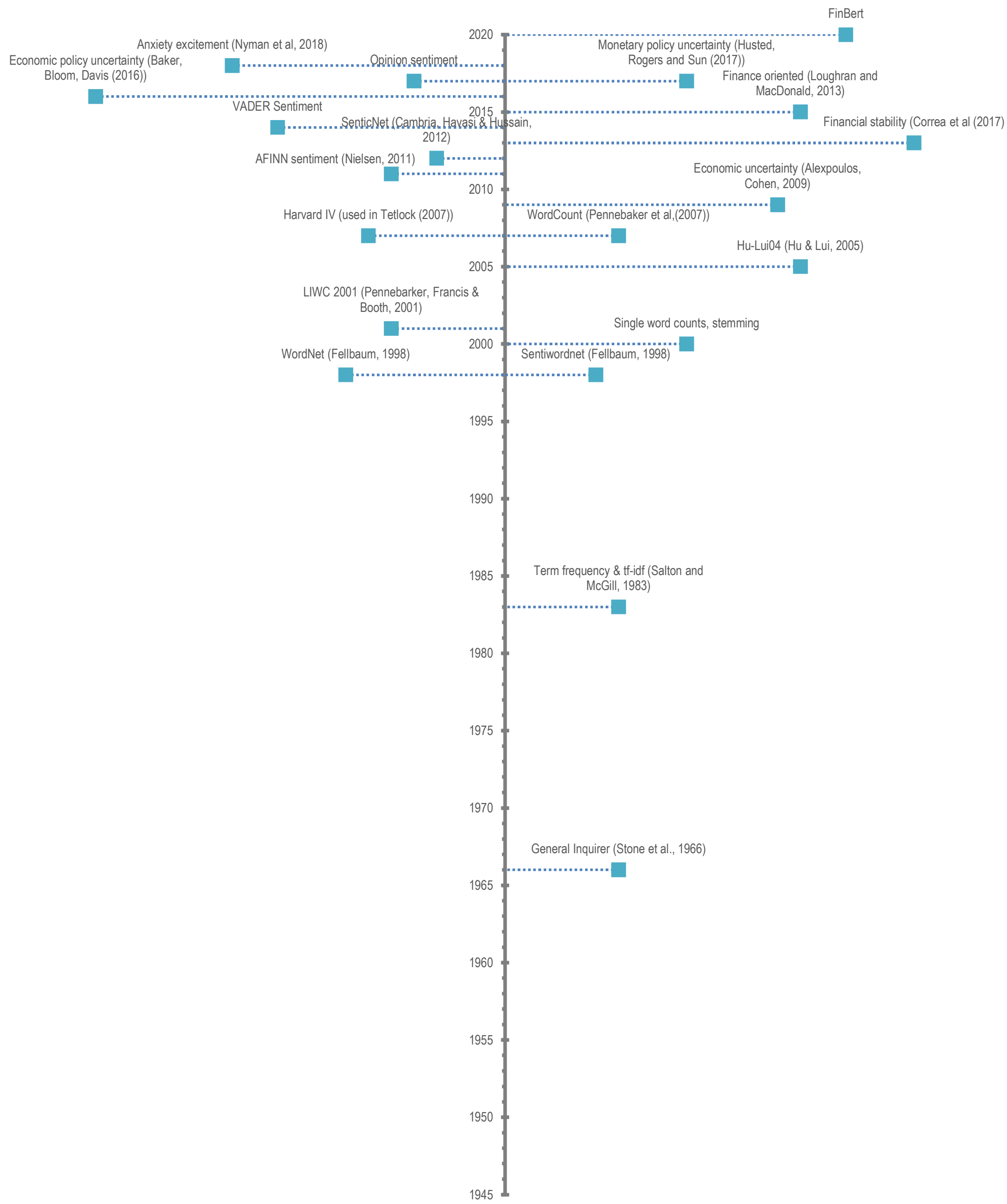


Figure 9: Timeline of the creation of sentiment scoring methods (lexical, rules-based, dictionary, term frequency, and count-based). Two things stand out. First, the first methods were established a long time ago. Second, as time passes, the rate of arrival of new techniques increase dramatically.

9.3 Statistical frequency models

9.3.1 Frequency counts

The simplest of rule-based approaches is frequency counting. This is based on the measurement of the frequency of terms or words in a passage of text. As Rao and McMahan (2019) explain, this is technically accomplished by aggregating terms across the one-hot matrix representation (which is a sparse matrix of each tokenised term along each axis that is set to one down the diagonal). The terms and words can then be mapped using a single term mapping to sentiment values or specific terms that are of interest in the particular study. The word counts are then weighted by article length, which provides the relative magnitude of the measure. A sensible approach would be to pre-process the data, including a stemming and/or lemmatization function that would then naturally group terms, which would then get the appropriate weighting in the aggregation. The frequency–inverse document frequency model is also denoted as td-idf. This is a simple model that shows how important a word is in a paragraph/corpus. This model is essentially the ratio of the frequency of the term by the number of terms in the paragraph of corpus under review. The more a word appears in a corpus of text, the higher the measure will be (Larsen and Thorsrud, 2019). One technique for measuring word similarity that can be used to improve on simple word counting is the comparison of word vector representations with cosine similarity. Goldberg and Hirst (2017) explains that this is efficient if I use the cosine angle (theta) between two vectors, as follows:

$$A \cdot B = |A||B| \times \cos \Theta \tag{73}$$

This function has handy properties for machine-learning in that the output lies between 1 and -1. So 'two-word' vectors will be classed as similar if the above equation is used to minimise the cosine angle between vectors.

9.3.2 Word / term vectors

The word and term vector construct include word vector representations: word vectors are used to translate text into a set of vectors of numbers that represent the meaning of a word. These vectors can be used to measure the level of association with other words or passages. These vectors allow for some powerful applications in NLP, such as understanding the context that a word was used in. Word vectors can be organised in a number of ways, including by frequency or by predictions. Examples of frequency vectors are the count vector, the ubiquitous tf-idf vector, and the co-occurrence vector. The use of word vectors for prediction can encompass methods like skip-gram neural network models. For instance, the well-known algorithm word2vec uses not one but two context models (CBOW and skip-gram) and goal functions (Negative-Sampling and Hierarchical Softmax).

9.4 Machine-learning models

Goldberg and Hirst (2017) explain that ML is especially good at problems that involve scenarios with a set of rules used for inputs that have an easily recognisable, or measurable set of outputs and are straight forward to annotate or label. Natural language processing is one such problem, with an unspecific set of rules that have large amounts of ambiguity and embedded context that can change meanings, hence the results. The words that make up a language are compositional, meaning that the meaning of a sentence or paragraph is more than the sum of its parts. One feature of language is that it evolves with time, but if you picture reading a newspaper in ten years, much of it will be incomprehensible owing to changes in the names of persons (such as the president of a country), buildings, companies, etc. This, as Goldberg and Hirst (2017) explain, leads to data sparseness in the modelling process.

A prominent method for predicting an audience's emotional reaction is a probabilistic model trained on massive amounts of text. These models are constructed from a training dataset based on human-labelled sentiment scores. The analysis improves by including all the context and complexity that the lexical model leaves out. When dealing with domain-specifics, it is crucial to use training data from that domain. The extra benefit of the ML approach is the possibility to use advanced techniques such as linear classifiers and deep learning approaches that can be configured to dynamically re-weight components when new data is added (learning ability). All ML models will be bound by the quality of the training data sets, which may contain bias and inaccuracy. This represents the greatest obstacle. It has been demonstrated that ML models improve the performance of forecasting not just in NLP applications, but also in many other use cases or domain applications. This additional power does not come without a price. All of the strategies involve large amounts of data and requires the creation of a comprehensive training database. This can be time-consuming and costly (CPU, gathering, scrubbing, wrangling, processing, scoring and storing data).

There are many techniques that have been used to apply to the problem of scoring text in unstructured data into sentiment based valence scores, these include:

- **Decision tree classifier**: decision trees are known to be relatively easy to train and are fast, in the simplest form, however this approach can have limited accuracy. Decision trees have been updated by pruning techniques and boosting to improve the accuracy. The tree based classifiers are very useful in cases where the speed of the algorithm is important whilst allowing for slightly less accuracy.
- **Linear and non-linear classifiers** : Simple regression and non-linear classifiers are well-known techniques in statistics, not formalised here, which can be utilised in simple NLP models.

9.4.1 *k*-nearest neighbour (KNN)

Under the umbrella term of machine-learning models, the KNN-algorithm is reasonably straightforward, non-parametric, and one of the simplest strategies (Pang et al., 2002; Mitkov, 2005). This method is used to categorise classes using the nearest neighbour of each data point, with each data point serving as a vote in the classification process. The number of neighbours is determined by k, this input represents the number of nearest training examples in a dataset, must be provided as an input to the model (k is a positive integer, typically small). Patel (2019) explains that this technique belongs to a class of methods known as 'lazy learners' since they do not require particular modelling, calibration (training), and are non-parametric, which is easy to use and fast to implement.

The KNN method is a learning method whereby the unknown data points in the sample are compared to the trained data. The comparison is made with distance as a measure using techniques such as Euclidean distance to Pearson or Spearman coefficients, these are the same measures used in correlation models. The data points can then use the smallest distance to data points in the training set to allocate a category classifier. The weakness in the techniques, however, is that as the k term grows, it becomes computationally burdensome. Each document in a corpus requires distance measurements to every other document, so assuming a 100k dimensional space, the routine would have a billion operations (Mitkov, 2005; Bengfort et al., 2018).

9.4.2 Naive Bayes

Naive Bayes (NB) is also a simple algorithm that is part of a simple set of probabilistic classifiers that rely on simplifying, naive assumptions. In this instance, it is assumed that features such as the lengths of the document are independent of each other (Kalamara et al., 2020). The Naive Bayes method, as described in detail by Bengfort et al. (2018), is predicated on Bayes' probability theorem, which is both straightforward and potent. Using only the previous tag and probabilities learned from the training data, it may anticipate the next tag of a piece of text. This technique is relatively easy in training and implementation and has been successfully used in many classification tasks, such as part-of-speech tagging (i.e., whether it is a noun, adjective, etc.). As it is a probabilistic classifier, a direct output will be to generate probabilities. Following Pang et al. (2002), I define the Naive Bayes in the context of NLP, where I assign a given document, g. I start by first looking to the classic Bayes formula, as:

$$P(h|g) = \frac{P(h)P(g|h)}{P(g)} \tag{74}$$

$P(g) \neq 0$, the numerator is a function of the conditional probability of event h taking place given that g, multiplied by the marginal probability of h. This is the probability of observing g and all terms are divided by the probability of observing g. Intuitively the model can be described as (prior $\times$ likelihood) $\div$ evidence and in the context of text classification, where f is the number of features in document d the number of times f_i appears in d is $n_i(d)$ and a class, $h^* = \arg\max_h P(c|d)$. This is formalised, following Pang et al. (2002), utilising Bayes classical rule:

$$P_{\mathrm{NB}}(c|d) := \frac{P(c)\Big(\prod_{i=1}^{m} P(f_i|c)^{n_i(d)}\Big)}{P(d)} \tag{75}$$

Pang et al. (2002) argues that even with the model assumption of independence that is not so in the real world, the model holds up well, even in cases where dependant features are part of the underlying data.

9.4.3 Maximum entropy

Maximum entropy (ME) is a model belonging to the exponential category of models. It is a more sophisticated model than NB. It is not, however, based on an assumption of independence. As Pang et al. (2002) explains, the model accounts for information entropy, and may allow for a performance improvement by comparison to NB when conditional independence assumptions are not met, as there are no specific assumptions about feature relationships. Following Pang et al. (2002), based on Naive Bayes (NB), I see that $P(h|g)$ takes the following exponential form:

$$P_{\mathrm{ME}}(h|g) := \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,h} F_{i,h}(g,h)\right) \tag{76}$$

We use λ as feature-weight parameters, $Z(d)$ is a normalisation function. For feature f_i and class h, $F_{i,h}$ is defined as:

$$F_{i,h}(g, h') := \begin{cases} 1 & n_i(g) > 0 \text{ and } h' = h \\ 0 & 0 \text{ otherwise} \end{cases} \tag{77}$$

Relationships and assumptions between features are not assumed in this model. This is in contrast to the NB model, which means this model could perform relatively well when assumptions around independence break down. Pang et al. (2002) explain that in this model, assumptions are kept to a minimum while distribution entropy is maximized, resulting inconsistent model data and the model.

9.4.4 Support vector machines

Support vector machines (SVM) is another technique that can be used to model classification and also be used as part of a statistical regression (Pang et al., 2002). This technique is part of the discriminant model. This is more powerful, and it has been shown to outperform NB techniques (Pang et al., 2002). The method is not based on probability, but rather a margin method where the training method uses hyperplanes, finding the optimal hyperplane to separate features (Cambria et al., 2017). The technique also makes use of a relatively simple method, as Aggarwal (2020) explains: for a model scoring two categories, each data point is measured on a hyperplane vector, and hyperplanes are used to represent classes; the algorithm will then seek to find a solution to the largest margin between hyperplanes that separates the classes.

The SVM technique ultimately determines hyperplanes that help to distinguish between points in the next class, it classifies points into two categories. To classify a data point is as simple as establishing on which side of the hyperplane it lies (in the two-class case). Yu and Nwet (2020) explain that the SVMs require little training data to calibrate and require less memory to build than the KNN. However, SVM training on large data sets takes more time than other methods, making it less viable in some cases. Aggarwal (2020) explains that a basic SVM model is essentially a constrained optimisation problem, referred to as a cutting-plane method, and there are many formulations that include the use of kernels, using regularised optimisation, logistic regression using an L_1 or L_2 regularised logistic regression, and LASSO[7] variable selection.

[7] least absolute shrinkage and selection operator (LASSO).

An example of the classical SVM method is formulated below, where $g_j \in \{-1, 1\}$ is the correct class of document (which in this case is negative or positive) for the target variable. $\bar{X}$ is a row vector containing, in this instance, the rows of document data and $\bar{W}$ is a column vector resultant weights, following Aggarwal (2020) the L_2-loss SVM formulation is:

$$J = \frac{1}{2}\sum_{n}^{i=1} \max\left\{0, (1 - g_i[\bar{W} \cdot \bar{X}_i^T])\right\}^2 + \frac{\lambda}{2}||\bar{W}||^2 \tag{78}$$

J is the objective function. For interest, I formulate another popular formulation, following Aggarwal (2020) the Hinge-loss SVM:

$$J = \sum_{n}^{i=1} \max\left\{0, (1 - g_i[\bar{W} \cdot \bar{X}_i^T])\right\} + \frac{\lambda}{2}||\bar{W}||^2 \tag{79}$$

As one can see, these methods are similar, but when compared to the least squares method, it is known that well-defined, or separated data, is not penalised in the loss function, which is not the case for least squares regression (Aggarwal, 2020).

9.4.5 Generative models

One problem with discriminant models is that they cannot be used in the absence of direct observation to characterise complicated interactions between dependent and independent variables, for example SVMs (Cambria et al., 2017). In the context of opinion and sentiment modelling, generative models' ability to reveal hidden relationships between variables is an asset.

The N-gram models are the most basic generative models, and they have been widely used in phrase similarity mapping, speech recognition, and text query similarity for documents. Language probability models are specified with distributions over sequences of text. Cambria et al. (2017) formalise a model of a given set of words $(w_1, \ldots, w_{k-1})$ as a $P(w_k|w_1, \ldots, w_{k-1})$ multinomial distribution over words as follows:

$$\begin{aligned} P\Big(w_1, w_2, \ldots, w_n\Big) &= P\Big(w_1\Big)P\Big(w_2|w_1\Big)P\Big(w_3|w_2 w_1\Big)\ldots P\Big(w_n|w_1 w_2, \ldots, w_{n-1}\Big) \\ &= \prod_{k=1}^{n} P\Big(w_k|w_1, \ldots, w_{k-1}\Big) \end{aligned}$$

This is built using the chain rule of two or more random variables. The chain rule connects the joint probability of a sequence of words to their conditional probability given the preceding words in the sequence (Cambria et al., 2017). The reader may see that the model is burdened because the probability of a word is then conditional on whatever sequence of words that was previous seen in the sequence, which get large, quickly. On a practical note the volume of text required to inform the probability matrix gets large very quickly, it grows exponentially. I therefore start to look at simplifying assumptions.

N-gram models. N-gram models are the first and possibly simplest solution to this, the probability is approximated by only using a section of the prior words in the sequence and not all the words. When the set of words compared to the previous word only, this is then a Markov model (Cambria et al., 2017). On the other-hand, if I assume that the current word is totally independent of prior words, then I have a 'unigram' model of 'bag-of-words' model, denoted as:

$$P\Big(w_1, w_2, \dots, w_n\Big) = \prod_{k=1}^{n=1} P\Big(w_k\Big) \tag{80}$$

In this model, as I will review of later, the order of the words is also not important. The n-gram models (bigram, trigram) are used to capture dependencies between words.

Latent Semantic Indexing. Latent Semantic Indexing (LSI) in this scenario is a document is modelled as a collection of latent topics. Generative models are used to reveal these hidden patterns in the data. Each concept in probabilistic Latent Semantic Indexing (pLSI) is represented as a multinomial distribution over the words, which allows for more accurate modelling of the data. Cambria et al. (2017) explain that the generative process in the pLSI is not complete as it does not deal with the document-level generative process. Recall the simplifying assumption that topics z for words in a document g are independent of the document index. So the joint probability of a document d, and the underlying words w, and respective topic assignments (Blei et al., 2003).

$$P\Big(g, w_1, w_2, \dots, w_N\Big) = P(g) \prod_{i=1}^{N} \sum_{z_i} P\Big(w_i|z_i\Big) P\Big(z_i|g\Big) \tag{81}$$

w_i are words, indexed from $i \dots N$, of sequence of word in a document, M of length N, that belongs to a corpus, G, of documents. The generative process will resolve for the latent topics k, each of which is modelled using the multinomial-distribution.

Latent Dirichlet allocation (LDA). The LDA is a model that reduces a large volume of text into a smaller set of words to best describe the document as a set of representative topics. Larsen and Thorsrud (2019) explain that LDA models are a form of unsupervised learning model, very useful at separating words into clusters of topics. The term 'Dirichlet' is from the use of a conjugate Dirichlet prior used to draw the topics. This method is very useful in providing labels to where apriori categorisation into topics is needed (Bengfort et al., 2018). Tran-The (2020) explains that for the purpose of generating textual topics and finding recurrent themes throughout a body of documents, LDA is a popular generative probabilistic model. In each document, the proportion of topics, $p(z|\theta, g)$ is modelled as a multi-dimensional vector, θ, is drawn from a Dirichlet distribution with a concentration parameter, α. The gamma function is denoted as Γ. The Dirichlet function as denoted here, where I follow Blei et al. (2003).

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_i^{\alpha_i - 1} \cdots \theta_k^{\alpha_k - 1} \tag{82}$$

To formalise the model here, I again follow Blei et al. (2003), where they initially define the model with further, simplifying, assumptions. Second, the probabilities are estimated and set a fixed quantity to words as parameter matrix B $(k \times V)$, in which $B_{(i,j)} = p(w^j = 1|z^i = 1)$. With the given parameters for α and β, the sets of words, $\mathbf{w}$, the set of $\mathbf{z}$ topics, of a given joint topic mixture, θ, and number of words/topics, N, is given by (Blei et al., 2003).

$$p\Big(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta\Big) = p\Big(\theta|\alpha\Big) \prod_{n=1}^{N} p\Big(z_n|\theta\Big) p\Big(w_n|z_n, \beta\Big) \tag{83}$$

The modelling can be fine-tuned using the hyper-parameter which is the number of topics. Blei et al. (2003) explain that in the context of risk disclosures, a low number of topics will produce topics which are more broad, showing key risks reported by most firms such as 'currency risk' or 'supplier risk'. Blei et al. (2003) explains that this solution will not be stable one once number of topics start to get larger (say >50 topics).

9.4.6 Deep learning and neural networks

Introduction to neural networks. Neural networks (NN) have proven to be flexible enough and capable to dealing with the challenges inherent in NLP problems (Goldberg and Hirst, 2017). A key functional component of neural network architecture is an embedding process which provides mapping of sparse discrete symbols to continuous vectors, thus enabling mathematical functions and processes to be operated on. Goldberg and Hirst (2017) explain that a sub-category of machine-learning including the Deep Learning techniques and the original architectures were called NN but have been recently been resurfaced as Deep Learning, or Deep NN. The 'learning' term stems from the mathematical parametrization of layers of differentiable functions, which, in the case of deep learning are due to the layers of differentiable function that are chained together (Goldberg and Hirst, 2017). Goldberg and Hirst (2017) explain that although machine-learning techniques are all trained, or learning based, on past data and larger volumes of the data is key to making more accurate predictions; deep-learning techniques, however these models once they have converged it correctly represents the data.

The resurgence of the techniques is largely linked to the improvement and cost in technology as the techniques have been in existence for a long time. For example, the first machine-learning was introduced by Rosenblatt in 1957. This process has also been likened to the processes in brain cells in Rosenblatt's paper this analogy the brain cell (neuron) as used to describe the signal processing for image recognition. The analogy that is often used is a neuron cell, with dendrites (signal interpreters), sending firing electric signals via the axon. If we extend that analogy, not all dendrites are created equal, they can be more sensitive and change dynamically over-time. Lane et al. (2019), amusingly, end the analogy as a 'lot of hand waving about biology and electric current', goes onto to introduce the more important parts of the concept. I will follow Goldberg and Hirst (2017), where mathematical descriptions are used to formalise these models.

So, as in other fields, data is used to identify independent variables (regressor in statistics, factors in finance or features machine-learning) to predict a dependant variable (target variables, or regress and statistics) in the model (or algorithm) is fed into a model. In the case of NN however, this only the first layer, which provides further information into another layer of models and this process improves the predictions. NN are often termed 'black boxes' due to hidden layers are not often reported as they form part of a larger algorithm. Mathematically, I would like to express a neural network problem by starting with a simple linear problem adding in the 'layers'. So, following Goldberg and Hirst (2017), I denote a simple linear model, which is the simplest of neural network models:

$$\mathrm{NN}_{\mathrm{LM}}(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{W} + \boldsymbol{b} \tag{84}$$

where $\boldsymbol{W}$ is the parameter matrix and $\boldsymbol{b}$ is a term used to track the bias. This is a single layer perceptron, where NN more than one layer, they are multi-layered; thus the following is a multi-layered (two layered) perceptron:

$$\mathrm{NN}_{\mathrm{MLP1}}(\boldsymbol{x}) = g(\boldsymbol{x}\boldsymbol{W}^1 + \boldsymbol{b}^1)\boldsymbol{W}^2 + \boldsymbol{b}^2 \tag{85}$$

The activation function g is non-linear, which is an important feature allowing for full flexibility and non-linearity models. This can then be further layered by adding another layer and activation function:

$$\mathrm{NN}_{\mathrm{MLP2}}(\boldsymbol{x}) = g^2(g^1(\boldsymbol{x}\boldsymbol{W}^1 + \boldsymbol{b}^1)\boldsymbol{W}^2 + \boldsymbol{b}^2)\boldsymbol{W}^3 \tag{86}$$

This equation is possibly clearer when decomposed into separate functions:

$$\begin{aligned}
\mathrm{NN}_{\mathrm{MLP2}}(\boldsymbol{x}) &= \boldsymbol{y} \\
\boldsymbol{h}^1 &= g^1(\boldsymbol{x}^1\boldsymbol{W}^1 + \boldsymbol{b}^1) \\
\boldsymbol{h}^2 &= g^2(\boldsymbol{h}^1\boldsymbol{W}^2 + \boldsymbol{b}^2) \\
\boldsymbol{y} &= \boldsymbol{h}^2\boldsymbol{W}^3
\end{aligned} \tag{87}$$

From what I can now see from the final representation, the layers are evident. Neural nets can have multiple layers these are known as deep learning NN. The first two layers are hidden; the last layer produces the result and is known as the output layer. This is a flexible framework and the layers can be achieved in different ways to achieve objectives. The different constructs are discussed further along in the study and this includes such as convolutional or pooling layers. The layers of matrices and bias terms are collectively referred to as parameters and denoted here as Θ.

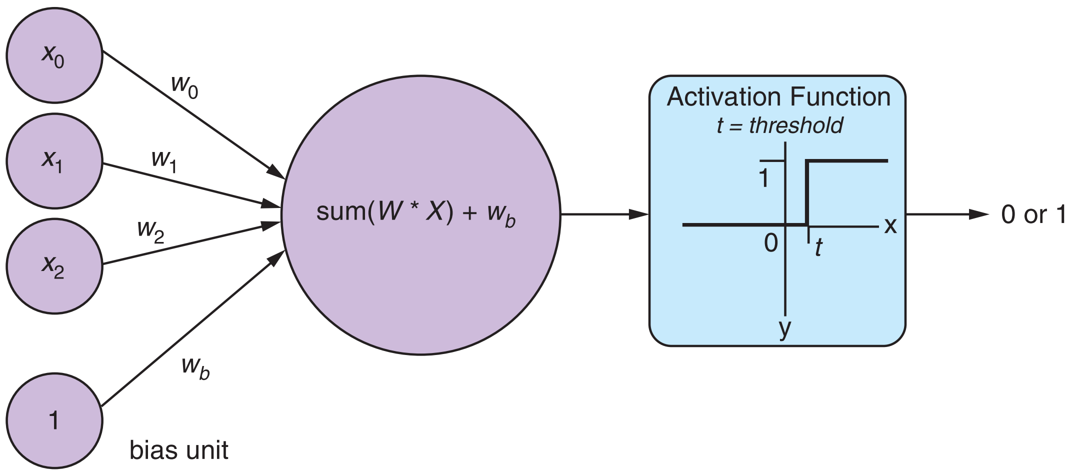


Figure 10: Source: Lane et al. (2019), pg. 158. Basic perceptron, where X is the collection of each feature, x_i. Associated weights, w_i, for each feature. The final block, the activation function, which is introduced in more detail below.

The activation function, which in the case of the a neuron will be one (fire a signal) or zero (not fire a signal), based on threshold t, for all instances in the analysis i:

$$f(w) = 1 \text{ if } \sum_{i=0}^{n} x_i w_i > t \text{ else } 0 \tag{88}$$

A key innovation in neural networks is backpropagation, which allows for the combined learning of multiple models (neurons) simultaneously, allowing updating learnings to be 'cross-pollinated'. This process sets layers where the first neuron feeds the next layer of neurons with information by updating parameter weights, w_i, and this contributes to the error. This architecture or model design, sets up an intermediate layer where the information is fed into, it is not directly observed in the model features. This is where the term 'hidden-layer' stems from neural networks (Lane et al., 2019).

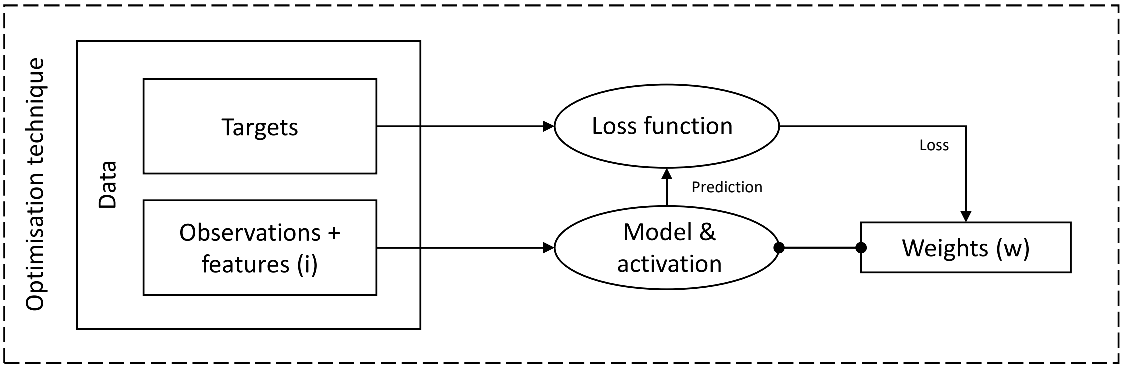


Figure 11: The architecture or model schematic describes the basic components of a neural network model. This includes the underlying data targets and features, the model and activation function, to the loss function and resulting weights (parameters). This construct provides a good concept flow for the descriptions below.

This is a good in time to introduce the loss function and explain the overall objectives of the neural network modelling framework. Lane et al. (2019) posit that the error, the difference between a predicted and actual value for each point in the analysis, $e_{x_i} = |Y - f(x_i)|$, the neural network model objective is to minimise this error, by way of a minimisation of the 'loss function', $J(x) = \min \sum_{i=1}^{n}(e_{x_i})$. Once this objective is solved the user can consider that this model will be the best representation of the data. The econometrics field will recognise the model as functionality similar to an autoregressive model and potentially a vector error correction function. A key process in the neural network design the introduction an appropriate loss function (and optimisation techniques).

Sigmoid function: Until now the activation function has been set up so as to return a binary result ([0,]) using a sigmoid function, $S(x) = \frac{1}{1+\exp^x}$. This is really just the inverse of the well-known logistic function and this approach is also used prevalent in credit modelling. Neural network activation function requirements are that the functions is differentiable so that key mathematical techniques such as the chain rule can be applied.

Hyperbolic tangent function: Other common non-linear activation strategies include the hyperbolic tangent function (tanh), which is cosmetically similar, with a bit of transformation is mathematically similar and the output is also S-shaped, with values strictly between -1 and 1 (Goldberg and Hirst, 2017). The tanh function is defined as, following Goldberg and Hirst (2017): $\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}$.

Rectified linear units (ReLu): Another, very simple, activation function is the ReLu activation function, which floors each unit to zero below a certain level, $x < 0$ then 0, else x, otherwise expressed as (following Goldberg and Hirst (2017)): $f(x) = \max(0, x)$. This function has provided excellent results and surprisingly has been introduced more recently than other functions (Rao and McMahan, 2019). An issue can arise where the activation function does not provide any information if it stays below zero. This is elevated with the use of the Leaky ReLu and Parametric ReLu. Leaky ReLu is a variant of the ReLu that has a small constant gradient, β, for values below zero which will solve for the hard floor set at zero for RELU. Rao and McMahan (2019) details the LRelu / PReLu activation function: $f(x) = \max(x, \beta x)$.

Softmax: This function is useful when working with categorical data, like the sigmoid function it provides an output range of 0 to 1, the values of predicted values will sum to one. Goldberg and Hirst (2017) explains that this make it interpretable as a discrete probability distribution of the observations of the event over 'i' different events, as follows (following Rao and McMahan (2019)):

$$\text{SoftMax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}} \tag{89}$$

Exponential Linear Unit (ELU): The ELU function is known to converge quickly and provides high accuracy. This function also caters for negative results, whereas the ReLu is not a smooth function and transitions to zero abruptly, this function slowly moves towards the lowest value and denoted as follows:

$$f(x) = \max\Big(x, \alpha(e^x - 1)\Big) \tag{90}$$

Although this is not an exhaustive list, I can see that there a numerous types of function and careful consideration is needed when selecting the correct activation function. Loss functions are an important component in the neural network modelling techniques. This logic is consistent in that the predicted value and the target are compared and the measurement of these differences are known as loss functions, which are also used in evaluating pure statistics models. Thus I will only list an briefly introduce common loss functions here:

Hinge: The hinge function is used for binary classification models, such as Logistic regression. The function uses the following formula that is sign specific, as follows:

$$L_{hinge}(y, \hat{y}) = \max(0, 1 - y \cdot \hat{y}) \tag{91}$$

As Goldberg and Hirst (2017) points out, the loss is zero if the predicted value, $\hat{y}$, and target value, y are the same sign. Larger values signify worse prediction in the model, where are fair model will requires a measure less than 1.

Mean squared error (MSE): MSE is broadly used in regression problems for continuous dependant (target) and independent (feature) variables. It is measured as the mean of the difference between predicted and targeted, summed and squared.

$$\mathrm{MSE}_{(y_i, \hat{y}_i)} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \tag{92}$$

Mean absolute error (MAE): MAE is similar to the MSE above, where the differences are measured in absolute and then added.

$$\mathrm{MAE}_{(y_i, \hat{y}_i)} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{93}$$

Cross-Entropy: Cross-Entropy is a another common loss function that has two variants, binary and categorical cross entropy loss functions (Rao and McMahan, 2019). In the case of binary cross entropy, also known as logistic loss function, it is assumed the activation function is a sigmoid function. The binary classification is set up as if $\hat{y} < 0.5$ then 0, and when $\hat{y} > 0.5$ this binary indicator is 1. The binary cross entropy model is defined as, following Goldberg and Hirst (2017):

$$L_{\mathrm{LCE}}(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}) \tag{94}$$

The categorical cross-entropy loss function is also known as the negative log-likelihood and is assumed to have an underlying activation function set as a Softmax function. This is used to assess models of multi-class categorisation (Rao and McMahan, 2019). Following Goldberg and Hirst (2017), the categorical cross-entropy measures the dissimilarity between predicted and target, it is defined as:

$$L_{\mathrm{BCE}}(y, \hat{y}) = -\sum_{i} y_i \log(\hat{y}_i) \tag{95}$$

As is that case on the activation functions, this study is not a comprehensive review of loss functions, but rather to give the reader an idea of the types of function and provide an overview.

Backpropagation: This is an important component that applies to NN, and this technique is also known as backward error propagation and has become a standard in training neural network models and is based on the chain rule. This process has been adopted by many domains as it is easy to program, fast and simply based on the mathematics chain rule that is used to decompose the loss by using the chain rule, allowing the direct connection between change in loss and partial derivative with respect to weight. The initial weights are random and serve only to initiate the first error calculations as a start to the optimisation, which is performed recursively. The error term is denoted as e here. Backpropagation is the process of systematically calculating the derivatives of a complex function used in a neural network; the algorithm is an extension to the gradient descent models which updates weights with the aim finding a local minimum of the error function. Backpropagation is a supervised learning method that calculates the loss function gradient by requiring a known output for each input value. The weights are generated in the output layer first from the initial feed forward model and then used to update the weights of the other layers, starting with final layer and proceeding backwards until the first layer is updated. Rojas (1996) explains that the backpropagation step provides an implementation of the chain rule. The extended optimisation function uses weights from the primary optimisation as a set of inputs against the result of the loss function. Rojas (1996) emphasises that it is a continuous differentiable function used to minimise e in a gradient descent model, of n weights, $w_1 \cdots w_n$, in the neural network model.

$$\nabla e = \left(\frac{\delta e}{\delta w_1}, \frac{\delta e}{\delta w_2}, \cdots, \frac{\delta e}{\delta w_n} \right) \tag{96}$$

This results in a function to update the weights of the network by the following increment, where γ is the learning rate governing the length change in increment, following Rojas (1996):

$$\nabla e = -\gamma \frac{\delta e}{\delta w_i} \text{ for } i = 1, \cdots, \ell \tag{97}$$

This incremental change in weights seeks to reduce the loss or error value, e, minimising as much as possible.

Regularisation: Regularisation is a way that a learning algorithm is amended so as to not impact training error whilst reducing generalisation error. The process of finding a solution to this problem is as crucial as optimisation. There are many solutions these can be either implicitly or explicitly determined. Regularization is a key consideration in AI models such as neural nets, coupled with optimisation. These components are vital, they are the reason the methods are so powerful and they represent the data accurately. The regularisation is a way of controlling the complexity parameter whilst still controlling for over fitting.

Optimisation techniques: At this point, the problem has not been solved which is in the model training and this also provides weights, that are 'learned' (Goldberg and Hirst, 2017). Problems that cannot be solved using closed form (which is found in most real world problems) are achieved via an optimisation program. This full solution, including the data, parameters, model, activation function, loss function and regulatorisation function, requires an optimisation to solve for the weights, or using machine-learning terminology to 'train' the model. Optimisations are set up to solve for the weights by minimising the error of prediction. Optimisation is a very deep field and central to the operations research field. I review a few of the more prominent methods here, by introducing the concepts and detail around optimisation techniques for neural nets. I start with a simple linear model, building it up to the a more advanced gradient based methods (this is only a summary view, for detail or complete review, please refer Aggarwal (2020)).

Linear model (closed form, normal, linear regression): Many applications of machine-learning involve optimisation of linear systems, partly due of the ease with which one can solve the problem (in closed form). It is also due to number of problems that can be handled by a linear system, or a combination of linear systems, which is explained by Cornuejols and Tutuncu (2007) for a use-case of stochastic programming and simple linear problems. As Aggarwal (2020) points out, solving linear programs is also known as linear regression and is a fundamental machine-learning problem. This problem is a solution for a set of linear equations, expressed as $Ax = b$, as it relates to a linear model of $y = Ax$. One way, a well used approach, will be to minimise the following objective function $J = ||b–\boldsymbol{Ax}||^2$, where $r = ||b–\boldsymbol{A}x||^2 = ||\sum_{j=1}^{n} a_j x_j - b||^2$ is the measure of the residuals, setting r to zero solves the linear equation. This is the classical least squares approach which seeks to fit a function in which the differences between the predicted value and observed data (measured as the squared difference) is minimised. Aggarwal (2020) notes that the equivalent approach would be to use a geometric approach and calculus. The assumption is that A is invertible and the observation are linearly independent. This is a task which relies on the closest point from a hyperplane to a specific point is always the orthogonal of the hyperplane itself, I detail here following Aggarwal (2020):

$$
\begin{aligned}
\text{minimise } & ||\boldsymbol{A}\bar{x} - \boldsymbol{b}||^2 \\
\text{by setting } & \boldsymbol{r} = \boldsymbol{A}^T(\bar{x} - \boldsymbol{A}) \text{ to } 0 \\
& \boldsymbol{A}^T\boldsymbol{A}\bar{x} = \boldsymbol{A}^T\boldsymbol{b} \\
& x = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}
\end{aligned}
$$

Equivalently one could solve x is orthogonal to span($\boldsymbol{A}$). This is a good example of the optimal solution being found in the closed form (Aggarwal, 2020). This is a fortunate result used in linear regression problems where the observations are linear and independent. Aggarwal (2020) explains that unfortunately this is not always the case and this would require one to consider different approaches such as the gradient descent approach.

Gradient descent methods: Gradient descent models are important in machine-learning. These methods are put to use when the function cannot be solved by means of a closed-form solution, such as the linear regression using the normal equations. Aggarwal (2020) delineates that, when the gradient descent method is used, a simple inversion is replaced by a computational algorithm. Algorithms such as this are initiated with a random number, $\bar{S}$ and incremental changes are parsed. If the objective function is lower than before (with respect to the negative derivative of the objective function, bringing it closer to zero), then it is then repeated with the objective of finding a global minimum value. This is shown below:

$$\text{Linear normal : } \left[\frac{\partial J}{\partial w_1} \dots \frac{\partial J}{\partial w_d}\right] = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$$
$$\text{Gradient descent :}[w_1 \dots w_d]^T \Leftarrow [w_1 \dots w_d]^T - \alpha \left[\frac{\partial J}{\partial w_1} \dots \frac{\partial J}{\partial w_d}\right]^T = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$$
$$\text{Gradient vector :}\nabla J(\mathbf{A}) = [w_1 \dots w_d]^T \Leftarrow [w_1 \dots w_d]^T - \alpha \left[\frac{\partial J}{\partial w_1} \dots \frac{\partial J}{\partial w_d}\right]^T = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b} \tag{98}$$

One can see the difference between the first equation (linear normal) and the second, which is the equivalent statement for gradient descent function, that the solution approaches the optimal w with incremental changes α. Aggarwal (2020) outlines that one can therefore simply write a gradient descent model as follows:

$$A \Leftarrow A - \alpha\nabla J(A) \tag{99}$$

The size of the incremental change is referred to as the 'learning rate' and determined by the size of α. The gradient descent process will continue until the measured benefit from each increment is no longer material enough for it to be useful, the solution is said to 'converge'. It is at this point that the solution is approximately equal to the optimal solution.

Stochastic gradient descent (SGD): This method is also an iterative approach to minimise the objective function with respect to the model parameters. This is achieved by moving in the opposite direction from the objective function gradient. The SGD function, unlike the vanilla gradient descent method, does not pull all the data into each iteration whilst it is learning. At each iteration a new parameter set is then updated. The sample is based on the batch of similar gradient instance in the data and is reshuffled with each iteration, hence the term 'stochastic'. Relatively, the descent path is not smooth, it appears noisy, but the reduction in redundancy (due to targetted samples) enhances the speed of convergence. The speedier convergence means it is often referred as the 'Online' SGD (Goldberg and Hirst, 2017). Ruder et al. (2017) point out that due to the change in parameter at each iteration, the later change in path reduces the chance of getting caught in a local minimum, which is a big benefit. The stochastic gradient descent model will then perform a parameter update for each iteration in training, x^i, for each label, y^i, denoted as follows (Ruder et al., 2017):

$$A \Leftarrow A - \alpha\nabla J(A; x^{(i)}; y^{(i)}) \tag{100}$$

The algorithm of a simple SGD (Online) per Goldberg and Hirst (2017) is set out in Table 3.

Online SGD algorithm
Input:
◦ Function of $f(x, \Theta)$ parameterised with parameters Θ
◦ Training set of inputs $x_1 \dots x_n$ and expected outputs $y_1 \dots y_n$
◦ Loss function l
While *stopping criteria are not met* **do**:
Sample training example x_i, y_i
Compute Loss $l(f(x_i, \Theta), y_i)$
$\hat{g} \leftarrow$ gradients of $l(f(x_i; \Theta), y_i)$ w.r.t. Θ
$\Theta \leftarrow \Theta - n_t \hat{g}$
return Θ

Table 3: Steps and flow of the Online stochastic gradient descent algorithm.

Mini-batch SGD: Mini-batch SGD is an enhancement of the SGD method. The loss model noise can be reduced by using a sample of the gradients examples when calculating the error estimates (Goldberg and Hirst, 2017). This method shows a general improvement in the accuracy and also allows for an improvement in computational efficiency. The mini-batch SGD will then perform a parameter update for each iteration in training, x^i, for each label. For each mini-batch of n training samples, I follow Ruder et al. (2017), denoted as:

$$A \Leftarrow A - \alpha \nabla J(A; x^{(i:i+n)}; y^{(i:i+n)}) \tag{101}$$

It may be interesting to contrast the mini-batch SGD algorithm, within the algorithm below, following Goldberg and Hirst (2017). One can see that aside from the subset of information used, this algorithm performs the same steps.
Batching in models is designed for greater efficiency in the optimisation. Batching is to do with the way data is fed at each iteration of the loop. The first is the batch learning technique. This method is very slow as it minimises the error cost function (the loss function) over the whole of dataset the data in one batch. The next, a more efficient method, is to break up the data set into the small sub-sets minimising these mini-batches; hence the name

Mini-batch SGD algorithm
Input:
○ Function of $f(x, \Theta)$ parametrised with parameters Θ
○ Training set of inputs $x_1 \ldots x_n$ and expected outputs $y_1 \ldots y_n$
○ Loss function l
While *stopping criteria are not met* **do***:*
Mini-batch training sample m examples $(x_1, y_1), \ldots (x_m, y_m)$
$\hat{g} \leftarrow 0$
for $i = 1$ to m do:
Compute Loss $l(f(x_i, \Theta), y_i)$
$\hat{g} \leftarrow \hat{g}$ + gradients of $\frac{1}{2}L(f(x_i; \Theta), y_i)$ w.r.t. Θ
$\Theta \leftarrow \Theta - n_t\hat{g}$
return Θ

Table 4: Steps and flow of the Mini-batch stochastic gradient descent algorithm.

mini-batch gradient descent. When the batch size equals one, this is then referred to as full online-learning in which weights are updated for every data input. This is also known as stochastic gradient descent. If one considers that this technique has to do with the data feed and not the optimiser, this can be applied to any optimisation technique. The three modes are full-online, mini-batch and full-batch.

Momentum based gradient descent algorithms: This technique is inspired by the 'local minimum' or 'hill climbing' technique introduced by Currey (1944). These techniques are based on gradient functions that have further enhancements which again deal with speed of convergence (as a general goal), being trapped in suboptimal local minima, choosing the correct learning rates especially when it comes to sparse data (NLP is not an exception to this) and lastly the learning rate scheduling (such as annealing) (Ruder et al., 2017). The ravine type surface (such as an elliptical surface) is an issue for an SGD algorithm. The descent path makes little forward progress relative to the zigzag movements in a local minima. The momentum enhancement places more reliance on the longer term direction of consistent movement. The addition of a momentum term, γ, provides impetus and counterbalance the oscillations. The γ-term is multiplied by the previous step vector v_t and the term itself is a factor, which is often set in the region of 0.9 (Ruder et al., 2017). This is denoted as:

$$v_t = \gamma v_{t-1} + \alpha \nabla J(A)$$
$$A = A - v_t$$

As Aggarwal (2020) explains, this process imports the previous velocity from the iteration, providing the necessary direction and momentum to speed up the convergence and avoid getting slowed or stuck in a flat region or local minima. A drawback is that the algorithm can overshoot the minima, but with the correct γ term, this will not be an issue.

Nesterov Accelerated Gradient Descent (NAG): The NAG is a slightly smarter algorithm that will have information on when to slow down. This algorithm provides a forward view by enhancing the gradient based on the future expected parameters (restated by the momentum term):

$$\begin{aligned} v_t &= \gamma v_{t-1} + \alpha \nabla J\Big(A - \gamma v_{t-1}\Big) \\ A &= A - v_t \end{aligned} \tag{102}$$

This enhancement will ameliorate for an overshoot, not completely but it will make an improvement (Ruder et al., 2017).

Adagrad and Adadelta: Aggarwal (2020) explains that the Adagrad algorithm is adaptive in that is keeps track of the size of the frequency of parameters, providing for larger updates of infrequent parameters. This is an effective enhancement when working with sparse matrices, which is, as Aggarwal (2020) explains, why it was used to train GloVe word embeddings for the NLP. At each iteration an update is made with respect to the objective function, J, which includes the impact of the squared magnitude:

$$A_i \Leftarrow A_i + \left(\frac{\partial J}{\partial w_i}\right)^2, \tag{103}$$

each parameter is updated as follows:

$$w_i \Leftarrow w_i - \frac{\alpha}{\sqrt{A_i}} \left(\frac{\partial J}{\partial w_i}\right)^2 \tag{104}$$

The update to the learning rate, α, at processed each time step whilst the algorithm aggregates the sum of past gradients with respect to the parameters (Ruder et al., 2017). The benefit of this algorithm is that it does not require manual tuning of the learning rate, γ

parameter. It is an accumulation in the denominator, which has the effect of monotonically reducing the learning rate. This challenge of an ever decreasing learning rate is resolved by the Adadelta and RMSprop algorithms (Aggarwal, 2020; Ruder et al., 2017). The Adadelta method will simply aggregate a sample or window instead of aggregating the sum of past gradients with respect to the parameters.

RMSprop: As Aggarwal (2020) explains, this algorithm has the same aim as the AdaGrad, with the exception of the sum of past gradients. Such is not simply aggregated but uses an exponential decaying average of past returns. This is a slightly different from the Adagrad formation, shown in Equation 103, denoted as follows:

$$A_i \Leftarrow \rho A_i + (1-\rho)\left(\frac{\partial J}{\partial w_i}\right)^2 \tag{105}$$

where ρ is the running squared aggregate. The advantage here is that the older gradients decay over-time, where the oldest of them are not material.

Adam. Adam is a very popular algorithm as it combines the benefits of both worlds, in which the RMSprop algorithm has the momentum component added (Ruder et al., 2017; Aggarwal, 2020). It is an adaptive learning process in which rates are paramterised for each parameter. There are two key differences between RMSprop and Adam, the gradient is replaced by an exponentially smoothed decay parameter, as when following Aggarwal (2020):

$$A_i \Leftarrow \rho_f F_i A_i + (1-\rho_f)\left(\frac{\partial J}{\partial w_i}\right)^2 \tag{106}$$

The learning rate is then updated for each iteration of t:

$$w_i \Leftarrow w_i - \frac{\alpha}{\sqrt{A_i}} F_i \tag{107}$$

The last difference is that the learning rate is now governed using exponential smoothing and incorporating the momentum:

$$\alpha_t = \alpha\left(\frac{\sqrt{1-\rho^t}}{1-\rho_f^t}\right) \tag{108}$$

Ruder et al. (2017) and Aggarwal (2020) note that the Adam algorithm is a successful implementation of the adaptive models such as RMSprop. The Adam algorithm has the benefit of momentum enhancements. The final enhancement stems from the Adam incorporating the NAG model momentum enhancement, this makes it superior to the vanilla momentum enhancements.

Important optimisation techniques for NN can also include Newton methods (e.g. Newton-Raphson Method), dynamic programming (e.g. Bellman equations), stochastic methods (e.g. stochastic linear programming), evolutionary algorithms and genetic algorithms. Please refer to Kiranyaz (2013); Rojas (1996), Sra et al. (2011), Aggarwal (2020) for further reading and detail on methods applied to building neural network frameworks. A more general review of optimisation techniques, to name a few, can be found in Kwon (2013), Cornuejols and Tutuncu (2007) and Rao (2009).

Convolutional neural network (CNN): Goldberg and Hirst (2017) explain that CNN's are named as such due to the functional sliding over data or 'convolving' over a window of data within the data under investigation (Goldberg and Hirst, 2017). Can be used for NLP tasks and in a simple CNN model words are processed and compiled into a word embedding matrix, using a non-linear learned function to extract important information using sliding windows over the text. Bengfort et al. (2018) explain that CNN's feature architecture which builds up a feature map by combining a convolutional layer and multilayer perceptrons; it also features a pooling stage to collect and store the most informative features. CNN have their beginnings in signal processing and image recognition, which explains the terminology aligned to the computer vision and signal processing. The CNN architecture is really effective when put to use in classification tasks and image data modelling. The convolutional layers with backward reference to words in a sequence are a strength showing promising results in text classification, sentiment analysis, POS tagging, question answering and more (Goldberg and Hirst, 2017). Lopez and Kalita (2017) explain that CNN's are really quite similar to ordinary NN in that they both have weights generated from a learning process and biases. The main difference is that NN's have each input neuron is connected to each output neuron, whereas CNN has additional layers of convolutions and non-linear activation functions.

CNN Basics: The CNN architecture is made up of a few key steps or actions which is detailed below.

- **Convolution**. This is the first step in the process, that involves a data window iterating through the larger data table. This step requires specifying the stride length (determines the degree of reduction), depth (number of filters needed) and zero-padding needs. Zero-padding inserts zeros around the border of the matrix (often used in image recognition). At each stride matrix multiplication will be performed and results stored. Using the correct terminology, the window is said to convolve over the data with a stride size which can drive the 'blurring effect' in image processing. The kernel and the data sample are multiplied, using matrix mathematics, and the answer stored for each iteration, or stride, forming a new, reduced, matrix of data that retains the high-level features of the data and the spatial relationships are maintained Lane et al. (2019).

- **Feature extraction - Non linear activation** is the layer that is introduced to extract salient information by using an activation function, most often the RELU. Practically, this function that is used in a pairwise matrix multiplication as part of the convolution. When coupled with a kernel or filtering mechanism, is then known as a convolutional net, or Covnet (Lane et al., 2019). The kernel partially plays the role of reducing the size of the information in the larger data set (where image recognition data is very large), keeping the pertinent information needed for making the prediction (Goldberg and Hirst, 2017).

- **Pooling** is used to combine the results of the convolution in a vector and has an important role, to control the size of the output to a fixed size vector or matrix (Lopez and Kalita, 2017). The convolutions layer is also very effective at data reduction whilst retaining the most salient information (Lopez and Kalita, 2017). The aggregation or pooling is done by using the max function, average function or dynamic pooling (which take difference pooling functions depending on the location of the matrix locations). This pooling technique can be useful in domains where the position can make a difference (such as the centre of an image or the end of a sentence).

- **Classification layer**. The results of the pooling layer are fed back into the neural network for modelling predictions (Goldberg and Hirst, 2017). The modeller can use of a Softmax function as the last layer and this would allow for the multiple category or object classifications, as part of a fully connected layer is a traditional multi layer perceptron model. The Softmax procedure takes the values of the input vectors and returns a vector with values between zero and one (representing the probabilities of taking a value).

Mathematical representation of a CNN convolution The following is a convolution operation in one dimension. If one considers a passage of text, which is a collection of word

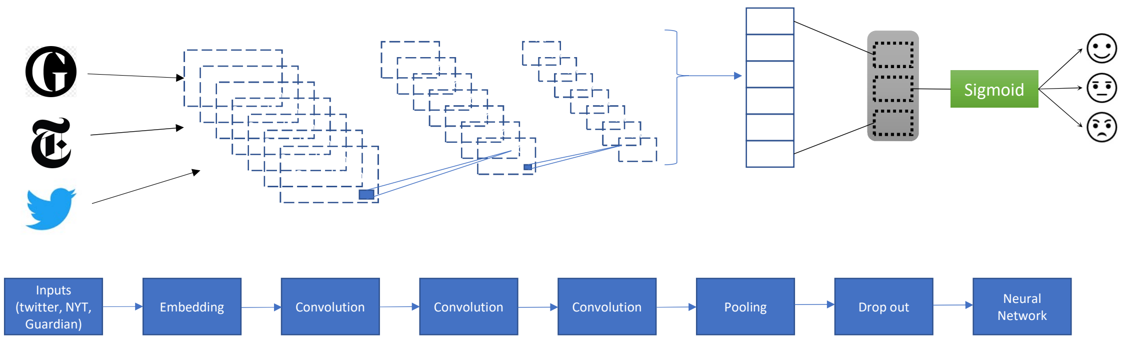


Figure 12: The above illustration represents a simple CNN that is based on text analysis, which could be a sentiment analysis based on the same data sources and sentiment analysis output in this study.

strings, defined as $T_{1:n} = T_1, \ldots, T_n$. The convolution operation is an iterative process of a sub-sample of a matrix of width k, progresses at a given stride length, until each position has been covered. The operation will also be filtered using a function designed to extract features and information. With each convolution I create a vector, T. I follow Goldberg and Hirst (2017) in defining the convolution:

$$\begin{aligned} p_i &= g(\mathbf{x}_i \cdot \mathbf{u}) \\ \mathbf{x}_i &= \oplus(\mathbf{T}_{i:i+k-1}) \\ \mathbf{p}_i &= g(\mathbf{x}_i \cdot \mathbf{U} + \mathbf{b}) \end{aligned} \tag{109}$$

b is a bias vector that is added. The resultant output is a vector, p_i for each of the filter value, u_i and the $\oplus(\mathbf{T}_{i:i+k-1})$ function is the aggregation of vectors from all the rolling windows, $\mathbf{T}$.

Pooling: The pooling function is set up to combine the vectors from the convolution by aggregation into a single vector, $\mathbf{c}$. The vector of salient information will be sent to the network layers for training weights that can then be used in predictions. Pooling requires an aggregation technique, the most common use is max-pooling. This is effective in highlighting the most salient features in each window by only taking the max value from $\mathbf{p}_i$,

$$\mathbf{c}_{[j]} = \max \mathbf{p}_{i[j]} \tag{110}$$

$\mathbf{v}_{i[j]}$ denotes the j^{th} component of $\mathbf{v}_i$. Average pooling is another popular technique which is simply an average of the underlying vector $\mathbf{v}_i$:

$$\mathbf{c} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{v}_i \tag{111}$$

Recurrent neural networks (RNN) Bengfort et al. (2018) describe RNN in the context of modelling sequential language data, not to be confused with recursive neural networks (RecNN) which is an architecture composed of structured inputs that are handled best with a tree or tree-like architecture. The RNN makes use of a recursive process of applying a calculation that is conditional on prior iterations with information retained as the sequence is iterated on. In the context of language processing, bag-of-words (BOW), even the more advanced Continuous BOW (CBOW), fail to accurately capture the meaning of the phrases or words being analysed when sentences or the order of the words, or even the context of a group of sentences are not picked up (Goldberg and Hirst, 2017). RNN's are designed in such a way that when information is fed in sequence and stored for referencing (a form of memory), it becomes more sensitive than lexical or BOW approaches. The sequencing of data in the RNN model retains the word order and POS of the sentences and words. In addition, the model tracks long-running dependencies between words as part of the inbuilt architecture (Goldberg and Hirst, 2017). Later in this chapter, I will discuss how one model, long short-term memory (LSTM), has been prove particularly successful.

RNN models can be broadly grouped by their function; the focus in this study is on acceptor models, rather than encoder and transducer models. This construct will sequentially take word by word sequences with the aim of predicting the sentiment of the text. I start with a simple RNN (sRNN) abstraction, vectors are denoted as $x_i, \ldots, x_j$ as $x_{i:j}$. This model returns a vector $y_n = \text{RNN}(x_{1:n})$. Such can be expressed as, following Goldberg and Hirst (2017):

$$\begin{aligned} \text{RNN}(x_{1:n}; S_0) &= y_{1:n} \\ y &= O(S_i) \\ s_i &= R(S_{i-1,} x_i) \end{aligned} \tag{112}$$

This output function can then be embedded for further prediction. The RNN is defined by recursion, with S_{i-1} as an input. The input vector, x_i will returns an output vector, s_i, where s_0 is usually omitted. Finally s_i is translated or mapped into the final output vector y_i using a simple deterministic function, $O(\cdot)$.

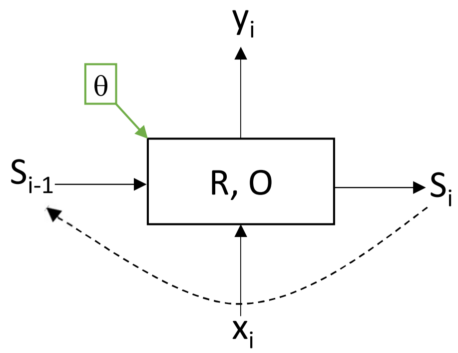


Figure 13: The above illustration represents a simple RNN schematic (Goldberg and Hirst, 2017).

When I consider the same representation mathematically, if we are to 'unroll' the notation, the stacked functions, this can then be represented by R and O, representing the way in which the network is structured:

$$\begin{aligned} S_3 &= R(S_2, x_3) \\ &= R(R(S_1, x_2), x_3) \\ &= R(R(R(S_0, x_1), x_2), x_3) \end{aligned} \tag{113}$$

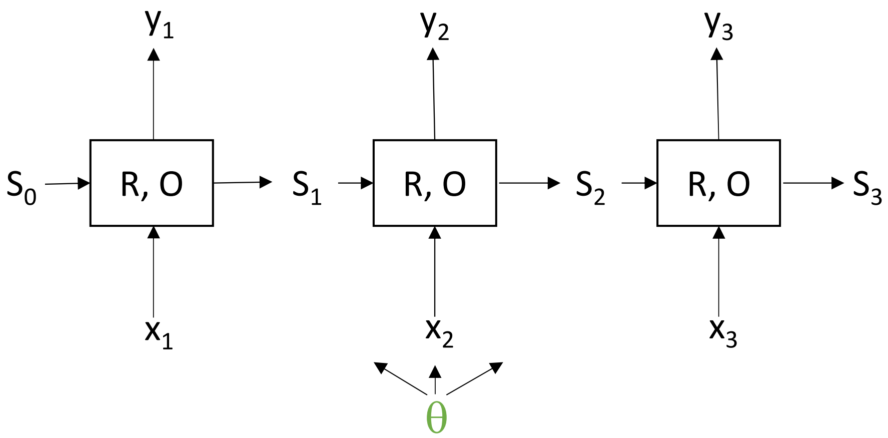


Figure 14: The above illustration represents a simple RNN that is unrolled, like a net that has been rolled out and laid down (Goldberg and Hirst, 2017). Note that the θ function is present in each node, as illustrated by the arrow pointing to each node from below.

Goldberg and Hirst (2017) explain that training the RNN involves a computational graph for each node in the sequence, which is coupled with a loss function for the whole unrolled network and this then ‘backpropogated’, (Goldberg and Hirst, 2017). Lane et al. (2019) elucidate that long short-term model (LSTM) introduces a state variable to correlate meaning related to words which may be far apart in a sentence, or even in other sentences.

There is an enhancement to the RNN which is relevant in an NLP use-case. The fixed window of words which are considered at each step is opened up to include words that are further down the sequence of the sentence. This is a bidirectional RNN (biRNN) and is applicable for sentiment analysis (given that I am not trying to predict a future word that is already in the observation window) (Goldberg and Hirst, 2017). The mechanism for this method involves having an RNN for the previous words in the sentence and a second RNN, which reviews the words in reverse order. The algorithm then solves for each RNN independently with results that are concatenated.

Recursive neural networks (RecNN): RecNN is a series of interconnected neurons forming a network of interacting connections. Goldberg and Hirst (2017) note that a RecNN is a natural and desirable approach to use in language processing, providing good mechanisms with which to model sequential data. RNN networks are chain structures, RecNN on the other hand, are based on hierarchical data models with a tree structure. Bengfort et al. (2018) comment that model hidden layers that make use of activation functions, coupled with a dimensionality compression function to match requirements. Goldberg and Hirst (2017) write that RecNN models are one of the more promising techniques to come from machine-learning in the domain of natural language processing. The technique does not rely on the Markov assumptions which many of the work-horse models in natural language processing have relied on, whilst taking word order into account. A significant advantage of the model is that it can serve as a source of data for other models, as an input, such as the required fixed vectors needed for the feed forward modelling to predict the next word in a phrase. The LSTM model is an extension to the RecNN which has memory component for solving the problem of vanishing gradients. It has gates, the first is an input gate that stops or accepts updates. The second gate trims the neurons that are not needed (determined by immaterial weights resulting from the algorithm) and lastly a control gate for outputting information.

The next important architecture, or RNN design, deals with gradient issues such as vanishing gradients and difficulties of training the simple RNN. Hochreiter and Schmidhuber (1997) designed a gated architecture called Long Short Term Memory (LSTM) which is highly effective. As Lane et al. (2019) remark, the LSTM model was set up in order to overcome challenges such as vanishing gradients. Goldberg and Hirst (2017) underline that the simple RNN operates each step with infinite memory, it is not controlled. This brings the introduction of gates, which control what is passed into memory and what is excluded from the analysis. Goldberg and Hirst (2017) explain that the use of a binary vector ($g \in 0, 1^d$) will serve the role of a gate. By way of element wise multiplications, useful information can be let through ($g = 1$), whilst useless data can be excluded ($g = 0$). However this simple component still requires a trigger to move from the open state, $g = 1$, to a closed state. An available solution is to use the sigmoid function, which can easily be mapped to a binary result, before being trained using gradient-based methods.

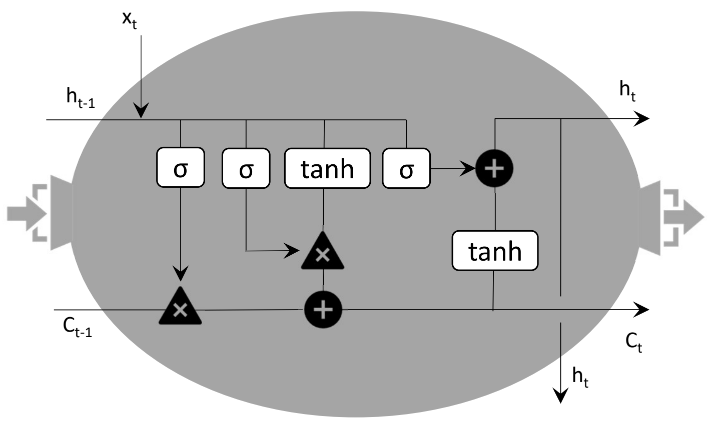


Figure 15: The above illustration is a schematic of the LSTM model, which has been adapted from Alom (2019).

A model that has risen to prominence that uses the gating mechanism is the Long Short Term Memory (LSTM), Goldberg and Hirst (2017) argue that LSTM models are the greatest contribution to natural language that has come from the statistics domain (following Goldberg and Hirst (2017)):

$$
\begin{aligned}
s_j &= R_{LSTM}(s_{j-1}, x_j) = [c_j : h_j] \\
c_j &= f \odot c_{j-1} + i \odot z && \text{[Memory cell]} \\
h_j &= o \odot \tanh(c_j) && \text{[Hidden layer]} \\
i &= \sigma(x_j W^{xi} + h_{j-1} W^{hi}) && \text{[Input gate]} \\
f &= \sigma(x_j W^{xf} + h_{j-1} W^{hf}) && \text{[Forget gate]} \\
o &= \sigma(x_j W^{xo} + h_{j-1} W^{ho}) && \text{[Output gate]} \\
z &= \tanh(x_j W^{xz} + h_{j-1} W^{hz}) \\
y_j &= O_{LSTM}(S_j) = h_j && (114)
\end{aligned}
$$

The forget, input and output gate values are calculated by combining the latest and previous state data, x_j, and h_{j-1} respectively, then parsed through an activation function (in this case a sigmoid function mapping to a binary outcome). The update variable is also a combination of x_j, and h_{j-1} but passed through a tanh activation function. The memory cell, c_j, is a function of the forget and input gates. The input gate will determine what new information enters downstream functions ($i \odot z$), the forget gate determines the retained level of previous information ($f \odot c_{j-1}$). The final step determining h_j, or y_j, involves an activation function, tanh and the output from the memory function, c_j.

LSTM architecture is currently the most powerful with acceptable modelling results in many domains (Goldberg and Hirst, 2017). Albeit a complex set up, the elegance of LSTM architecture is that trained neural nets contain self contained rules for state memory (Lane et al., 2019).

9.5 Examples of NLP and sentiment analysis in financial literature

Market instruments and securities are early targets for unstructured text modelling, as demonstrated by Antweiler and Frank (2004). By investigating the claim that online discussion forums for the financial sector can affect the market, they hope to shed light on this topic. Where they review 1.5 million messages extracted from Yahoo! Finance and Raging Bull with reference to 45 companies found on the Dow Jones internet index. When The Wall Street Journal was used as a control, the effect of these messages on stock prices was found to be significant but economically insignificant. Antweiler and Frank (2004) find, interestingly, that disagreement among the people posting is also associated with increased trading volume.

According to the findings of Bollen et al. (2011), using certain variables of public sentiment can greatly enhance forecasts of the index value of the Dow-Jones Industrial Average. According to a recent article by Mudinas et al. (2019), sentiment analysis is well-suited for predicting future economic performance, as well as identifying and understanding human emotions and behavioural shifts (such as happiness, rage, and sadness). Analysis of economic news and consumer sentiment by Shapiro et al. (2020) demonstrate that positive shocks to sentiment based on a large corpus of financial newspapers are followed by rises in output, interest rates, and consumption while lowering inflation.

For investing timing techniques using the MSCI World Index, Beckers (2018) indicates that news-based sentiment delivers a more consistent signal than sentiment gleaned from social media. According to a study by Beckers (2018), sentiment signals gleaned from news and social media can be used to predict the performance of major stock markets in the months ahead. Shapiro and Shapiro (2019) measure tonality by measuring sentiment in the U.S. Federal Open Market Committee's (FOMC) internal discussions, then use this to create an index, which is used as a proxy for a loss estimate. Using textual data from the public archives stored on the Federal Reserve Board of Governors website, the sentiment model measures style and tone in speaking by different members' speeches and remarks from 1976 to 2013, allowing for heterogeneity [8]. The speeches and remarks are then filtered to only include sentences and remarks with one economic-related term defined in the Oxford Dictionary of Economics (ODE).

Non-traditional data sources like Twitter and Stock-Twits where used by Liew and Budavari (2017) to index company-specific tweet sentiment. Azar and Lo (2019) investigated the relationship between Stock-Twits and Twitter as it relates to liquidity measures. They also point out that negative social media sentiment has a two-fold greater liquidity effect than positive sentiment, which is consistent with the idea that fear and panic have a greater impact on markets than manias or booms. Azar and Lo (2019) show that the Twitter data contains information and can be used to predict returns and volatility by using a regression on the Twitter data and controlling for other factors (Fama-French factors), all based on excess returns.

Recently, Karagozoglu and Fabozzi (2017) used social media sentiment as an input for VIX volatility trading and discovered that the information available on social media provided for a meaningful trading strategy that outperformed the benchmark. Liew and Budavari

[8] https://www.federalreservegov/monetarypolicy/fomc_historical.htm

(2017) view is that idiosyncratic risk can be broken down into two components, noise and social sentiment. Liew and Budavari (2017) explain that Stock-Twits allowed them access to their historical data. Liew and Budavari (2017) find lots of evidence of information at the security level from social media information sources. These applied to time series data are significant in explaining security returns, even after being controlled for traditional factors of Fama-French.

Liew and Budavari (2017) report that Twitter users with greater followers numbers have a stronger influence on returns. Sul et al. (2016) conclude in their work that S&P500 returns and sentiment polarity drawn from Twitter are positively correlated. What is also recognised is that Twitter users with more followers have a influence on returns. Erlwein-Sayer (2018) reviews the relationship of the economic fundamentals that are seen by many market players as drivers of sovereign spread, their study finds that the yield spread forecasting can be enhanced by using a daily news sentiment. Erlwein-Sayer (2018) continue in their conclusion, that macro-economic news sentiment and sovereign bonds of the five European countries are significantly correlated (especially in the case of yields spread, spread changes and volatility of spreads). In bull markets, bonds spreads are negatively correlated to the volume of positive news, whereas positive correlation are found for news volumes in the case of bear markets.

Using newspaper articles from economic and financial sources, Shapiro et al. (2020) created a new economic consumer sentiment time series spanning January 1980 to April 2015. The objective of their study is to compare the accuracy of the time series sentiment index to well-known survey-based consumer sentiment indexes developed by the University of Michigan and Conference Board. The findings of Kalamara et al. (2020) indicate that linguistic analysis of newspaper material considerably enhances the forecasting of important macro-economic variables. This study examined variables such as GDP, inflation, and employment. Kalamara et al. (2020) discovered that the utilised methodologies were particularly effective during times of economic stress; they examine the power of news text for sentiment and uncertainty signals. They discovered that news sources contained more robust signals for evaluating sentiment than economic uncertainty. Kalamara et al. (2020) find that simple methods which target text based variables do not perform as well as the combination of linear and non-linear supervised machine-learning algorithms.

Nyman et al. (2018) show that sentiment build-ups prior to a financial collapse are linked to sentiment; and narratives also play a role in financial markets. The analysis of these researchers centred on two measures – excitement and anxiety – whose difference expressed as a percentage of size forms the sentiment index as a time series. Souleles (2004), Carroll et al. (1994), and Bram and Ludvigson (1998) are good examples of studies showing how sentiment measures have predictive power for individual spending.

Kalamara et al. (2020) maintain that the news text substantially improves forecasts of variables such as inflation, gross domestic product, and employment levels. The researchers point out that this will assist in augmenting less frequent and more expensive survey methods currently in use. Kalamara et al. (2020) discuss that newspapers are an independent source of information and potentially capture the influence of 'animal spirits'. They test the different techniques of extracting information from text by comparing word counts, which perform well considering how basic the method is, to methods that use a dictionary of financial stability words and perform better. The most powerful of the methods is based on machine-learning methods trained on text-derived regressor variables, which outperform the other methods. Non-linear NN ultimately performs the best when considering dependent variables, sources of information, and time in each analysis.

Garcia (2013) used data from The New York Times to model the long history of stock prices from 1905 to 2005. They demonstrate that stock price prediction during market downturns can be attributed in part to news sentiment, after accounting for standard variables. A one-standard-deviation shock to the news sentiment indicator is followed by a decrease in the Dow Jones Industrial Average, suggesting that the data is concentrated at times of extreme market volatility. Manela and Moreira (2017), on the other hand, used data from The Wall Street Journal front page to investigate the relationship between how a news-based uncertainty measure performs. Manela and Moreira (2017) observe that, over the long-run (1890-2009), their uncertainty measure predicts higher future returns in normal times, which is also predictive of economic disaster due to an increase in the uncertainty measure just before the business cycle change.

In another example of text based modelling by Lang et al. (2017), they show that buyer emotions can potentially be affected by news sentiment, ultimately affecting the prices of US commercial real estate. They compare the use of a NB and a SVMs against 40,000 news articles extracted by filtering news data from 2005 to 2015. Fundamental economic variables alone do not explain commercial real estate returns (Lang et al., 2017).

Larsen and Thorsrud (2019) take a slightly different approach, examining the impact of Norwegian newspaper articles and the ability of news text sentiment to predict economic variables. They reveal that many topics have forecasting ability and predictive power for key macro-economic variables. Unexpected changes in the news index have been linked to long-term effects and changes in asset prices, including effects on credit and borrowing.

10 Implementation of sentiment modelling

When applied to text written in a natural language, sentiment analysis can determine if the writer is being positive or negative. NLP techniques are largely focused on unstructured sources of information that require a degree of pre-processing in order to get the information in a usable format for modelling. Text data can then be classified into sentiment categories such as positive, negative, or neutral; this is often added to a scoring/rating model that generates valence (a rating from 1–5). A valence score can be generated using different methods, such as lexical analysis, simple statistics, or frequency based counting, or more advanced techniques such as machine-learning. Lexical models can range from simple word-based scoring systems that apply scores to words based on a list of positive and negative words (such as Textblob), to more advanced systems allocating valence scores and a secondary set of heuristic rules. Some of the tools can now review in the context of the word within a sentence, affecting the sign and scale of the valence score. VADER is one such tool; it is a free Python package that I use in this analysis.

Given the magnitude and scale of the data and modelling in this exercise, the next section I dedicated to the process I undertook to collect, transform, pre-process, structure, and ultimately analyse the sentiment in the news and social media text. This is an important foundation for more advanced modelling. What is important to highlight, often, is the extent of the exercise that collecting Twitter data was. This included six to eight hour checking that the process of collecting streaming tweets continued, for more than a year. The data was voluminous, billions of lines long, that was largely filtered out (the process and net result I detail later in the chapter).

10.1 Alternative data extraction and modelling process

The approach I took in this study was to independently create a solution for market related signals, using readily available information, modelling frameworks and tooling. This approach requires that I independently researched, sourced, cleaned and collated, signal processed and aggregated data for each of the three regions in this study. I also confirmed an appropriate business cycle proxy as a dependent variable. The platform to deal with data was fully codified by myself using data extracted from APIs into a Google Colab environment or onto my local machine using a data base. A significant proportion of my time was invested in cleaning up the unstructured data, some sources were more difficult to deal with (I explained the challenges of working with Twitter data). Even when data was relatively easy to work with, for example the news sources, the challenge of dealing with large volumes of unstructured data should not be underestimated. Stemming and cleaning data required many iterations of programming and review to ensure that the data was adequately processed and the natural language programming techniques are appropriate for the investigation in this study. I found the only way to really understand the techniques, was to make use of the technique and observe the results. I summarised the net results of the favoured techniques only, I note here that many were not reported in the final study (for instance CNN and LDA analyses). The same is true for the modelling for prediction of the business cycle proxy, many techniques failed (for instances making use of the data as a corpus and directly modelling the dependant variable). The effort and time required for this final result set in this study was significant. I organised this analysis in a way that is logical and follows the lifecycle of the data in this study, I illustrated this in Figure 16.

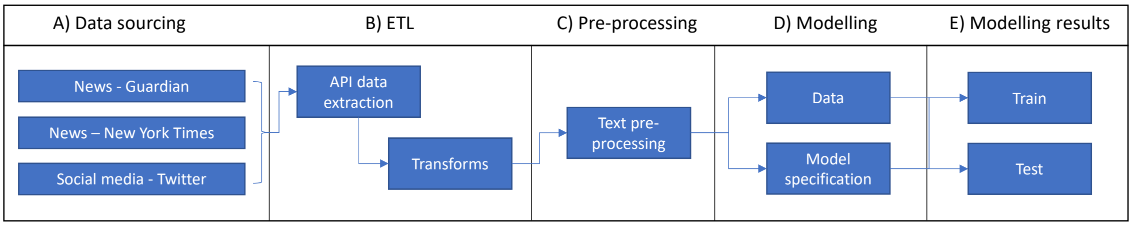


Figure 16: The above process details the high-level steps for the data and modelling processes using a simple process flow diagram. This type of analysis is an effective way to describe the interrelated processes that are needed for the data and modelling in this study. This financial and economic data have been included in the diagram for completeness, as they are part of the larger study but are not detailed until later. The diagram is annotated in a way that aligns with the document write-up that follows.

10.1.1 Data sourcing

Independent variables sourcing. The following sections provide a summary of the types of data and respective sources that are extracted from the alternative data sources, including

Twitter, The New York Times, and The Guardian.

Category	Source	Access	Description of data source	Start date	Article count
UK News	The Guardian	Free API	The Guardian newspaper is a daily broadsheet format that has been published since 1821 Guardian provides full access to news articles freely available via API.	Jan'1980	531,408
US News	The New York Times	Free API	Daily morning newspaper from New York city, it has a long history, founded in 1851. It is important newspaper that is regarded as the 'newspaper of record'. New York times make available freely data via an API.	Jan'1980	2,379,005
Social media	Twitter	Free API	Microblogging site and application, allowing free publication of 140 character posts called tweets. Twitter make streaming data available via the Twitter API version 2.	Sept'2020	2,347,319
Local news sentiment	GDELT	Free API	Supported by Google Jigsaw, this services hosts live monitors of the world's media including web articles, blogs, newspapers and news sites.	Jan'2017	7,434,976

Table 5: Description of alternative data sources, article counts and start date. I have not specifically listed it in the table, but the end date for all sources is December 2020.

Dependant (target) variables sourcing The indexation of economic variables is a complex, data-intensive task. There are different indices that are available in a particular region, but these approaches are only local. In this study, I need to select an index that best represents business sentiment at a given point in time. This needs to be consistently measured across the three regions in this study. The principles I applied for index selection are as follows:

- Independent indexation and data sourcing (i.e. not using own data or indexation approach).
- Globally recognised source of data.
- Consistent methodology across the regions.
- Freely available data that is updated in a timely fashion.
- Final choice of index will be the index that best aligns to the business segment under investigation. In this case the index that best represents the small-to-medium enterprise segment.

The OECD is a global organisation that specialises in econometric indexes. The methods of calculation and underlying philosophy for how they construct their indices and composite indices. The methods are published and transparent (https://www.oecd.org/ sdd/42495745.pdf). The target variables in this analysis are sentiment indices that have been provided by the OECD database service and can be found in the main economic indicators section, (OECD, 2021). I make use of Quandl data API to access and retrieve the OECD data indices.

OECD index name	**Index code**	**OECD index description:** sourced and directly quoted from OECD (2021) and the same index is drawn for each of the three regions selected in the study (GBR, USA, RSA).
The Business Confidence Index	BCI	This business confidence indicator provides information on future developments, based upon opinion surveys on developments in production, orders, and stocks of finished goods in the industry sector. It can be used to monitor output growth and anticipate turning points in economic activity. Numbers above 100 suggest increased confidence in near future business performance, and numbers below 100 indicate pessimism towards future performance.
The Composite Leading Indicator	CLI	This index is intended to provide early warning signs of business cycle turning points by displaying fluctuations in economic activity around its long-term potential level. CLI's show short-term economic movements in qualitative rather than quantitative terms.
Consumer confidence index	CCI	This consumer confidence indicator provides an indication of future developments in households' consumption and saving, based upon their answers regarding their expected financial situation, their sentiment about the general economic situation, unemployment, and their capability of saving. An indicator above 100 signals a boost in the consumers' confidence towards the future economic situation, as a consequence of which they are less prone to save and more inclined to spend money on major purchases in the next 12 months. Values below 100 indicate a pessimistic attitude towards future developments in the economy, possibly resulting in a tendency to save more and consume less.

10.1.2 Extract, transform and load (ETL)

The following table represents the steps that are followed in the coded data extraction script in Python. Each script differs slightly depending on the underlying data source. Rather than include the scripts, I represented this process in a table below, with each data source indicating whether this is included in that particular script. What is noticeable, relative to the other sources, is the extra steps required for Twitter extraction (i.e., streaming data for a full year to collect tweets is a slightly larger task than one extraction of historical news data, as is the case with The New York Times and The Guardian).

Section	Data engineering steps Sources ⟶	Guardian	NYT	Twitter	GDELT
Extraction	• List dates	✓	✓	✓	✓
	• List search terms (refer to 26.4)			✓	✓
	• Call API (using API key) in a loop for each date	✓	✓	✓	✓
	• Collect tweets from Twitter API streaming (on-going) and occasionally save a new JSON file (size dependant)			✓	
	• Filter on business, markets and finance sections	✓	✓		
	• Filter by Language = ‘ENG’			✓	
	• Extract required fields	✓	✓	✓	✓
	• Concatenate text with extended tweets using tweet mode = ‘extended’ and concatenate ‘full_text’ field			✓	
	• Exclude retweets by filtering for ‘RT @’ and excluding tweets with retweet_ind set as positive			✓	
	• De-duplicate data, removing all duplicate tweets			✓	
	• Save JSON files returned from API endpoint	✓	✓	✓	✓
Sentiment Scoring (1)	• Using Textblob and VADER, scored text and retrieve sentiment scores	✓	✓	✓	✓
Transform and save	• Transform dictionary file and append as CSV	✓	✓	✓	✓

10.1.3 Pre-processing of unstructured data

The next step is the pre-processing of the natural language data, following the process that was described in Section 8.2. This is the key step in preparing the data for the sentiment scoring process. Some of the steps may not seem necessary, such as removing @RT from the Twitter text, this being inconsequential for the net result; however it does clean up the text and reduce the overall size of the data. The language text is first scored using the sentiment packages (VADER and Textblob) and then scored after the pre-processing steps have been concluded. The second round of scoring is conducted using the exact same technology as the first, but performed on the raw text that has not been preprocessed. What will be of interest is the difference in sentiment source, which I will investigate later (i.e., testing the impact of preprocessing).

The following processes were used for each of the data sources: Pre-processing (removal of special characters and punctuation, stemming, vectorisation using the Keras framework). Sentiment scoring using Python TextBlob and VADER; and lastly feature engineering (rolling variables, exponentially moving weighted average and rolling percentiles).

Section	**Data engineering steps** **Sources** ⟶	**Guardian**	**NYT**	**Twitter**	**GDELT**
Pre-processing	• Remove the following: special characters, single characters, any remaining 'RT @' or '@RT', remove 'HTTPS', replace double spacing with single, exclude emoticons	✓	✓	✓	✓
	• n-gram analysis: create bi-grams using NLTK and thereafter extend to tri-grams using NLTK	✓	✓	✓	✓
	• Stemming: stem remaining text using NLTK	✓	✓	✓	✓
	• Vectorization: vectorisation of the remaining text using the Keras framework	✓	✓	✓	✓
Sentiment Scoring (2)	• Using Textblob and VADER, score pre-processed and stemmed text and retrieve sentiment scores	✓	✓	✓	✓
Feature engineering	• Summarise Textblob and VADER outputs, including volatility, rolling averages, EMWA, percentile trends (of average score of 3 month window of 75th and 25th percentile). Repeat for both raw text and text that has been stemmed and pre-precessed	✓	✓	✓	✓

10.1.4 Scaling variables

Scaling is an important process used to ensure the training and target variables are on the same measurement scale. Scaling can be important, as variables that are fitted in models at different scales will not contribute equally and can therefore introduce bias into the final model function. The process has to do with changing the range of a variable, which does not affect the shape of the distribution or the trend of the time-series data over time. Scaling is often done by transforming the variable range to be between 0 and 1 or between -1 and 1. Min-max scaling is one of the most commonly used and simplest methods in machine-learning problems. The method is as easy to perform as it is to reverse (which is needed for models to produce a prediction). Min-max scaling is defined as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{115}$$

The Python scikit-learn pre-processing package has a scaler function called MinMaxScaler. I detail the key algorithm below, and further detail of the function can be found on scikit-learn technical pages [9]. The second technique for dealing with data on different scales is standardisation. Using the formula below, the standardisation technique is applied to normally distributed variables and scales by setting the mean to zero and the standard deviation to one. Standard scaling is defined to be, where μ is the mean and σ is the standard deviation:

$$z = \frac{x - \mu}{\sigma}$$

To summarise, the dependent variable is transformed to account for autocorrelation in sentiment variables before being standardised and scaled for plotting with the StandardScaler. The result is a range between negative four and four standard deviations from the mean. The Python scikit-learn pre-processing package has a scaler function called StandardScaler. Algorithm detail and function can be found on scikit-learn technical pages [10].

10.2 Modelling data descriptive analytics

In the following section, I describe the variables used to build the models, both the target variables and independent variables.

[9] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

[10] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html).

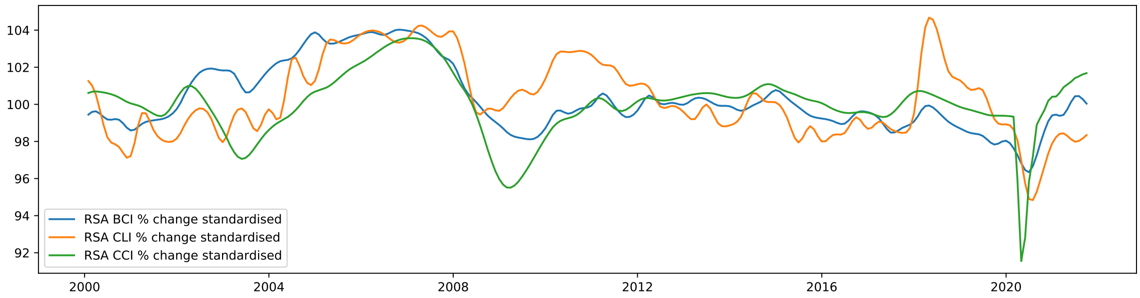


Figure 17: The sentiment indices I selected are available across different regions using a standard methodology, the OECD methodology. The South African CCI, CLI, and BCI indices are plotted in a time series above using the month end values, these numbers have been standardised (therefore will lie between 0 and 100). Significant downturns in sentiment can be seen, first during the GFC around 2008, and more recently, with the COVID pandemic beginning in 2020.

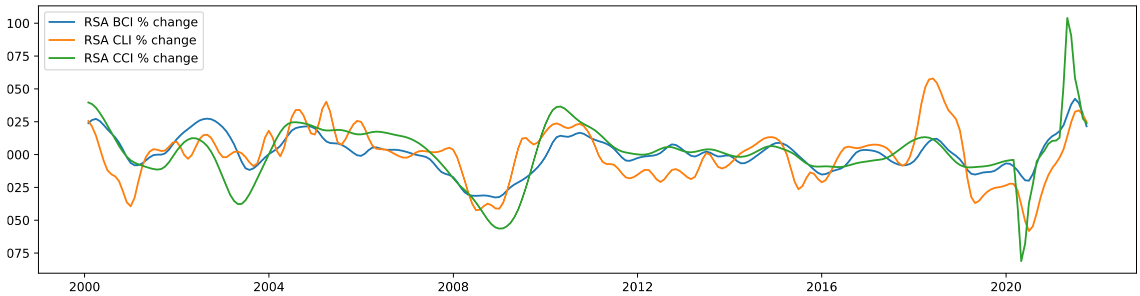


Figure 18: South African CCI, CLI, and BCI indices are transformed to year-on-year percentage change and re-plotted in the above time-series using month-end values, these numbers have been standardised (therefore will lie between 0 and 100).

10.2.1 Target variables

In the following section is an analysis to highlight salient features in the trends displayed in the target variables, the OECD sentiment indices. I plotted a time-series chart for each region and included the raw OECD data for each of the three sentiment indices (BCI, CCI, and CLI); refer to figures 19, 21 and 22. The first discernible feature is the distinct periods of negative downturn or drawdown in the sentiment indices between 2008 and 2020, corresponding to the Great Recession and the more recent COVID pandemic. Secondly, all three indices in all regions showed a sharp increase in and around the time of the announcement of COVID vaccine approvals in 2020.

A review of each regional index reveals that CLI sentiment has a consistent trend across

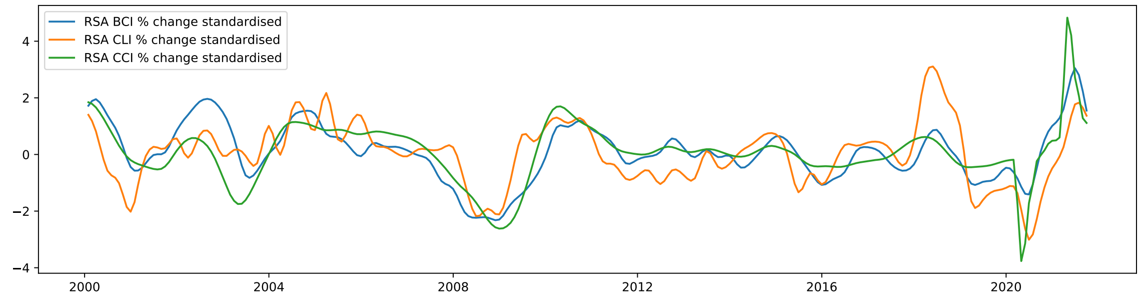


Figure 19: The above figure demonstrates an example of scaling by standardisation on the RSA BCI sentiment variable (therefore will lie between 0 and 100). The same sentiment indices have now been transformed using the standard normalisation technique. The time-series dynamics are not impacted by this technique, which is evident when comparing to Figure 20.

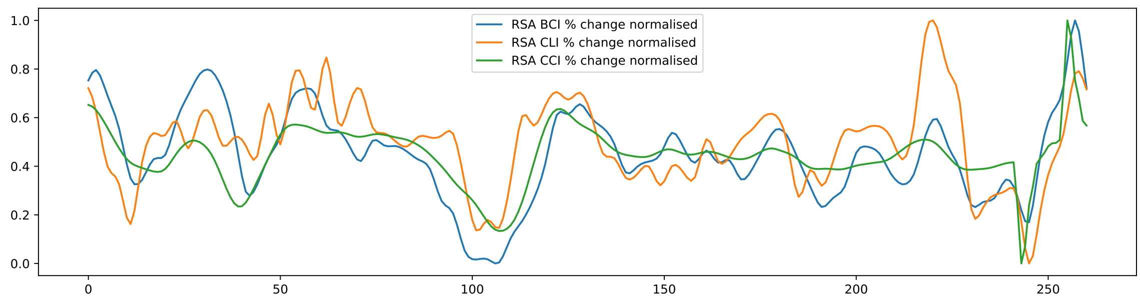


Figure 20: The above figure demonstrates an example of scaling, using MinMax scaling on the RSA BCI sentiment variable. What is clear is that even though the scale is transformed to lie between 0 and 1, the trend has been retained. An attractive feature of scaling is the ease of plotting multiple variables on the same axis, which allows for direct comparability.

regions (UK CLI index correlation of 72% in RSA and 75% in USA, respectively). If I observe the CCI index, this index appears to reflect more localised events, such as the RSA CCI index showing a sharp increase in 2018 (refer to Figure 17) which aligns with the election of Cyril Ramaphosa as South Africa's president (the UK and US did not experience the same upturn). Similarly, the UK CCI shows a sharp decline in 2016, around the time of the Brexit vote (refer to Figure 21). Knowing the test data will include the recent pandemic (by design), it will be interesting to see whether sentiment scores from news and social media data will have power to predict an event which is a human crisis, where the models are trained from data which is based on an event born of a GFC (training data will not include the period when the pandemic took place).

The final decision on which sentiment value to use was taken by considering which variable reflects sentiment that is linked to the small-to-medium business segment. The sentiment variable that is closest to this is the OECD BCI, or business confidence index (BCI).

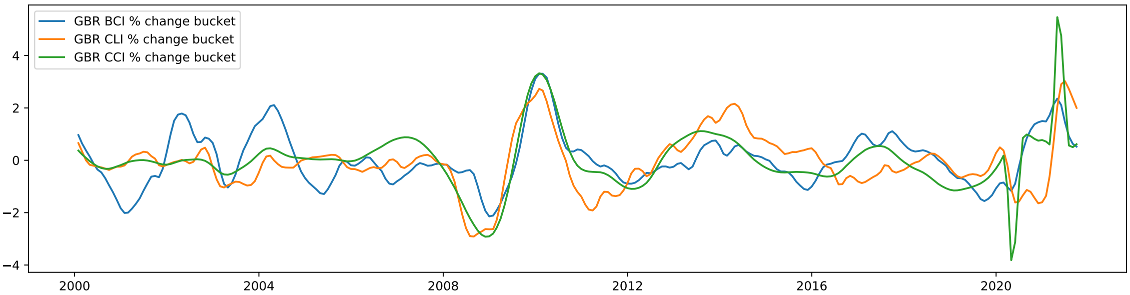


Figure 21: Trend analysis for UK target variables, including OECD BCI, CCI, CLI variables.

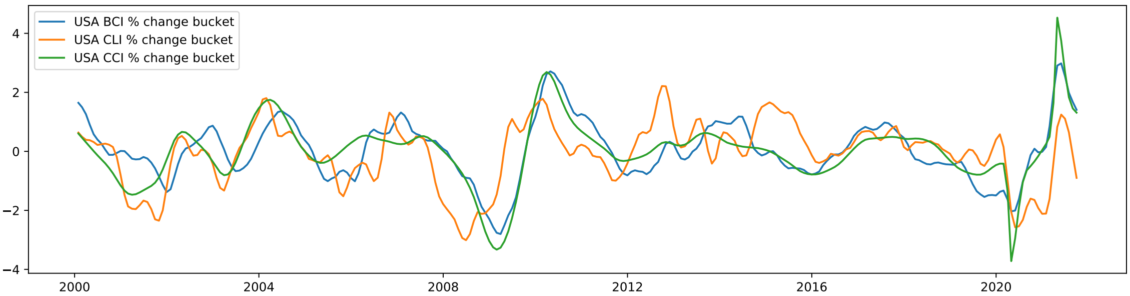


Figure 22: Trend analysis for USA target variables, including OECD BCI, CCI, CLI variables.

10.2.2 Training data and factor analytics

The following section is a brief analytical review of the lexical analysis and variables created from the process of feature engineering performed on the lexical sentiment scores. I also compared the lexical results to the target variables to get an indication of the predictive capacity of the factor data. I start with the summary of the data sources in Figure 23, measured by the number of rows of data in the final training and test data. The significance of the number of data points from Twitter is interesting on two counts. Firstly, this collection of Twitter data only spans one year (the other sources include more than 12 years), and secondly, the data excludes all the rows (or tweets) that are retweets, duplicates, or in foreign languages. On the other hand, the size of text carried in each row of news data is significantly larger (as the comparative run times on the lexical modelling, stemming, and vectorization will have proven). The length of the natural language varies by the different sources (The New York Times, The Guardian and Twitter), from an average length of tweets in the Twitter data at 176 characters, up to an article length in The Guardian being 4926 characters long on average.

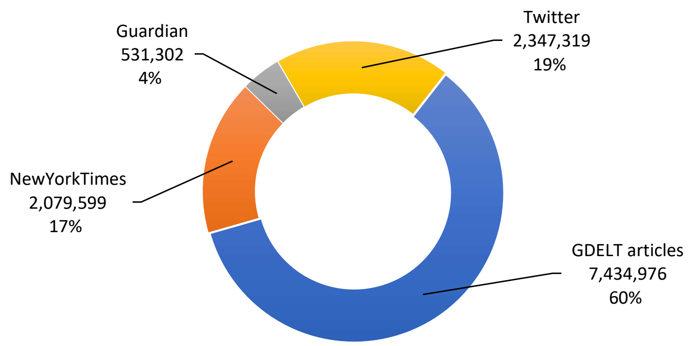


Figure 23: Key data sources and their respective contributions to the training data table The most significant source of information is the GDELT data set.

The analysis reported in Table 6, is a review of the news text processed using the Textblob and VADER lexical engines. The scoring was performed on the pre-processed data and then re-performed using data that had been stemmed and n-gram analysis completed. The n-gram and stemming have clearly had an effect, with negative and positive sentiment returning higher sentiment valence measures. The mean shifts from negative to positive for VADER, and this can be largely explained by the improved mapping of words from the lexicon (a central task in NLP). Importantly, the trend analysis (refer to Figure 24) will show that the lexical engine results are very similar since 2007.

Guardian	μ	σ	**Min**	**Max**
VADER compound clean	9.7%	7.6%	-27.0%	14.2%
VADER compound clean stem	15.2%	6.5%	-14.4%	29.3%
Textblob polarity clean	6.4%	4%	5.1%	9.0%
Textblob polarity clean stem	7.1%	5%	5.5%	9.8%

Table 6: The Guardian lexical scoring summary table.

If I compare the benefit of creating factor indices with full processing (including stemming and cleaning) and compare the results of the trend of the indices, see Figure 24, the difference is small. The Pearson correlation of VADER sentiment scores between stemmed text and raw text is 83% and 77% for VADER and Textblob, respectively. This difference drops even further from 2007 (the correlation moves to 97% and 96% for VADER and Textblob, respectively). This shows that full processing, such as stemming and n-gram, does not dramatically affect the results. If I consider the time taken to execute the stemming program, it can be seen as an optional step for articles with lots of text, such as the data from The Guardian.

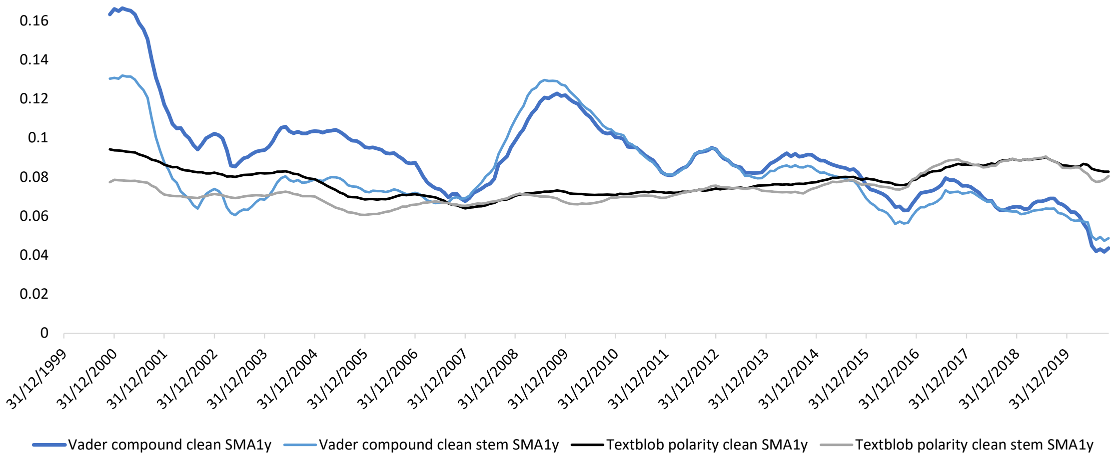


Figure 24: Textblob and VADER score raw data based upon stemmed text plotted for Guardian text. The Pearson correlation of VADER sentiment scores between stemmed text and raw text is 83% and 77% for VADER and Textblob, respectively.

The Textblob sentiment trend is flatter than the VADER sentiment trend, arguably, the Textblob sentiment trend does not show much activity, less than one may have expected over two periods of significant market stress and another of human stress in the COVID pandemic. VADER mean based sentiment variables are plotted on Figure 25, on The Guardian and New York Times, which is quite interesting to review, where both show a similar sentiment trend toward the middle of the GFC, January 2008 onwards. What is different is in the run-up to the GFC, where The New York Times shows a significant and sharp decrease and The Guardian, from a low base, shows a gradual decrease from 2005 to 2008. Both indices show significant deterioration from the end of 2019 and into 2020, in particular. The New York Times shows a significant increase in the sentiment score as the effects of a successful COVID vaccine are announced and rolled out.

Now I review the rolling volatility of VADER scores. The 75th percentile of the VADER scores shows a distinct increase around the time of the GFC and again around the time

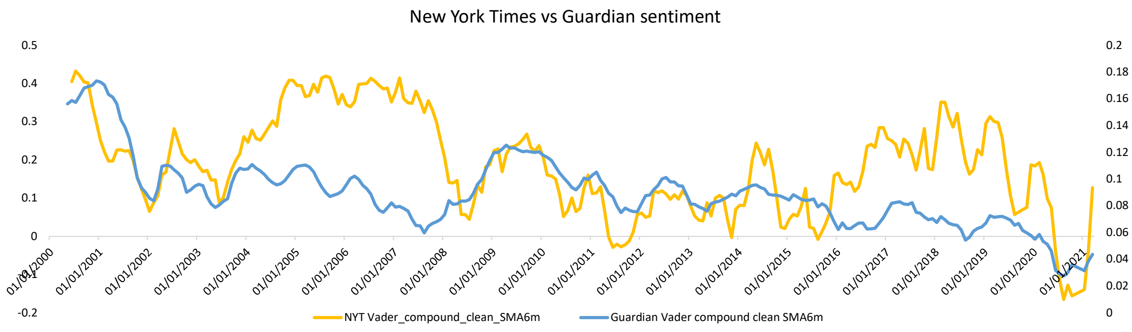


Figure 25: This figure is of the rolling six-month average of the average sentiment, which was generated using the Python VADER package and reported by both The New York Times and The Guardian.

of the pandemic. This indicates significant changes in the aftermath of the GFC and the pandemic, as shown in Figure 25. The second variable of interest on the Figure 26, is the 75th percentile of VADER sentiment scores, which shows an increase around the GFC but not during the pandemic.

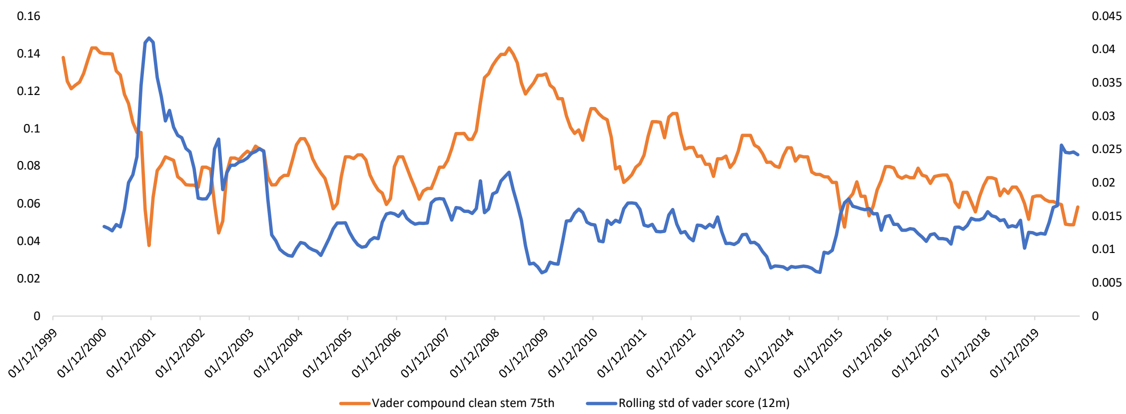


Figure 26: The above figure features variables that have been created from the VADER sentiment results, including the 75th percentile of VADER sentiment scores that are aggregated and, secondly, the volatility of VADER scores.

Figure 27 is a plot of the sentiment scores that are generated by GDELT data. GDELT data is a time-series reaching back to 2017. The grey line represents the average scores from GDELT by extracting data based on a set of search terms. The green line the same sentiment measure weighted by the relative volume of articles (GDELT returns the number of articles relating to the search term). Both lines indicate a significant decline during the pandemic. GDELT is the only sentiment data source included in this study that accurately captures the true of difficulties during the epidemic that I would expect to see from sentiment scoring. The GDELT data project's comprehensive coverage of information and news outlets, from national and international news outlets to popular blogs, enables readers to get insight into the true sentiment at the business level. The increasing coverage can only be seen as beneficial, especially when the less prominent local news sources may not be as closely controlled in terms of their content and information emerges through sentiment analysis.

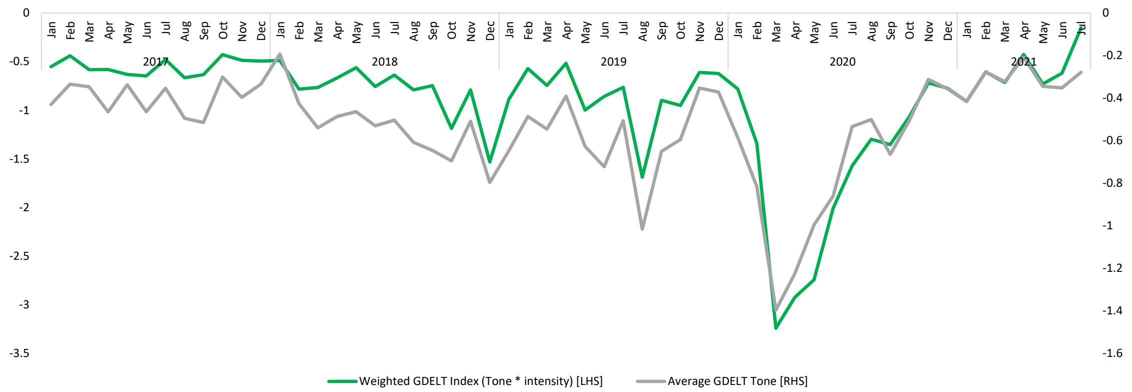


Figure 27: GDELT tone and intensity-weighted tone GDELT data shows a significant drop in sentiment around the time of the COVID pandemic, more so than other trend variables.

As a contrast, when I review the Twitter data, which does not span a period that is long enough to make sense of the longer-term trends, what I can see is a slow increase in VADER sentiment scores during the holiday season, with a sharp decrease in early 2020, which flattens. The longer trend (12 month line) shows a gradual improvement to tweets sentiment over the period of investigation; this is indicated by the light blue line continually trending slowly up.

10.2.3 Definition of training and test data

The following list sets out the key principles and techniques used to structure the modelling data set, and the method to allocate the data into training and test data sets:

- **Test/train spit.** Two thirds of the data is use as training data, which ended up being 68% of the data, to ensure the data could be factorized for use in the Keras framework.
- **Data range.** Training data starts in January 2000 and ends in March 2014. Test or

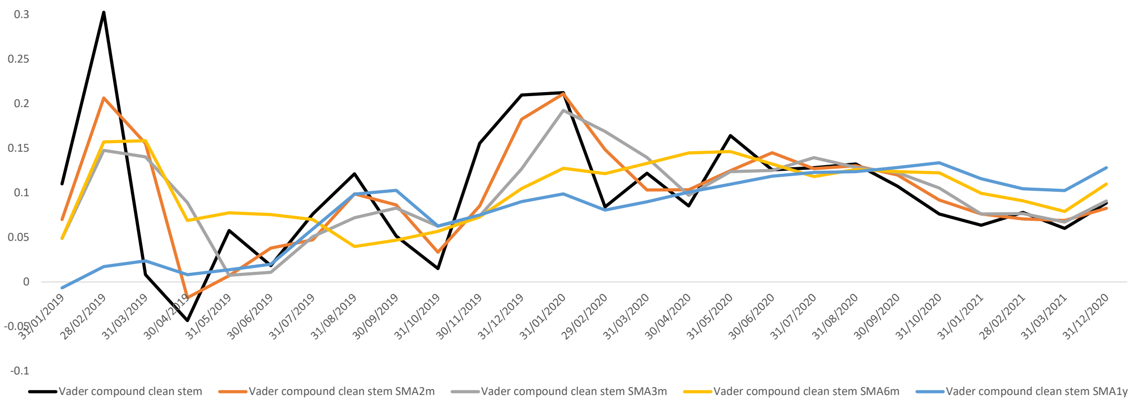


Figure 28: Twitter VADER scores, including rolling average scores with trailing windows of 1, 3, 6, and 12 months. Interestingly, the Twitter data (as defined by the Twitter handles listed in the appendix 26.4) does not show a significant change over the period of investigation.

validation data starts in April 2014 and ends in February 2021.

- **Predicted variable horizon.** Target variable is set as the next month's sentiment index value, t+1 months.
- **Trading window.** To keep a realistic trading window, ten days prior to sentiment variable change, I filter out these rows of news/tweet data.
- **Factors.** Simple factors were created based on sentiment scores from VADER. Factors included 30 day. rolling percentile scores (25^{th}, 75^{th}), 30 day rolling volatility of score, index of filtered terms, frequency of filtered terms (r index, terms such as default, recession, liquidity shock BUT not using pandemic nor coronavirus).
- **Prominent factors.** The r-index and 75^{th} percentile VADER score, and mean VADER score.
- **Goodness of fit.** The following assessment criteria have been used to assess the predictive power of models in this section.

  ◇ **Visual inspection** is a significant secondary review criterion, with a particular emphasis on the test predictions timeliness of sentiment direction shifts and on dramatic negative events (including corona virus pandemic and the GFC).

  ◇ **Review metrics.** Alexander (2008a) explain that distance metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) commonly used for assessing both the in-sample model fit and out-of-sample predictions, or forecasts, for time series models. Both measures produce negatively oriented scores, and as such, a model with a lower score produces a better result. The metrics are defined, following Alexander (2008a):

$$\begin{aligned} \text{RMSE} &= \sqrt{T^{-1}\sum_{t=1}^{T}(r_t - \hat{r}_t)^2}, \\ \text{MAE} &= T^{-1}\sum_{t=1}^{T}|r_t - \hat{r}_t|, \end{aligned} \tag{116}$$

where r denotes the realised target variable and $\hat{r}$ the predicted target variable. As I can see from equation 116 above, the RMSE function squares residuals, which, relative to MAE, punishes larger variances in the scored metric and is therefore always expected to be larger than the MAE metric. This difference in results can provide further insight into the model residuals. So, when the residuals have a large range of values, the greater the difference will be between the RMSE and MAE metrics. If the residuals are uniform in size (not likely), the RMSE and MAE will be equal.

To model economic variables (such as GDP, employment, sentiment, and price indices), one must be cognisant of data time series exhibiting sluggishness or inertia, which manifests as autocorrelation, (Gujarati, 1999). According to Gujarati (1999), the effect of business cycles is that variables will exhibit 'momentum'. RMSE and MAE can also be utilised to evaluate the autocorrelation of the anticipated variable as well as the empirical autocorrelation of the time series data.

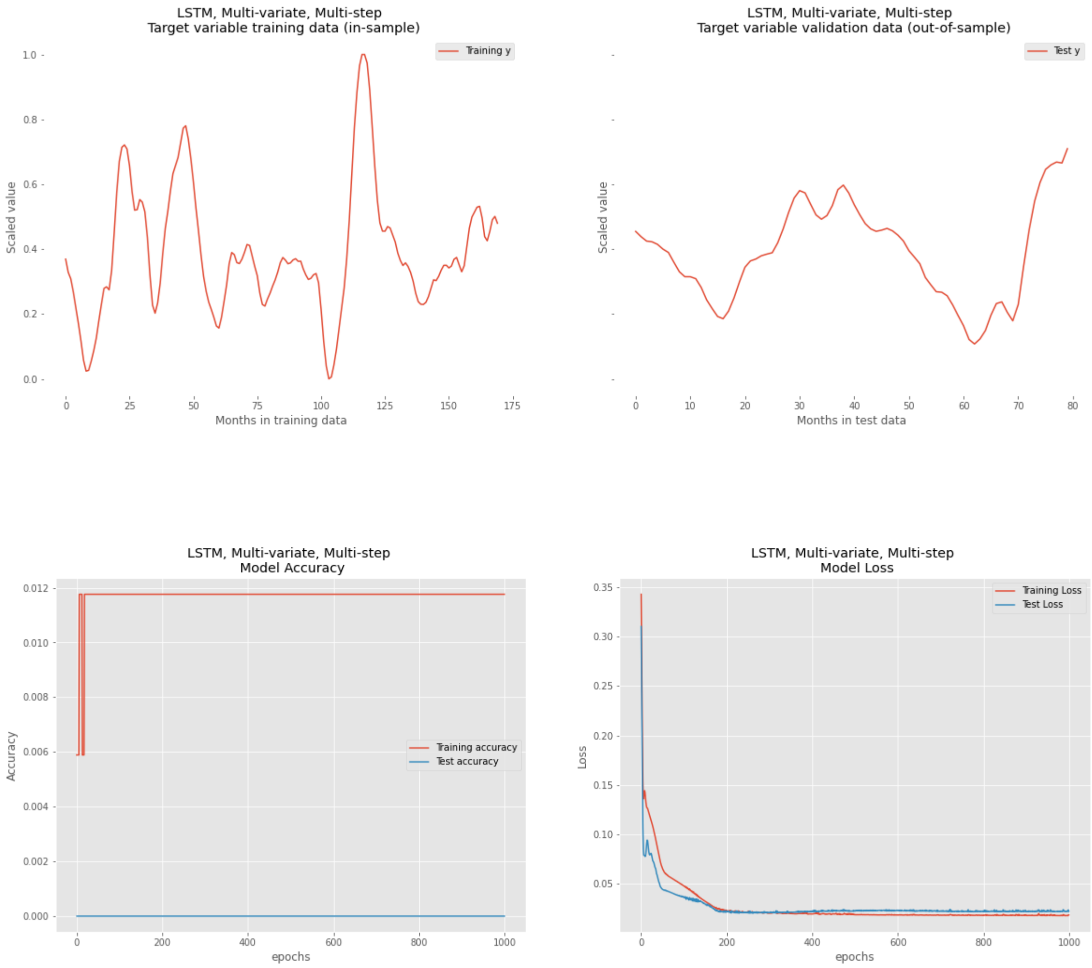


Figure 29: Training and test review are also enabled by the design of a set of standard charts. The first set of charts above is a plot of the target variable, and beneath it is a set of model performance and model loss charts that plot the results from the modelling of each epoch.

10.3 Machine-learning model parameters and architecture

The following table is a summary of the modelling techniques, inputs and technologies.

Model features	Machine-learning
Model objectives	Test the effectiveness of using alternative data streams, (including news and social media) and utilising more sophisticated machine-learning techniques to forward predict globally recognised sentiment indices.
Variables	Independent: text data for each of the data sources, factors derived from VADER and Textblob scores such as rolling average (1m, 3m 6m, 1 year), volatility of VADER/Textblob scores, 75^{th} and 25^{th} percentile of rolling sentiment, article count based on filters (such as search for the word 'depression') and ARIMA terms such as prior month target. Dependant: Market Sentiment proxied by OECD sentiment indices (CCI, BCI and CLI), modelled using month end data.
Data source	News based data (The Guardian and New York Times), GDELT, Social media, Twitter and Sentiment scores from lexical models.
Tools	Python, VADER, Google Colab, Textblob, Keras modelling framework, statsmodel package.
Models tested	Neural Networks, Recurrent Neural Networks (Long-Short-Term-Memory) models; Recursive Neural Networks, and Convolutional Neural Networks, ARIMA model.
Lagged variables	Single-period models are as such and not lagged, but multi-period models are lagged for 3 months of data using all independent variables. This decision was based on an inspection of the p-ACF and ACF numbers of lags that feature autocorrelation.
Hyper-parameters	
Activation function	Sigmoid
Optimisers	Adam
Epochs	1000
Train/Test split	68%, database size symmetry dimension that 66.667% does not allow for, using 68% worked.
Testing framework	
Visual	Graphical analysis of the out-of-sample test results is plotted as a time series of the actual versus the predicted value. Autocorrelation is identified and managed using the autocorrelation function (ACF) and the partial ACF.
RMSE	Root Mean Square Error, negative parameter which requires selecting models with lowest value of the out-of-sample test results. A proposed 6% RMSE and 5% MAE will be used to classify good models from models that would not be selected.
MAE	Mean Absolute Error, negative parameter which requires selecting models with lowest value of the out-of-sample test results.

Table 7: Sentiment models description, objectives, variables and data sources.

The following report is a mix of model inputs, model architecture summary, and accuracy statistics that have been compiled using Python and Keras, which I stored for a detailed archive of modelling results and metadata.

```
Training and test data dimensions:
Train X: (170, 6, 7) Test X: (80, 6, 7) Train y: (170,) Test y: (80,)
6/6 [==============================] - 0s 3ms/step
Timestamp: NN modelling completed on 29/07/2021 at 08:36:05
Mode reference: "sequential_6A"
_________________________________________________________________
Layer (type)   Output Shape  Param #
=================================================================
conv1d_1 (Conv1D)  (None, 5, 64)  960
_________________________________________________________________
max_pooling1d (MaxPooling1) (None, 2, 64)  0
_________________________________________________________________
flatten_1 (Flatten)  (None, 128)  0
_________________________________________________________________
dense_12 (Dense)  (None, 50)  6450
_________________________________________________________________
dense_13 (Dense)  (None, 1)  51
=================================================================
Total params: 7,461 | Trainable params: 7,461 | Non-trainable params: 0
_________________________________________________________________
Model Name: NN, Multi-Factor, Multi-step completed: 29/07/2021 08:36:05
X Variables included: 'VADER_compound_clean_stem', 'VADER_compound_clean_stem_75th_filtered',
'Article_count_30_filtered', 'GDELT_weighted_tone_augmented', 'Article_count_30_filtered_scaled,
'OECD_CLI_MEI_GBR2_scaled'

Target variable (Y): OECD_CLI_MEI_GBR2_scaled
Epochs = 1000 | Timesteps = 3 | Train/Split = 0.68
Train Accuracy [Root Mean Squared Error]: 0.023
Test Accuracy [Root Mean Squared Error]: 0.069
Train Accuracy [Mean absolute error]: 0.036
Test Accuracy [Mean absolute error]: 0.049
Saving model results: 29/07/2021 08:36:06
```

The example results above are of a testing model that was executed on July 27, 2021. It was based on a CNN architecture and predicts the UK CLI index (OECD CLI MEI GBR) using multiple predictive factors.

10.4 Modelling results, testing and accuracy

I detail the results of the modelling to understand the most appropriate models and data sources.

10.4.1 Results following a progression in modelling technique

To fully understand the relative strengths of modelling techniques, combinations of data sources, and hyper-parameters, the following exercise is based on running iterations of the models using planned scenarios (in the same style as a grid search with multiple models). The model types are a combination of single factor with a single time step, multi-factor, and multi-step. The multi-step feature has been established in a prior analysis. The analysis was performed on time steps ranging between two and six months. The modelling results showed an improvement to the RMSE and MAE results by increasing the time steps to three months; thereafter, the results did not improve. The autocorrelation function (ACF) validates this conclusion. In summary, the ACF plot is one of a set of common methods for identifying autocorrelation characteristics in data. The ACF measures the coefficient of correlation between two successive values in a time series and illustrates the relationship between time-series data and its lag variable. The chart is a device for identifying autocorrelation characteristics. Autocorrelation is highlighted when lagged variables go outside of the band (highlighted in pink). This procedure requires iteration (using lag terms, differencing, and moving averages) until the autocorrelation effects are tolerable. For this modelling study, a three-period time lag based on the percentage change of the emotion variable was sufficient (the modelling review process requires that each time you make a change to the model variables, the autocorrelation function (ACF) and partial ACF must be re-performed and reviewed). One of the modelling strategies includes the introduction of variables to deal with autocorrelation (ARIMA terms); this was performed to determine whether the inclusion of the lagged target variable enhances the final model scenario.

The final results of each modelling scenario were generated and stored. Considering there was a huge volume of analytics and charts, this amounts to well over 100 A4 pages per scenario. I only selected the summary views that can best answer the questions raised in the study. The following is a summary result view of machine-learning techniques based on visual inspection of in-sample and out-of-sample charts, in Figure 32. In the charts that following, I summarised the results from the basic feed-forward neural network model to the

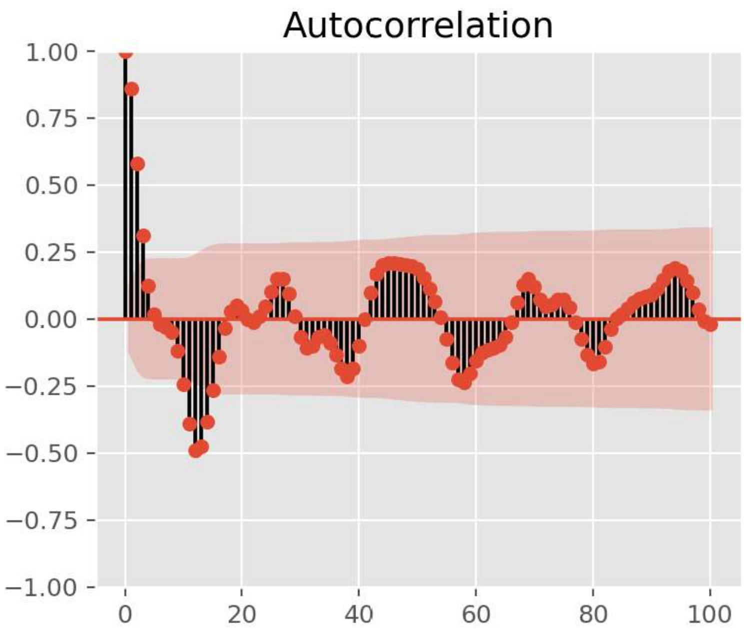


Figure 30: Autocorrelation function (ACF) for the scenario of GBR Guardian modelling. This analysis is predicated on a goal variable representing the standard deviation of the percentage change in the BCI sentiment value.

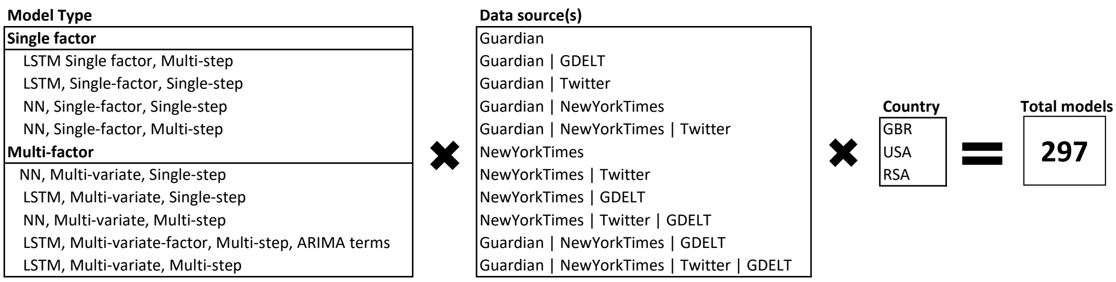


Figure 31: Modelling scenarios are used to generate results from different model types. This requires iterating data source combinations for each region and model types investigated. Importantly, these models are only those that were executed and returned reasonable results (or any results) across all scenarios.

models with more structure such as such as the Long Short Term Memory (LSTM) model, see Figure 32.

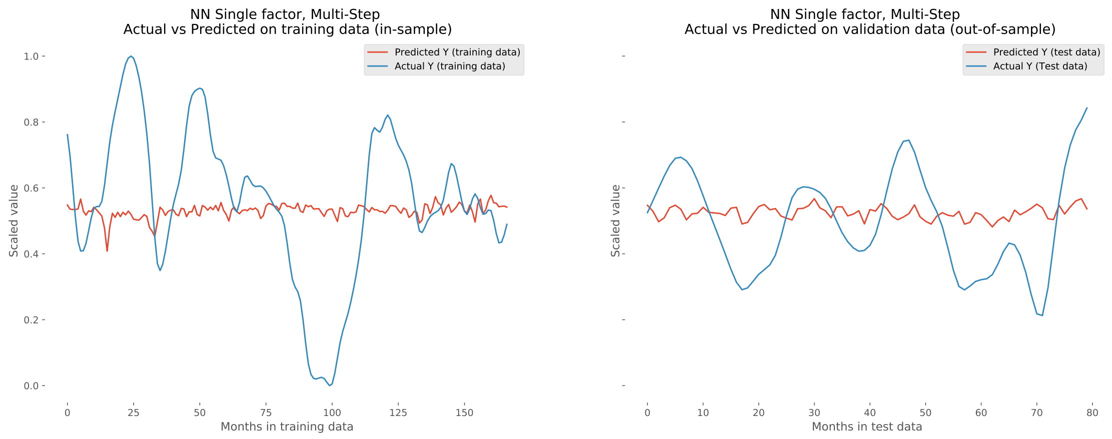


Figure 32: Multi-factor neural network model with no lag variables (timed steps). Visual inspection reveals a model that fails to predict sentiment movements both in and out of sample.

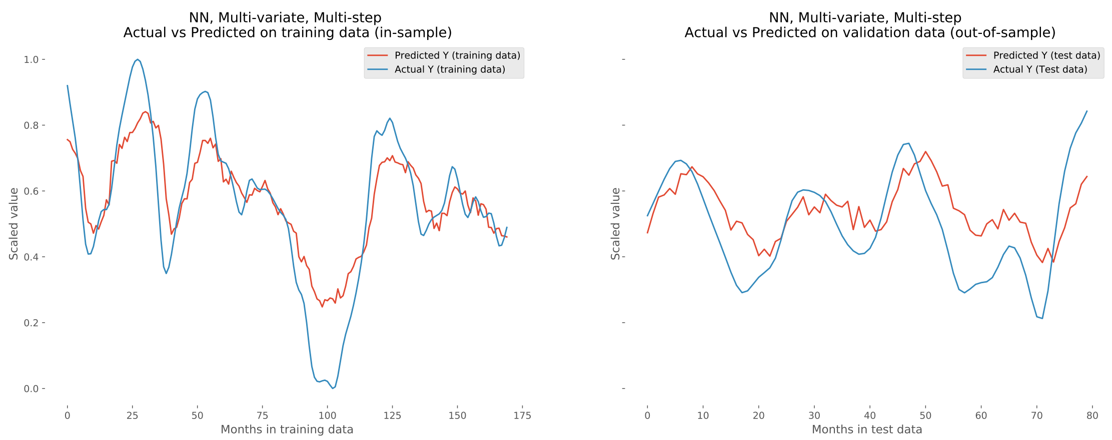


Figure 33: Neural network model with multi-factor and multi-lag variables (timed steps). With reasonably good directional predictions and timing of sentiment changes, the in-sample model shows real promise. The out-of-sample model performance shows results that diverge in more recent time periods, but the directionality and timing of significant shifts are reasonable. Refer to the Appendix 26.9 for a more complete set of modelling results covering the different regions.

10.4.2 Summary modelling results

The first analysis of the results tests the effect of increasing the number of data sources and reviewing them on the category of factors. The results in the bar chart, Figure 35, measure the out-of-sample RMSE of the test database; that is, the out-of-sample data.

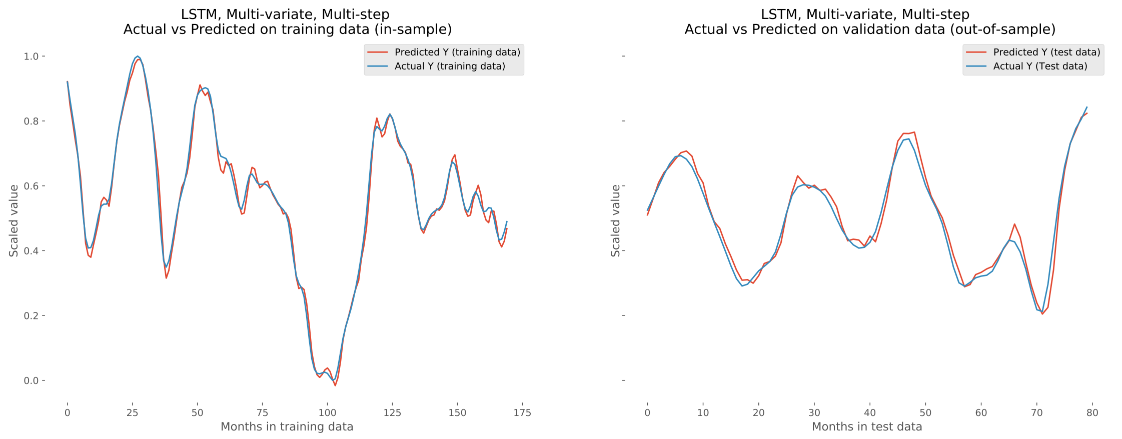


Figure 34: LSTM modelling that includes multi-variate, multi-lagged variables (times steps). The final model I reviewed in this instance has a highly favourable modelling result, with RMSE and MAE lower than 6% and 5% respectively. The timing, directionality, and scale of peak-to-trough are accurate.

It is clear from the modelling result (in Figure 35), that including additional data sources does not improve the RMSE; it actually increases the RMSE on average. By contrast, by including further factors, the modelling results improve significantly, and this is true for each of the three regions. The underlying detail will be further analysed by reviewing each of the respective regions below. I continue the overall summary analysis by comparing the effects of various modelling techniques while maintaining the single- and multi-factor split. This analysis aims to isolate the impact of modelling technique in relation to single or multiple factors in the model, based on out-of-sample RMSE. The analysis reveals what drives large parts of the accuracy in the modelling.

Figure 36 reports the full result set for both the RMSE and MAE, together with the in-sample and out-of-sample data. This analysis has been broken down to show the average results for single- or multi-factor, as well as single- or multi-step model scenarios. As expected, the RMSE results will be higher than the MAE due to the residuals being squared, that contributes a greater weight to the final RMSE in comparison with the same errors contributing to the MAE result. The multi-factor, multi-step models produce out-of-sample errors that are on average close to the 6% RMSE and 5% MAE thresholds, respectively.

The modelling technique has less of an impact than the combination of multi-factor analysis and modelling technique. I show this result in the left part of the bar chart in Figure 37. This result is approximately equal when we review across the modelling techniques. This

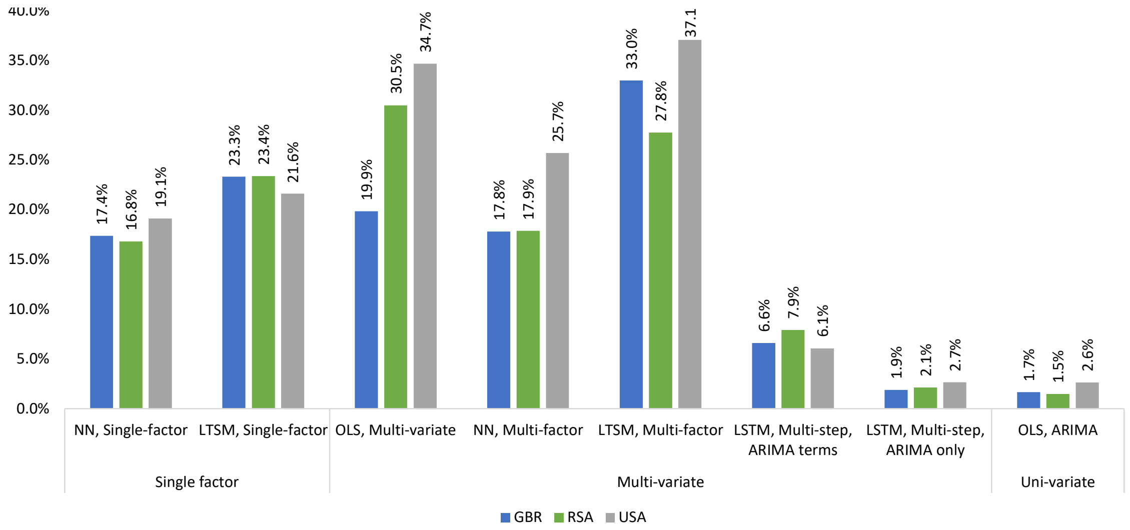


Figure 35: Out-of-sample RMSE test results broken down by factor category, number of data sources and region.

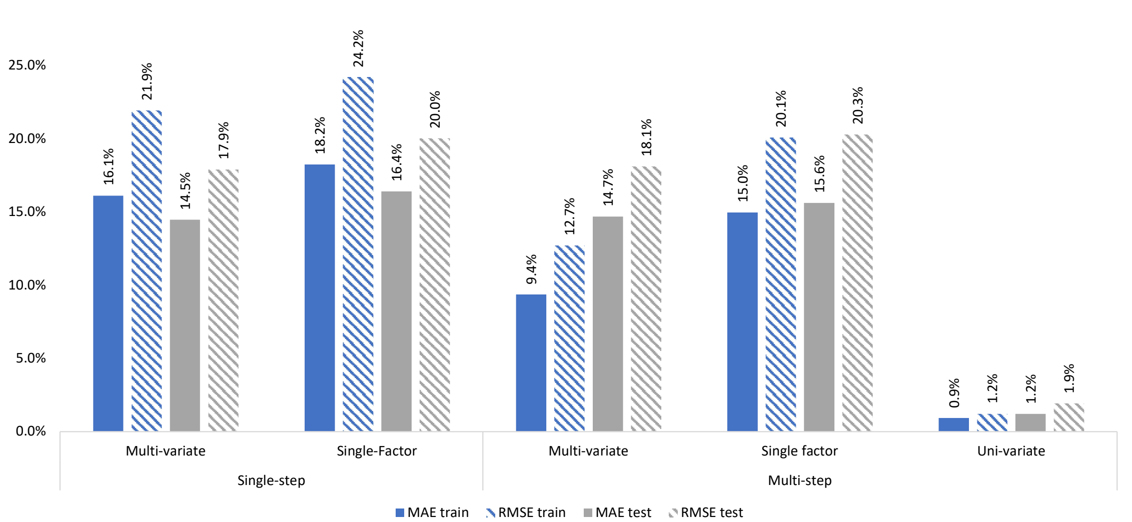


Figure 36: Reports the results of the modelling using RMSE and MAE for both in-sample and out-of-sample, broken down by step category (single or multi-step) and factor category (single or multi-step).

indicates that modelling technique alone has little impact. The results on the right half of Figure 37 show the impact of modelling using different techniques and single- or multi-step time lags. In a multi-factor and multi-time-step model, both the neural network model (circa 40%) and the LSTM model shows a significant reduction in RMSE, with the LSTM showing a significant drop in RMSE of around 40%. The LSTM model with multi-factor and multi-step models produces out-of-sample errors that are greater than the 6% RMSE and

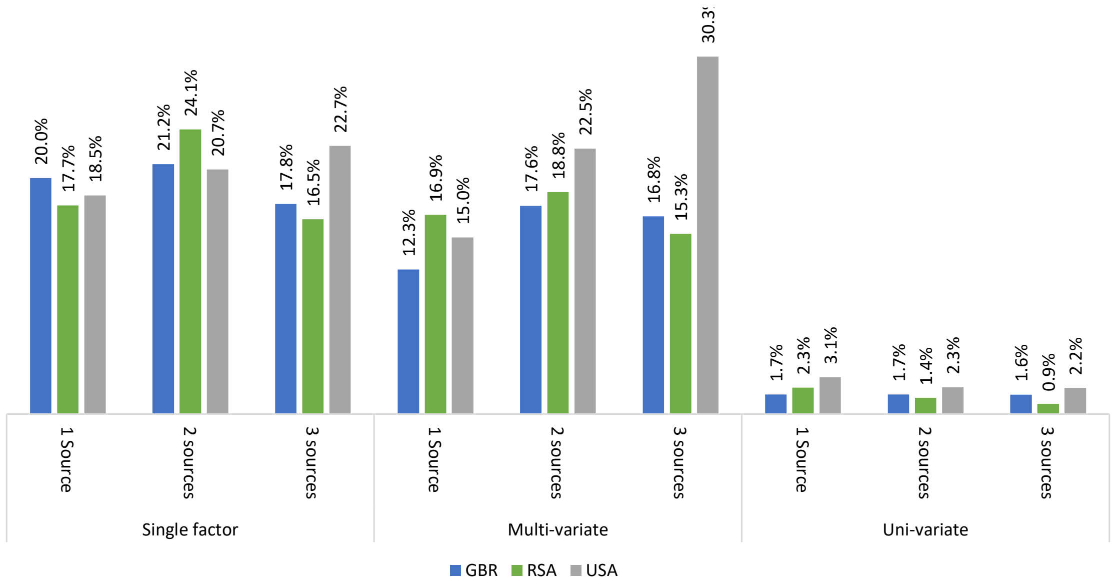


Figure 37: Out-of-sample RMSE test results broken down by factor category, single/multi factor category, model type and region.

5% MAE thresholds, respectively.

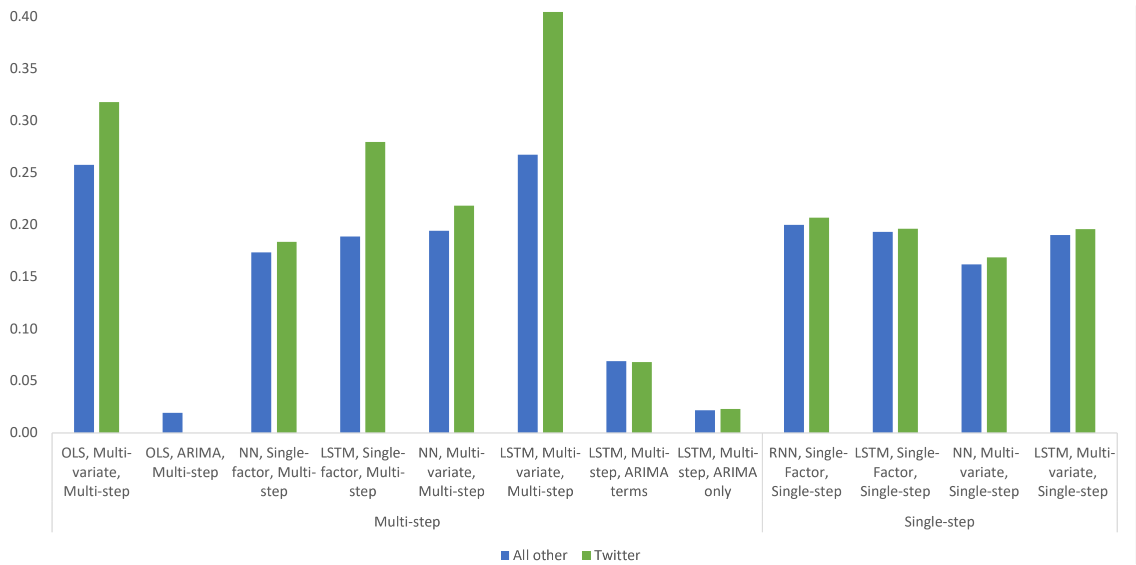


Figure 38: RMSE results where I compare the effects of including Twitter in the modelling data. In all the modelling categories, the results are worse off for including Twitter data. The multi-step model using LSTM is the best performing model that uses the NLP sentiment signal data.

In the next sections, I review the results for each of the specific locations and go into more granular detail around the data sources.

10.4.3 UK modelling results

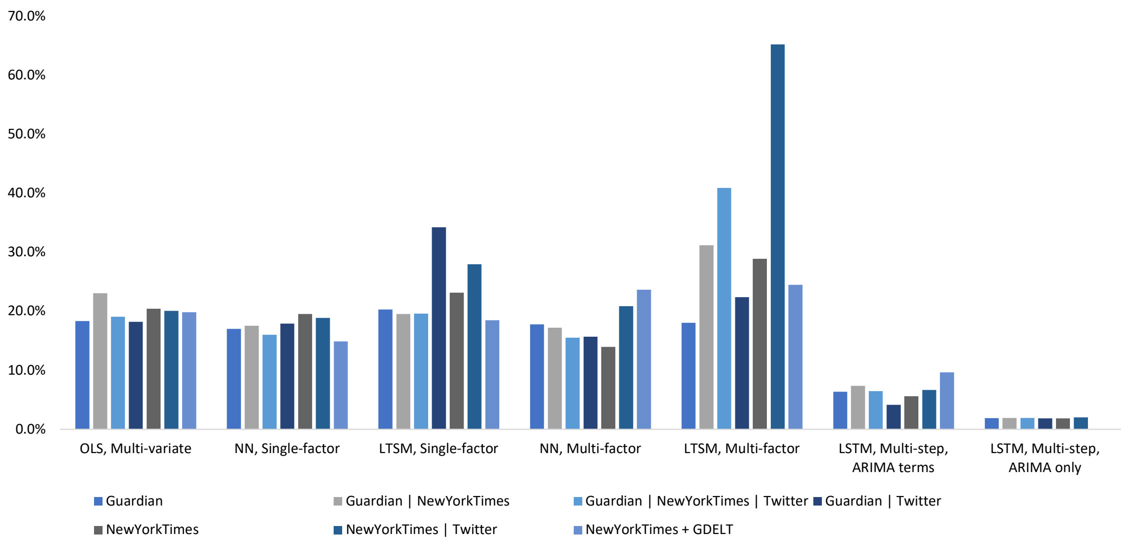


Figure 39: RMSE results from the UK modelling exercise. What is clear from the chart is that LSTM, multi-step, and ARIMA terms improve the results. The inclusion of Twitter data, even with the shorter observation period, only adds noise to the model results.

The following analysis is constructed solely on the basis of the data collected from the UK BCI index. I will be able to generate a more granular review and analytics. In Figure 39 I go over the detailed model types as well as the underlying data sources, both single- and multi-step. When I review the single-step model results, the best results are from those models based on a single news data source. What is also clear is that the inclusion of Twitter data impairs the results (in every case). If I review the multi-step modelling and focus on the best results by model type (LSTM without ARIMA terms), a single new data source again proves to be the best configuration. Surprisingly, the best combinations are The New York Times and The New York Times + GDELT, followed by The Guardian and The Guardian + GDELT. When the Twitter data is added, the result is impaired, and lastly, the combination of more than two data sources also impairs the result. The Guardian's single data source produces the most consistent results. A final observation is that the combination of the two primary news data sources (New York Times and The Guardian) produces a worse RMSE result than using either of the single data sources alone.

10.4.4 USA modelling results

The following analysis is based on the results for the USA BCI index only. I will also be able to generate more granular review and analytics.

In Figure 41 I examine single/multi-step model types and data sources. Single-news data source single-step models perform the best. Twitter data also degrades outcomes (in every case). If I analyse multi-step modelling and focus on model type (LSTM without ARIMA terms), single news data source again performs better. The New York Times and GDELT are the finest, followed by The Guardian and GDELT. Twitter data affects the result, as does using more than two data sources. The Guardian has the most consistent data. Finally, the combination of the two key news data sources (New York Time and The Guardian) yields a higher RMSE than either data source alone.

10.4.5 RSA modelling results

Only the results of the RSA BCI index are used in the following analysis, which you may find below. I will be able to produce a more granular assessment and statistics with the help of this filter. In Figure 43 I reviewed at the underlying data sources too and the detailed model types, separating them into single and multi-step categories. When I look at the results of

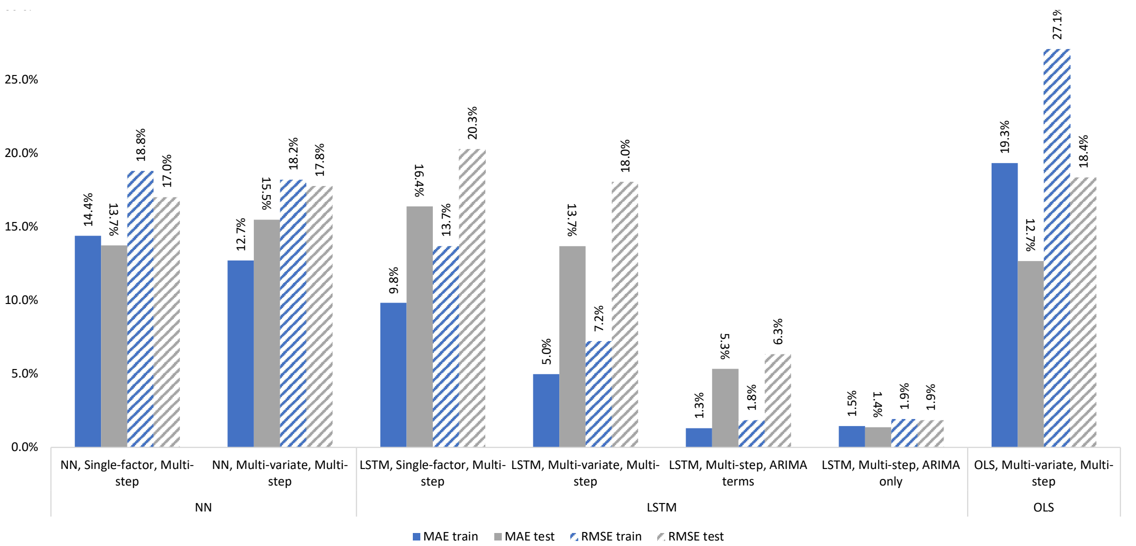


Figure 40: UK in-sample and out-of-sample MAE and RMSE testing results are reported above. What is clear is that the LSTM model, multi-step and multi-factor, with no ARIMA terms, performs best, with RMSE and MAE values lower than the proposed 6% and 5% thresholds.

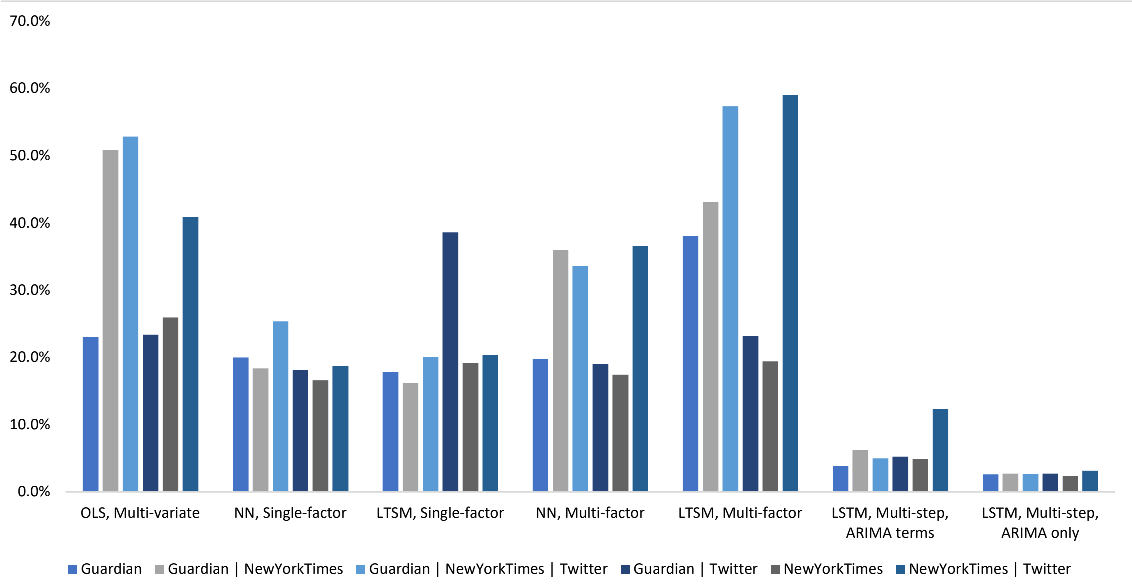


Figure 41: RMSE results from the US modelling exercise. Both the results are included (excluding Twitter data and including Twitter data). What is clear from the chart is LSTM modelling, multi-step and ARIMA terms improve the results. The inclusion of Twitter data, albeit over shorter time periods, provides significant noise to the modelling.

single-step models, I find that the models based on a single news data source produce the most accurate predictions. It is also clear from the results that including data from Twitter brings about a degradation in the outcomes (in every case).

When I look back at the multi-step modelling and concentrate on the best results according to model type (LSTM without ARIMA terms), I find that using a single data source for the news is, once again, the optimal set up. It is interesting to see that The New York Times and The New York Times combined with GDELT produce the best results, followed by The Guardian and The Guardian combined with GDELT. The outcome is degraded each time the data from Twitter is added, and additionally, the use of more than two different data sources also has the effect of degrading the outcome. The results obtained from The Guardian's one and only data source are the ones that are the most reliable. A last observation is that making use of both of the key news data sources (New York Time and The Guardian) together generates a result with a higher RMSE than making use of just one of the data sources on its own.

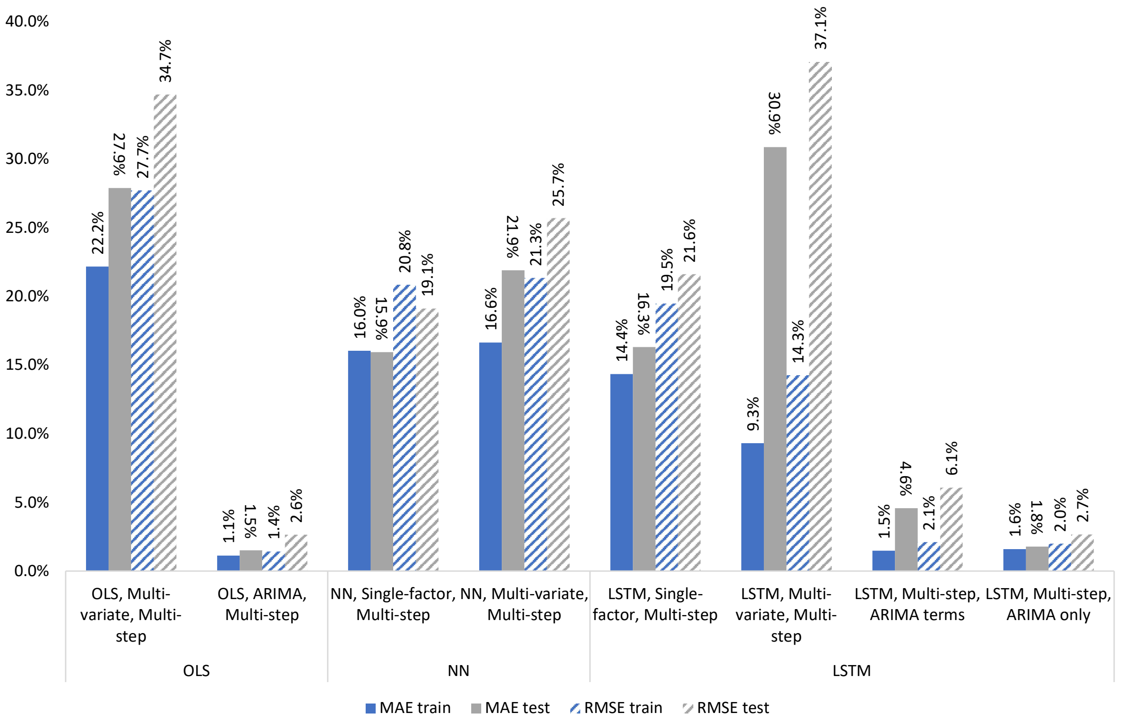


Figure 42: The findings of both in-sample and out-of-sample RMSE and MAE tests in the US are presented up top. Results for both RMSE and MAE are below the specified thresholds of 6% and 5%, respectively, suggesting that the LSTM model, which is both multi-step and multi-factor, without any ARIMA components, performs best.

11 Summary of analysis

This section has been a revealing investigation into the use of alternative data and contemporary modelling techniques that extract macro and fundamental risk factors. In terms of review, I inspected the incremental benefit of incorporating each technique, revealing the true benefit. I separately explored the data types, methods, and how one would use factors to deal with autocorrelation, thereafter reviewing each part to understand its impact. The greater part of this review is investigating the relative benefits of using various sources of alternative data, both reviewing the difficulty in handling the data and the information the respective data sources hold.

When I reviewed the power of alternative data for use in building signals, I found the following. The statistics from the news are easily available and offer a comprehensive history (e.g. The New York Times has provided free access since 1900). According to my analysis,

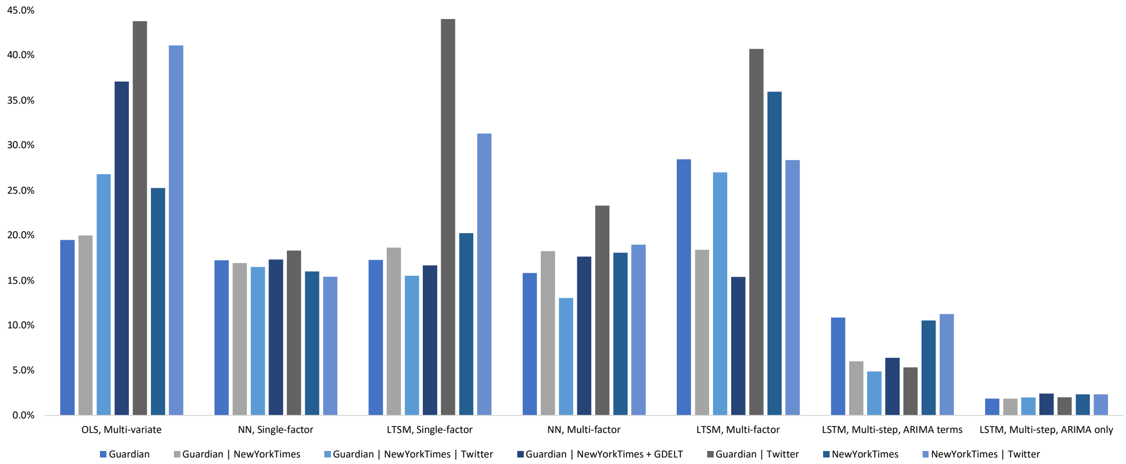


Figure 43: The results of the RMSE exercise that was performed with the RSA model. Both of these results are included (those that do not include Twitter data as well as those that do include Twitter data). The graphic makes it abundantly evident that LSTM modelling, multi-step, and ARIMA terms all contribute to improved results. The inclusion of Twitter data, even if it was collected over shorter time periods, adds a considerable amount of noise to the modelling.

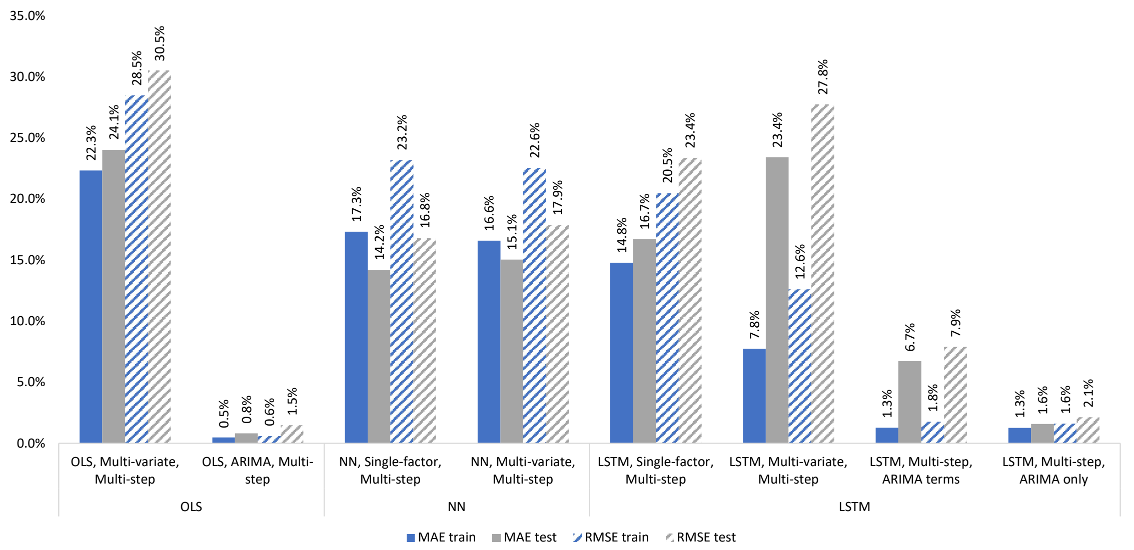


Figure 44: RSA in-sample and out-of-sample MAE and RMSE testing results are reported above. What is clear is that the LSTM model, multi-step and multi-factor, without any ARIMA terms performs the best, with RMSE and MAE results lower than the 6% and 5% proposed threshold.

the information that is presented in the news possesses predictive power. Collecting Twitter data can be a challenging endeavour (streaming data can be obtained without charge, but historical data draws might be greatly slowed down and/or require payment). It was difficult to maintain and clean the data that was based on filtered terms, which resulted in minimal proof of its predictive value (albeit over shorter time periods). The data from Twitter provides a great deal of noise even over the short horizon of investigation. On the other-hand, the signals provided by Google's GDELT has shown to provide a significant dip in measured sentiment during the pandemic that other news sources did not display. News sources provide relevant signals, especially when reviewed within a particular geography.

When it comes to my investigation of modelling techniques for the time-series data, machine-learning models, in particular LSTM, are a good choice. The use of ARIMA modelling in neural network frameworks and lexical outputs as factors are examples of hybrid modelling considerable improvements over machine-learning models used alone. Hybrid modelling strategies, which make use of the judgements of subject matter experts, will produce variables that have predictive value and be sensible to the domain user in question. All said, the best results was based on univariate ordinary least squares model using ARIMA terms (please keep in mind, the purpose here is to consider alternative data, however a benchmark model is of interest too).

I would like to provide a view on the tooling that I used during this analysis by design readily available at no material cost, or free of charge. This was greatly encouraging – I found that quick access to powerful modelling can be accomplished with the help of frameworks such as Keras and Python notebooks. These frameworks enable hybrid modelling systems that are easy to use and are free of charge. Google Colab has a low cost GPU option; prevents the need make a considerable financial investment to purchase GPU hardware.

In summary, the analysis was interesting and revealing. Given the scope of this study and the information available in a full newspaper history, I maintain that I only scratched the surface of its capability. The advent of new technology to wrap up existing techniques into freely available frameworks at a fraction of the cost is compelling. I am still left with the view that unless one has a specific use case against which the extraction of information using NLP technologies and utilizing highly advanced machine-learning packages may need careful consideration. It is still an onerous task to curate the data and extract value from it. I also found that financial data features, firstly autocorrelation, have the most significant overriding impact for modelling. This draws me to a natural conclusion for finance, domain knowledge and understanding of techniques is the real benchmark or departure point for these new techniques and alternative data sources.

Part IV

Private debt and credit-risk measurement

12 Introduction

Anson et al. (2011) define credit-risk as instruments issued by corporations or entities that may default, where outstanding money owed to the investor is at risk of default. These models clearly indicate for an investor to bear financial risk and invest in this instrument they seek additional incentives over and above the risk-free rate of return. The extra yield is usually measured as a spread charged over a reference curve, in the case of a private credit loan the reference rate could be the base lending plus an additional spread. For example, for a bond issued by a corporation, the reference curve is usually the swap curve. Traditionally, the credit portfolio risk can be understood by observing the level of credit spreads. The average spread difference is 1.4%, which indicates that private debt with similar losses would thus carry a high reward. I focus on private debt market that this is not traded in liquid markets and risk are not thoroughly understood. In the following section I will detail the core credit-risk theory and models which will assist me in approach the building out of the most appropriate model for private credit, whilst considering asset-pricing features such as the credit puzzle. I will also analyse the available credit data and start to build the core parameters that are needed for simulation of asset class returns. In this analysis, I also directly investigate the use of the DE-CCI proxy data that I finalised in part III.

12.1 Preliminaries: credit modelling parameters

The Financial Account Standards Board (FASB) for credit-risk and from the statutory perspective, the International Accounting Standards Board (IASB) have also been updated with the introduction of the International Financial Reporting Standards number nine (IFRS9) for the quantification and reporting of loan loss provisions (Engelmann, 2021). Both Basel II and IFRS9, in the three applications of risk management (minimum capital, loan loss provisions, and stress testing), require internally estimated credit-risk estimates (Engelmann, 2021). The three key parameters are loss given default (LGD), default probability (PD), and exposure at default (EAD). Please refer to Table 9 for a good visual definition of TTC and PIT PD.

The Basel II accord is a prescriptive framework for banks that standardise credit measures as part of their risk and capital management framework. This framework directly impacts bank capital requirements and earnings to satisfy regulatory (Basel II) and statutory (IFRS) requirements for products and risk parameters. As these standards define the core building blocks for credit-risk measurement and have been broadly adopted in banks and elsewhere, I list them below.

12.2 Probability of default

Trueck and Rachev (2009) explain that all banks using internal ratings for Basel II use a Probability of default (PD) to estimate the risk for borrower (or obligors[11]) default. The PD defined by Basel II is a conservative long-run average PD estimate, based on historical experience and supported by empirical validation. The Basel Committee on Banking Supervision defines that a default is considered to have occurred with regard to a particular obligor when:

- obligor is considered to not be capable of paying its credit obligations in full, also not by making use of collateral held.
- obligor has not paid (past due) on a material payment by more than 90 days.
- overdrafts or revolving facilities are equivalently past due when a client has been in excess of the agreed limit by more than 90 days.

In financial circles lending and credit-risk have standard definitions now, the likelihood that a borrower will not meet their financial obligations and debt over a defined period of time is known as the probability of default (PD). The time horizon for the PD is an important assumption in credit system core approaches are:

- **Point in time (PIT) PD** - one year estimate of the current expected PD based on current economic conditions. A PIT credit-risk measure takes into account all relevant data as of a specific date in order to capture the PD for a client over a specific period of time.
- **Long-run or through the cycle (TTC) PD** - based on the average long-run PD that take into consideration a full business cycle.

[11] A legal term: a person who owes or undertakes an obligation to another by contract or other legal procedure.

Traditionally, the PD has been the central focus for credit-risk measurement. The advent of Basel II meant that other key measures got formalised for all banks. The following section describe the other key measures, starting with the exposure at default.

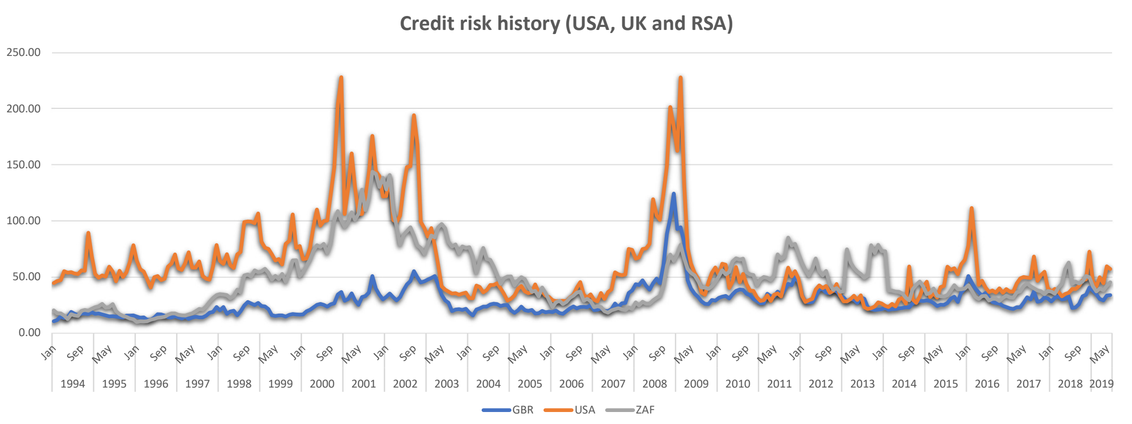


Figure 45: Comparison of PD across country for one year PD.

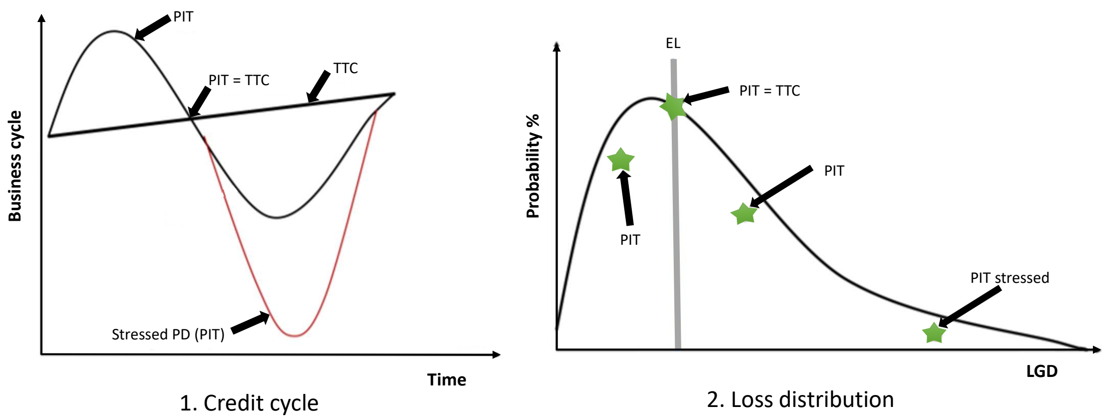


Figure 46: Illustration of a Potential for Default (PD) showing a representation of PD measure. The left panel shows how the PD measures vary through-the-cycle. The right panel explores the same concept by illustrated through the VaR loss distribution.

12.2.1 Exposure at default

This measure was set to account for cash drawdowns an obligor is likely to have made in the run-up to their default. The exposure at default (EAD) is an adjustment to the loan exposure to account for any further drawdowns up to the point of default (Trueck and Rachev, 2009). I make the assumption that the EAD is 100% in this analysis, which is a conservative but appropriate assumption.

12.3 Loss given default

The loan loss given default (LGD) metric is used to determine the extent of loss that the bank will incur if the loan goes into default. Losses are not always linear with respect to loan-to-value for the LGD; the loss is largely related to contract-specific characteristics that include the value of the underlying collateral, the loan-to-value ratio, and the seniority of the claim. In times of stress, the market liquidity of the collateral can heavily influence the resultant losses, we know this from the experience in GFC. The LGD might be any value from 0% to 100% if no workout fees have been incurred by the bank. There are various types of LGDs; I list a few of the recognised methods:

- **Implied market** is based on the credit spreads have been measured and reported in the market place.
- **Market** on the other-hand is based on the residual value of market debt instruments after default.
- **Work-out** is based on the actual residual value after accounting for the sale of collateral held coupled with the cost of the recovery process.

According Araten et al. (2004), the benefit of a work-out recovery is that the true value of recovery is revealed, whereas the market LGD is further influenced by demand and supply, liquidity and risk aversion, and taxes. Because the market LGD has been found to systematically underestimate the true recovery rate, standard market practice gives the LGD three ways to account for the business cycle (again, recognising the LGD's instability over time). The three methods are as follows:

- **Point in time (PIT)** - a one year estimate of the current expected LGD based on economic conditions.
- **Long-run or through the cycle (TTC)** - based on the average long-run LGD that take into consideration a full business cycle.
- **Down-turn (DT)** - calibrated from facility and loan loss data from stressed markets.

Trueck and Rachev (2009) explain that senior claims start with the LGD value of 45% and for subordinated exposures this value is closer to an LGD of 75%. These values will then be scaled in accordance to collateralization levels, with separate levels used for real estate. In a perfect world one would have an extensive modelling data set, rich in feature to support LGD parametrisation and then specify an advanced multi-variate LGD modelling approach with own loss data. This approach can be used to calibrate at an exposure level, factors may include sub-industry, product or legal characteristics and geographical location (Trueck and Rachev, 2009). Keijsers et al. (2018) point to studies and loss data that show the variation over the cycle. Cycles can affect bank capital requirements by more than a factor of two, which is significant. The most significant portion of this change is driven by the LGD fluctuations over the cycle.

13 Rating and risk modelling techniques

The development of credit modelling began with the first models to aid in credit judgement with active signals of potential default, followed by models such as Altman's Z-score (Benzschawel, 2017). In this development, there was an advancement in real-world modelling that was not as useful to broker dealers to hedge their credit exposures, as they did not have the correct data to calibrate models. Benzschawel (2017) explains that due to a general lack of data to build structured models, practitioners developed an easy way to apply the general credit-risk framework by using a simplified reduced-form modelling framework (Benzschawel, 2017). In the following sections I summarise key models and frameworks used in credit-risk modelling.

Numerous studies and commercial models (including reduced-form models like CreditRisk+, structural models like the Merton model, and banking scoring models) have been developed to assess the risk in a credit portfolio. As a result, we know that the risks associated with a credit loan portfolio fluctuate over time and resemble the credit cycle. Both business results and overall economic activity affect the credit cycle. Recent studies have concluded that the collateral levels and recovery values are not independent from the credit cycle (Moody's (2010), Han (2019) and Fischer et al. (2019)). Exciting research is being conducted on news-based projections of macro-economic indicators (Feuerriegel and Gordon, 2019) and variable selection strategies for credit recovery models (Nazemi and Fabozzi, 2018). This is a brief overview of credit modelling technique and it is by no means exhaustive, I grouped them into four distinct classes of credit models.

13.1 Rating agencies

Trueck and Rachev (2009) explain the rating agency services had a recorded start in 1860, with formal debt and government ratings. The first ratings used to be allocated for a 75+ year horizon. In today's finance world, rating agencies are designated and referred to as nationally recognised statistical rating organizations, which include such well-known names as Moody's, S&P as well as Fitch (there are more, but not covered here). Trueck and Rachev (2009) note that S&P and Moody's have very similar rating approaches. The ratings are issued with specific rating bands that can be summarised as investment grade or sub-investment grade. Corporates will generally approach a rating agency prior to a new debt issue, where rating agencies provide two primary ratings:

- **Issuer credit ratings**: These ratings are an assessment of the overall capacity to meet obligations
- **Issue-specific credit ratings**: are assessments on a specific obligation, whether the structure and cash-flows and obligors will meet the obligation with a measure of rank, or rating category

Differences between the issuer and issue specific ratings are referred to as 'notching'. Where security issues are said to be 'notching' up relative to the issuer rating. Trueck and Rachev (2009) add that ratings are based on a lot of underlying measurement factors, set into two categories: firstly, financial risk (financial factors, policies, profitability, structure, protection of cash flows, and agility) and secondly, business risk (including, industry factors, competition, marketing, technology, efficiency, regulation, and management). Other factors are captured by sovereign, currency, and emerging markets factors, which are considered (Trueck and Rachev, 2009).

Alexander (2004) explains that listed debt is rated by public rating agencies. There are three main agencies that supply ratings for listed debt instruments. The ratings include details on the quality of the debt and the obligor's ability to service that debt. The role of the rating agency as an independent party that lowers information barriers caused by information asymmetry (the company has more information about its ability to service debt than the prospective investor) is critical. I am only focused on private credit, so this brief introduction be enough to place context around the modelling later on in the chapter.

Rating agency scales mapping		
Fitch	**S&P**	**Moody's**
Investment grade		
AAA	AAA	Aaa
AA+	AA+	Aa1
AA	AA	Aa2
AA-	AA-	Aa3
A+	A+	A1
A	A	A2
A-	A-	A3
BBB+	BBB+	Baa1
BBB	BBB	Baa2
BBB-	BBB-	Baa3
Sub-Investment grade		
BB	BB	Ba2
BB-	BB-	Ba3
B+	B+	B1
B	B	B2
B-	B-	B3
CCC	CCC	Caa
CC	CC	Ca
C	C	C
D	D	D

Table 8: The above ratings scale comparison table between Moody's, S&P and Fitch long term ratings

13.2 Scoring models

Linear probability models, logit models, probit models, discriminant analysis, NN and, more recently, ensemble models covered in the credit scoring approaches of Trueck and Rachev (2009). Interestingly, neural network models are not new and have been used in credit scoring since the early 1990s, but continue to be hampered by the fact that attribution to the factors that drive credit scores is not systematically possible (Trueck and Rachev, 2009).

Early credit scoring was formalised by Altman (1968), this method was to score a company based on its financial ratios (including sales/total sales, retained earnings/total earnings and market capitalisation / debt) on a Z-score. Large retail and commercial banks have made use of scoring techniques in the sanctioning process for a long time, at varying levels of sophistication. The advent of Basel II formalised the modelling of credit-risk, in particular for large homogeneous portfolios such as retail credit loan books, which has become reasonably standardised. There are various methods used, but the ratings are most commonly established by using factors that predict obligor default in a Logit or Probit model, which is then mapped to a rating table.

13.2.1 Probit models

The Probit model's method can be thought of as being analogous to that of a multivariate non-linear function, where Y is the dependent variable and X is the independent variables or vector of variables, following Trueck and Rachev (2009):

$$P(Y = 1|x_1, \dots, x_k) = f(x_1, \dots, x_k). \tag{117}$$

The Probit model makes slightly different assumptions:

$$P(Y = 1|x_1, \dots, x_k) = \Phi(x_1, \dots, x_k) \tag{118}$$

where p_i is the default probability for obligor i. the difference in the Probit function is that it is based on a cumulative standard normal distribution (Φ), that transforms data in the regression into an interval space [0,1], a handy result for credit modelling as we are solving for PD range is between [0,1].

$$\text{Probit} \longrightarrow p_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_1,\dots,x_k} e^{-\frac{z^2}{2}} dz \tag{119}$$

13.2.2 Logit models

Logit models, or logistic regressions, are quite similar, where the mapping, or mathematical transform, of the same factors is via the Logit transformation and calibrated using this method in a process called a logistic regression, with β loadings calibrated from historical data of X, following Alexander (2004):

$$\text{Logit} \longrightarrow p_i = \frac{1}{1 + \exp(\beta_0 + \beta_1^i X_1, \dots, +\beta_n X_n^i)} \tag{120}$$

Trueck and Rachev (2009) explains that the Logit and Probit functions will produce results that are similar, in particular in rank. There will be differences in the extreme tails, though. The interpretation of the Probit results is easier in that the coefficients are already in a metric to directly compare with the standard normal cumulative distribution. For the Logit function, the regression results for company i must be translated into a prediction of defaulted status ($y = 1$) or non-defaulted status ($y = 0$). This is achieved with a default threshold K, where:

$$Y_i = \begin{cases} 0 & \text{if } z_i \leqslant k_1 \\ 1 & \text{if } z_i > k_1 \end{cases} \tag{121}$$

this approach can be extend to a series of categories and this can be extended to rating categories so as to align to rating agencies rating buckets.

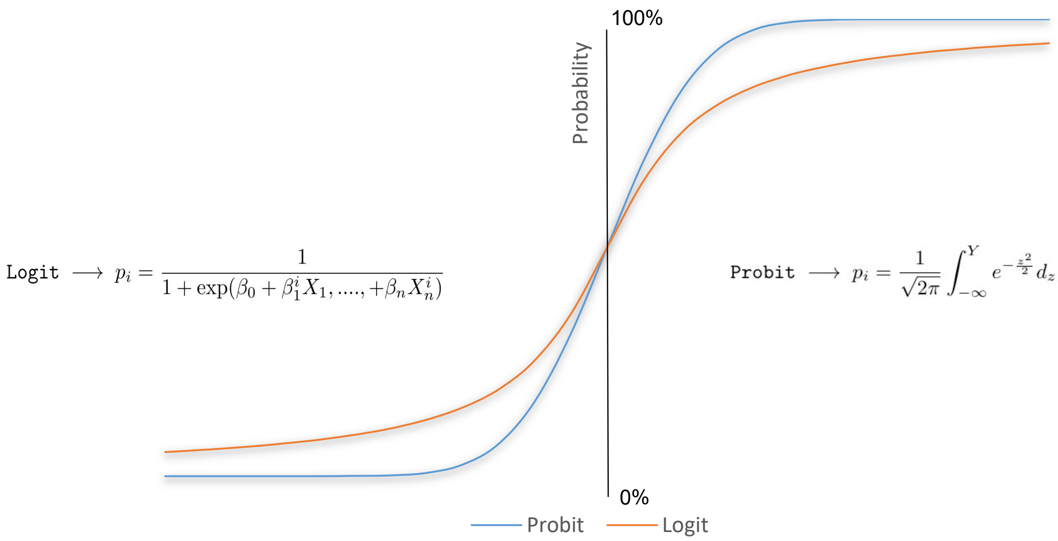


Figure 47: The graph above depicts the Logit and Probit cumulative distribution, highlighting the distinctive S-curve that allows for a ranking area that shifts from one binary value to another over a short range in the x-axis. The illustration is set to demonstrate the s-shaped curve and the relative difference in the tails between the two approaches, where the Probit model has wider tails.

13.2.3 Z-score (Altman)

Altman explored the use of a multivariate regression, using a multiple discriminant approach. Altman's Z-scores were calibrated on historical data that included company defaults using a discriminant model, which was readily adopted in the 1980's. The model solves for linear combinations using the following equation, per Trueck and Rachev (2009):

$$Z = w_0 + w_1 X_1, \ldots, w_n X_n \tag{122}$$

Z is the discriminant score, often called the Z-score. w_0 is the constant and the remainder of the w-weights, $X_i (i = 1, \ldots, n)$ are coefficients that minimise the error in the regression, but are in practice used to 'load' the independent variables or factors, X in the equation. Altman (1968) discovered that bankrupt firms that defaulted within a year can be ranked by looking at the following key accounting ratios:

- Working capital/total assets ($w_1 = 0.012$)
- Retained earnings / total assets ($w_2 = 0.014$)
- Earnings before interest and taxes / total assets($w_3 = 0.033$)
- Market value of equity/book value of total liabilities ($w_4 = 0.006$)
- Sales / total assets ($w_5 = 0.999$)

The model was used for credit lending decisions, where the Z-score was compared to some cut-off level (originally 1.8), with higher scores implying better credit worthiness and lower scores implying difficulty in meeting the cut-off criteria for lending.

13.3 Structural models

In the following section I provide a brief overview of structural credit-risk models. Unlike in reduced form, these models are defined by their view on how to calculate the PD for a firm based on the value of its assets and liabilities. So certainly targetting corporate exposures where the modelling technique requires lots of data.

13.3.1 Merton model

These credit-risk models, often known as 'Merton' models, are characterised by their extensive reliance on assumptions about the company's financial statements. Default is treated as an endogenous variable in the models, with its eventual value tied to the firm's balance sheet. The approach is predicated on an appreciation of a company's balance sheet's resilience under pressure. To put it another way, this is the level of the company's assets below which the company will be unable to continue operations, and this level is established as a proxy for the point at which the company will default.

Modigliani and Miller (1958) laid the groundwork for what would become the famous Merton model by providing the necessary bridge between debt and equity. Investor indifference toward debt and asset financing operations, as described by Modigliani and Miller (1958). Following on from this work, Black and Scholes (1973) established their well-known options theory, which shows how a firm's value relates to its total debt.

Structural models are the first type of credit models, and these models all stem from Merton's seminal 1974 paper as it connected how equity and debt levels are related to the value and volatility of a firm's assets under specific circumstances. Many subsequent models relied heavily on this assumption and approach. The core underlying theory that the Merton model is based on is option theory, as it is applied to a specific single firm's value of assets and liabilities. This analysis then models the contingent claims on the firm. The equity value of the firm can be thought of as the value of a call option on the firm's assets, and this option value can be used to calculate the firm's value. This is known as a default when the firm's assets are worth less than its liabilities. Moody's famous commercial KMV model, which is rooted in Merton theory, is widely used and well-known. In this scenario, the default probability is assumed to be proportional to the firm's balance sheet assets value, structure, and volatility (Alexander, 2004).

As a result of the development of Moody's KMV, also known as the CreditEdge model, I now have a way to correlate the time to default with the likelihood that default will occur, known as the expected default frequency (EDF). The function was based on segmenting firms into distance-to-default buckets and the default experience for years n=1 to n=5 and placing a floor and cap of 0.05% and 20% for default mappings, respectively. Moody's KMV has been a huge commercial success. The one-factor Merton model is based on asset returns, r_i per borrower i, that are fully described by a single risk factor for each firm. To begin with, the Merton model posits that a zero-coupon bond represents the value of debt with face value D that matures at time T, and models the asset value of the firm as S_t for each time period t. As soon as $D_t \leqslant E_t$, the obligor's asset's value, X_t, becomes negative. Assuming a geometric Brownian motion (GBM) model, this approach treats the obligor's asset value, A_t, as stochastic. This relationship is denoted in a stochastic differential equation:

$$dA_i = \mu A_i dt + \sigma_i A_i dx_i \tag{123}$$

where x is a Wiener process or Brownian motion. Solving this stochastic differentiable equation (as described in Hu and Liu (2004)), showing the value of firm A asset as, following Chatterjee (2015):

$$A_t(T) = e^{A(0)+\mu_i T - \frac{1}{2}\sigma_i^2 T + \sigma_i \sqrt{T} X_i} \tag{124}$$

If a firm is said to default when assets are less than debts values, $A_i(T) \leqslant B_i$, then the probability of such an event is, again following Chatterjee (2015):

$$P[A_i(T) \leqslant B_i] = P[X_i < c_i] = N(C_i) = p^* \tag{125}$$

where N is the cumulative distribution function and is a field derived from (124) above, this is an indication of the probability of a single obligor default. If we consider more than one obligor at a time, we impose a correlation structure between obligors. We assume that X_i is correlated with correlation coefficient ρ. Thus a firms value can now be seen as a combination of an idiosyncratic risk factor ε_i and another single factor, the systemic risk factor Z that is linked to exogenous shocks. The systematic risk factor and asset return have a correlation, denoted by ρ. c is the defined level of asset value, that can be shown to be a function of the TTC PD and default threshold is defined as default when $X_i \leqslant c$. Given that the firms are assumed to be normal and similarly correlated, X_i can be represented as the sum of S and Z that represent the systematic and idiosyncratic component respectively and I detailed this model below:

$$r_i = S_t\sqrt{\rho} + \sqrt{1-\rho_i}Z_i \tag{126}$$

where $\sqrt{1-\rho_i}Zi$ can be seen as the firm's idiosyncratic risk and $\sqrt{\rho_i}Z$ is a firm's exposure to the systematic risk factor, that is assumed to be normally distributed, $Z \sim N(0,1)$.

To summarise, structural models have proven to be invaluable in both academia and industry. There are a few issues though, which include the required processing power and expense of the models; an overly simplistic view based on debt and equity (which misses important underlying events such as M&A activity) and the extent of the information required to support the models. Benzschawel (2012) explain that the Black and Scholes model is so detailed, with so many data attributes, it is difficult to support with real data. The data to support key measures are not directly observable, notably firm asset values.

13.3.2 CreditPortfolioView (McKinsey)

Another reduced-form model is CreditPortfolioView, this model can be classed a hybrid macro-simulator. The modelling approach has a credit transition matrix at the centre that

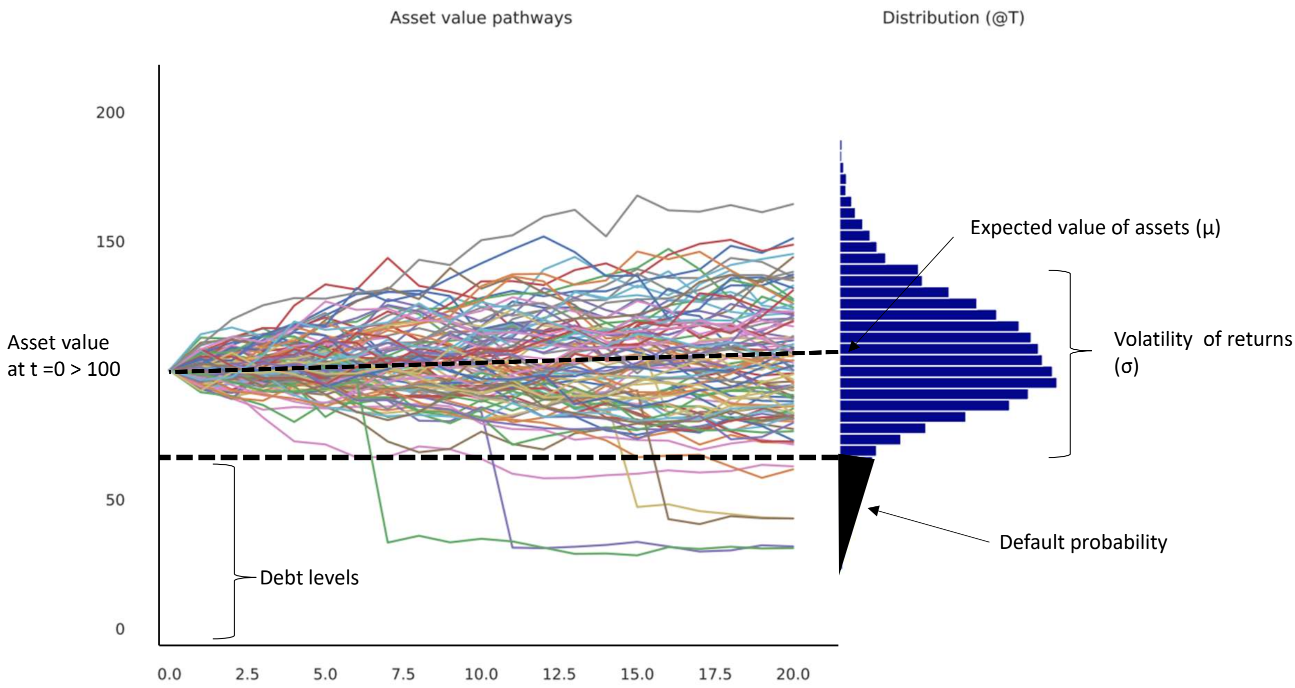


Figure 48: Merton's model is based on stochastic asset values that are simulated over a holding period. These are compared to a reference debt level, and the distance from that debt level is modelled by using underlying data points, which are plotted as a histogram on the right part of the illustration. This is a well-known illustration, however I modelled this data a plotted a chart that is an adaption of Benzschawel (2012) on page 82.

varies through the business cycle. In times of economic stress the output results in default probability increases, this is seen in the data as credit downgrades increase and upward migrations decrease (Alexander, 2004). The model uses a Logit regression model to establish conditional PD denoted as $p_{j,t}$ in period t and counter party grade, j (such as speculative grade), following Trueck and Rachev (2009):

$$p_{j,t} = \frac{1}{1 + e^{-y_t}} \tag{127}$$

The index $y_{j,t}$ is based on a multi-variate regression time-series model, denoted per Trueck and Rachev (2009) as follows:

$$y_t = \beta_0 + \beta_1 X_{1,t} + \cdots + \beta_m X_{m,t} + v_t \tag{128}$$

where X macro-economic variables determine the index y_t and error v_t is the error term. The macro-economic factors are modelled as an autoregressive stochastic process AR(2),

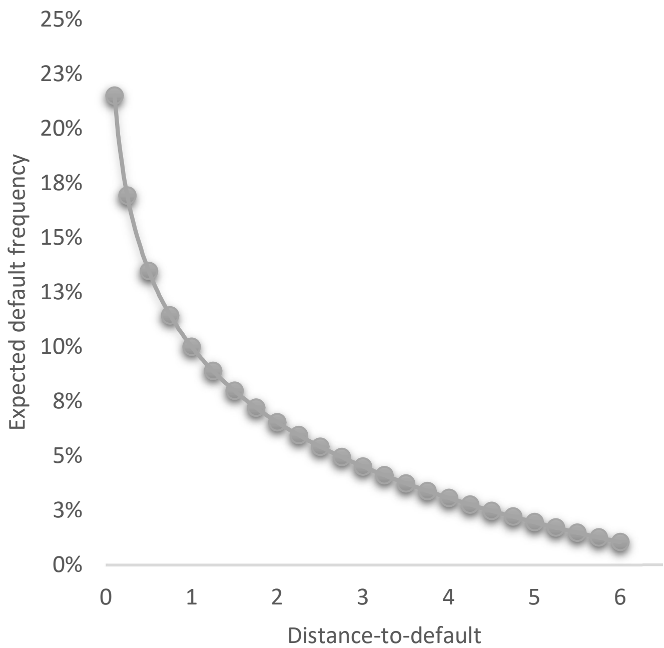


Figure 49: Moody's KMV is based on the Merton approach, where the final mapping considers the modelled distance-to-default and maps this back to an expected default frequency, as I demonstrated in the above chart.

where $e \sim N(0, \sigma_{e,k,t})$:

$$X_{k,t} = \gamma_{k,0} + \gamma_{k,1} X_{j,k,t-1} + \gamma_{k,2} X_{k,t-2} + e_{k,t} \tag{129}$$

This information can then be used to shift the transition matrix in a way which is linked to the changing economic conditions. Wilson (1997) explains that intuitively the conditional one year Markov transition matrices, I follow the notation from Wilson (1997):

$$M_t = \Pi_{i=1\ldots t} M \left(\frac{P_t}{\varphi P_t} \right) \tag{130}$$

M_t is the conditional cumulative rating distribution for all points over the horizon T. P_t is the realised default rate, φP_t is the average default rate and $\Pi_{i=1\ldots t}$ is the operator for multiplication. Therefore M_t will therefore shifts the probability mass closer to default when the ratio is greater than one. In the CreditPortfolioView approach, the transitional matrices are sub divided by industry, which improves the accuracy of the models.

13.4 Reduced form models

The next important category of credit models is called reduced-form models, and these models go by several names including rating-based models, intensity-based models, and actuarial models. In this framework, default and the associated recovery process are treated as exogenous. These models, although they cannot explain the reason for the default of a particular firm, provide a favourable match to market data.

13.4.1 Fons model

The first of these models was developed by Fons (1994) who modelled credit spreads and recovery rates in a risk-neutral setting. This model did not make use of firm information related to balance sheets or firm value. The model was based on Moody's corporate bond data that featured 473 issuer defaults. Fons (1994) model looked at how to value the bonds by incorporating the quality of credit affected as a function of spread, this is used as part of discount rate of the bond over time. An insight that Fons (1994) explains that credit rating curves show that low credit quality firms improve in credit quality as over time horizon, whereas the reverse is true for the highest rated firms. Trueck and Rachev (2009) explain the how the credit-risk is handled for a risk free bond and a bond that has credit-risk in following bond discounting equation, where a bond price $B(0, T)$ with face value B matures at time T, is as follows:

$$\underbrace{B(0,T) = Be^{-rT}}_{\text{risk free bond}}$$
$$\underbrace{B(0,T) = Be^{-(r+s)T}}_{\text{credit risky bond}} \tag{131}$$

The second equation also features a credit spread s and yield r. When we consider the credit spread or cost of credit we need to calculate the PD (which is equivalent to 1- rate of survival) that we denote as d_R is the default probability in a year. Another key metric is cumulative probability of default S_R for a bond of over time t is as follows, $S_R(t) = \Pi_{j=1}^{t}(1 - d_R(j))$. The final credit spread of a zero risk bond can be determined using the following formula:

$$s = -\frac{1}{T}\left[\sum_{t=1}^{T} S_R(T-1) \cdot d_R(t) \cdot \mu \cdot e^{-r(t-T)} + S_R(T)\right] \tag{132}$$

From the Fons model Jarrow et al. (2008) further enhanced the model to include Markov framework in a discrete time model. Wilson (1997) and Wilson (1998) advanced this approach by incorporating credit migrations using conditional transition matrices (I detailed this model in Section 13.3.2, which is a structural hybrid model).

13.4.2 Intensity models

The next set of reduced-form models are intensity models, these models focus on the relationship between the surprise of defaults and an underlying process that governs this default timing, as an exogenous input. This is referred to as the 'instantaneous rate of default', λ. This model is often expressed as a function of time, where default intensity $Pr \sim \lambda_t \Delta t$. Most often the default frequency is mathematically driven with the use of a Poisson process, so as to keep in tune with the discrete and rare events in a jump process. In order to account for the recovery rate φ as a percentage of the market value before default, Duffie and Singleton (1999) provide the following addition to the model:

$$R_t = r_t + \lambda_t(1 - \varphi) \tag{133}$$

13.4.3 CreditRisk+ (Credit Suisse)

Another commercial model was developed by Credit Suisse and called the CreditRisk+ model (CR+). This model is based on actuarial methods applied in insurance mortality modelling. According to Alexander (2004), one significant difference between structural models and the reduced form is that the default timing is treated as a surprise and is an absorbing state, to use Markov terminology. The Poisson distribution is used to control the rate of arrival of defaults. The models assume that defaults are independent and assigned as a probability p_i for each firm i, allowing for default portfolio aggregation that is developed for the probability generating function, following Trueck and Rachev (2009):

$$F_i(z) = (1 - p_i)z^0 + p_i z^1 + p_i(z - 1) \tag{134}$$

then the probability generating function (pgf) will produce for the portfolio:

$$F(z) = \prod_i (1 + p_i(z - 1)) \tag{135}$$

Using the approximation:

$$1 + p_i(z - 1) \simeq e^{p_i(z-1)} \tag{136}$$

and setting $\mu = \sigma_i p_i$ and working with a bit of simple algebra, the probability of n defaults is:

$$\Pr(n) = \frac{\mu^n e^{\mu}}{n!} \tag{137}$$

This approach is often used for credit modelling where data sparseness is an issue, which is the case in this study.

13.4.4 CreditMetrics (JP Morgan)

The last of the reduced-form models I discuss here is the CreditMetrics model. This was disclosed to the market by JP Morgan and Gupton in 1997. This approach uses the market value of a firm to determine the PD and bond value. This approach can be extended to all manner of securities; it relates to final market values data.

The CreditMetrics approach assumes that at the beginning of a time period t, if a firm is not in a state of default, the value of a bond or loan can be calculated by discounting the expected cashflows by including the credit-risk spread in the discounting function. So, the underlying assumption is that the value of a bond at time T is essentially tied to the rating at a point in time i. Future cash flows are discounted based on the credit spread and risk-free rate r. A distribution is provided by utilising the information in the rating probabilities at time. The current rating state i is given by a probability $P(X = i)$. This is compiled for all states i in t creates the transition matrix P. The rating data is then used to create a new probability transition rating matrix $p_i(t) = \delta_i P(t)$. This probability matrix allows for a forward price adjustment of the bonds, differentiating between default and non-default conditions. The bonds are essentially discounted using future adjusted information for non-defaulting bonds as follows:

$$B_j(t,T) = \sum_{k=1}^{t} C_k(1 + f^*(t,k))^{k-t} + \sum_{k=t+1}^{T} \frac{C_k}{(1 + f_j(t,k))^{k-t}} + \frac{B}{(1 + f_j(t,T))^{T-t}} \tag{138}$$

where f^* is the risk free rate and f_j is the forward rate. B is the nominal principle payment and C_k denotes nominal coupon in year k. For defaulted bonds, the following applies:

$$B_K(t,T) = R \cdot \left(\sum_{k=1}^{T} C_K + B \right) \tag{139}$$

where R is the expected recoverable amount in nominal cash flows. Based on the bond's seniority, the CreditMetrics methodology simulates the probability of recovery using a beta distribution. The final value of the bonds can then be aggregated at the portfolio level, and risk analytics can be applied. Knowing the resulting Distribution of Values (DoV as CreditMetrics term it) are not normal and feature a heavy tail in the loss side, so classic VaR analytics are not possible, one can make use of the Basel II terms and review unexpected losses. It is due to the use of equity values, rating, and bond information that the CreditMetrics approach is not a true reduced form but can be classified as a hybrid (Trueck and Rachev, 2009).

For modelling purposes, it is known that default rates, credit spreads, and the rating upgrade/downgrade ratio all vary through the credit cycle. Benzschawel (2017) explains that published one-year transition matrices can be used in a simple analysis for a known entity to understand changes to ratings at a point in time or a joint transition matrix analysis can be used in a VaR credit portfolio approach. The matrices are not very useful though for estimating spread volatility of expected losses. This has to do with rating being state dependant in stages of the credit cycle. Benzschawel (2017) emphasises that the current macro-economic setting needs to drive the updating of the default matrix using a known variable such as the upgrade/downgrade ratio (Benzschawel, 2017). A bond spread can be decomposed into a compensation for default and a spread that is related to its level of volatility. For a given transition matrix P_t the expected spread value for an given bond (in this example, AAA), can approximated as:

$$\hat{S}^{AAA} = \sum_{j \neq D} \frac{P^{AAA \to j}}{\sum_{j \neq D} P^{AAA \to j}} * \left(\bar{S}^j - \bar{S}^j_d\right) + \bar{S}^{AAA}_d \tag{140}$$

where $\hat{S}^{AAA}$ is the estimated bond spread, j is the credit state at time t. The transitioning from moving from AAA to j is $P^{AAA \to j}$, $\bar{S}^j$ is the market spread for a given state j and $\bar{S}^j_d$ is the default spread for state j at time t. Benzschawel (2017) explains that stochastic credit spreads can be generated using the Ornstein-Uhlenbeck and the current one year transition matrix as follows:

$$dp_t = \kappa(\mu - p_t)dt + \sigma dz \tag{141}$$

where μ is the term average default rate, p_t is the current default rate at time t and σ is the rating volatility and κ is the speed of mean reversion. This allows for ratings to be simulated and generated in a pattern that follows the term structure of credit ratings and volatility pattern. Benzschawel (2017) explain how this approach is useful in risk management for estimation of spread moves and credit related losses.

Credit spreads are not known to be explained by financial and/or macro-economic variables, and Benzschawel (2017) explain that an insignificant portion of the risk premium from markets is attributed to the actual cost of default. At times, the non-default side of the spread can be 10 times greater than the default spread (3.83% vs 0.36% respectively for BB- bonds). This is termed the 'credit-risk premium puzzle', which currently remains unsolved and is regarded as a big issue in credit modelling. Benzschawel (2017) goes further to explain that the bond spread can be decomposed to $s = s_d + s_\lambda$, this compensation is now split between default and spread volatility:

$$s_d = \frac{1}{T} ln\left[1 - (p_T \cdot LGD)\right] \quad \text{and} \quad s_\lambda = \lambda_t \sigma \tag{142}$$

This can then be translated into an expression of default spread only, using real world spreads, CAPM Sharpe ratios, risk neutral theory to convert PD to spreads:

$$s = \lambda_t \sigma - \frac{1}{T} ln\left[1 - (p_T \cdot LGD)\right] \tag{143}$$

This model, having inferred spread to volatility ratio, λ and utilising the recent credit spreads can allow for a market implied PD for an obligor or entity.

13.5 Credit dependence

I documented that correlation of asset class returns and PLC is not always stable through time. In times of extreme stress, one can see the correlation increasing, this is referred to as the leverage effect, otherwise referred to as 'correlation breakdown' (Buraschi et al., 2010). Evidence shows correlation is time-varying and also linked to the business cycle (Ledoit and Wolf, 2002).

13.5.1 Copulas

Fabozzi and Pachamanova (2010) explain that the Latin word 'copula' means 'link'. Not only are copulas methods simple and relatively easy to use, but the method can handle more general dependence between the variables being modelled. Trueck and Rachev (2009) explain that a copula is a random vector distribution function with standard uniform marginals. The copula function facilitates the connection between the distribution functions of two or more random variables and their marginal distribution functions. The copula was first formulated by Sklar (1959) and a fairly recent application of copula methods in finance is the work of Schweizer and Wolff (1981), and the well-known one-factor copulas by Vasicek (1987). These methods were more recently extended by Li (2000). As Trueck and Rachev (2009) explain this process allows for the joining of multivariate and univariate distributions in a very simple way, making statistical problems of dependence a lot simpler. As Fabozzi and Pachamanova (2010) explain, copulas are useful in how they separate the modelling of dependence from the modelling of a distribution. As Trueck and Rachev (2009) explain, all copulas have the following mathematical properties:

$$\begin{aligned}
&C(u_1, ...u_n) \text{ is increasing for each sequential part of } u_i \\
&C(1, ..., 1, u_i, 1, ..., 1) = u_i \\
&\text{For all } (a_i, ..., a_n), (b_1, ..., b_n) \in [0,1]^n \text{ with } a_i \leqslant b_i : \\
&\sum_{i_1=1}^{2} \cdots \sum_{i_n=1}^{2} (-1)^{i_i+\cdots+i_n} C(u_{1i_1}, \ldots, u_{ni_n}) \geqslant 0
\end{aligned} \tag{144}$$

where C is the copula, $u_{j1} = a_j$ and $u_{j2} = b_j$ for all $j \in 1, \ldots, n$. Let $X = (X_1, \ldots, X_n)'$ be a random vector of real values that have a dependence structure that can be fully catered for by the following function that is a joint distribution, following Trueck and Rachev (2009):

$$\begin{aligned}
F(x_1, \ldots, x_n) &= P(X_1 < x_1, \ldots, X_n < x_n) \\
&= P[F_1(X_1) < F_1(x_1), \ldots F_n(X_n) < F_n(x_n)] \\
&= C(F_1(x_1), \ldots, F_n(x_n))
\end{aligned} \tag{145}$$

As I denoted C as the copula function which is a joint distribution function that has uniform marginals and is a real valued function that transforms a multi-variate n dimensional function:

$$C : [0,1]^n \to [0,1] \tag{146}$$

For univariate function $G_i(u_i) \in [0;1]$, marginals that are uniformly distributed with a unit interval $[0,1]$ and $i \in N$, then there exists a copula function, C. $G_i(u_i)$ are marginal distributions, F_n is the joint cumulative distribution function, F_i^{-1} is the inverse of F_n and ρ_F is the correlation structure of F_n. The easier way of explaining a copula, for two variables x and y that are not normally distributed, x is mapped to u_1 per each percentile along its respective range, which then connects it to a standard normal distribution. The same process is used to transform y into a new variable u_2 that are as a linked using a bivariate normal distribution. The correlation, however, has given a dependence structure denoted as ρ. The random vector (u_1 and u_2), are assumed to follow a bivariate normal distribution, given by:

$$f(u_1, u_2) = \frac{1}{2\pi\sigma_X\sigma_X\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right) \tag{147}$$

where ρ is the correlation between u_1 and u_2, and $\sigma_Y > 0$ and $\sigma_X > 0$. There are many copula variations within a copula family. Figure 50 is a schema showing a 'copula family' and it summarises many of the main copulas used in the financial arena. The correlation characteristics of financial markets will aid in the selection of a coefficient.

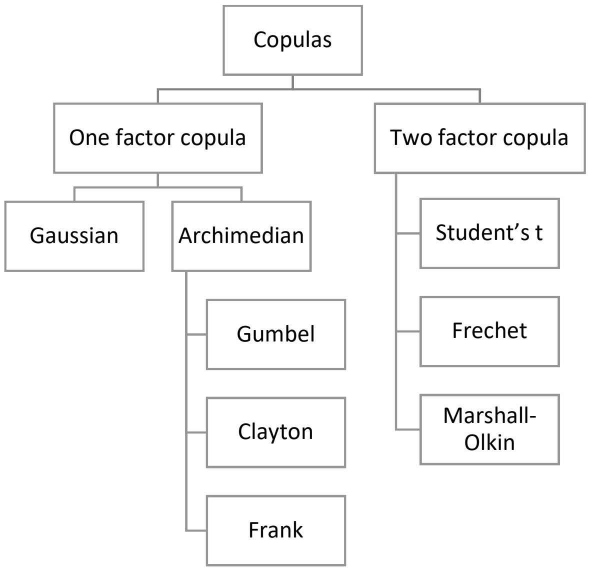


Figure 50: Schematic of the commonly used copulas, known as a copula family.

The chart on the left of Figure 51 presents data generated using a Gaussian copula and this shows the correlation that is set up to indicate that in extreme events, the correlation will increase. A copula that features an asymmetric correlation structure is the Clayton copula. This feature allows the modeller to mimic dependency features of asset-pricing data, such as correlation breakdown, in a flexible way.

From the Figure 51, I can see an illustration of how two very different correlation surfaces from two sets of asset returns in which the linear correlation is equal. The most extreme negative events show elevated correlation, this is evident chart of Figure 51, data generated using the Clayton copula. This simple component like approach used in copula technology is powerful and flexible.

Clayton copula and t-copula. As a result of the non-linearities in correlation structures between asset class returns that were experienced in the GFC, using a symmetrical copula such as the Gaussian copula may not be the most appropriate for financial asset class returns. Other copula functions, such as the t-copula, generate distributions that have increased

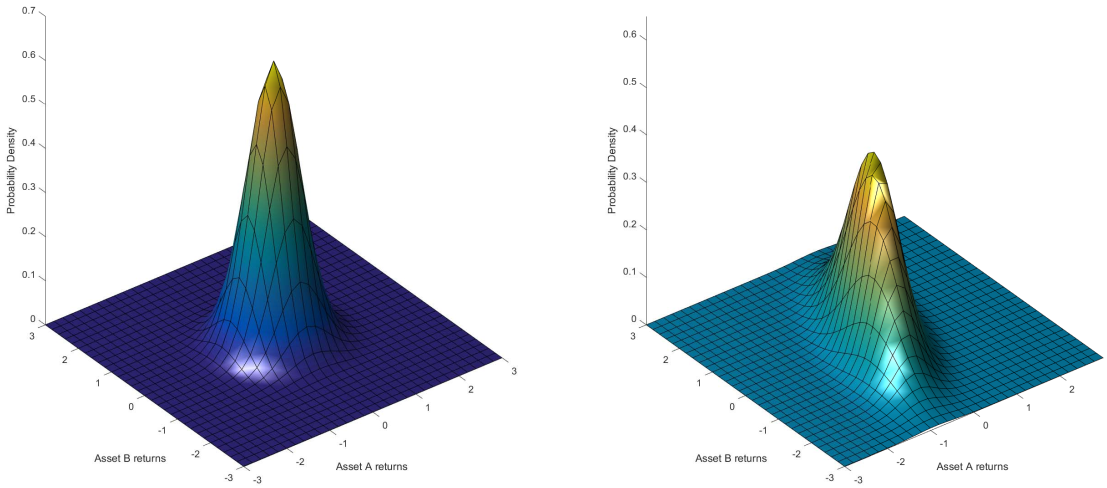


Figure 51: Independent bivariate distribution on the left and bivariate distribution with correlation on the right.

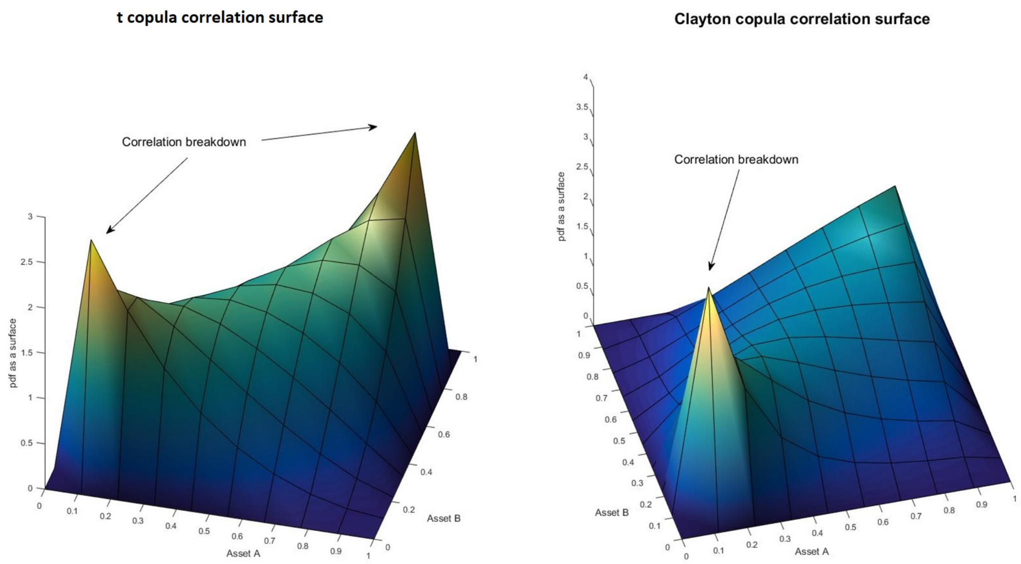


Figure 52: Correlation breakdown, comparison between t-copula and Clayton-copula. This surface is a great way to represent the level of correlation between asset as points in the credit cycle.

correlation in the deep tails, which refers to the extreme outer margins of a distribution (refer to Figure 51). If the financial returns in review also show asymmetrical correlation and we observe higher correlations in the left tail than the right tail, then the Clayton copula function can be used. The formulation of the t-copula is as follows per Munk (2011):

$$C(y_1, y_2) = t_{v,2}\Big(t_v^{-1}(y_1), t_v^{-1}(y_2); \rho\Big) \tag{148}$$

where t_v is the one-dimensional t-distribution that has v degrees of freedom. The formulation of the Clayton copula is as follows per Munk (2011):

$$C(y_1, y_2) = \left(\left[(y_1^{-\theta} - 1)^{\delta} + (y_2^{-\theta} - 1)^{\delta}\right]^{1/\delta} + 1\right)^{-1/\theta} \tag{149}$$

where $\delta \geqslant 1$ and $\theta \geqslant 1$. As a result of the Clayton copula being explicit functions, this makes the copula easier to handle, where the dependency structure for the Clayton copula is driven by changing θ, this value is $0 < \theta < \infty$. As θ increases, so the data follows a tighter correlation, in particular the lower tail.

A key market feature that I am interested in mimicking when simulating asset return paths is correlation breakdown, as was the case in the GFC for PD and LGD. To use an example, bonds and equity typically have a small to negative correlation; that was not the case during the GFC, a time of extreme market stress when both equity and bonds had simultaneous drawdowns, at a point in time reversing the correlation from negative to positive. This can result in the unfortunate situation of underestimating the risk in the portfolio (when the power of diversification cannot be relied on).

14 Literature surveys: risk premiums, regulations and cycles

14.1 Regulatory approach and modelling requirements (Basel & IFRS9)

Risk modelling in credit often follows the Bernoulli law covering discrete events, where y can take the value of 1 or 0 and $y = 1$ in credit-risk this indicates an obligor is in a default state (Trueck and Rachev, 2009). The Merton model is a core model that has been utilised for Basel II regulation. In this, default is linked to asset values against a set established threshold, D, representing the contracted value of its obligations, therefore $P(Y = 1) = P(V < D)$.

The Basel II framework (Gordy, 2003) defines asset value $Z_{i,t}$ of obligor i over time tenor t. This is the constructed as a single factor model, where $Z_{i,t}$ is associated with the Gaussian distribution having $\mu_{i,t} = 0$ and $\sigma_{i,t} = 1$, as I defined in (124). This construct, also termed the asymmetric single risk factor model features a systematic variable, X_t and an idiosyncratic variable ε and $\sqrt{\rho}$ is known as a factor loading of systemic risk, which measures the sensitivity to systematic risk (Trueck and Rachev, 2009). The probability of default is then denoted as $P(Y_{i,t} = 1) = P(Z_{i,t} < c_i) = \Phi(c_i)$. Trueck and Rachev (2009) explain that this is known as the unconditional probability of default, where Φ denotes the cumulative standard normal distribution function and a default threshold for firm i of c_i. The conditional probability of default is:

$$\begin{aligned} P(Y_{i,t} = 1|X_t = x) &= P(Z_{i,t} < c_i|X_t = x) \\ &= P(\sqrt{\rho}X_t + \sqrt{1-\rho}\varepsilon_{i,t} \leqslant c_i|X_t = x) \\ &= P\left(\varepsilon_{i,t} < \frac{c_i - \sqrt{\rho X_t}}{\sqrt{1-\rho}}|X_t = x\right) \\ &= \Phi\left(\frac{c_i - \sqrt{\rho}x}{\sqrt{1-\rho}}\right) \end{aligned} \tag{150}$$

with the model for a PD of an obligor confirmed for a portfolio of loans, assuming the same default threshold c, then depending on the state of the economy Y, the number of defaults follows a binomial distribution with $p(x) = \Phi[(c - \sqrt{\rho}x)/(\sqrt{1-\rho})]$:

$$P\left(\sum_{i=1}^{n} Y_{i,t} = k|X_t = x\right) = \binom{n}{k} p(x)^k (1 - p(x))^{n-k} \ (k = 0, ..., n) \tag{151}$$

This implicitly assumes independence between obligor defaults. Trueck and Rachev (2009) explain that by using the law of iterated expectations, the expected value of conditional probability of k defaults can assumed to be the same as the probability of k defaults:

$$\begin{aligned} P\left(\sum_{i=1}^{n} Y_{i,t} = k\right) &= \int_{\infty}^{-\infty} P\left(\sum_{i=1}^{n} Y_{i,t} = k|X_t = x\right)\phi(x)dx \\ &= \int_{\infty}^{-\infty} \binom{n}{k} \left(\Phi\left(\frac{c - \sqrt{\rho}x}{\sqrt{1-\rho}}\right)\right)^k \left(1 - \Phi\left(\frac{c - \sqrt{\rho}x}{\sqrt{1-\rho}}\right)\right)^{n-k} \phi(x)dx \end{aligned} \tag{152}$$

Trueck and Rachev (2009) explain that this model is quite simple to apply to a single loan to establish the VaR, right up to a portfolio of loans (i.e., VaR(99.9%) $= \Phi^{-1}(0.001)$). The expected loss of a loan is set as:

$$\mathrm{E}[\mathrm{L}_i|X_t = \alpha_{0.999}] = \mathrm{LGD} \cdot \Phi\left(\frac{c_i - \sqrt{\rho}\Phi^{-1}(0.001)}{\sqrt{1-\rho}}\right) \tag{153}$$

We recall that the unconditional PD is defined by:

$$\mathrm{PD}_i = P(Y_{i,t} = 1) = P(Z_{i,t} < c_i) = \Phi(c_i) \tag{154}$$

and following the Gaussian distribution allows one to infer the following:

$$\mathrm{PD}_i = \phi(c_i) \Leftrightarrow \Phi^{-1}(\mathrm{PD}_i) = c_i \tag{155}$$

The formula in (153) and (155) is the basis for calculating the capital formula for the capital requirements of Basel II for a single loan. This becomes for the following formula, referred to as the worst-case default rate (WCDR), as Trueck and Rachev (2009):

$$\text{WCDR} = \Phi\left(\frac{\Phi^{-1}(PD_i) + \sqrt{\rho}\Phi^{-1}(0.999)}{\sqrt{1-\rho}}\right) \tag{156}$$

where the final Basel II formula for capital calculations is 8% $\left(\frac{1}{12.5}\right)$ of RWA:

$$\text{WCDR} = \overbrace{12.5}^{8\%} \cdot \overbrace{\text{EAD} \cdot \text{LGD} \cdot (\text{WCDR} - \text{PD}) \cdot \text{MA}}^{\text{RWA}} \tag{157}$$

where MA is the maturity adjustment that is based on standard assumptions that are linked to the loan characteristics. The original formulation, is heavily dependent on asset correlation and was set to 20%. This approach was not approved by the market. The final decision was to use PD, as follows:

$$\rho(\text{PD}) = 0.12 \cdot \left(\frac{1-e^{-50\text{PD}}}{1-e^{-50}}\right) + 0.24 \cdot \left(1 - \left(\frac{1-e^{-50\text{PD}}}{1-e^{-50}}\right)\right) \tag{158}$$

this has subsequently been amended to the following function that is a very close approximation:

$$\rho(PD) = 0.12 \cdot \left(1 + e^{-50\text{PD}}\right) \tag{159}$$

where ρ is correlation, this is set to increase with an increase of PD of a firm (idiosyncrasies) of a firm and not systematic market factors. Therefore this implies that as a firm becomes less creditworthy, the correlation parameter decreases. This assumption was only set this way due to the fear that small to medium-sized companies would receive Basel II capital charge that was too high (Trueck and Rachev, 2009). Trueck and Rachev (2009) point out is that the internal ratings based (IRB) method is an incentive approach to adopt the Advanced-IRB modelling approach[12], but capital levels can in fact be higher than the standard assumptions, and a decrease in WCDR for an increase in PD due to correlation is questionable. I do question whether the blind use of banking techniques for asset management purposes, specifically loans to SMEs is sensible.

There are fundamental differences between the measures depending on the risk management framework, Basel II used long-run average PD and downturn LGD's, whereas the IFRS9 are all forward looking parameters. IFRS9 does not detail specifics on how to calculate the

[12]this approach requires a bank to internally develop the models using internal risk history and internally developed models

models, nor determinants for loan quality, this is left up to the bank to model internally Engelmann (2021). For a loan of tenor n years, with fixed interest rate z. Loan loss provision, as denoted by Engelmann (2021):

$$\text{LLP}_1 = cp_1 \cdot \ell_1 \cdot N_1 \tag{160}$$

which pertains to the loss category one, were cp_1 is the cumulative probability until the client default in year t. The loss in the instance of the default, ℓ_1 and finally, this is multiplied with the full loan amount or exposure, N_1. Where loans are materially impaired, the loan moves to an expected lifetime loss methodology, per Engelmann (2021):

$$\text{LLP}_2 = \sum_{l=1}^{n} \frac{1}{(1+z)^{t-1}} \cdot (cp_t - cp_{t-1}) \cdot \ell_1 \cdot N_t \tag{161}$$

which is effectively a probability weighted cash flow. When I consider the Basel II regulatory modelling needed to determine banking capital requirements, the same parameters are needed, PD, EAD, and LGD. However, these are slightly different in their modelling applications. The Basel II capital requirements are set from the capital formula from the original works in Gordy (2003), following Engelmann (2021).

$$\text{K}_{\text{min}} = \text{EAD} \cdot \text{LGD} \cdot \left(\Phi\left(\frac{\Phi^{-1}(\text{PD}) + \sqrt{\rho} \cdot \Phi^{-1}(0.9999)}{\sqrt{1-\rho}} \right) - \text{PD} \right) \tag{162}$$

Where the additional variables are ρ, the asset correlation, and the Φ is the cumulative standard normal distribution. This equation is a single risk factor model that represents the value at risk at a 99.9% level of the loss distribution. Key assumptions include that it is a large (homogenous) loan portfolio with no concentration and driven by one systematic risk factor. Here, the emphasis is on core parameters that need to stressed as part of the analysis. Each parameter has to be calibrated in a different way depending on the regulation under review. Please refer to Section 46 for the definitions of PD. Table 9 shows the various modelling assumptions used in the parametrisation process Engelmann (2021):

Unlike the single-horizon models used (e.g., one year for the PIT), the IFRS9 and stress testing models require a multi-year horizon. Further to this, using different underlying assumptions may cause alignment issues between the reserves held for asset provisioning and the capital requirements (given that the provisions are a subset of the reserves held). If the current macro-environment calls for lower provisions given the PIT nature of the modelling assumptions, this would have an impact on the overall capital held (given the requirement for TTC PD, EAD, and downturn LGD). The regulatory assumption is metered in the Basel

Reg framework	PD	LGD	EAD
Basel II/III	TTC	DT	TTC
IFRS9	PIT	PIT	PIT
Stress Test	DT	DT	DT

Table 9: Regulatory assumptions for parametrisation of PD, EAD, LGD.

II capital formula, refer to (162), where the provisions are subtracted from the amount held. To account for this mismatch, the following mechanism is used, following Engelmann (2021):

$$\hat{k}_{\text{min}} = K_{\text{min}} - \min(\text{LLP} - \text{EL}_{\text{B}}, 0.06 \cdot \text{RWA}) \tag{163}$$

where LLP represents the loan loss provision and EL_B represents the effective provision calculated using the Basel II parametrization of PD, EAD, and LGD. Engelmann (2021) explains that if the LLP is lower, additional capital will need to reserved, but where the LLP is higher, then the capital will be topped up by 6% of the Risk Weighted Asset (RWA) amount. Please refer to Engelmann (2021) for a good summary of the key literature in the development and refinement of the models, including TTC to PIT conversions and methods applied to meet the requirements in specific regulatory regimes (i.e., IFRS9, Basel, and stress testing). Engelmann (2021) points out that much of the literature is based on the assumption of a long data series with which to calibrate/parametrise the models, their study develops a method for shorter time horizons of data. The basis of the default probabilities for Basel II PIT-TTC framework is a one-factor risk model for a one-period log-return r of a borrowers credit-risk assets, following Engelmann (2021):

$$r = \sqrt{\rho} Z + \sqrt{1-\rho} \cdot \varepsilon \tag{164}$$

where Z and ε denote the independent standard normally distributed random variables, and ρ is the correlation between two assets. Z denotes the systematic risk factor, and with this formulation, this affects borrowers at the same time. r is designed in such a way that a default occurs when an asset falls below a certain threshold, Θ:

$$\text{PD} = P(r < \theta) \Rightarrow \theta = \Phi^{-1} PD \tag{165}$$

The default probability is conditional on the realisation z of Z, is:

$$\text{PD}_{\text{PIT}} = \Phi\left(\frac{\Phi^{-1}(\text{PD}) + \sqrt{\rho} \cdot \Phi^{-1} \cdot z}{\sqrt{1-\rho}}\right) \tag{166}$$

Engelmann (2021) explains the above formulation can be used to transform PD between TTC and PIT. The $PD(z)$ can be viewed as the conditional state of the economy that is aligned to the definition of the PIT PD, which changes with the state of the economy. The PD can be interpreted as the TTC PD:

$$\mathrm{PD}_{\mathrm{PIT}} = \Phi\left(\frac{\Phi^{-1}(\mathrm{PD}_{\mathrm{TTC}}) + \sqrt{\rho}\cdot\Phi^{-1}\cdot z}{\sqrt{1-\rho}}\right) \tag{167}$$

In order to utilise this framework the only outstanding variable is ρ, the asset correlations. Now this formalises the relation of PIT PD to TTC PD. An econometric model can be used to model the relationship between PIT and TTC, where C_t is the proxy for the realised PIT PD, then following Engelmann (2021):

$$\Phi^{-1}(C_t) = \beta_0 + \beta_1 X_{1,t-l} + \cdots + \beta_1 X_{k,t-l} + \varepsilon \tag{168}$$

This can then be used in (167) and solved using a number of methods, including the method of moments. With this formula, one can then imply the time series history for PIT PD or Z_t.

Smith and Balint (2019) interviewed private debt asset managers and specifically asked them to name the risks that they were most concerned about. The impact of the business and credit cycle on private debt assets emerged as their most highest ranked concern. This topic is a core part of this study and warrants a specific review of the current literature on the topic.

14.2 Credit premium

The ratio of bank financing to firm debt shifts over business cycles and this may play a role in the transmission of economic shocks (Drechsler et al., 2018). Debt restructuring in favour of bonds, as demonstrated by Drechsler et al. (2018), has been shown to mitigate the negative effects of economic cycle swings. As a result of the Volker rule and the Basel accords, which impose limits on how much leverage financial institutions can use, the availability of loans has shrunk. The real economy experiences more significant business cycle fluctuation when the composition relies heavily on bank-based finance.

Gilchrist and Zakrajšek (2012) investigate the relationship between corporate bond spreads and economic activity and involved the creation of a proxy for the credit spread index. They created an index, the ‘GZ spread’, for measuring economic activity based on the credit spread index. This spread has shown to be better in predicting economic activity than using a simple

BAA-AAA corporate bond spread, or a rated bond versus a treasury or commercial bill. The GZ spread is based on corporate bonds already in circulation, not on the primary market, over a long period (1973-2010). Gilchrist and Zakrajšek (2012) separate the GZ spread into two parts: 1) a part isolating the countercyclical movements in expected defaults; and 2) an excess bond premium, the difference between the credit spread and the empirical default measures. This construct then allows one to easily align with the so-called credit-spread puzzle, which captures an unexplained premium, potentially a liquidity premium that is time-varying (Gilchrist and Zakrajšek, 2012). Gilchrist and Zakrajšek (2012) conclude that the excess bond premium fluctuation is linked to pricing changes to default risk and that the excess bond premium movements can explain all the predictive information inherent in the GZ spread, when measured between 1985 and 2010.

Lin et al. (2020) argue that time-varying risk premium and asset-return predictability are linked. The driving force of this relationship is changing economic business cycles. They investigate the predictive power of the credit index spread for long- and short-run corporate bond returns. They report that the GZ spread has greater predictive power than the conventional term structure variables and default spread (Lin et al., 2020). The key is that the GZ index is linked to business cycles and captures time-varying bond risk premiums (Lin et al., 2020).

Gourio (2012) explains that the credit puzzle is shown empirically by the credit spreads being greater than the cost of credit, expressed as the potential for default multiplied by the loss due to that default (assumed to be 50%). The corporate bonds default with a probability of 0.4% and have experienced recoveries of around 50%, which would amount to a credit cost of 0.2%, whereas the BAA-AAA credit spread is 1%. The corporate bonds are therefore being priced at a discount and the investor can expect higher excess return premium, this again is the credit puzzle. Gourio (2012) find that defaults are primarily driven by insolvency and not liquidity shortfalls.

A well-known measure of the cost of credit in traded markets is shown by the option-adjusted spread (OAS). When compared to the actual cost of credit losses over time, the OAS is positive over time, by quite a margin. Evidence from median credit default swap (CDS) rates and expected losses points to a huge risk premia during stress periods such as 2002 and the GFC in 2009. Berndt et al. (2018) observe a net benefit for holding credit-risk assets from a lowest spread of 0.075% to a highest spread of 1%, which is more than a ten-factor increase. The premia increase if one take a cross-sectional view, with the higher rated (Aaa) bonds offering less of a risk premia than the lower rated bonds (Ca-C), with a median risk premia of 0.1% and 7% respectively. From a sector perspective, utilities bonds carry the highest risk premia and financial firms carry the lowest premia benefit (Berndt et al., 2018).

Refer to Berndt et al. (2018) for more examples of CDS and bond spillover effects connected to the IG-HY segmentation effect. Berndt et al. (2018) build a model to predict the corporate credit premia from CDS expected losses based. They use an econometric panel data regression using observable factors including equity volatility, interest rates, sentiment variables, liquidity factors, firm credit ratings, and sector dummies. Berndt et al. (2018) find that detailed Moody's ratings (i.e., in the format of Ba2) are significant factors for predicting cross-sectional spreads of credit-risk as measured by CDS rates. Giuzio et al. (2018) in their paper seek to decompose and attribute risk premium using a multivariate analysis using factors such as volatility, credit spreads, and interest rate spreads. Giuzio et al. (2018) build a portfolio to take advantage of low correlations. Giuzio et al. (2018) find that loan spreads are the most significant contributors and most persistent, and Giuzio et al. (2018) conclude that private debt has a complexity premium that is large enough and persistent enough to take advantage of and invest in. There is a substantial and persistent 1.4% differential between private debt and corporate bond premia at the same credit-risk, rating, and maturity (Giuzio et al., 2018). Interestingly, the private debt risk experience is lower than that of the corporate bonds. Giuzio et al. (2018) explain the difference in private markets can be attributed to information asymmetry, low transaction frequency, and at times, complexity and reasons we mention in this section.

14.3 Modelling approach to credit premium

Bonfim (2009) makes use of a rich database covering 30,000 Portuguese firms' credit-risk history and associate financial information, augmented with macro-econometric data to establish what the relative power of idiosyncratic and macro-economic data that explains credit-risk, in bank loan defaults. Bonfim (2009) bases their modelling framework on the Merton model, but makes two key amendments. The first is to allow a more flexible approach to idiosyncratic and systematic risk, supported by empirical evidence. These are modelled concurrently using the following model, which is a departure from the (124):

$$R_{i,t} = \Gamma X_t + \Delta Z_{i,t} \tag{169}$$

where Γ and Δ are vectors for the modelling parameters and following Bonfim (2009) can be modelled using a panel regression:

$$R_{i,t} = \alpha + \gamma X_t + \delta Z_{i,t} + u_{i,t} \text{ where } R_{i,t} \leqslant c_{i,t} \Leftrightarrow Y_{i,t} = 1 \tag{170}$$

where α is a modelling constant and $u_{i,t}$ the modelling error term. Γ and Δ are estimates of γ and δ respectively. The default threshold is $c_{i,t}$. Given $R_{i,t}$ is not observable, Bonfim (2009) defines the PD,$\lambda_{i,t}$ for each firm i and time t, as:

$$\lambda_{i,t}(X_t, Z_{i,t}) = (u_{i,t} \leqslant c_{i,t} - \alpha - \gamma X_t - \delta Z_{i,t} | X_t, Z_{i,t}) = F(\tilde{\alpha} + \tilde{\gamma} X_t + \tilde{\delta} Z_{i,t}) \tag{171}$$

where $F(\cdot)$ is the cumulative distribution of the error term, $-\gamma = \tilde{\gamma}$ and $-\delta = \tilde{\delta}$. Assumptions can then be made around the nature of these errors, which Bonfim (2009) keeps as a standard normal distribution function Θ. The model was then conclude to be a Probit model:

$$\lambda_{i,t}(X_t, Z_{i,t}) = \Theta(\tilde{\alpha} + \tilde{\delta} X_t + \tilde{\delta} Z_{i,t}) \tag{172}$$

The second difference in the modelling relative to the Merton model is in how the defaults are modelled, by way of a survival model using a duration model so as to completely describe the timing of defaults. Bonfim (2009) concludes from their modelling exercise that firm defaults are directly affected by the idiosyncratic variables such as firm structure, liquidity, sales performance and investment policy. Finally, when macro-economic variables are taken into account, the results show an improved modelling result that is independent from the idiosyncratic variables. This demonstrates that a firm must be evaluated not only for its characteristics but also for the economic conditions.

Duan and Zhu (2020) explain that the credit-to-GDP ratio and credit growth rate can act as early warning indicators for a GFC, in their analysis they explore the interplay between credit-risk and quantity of credit. They posit that quantity measures are forward-looking indicators for financial stresses. The Basel Committee on Banking Supervision (BCBS) also uses the country ratio of credit to GDP as a key measure for banks capital. Duan and Zhu (2020) point out that the classic indicators of credit such as default rate, insolvencies, bankruptcy, corporate default, and loan payment indicators are all ex-post and will therefore not give warning indicators for build-ups of macro-economic stresses. Duan and Zhu (2020) build an index, their credit cycle index (CCI) predicts PD based on firm characteristics and macro-economic factors. This model enables risk aggregation across firms, sectors, and countries. Berndt et al. (2018) develop a model for predicting the temporal variation, cross-sectional five year credit-risk premia. The modelling starts with a base model to explain the movements in time-series credit-risk premia using expected losses. The model is defined below, following Berndt et al. (2018):

$$\log\left(\frac{C_t}{\mathrm{ExpL}_t}\right) = \beta_0 + \beta_1 \log(\mathrm{ExpL}_t) + X_t\beta_X + Y_t\beta_y + \varepsilon_{i,t} \tag{173}$$

where the observed CDS rate C_t and the hypothetical CDS rate ExpL_t are modelled. X and Y are vectors of firm specific and macro-specific predictors, with coefficient vectors of β_X and β_Y respectively. The model is set up as a log-log model so as to ensure the effects of heteroscedasticity are dealt with. Naturally, ε_t^i is the random noise term.

Berndt et al. (2018) firstly model, as a base, the expected losses only as the predictor variables and this model reports an R^2 of 0.261. With the addition of firm-specific information (this includes dummy variables for alphanumeric rating) in the model, this result increases dramatically to 0.576 which is a significant improvement. Berndt et al. (2018) go further to include equity market volatility as a predictor, then adds the smirk from a volatility surface by linking changes in HY/IG ratings (recent upgrades or downgrades). The overall model reports an R^2 of 0.72. When market sentiment is included in this walk-forward modelling paradigm, the R^2 value increases to 0.81. Berndt et al. (2018) again highlight the movement in risk premia for HY firms is less correlated with macro-fundamentals than the IG firms. Berndt et al. (2018) conclude that the relationship between risk premia and macro-economic fundamentals, once controlled for firm and market-based variables, is clear and contributes to a powerful model. Berndt et al. (2018) discuss that the stressed periods (including 2002 and 2009) will have experienced marked decreases in capacity to bear risk at various levels. This will likely explain the larger increases in the risk premia relative to the expected losses. According to the findings of a recent study by Berndt et al. (2018), the ratio of risk premia to projected losses varies over time more for investment firms than for high yield firms.

14.4 Credit-risk and cycle modelling

Credit spreads are known to vary in conjunction with the credit cycle; this can be seen from rating agency data, CDS spreads, and defaults over time. A further feature is that volatility of credit spreads is more pronounced for riskier businesses and products. Now the difference in the yield is reflected by the increase in risk and does not directly translate to an increase in the returns. Spreads are understood to be composed of a credit premium and a liquidity premium. The credit premium is relatively stable over time, whereas the liquidity premium can be significant. According to Gilchrist and Zakrajšek (2012), much of the predictive power of the excess bond premium is related to fluctuations in the excess bond premium, and not changes in the underlying default risk. From 1985 to 2010, the excess bond premium is shown by Gilchrist and Zakrajšek (2012) to be the sole determinant of the GZ spread's predictive power. Lin et al. (2020) investigate the predictive power of the credit index spread for long and short run corporate bond returns. Lin et al. (2020) note that the index's link to future economic conditions gives the speculative grade component greater predictive power for future bond returns (even after controlling for macro-economic and policy variables). As a result, the GZ spread outperforms the default spread and other standard term structure factors in predictive power (Lin et al., 2020). The GZ index reflects time-varying bond risk premiums and is related to business cycles, which is an interesting finding (Lin et al., 2020). Drechsler et al. (2018) review this from another perspective, they highlight that the corporate debt composition varies over business cycles, where the use of bank finance versus corporate debt varies, and this can play a role in the transition of economic shocks. Drechsler et al. (2018) argue that changing the composition of debt in favour of bonds would have the affect of dampening the effects of a adverse business cycle fluctuations. Drechsler et al. (2018) show that the ratio of bank to bond finance falls in response to a shock from either a monetary or financial source, but it does increase following a negative supply shock. They show that changes in the corporate debt composition have a dampening effect on business cycles, but when the composition is based primarily on bank-based finance, this is associated with a stronger cycle where the contraction in the real economy is greater.

14.5 Dealing with dependence in credit-risk

Modelling a single loan is a good starting point for portfolio credit-risk, then creating a book of loans that are independent is relatively easy to attain risk statistics of your choice. Trueck and Rachev (2009) explain that as soon as we review the statistics of a portfolio of loans, how these statistics are aggregated requires more careful attention, as dependence factors in and the reliability of the statistic starts to reduce. Sometimes this effect is called the portfolio effect, in the next section I provide a targetted review as it pertains to PD and LGD dependence.

14.5.1 Modelling credit migrations

Trueck and Rachev (2009) model credit migrations using two methods: first, forecasting conditional transitional credit rating matrices using the correlation with the systematic risk factor and the idiosyncratic firm-specific factor, which is governed by a threshold correlation with the systematic risk factor. In their model they extend the Merton method to a multi-variate formulation as follows:

$$z_{ij} = \alpha(x + x_i) + \sqrt{1 - 2\alpha^2} \cdot \varepsilon_{i,j} \tag{174}$$

where x_i denote rating class specific factors and ε_{ij} represents the idiosyncratic factor. The x factors represent all the relevant economic factors relevant for rating changes.

The second method uses Gaussian and student-t copulas to demonstrate the impacts of dependence and portfolio choice. This exercise demonstrates that the higher the correlation, in terms of a VaR result for a portfolio, the results are significantly different for dependent portfolios, particularly in the case of the skewed heavy-tailed distribution such as the student-t.

14.5.2 Time varying LGD modelling

Sheikh (2018) explains that LGD models are now considered non-normal and often follow a beta distribution with a large proportion of losses close to zero and a distinct spike of losses around 100 percent (as losses are known to be higher than 100 percent). LGD models have been linked to PD correlation from the earlier forms of the modelling, including Frye (2000) who rescaled the LGD variable, as part a larger single factor risk model for the PD variable, as follows:

$$\text{LGD}_j = \mu_j + \sigma_j \text{PD}(X_{j,t,T}) \tag{175}$$

where μ_j and σ_j represent the mean and standard deviation of losses for a bank segment j, respectively. What is encouraging in this model is that PD-LGD correlation is taken care of, but the model also features a normal distribution assumption. This may be hard to justify. Sheikh (2018) notes that modelling approaches have often taken a risk factor approach, which indicates that the LGD is driven by such factors as:

- firm-specific factors
- instrument specific factors

- industry specific factors
- macro-economic factors

This amounts to a model that can be delivered in a VAR format, including sensitivities to the above listed factor risks. Although these models are flexible and relatively easy to calibrate (given the approach volume of data), these approaches still use a normal distribution as a basis which again is hard to justify. Sheikh (2018) argues that where the micro-prudential rule setting is perceivably stable (Basel II impacts), it can become macro-prudentially dangerous though. Key areas that can go undetected in a system that focuses on micro-rule setting will not detect correlation risks (both intra-parameter and intra-parameter). As banks regulatory and statutory requirements are not geared at picking up correlation (due to the simplified application of an asymmetric-single risk-factor model with a predetermined correlation variable) and concentration risks that can remain undetected, this is exactly why the regulatory bodies continue with economic capital monitoring for Pillar II. Sheikh (2018) also highlights that in the case of structured lending for banks, LGD factors are not considered, which are the most important risk drivers, and PD is really not as important.

14.5.3 PD LGD dependence (PLC)

Keijsers et al. (2018) provide an in depth analysis of loan losses on a vast banking default and loss database, the Global Credit Data consortium[13]. Keijsers et al. (2018) use a VAR model that is centred on a set of latent factors, that have been identified to affect loan losses. Vectors of factors include vectors for macro factors, loan factors and LGD factors. A credit concentration may not be revealed or fully understood until a time of crisis where the levels of credit losses are greater than expected. The combined stresses in PD and LGD are shown to take hold during a crisis, this is known as 'correlation breakdown' (i.e., the breakdown of the 'no correlation assumption'). The current Basel II accord standards for capital adequacy for banks do not incorporate PD and LGD correlation (Moody's, 2010). The downturn LGD measure is used and calibrated over periods of stress. However, evidence to the contrary is overwhelming and points out that recovery in a default event is linked to macro-economic conditions. The relationship can be understood by considering that during times of economic stress, the recovery sale is into a market in which the firm's peers are also stressed, arguably

[13] review https://globalcreditdata.org/about-gcd/ for more information relating to this banking consortium of credit-risk data

affecting the liquidity in that market (Moody's, 2018).

Moody's (2010) tests the two methods to account for this issue, firstly by using a stressed LGD that is calibrated with adverse economic conditions (much like the downturn LGD used in Basel) and secondly by expanding the model to incorporate the PD-LGD correlation (PLC). Moody's (2010) find that the downturn LGD model and stressed LGD model are not conservative enough, especially in the tail of the distribution (which is where the unexpected loss calculations are based). In this article, a PLC was used to determine the portfolio's final risk, reducing risk by as much as 40% compared to a portfolio assessed with a conventional LGD. Ignoring this could materially underestimate credit-risk in the portfolio. This is a specific detail that will be analysed from the perspective of investment portfolio losses attributable to the correlation between the PD and LGD bands.

The approach of Ozdemir (2017) to PLC is looking at the impact of the correlation of PD and LGD during times of stress. Ozdemir (2017) notes that the default correlation assumption and parameter in the Merton model is constant, based on a cross-section across obligors and through time. Ozdemir (2017) explains the model used to cover for correlation between parameters, PD and LGD, was modelled by observing the asymmetric correlation pattern, as credit-risk is sensitive to factors and indicate higher levels of correlation during stressful times. This is also known as a correlation breakdown. Ozdemir (2017) models this process as stochastic and makes use of a logistic regression:

$$R^2 = \frac{1}{1 + e^{\alpha + \sigma \cdot f(*) + m + \mu_t}} \tag{176}$$

where $f(*) = f(X_t^1, X_t^2, ..., X_t^J))$ are the observable factors, such as macro-economic factors or instrument detail. The core of Keijsers et al.'s study is to understand fluctuations in loan losses and their links to the macro-economic business cycle (with adequate data to control for seniority, product type, product, counterparty, and domicile). The Keijsers et al. (2018) model shows how LGD and PD variation sources change during the business cycle. There are two layers: first, the latent factors can be seen as the driver of time variation and dependence among variables. The latent variable structure has been shown to cover the specific interplay between variables (Keijsers et al., 2018). The second captures the LGD specifics from a panel structure of loss observations, allowing for time variation and cross-sectional effects to be collected simultaneously. The model used in Keijsers et al. (2018) considers macro-economic variables, default rates, and LGD with latent factors, using mixed distributions to create the bimodal shape. The losses may be severe or not, but this will be determined by a set of factors, that are cyclically linked. Keijsers et al. (2018) show that there is both a macro-based business cycle link and another function that is default specific, capturing the variation in credit cycle. Keijsers et al. (2018) found that LGD varies across industries due to factor sensitivity rather than specific cycles. SME's show higher cyclical variability than large corporates do. Losses on collateral loans are markedly lower, however, they fluctuate more than unsecured lending over the credit cycle. Keijsers et al. (2018) point to studies and data that show the variation over the cycle can create significant changes in capital requirements. Changing the capital requirements by a factor of more than two, and a significant portion of this change is driven from the LGD fluctuations over the cycle. Keijsers et al. (2018) determine that the time variation in defaults, LGD, and macro variables is sufficiently captured by the model that includes the macro-factor and loan factor (but not the default factor).

Keijsers et al. (2018) note some stylised facts, which are largely reinforcing what is intuitive, that unsecured loans have a lower LGD, larger borrowers (corporate) suffer lower losses than SME do, and in terms of industrials, financials feature higher losses than industrials or consumer staples. From a business cycle perspective, loans with collateral are more affected by business cycles to a higher degree, and when I consider the industrial sector it is the most stable and financials are the most affected. According to Keijsers et al. (2018), calculating economic capital by reviewing time variation in LGD implies a much higher capital requirement (measured by an expected loss number). The expected loss moves from 0.61 (0.03%) to 2.37 (0.12%). During the GFC, the time-varying LGD produces an economic capital number of 5.64 versus the constant downturn LGD of 5.03. Keijsers et al. (2018) conclude that default measures explain changes to the credit cycles during times of high-stress environments, whereas the business cycle is explained by macro-factors.

15 Credit-risk parameter modelling

The modelling of credit-risk for the purpose of investment strategy is started in the following section. The modelling will be primarily descriptive and inferential in nature, with the goal of determining the level of credit-risk for a set of counterparties. The data has been supplied for research purposes by NUS-CRI [14]. This data is used as the primary data source

[14] The Credit Research Initiative (CRI) is a non-profit undertaking under the Asian Institute of Digital Finance (AIDF) of the National University of Singapore (https://nuscri.org/en/about/).

and represents the list of investable entities against which direct loans are provided. These loans form the private debt building block for pension requirements. Another significant set of data is the sentiment indices used in part III. The remainder of the data requirement for this study is collected from academic texts in the form of parameter assumptions, such as PD or LGD.

Credit parameter programming I coded the credit analysis using Python and Python packages as part of a Colab hosted program that I coded. I provided details of the address of the Google Colab files in the appendix; Section 30.

15.1 Catalogue of parameters and sources of data

In Table 10, I provided a description of the key data attributes used in this modelling analysis. The first table is a summary of available data sources.

Category	Parameter	Description of data source
Sentiment	OECD BCI	The sentiment proxies sourced the forecasting risk appetite in part III. This includes the sentiment variables from the survey approach from the OECD (BCI, CCI, and CLI) that have been generated using TextBlob and Vader. In this chapter, instead of concerning myself with how well newspapers can predict the sentiment variables (BCI, CCI, and CLI), I will now consider these as modelling factors for assessment in modelling credit-risk spreads for the next twelve months.
	Vader TextBlob	The sentiment proxies I generated by The Guardian newspaper is used in the UK, The New York Times is used in the US, and GDELT-augmented sentiment is used in the RSA. These combinations are based on the best performing sentiment analysis from Chapter III.
Credit	SRI	NUS-CRI Probability of Default (NUS-CRI PD) is a daily measure that is a forward-looking PIT PD. The data covers all exchange listed issuers worldwide (80,000). The PIT is reported at points ranging from 1 month to 5 years. Analytics are easily aggregated, however, aggregate measures are also provided by country, sector, or region.

Table 10: Description of alternative data sources, article counts and start date.

15.2 Simulation parameters inferential analysis

The following data and academic analysis are intended to establish the most appropriate parameters for the simulation of private debt assets as part of a multi-asset economic scenario generator (ESG). The data on private debt is sparse and for the most part, this is a real-world feature. The parameters in this study will best represent the true value of assets enabled by direct loans to corporations and small and medium-sized businesses. The analysis will be partly empirical, using NUS-CRI data and I provide an analysis is a summary of parameters from academic text that can be considered direct inputs and/or reference benchmarks, keeping the overall analysis reasonable.

The first analysis summary shows the summary statistics of the twelve-month SRI EDF using the NUS-CRI PD data. The largest proportion of firm data comes from the USA, followed by the United Kingdom and, understandably, South Africa. I also summarised the number of rows of data (DB rows). The mean PD value is highest for the USA and lowest for the UK (this is a unweighted average and not weighted in accordance with the economy). The sentiment and credit cycle indicators are both scaled using the same approach taken in Section 10.1.4.

Domicile	**Entities / Grade**	**Row count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
GBR	1,439	454,488	0.4%	1.5%	0.0%	0.0%	0.1%	0.4%	92.1%
	Investment	392,298	0.3%	0.6%	0.0%	0.0%	0.1%	0.3%	62.6%
	Sub-investment	13,961	2.6%	4.0%	0.0%	0.6%	1.4%	3.0%	69.1%
USA	4,837	1,751,433	0.9%	3.6%	0.0%	0.0%	0.1%	0.5%	99.7%
	Investment	1,487,428	0.4%	1.0%	0.0%	0.0%	0.1%	0.4%	96.9%
	Sub-investment	73,326	5.6%	8.1%	0.0%	1.2%	3.0%	6.6%	99.7%
RSA	327	102,705	0.8%	2.2%	0.0%	0.1%	0.2%	0.6%	74.8%
	Investment	83,466	0.4%	0.6%	0.0%	0.1%	0.2%	0.5%	26.2%
	Sub-investment	7,741	3.9%	4.6%	0.0%	1.4%	2.5%	4.6%	65.8%

Table 11: Long-run credit parameters for Economic Scenario Generation. These results are based on the results of a data analysis of SRI data. I summarised the NUS-CRI PD data, categorised it by country, investment grade and also provides descriptive data points, including number of entities, rows of data and summary statistics of the NUS-CRI 12m PD. Notably, this shows a significant proportion of data as investment grade, possibly not representative of the economy at large.

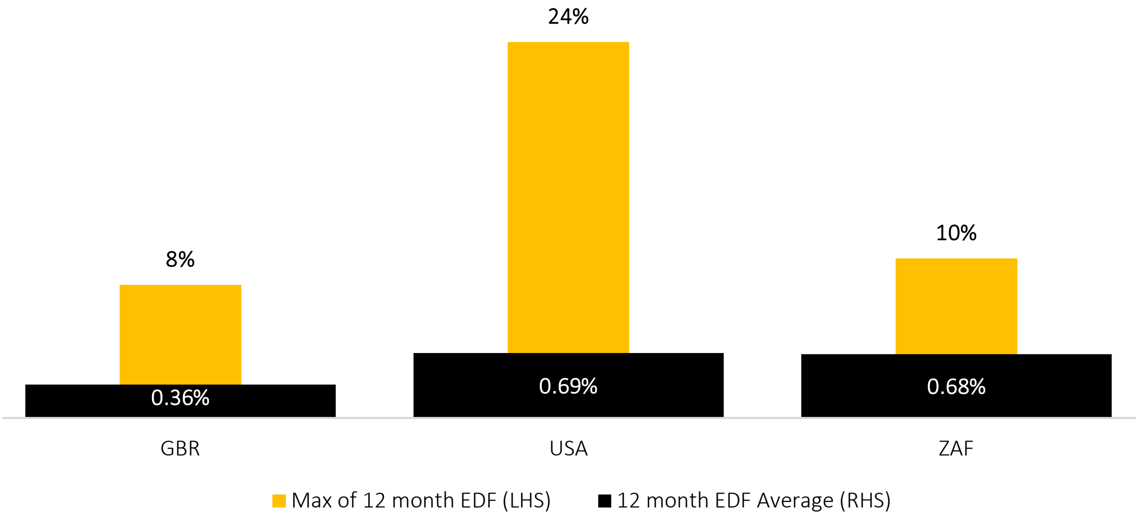


Figure 53: The bar chart is a country-level visual of the degree to which companies change PD from the average value to the maximum value; this is a proxy for an average TTC versus downturn PD. The USA is showing a far higher value (24%) than the UK or RSA (8% and 10% respectively). A visual comparison of the difference between the average monthly credit spread values and the maximum values for each country, as well as a summary of the underlying sectors. The US features the largest range between what can be considered TTC PD and DT PD.

15.2.1 Credit cycle index (CCI) and diagnostic extrapolation (DE)

The following empirical analysis explores the relationship between the proxy for diagnostic extrapolation, the scaled sentiment variable and credit spread. The credit cycle index is proxied using the twelve month forward credit spread, which is then compared to our sentiment variables in Table 10. The parameters from that analysis are core to the simulation of data in Part V, the portfolio construction and optimisation part of this study. Instead of looking for subtle correlations using advanced analytical techniques on limited data. Which is the method for analysis will define parameters for modelling that have already been checked for reasonability using research of the academic literature and market participant public information. I consider the visualisations in selecting the most appropriate index for forward credit spread simulation.

GBR CCI-DE analysis The following analysis is based on the results for the UK BCI index only. With this filter, I will be able to review the credit dynamics for each country and the relationship of the DE proxy to the forward credit spread. In the UK, reviewing the OECD BCI index as time-series rolling six-month index, it is the Textblob data that performs the best. This index appears to be well connected, with a clear negative relationship in times of stress. I can see a distinct drop in sentiment just before and during the GFC in 2009, as well as from 2002 to 2004. The sentiment levels were high in the run-up to the GFC, in 2016, and in the more recent period, beginning in 2019. The favourable feature of this time series is the sharp drops in sentiment around shock periods, with well-timed turning points.

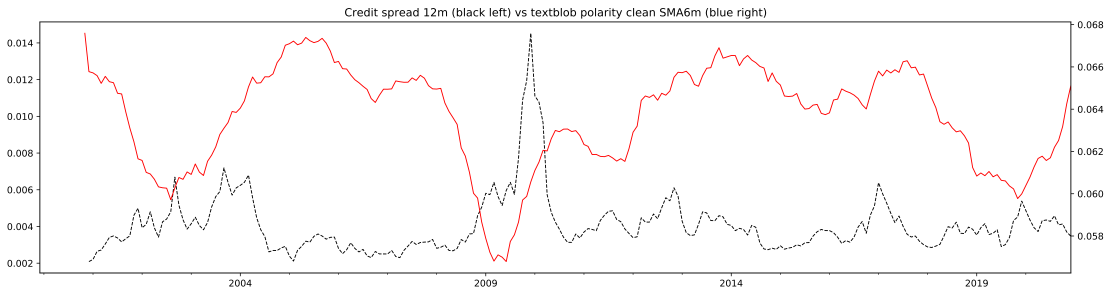


Figure 54: UK Textblob based on The Guardian data plotted against the 12 month UK credit spread 12 months forward. Please note that the data is effectively lagged, hence the negative relationship, that may on inspection be interesting.

This analysis has been performed for using the outputs from the my analysis using Textblob, and Vader on data sources such as The Guardian and The New York Times (please refer to III. I reviewed factor variables as the analysis was completed for OECD BCI, Textblob, and Vader (please refer to the appendix; Section 26). The analysis includes a time-series chart for each factor and the complete series of XY charts for the term structure. The variables that show the most consistent relationship over the term structure and the clearest pattern on the XY chart have been selected and are therefore reported in the body of this thesis. In Figure 55 there is clear evidence of a negative relationship between the Textblob variable and the 12 month forward credit spread in the UK. This relationship continues through the term structure of the credit spread and is still in place for the next 12 months (far right chart). A clear feature in this data is the tight grouping of results with low credit spreads and DE (as measured by Textblob sentiment valence). The relationship appears to increase in variability, albeit still negative, as DE is lower. This relationship will now be summarised and used in the simulation of asset pricing in Chapter V.

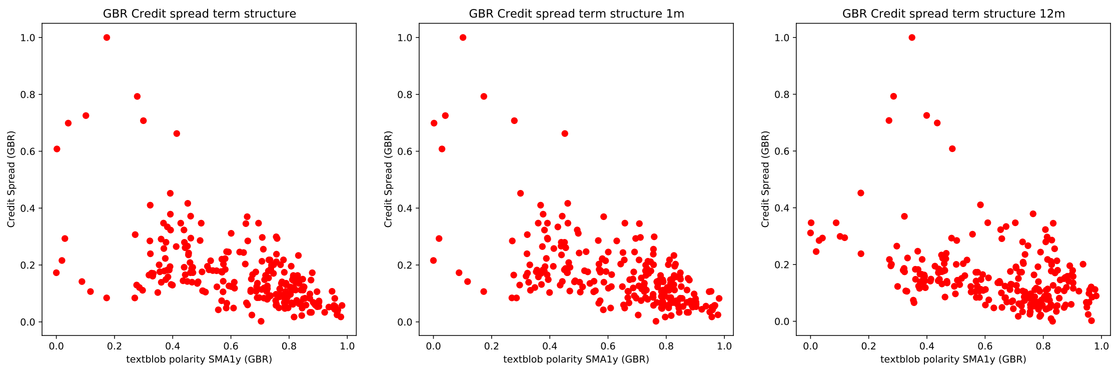


Figure 55: UK Credit spread as it relates to business Textblob. Both variables have been scaled for this analysis. The XY chart indicates a negative relationship that extends through the term structure of credit spread. Note the tighter grouping of results in the high-DE to low-credit spread segment of the chart.

USA CCI-DE analysis The following analysis is based on the results for the USA BCI index only. I will also be able to generate a more granular review and analytics. I will also be able to review the credit dynamics for the RSA based only on the relationship of the DE proxy to the forward credit spread. In the RSA analysis, I reported the US OECD BCI index factor only in this chapter; other factors are reported in the appendix; Section 26.9).

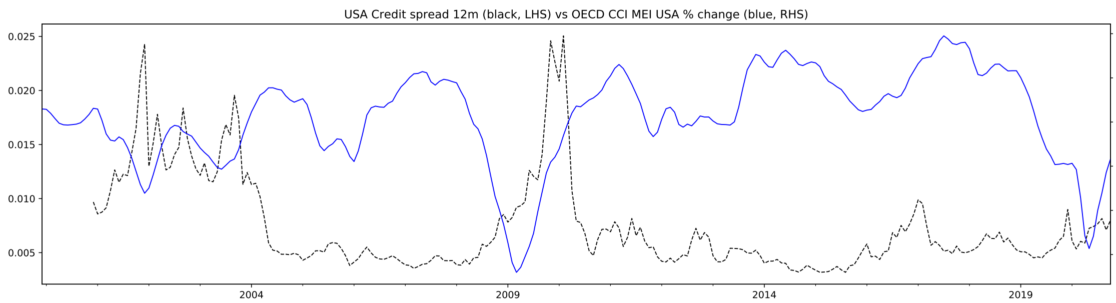


Figure 56: US OECD BCI (as the DE proxy for US) data plotted against the 12 month UK credit spread 12 month forward.

US OECD BCI, which is promoted as a USA DE index, appears to be well correlated, with a clear negative relationship in times of stress. I can see a distinct drop in sentiment just before and during the GFC in 2009, as well as from 2002 to 2004. The sentiment values are consistently high in the run up to the GFC, 2016 and the more recent period, 2019 onwards. The favourable feature of this time series is the sharp drops in sentiment around shock periods, with well-timed turning points.

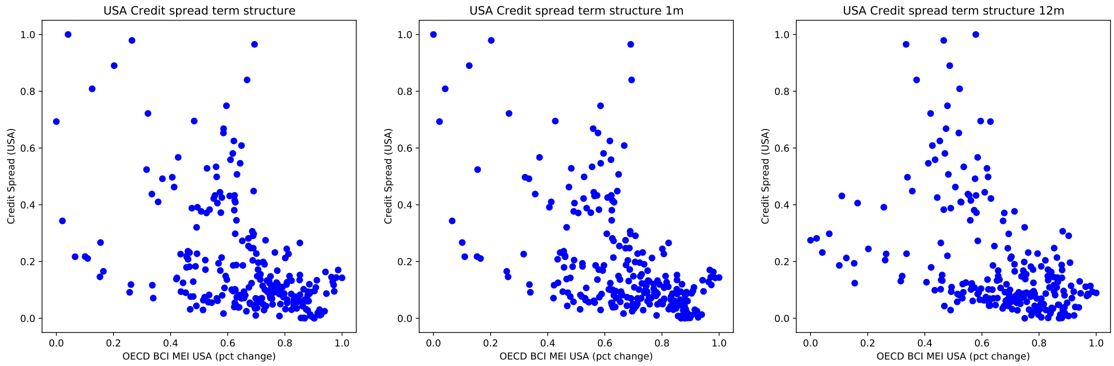


Figure 57: US credit spread as it relates to business: US, OECD BCI, DE proxy. Both variables have been scaled for this analysis. The XY chart indicates a negative relationship that extends through the term structure of the credit spread. Note the tighter grouping of results in the high-DE to low-credit spread segment of the chart.

RSA CCI-DE analysis The following RSA analysis is based on the results for the Vader index performed on The Guardian and The New York times data only. I will also be able to generate a more granular review and analytics. With this filter, I will be able to review the credit dynamics for the RSA based only on the relationship of the DE proxy to the forward credit spread. In the RSA analysis, I reported the RSA Vader index factor only in this chapter; other factors are reported in the appendix; Section 26.9).

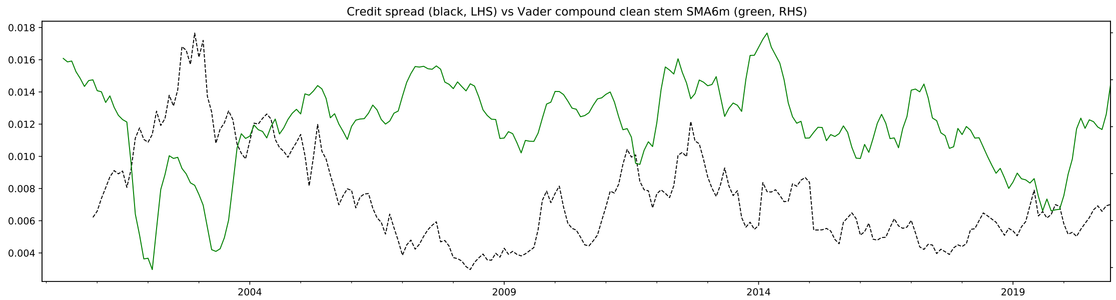


Figure 58: RSA Vader index based on data from both The Guardian and The New York times (as the DE proxy for US) data plotted against the 12 month RSA credit spread 12 month forward.

The Vader index, which is promoted as a RSA DE index, is plotted in the Figure 58 shows a negative relationship between the variables and a concentration in times of DE stress. I can see a distinct drop in sentiment just before and during the GFC in 2009, as well as from

2002 to 2004. Sentiment levels were consistently high in the run-up to the GFC in 2016, and in the more recent period, beginning in 2019. The favourable feature of this time series is the sharp drops in sentiment around shock periods, with well-timed turning points.

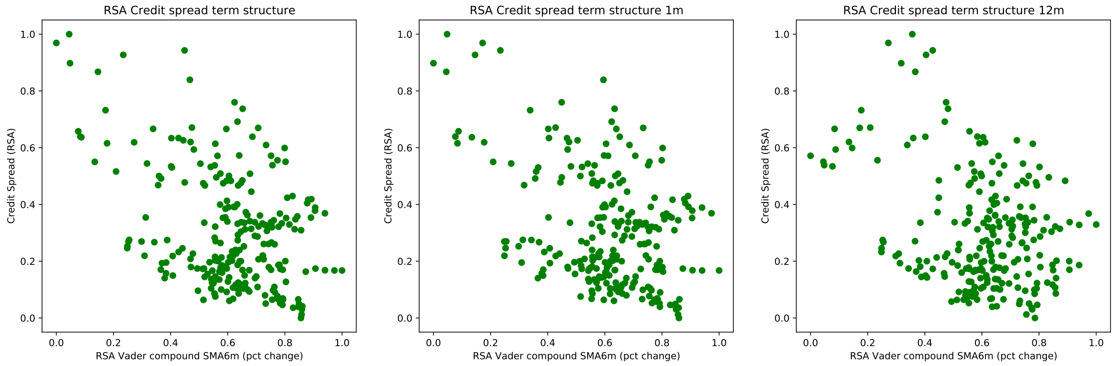


Figure 59: RSA Credit spread as it relates to the RSA VADER Index based on data from both The Guardian and The New York times. Both variables have been scaled for this analysis. The XY chart indicates a negative relationship that extends through the term structure of credit spread. Note the tighter grouping of results in the high-DE to low-credit spread segment of the chart.

15.2.2 Transform variables - PD distribution

UK PD distribution analysis The distribution of default entity risk is measured as the twelve-month forward distribution of realised SRI-EDF ratings. The framework relates to economic states as they are defined by the CCI-DE. In this analysis, I provided a perspective based on the company EDFs as they are mapped to CCI-DE deciles.

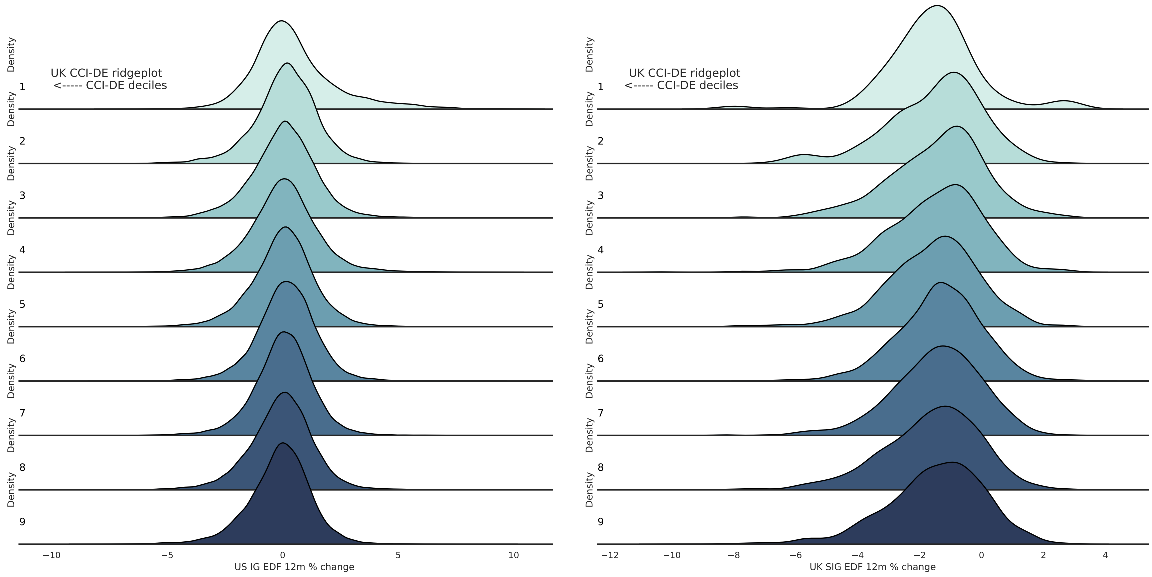


Figure 60: Joint plot showing the distributions of UK investment-grade (IG) SRI EDF and sub-investment-grade (SIG) SRI EDF ridge plots, respectively. In the case of the IG distribution, it is clearly more stable than the CCI-DE deciles and has a tighter distribution. The SIG ratings report a less stable distribution over the CCI-DE deciles, where the EDF shows a significant hump in the tail in CCI-decile = 0, thereafter the distribution is skewed in the opposite tail.

US PD distribution analysis The distribution of default entity risk is measured as the twelve month forward distribution of realised SRI-EDF ratings. the framework relates to economic states as they are defined by the CCI-DE. In this analysis I provided a view based on the company EDF's as they are mapped to CCI-DE deciles.

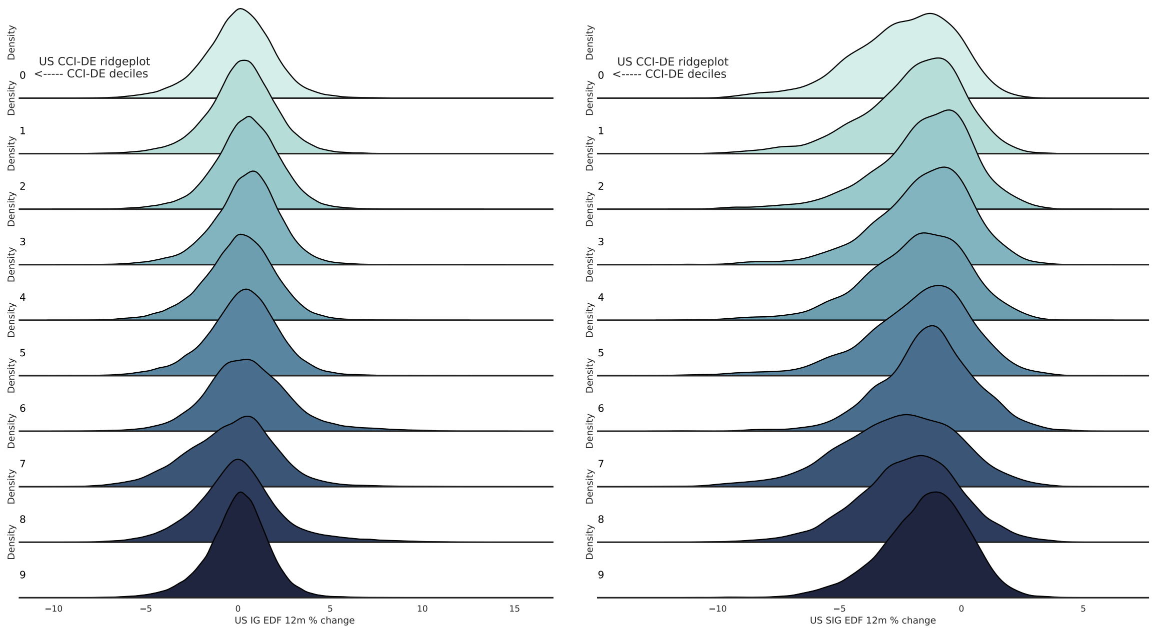


Figure 61: Joint plot showing the distributions of the US investment-grade (IG) SRI EDF and sub-investment-grade (SIG) SRI EDF ridge plots respectively. In the case of the IG distribution is clearly more stable over the CCI-DE deciles and has a tighter distribution. The SIG ratings report a less stable distribution over the CCI-DE deciles, where the EDF shows a significant hump in the tail in CCI-decile = 0, thereafter the distribution is skewed in the opposite tail.

RSA PD distribution analysis The distribution of default entity risk is measured as the twelve month forward distribution of realised SRI-EDF ratings. The framework relates to economic states as they are defined by the CCI-DE. In this analysis I provided a view based on the company EDF's as they are mapped to CCI-DE deciles.

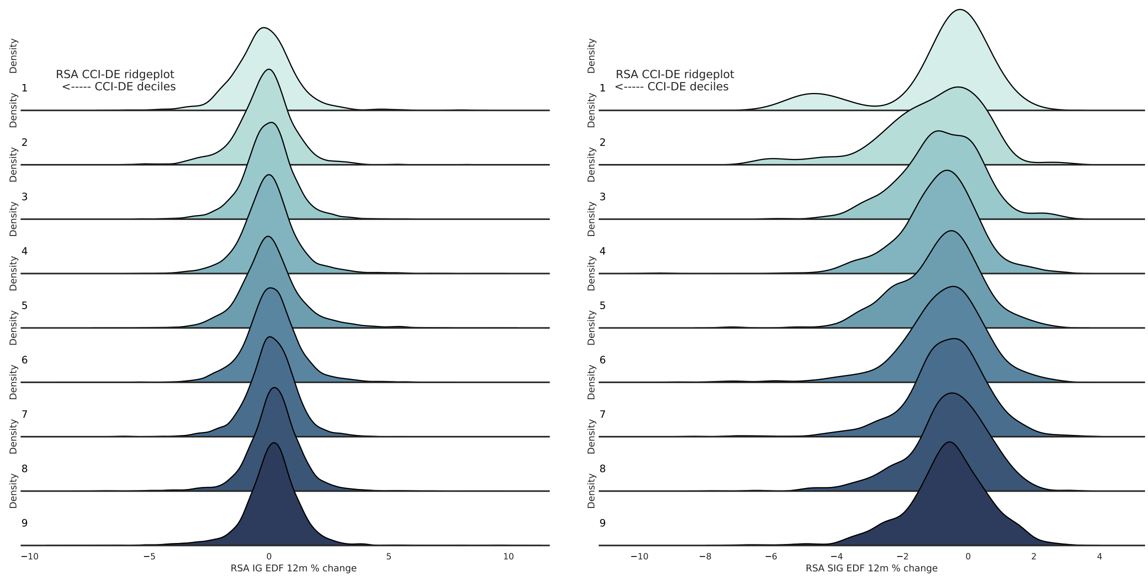


Figure 62: Joint plot showing the distributions of RSA investment-grade (IG) SRI EDF and sub-investment-grade (SIG) SRI EDF ridge plots, respectively. In the case of the IG distribution, it is clearly more stable than the CCI-DE deciles and has a tighter distribution. The SIG ratings report a less stable distribution over the CCI-DE deciles, where the EDF shows a significant hump in the tail in CCI-decile = 0, thereafter the distribution is skewed in the opposite tail.

16 Summary of analysis

Credit-risk modelling is a key part of this study. In this section I reviewed the underpinning theory of the core credit-risk models. This allowed for a deeper understanding of the underlying assumptions that are used for banking, which are not ideal for the purposes of asset management and portfolios that are not large, nor homogenous. In this section I investigated the key methods for credit-risk modelling, detailing the key underlying assumptions and modelling frameworks with the aim of detailing the most appropriate modelling technique for use in this study.

This is an important section in which I established the core methods, models and theories needed for credit modelling. I started by defining the key variables (PD and LGD) and their measurement frameworks (PIT, TTC and DT). I provided a detailed review of the Basel II banking methodology for understanding portfolio credit-risk. I surveyed and summarised the core modelling techniques and their history in developing the commercial solutions that banks use today. Modelling includes the approaches taken to retail and commercial banking models such as Logit models and Altman's Z-score. I detailed the data rich modelling of KMV Moody's to the slender reduced form models, that have become a vital method for this study. The review in banking modelling was noteworthy when I built a portfolio of loans in which the modelling in banking appears ubiquitous. Understanding the models is essential, as these theories all too often being based on an assumption of large portfolios. One key assumption that is not solid is the independence of PD and LGD in which I detailed a key method to model if there is a break down in assumptions. I went further into detail when there is also a breakdown in assumption in stressed markets, identified as correlation breakdown.

In the last section, I modelled the relationship between my sentiment measure based on news signals, as a simple measure of reaction based on news sentiment, confirming the time series relationship with the expected credit spreads. I made use of the aggregate CCI proxies from Part III, to understand whether there is more information when compared with the credit data, when such was based on a large sample of companies and corporates. The distribution of the expected credit shows a widening of distribution and longer tails as the cycle indicator moves toward as stressed scenario.

Part V

Economic scenario generator and portfolio modelling

The next section describes the designs of a suitable economic scenario generator (ESG) that takes into account the complexity of the private debt asset class, contends with the phases of the credit business and is part of an integrated market data generator. Following that, I use the scenarios to build a portfolio choice solution that includes an analysis of reinforcement learning (RL) strategies for sequential planning.

17 Introduction

I will now describe the asset-pricing models by defining the mathematical models I used. I also illustrated the model using charts to elucidate the underlying intuition for the benefit of the reader. The aspect part of the model is the embedded credit-risk simulation technique. This technique allows me to configure a reasonable set of return and cost pathways for the private debt building blocks. I have taken special care to explain the use of copula techniques that reflect the correlation between PD and LGD. This model is dynamic, depending on the current economic cycle, and my innovation is to utilise the forward view of distributions using the sentiment based information that is cycle-dependent (outputs from the modelling in Part IV).

To better understand the potential future scenarios and dynamics of the economy and capital markets, ESG models are utilised. The output can then be effectively utilised for risk-based decisioning strategies, giving decision makers a good view of where their risks lie. The Society of Actuaries (2016) defines an ESG as a software tool set up to simulate future economic scenarios that drive prices for financial assets. Although the mathematical approaches and economic concepts can be identical, ESG models have a different framework from other economics and finance models, the ultimate purpose being different. These are interest rate sensitive and are an integrated view of the asset class universe, providing a full view of potential extreme values for both assets and liabilities.

18 Asset price scenario generation

The output from the ESG is a set of simulated interconnected paths for each asset type specified. Each path represents a possible and plausible economic outcome of an asset over time. A full set of paths is vital. When used correctly, such can be crucial to asset managers (such as a defined-benefit pension fund manager) in making investment decisions (Society of Actuaries, 2016). The ESG will output a set of integrated plausible asset/security distributions across various scenarios interrelated in a meaningful way (i.e., correlations are considered). The working parts of the ESG model are structured such that they cater to correlation; this is made possible by means of a hierarchical cascade structure. At the core of the ESG is an interest-rate projection, and many other components can be built around this, such as risk premia models. Society of Actuaries (2016) explains that ESG models are usually set up to perform on risk-neutral and real-world modelling frameworks, depending on the intended product or trading environment (in my case, a real-world projection). A high-level taxonomy of models as illustrated below:

- These models can be categorised firstly by whether they have been developed using discrete time or continuous time. Continuous time models are most often in fixed-income markets as they allow for analytical tractability, especially when pricing derivatives.
- The second consideration is whether the model is arbitrage-free or an equilibrium model. Equilibrium models consider the utility of the investor and a full description of the economy. This method presupposes economic equilibrium, in contrast to the arbitrage-free models which assume no arbitrage opportunities. Instead, these models make assumptions about the stochastic behaviour of interest rates and the price of market risk. This is a subtle but important difference.
- Thereafter, the models are developed using either a single or multiple factors. The one-factor models, often referred to as one factor short-rate models, are based on the short rate, which describes the corresponding term structure. Single-factor models assume that all the information about a term structure is captured by one specific factor, the short-term interest rate.

When I consider the more advanced requirements for financial assets, in which simulations are generated for use in risk management, investment planning, and pricing, I consider the following specific stylised features, also detailed in the Society of Actuaries (2016):

- **Fixed-income** Short-term bonds are more likely to produce lower yields than long-term bonds. This can be reversed in the case of an inverted bond curve. The reverse relationship is true for fluctuations over a month, in which a long-term bonds are more stable than short-term bonds. Even when compared with other parameters, term spreads have shown strong predictive ability, as argued by Engstrom and Sharpe (2018). Since the GFC, term structure models have been performing especially poorly; and economic aspects have been mainly stripped out of macro-finance models, as evaluated by Kucera (2017).

- **Rates** are not zero-bound and, in today's financial world, can be negative. More recently, this has started to reverse, with rates moving up in key economies.

- **Corporate credit** spreads are expected to decrease as the rating of the underlying bond improves. This proportion of the spread is not only due to credit; it is partly attributed to the liquidity factor. This approach is considered important when reviewing private debt as an asset class.

- **Default factors** that affect the PD (more than rating alone) include firm, industry sector and country.

- **Equity returns** are expected to have a higher range and average value than those seen in fixed-income, with volatility varying over time.

- **Economic recession** and fluctuation tend to disrupt credit spreads and the probability of default.

- **Correlation** between variables is also not stable, one can expect this to change over an economic cycle.

In summary, a good ESG model will find the right balance between simplicity and modelling of critical historical features, and must be shown to capture extreme events. This approach is preferable to using an over-fitted VAR model that fits the data but does not have well-thought-through-dynamics that are modelled and can be explained (Society of Actuaries, 2016). Going forward, I will detail the ESG model I built to generate asset-pricing paths.

ESG programming I coded the full ESG model, this model provides asset-pricing pathways needed in the portfolio construction Section 19. I utilised Python and Python packages as part of a Colab hosted program. I provided details of the address of the Google Colab files in the Appendix; Section 30.

18.1 Mathematical preliminaries

I summarised typical features of the pricing model, providing context around use this model to generate asset-class data. I will now start to formalise the mathematical model I use as part of a simulation program. Far from trivial, an important step is to appropriately describe the mathematical model I use to generate the simulated pathways. The ESG model has primarily focussed on asset class return scenarios used to describe hidden risks and opportunities as asset pathways, determined by a reasonable set of input parameters.

18.1.1 Discretised probability space

I consider a filtered probability space that satisfies regular conditions; this is formalised using measure theory. Measure theory becomes important when I am working with continuous or discrete variables and distributions; the probability theory is a special case of measure theory. Foundational components, using measure theory, for a probability space include the "triple" that consists of $\Omega, \mathcal{F}$, and $\mathbb{P}$. To explain further, Ω denotes the sample space, $\mathcal{F}$ denotes a sigma-algebra (σ-algebra), and $\mathbb{P}$ denotes the probability measure. The triple is critical when defining a sufficient statistic for random processes, explained here:

- Ω is the sample space, the set of possible outcomes or state space. An element is where $\omega \in \Omega$ represents a specific realization from among all possible uncertain objects of the model. An event is a subset of Ω, which in this case is a collection of events.
- $\mathcal{F}$ is a sigma algebra, or σ-algebra on Ω, that is a collection of events of subsets Ω with the properties:
  - ◇ $\emptyset \in \mathcal{F}$, this is the unique set having no elements and its size is zero
  - ◇ $\Omega \in \mathcal{F}$
  - ◇ For any set $F \in \mathcal{F}$, the compliment F^c is also in $\mathcal{F}$
  - ◇ If $F_1, F_2, ... \in \mathcal{F}$, then the union $\bigcup_{n=1}^{\infty} F_n \in \mathcal{F}$
- $\mathcal{F}$ is the collection of all events that can be assigned a probability.

- Lastly $\mathbb{P}$ is the probability of an event on $\mathcal{F}$, for each of the events that are part of the σ-algebra that are to have a probability associated with it. $\mathbb{P}$ probability measure is a function $\mathbb{P} : \mathcal{F} \longrightarrow [0, 1]$ with $\mathbb{P}(\Omega) = 1$ and the property that $\mathbb{P}(\bigcup_{m=1}^{\infty} A_m) = \sum_{m=1}^{\infty} \mathbb{P}(A_m)$ for any sequence $A_1, A_2, \ldots$ of disjoint events each in $\mathcal{F}$.

According to Munk (2011), in the case of multi-period models, the state space is a consideration of all factors and all conceivable combinations of events for all periods in time specified $t \in T$, where $t \in [0, T]$ simply where $T > 0$. This leads to a tremendously vast state space.

18.1.2 Wiener process to generate random numbers.

For a modeller to denote random processes, a special notation is used. For this study, I used a well-known process, the standard Wiener process. It is a key part of the modelling for many asset-pricing models, also known as Brownian motion. Fabozzi and Markowitz (2011) define the Wiener process as W_t, reflecting the following properties:

- For any time $s < t$, the difference $W_t – W_s$ is a normal random variable with mean zero and variance $(t–s)$. The difference can be expressed as $\sqrt{t - s} \cdot \hat{\varepsilon}$, where $\hat{\varepsilon}$ is a standard normal random variable
- For all time $0 \leqslant t_1 \leqslant t_2 \leqslant t_3 < t_4$, the differences $W(t_2)–W(t_1)$ and $W(t_4)–W(t_3)$ are independent random variables, said differently these increments are independent
- The value of the Wiener process at the beginning is almost surely zero, $P(W(t_0) = 0) = 1$

The properties above characterise the Weiner process, in which these properties can be considered a mathematical formula representation of a Brownian motion. I make use of same mathematical notation in developing stochastic differential equations (SDE). Such can be use to provide a model of Brownian motion.

$$dX_t = \mu dt + \sigma dW \tag{177}$$

In order to denote return increments from discrete to infinitesimal time steps, Δ change to d. The Wiener process is a form of Markov stochastic process that is frequently used to research Brownian motion. This is an important factor in traditional asset pricing, and it is the random value or random element used to express uncertainty.

18.1.3 Random walk models

With a random number generated, we also require a model that appropriately generates the asset pricing with predefined known market characteristics. The mathematics foundations are defined, I now consider random number process and the limits imposed by specific random walk models. Campbell et al. (1997) have categorised these random walk models into the following classes:

- *Random walk I (RW1): Independent identically distributed (IID) increments.* The random walk models are simplest because they only require that the random numbers, or increments in the asset-pricing path, be independent and identically distributed (functions of the increments must also be independent). This is the only one of the categories that results in a Weiner process.
- *Random walk II (RW2): Independent increments.* If I soften assumptions around the persistence of increments in RW1, the error term also may allow for heteroskedasticity in RW2. The key is that increments are independent, not identically distributed.
- *Random walk III (RW3)*: This is the weakest form of random walk, in which the independence assumptions are also relaxed. In the realm of random walks, this class has the widest applicability. The RW version used in this research is one of the most extensively researched and tested types of market model.

I defined the core elements of the mathematical methods and statistical assumption sets before proceeding with the ESG model development.

18.2 ESG model logic and parameters

To introduce the model, before I denote the model mathematically, I explain the model logic or intuition, using simple formulas and asset paths to illustrate the key asset-pricing mechanisms. Central to the risk premium model is the net benefit. The net benefit is the return in addition to a risk-free cash asset, from investing in a specific systematic risk premia in a specific domicile (in our analysis: UK, USA, RSA). In the case of a private debt asset, I am really introducing two systematic risk premia: first, the benefit from default risk, and second, the systematic benefit from bearing the liquidity risk. In this analysis, the risk premia are expressed as a percentage return per year; and the additional cost is expressed as a percentage cost per year, hence the requirement that the strategy be net positive. Therefore, capture the private debt strategy in a formula:

$$\mathrm{E}[C] = [\mathrm{rate}]_{\mathrm{rf}} + [\text{private debt premium}]_i - [\text{expected losses}]_i - [\text{fees}]_i \tag{178}$$

i represents asset class risk factor and rf denotes a risk-free asset. To expand, the expected cost of credit can be further broken down into its component parts. The expected loss is defined as the "potential for default" multiplied by the magnitude of loss given said default ($E[Loss] = PD \cdot LGD$). The relationship can then be restated as follows:

$$\mathrm{E}[C] = [\mathrm{rate}]_{\mathrm{rf}} + \underbrace{[\text{default premium}]_i + [\text{liquidity premium}]_i}_{\text{private debt premium}} - \underbrace{[PD \cdot LGD]_i}_{\text{expected loss}} - \text{fees} \tag{179}$$

I will later detail how this model caters for difference cycle effects, correlations whilst considering different asset classes.

18.2.1 Parameter illustrations

In this section, I introduce key features of asset-price simulation in a set of charts for this analysis that can also be used to report results. I introduce concepts incrementally to explain the model and in developing the model logic. First, I demonstrate the value of an arbitrary portfolio of over-time. The portfolio is invested in a risk-free asset, such as cash, that accumulates interest over the full horizon.

- **Risk-free asset (rate$_i$)**: a core underlying building block in the simulation of asset-prices is the rate of return received on risk-free assets. The cash value of the portfolio nicely demonstrates the cumulative compounding of returns over time.

- **Asset pricing mean reversion**: this view reports only the rates that the portfolio is receiving in return over the investment horizon. The rate increases over time and flattens out to a steady rate. This is a key underlying concept that I stylise in the simulation of asset prices in this analysis: that the current rate will converge toward a long-run average over time, known as 'mean reversion'. In the illustration, in Figure 64, the current rate starts close to 3% per annum and evens out 15% per annum (note the chart y-axis is reported in monthly rate return). Another consideration is how rapidly the rate of return converges on the long-run rate; this is known as the rate of convergence, or speed of mean reversion and denoted by (κ).

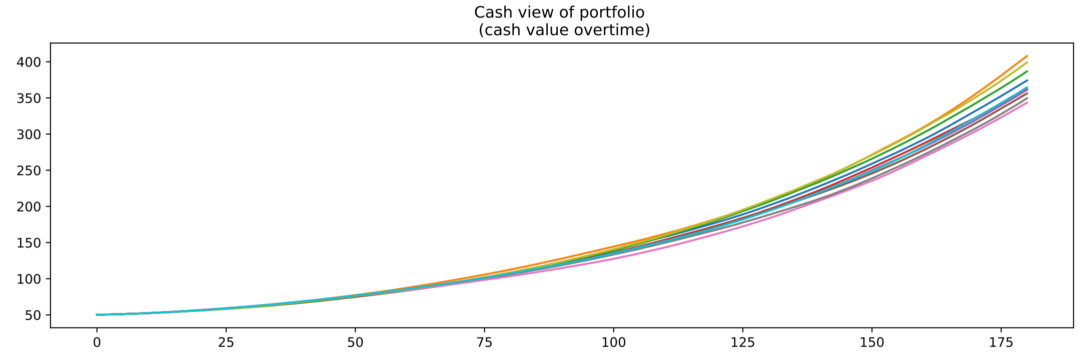


Figure 63: This view is a demonstration of the compounding effect of invested cash in continuously compounding markets, the upward curved index of portfolio value. The colours in the charts only serve to highlight different paths or scenarios.

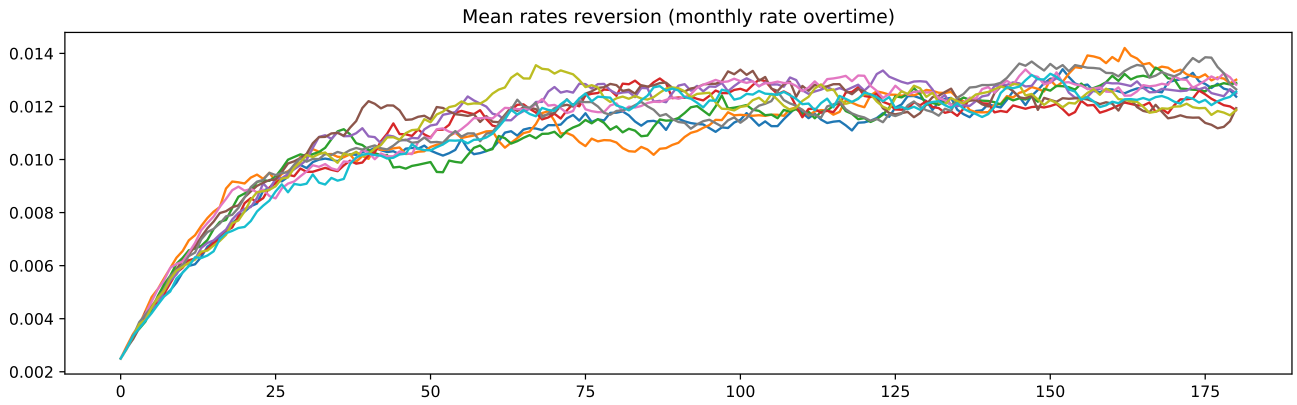


Figure 64: Simulation of rates, showing the mean reverting nature of rates. Another key feature that has been introduced to this stylised analysis is the volatility factor, which is the cause of the random nature of the return series.

- **Asset pricing correlation**: In this study, we expect that asset prices are also driven by correlation between the assets; this is therefore a multidimensional approach to modelling the assets. This is not a difficult process and is standard for asset class simulations.

- **Credit jumps**: using a jump diffusion model introduced by Merton (1976) and adopted for use in a credit-risk reduced form model, a sudden change in asset value linked to the default of the company is driven by a stochastic process called a jump. The jump mechanism in this context drives a significance change firm value. So the frequency of each jump is a key parameter (and how it functions across an economic cycle).

- **Loss given default (LGD)**: This is the loss associated with each credit jump. The loss variable is always negative, however, its impact is latent, as it is multiplied with the jump variable, which is mostly zero until a jump arrives.

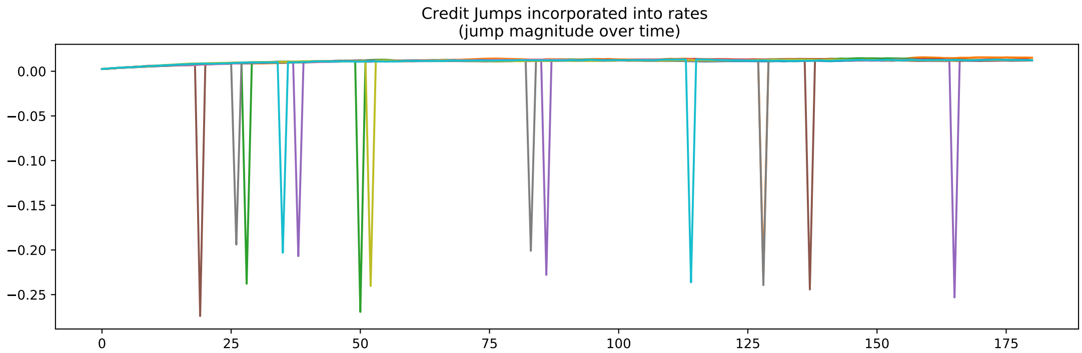


Figure 65: For credit-risk, the jumps are associated with a significant loss, which, varies between -20% and -25%. What is notable, is the extent of a credit loss relative to the rate of return in each period; the rates of return in Figure 64 pale in comparison to the losses associated with a default event, or jump. The change in value, can be clearly seen in this rate chart, that is now augmented with the value change due to credit-risk jumps. Basically the investor sees that value of her asset instantaneously fall when a default occurs.

- **Credit dependence**: An important theme in this study is time variation of the credit cycle. I also detailed how the diagnostic extrapolation (DE) is connected to business cycles and credit-risk cycles. The time-varying dynamics capture variation in the credit/business cycle per the following parameters:

  ◇ DE impacts systemic risk as it drives the PD variable and respective variability.

  ◇ DE that is a connection to herding behaviour is connected to a resulting credit cycle. The DE drives a correlation factor between sentiment and the credit cycle, impacting the θ in the Clayton copula. The net impact is that the losses associated with a jump is also expected to increase due to herding behaviour. This is because assets are expected not to receive their full value in a sale during stressed times. Liquidity in the asset markets is anticipated to be reduced.

The cash value of a portfolio is one way of summarising a journey; while a histogram or density chart can be used to summarise a distribution. A time-series plot can be used to summarise the path of the value of the asset. Figure 67 conveys the paths and distributions

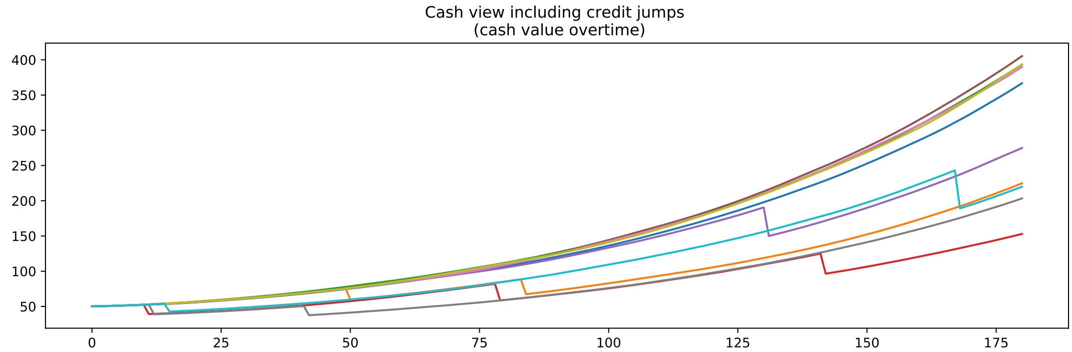


Figure 66: This is a final illustration of the progression of a few indexed asset pathways. This chart also demonstrates the impact of a default when it occurs (an example is the purple line showing the distinct negative jump just after the 125^{th} time interval). Bear in mind, these stylised paths are not derived from real calibrated values. Such does demonstrate the impact on value that a credit-risk event has relative to the slower incremental increases from cash flows (this is true in real-world data). This analysis also features a few paths that show a second or third default. Such may be rarer in the real world (the rate of arrival may not be as high). In all cases, the net impact over the full horizon is significantly different when a credit default has occurred. The colours in this chart only serve to illustrate the different scenarios or asset pricing paths.

together with the resulting distribution, to the right of the path plot, shows a marginally skewed distribution. The results of the simulation clearly indicate the effect of credit losses, where certain paths illustrate a clear loss in value. As it has been explained, this loss is a function of default intensity and the magnitude of the loss. Figure 68 separately illustrates PD on the left and the LGD distribution on the right. Both distributions vary over the credit cycle. Such can affect the shape (skewness), the median value, and the standard deviation. The LGD distribution in this study follows the beta distribution, with high proportions of loss reported closer to each tail (0% loss and 100% loss). The typical U shaped distribution of credit losses is illustrated to the right of the Figure 68.

As I explained earlier, the impact of the PD and LGD correlation is attenuated in favourable markets and heightened in stressed markets. This feature has also been referred to as "correlation breakdown", which negatively correlated assets become more correlated in stressful markets.

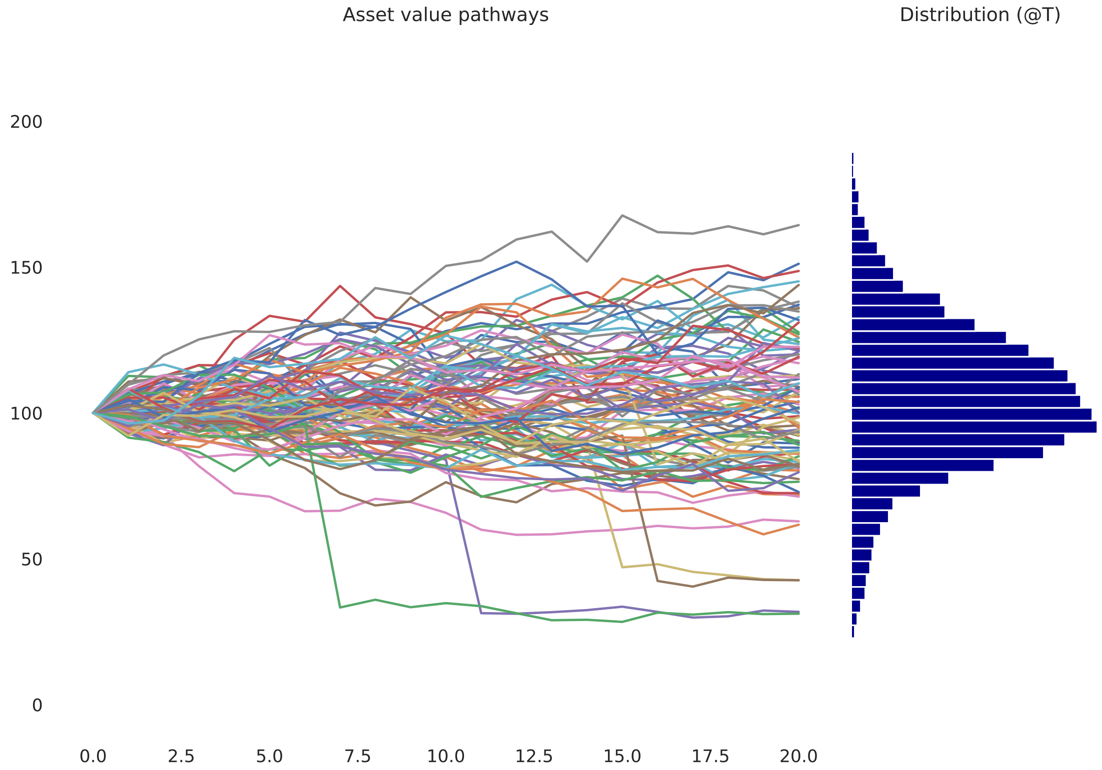


Figure 67: Path plots on the left, featuring asset path scenarios with a step change in value, due to the loss incurred from a credit event (jump default).

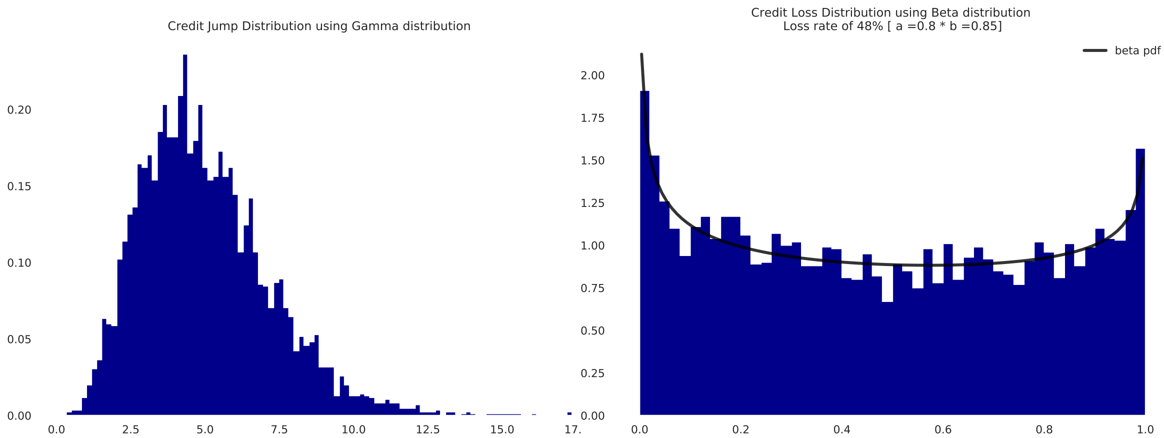


Figure 68: PD and LGD distributions for defaults and loss expectations respectively. The PD distributions is skew and in this illustration is a gamma distribution. The LGD is based on a beta distribution.

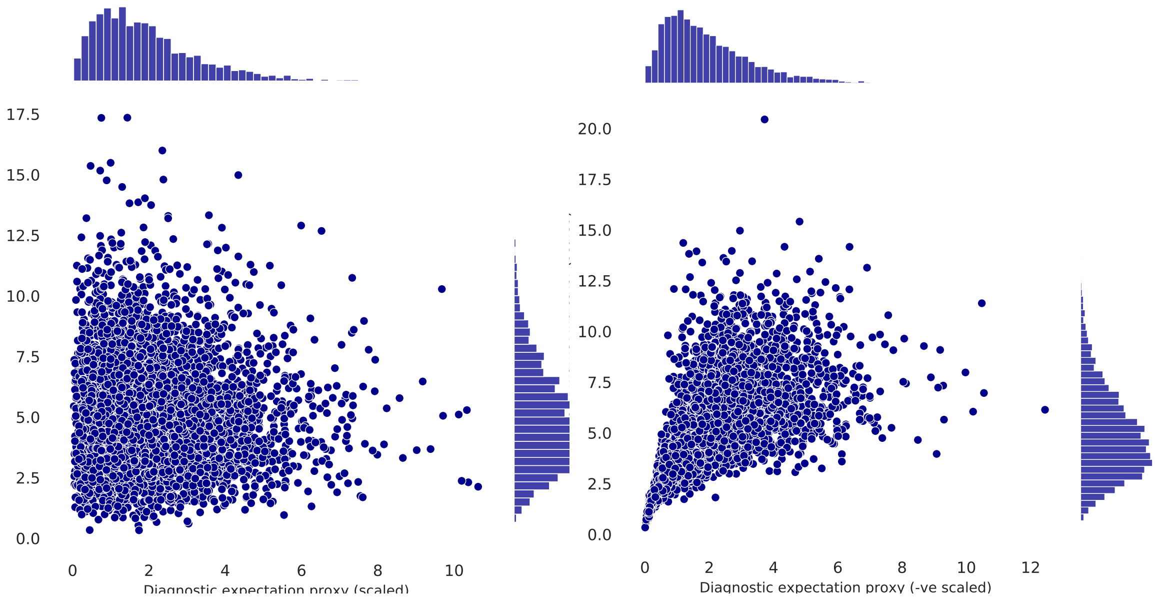


Figure 69: Joint plot of the DE-CCI and the PD distributions with two different values for the Clayton copula θ variable. To account for the different correlation, the chart on the right is an illustration of a correlation breakdown (described above). Here the marginal distributions are the same as in the left-hand chart; however the concentration of losses is shown in right chart.

18.3 ESG mathematical formulation

What is well-established in financial markets is that the parameters exhibit mean reversion, the phenomenon in which returns revert to the long-term mean. In this research, I add a discrete mean-reverting process to the ESG model, similar to a process in the continuous space is known as the Ornstein-Uhlenbeck process (Munk, 2013). This model belongs to a RW3 model category from Campbell et al. (1997) as the model assumptions account for endogenous variables linked to market sentiment.

18.3.1 Risk free asset pricing model

The risk-free asset is one with a certain future return and no risk of loss and is denoted in this study as:

$$\frac{dp}{p} = r(t)dt \tag{180}$$

Where r is the risk free rate and seen as a deterministic function. This represent the change in yield in the market and as this is, by comparison to other assets, not going to feature many risk types that reflect its value, hence the simple representation of rate r over time.

18.3.2 Liquid asset model

The traded liquid asset is captured by the structure of the geometric mean reversion (GMR) model. This model is used in a standardised way for equity, fixed-income, property and gold returns. Parts of the model follow a similar approach in this modelling taken by Uhlenbeck and Ornstein (1930) however this model is set in discrete time, following Fabozzi and Pachamanova (2010), as follows:

$$dr_t = \kappa\left(\xi - r_t\right)dt + \sigma\sqrt{r_t}dW_t \tag{181}$$

where ξ represents the speed of adjustment and correlated multi-variate at the asset classes level. The magnitude of ξ is positively related to the speed that the process takes to return to the long-term average.

18.3.3 Illiquid private credit model

This model is a natural extension of the GMR while keeping the analytical tractability of bond prices. Credit-risk is dealt with by incorporating stochastic jump into the GMR model. Jump models have been widely utilised in finance to estimate asset pricing, especially for options.

The jump diffusion models are made up of two parts: a jump portion and a diffusion part. The diffusion process is driven by the beta distribution and this determines the extent of the expected loss (a value between 0% and 100%). On the other hand, the Poisson process determines the jump that allows one to simulate abrupt and unexpected price increases in the underlying asset. In my model, this change is due to a realised default. The beta distribution approach to this model is also widely, Moody's (2007) even providing the average α and β variables used. Notice also that the jumps is positive; however, it is subtracted from the asset value for all-time period where a jump indicator is greater than one. For each asset class, i that is effected by the DE proxy, Θ, the asset-pricing dynamics are determined by the following stochastic jump diffusion model:

$$dr_{t,\Theta} = \kappa\left(\xi - r_{t,\Theta}\right) dt + \sigma_\Theta dW_t - dJ_{\lambda_\Theta} \cdot \Upsilon_{\alpha_\Theta,\beta_\Theta} \tag{182}$$

where the mean reversion function, denoted by $\kappa\left(\xi - r_t\right)$, is driven by the speed of mean reversion κ, long-term asset returns ξ and current asset prices r_t. The loss Υ associated with the default and jump intensity dJ are both connected to the diagnostic extrapolation proxy, the CCI level denoted as Θ. The jump model is a component of the pricing model. The jump model is based on the Poisson process driven by the intensity of default. This Poisson process is often used in asset-pricing methods, in which asset returns are the occurrence of spikes in returns, often asymmetrically (I detail this in Figure 183).

18.3.4 Jump frequency, intensity model (PD

This technique has been often used in academic studies and by market-based practitioners, when dealing with reduced-form credit-risk models. The default risk is modelled by introducing a stochastic jump, randomly using a Poisson distribution that impacts the value of the asset (in this case negatively). This technique can therefore mimic the spike in asset value after a default. The Poisson process models the value spike by randomly sampling in accordance with a frequency directly related to the expected PD, the jump intensity, λ in this model. If necessary, the magnitude of the jump, Υ, representing the loss incurred, can be modelled using an appropriate distribution (often a beta distribution for loss-given default). For a time-homogeneous Poisson process with intensity, λ, and for $t > 0$, the number of arrivals of a jump, J_t, in the case of N_t, is given by:

$$P_{N_t}(n) = \frac{\lambda_t^n e^{-\lambda_t}}{n!} \tag{183}$$

18.3.5 Loss associated with default (LGD, Υ).

The magnitude of the jump in this stochastic process, drives the magnitude of the shock Υ and is akin to the loss given default, LGD, in bank credit modelling. When multiplied by the shock arrival rate N which is akin to the PD, please refer to (183). In this study I make use of the very flexible beta distribution, as this has often been used for LGD modelling, in particular for simulation based approaches. This includes Moody's KMV portfolio manager and the CreditMetrics model. The beta distribution only requires two variables, the beta distribution's centre parameter, α_d, and a shape parameter β_d. The beta distribution probability density function is defined as:

$$f(x) = \frac{X^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{184}$$

where the function is normalised in B, ensuring that the function lies within the range of zero to one. The mean of the distribution is calculated to be:

$$\mu = E(\text{LGD}) = \frac{\alpha}{\alpha+\beta} \tag{185}$$

The variance of the distribution is calculated as:

$$\sigma^2 = \frac{\mu(1-\mu)}{\phi} \tag{186}$$

where $\phi = \alpha + \beta$. I made assumptions around the shape of the distribution (the standard shape is a bi-model u-shaped distribution) and I assume a mean and variance that aligns with the standard assumption that Moody's utilise in their LGD models. The beta distribution is also used with a mean equal to 50%; and a standard deviation equal to 26% (Moody's, 2007). The standard assumptions are followed for the DE proxy that represents a standard market (CCI=50%). More significant changes are applied in the case of an expected economic shock or boom (in which CCI=0% and 100% respectively). Please refer to the remaining assumptions in Table 15.

18.3.6 Jump and loss dependence (Clayton copula) model

Macro based LGD models deal with parameter correlation to the PD variable forms of the modelling, including that of Frye (2000) who rescaled the LGD variable, as part of a larger single factor risk model for the PD variable. In my model both the LGD and PD are affected by a common variable linked to expected credit cycles. The formulation of the Clayton copula is as follows per Munk (2011):

$$C(y_1, y_2) = \left(\left[(y_1^{-\theta} - 1)^{\delta} + (y_2^{-\theta} - 1)^{\delta}\right]^{1/\delta} + 1\right)^{-1/\theta} \tag{187}$$

where $\delta \geqslant 1$ and $\theta \geqslant 1$. Where θ is governed by my CCI diagnostic expectation proxy Θ. As a result of the Clayton copula and Gaussian copula being explicit functions, these are largely easier to handle than the Gaussian copula. The dependency structure is determined by θ, where $0 < \theta < \infty$. The higher the value of θ causes a tighter correlation pattern, in particular the lower tail.

19 Portfolio construction design and implementation

Optimisation in finance is a broad and challenging area of research. In this thesis my interest in the use of an optimisation model that is a rapid and stable algorithm, which has at its core a control for downside risk. The algorithm needs to be based on the same downside measure we are controlling in this problem, and designed for sample paths. Chekhlov et al. (2004) introduced an optimisation technique that is directly related and robust, as their popular method in Rockafellar and Uraysev (2002). This method is termed the conditional drawdown at risk (CDAR) portfolio optimisation. In the following sections, I provide a detailed description of the optimisation modelling approach, the parameters used, and the results of the optimisation.

19.1 Portfolio optimisation modelling

Chekhlov et al. (2004) introduced a family of risk measures that deal with drawdown. This is a key risk measure in both asset management and for their client portfolios. Examples are pension fund and insurance investing, both set against a benchmark, and for those investments linked directly to a liability. The drawdown measure is often used to state a maximum risk level acceptable in the writing of mandates. Maximum drawdown (MDD) is defined as measure of an asset price historical high (often measured locally over a sensible time horizon, which is closer to the investors investment time horizon). As Chekhlov et al. (2004) explain, the CDAR is the measure of the average of the worst drawdowns over a sample set, that are greater than a threshold selected by the user (α is often set at 5%). For example, over a selected time horizon, the CDAR is then the mean of the worst (1-α) percentage of drawdowns experienced. Chekhlov et al. (2004) go further to explain that the Rockafellar and

Uraysev (2002) method for optimisation using CVAR is linear, and can be directly used for this problem. The problem has been cast as a linear problem that has been broken down into sections by using a linear piecewise convex model. This then means that the problem can be solved using linear programming. Such has the advantage of being rapid, stable, and robust. I follow Chekhlov et al. (2004) for the formulation of the model, in which $x = (x_1, x_2, \ldots, x_n)$ represented the asset classes weights for m asset classes. The drawdown function is defined as:

$$D(x,t) = \max_{0 \leqslant \tau \leqslant t} \{w(x,\tau)\} - w(x,t) \tag{188}$$

where the drawdown function is the peak maximum value of the asset price over the time horizon of:

$$\max_{0 \leqslant \tau \leqslant t} \{w(x,\tau)\} \tag{189}$$

t less the current value at the time. Let $\zeta_\alpha(x)$ represent the threshold and CDAR optimisation is represented by the following linear function:

$$\begin{aligned} &\max_{x \in X} \frac{1}{dC} y_N \cdot x \\ &\text{s.t. } \zeta + \frac{1}{(1-\alpha)N} \sum_{k=1}^{N} [\max\{y_j \cdot x\} - y_k \cdot x - \zeta]^+ \leqslant VC \end{aligned} \tag{190}$$

d is the number of years and C is the capital value of the investment. Capital values is measured as CAGR in our example over the investment horizon of $[0, T]$, that has been set at 1 and 5 in this study (given the mean reversion, the first year is interesting, relative to a sensible time horizon of 5 years for a long term investment problem). $R(x)$ is the expected return, measured as the inner product, $\frac{1}{dC} y_N \cdot x$, y represents the cumulative return and V represents the allowable losses in capital value.

Optimisation programming I made use of Chekhlov et al. (2004) for the portfolio modelling, Chekhlov et al. supply the optimisation logarithm in github[15]. I utilised this program to run the optimisation as part of the larger colab-hosted program that I coded; and I provided details in the Appendix; Section 30.

19.2 Portfolio construction implementation

This section provides a summary and high level process for the optimisation modelling and portfolio construction.

[15]https://riskfolio-lib.readthedocs.io/en/latest/portfolio.html

19.2.1 Investable asset classes

The universe of asset classes is the foundation from which all portfolios can be constructed.

Assets	Description	GBR	RSA	USA
Cash and Fixed-income				
Cash and equivalents	Defined by U.S. GAAP & IFRS, cash and cash equivalents, cash convertible within 3 days, e.g. commercial paper, short term government bonds.	✓	✓	✓
HY Bonds, Sov Bonds, Inv Bonds	Fixed-income assets featuring payments referred to coupons, plus the original principle returned at term. Coupons rates are agreed on schedule and bought in secondary markets from Sovereign and corporate issuers.	✓	✓	✓
Growth assets				
Equity	Equity investments, public direct company share holdings. Return from the change in price of shares and dividends that the company pays out.	✓	✓	✓
Property	Real Estate Investment Trusts (REIT), returns from income producing real estate with earnings from properties and rentals distributed to shareholders.	✓	✓	✓
Alternatives				
DC Inv	Direct holding loans as direct credit (DC), held to maturity. Priced with referencing rate, such RSA JIBAR, plus a spread for expected costs (primarily default and liquidity) and a premium.	✓	✓	✓
DC Sub-Inv	DC Sub-Inv represents higher risk loans. The loan pricing will need to be priced far higher to cover for the increased risks and associated costs.	✓	✓	✓
Gold	Represented by the gold price can be accessed via an exchange traded fund.	✓	✓	✓

Table 12: Definition of asset classes used in this analysis

19.2.2 ESG input parameters

The data used for asset class input parameters has been directly sourced from significant global asset managers' long term capital market assumptions. This includes (JPM (2022),TPriceRowe (2022), GSAM (2022), Investec (2022), NorthernTrust (2022), PGIM (2022), CTI (2022), MorganStanley (2022), Invesco (2022), Robeco (2022), BlackRock (2022), AQR (2022), Schroders (2022) and NinetyOne (2022)). There is a great degree of variability and I have taken care to regulate such that the numbers are plausible. Because of the size of the assets invested by the asset managers I listed above, I believe the use of these capital market assumptions to be representative. I see these parameters as a reasonable, representative set of parameter inputs informed by a market view and historical data.

As I described, the ESG modelling parameters are representative and are forward-looking parameter assumptions that the investment house would utilise in their investment process. I make use of these parameters as they have already incorporated multiple levels of review and scrutiny prior to being published (it is a house view). The key to an ESG model is to ensure that the input assumptions are reasonable and that they follow the SOA guidelines in design.

Figure 70: Mean reversion is shown in the chart, where the economy in this chart starts from a bust state (lowest 10% of historical observations), showing the asset class tendency to eventually slow down or revert back to its long-term average. This can be represented in a chart as a trend line that shows the economy moving upwards, but eventually levelling off or dropping back down to its historical average.

Based on the review of the list of asset manager capital market assumptions, the following key long-term parameters have been used in this analysis:

Region	**UK**		**USA**		**RSA**	
Metric ⟶	Return	Volatility	Return	Volatility	Return	Volatility
Cash & Income						
Cash	1.5%	0.6%	2.0%	0.7%	6.0%	3.0%
Sov Bonds	2.3%	2.2%	2.6%	3.5%	7.5%	4.8%
Inv Bonds	2.6%	4.3%	3.2%	4.5%	8.0%	6.3%
Growth assets						
Equity	6.8%	17.2%	6.0%	18.0%	13.0%	22.0%
Property	6.9%	15.5%	5.4%	16.0%	12.5%	19.5%
Alternatives						
DC Inv	5.8%		5.2%		8.7%	
DC Sub-Inv	7.75%		7.15%		13.0%	
Gold	5.0%	14.6	5.0%	14.7	9.0%	16.5

Table 13: Long-run asset-class input parameters (DC = Direct credit / Private debt).

My ESG model accounts for temporal dynamics of asset paths, and considers that asset classes have completely different assumptions for different stages of an economic/credit cycle using an index, plotted in Figure 71.

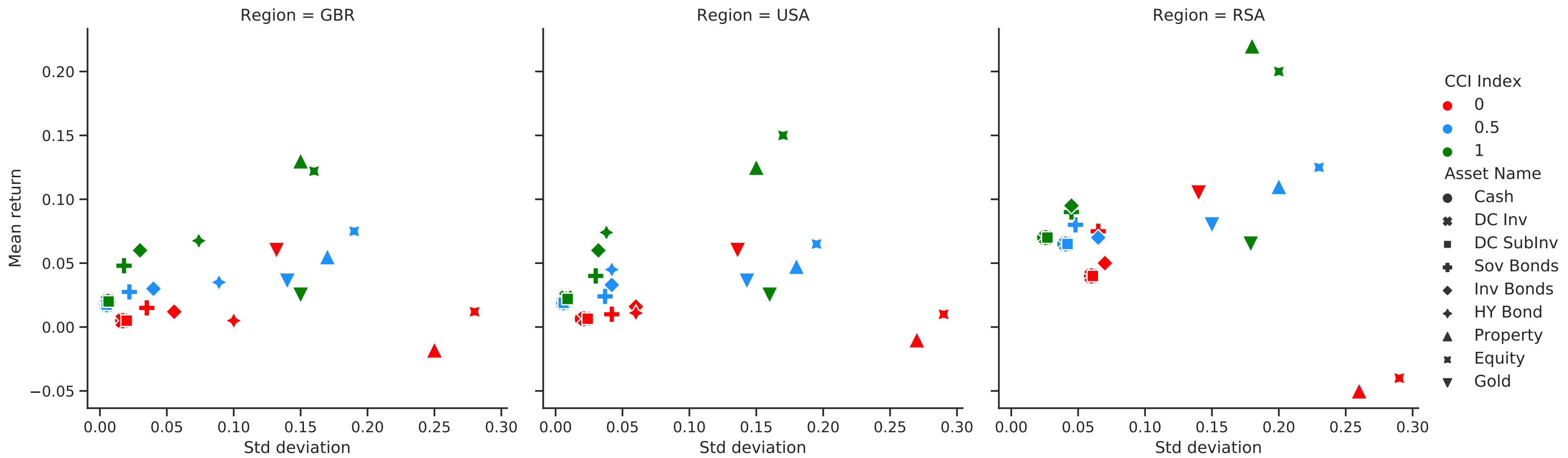


Figure 71: Return and Standard deviation assumptions are determined by reviewing a broad range of global investment managers. In this analysis I use a colour scheme to show the distinct CCI regimes, where 0% represents the worst (bust = red) credit cycle experience to CCI 100% which is signifying boom economic regimes (green). The inputs were divided by region and CCI regime (0%, 50% = E[R], and 100%). The shapes in the chart distinguishes the specific asset classes. This analysis shows some expected stylised features of asset classes, the expected risk and the return features of asset classes over the economic cycle. This layout also allows the reader to compare reasonability across regions, asset classes and cycle. The bust scenario indicates a relatively poor return from growth assets but a decrease across all asset classes. Emerging market returns are expected to be higher risk and higher return, which the Region = 'RSA' assumptions reflect.

Correlations are also provided by asset managers where available and that can be mapped from available information; the primary source of correlations in this analysis is from JPM (2022). The private debt correlation assumptions are implied from the cash assumptions, as this ESG mode builds on an expected cash plus premium plus the effects of defaults (dealt with separately in the model).

UK Assets	**Cash**	**Agg Bonds**	**FI corp Inv**	**FI corp Long**	**Equity**	**Property**	**DC Inv**	**DC Sub Inv**	**Gold**
Cash	1.00								
Agg Bonds	0.18	1.00							
FI corp Inv	-0.15	0.55	1.00						
FI corp Long	-0.13	0.49	0.79	1.00					
Equity	-0.12	0.06	0.44	0.45	1.00				
Property	-0.26	0.03	0.05	0.40	0.39	1.00			
DC Inv	0.84	0.17	0.02	0.02	-0.12	-0.20	1.00		
DC Sub Inv	0.54	0.17	0.02	0.02	-0.12	-0.20	0.85	1.00	
Gold	0.17	0.27	0.15	0.13	-0.06	-0.28	0.36	0.30	1.00

Table 14: UK asset correlation assumptions used in the ESG asset simulation, please refer to RSA and USA regions correlation matrices at Tables number 44 and 45, respectively.

19.2.3 Transform variables - PLC/LGD

The following section illustrates the key parameters used in simulating private credit assets in a dynamic time-series framework. The business cycle framework, as it has been described in a diagnostic extrapolation model proxied by the CCI parameter that I modelled in Part III.

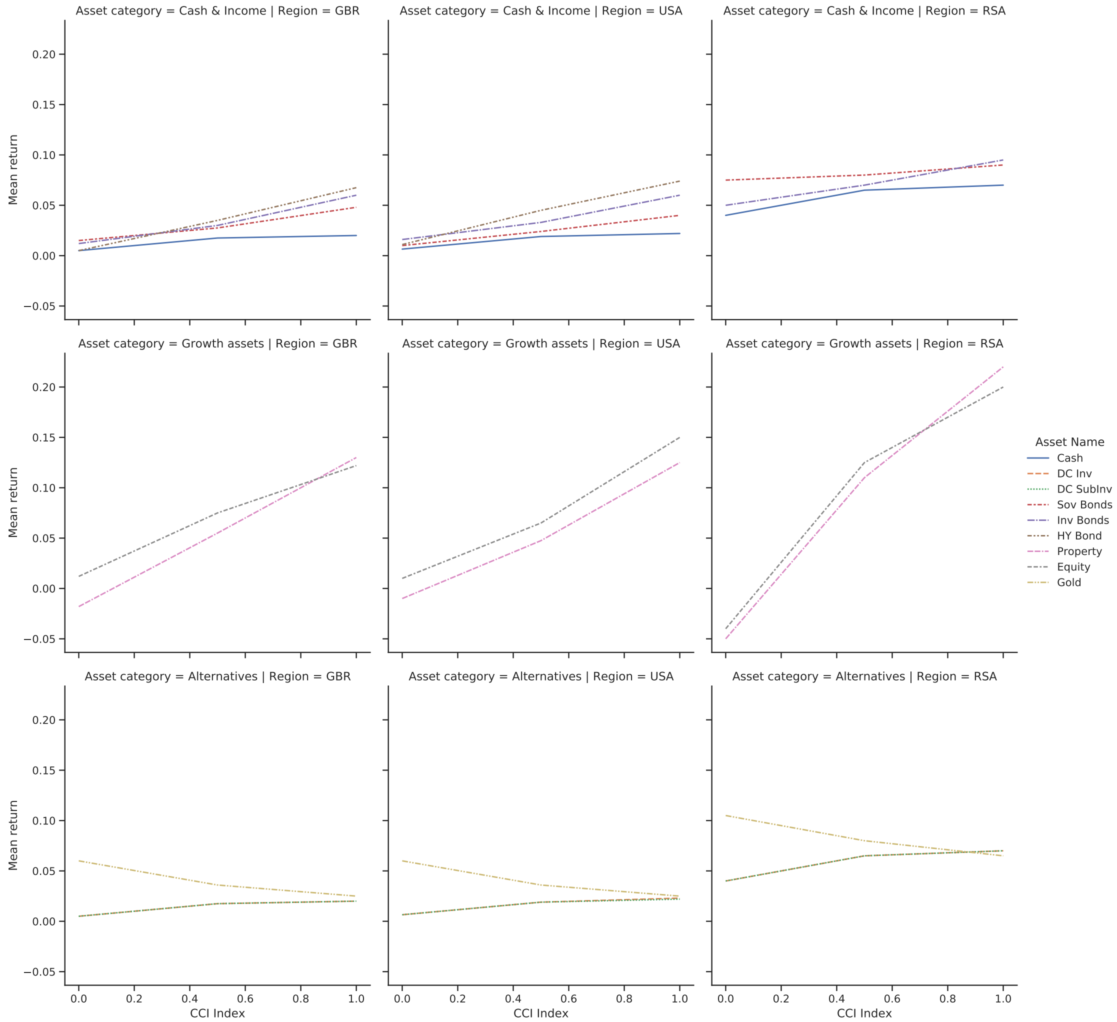


Figure 72: Return input assumptions, showing the changes in assumption per CCI assumptions, where 0% represents the worst credit cycle experience to CCI 100% which is signifying bull regimes. Key simulation and transform variables have been illustrated in the charts above. The PD and standard deviation of the PD variable is positively related to DE-CCI. The loss impact, LGD is also positively correlated to the DE-CCI. The Clayton copula theta variable drives the presence of PLC over the credit cycle. This will simulate the extent to which variables are clustered to reflect the clustering indicated by the analysis in Section 15.2.1, showing a concentration in a favourable DE-CCI and low credit spread.

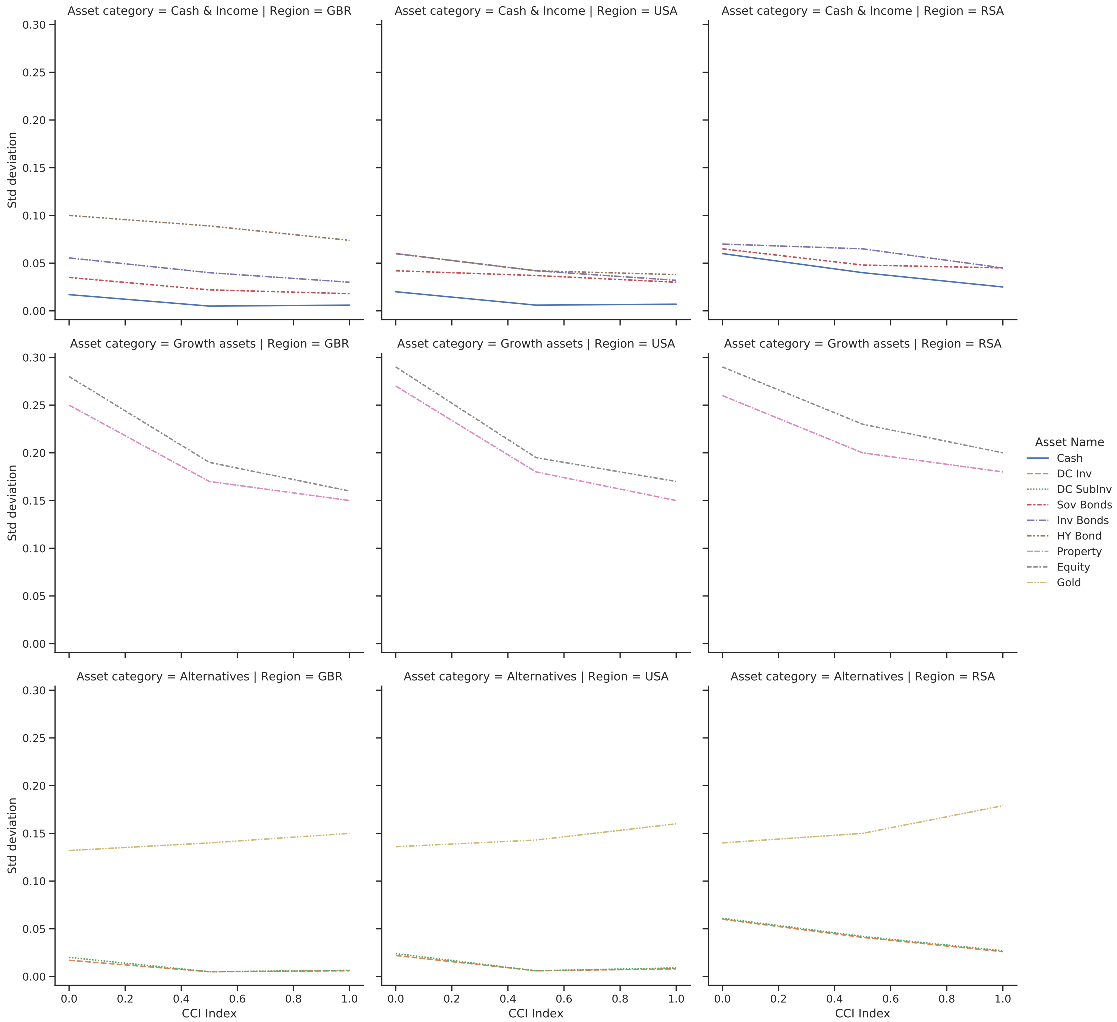


Figure 73: Standard deviation input assumptions, showing the changes in assumption per CCI assumptions, where 0% represents the worst credit cycle experience to CCI 100% which is signifying bull regimes.

19.2.4 Simulation outputs and results

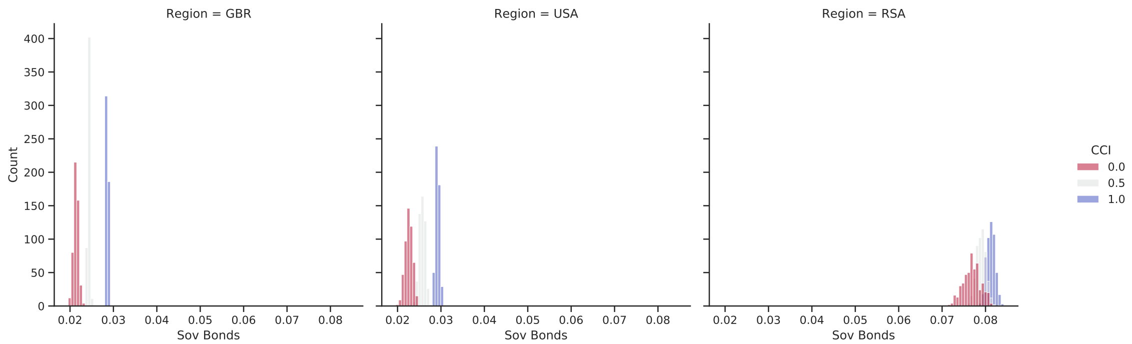


Figure 74: Distribution of sovereign bond returns output from the ESG simulation, reported for each expected phase of an economic cycle. Two clear patterns emerge, the distribution is broader in the bust than boom, and the return expectation is naturally higher in boom periods (CCI=1).

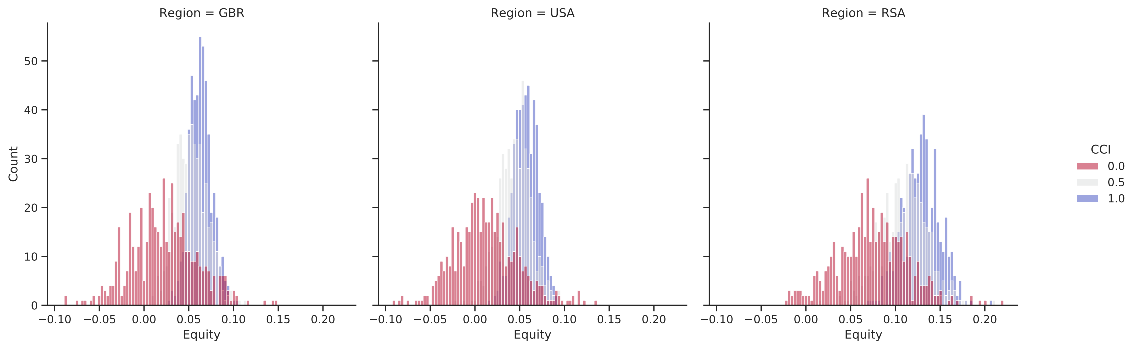


Figure 75: Distribution of equity five year CAGR returns output from the ESG simulation, reported for each expected phase of an economic cycle. Two clear patterns emerge, the distribution is broader in the bust than boom, and the return expectation is naturally higher in boom periods.

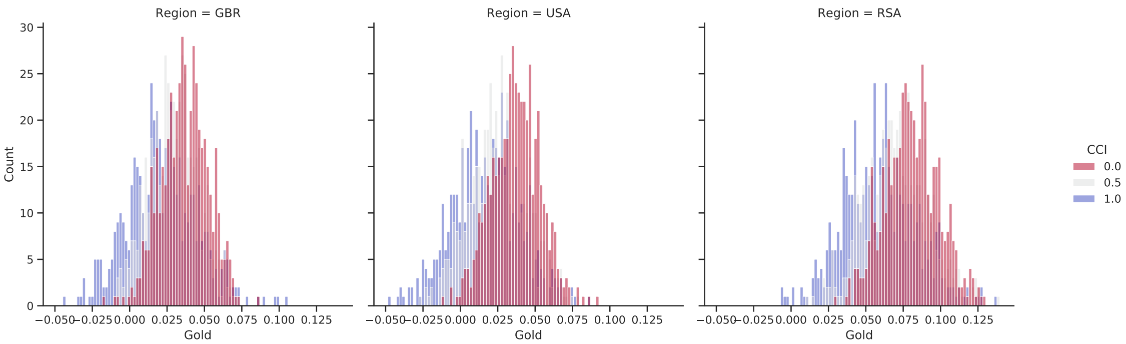


Figure 76: The pattern for gold is slightly different as it is known to feature strong returns when the market is in stressed conditions.

19.2.5 Credit parameters

Metrics ⟶ by region	CCI bands	PD mean	LGD Beta (alpha)	LGD Beta (beta)	Clayton theta	Credit and liquidity Premium
UK	**Investment grade**					
	0%	6.0%	10.0%	4.00%	0.4	4.30%
	50%	0.5%	0.70%	0.85%	2.0	4.30%
	100%	0.1%	0.60%	20.0%	9.0	4.30%
	Sub-investment grade					
	0%	9.0%	10.0%	2.50%	0.4	7.00%
	50%	5.0%	0.80%	0.85%	4.0	7.00%
	100%	2.0%	1.50%	13.0%	9.0	7.00%
USA	**Investment grade**					
	0%	7.0%	10.0%	4.00%	0.4	2.70%
	50%	0.5%	0.70%	0.85%	2.0	2.70%
	100%	0.1%	0.60%	20.0%	9.0	2.70%
	Sub-investment grade					
	0%	10.0%	10.0%	2.50%	0	7.00%
	50%	5.0%	0.8%	0.85%	4	7.00%
	100%	2.0%	1.5%	13.0%	9	7.00%
RSA	**Investment grade**					
	0%	9.0%	10.0%	4.00%	0	3.20%
	50%	0.7%	0.7%	0.85%	2	3.20%
	100%	0.2%	0.6%	13.0%	9	3.20%
	Sub-investment grade					
	0%	12.0%	10.0%	2.50%	0	4.15%
	50%	7.5%	0.8%	0.85%	4	4.15%
	100%	2.8%	1.5%	13.0%	9	4.15%

Table 15: Investment-grade and sub-investment-grade assumptions that includes the PD and LGD parameters.

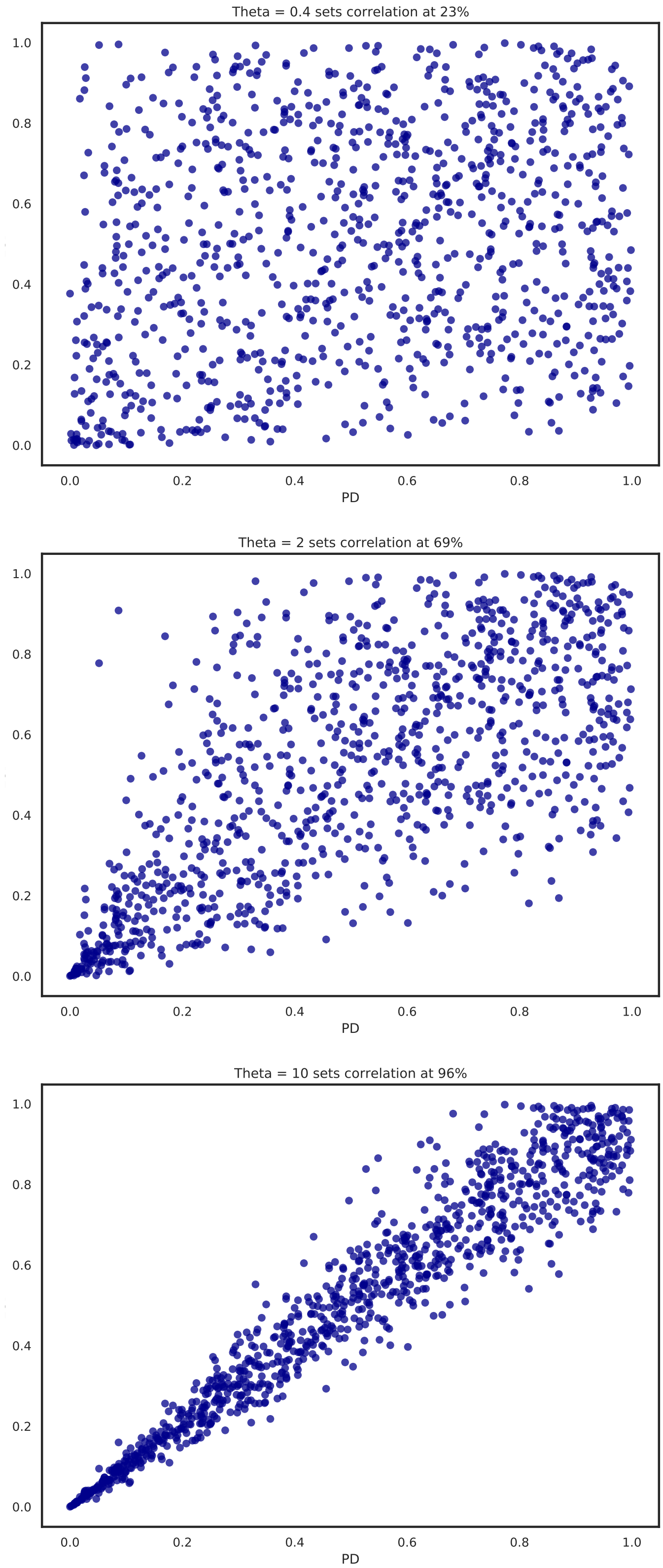


Figure 77: Having selected the Clayton copula, I use the input parameter theta to determine the correct level of linear correlation. In the chart I have shown what the effect of increasing the Clayton copula theta-variable is and also reported what the resultant linear correlation coefficient looks like.

Impact of parameter scenarios on credit-risk expected loss. ESG simulations can be a complex build, with modelling features set up to cascade with an embedded integrated cost of credit part. This includes an inverse transform sampling method to build the PD distribution that is incorporated into the loss distribution as the Poisson λ parameter. λ is dependant on the CCI parameter. This is cascaded to a dynamic cycle dependant Beta LGD, all the while dependant on credit cycle states. When simplified, results in lower expected loss values, as table 17 demonstrates, by progressively simplifying modelling approach. This is done by running scenarios and storing the results after removing features of the model. The final model uses a flat LGD and λ.

DC Inv CCI=0%	**A. full estimate**	**B. no copula**	**C. no copula, flat LGD**	**D. flat PD, flat LGD and no copula**
PD	YOY distribution	YOY distribution	YOY distribution	Point
LGD	distribution: beta	distribution: beta	distribution: single beta	Point
Copula	Yes	No	No	No

Table 16: Credit modelling assumptions for demonstrating the effect on expected losses.

From this analysis, the real risk in credit-risk modelling challenge is understanding the full extent of asymmetry in losses. If this is not correct, the risk is unanticipated losses in investment portfolio, with no chance of an upside return. In the following analysis, I show the impact on expected losses by removing modelling features.

UK CCI=0%	**A. full estimate**	**B. no copula**	**C. no copula, flat LGD**	**D. flat PD, flat LGD and no copula**
IG EL	29%	27%	25%	22%
SIG EL	37%	36%	34%	33%

Table 17: Simulated expected losses by incrementally reducing the simulation model features.

Table 17 shows a gradual decrease in expected risk for the UK Sub-Inv and UK Inv assets, which is true for the other regions. Losses will mostly occur during the bust scenario (CCI=0%), leaving the investor with unanticipated losses at risky periods. I feel this conservative approach to modelling is warranted, especially as measurements are not based on directly observable data (both in this study and for many asset managers).

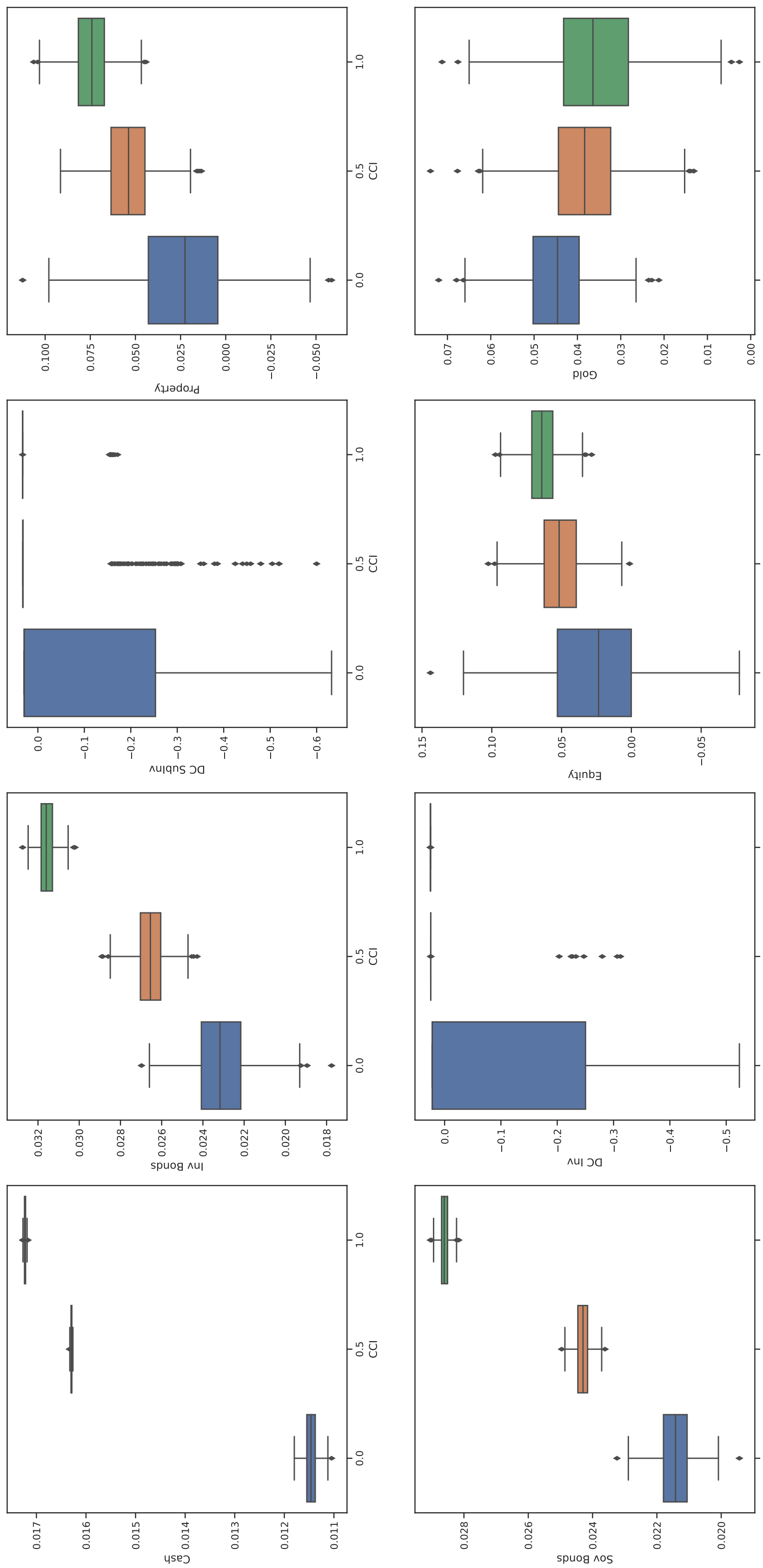


Figure 78: GBR Asset class CAGR distributions from one year of investing, split by CCI category.

Summarising the results of the simulation An ESG data model has the benefit of providing lots of integrated data to analyse, however, I needed to carefully consider what part of the data to study and how to summarise it. So in practice, the simulations are set up in a nested loop structure, with each draw generating a set of asset paths over the defined investment horizon. Always using the correlation structure imposed on the assets. Specific results are then summarised in a table structure that can be used in the analysis and portfolio modelling and to review the variables that have been modelled for correctness.

Asset ⟶ **Stats**	**Country**	**CCI**	**Scenario #**	**Cash**	**DC Inv**	**DC Sub Inv**	**Equity**	**Gold**	**HY Bond**	⋯
# defaults	UK	0.5	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
average default	UK	0.5	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average return	UK	0.5	1	0.00	0.02	0.00	0.00	0.04	-0.01	0.00
Standard deviation	UK	0.5	1	0.01	0.02	0.02	0.11	0.12	0.08	0.04
Final value	UK	0.5	1	106.5	125.3	126.4	131.9	145.2	112.2	112.3
Final CAGR	UK	0.5	1	0.01	0.05	0.05	0.06	0.08	0.02	0.02
Start Value	UK	0.5	1	100	100	100	100	100	100	100
⋮						⋮	⋮			
Looping ↻						[next	scenario]			

Table 18: Each loop in the simulation reports a set of rates and asset pathway indexes, as illustrated in the respective charts. The loops are then set for each scenario, CCI variable, and country (UK, USA, RSA), using the full set of parameters.

19.2.6 Portfolio optimisation specification

The following table is a summary of the modelling techniques, inputs and technologies that have been used in the optimisation of the portfolios and building blocks.

Model features	
Optimisation model:	CDAR optimisation
Returns:	CAGR measured over the full 5 year horizon
Risk parameter:	CDAR
Time Horizon	Five years
Model objectives:	Maximal Sharpe ratio
Constraints	Long-only, no other constraints
Portfolio building blocks	
Fixed-income (pre-credit):	Sov bonds and inv bonds
Credit Including:	Sov bonds, Inv bonds, DC inv and DC sub-inv
Profit seeking portfolio:	Property, equity, gold (set at 30% of portfolio)
Tools	
Data:	ESG simulated returns
Python packages	Python: riskfolio, Mosek, CPLEX, pyplot, matplotlib, plotly, numpy, pandas

Table 19: ESG optimisation model description, objectives, tools and data sources.

19.3 Portfolio construction results and analysis

Learning how various asset categories (growth, fixed-income and alternatives) and asset classes weightings affect one another, how their performance and risk compare, and how these factors relate to an investor's goals is essential when constructing a portfolio. In the following section I describe how the building block portfolios are made up from two blocks. The first building block is made up of the credit and fixed-income assets and the second block consists of the growth related assets. This portfolio is then combined with an optimised growth portfolio into the final SAA portfolio. I lastly provide a view of the TAA portfolio which can be created through the use of my forward looking DE-CCI indicator.

19.3.1 Credit and income building block

The credit building block is optimised on the five year CAGR asset class simulation scenarios. I make use of the CDAR technique in an unconstrained optimisation. The challenge for balancing the credit block is to set the minimum level of liquidity and credit default premium such that it covers the expected cost of credit.

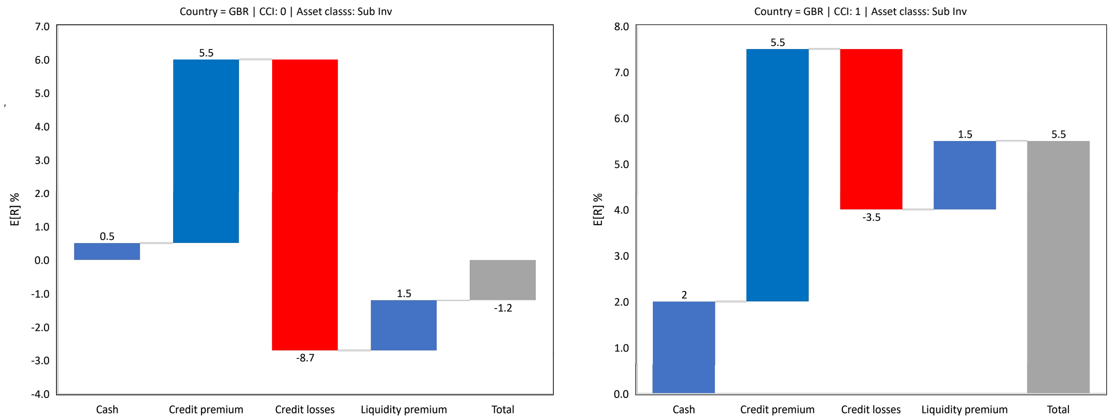


Figure 79: This illustration confirms that the cost of credit is expected to increase dramatically over the credit cycle. The excess earned in boom times needs to shield the portfolio from the excess losses that investors can experience in a stressed market.

I created a method to directly review the net impact on the portfolio with private credit introduced, relative to pure fixed-income (can be seen in Table 20). This is a comparison of two sets of portfolio results, both optimised in exactly the same optimisation program, with one including the private credit assets. The ideal scenario is that the cycle balanced credit and fixed-income block out performs the fixed-income only scenario. From Table 20, in all cases the CCI=0% produces returns that are inferior to the fixed-income only scenario, for both the CAGR over one and five years. The reverse is expected for both CCI=50% and 100% scenarios. Overall the use of private credit, given that private credit loans can be sanctioned with these risk levels (lending rates are implied by cash plus default and liquidity premium). For both the CCI=50% and 100% scenarios, a fairly high excess return is on offer for a fixed-income asset, making this proposition compelling.

The distribution charts below are an interesting visual that shows a comparison of the RSA

Credit block	CAGR (1YR)		CAGR (5YR)	
Portfolio returns	Excl. private credit	Incl. private credit	Excl. private credit	Incl. private credit
CCI=0%	**3.9%**	**3.6%**	**4.2%**	**3.4%**
UK	1.9%	1.7%	2.2%	1.8%
RSA	7.6%	6.9%	7.7%	6.1%
USA	2.2%	2.2%	2.7%	2.2%
CCI=50%	**4.9%**	**6.0%**	**4.5%**	**4.9%**
UK	2.9%	3.7%	2.6%	2.4%
RSA	8.4%	9.3%	8.0%	7.8%
USA	3.3%	4.9%	3.0%	4.6%
CCI=100%	**6.4%**	**6.5%**	**3.8%**	**4.9%**
UK	4.6%	4.7%	3.0%	2.9%
RSA	9.7%	9.9%	8.5%	8.5%
USA	4.8%	5.0%	0.0%	3.3%

Table 20: Fixed-income and credit building block expected returns.

fixed-income and credit building block expected returns over the three CCI scenarios. Using the RSA portfolio to demonstrate, I start with the boom time (CCI=0%) in Figure 80, where the impact of credit losses to the portfolio is relatively limited which is at a level that does not detract from the portfolio upside (additional earnings from the credit and liquidity premium). Importantly, there are still defaults and losses that are visible on the far left in green.

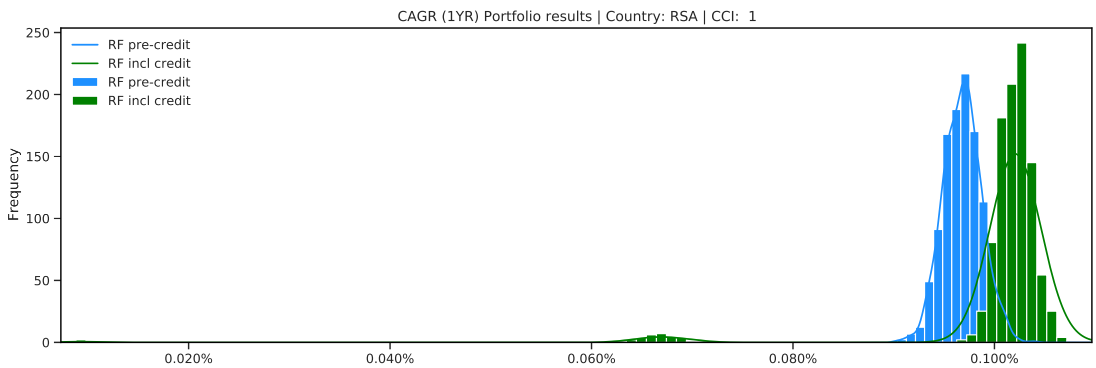


Figure 80: CAGR (1YR) CCI=100%: RSA.

As the cycle moves into a more stable setting (refer to Figure 81), I can see the impact of the losses is more significant, but can also see the bulk of the green portfolio attains higher levels of return (further to the right). Finally, as the portfolio moves into a full stress scenario, CCI=0% (Figure 82), the proportion of the portfolio in default is higher and the extent of the defaults are greater (shown by taller green bars on the far left).

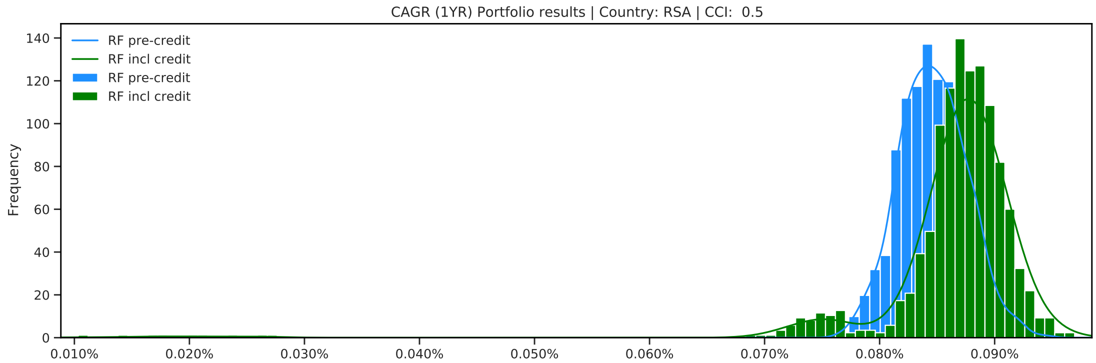


Figure 81: CAGR (1YR) CCI=50%: RSA.

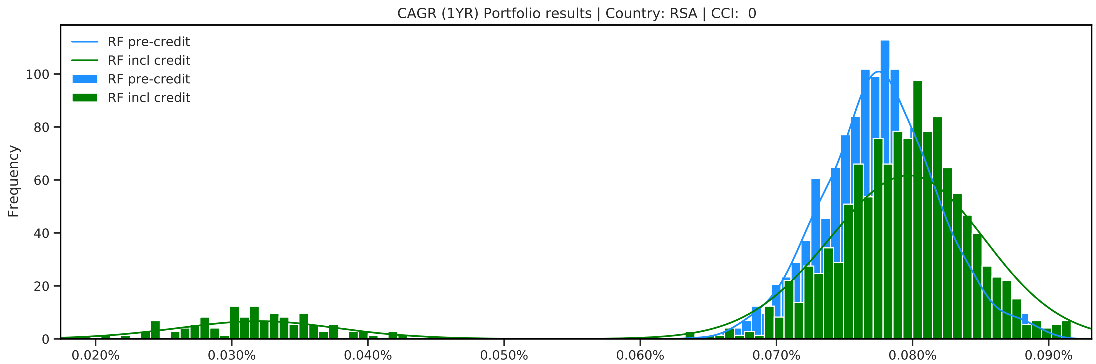


Figure 82: CAGR (1YR) CCI=0%: RSA.

19.3.2 Performance seeking portfolio (PSP)

The PSP portfolio is also optimised using the same program but the assets included are equity, property and gold only. This results in an expected portfolio result for the one and five year horizon in Table 21 below. The results are as expected (given the simulation input parameters), where the returns start higher in good years (CCI=0%) and revert toward the mean over the 5 year investment horizon. RSA, as expected, produces higher levels of compound return. What is also true, given the assumptions that this model hold, returns after a crisis (CCI=0%) require longer than 5 years to recover the losses.

PSP portfolio	CAGR (1YR)	CAGR (5YR)
UK / CCI ↓	**5.9%**	**5.0%**
0%	4.4%	3.7%
50%	5.6%	5.0%
100%	10.2%	6.6%
RSA	**10.5%**	**10.1%**
0%	9.5%	7.9%
50%	9.8%	10.2%
100%	17.3%	11.8%
USA	**5.2%**	**4.3%**
0%	4.3%	3.6%
50%	4.8%	4.2%
100%	9.3%	5.4%

Table 21: Profit seeking portfolio results over the one and five year horizon.

19.3.3 Strategic asset allocation Static portfolio results

With the credit and growth portfolio optimised and having balanced premium levels (this requires numerous runs of the simulation with setting the premium levels to the appropriate level, to establish the minimum level), I can again see the impact of including/excluding the credit from the portfolio. The strategic asset allocation (SAA) portfolio is based on the CCI=50%, as this portfolio would carry the long run assumptions and typically an expected return for the current market. The results are then calculated off the full scenario set.

SAA static	**CAGR (1YR)**		**CAGR (5YR)**	
Portfolio returns	**Pre-credit**	**Credit**	**Pre-credit**	**Credit**
UK / CCI ↓	**2.96%**	**2.97%**	**2.68%**	**2.62%**
0%	0.95%	0.86%	1.83%	1.64%
50%	2.96%	2.98%	2.70%	2.65%
100%	4.96%	4.99%	3.35%	3.34%
RSA	**7.21%**	**7.26%**	**7.00%**	**6.95%**
0%	3.93%	3.44%	5.89%	5.10%
50%	7.30%	7.40%	7.03%	7.06%
100%	9.78%	9.92%	7.83%	7.96%
USA	**2.99%**	**3.19%**	**2.68%**	**2.80%**
0%	1.17%	1.11%	1.79%	1.53%
50%	2.97%	3.19%	2.71%	2.86%
100%	5.00%	5.25%	3.34%	3.55%

Table 22: Summary of expected portfolio returns from the SAA portfolio.

19.3.4 Tactical asset allocation with sentiment signals

Given the DE-CCI proxy is a forward leading indicator, I would make use of this information to directly affect the liquid portfolio, the growth assets. This would mean that the portfolio manager would tilt the portfolio, as part of the tactical asset allocation (TAA) to capture the gains of gold in stressed times relative to the normal periods and also tilt toward the higher expected return profiles for future periods of expected gains (CCI=100%).

TAA Tactical	**CAGR (1YR)**		**CAGR (5YR)**	
Portfolio returns	**Pre-credit**	**Credit**	**Pre-credit**	**Credit**
UK / CCI ↓	**3.76%**	**3.80%**	**3.24%**	**3.17%**
0%	2.63%	2.59%	2.66%	2.61%
50%	3.63%	3.67%	3.24%	3.17%
100%	5.88%	6.06%	3.79%	3.74%
RSA	**9.10%**	**9.25%**	**8.67%**	**8.69%**
0%	8.14%	7.89%	7.77%	7.04%
50%	8.92%	9.03%	8.71%	8.75%
100%	11.55%	12.36%	9.29%	9.91%
USA	**3.79%**	**4.01%**	**3.36%**	**3.67%**
0%	2.85%	2.83%	2.95%	2.94%
50%	3.63%	3.81%	3.35%	3.57%
100%	6.03%	6.80%	3.91%	5.29%

Table 23: Summary of expected portfolio returns from the TAA portfolio with active delta.

The difference in returns is reported in this active delta portfolio. The gains are primary from capturing the effects of timing using signals, whilst capturing the net gains from holding private credit in the portfolio.

19.4 Modern portfolio-planning design and implementation

In the following section, I complete the analysis by providing a review of a powerful machine-learning technique for sequential decision making, where portfolio planning is a good example. This framework is both flexible and capable of high power optimisation techniques. I provide a novel example that is a building block for future innovation, however, this is a key and important method as it directly connects the asset-pricing model theory for optimal asset

Active delta	**CAGR (1YR)**		**CAGR (5YR)**	
Portfolio returns	**Pre-credit**	**Credit**	**Pre-credit**	**Credit**
UK / CCI ↓	**0.80%**	**0.83%**	**0.56%**	**0.55%**
0%	1.68%	1.73%	0.83%	0.98%
50%	0.68%	0.68%	0.54%	0.52%
100%	0.92%	1.06%	0.44%	0.40%
RSA	**1.89%**	**1.99%**	**1.68%**	**1.74%**
0%	4.21%	4.46%	1.89%	1.94%
50%	1.61%	1.63%	1.68%	1.69%
100%	1.77%	2.44%	1.46%	1.95%
USA	**0.80%**	**0.82%**	**0.68%**	**0.88%**
0%	1.67%	1.73%	1.16%	1.41%
50%	0.66%	0.62%	0.64%	0.70%
100%	1.03%	1.55%	0.57%	1.74%

Table 24: Summary of expected portfolio active returns from the TAA portfolio.

pricing and portfolio optimisation.

19.4.1 Context for my portfolio choice problem

Pension fund investments. The Society of Actuaries (2016) explain that the pension fund investment is based on a larger range than that in the insurance framework due to the different regulatory constraints in place. On the liability front, there are more factors that concern the liability values, such as wage inflation, consumer price inflation, employment factors, retirement indicators, and disability rates, all of which are affected in a cash-flow analysis.

19.4.2 Reinforcement learning introduction

Reinforcement learning (RL) is a general class of algorithms in machine-learning and is a natural fit for the kind of problem in which the end result or reward is a function of time; such can be delayed from the point in time at which the decision was made. As part of a core long-term investment program, such as our pension fund problem, I use a deep reinforcement learning algorithm to solve the portfolio choice problem. The core of RL is to learn from experience and make improvements. As I will explain, the core component of this modelling framework for sequential problems and stochastic planning, such as long-run asset allocation, is successfully modelled using an Markov decision process (MDP) at its core (Silver, 2015). This modelling framework became the de-facto model for sequential planning. The introduction to the RL framework follows Silver (2015). Chen (2019) explains that RL models can be categorised in two approaches:

- **Model-based RL** that takes into account the experience from Markov, or conditional probabilities and a policy is utilised. Dynamic programming (also known as policy iteration and value iteration) is a classic example of a model-based algorithm. Such uses the model's predictions or distributions of the subsequent state and reward to choose the best course of action. In Dynamic Programming, in particular, the model must include transition probabilities between states and predicted rewards for each given state and action combination. Keep in mind this is not typically a learnt model.
- **Model-free RL.** The model free RL provides various policy functions; and learning takes place by means of a search function, such as a gradient ascent. Model-free RL algorithms, include methods such as the Monte Carlo Control, SARSA, Q-learning, and Actor-Critic, learning solely via sampling from experience. Instead of on computer-generated forecasts of the future state and reward to guide their actions, they depend on data collected directly from the environment (although they might sample from experience memory, which is close to being a model).

This category is strongly aligned with the machine-learning approach. The RL is a construct that makes use of existing techniques. The elements include an agent, a reward (or loss), and an environment. By interacting with the environment at each state, the agent recognises some reward and thus learns. I briefly introduce these components below.

19.4.3 Reinforcement learning model component implementation

Markov decision process. Given the uncertainty, or stochastic nature of asset classes in financial markets, it is natural to use a Markov decision process (MDP). The MDP is an extension of a Markov chain that allows for actions partly under the control of a decision-maker, or agent. A key component of the RL model formulation is the MDP. In my model setup, the agent interacts at discrete time periods with the environment (also known as the 'agent state'). At each point this is known as the state, S_t at time t. The agent will consider the state and based on this information, will take an action A_t. The net result of the action,

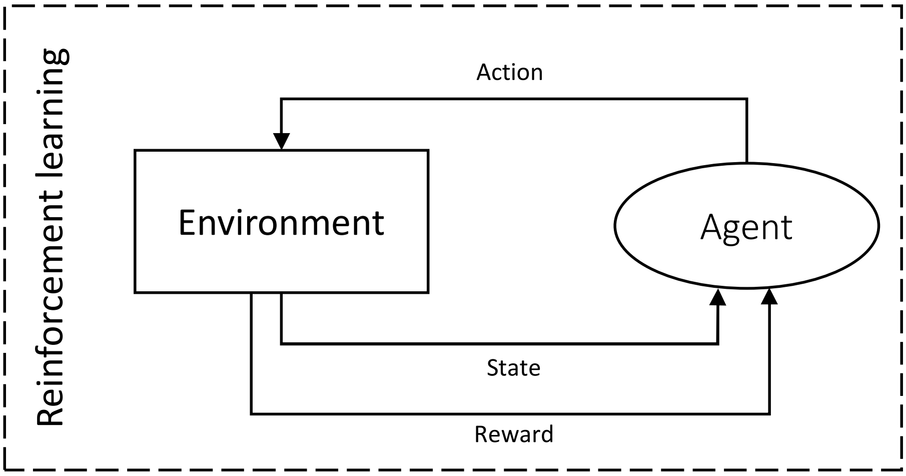


Figure 83: The goal of a reinforcement learning model, demonstrated in this chart, is to develop an agent, aware of its environment, that performs actions at a sequential state and assesses the value of a reward (or loss) with the aim of maximising this reward.

A_t, is that the agent receives a reward, R at $t+1$. The key with the MDP is that the agent gains an awareness of its own state due to the MDP, thus applying a preference of actions into the next so-called "decision epoch" (Chen, 2019). Remembering that the agent will seek to make a cumulative set of decisions that optimise the cumulative loss function. The goal of MDP is to find the best policy, one that attains the highest reward. The MDP is best described as the probability p of each reward r and state s combination, after having taken an action a:

$$p(s', r|s, a) = \Pr(S_{t+1} = x', R_{t+1}|S_t = x, A_t = a) \tag{191}$$

As Chen (2019) explains, the goal of an MDP is to solve for the optimal policy π^* that translates to the maximum reward from executing the policy. RL is a process of learning in which problems are solved by testing each state, accounting for uncertainty and differing scenarios. For agents basing policy learning on simulations, this is known as offline learning. A policy may be defined:

$$\pi^* \doteq \arg\max_{\pi} v_\pi(s) \quad \forall s \in S \tag{192}$$

where $v_\pi(s)$ is the value function that is the result of actions (a) from an agent following a policy π at each state s.

State space. As part of an MDP, a state space S contains all states s, that an agent can transition into, so $s \in S$. This will be defined for each time period t over the horizon K, so $t \in 0, 1, 2, 3, \ldots, K$.

Action space. The set of actions, $a \in A^K$, is used to control the system state. Each action a taken at time t is the action space, A. Chen (2019) explains this is dependent on the agents current state, denoted as $A(s_t)$.

Transition function. Once an action a has been applied to a state s, the system will transition from state s to a new state s'. This is often associated with the probability of moving from one state to another, across a range of variables, describing a probability distribution. In this MDP, the key is that the system is Markovian in that it depends on historical states, not just the current state.

Value function. This is the prediction of the future reward. As the agent progresses through each epoch, or time period in the horizon K, there is an action required for each state. This is captured in a decision rule δ_t, which maps each $s \in S$ into action $a \in A$.

Reward function. Is where $v_\pi(s)$ is known as the Value function (Chen, 2019). The expected reward in future is therefore simply defined as:

$$v_\pi(s) \doteq \mathrm{E}_\pi\left[\sum_{i=t+1}^{K} R_i \middle| S_t = s\right] \tag{193}$$

For the MDP process to converge, γ is used and lies between zero and one, $0 < \gamma < 1$ and this serves as a discount factor. By splitting out the first state in the value function (dependent on policy π), as follows (Chen, 2019):

$$v_\pi(s) \doteq \mathrm{E}_\pi\left[\sum_{i=t+2}^{K} \gamma^{i-t-2} R_i \middle| S_t = s\right] \tag{194}$$

The function can also be extended to work for state-action pairs (s, a) as follows:

$$q_\pi(s, a) \doteq \mathrm{E}_\pi\left[\sum_{i=t+1}^{K} \beta^{i-t-1} R_i \middle| S_t = s, A_t = a\right] \tag{195}$$

this is also known as the action-value-function for policy π. Summarised by isolating the sum of expected rewards, G_t, denoting the sum of reward $\sum_{i=t+1}^{K} \gamma^{i-t-1} R_i$, this changes the value function to:

$$v_\pi(s) \doteq \mathrm{E}_\pi\left[R_{t+1} + \gamma G_{t+1} \middle| S_t = s\right] \tag{196}$$

Bellman equations. Now, moving on, Chen (2019) explains the connection to the Bellman equation, which I use to set in the context of a reinforcement learning model. This requires describing the second term, letting the policy $\pi_{a|s}$ be the probability of action a, for states $S_t = s$ and letting reward $R_t = r$ and the next state be determined by function $p(r, s'|s, a)$, this equation becomes, following Chen (2019):

$$\begin{aligned}
&\sum_{a \in A(s)} \pi(a|s) \mathrm{E}_\pi \left[R_{t+1} + \gamma G_{t+1} \middle| S_t = s, A_t = a \right] \\
&= \sum_{a \in A(s)} \pi(a|s) \sum_{s'} \sum_{r'} p(r, s'|s, a) \left[r + \gamma E_\pi [G_{t+1} | S_{t+1} = s'] \right] \\
\Rightarrow v_\pi(s) &= \sum_{a \in A(s)} \pi(a|s) \sum_{s'} \sum_{r'} p(r, s'|s, a) \left[r + \gamma v_\pi(s') \right] \qquad (197)
\end{aligned}$$

As Chen (2019) explains, Equation 197 is the Bellman equation for $v_\pi(s)$, showing that the optimal rule S_t is the choice of actions that will maximise the expected reward plus the discounted future state. There are several ways to solve an MDP, including linear stochastic programming as part of a dynamic programming framework. However, the standard method for large systems is reinforcement learning (Abrate et al., 2021). Chen (2019) explains that the collection of decision rules is referred to as a policy π and the ultimate goal in an MDP is to find the optimal policy. The optimal policy can be achieved with a model which will solve it directly or by making use of a model-free environment. This environment does not rely on the perfect model; but rather uses simulation and policy iteration to find the optimal model, known as iterative policy evaluation. The final step is to refine the policy, thus finding the best model. This is often achieved through a combination of exploitative and exploration techniques that are part of a deep reinforcement learning solution.

19.4.4 RL model definition and implementation

Now I deal with a defined-benefit asset liability problem, in which we are looking to understand the best policy for investing, selecting between two asset building-blocks options (fixed-income including private debt and a portfolio-seeking portfolio), given different points in the economic cycle. Asset values, are however, compared to the liability values. This is a dynamic problem as the following are core drivers of liabilities, following Shang (2015).

- **Discount rate**. The future value of the portfolio is specifically affected by the discount rate. This is often the sovereign rate of debt reflective of the cost of finance in the market. This is also an investable asset on the asset side, thus becoming an integrate analysis. This is a key driver in this study.

- **Inflation**. This is particularly important when we consider medical benefits payments, where the cost of medical is more closely linked to the consumer price index. This is not a focus of this study.
- **Claims rate**. This is affected by the mortality rate, morbidity rate, the drawdown of pension amount and more closely aligned with insurance asset owner characteristics. This is not a focal point in this study.
- **Policy behaviour**. Such as lapse rate as part of a long-term insurance policy. This is not a feature of this analysis.

The following table is a summary of the features of the reinforcement learning Q model, inputs, and technologies that have been used in the optimisation of a portfolio planning model. The purpose of the model is to understand how to fairly allocate the fixed-income portfolio and the PSP portfolio. This has been provided per a model-free environment, using the Bellman equation.

I defined a reinforcement model to assist with how we understand the optimal investment strategy, π, based on S reflects the states of at the time decisions (π) by the agent. In our analysis the states are the CCI levels that I have been modelling. The strategy is defined as the actions that maximise the reward function $Q^*(s, a)$, where a is the available actions, determined by the strategy $\pi^*(s)$, following Shang (2015):

$$\pi^*(s) = \max_a Q^*(s, a) \tag{198}$$

Reinforcement learning programming I coded the RL model using the Python logic in the Bellman function from https://www.datahubbs.com/reinforcement-learning-markov-decision-processes/. The ESG model pathways provides asset-pricing pathways needed in to create the transition probability matrices. I also utilised Python and Python packages as part of a Colab hosted program that I coded. I provided details of the address of the Google Colab files in the Appendix; Section 30.

19.5 RL modelling results

An important consideration in ALM modelling, in the case where goal is to understand the assets relative to the liabilities, when we consider how the liabilities will vary, and this is to

Model features	
Optimisation model:	Bellman optimisation
Returns:	Distribution of returns over a one year horizon, analysis covers the full 5 year horizon. Liabilities are configured as part of this analysis.
Risk parameter:	Risk adjusted over full distribution
Time Horizon	Five years
Model objectives:	Expected Returns over full distribution
Constraints	Long-only, no other constraints
Portfolio building blocks	
Credit Including:	Sov bonds, Inv bonds, DC inv and DC sub-inv
Profit seeking portfolio:	Property, equity, gold
Tools	
Data:	ESG simulated returns
Python packages	Python: numpy, pandas matplotlib, sklearn LinearRegression, scipy.stats

Table 25: ALM Bellman optimisation model description, objectives, tools and data sources.

do with the horizon of the cash-flow expectations.

The analysis shows that when we consider the planning problem in an ALM setting, the reinforcement techniques that I applied show that even considering the downside distribution, an investor would benefit from investing in performance seeking portfolios. This is understandable if we consider that a lot of the portfolio risk is reduced by the diversification benefit from the inclusion of gold and property into a portfolio seeking portfolio. We have already established that private credit can improve the fixed-income building block. Now the modeller can base the modelling on a risk budgeting approach, where the level of profit seeking portfolio contribution is maximally allocated for a given constraint of risk (that may be determined as volatility, drawdown and liquidity). From a portfolio planning and construction perspective, this is a simple balancing process of risk consumption with an asset owner.

20 Summary of analysis

The models used in credit have been well researched and documented: this is clear from my broad overview, I also detailed the context for incorporating credit modelling from the

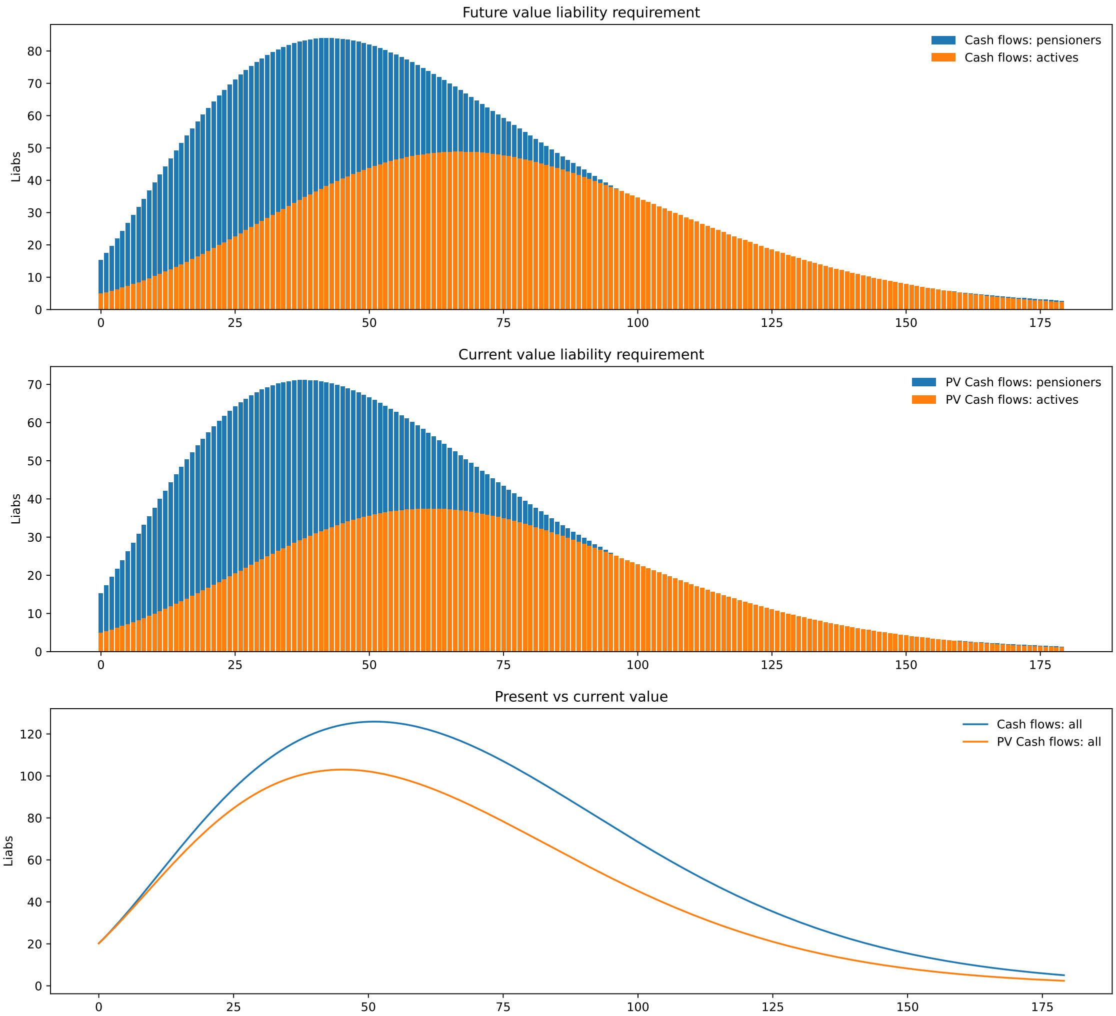


Figure 84: Defined-benefit pension liability profile, showing a typical set of liabilities for pensioners and those pension members still in the workforce (actives). This liability profile has been modelled using a simple Chi-Squared distribution.

investment management perspective. To solve the problem of asset allocation for longer term asset allocation, that considers a credit cycle, I provided a framework that can be used to understand the relative risks of private debt asset building blocks relative to core assets in an SAA. The results of this analysis show that there is indeed a benefit to a correctly priced portfolio of private credit as part of a strategic investment program. Simulation of asset classes, making use of models that cascade information in a such way that both the dynamics of the cost of credit, whilst accounting for cyclical effects on all asset classes, can be handled (as these are integrated). As with all portfolios with limited data, the experts' role in the input assumptions is vital; but equally important are the features incorporated into the

Region	CCI	25%	50%	75%
GBR	0%	1.81%	1.90%	1.98%
GBR	50%	2.77%	2.81%	2.84%
GBR	100%	4.23%	4.26%	4.28%
RSA	0%	7.91%	8.21%	8.56%
RSA	50%	8.51%	8.69%	8.88%
RSA	100%	9.30%	9.43%	9.59%
USA	0%	1.52%	1.65%	1.78%
USA	50%	2.54%	2.64%	2.74%
USA	100%	3.72%	3.79%	3.85%

Table 26: This table reports the liability discounting percentiles (25^{th}, 50^{th} and 75^{th}). The rates are based on asset manager capital market assumptions. The interesting factor here is that the variance in rates is expect due to changes in the CCI rather the percentiles of rate within a CCI percentile. Please refer to Section 29 in the appendix to see further illustrations of the liability distributions.

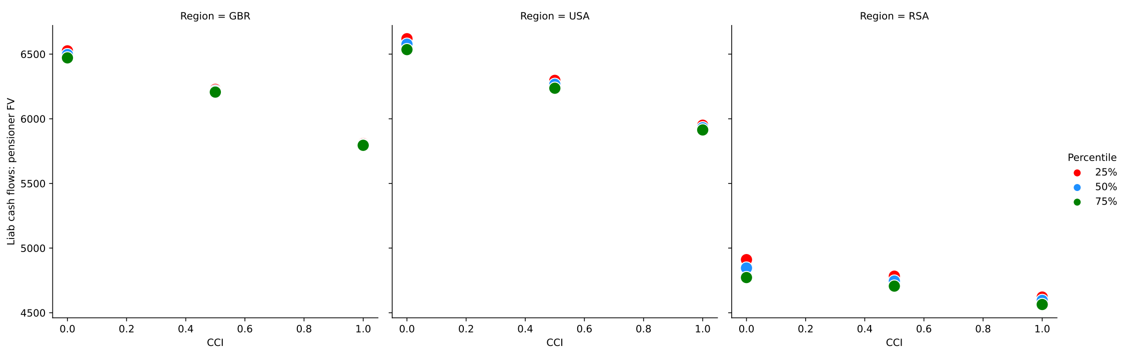


Figure 85: Once the liability cash flows are discounted, using the different scenarios reporting in Table 26, the effects to liability values is clear. When interest rate are expected to be higher (CCI=0), the liability values are lower and vice versa. This is in contrast to the asset values that tend to be higher in boom times (CCI=1).

modelling of the credit costs. As I demonstrated, if the true cost of credit is not understood, the investor will effectively import a structured loss into the portfolio with no real chance of recovering that loss. In particular, when cycles affect many parameters simultaneously, there is potential for significant changes to the expected losses in credit. Moreover, an informed

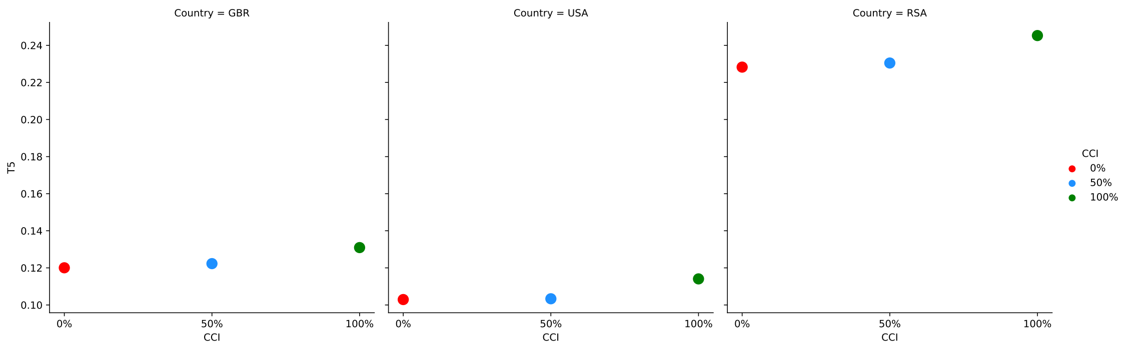


Figure 86: Asset returns Bellman Q values. This chart indicates the optimal values from using the Bellman function to establish the optimal fixed mix allocation of the portfolio seeking portfolio. In each case, this short term analysis yields the same result, 100% allocation to the portfolio seeking portfolio. What is also clear in this analysis is that the variance in portfolio value is attributed to changes in the cycle indicator, as we would expect, the lowest values are in bust periods (CCI=0). The implication here is that, unless one has a risk constraint, the best allocation strategy is to add as much portfolio seeking asset as possible.

signal that is related to the credit cycle allows the investor to take advantage of an active allocation program for liquid assets. I demonstrated the value this can offer the portfolio. The sub-investment portfolio should allow the lender to access greater premiums (base rate plus premiums) than the investment portfolio would be able to bear (obligors may well have choice), but again, this is dependant on the modellers understanding the correct expected cost of credit.

One final thought, the simulation approach allows the modeller (of both credit and investment) to integrate a complete analysis of their investment. Because the underlying correlations have a framework on which to model, the analysis can be implemented in a way that is repeatable and aggregated. The modeller would not need to use credit assumptions that are designed for banking portfolios with the assumptions of large homogenous portfolios and banking risk management (Trueck and Rachev, 2009; Engelmann, 2021) and therefore not have to make adjustments for concentration and correlation.

Part VI

Conclusion

In conclusion I will highlight my innovation and development for the key topics of the study. I believe that prior research and up to date methods have been thoroughly and clearly demonstrated in this study. I put together an appropriate framework showing how to include private debt assets in a SAA program in a meaningful way, whilst benefiting from contemporary methods and toolkits. I summarised the parts of this study below.

21 Macro-finance and credit cycles

Business cycle theory, as it pertains to asset pricing and macroeconomics, is a complex and interesting field. For investment managers, the data shows that it is clearly a fundamental driver of returns in the market. As I show from models in macro-finance, there are powerful frameworks that seek to answer questions regarding what drives the fluctuation in asset prices. I found these to be helpful constructs for thinking about modelling when the modeller has limited data. My literature review shows that the correct approach is not yet settled. Depending on the modelling context, various different frameworks can be appropriate. Thus important decision-making entities, such as monetary authorities, rely on varied models, however really, some of the model assumptions require an update. Chen et al. (2021) observe that sentiment is influenced by macro-variables which, in turn, influences macro-variables. For this reason sentiment is seen as the bridge between financial markets and macro-variables. We understand from a review of the literature that credit cycles and credit puzzles can be predicted with reasonable certainty by understanding how sentiment is linked. An important feature of this predictability is the human component, I investigate the theories of how humans extrapolate information and collectively react to market settings based on the recent past and on sentiment.

We must understand how business cycle peaks or troughs are amplified by diagnostic expectations, this cycle amplification further adding to sentiment. A spiral is then set up, which in a case of a stress event is based on fear. This highlights how the human bias cannot be completely ignored in macro-economic modelling and especially in pricing of financial assets. Cochrane (2017) explained it very well, that recessions are not times at which individuals just happen to consume less, having the view that the future is a better time to consume. Rather the individuals are in a state of real fear and are concerned about losing a job or seeing their business fail. Another key learning is what happens in a crisis, and the residual excess held for caution as part of the investment risk premium, which goes some way to describe asset puzzles. Taking into account these powerful factors and their respective links to credit cycles is a vital theme in this study. An investment manager utilizes asset-price modelling to predict market prices, focusing on predicting market trends rather than solely focusing on exogenous shocks. I show that existing macro-economics frameworks and asset-pricing models can effectively cater to these features and models using endogenous variables. In this study I explained the various different modelling frameworks and their merits. Lastly, I suggested a preferred approach in the context of private debt.

This section is finalised by introducing the core concepts for individual optimality and reinforcement learning techniques. These models are in-fact based on the same mathematical modelling, MDP and the Bellman equations are central to both models that optimise a value policy at a future date. Such methods are primarily used in the modelling of ALM portfolio approaches that I utilise in the portfolio modelling sections in Section 19.1. The key aspects of macro-economic links between the real economy and financial markets as they relate to forward-looking risk signals for private debt have been researched and I provided an appropriate framework to incorporate in an SAA program.

22 Sentiment, NLP and ML

This section is my investigation of how to effectively build indices for use in business cycle modelling and investment management of private debt as an asset class. I made use of freely available information and tools only, I found this approach revealing, in understanding the use of alternative data and tools such as Google's Colab. I invested a great deal time and effort in fully understanding the practical application of contemporary modelling techniques, such as ML and NLP for factor modelling, ascertaining their relative strengths. Part of this was to review the alternative sources of data and how much information the respective data sources held. From the analysis I found news data to be easy to work with, whereas collecting Twitter data can be seriously challenging, required streaming, a disproportional amount of cleaning for use in NLP with little to no benefit in the modelling. By contrast, the signals provided by Google's GDELT provided the clearest sentiment dip during the pandemic that other news sources did not display (a fascinating analysis will become available as GDELT data history lengthens and time-series analysis becomes more realistic). Overall, news sources provide relevant signals, especially when reviewed within a particular geography.

When I compared modelling techniques in a time series modelling review using my DE-CCI proxy to forecast sentiment data, I found that LSTM models were a stable and therefore a good choice. What was also clear is how much impact well-known financial data features such as autocorrelation has on the modelling, compared with the impacts due to advanced modelling techniques. The use of simple ARIMA factors as part of the ML framework provided considerable improvements to modelling results, certainly when compared to machine-learning models alone. This analysis highlights the limited data volume needed to improve the predictivity of ML models, despite their attractiveness and power. This was reflected in that the best modelling result in my analysis was the combination of an ordinary least squares model with ARIMA factors.

From the perspective of natural language processing to extract information from a volume of unstructured data, I believe that this analysis was merely scratching the surface. Given time and a specific exploratory focus on the unstructured data, freely available data has the potential to reveal more information. To summarise, I investigated contemporary NLP techniques for measuring sentiment and have successfully applied these techniques to identify forward-looking signals for credit-risk measurement and business-cycle dynamics, using related social media and news. To conclude on this section, I recognise an opportunity to reveal information using NLP on alternative data sources, whilst respecting that financial data features, primarily autocorrelation, have significant bearing on how one should approach the modelling of financial time series.

23 Credit modelling

In this important section I establish the core methods, models and theories needed for credit modelling of private debt. I start with defining the key modelling parameters (PD and LGD) and their measurement frameworks (PIT, TTC and DT). I provide a detailed review of the Basel II banking methodology and assumptions coupled with core credit modelling techniques in use today. I have shown in my research that a key assumption in these banking models may not be appropriate for asset management and specifically SME lending. This can be an issue when this is applied to unlisted private debt in the form of loans. One specific assumption of real concern is that the assumption of independence between the PD and LGD parameters over the different phases of the credit cycle. I go further into detail when there is also a breakdown in assumptions in stressed markets, known as correlation breakdown.

This study has much to do with establishing a robust modelling framework in asset management for sparse data modelling environments. I found that the reduced form credit modelling has been used in very similar modelling studies and is well suited. I again show that there are methods employed in this modelling environment that allow for endogenous variables to account for different economic phases or parts of the credit cycle, whilst also accounting for a financial market mean reversion in asset prices. I also found that the simplification brought about by copula modelling, in particular the Clayton copula, is appealing. This modelling technique allows for a relatively easy way of reflecting important features of the relationship between PD and LGD in a stressed market environment (such as correlation asymmetries).

I modelled a set of CCI proxies, in Part III, that have been used to investigate which proxy best differentiates the credit cycles and a representative sample of company level credit data for each of the three regions. Specifically, I model the relationship between my sentiment measure based on news signals, as a simple measure of reaction based on news sentiment confirming the time series relationship with the expected credit spreads. It turns out that the NLP DE-CCI proxies perform better than the survey based OECD CCI data. The final distribution of the expected credit shows the forecast of widening of distribution and longer tails as the cycle indicator moves toward a stressed scenario, as determined by the CCI proxies modelled in Part III.

24 ESG and portfolio summary of analysis

I defined and built an ESG model to simulate future economic scenarios that drive prices between asset classes, including private debt, and liabilities. These asset classes all have interest rate sensitivity at their core and are based on a cascaded modelling approach that provides a comprehensive view of potential extreme values for both assets and liabilities. This modelling framework is appropriate for use into building a picture in the SAA modelling and portfolio construction phase.

I also provided detailed research on how to model private credit assets appropriate from the investment management perspective. It is clear in my analysis that modelling credit to account for correlation of PD and LGD is crucial; this gives the modeller a more accurate view of the risk, which is conservative. Part of this challenge for asset management planning is to solve the problem of asset allocation for longer term asset allocation, that considers a credit cycle. I provided a framework that can be used to understand the relative risks for private debt asset building blocks that evolves as assets mean revert. At centre of this is a credit premium that is dynamic over the credit cycle.

I conducted extensive research and developed a modelling framework that gives a modeller confidence that the credit assumptions are not designed for banking portfolios. Banking credit modelling frameworks are based on a core assumption of large homogenous portfolios formulated for use in bank capital management functions. I also included my DE-CCI proxy as the forward looking signal. This signal is used to assist the portfolio manager to identify the onset of a credit cycle shift, reacting to this accordingly. I showed that active portfolios using this information are likely to attract higher rewards.

The final part of my analysis is to review the building block allocation from the perspective of ALM planning. The problem is formulated as a policy problem using transition matrices that inherently rely on the MDP, solved using the Bellman equation. I employed the machine-learning reinforcement framework to solve the final problem. Across the cycle, the assumption set that I developed show that the portfolio construction requires maximal levels of PSP portfolio allocation. To put this into context, the results are totally unconstrained. Results should be seen as the allocation in which investors do not have specific risk criteria, such as volatility or drawdown constraints.

To summarise, the adoption of private credit is not only possible, but is a benefit from a factor modelling perspective. The return that this type of portfolio introduces is not, for structural reasons correlated to with other investment asset classes. The risk premium is available and dynamic over a credit cycle and when it is modelled appropriately, it can be efficient. The market signals that I researched are powerful in that they bring together the dynamics from a macro-finance perspective that can be used as information for all asset classes in the universe.

25 Limitations and opportunities for further research

This study has its natural limitations and some areas that I see are areas that can be interesting to research. I listed these below.

- **Credit data shortage.** This study is based on freely available credit risk data; it does not include any information on credit losses from the sale of collateral at different points in the business cycle. Data companies have rich credit data based on bank lending, with a reasonable time-series across different geographies and product types. The use of this data for research is available, but not to all. An area for future research would be to understand the differences in losses between credit exposure for tradable and non-tradable companies, I believe this would be an interesting avenue of research.
- **Liquidity data.** No specific information is used to model of liquidity for SMEs at various points in the credit cycle. The liquidity measures are a significant portion of the cost of credit as part of the risk premium. Portfolio construction with added information on liquidity risk would an important risk characteristic to understand and monitor, yielding insight to the portfolio construction process.
- **ALM modelling.** ALM modelling in this study is not performed using historical cash flows and liability profiles, which would be propriety information from a pension fund or life book at an insurer. Monitoring the integrated risks over the various points of the credit risk cycle would be an appealing avenue of research and a true test of the machine learning technology when applied to ALM problems.
- **South African news data.** The news publications in this study are not specific to South Africa when I extract data for NLP analysis. Data for South African newspapers are not accessible with a reasonable historical window.

Possible areas of future research:

- **NLP news signal mining.** Information that is held in news data is not mined to its full potential. Obligor specific research that investigates the cycle impacts and how this can be understood from an asset pricing perspective would provide for more information to model the risk premia.

- **GDELT signal mining.** The utilisation of GDELT in this study was high-level and for specific purposes. A full investigation of the available information in GDELT, especially if it pertains to specific security names and comparison of the securities for selection by using cross sectional analytics would provide interesting avenues of research. The use of this information to build credit and market cycles would be a very interesting area of research. Please refer to GDELT (2020) for meta-data on the GDELT database and publications based on the use of GDELT data. A good example of GDELT data in use for a sovereign bond study is found in Consoli et al. (2020).

- **Machine learning techniques to further inform risk premia.** The task of understanding risk premia will always be a key factor in SAA analysis and modelling. Further research and methods to understand these factors, such as Iworiso and Vrontos (2020) can provide interesting steer for further research. If coupled with data from credit consortiums; coupled with the appealing data that is now available in GDELT, will be more interesting as the length of time-series data increases.

- **Private data attribution.** Both the methodologies and available data for private companies is improving. Given the overarching focus of private assets in this study, further improvements by way of private data or risk attribution methods, such as in Brown et al. (2020), will be interesting and provide value from a market practitioner perspective.

Part VII

Bibliography

References

Abdulrahman Naef Farhan, A. and Varghese, P. (2018). Facebook utilization and Arab spring movement: A study among Yemeni youth. *International Journal of Social Sciences and Management*, 5:5.

Abrate, C., Angius, A., De Francisci Morales, G., Cozzini, S., Iadanza, F., Puma, L. L., Pavanelli, S., Perotti, A., Pignataro, S., and Ronchiadin, S. (2021). Continuous-action reinforcement learning for portfolio allocation of a life insurance company. In Dong, Y., Kourtellis, N., Hammer, B., and Lozano, J. A., editors, *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, pages 237–252, Cham. Springer International Publishing.

Afanasyeva, E., Jerow, S., Lee, S. J., and Modugno, M. (2020). Sowing the seeds of financial imbalances: The role of macroeconomic performance. Finance and Economics Discussion Series 2020-028, Board of Governors of the Federal Reserve System (U.S.).

Aggarwal, C. (2020). *Linear Algebra and Optimization for Machine Learning*. Springer International Publishing, New York.

Akkaya, C., Wiebe, J., and Mihalcea, R. (2009). Subjectivity word sense disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, page 190–199, USA. Association for Computational Linguistics.

Alexander, C. (2004). *Volume I: Finance Theory, Financial Instruments and Markets*. PRMIA pulications, Wilmington, DE.

Alexander, C. (2008a). *Market Risk Analysis II: Practical Financial Econometrics*. Wiley, Chichester.

Alexander, C. (2008b). *Market Risk Analysis IV: Pricing, Hedging and Trading Financial Instruments*. Wiley, Chichester.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609.

Ang, A. (2014). *Asset Management: A Systematic Approach to Factor Investing*. Oxford, New York, 1st edition.

Ang, A. and Piazzesi, M. (2003). A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. *Journal of Monetary Economics*, 50(4):745–787.

Anson, M. J., Fabozzi, F. J., and James, F. J. (2011). Asset classes, alternative Investments, Investment Companies, and Exchange Traded Funds. In Fabozzi, F. J. and Markowitz, H. M., editors, *The theory and practice of investment management*, pages 45–78. John Wiley & Sons, Hoboken, New Jersey, 1st edition.

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294.

AQR (2022). *AQR: 2022 Capital Market Assumptions for Major Asset Classes*. [available online at https://www.aqr.com/-/media/AQR/Documents/Alternative-Thinking/AQR-Alternative-Thinking-1Q22-Capital-Market-Assumptions.pdf, accessed on 03 Nov 2022].

Araten, M., Jacobs, M., and Varshney, P. (2004). Measuring LGD on commercial loans: an 18-year internal study. *The Journal of the Risk Management Association*, 2:28–35.

Arrow, K. J. (1965). *Aspects of the Theory of Risk-Bearing*. Yrjo Jahnssonin Saatio. Reprinted in: Essays in the Theory of Risk Bearing, Markham Pubishing Company.

Azar, P. D. and Lo, A. W. (2019). The wisdom of Twitter crowds: Predicting stock market reactions to FOMC meetings via Twitter feeds. *The Journal of Portfolio Management*, 42(5):123–134.

Azariadis, C. (2018). *Credit cycles and business cycles*. Review, Federal Reserve Bank of St. Louis.

Back, K. E. (2010). *Asset Pricing and Portfolio Choice Theory*. Oxford University Press, New York.

Back, K. E. (2017). *Asset Pricing and Portfolio Choice Theory, Revised edition*. Oxford University Press, New York.

Baker, H. K. and Filbeck, G. (2013). *Alternative Investments: Instruments, Performance, Benchmarks, and Strategies*. John Wiley & Sons.

Bansal, R. (2007). Long-run risks and financial markets. NBER Working Papers 13196, National Bureau of Economic Research, Inc.

Bansal, R. and Zhou, H. (2002). Term structure of interest rates with regime shifts. *The Journal of Finance*, 57(5):1997–2043.

Barro, R. J. (2006). Rare disasters and asset Markets in the twentieth century. *The Quarterly Journal of Economics*, 121(3):823–866.

Bass, R., Gladstone, S., and Ang, A. (2017). Total portfolio factor, not just asset, allocation. *The Journal of Portfolio Management*, 43(5):38–53.

Beaudry, P. and Portier, F. (2014). News-driven business cycles: Insights and challenges. *Journal of Economic Literature*, 52(4):993–1074.

Beckers, S. (2018). Do social media Trump news? The relative importance of social media and news based sentiment for market timing. *The Journal of Portfolio Management*, 45(2):58–67.

Bellman, R. (1984). *Eye of the Hurricane: An Autobiography*. World Scientific Publishing Company.

Bengfort, B., Bilbro, R., and Ojeda, T. (2018). *Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning*. O'Reilly Media, Inc., Sebastopol, California, 1st edition.

Benrud, E. (2011). Portfolio selection. In Baker, H. K. and Filbeck, G., editors, *Alternative Investments: Instruments, Performance, Benchmarks, and Strategies*, pages 37–52. John Wiley & Sons, Hoboken, New Jersey, 1st edition.

Benzschawel, T. (2012). *Credit Modelling: Facts, Theory and Applications*. Risk Books, London.

Benzschawel, T. (2017). *Credit Modelling: Advanced Topics*. Risk Books, London.

Bernanke, B. and Gertler, M. (1998). *The financial accelorator in a quantitative business cycle framework*. NBER Working Paper No. 6455.

Berndt, A., Douglas, R., Duffie, D., and Ferguson, M. (2018). Corporate credit risk premia. Working Paper 24213, National Bureau of Economic Research.

Black, F. (1972). Capital market equilibrium with restricted borrowing. *The Journal of Business*, 45(3):444–455.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654.

BlackRock (2022). *Black Rock: Capital Market Assumptions.* [available online at https://www.blackrock.com/institutions/en-us/insights/charts/capital-market-assumptions, accessed on 03 December 2022].

Blanchard, O. (2018). On the future of macroeconomic models. *Oxford Review of Economic Policy*, 34(1-2):43–54.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(1):993–1022.

Boiy, E. and Moens, M.-F. (2009). A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval*, 12:526–558.

Bollen, J., Mao, H., and Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Bonfim, D. (2009). Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics. *Journal of Banking & Finance*, 33(2):281–299.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2018). Diagnostic expectations and credit cycles. *The Journal of Finance*, 73(1):199–227.

Borio, C. (2012). The financial cycle and macroeconomics: What have we learnt? *BIS Working Papers No 395., Bank of International Settlements.*

Borio, C., Drehmann, M., and Xia, D. (2019). Predicting recessions: Financial cycle versus term spread. *BIS Working Papers No. 818 , Bank of International Settlements.*

Bradley, M. M. and Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings.* Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Bram, J. and Ludvigson, S. (1998). Does consumer confidence forecast household expenditure? A sentiment index horse race. *Economic Policy Review*, 4(Jun):59–78.

Brown, G. W., Ethridge, F., Johnson, T., and Keck, T. (2020). *Private Portfolio Attribution Analysis*. Social Science Research Network.

Buraschi, A., Porchia, P., and Trojani, F. (2010). Correlation risk and optimal portfolio choice. *Journal of Finance*, 65(1):393–420.

Cambria, E., Das, D., Bandyopadhyay, S., and Feraco, A. (2017). *A Practical Guide to Sentiment Analysis*. Springer Publishing Company, Incorporated, 1st edition.

Campante, F., Sturzenegger, F., and Velasco, A. (2021). *Advanced Macroeconomics*. LSE Press, London.

Campbell, J. Y. and Cochrane, J. H. (1999). By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy*, 107(2):205–251.

Campbell, J. Y., MacKinlay, A. C., and Lo, A. Y. (1997). *The Econometrics of Financial Markets*. Springer, New York.

Carroll, C. D., Fuhrer, J. C., and Wilcox, D. W. (1994). RATS code for does consumer sentiment forecast household spending? If so, why? QM&RBC Codes, Quantitative Macroeconomics & Real Business Cycles.

Chambers, D. R., Anson, M. J. P., Black, K. H., and Hossein, K. (2015). *Alternative Investments: CAIA Level I, 3rd Edition*. John Wiley & Sons.

Chatterjee, S. (2015). *Centre for Central Banking Studies: Modelling credit risk*. Bank of England.

Chekhlov, A., Uryasev, S., and Zabarankin, M. (2004). Portfolio Optimization With Drawdown Constraints. In Pardalos, P. M., Migdalas, A., and Baourakis, G., editors, *Supply Chain And Finance*, World Scientific Book Chapters, chapter 13, pages 209–228. World Scientific Publishing Co. Pte. Ltd.

Chen, K.-C. (2019). *Artificial Intelligence in Wireless Robotics*. River Publishers.

Chen, Z., Lien, D., and Lin, Y. (2021). Sentiment: The bridge between financial markets and macroeconomy. *Journal of Economic Behavior & Organization*, 188:1177–1190.

Cochrane, J. H. (2001). *Asset pricing*. Princeton University Press, Princeton, New Jersey, United States.

Cochrane, J. H. (2005). John H. Cochrane, Asset Pricing (Revised Edition), Princeton University Press, Princeton and Oxford (2005). *Journal of Economic Behavior & Organization*, 60(4):603–608.

Cochrane, J. H. (2017). Macro-Finance. *Review of Finance*, 21(3):945–985.

Connolly, R., Dubofsky, D., and Stivers, C. (2018). Macroeconomic uncertainty and the distant forward-rate slope. *Journal of Empirical Finance*, 48:140–161.

Consoli, S., Pezzoli, L. T., and Tosetti, E. (2020). Using the GDELT Dataset to Analyse the Italian Sovereign Bond Market. In Nicosia, G., Ojha, V., La Malfa, E., Jansen, G., Sciacca, V., Pardalos, P., Giuffrida, G., and Umeton, R., editors, *Machine Learning, Optimization, and Data Science*, pages 190–202. Springer International Publishing.

Cornuejols, G. and Tutuncu, R. (2007). *Optimization Methods in Finance*. Cambridge, New York.

Correa, R., Garud, K., Londono, J. M., and Mislang, N. (2020). Sentiment in central banks' financial stability reports. *Review of Finance*, 25(1):85–120. eprint: https://academic.oup.com/rof/article-pdf/25/1/85/36302154/rfaa014.pdf.

CTI (2022). *Columbia Threadneedle: Solutions Enhanced: Capital Market Assumptions 2022*. [available online at https://www.columbiathreadneedle.co.uk/en/inst/insights/solutions-enhanced-capital-market-assumptions-2022-making-a-transition/, accessed on 03 December 2022].

Currey, H. B. (1944). The method of steepest descent for non-linear minimization problems. *Quarterly of Applied Mathematics*, 2(3):258–261.

Dang, N. C., Moreno-García, M. N., and De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3):483.

Dees, S. (2016). Credit, asset prices and business cycles at the global level. *Economic Modelling*, 54:139–152.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv, eprint (1810.04805).

Dobbs, R., Koller, T., Lund, S., Ramaswamy, S., Harris, J., Krishnan, M., and Kauffman, D. (2016). *Diminishing returns: Why investors may need to lower their expectations*. McKinsey and Company, McKinsey Global Institute [available online at https://www.mckinsey.com, accessed on 25 October 2017].

Dobrescu, M., Badea, L., and Paicu, C. (2012). Business cycle theories and their relevance to the current global crisis. *Procedia - Social and Behavioral Sciences*, 62:239–243.

Dor, A. B., Dynkin, L., Hyman, J., and Phelps, B. D. (2012). *Quantitative Credit Portfolio Management: Practical Innovations for Measuring and Controlling Liquidity, Spread, and Issuer Concentration Risk*. John Wiley & Sons, New York.

Drechsler, I., Savov, A., and Schnabl, P. (2018). A Model of Monetary Policy and Risk Premia. *The Journal of Finance*, 73(1):317–373.

Duan, J.-C. and Zhu, Y. (2020). Credit risk cycle indices: Properties and macroprudential policy. Technical report, National University of Singapore.

Duffie, D. and Singleton, K. (1999). Modeling the term structure of defaultable bonds. *Review of financial studies*, 12:687–720.

Durdu, C. B. and Zhong, M. (2021). Understanding bank and non-bank credit cycles: A structural exploration. *BIS Working Papers No 919, Bank for International Settlements*.

Engelmann, B. (2021). A simple and consistent credit risk model for Basel II/III, IFRS 9 and stress testing when loan data history is short. Technical report, Working paper.

Engstrom, E. and Sharpe, S. A. (2018). *(Don't Fear) the Yield Curve*. Social Science Research Network, SSRN Scholarly Paper.

Epstein, L., Farhi, E., and Strzalecki, T. (2014). How much would you pay to resolve long-run risk? *American Economic Review*, 104(9):2680–97.

Erlwein-Sayer, C. (2018). Macroeconomic News Sentiment: Enhanced Risk Assessment for Sovereign Bonds. *Risks*, 6(4):1–27.

Evgenidis, A. and Malliaris, A. G. (2022). Monetary policy, financial shocks and economic activity. *Review of Quantitative Finance and Accounting*, 59:429–456.

Fabozzi, F. J., Kolm, P. N., Pachamanova, D. A., and Focardi, S. M. (2007). *Robust Portfolio: Optimization and Management*. John Wiley & Sons, Hoboken, New Jersey.

Fabozzi, F. J. and Markowitz, H. M. (2011). *The Theory and Practice of Investment Management*. John Wiley & Sons, Hoboken, New Jersey.

Fabozzi, F. J. and Pachamanova, D. (2010). *Simulation and Optimization in Finance*. John Wiley & Sons, Hoboken, New Jersey.

Fabozzi, F. J. and Pachamanova, D. A. (2016). *Portfolio Construction and Analytics*. John Wiley & Sons, Hoboken, New Jersey, 1st edition.

Falato, A. and Xiao, J. (2020). Credit markets, learning, and the business cycle. In *Credit Markets, Learning, and the Business Cycle*.

Fama, E. F. (1970). Efficient capital markets: Review of theory and empirical work. *The journal of Finance*, 25(2):383–417.

Fellbaum, C. (1998). A semantic network of English: The mother of all WordNets. *Computers and the Humanities*, 32(2):209–220.

Feuerriegel, S. and Gordon, J. (2018). Long-term stock index forecasting based on text mining of regulatory disclosures. *Decision Support Systems*, 112:88–97.

Feuerriegel, S. and Gordon, J. (2019). News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions. *European Journal of Operational Research*, 272(1):162–175.

Filardo, A. J., Lambardi, M., and Raczko, M. (2019). *Measuring financial cycle time: Staff Working Paper No. 776*. Bank of England.

Fischer, M., Köstler, C., and Jakob, K. (2019). Modeling stochastic recovery rates and dependence between default rates and recovery rates within a generalized credit portfolio framework. *Journal of Statistical Theory and Practice*, 10(2):342–356.

Fons, J. S. (1994). Using default rates to model the term structure of credit. *Financial Analysts Journal*, 50(5):25–33.

French, C. W. (2003). The Treynor capital asset pricing model. *The journal of investment management*, 1(2):60–72.

Frye, J. (2000). *Collateral Damage: A Source of Systematic Credit Risk* . Federal Reserve Bank of Chicago, Risk Magazine.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300.

GDELT (2020). *The Empirical Use of GDELT Big Data in Academic Research*. Google GDELT report - GLOBE – The European Union and the Future of Global Governance.

Gilchrist, S. and Zakrajšek, E. (2012). Credit spreads and business cycle fluctuations. *American Economic Review*, 102(4):1692–1720.

Giuzio, M., Gintschel, A., and Paterlini, S. (2018). The components of private debt performance. *The Journal of Alternative Investments*, 20(4):21–35.

Goldberg, Y. and Hirst, G. (2017). *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers.

Gordy, M. (2003). A risk-factor model foundation for ratings-based bank capital rules. *Journal of Financial Intermediation*, 12(3):199–232.

Gourio, F. (2012). Disaster risk and business cycles. *American Economic Review*, 102(6):2734–2766.

Gourio, F. (2013). Credit risk and disaster risk. *American Economic Journal: Macroeconomics*, 5(3):1–34.

Greenwood, R., Hanson, S. G., and Jin, L. J. (2016). A model of credit market sentiment. *Harvard Business School Working Paper*, 17(015):1–44.

GSAM (2022). *Goldman Sachs Asset Management: Market Insights*. [available online at https://www.gsam.com/content/gsam/uk/en/advisers/market-insights.html, accessed on 03 December 2022].

Gujarati, D. (1999). *Essentials of Econometrics*. McGraw-Hill economics series. Irwin/McGraw-Hill.

Gulen, H., Ion, M., and Rossi, S. (2019). Credit cycles, expectations, and corporate investment. CEPR Discussion Papers 13679, C.E.P.R. Discussion Papers.

Gupta, R., Cepni, O., and Ji, Q. (2021). Sentiment Regimes and Reaction of Stock Markets to Conventional and Unconventional Monetary Policies: Evidence from OECD Countries. Working Papers 202126, University of Pretoria, Department of Economics.

Gupton, G. M. (1997). *Credit Metrics - An publically disclosed internal JP Morgan technical document*. JP Morgan.

Gârleanu, N. and Panageas, S. (2015). Young, old, conservative, and bold: The implications of heterogeneity and finite lives for asset pricing. *Journal of Political Economy*, 123(3):670–685.

Hall, R. E. (1988). Intertemporal substitution in consumption. *Journal of Political Economy*, 96(2):339–357.

Han, C. (2019). Modeling severity risk under PD–LGD correlation. *The European Journal of Finance*, 23(15):1572–1588.

Hansen, L. and Jagannathan, R. (1991). Implications of Security Market Data for Models of Dynamic Economies. *Journal of Political Economy*, pages 225–262.

Hens, T. and Rieger, M. O. (2010). *Financial Economics: A concise introduction to classical and behavioural finance*. Springer, New York.

Herwartz, H., Ochsner, C., and Rohloff, H. (2020). The credit composition of global liquidity. MAGKS Joint Discussion Paper Series in Economics 15-2021, Philipps-University Marburg, School of Business and Economics, Marburg.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.

Houweling, P. and van Zundert, J. (2017). Factor investing in the coporate bond market. *Financial Analyst Journal*, pages 100–115.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, page 168–177.

Hutto, C. and Gilbert, E. (2014). *VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text.* Ann Arbor, MI.

Invesco (2022). *Invesco: 2022 Long-Term Capital Market Assumptions.* [available online at https://www.invesco.com/apac/en/institutional/insights/multi-asset/long-term-capital-market-assumptions.html, accessed on 18 October 2022].

Investec (2022). *Investec Wealth and Investment: Global Wealth and Investment View.* [available online at https://www.investec.com/content/dam/south-africa/myself/invest/global-investment-view /global-view-pdfs/Investec-Global-Wealth-and-Investment-View-Q1-2022.pdf, accessed on 03 December 2022].

Iworiso, J. and Vrontos, S. (2020). On the directional predictability of equity premium using machine learning techniques. *Journal of Forecasting*, 39(3):449–469.

Jarrow, R., Lando, D., and Turnbull, S. M. (2008). A markov model for the term structure of credit risk spreads. In *Financial Derivatives Pricing Selected Works of Robert Jarrow*, chapter 18, pages 411–453. World Scientific Publishing Co. Pte. Ltd.

JPM (2022). *J.P. Morgan Asset Management: Portfolio insights, 2023 Long-Term Capital Market Assumptions.* [available online at https://am.jpmorgan.com/content/dam/jpm-am-aem/global/en/insights/portfolio-insights/ltcma/ltcma-full-report.pdf, accessed on 03 December 2022].

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., and Kapadia, S. (2020). Making text count: economic forecasting using newspaper text. *Bank of England: Staff Working Paper No. 865.*

Karagozoglu, A. K. and Fabozzi, F. J. (2017). Volatility wisdom of social media crowds. *The Journal of Portfolio Management*, 43(2):136–151.

Katona, G. (1947). Contribution of psychological data to economic analysis. *Journal of the American Statistical Association*, 42(239):449–459.

Keijsers, B., Diris, B., and Kole, E. (2018). Cyclicality in losses on bank loans. *Journal of Applied Econometrics*, 33(4):533–552.

Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money*. Macmillan. 14th edition, 1973.

Kiranyaz, S. (2013). *Multidimensional particle swarm optimization for machine learning and pattern recognition*. Springer-Verlag, Berlin, Heidelberg.

Kiyotaki, N. and Moore, J. (1997). Credit cycles. *The Journal of Political Economy*, 105(2):211–248.

Kiyotaki, N. and Moore, J. (2019). Liquidity, business cycles, and monetary policy. *Journal of Political Economy*, 127(6):2926–2966.

Kokoszka, P. (2023). *FCA set to kick off review of private market valuations*. IPE International Publishers Limited [available online from https://www.ipe.com/fca-set-to-kick-off-review-of-private-market-valuations/10069217.article, accessed on 19th November 2023].

Kucera, A. (2017). *Interest rates modeling and forecasting: Do macroeconomic factors matter?* IES Working Paper.

Kurdi, M. Z. (2016). *Natural language processing and computational linguistics: Speech, morphology and syntax*, volume 1. LTSE and John Wiley & Sons.

Kwon, R. H. (2013). *Introduction to Linear Optimization and Extensions with MATLAB®*. CRC Press.

Lane, H., Hapke, H., and Howard, C. (2019). *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Manning Publications.

Lang, M., Ruscheinsky, J., and Hausler, J. (2017). A contemporary sentiment analysis approach: Algorithm-based analysis of news items within the direct real estate market in the US. *European Real Estate Society (ERES)*.

Larsen, V. H. and Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, 210(1):203–218.

Ledoit, O. and Wolf, M. (2002). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621.

Li, D. (2000). On default correlation: A copula approach. *Journal of Fixed Income*, 9:119–149.

Liew, J. and Budavari, T. (2017). The "sixth" factor: A social media factor derived directly from tweet sentiments. *The Journal of Portfolio Management*, 43(3):102–111.

Lin, H., Tao, X., Wang, J., and Wu, C. (2020). Credit spreads, business conditions, and expected corporate bond returns. *Journal of Risk and Financial Management*, 13(2):20.

Lintner, J. (1961). The valuation of risk assets and the selection of risky investments in stock portfolio and capital budgets. *Review of Economics and Statistics*, 47(1):13–37.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Lopez, M. M. and Kalita, J. (2017). Deep Learning applied to NLP.

López-Salido, D., Stein, J. C., and Zakrajšek, E. (2017). Credit-market sentiment and the business cycle. *The Quarterly Journal of Economics*, 132(3):1373–1426.

Manela, A. and Moreira, A. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics*, 123(1):137 – 162.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1):77–91.

Martellini, L. and Milhau, V. (2017). *Factor Investing: Welfare improving or marketing fad?* [available online from https://thehedgefundjournal.com/factor-investing/, accessed on 4th February 2017].

Mehra, R. (2007). The Equity Premium Puzzle: A Review. *Foundations and Trends in Finance*, 2(1):1–81.

Mehra, R. and Prescott, E. C. (1985). The equity premium: A puzzle. *Journal of Monetary Economics*, 15(2):145–161.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1):125–144.

Minsky, H. P. (1978). The financial instability hypothesis: An interpretation of Keynes and an alternative to “standard” theory. *Nebraska Journal of Economics and Business*, 16(1):5–16.

Mitkov, R. (2005). *The Oxford Handbook of Computational Linguistics (Oxford Handbooks)*. Oxford University Press, Inc., USA.

Modigliani, F. and Miller, M. H. (1958). The cost of capital, corporation finance and the theory of investment. *The American Economic Review*, 48(3):261–297.

Moody's (2007). Special comment: Back-testing moody's LGD methodology. *Moody's Analytics, from Moody's KMV*, page 10.

Moody's (2010). Implications of PD-LGD correlation in a portfolio setting. *Moody's Analytics: Modeling Methodology*.

Moody's (2018). Features of a lifetime PD model: Evidence from public, private, and rated firms. *Moody's Analytics: Modeling Methodology*.

MorganStanley (2022). *Morgan Stanley Wealth Management, Global Investment Commitee: Annual Update of GIC Capital Market Assumptions*. [available online at https://www.morganstanley.com/assets/pdfs/2d9493c3-822f-4f18-8c28-ba3ad25e8473.pdf, accessed on 08 September 2022].

Mudinas, A., Zhang, D., and Levene, M. (2019). Market trend prediction using sentiment analysis: Lessons learned and paths forward. *In: Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM'18)*.

Müller, K. and Verner, E. (2021). Credit allocation and macroeconomic fluctuations. *Available at SSRN: https://ssrn.com/abstract=3781981*.

Munk, C. (2011). *Fixed Income Modelling*. Oxford University Press, Oxford.

Munk, C. (2013). *Financial Asset Price Theory*. Oxford University Press, Oxford.

Nakamura, E., Steinsson, J., Barro, R., and Ursúa, J. (2013). Crises and recoveries in an empirical model of consumption disasters. *American Economic Journal: Macroeconomics*, 5(3):35–74.

Nazemi, A. and Fabozzi, F. J. (2018). Macroeconomic variable selection for creditor recovery rates. *Journal of Banking & Finance*, 89:14–25.

Nimisha (2010). *Alternative Credit*. Towers Watson, [available online at https://www.towerswatson.com, accessed on 14 June 2017], 3rd edition.

NinetyOne (2022). *Ninety One: Capital Market Assumptions*. [available online at https://ninetyone.com/en/insights/capital-market-assumptions-podcast-march-2022, accessed on 02 December 2022].

NorthernTrust (2022). *Northern Trust: Capital Market Assumptions: Five year outlook, 2023 addition*. [available online at https://www.capitalmarketassumptions.com/, accessed on 08 September 2022].

Nowzohour, L. and Stracca, L. (2017). More than a feeling: Confidence, uncertainty and macroeconomic fluctuations. Working Paper Series 2100, European Central Bank.

Nyman, R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P., and Smith, R. (2018). News and narratives in financial systems: Exploiting big data for systemic risk assessment. Bank of England working papers 704, Bank of England.

OECD (2021). *OECD Economic Indicators*. OECD.org, [available online at https://data.oecd.org/economy.htm - select 'Leading indicators', accessed on 7 August 2021].

OFCOM (2020). *News Consumption in the UK: 2020*. Jigsaw Research.

Ozdemir, B. (2017). *Adapting to Basel III and IV: Re-engineering capital, business mix and performance management practices post-crisis*. Risk Books, London.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). *Thumbs up? Sentiment classification using machine learning techniques*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

Patel, A. A. (2019). *Hands-On Unsupervised Learning Using Python*. O'Reilly Media, Inc., Sebastopol, California, 1st edition.

Pedro Bordalo, Nicola Gennaioli, A. S. and Terry, S. J. (2021). *Real Credit Cycles*. [available online at https://www.nber.org/papers/w28416, accessed on 13 November 2021].

Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., and Booth, R. (2007). The development and psychometric properties of LIWC2007. *LIWC2007 Manual*, pages 1–22.

Pennebaker, J., Francis, L., and Booth, R. (2001). Linguistic inquiry and word count: LIWC2001. *LIWC Operators Manual 2001*.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Pesaran, H. M., Schuermann, T., Treutler, B.-J., and Weiner, S. M. (2003). Macroeconomic dynamics and credit risk: A global perspective. Technical report, DAE Working Paper No. 0330.

PGIM (2022). *PGIM: 2022 Q3 Capital Market Assumptions*. [available online at https://www.pgim.com/investments/article/2022-q3-capital-market-assumptions, accessed on 15 October 2022].

Pozzi, F. A., Fersini, E., Messina, E., and Liu, B. (2017). *Sentiment Analysis in Social Networks*. Cambridge University Press.

Prabheesh, K. P., Anglingkusumo, R., and Juhro, S. M. (2020). The dynamics of global financial cycle and domestic economic cycles: Evidence from India and Indonesia. *Economic Modelling*.

Prado, M. L. D. (2018). *Advances in Financial Machine Learning*. John Wiley & Sons, Hoboken, New Jersey, 1st edition.

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica*, 32(1/2):122–136.

Preqin (2017). *Preqin Investor Outlook: Alternative Assets*. Preqin Ltd, [available online at https://www.preqin.com, accessed on 17 September 2017], h1 2017 edition.

Preqin (2018). *Preqin Investor Outlook: Alternative Assets*. Preqin Ltd, [available online at https://www.preqin.com, accessed on 11 July 2019], h2 2018 edition.

Preqin (2019). *Preqin Investor Outlook: Alternative Assets*. Preqin Ltd, [available online at https://www.preqin.com, accessed on 11 July 2019], private debt sample pages edition.

Preqin (2022). Preqin investor outlook: Alternative assets. *2022 Preqin Global Alternatives Reports*.

Rao, D. and McMahan, B. (2019). *Natural Language Processing with PyTorch*. O'Reilly Media.

Rao, S. S. (2009). *Introduction to Optimization*. John Wiley & Sons, Ltd.

Robeco (2022). *Robeco: Outlook 2022-2026 - 5-year Expected Returns*. [available online at https://www.robeco.com/latam/en/themes/expected-returns/2022-2026/, accessed on 03 December 2022].

Rockafellar, R. and Uraysev, S. (2002). Optimization of conditional value at risk. *Journal of Banking & Finance*, 26:1443–1471.

Rojas, R. (1996). *Neural Networks: A Systematic Introduction*. Springer-Verlag, Berlin, Heidelberg.

Roll, R. and Ross, S. A. (1980). An empirical investigation of the arbitrage pricing theory. *The Journal of Finance*, 35(5):1073–1103.

Roll, R. and Ross, S. A. (1984). The arbitrage pricing theory approach to strategic portfolio planning. *Financial Analysts Journal*, May-June:14–26.

Roncalli, T. (2017). Alternative risk premia: What do we know? *Factor Investing*, pages 227–264.

Rosenblatt, F. (1957). The perceptron: A perceiving and recognizing automaton. *Report -1, Cornell Aeronautical Laboratory*, 01:85–60.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):343–362.

Ruder, S., Vulić, I., and Søgaard, A. (2017). A survey of cross-lingual word embedding models. *arXiv:1706.04902*.

Sahin, S. (2021). The effect of consumer confidence on aggregate household consumption and leverage. *Journal of Management and Economics Research*, 19:62 – 87.

Schroders (2022). *Schroders Asset Management: 10-year return forecasts (2022–31)*. [available online at https://www.schroders.com - 10 year return forecasts] .

Schularick, M., Òscar Jordà, and Taylor, A. M. (2019). The Total Risk Premium Puzzle? Working Paper Series 2019-10, Federal Reserve Bank of San Francisco.

Schweizer, B. and Wolff, E. F. (1981). On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9(4):879 – 885.

Shang, K. (2015). *Deep learning for liability driven investment*. Society of Actuaries [available online from https://www.soa.org/, accessed on 17 November 2022].

Shapiro, A. H. and Shapiro, D. W. (2019). Taking the fed at its Word: A new approach to estimating central bank objectives using text analysis. *Federal Reserve Bank of San Francisco Working Paper*, Working Paper 2019-02(01):1–73.

Shapiro, A. H., Sudhof, M., and Wilson, D. (2020). Measuring news sentiment. *Federal Reserve Bank of San Francisco Working Paper*, Working Paper 2017-1(01):1–49.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442.

Sheikh, A. (2018). *Firm-wide Stress Testing and Economic Capital*. Risk Books, London.

Silver, D. (2015). *Lectures on Reinforcement Learning*. [available online at https://www.davidsilver.uk/teaching/, accessed on 03 September 2022].

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris*, 8:229–231.

Smith, N. and Balint, A. (2019). *Financing the economy: The future of private credit*. AIMA, Alternative Credit Council. available online at [https://acc.aima.org/resources/research/financing-the-economy.html], accessed on 19 February 2020.

Society of Actuaries (2016). *Society of actuaries (SOA) – Economic Scenario Generators – A Practical Guide*. Society of Actuaries [available online from www.soa.org/globalassets/assets/Files/Research/Projects/research-2016-economic-scenario-generators.pdf, accessed on 22 February 2022].

Souleles, N. S. (2004). Expectations, heterogeneous forecast errors, and consumption: Micro evidence from the Michigan consumer sentiment surveys. *Journal of Money, Credit and Banking*, 36(1):39–72.

Sra, S., Nowozin, S., and Wright, S. (2011). *Optimization for Machine Learning*. The MIT Press.

Stiglitz, J. E. (2017). *Macro-economic management in an electronic credit financial system, National Bureau of Economic Research, Working Paper no. 23032*. [available online at http://www.nber.org/papers/w17182, accessed on 08 October 2022.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). *The general inquirer: A computer approach to content analysis*. M.I.T. press.

Sul, H., Dennis, A., and Yuan, L. (2016). Trading on Twitter: Using social media sentiment to predict stock returns: Trading on Twitter. *Decision Sciences*.

Tenreyro, S. (2018). *Models in macroeconomics. Speech given by Silvana Tenreyro, External MPC Member, Bank of England*. Bank of England.

Tim Loughran and Bill McDonald (2014). Measuring readability in financial disclosures. *Journal of Finance*, 69(4):1643–1671.

Towers Watson (2018). *Finding value in private debt - UK paper*. Towers Watson, [available online at https://www.towerswatson.com, accessed on 2 January 2020].

TPriceRowe (2022). *T Price Rowe: Capital Market Assumptions, 2022*. [available online at https://www.troweprice.com/content/dam/ide/articles/pdfs/2022/q1/capital-market-assumptions.pdf, accessed on 03 December 2022].

Tran-The, T. (2020). *Modeling institutional credit risk with financial news*. arXiv.org, 2004.08204.

Trueck, S. and Rachev, S. T. (2009). *Rating Based Modeling of Credit Risk: Theory and Application of Migration Matrices*. Elsevier, London, 1st edition.

Uhlenbeck, G. E. and Ornstein, L. (1930). On the theory of Brownian motion. *Physical Review*, 36:823–841.

Varian, H. (1999). Estimating risk aversion from Arrow-Debreu portfolio choice. *Econometrica*, 56(4):973–979.

Vasicek, O. (1987). *Probability of loss on a loan portfolio*. Risk.net.

Whyte, A. (2018). *The Investors Dominating Private Debt*. Insitutuional investor, [available online at https://www.institutionalinvestor.com/article/b180kjyqfmr30y/the-investors-dominating-private-debt, accessed on 2 January].

Wilson, T. C. (1997). Portfolio Credit Risk (I). *RISK magazine*.

Wilson, T. C. (1998). Portfolio Credit Risk. *FRBNY Economic Policy Review*, 04(03).

Yu, T. and Nwet, K. T. (2020). Comparing SVM and KNN Algorithms for Myanmar News Sentiment Analysis System. In *Proceedings of 2020 the 6th International Conference on Computing and Data Engineering*, ICCDE 2020, page 65–69, New York, NY, USA. Association for Computing Machinery.

Zarnowitz, V. (1992). *Business Cycles: Theory, History, Indicators, and Forecasting*. University of Chicago Press.

Çelik, S., G. D. and Isaksson, M. (2020). *Asset pricing and portfolio choice theory*. OECD Capital Market Series, Paris. available online at [www.oecd.org/corporate/Corporate-Bond-Market-Trends-Emerging-Risks-and-Monetary-Policy.htm], accessed on 19 February 2020.

Part VIII

Appendix

26 Natural language processing

26.1 Social media data sourcing

Twitter data is the source of information from social media sources. The data is accessed via the Twitter API version 2 using Python and the Tweepy package. Tweepy is a pre-customised package to draw data using Twitter API in a relatively quick and easy fashion, which suits a use case of NLP modelling well without needing to make use of the full capability of Twitter's API v2. Two mechanisms were used for extracting or filtering the tweets from Twitter. Firstly using search terms, the following string based filters have been set up to isolate those topics which will be filtered from Twitter, using the string search endpoint functionality as part of the Twitter API 2.

26.2 Twitter table structure

Field name	Description and example	Format
Twitter Table: Followers		
id	Unique Twitter user key	Integer
create_at	Creation time of Tweet example: : ‘Wed Oct 10 20:19:24 +0000 2018’	datetime
screen_name	Name of follower as it appears on the Twitter screen	String
name	Name of follower	String
description	Description that the user has written describing themselves	String
followers_count	Total number of followers this tweet profile carries	Integer
friend_count	Total number of friends this profile has	Integer
favourites_count	Total number of tweets that the user profile has liked	Integer
list_of_followers	Twitter id list of all the followers of this particular follower	Integer
list_of_friends	Twitter id list of all the friends of this particular follower	Integer
coordinates	coordinates of the user for users exposing location whilst using the Twitter application [-77.119759,38.791645]	geoJSON
Twitter Table: Twitter Tweets		
id_str	String representation of the unique identifier for this Tweet.	String
id_user	Unique Twitter user key [1050118621198921728]	Integer
tweet_text	Text captured in the tweet by the user profile	String
date	Datetimestamp marking the time of the data draw	Datetime
retweet_count	Count of the number of retweets	Integer
retweet_status	Indicator for retweets, where users can broadcast tweets from other users	true/false
lang	Language of the user - default set by profile user (lang = ‘en’)	String
truncated	Indicator for tweets longer than 140 characters, which are truncated and end with an ellipsis, ...	Boolean

Table 27: Social data description, keys, field names and data sources.

26.3 Twitter search terms

Search terms		
ACWI	FOMC	Non-Farm Payrolls
AllShare	FTSE100	production manufacturing
FTSE350	GDP	purchase managers index
10 year bond	gross domestic product	real business cycle
short rate expectations	housing starts	retail sales
balance of trade	interest rates	S&P500
business cycle	market sentiment	stock market performance
Consumer Price Index	market volatility	unemployment rate

Table 28: Twitter search terms used to filter data from Twitter.

26.4 Twitter handles

Streaming tweets, using Twitter's free API. Tweets are filtered from the streaming data, based on search terms and 50 prominent economics and market profiles and news profiles. I captured streaming data from 01 September 2020 to 30 August 2021, using Python and the Twitters API endpoint accessed using package. In the tables that follow, I detailed the tables structure, then search terms for this study and lastly the twitter handles I followed for data extraction.

Name	Screen Name	Location	Description	Follower Count
Adam Posen	AdamPosen	Washington, DC	President, Peterson Institute for International Economics; proclaims on monetary, fiscal and trade policy, advises investors.	43,060
Al Jazeera	AJEnglish	Doha, Qatar	English, Hear the human story and join the discussion.	6,539,036
Amine Ouazad	amine_ouazad	Canada	Econ prof @HEC_Montreal. Researcher in urban, real estate, climate, finance.	2,350
Annie Lowrey	AnnieLowrey	San Francisco, CA	Staff writer for @TheAtlantic, author of 'Give People Money', new Californian. Say hi on annie@theatlantic.com.	97,285
Antonio Fatas	AntonioFatas	Singapore	Professor of Economics at INSEAD, an international business school with campuses in Singapore, France and Abu Dhabi.	5,942
Atif Mian	AtifRMian	Princeton	I do finance and macro ... with data, at Princeton.	
Austan Goolsbee	Austan_Goolsbee	Chicago	Econ prof at U.Chicago's Booth School of Business and former Chairman of the Council of Economic Advisers.	101,333
Barry Eichengreen	B_Eichengreen	Berkeley, California	George C. Pardee and Helen N. Pardee Professor of Economics and Political Science, University of California, Berkeley, NBER Research Associate, CEPR Res Fellow	35,945
Barry Ritholtz	ritholtz	NYC	Welcome to the Dopamine factory! Chair/CIO of RWM https://t.co/n78eQEY6QZ Masters-in-Business podcast/radio host Director of Twitter Cognitive Dissonance	171,920
BBC News	BBCBreaking	London, UK	Breaking news alerts and updates from the BBC.	46,885,235
Ben Casselman	bencasselman		Econ/business/data reporter for @nytimes. Formerly: @fivethirtyeight, @WSJ. Adjunct @newmarkjschool	63,794
Ben White	morningmoneyben	NYC	POLITICO Chief Economic Correspondent and Morning Money columnist.	116,392
Ed Dolan	dolanecon	Northwest Michigan	Sr. Fellow @NiskanenCenter. Healthcare, environment, poverty, lots of other stuff. Author of intro econ text by BVT	4,772
Edward Harrison	edwardnh		Credit Writedowns / Real Vision	33,893
Erik Brynjolfsson	erikbryn	Stanford	Director @DigEconLab; Jerry Yang and Akiko Yamazaki Professor and Ralph Landau Senior Fellow @Stanford; Co-author @2MABook and @MPCBook. https://t.co/8Ounj6hmzw	188,710
Evan Soltas	esoltas	Cambridge, MA	PhD student @MITEcon, interested in public finance and labor economics.	9,822

Table 29: Twitter handles table (1/8).

Name	Screen Name	Location	Description	Follower Count
Betsey Stevenson	BetseyStevenson	Ann Arbor	Former Member of the President's Council of Economic Advisers and Chief Economist at Labor. Current academic economist at Michigan. Always an economist at home.	47,864
Bloomberg	business	New York and the World	The first word in business news.	6,828,441
Boris Johnson	BorisJohnson	United Kingdom	Prime Minister of the United Kingdom and @Conservatives leader. Member of Parliament for Uxbridge and South Ruislip.	3,179,577
Brian Romanchuk	RomanchukBrian	Montréal, Québec	Writes on bond market economics	2,862
Cameron Crise	5thrule	Westport and Gotham	Macro strategist at Bloomberg writing for the Markets Live blog. Former macro HF PM. Always on the hunt for pink flamingos.	6,981
Catherine L Mann	CLMannEcon	New York City	Global Chief Economist, Citi. International, analytical, policy (Fed, OECD), research (PIIE, Brandeis), private sector (Citi). My views. Loves Cocker Spaniels	7,056
Catherine Rampell	crampell	New York City	Syndicated op-ed columnist @washingtonpost, commentator @cnn, special correspondent @newshour. Previously econ/theater NYT. Econ, immigration, tax, politics etc	184,625
Charlie Bilello	charliebilello		Founder and CEO of Compound Capital Advisors. Sharing ideas that compound over time. https://t.co/1zNIPqZdEH	199,826
Chris Dillow	CJFDillow	Oakham	One of Rutland's most prominent Marxist economists.	12,530
Christophe Barraud	C_Barraud	Nice/Paris	FR Chief Economist, Strategist — PhD — Bloomberg Top Forecaster of the US [2012-2019], EU [2015-2019] and CN Economy [2017-2019] — RT does not equate to endorsement	74,554
Cullen Roche	cullenroche	Encinitas, CA	Former mail delivery boy turned multi-asset investment manager, author, Ironman & chicken farmer. Probably should have stayed with mail delivery....	53,571
Dani Rodrik	rodrikdani	Cambridge, MA	Economist at Harvard Kennedy School	161,575
Daniel Lacalle	dlacalle_IA	London, UK	PhD #Economist, #Author. Chief Economist @TressisSV. #Professor @IEbusiness @IEB_Spain @UNED. #Advisory @frdelpino. Married, 3 kids https://t.co/Cfm6rqzR1F	90,463
Darrick Hamilton	DarrickHamilton	Brooklyn, NY	Henry Cohen Professor of Economics and Urban Policy and University Professor @TheNewSchool	37,244

Table 30: Twitter handles table (2/8).

Name	Screen Name	Location	Description	Follower Count
Dave Giles	DEAGiles	L'Amable, ON Canada	Tame econometrician; professor retired from @UVic; mild-mannered ex-blogger; gone fishin'.	3,126
David Andolfatto	dandolfa	St Louis, MO	Construction worker turned academic turned central banker. Opinions expressed here are my own, not St. Louis Fed nor U.S. Fed Reserve System.	17,934
David Boaz	David_Boaz	Washington DC	Executive Vice President of The Cato Institute; Author, The Libertarian Mind; opinions are my own.	23,538
Dean Baker	DeanBaker13	Utah	Senior economist at the Center for Economic and Policy Research (@ceprdc). Blog– Beat the Press (@beat_the_press)	59,508
Diane Coyle	DianeCoyle1859	London, Cambridge	The Enlightened Economist. Bennett Professor of Public Policy, University of Cambridge.	26,773
Donald J. Trump	realDonaldTrump	Washington, DC	45th President of the United States of America	88,649,448
Dr. Jennifer Doleac	jenniferdoleac	Austin & College Station	Economics professor @TAMU. Director @JusticeTechLab. Host of the @ProbCausation podcast. I study crime & discrimination.	34,845
Dr. Shelly Lundberg	ShellyJLundberg		Economist; Broom Professor of Demography @ucsantabarbara; Mostly about women in economics	14,804
Five Minute Macro	5_min_macro		Macro Trader; tweeting on rates, FX, equities, commods, life's rich pageant...in roughly that order. Not looking to set the world to rights in 140 characters	18,856
Forbes	Forbes	New York, NY	Official Twitter account of Forbes, the world's leading voice for entrepreneurial success and free enterprise.	16,457,772
Frances 'Cassandra' Coppolas	Frances_Coppola	United Kingdom	Finance & economics writer and commentator. Author of 'The Case For People's Quantitative Easing' (Polity Books). Sings a bit too. Always right, never believed.	57,096
Gita Gopinath	GitaGopinath	Washington, D.C.	Chief Economist of the International Monetary Fund. On leave of public service from Harvard University's Economics Department. Views are my own.	80,145
Heidi Hartmann	HeidiatIWPR	Washington, DC	Founding President @IWPResearch: policy issues, gender analysis & women in economics. Distinguished Economist in Residence, American Univ. First Generation.	1,989

Table 31: Twitter handles table (3/8).

Name	Screen Name	Location	Description	Follower Count
Heidi Shierholz	hshierholz	Washington, DC	Director of Policy, @EconomicPolicy. Former Chief Economist at the US Dept of Labor. Bike commuter, backyard beekeeper. Tweets my own. She/her.	13,492
James Picerno	jpicerno		Editor (US Business Cycle Risk Report), author (Quantitative Investment Portfolio Analytics In R), blogger (https://t.co/prQ9D8e8ID)	1,671
Jared Bernstein	econjared	Washington, DC	CEA Member for President Biden. Former Chief Economist & Adviser to VP Biden during the Obama-Biden administration.	72,989
Jeff Deist	jeffdeist		Ex uno plures.	20,262
Jesse Colombo	TheBubbleBubble	Dallas, TX	Economic analyst, Zero Hedge & Forbes contributor. Warns about dangerous bubbles. Recognized by the London Times for predicting the Global GFC.	168,975
Jessie Handbury	jessiehandbury		Economist and Assistant Professor of Real Estate at Wharton	1,076
Jim Cramer	jimcramer	New York City	Founder of @TheStreet & I run charitable trust portfolio https://t.co/0UYF2L7v5y. I also host @MadMoneyOnCNBC & blog daily on https://t.co/sGTJX6GWO2. Booyah!	1,451,582
Jodi Beggs	jodiecongirl	Cambridge, MA	Behavioral Economist, data scientist, Economists Do It With Models, We the Economy, Homer-Economicus.	30,444
Joe Biden	JoeBiden	Wilmington, DE	President-elect, husband to @DrBiden, proud father & grandfather. Ready to build back better for all Americans.	21,110,233
Joe Weisenthal	TheStalwart	New York City	Co-host of the Odd Lots podcast and 'What'd You Miss?' on Bloomberg TV. Editor. Chess and Bakersfield fan.	233,371
John Burn-Murdoch	jburnmurdoch	Doncaster London	Stories, stats & scatterplots for @FinancialTimes — Daily updates of the coronavirus trajectory tracker	314,890
John Van Reenen	johnvanreenen	Cambridge, MA	Professor in MIT Economics & Sloan. OBE. Jansson Award winner. Fellow of British Academy & Econometric Soc #FBPE	27,197
Jonathan Portes	jdportes	London	Professor of Economics, King's College London; Senior Fellow, UK in a Changing Europe. Personal views only.	71,123
Mark Thoma	MarkThoma	Eugene, Oregon		44,839

Table 32: Twitter handles table (4/8).

Name	Screen Name	Location	Description	Follower Count
Mariana Mazzucato	MazzucatoM	London	UCL Professor, Director of Institute for Innovation & Public Purpose. Author of Entrepreneurial State; Value of Everything. 4 kids keep me laughing. Bread maker	145,694
Marie Mora	Marietmora		Provost/Exec VC for Academic Affairs @UMSL; Prof of Econ. @PAESMEM recipient; @FocusEconomics Top 75 Econ Influencer; Econ_Rockstar. RT isn't an endorsement.	4,237
Mark Thornton	DrMarkThornton		Austrian School Economist at the Ludwig von Mises Institute, Skyscrapers and Business Cycles, not afraid of deflation	15,646
Matthew E. Kahn	mattkahn1966		Bloomberg Distinguished Professor of Economics and Business at Johns Hopkins University. @jhu_cities , https://t.co/qisqzsF9FO	5,737
Michael Batnick	michaelbatnick		Long-distance reader	124,400
Mike 'Mish' Shedlock	MishGEA	Crystal Lake, IL		21,706
Miles Kimball	mileskimball	Superior, CO	Eaton Professor of Economics at the University of Colorado, Bloomberg columnist, and independent blogger on economics, politics, religion & fighting obesity	21,201
Mises Institute	mises	Auburn, Alabama	Promoting Austrian economics, freedom, and peace in the tradition of Ludwig von Mises through research, publishing, and education. IG: misesinstitute	128,298
Mises Media	mises_media	Auburn, Alabama	Featured Audio & Video from the Mises Institute—the research and educational center of classical liberalism, libertarian political theory, & Austrian economics.	13,044
Mohamed A. El-Erian	elerianm	USA	President, Queens' College, Cambridge University. Chief Economic Adviser, Allianz. Chair, Gramercy Funds Mng. Wharton Professor. Lauder Institute Senior Fellow	355,219
David Wessel	davidmwessel	Washington, D.C.	Director, Hutchins Center on Fiscal & Monetary Policy, Brookings. Contributing correspondent, Wall Street Journal. Views in tweets are strictly my own.	86,284
Nate Silver	NateSilver538	New York	Editor-in-Chief, @FiveThirtyEight. Author, The Signal and the Noise (https://t.co/EYTxvN6BLY). Sports/politics/food geek.	3,665,986
New York Fed	NewYorkFed	New York City, USA	Serving the Second District and the Nation. Tweets from President John Williams are signed -JCW.	169,906

Table 33: Twitter handles table (5/8).

Name	Screen Name	Location	Description	Follower Count
Noah Smith	Noahpinion	San Francisco, CA	Bloomberg Opinion writer. Founder, https://t.co/j9CinPoHlo. Writes about economics, tweets about rabbits. Also blogs at https://t.co/OBEFRT6Hei	187,073
Nouriel Roubini	Nouriel	New York	Stern School NYU Prof; Roubini Macro Associates CEO; @RosaRoubini CoFounder; Crisis Economics author; Time100, FT, FP, Forbes, Coindesk Most Influential Awards	512,360
Owen Zidar	omzidar	Princeton, NJ	Professor of Economics and Public Affairs at Princeton.	8,498
Paul Krugman	paulkrugman	New York City	Nobel laureate. Op-Ed columnist, @nytopinion. Author, 'The Return of Depression Economics', 'The Great Unraveling', 'Arguing With Zombies', + more.	4,646,837
Faisal Islam	faisalislam	London(head) Manchester(heart)	Economics Editor, BBC. Fin crisis book @theDefaultLine .. Brexitologist. Host award-winning 2016 EUref TV interviews & 2017 live GE debate. United ST.	365,872
Pedro Nicolaci da Costa	pdacosta	Washington, DC	Federal Reserve & economy correspondent at Market News International @MNINews Previously: @Reuters, @WSJ, @PIIE. I see Fed people. Opinions my own. Wear a mask.	139,631
Peter Brandt	PeterLBrandt	CO, MN and AZ	Futures/fx career trader since 1975. Author and publisher of the Factor Report. I Tweet charts w/o comments & about stuff I've learned the hard way.	386,559
Prof. Steve Hanke	steve_hanke	Baltimore & Paris	Economist @JohnsHopkins — Sr Fellow & Director #TIedCurrencies Project @CatoInstitute — @NRO — FX & Commodity Trader — Reagan White House — Views are my own	313,034
Quant Insight	Quant_Insight	London, UK	Analyses data using science and technology to understand markets. Not investment advice. Look Up.	513
Rachel Glennerster	rglenner	London, England	Development economist, researcher, FCDO Chief Economist. Tweet on evidence, econ policy, education, life. Tweets my personal views not FCDO policy.	22,616
Joseph E. Stiglitz	JosephEStiglitz	New York	The official account of Joseph E. Stiglitz, Nobel laureate economist based @Columbia University. President @policy_dialogue.	337,910
Justin Wolfers	JustinWolfers	Ann Arbor, MI	Professor @UMichEcon & @FordSchool — @NYTimes contributor — Senior Fellow @BrookingsInst & @PIIE — Intro Econ textbook author — Think Like an Economist podcast	206,827
Larry Mishel	LarryMishel	DC	Distinguished Fellow & former pres, EPI. Dad(4), husband, grandpa(3), Bella (pure mutt). Economist.	7,059

Table 34: Twitter handles table (6/8).

Name	Screen Name	Location	Description	Follower Count
Lars Christensen	MaMoMVPY	Copenhagen, Denmark	International economist, 'Money Doctor', sports analytics nerd, research associate Copenhagen Business School, lacsen@gmail.com +45 52 50 25 06	11,139
Lars P Syll	LarsPSyll	Sweden	Professor at Malmö University. Primary research interest - the philosophy, history and methodology of economics.	1,998
Leah Boustan	leah_boustan		Economic history @Princeton. Stay-at-home mom x3. Jew. Talmud @leah_dafyomi.	27,136
Liz Ann Sonders	LizAnnSonders	New York	Chief Investment Strategist, Charles Schwab & Co., Inc. Disclosures: https://t.co/nswxFWxPYE	155,609
tylercowen	tylercowen		new book *Big Business*, https://t.co/rQrQ30aCGt, Conversations with Tyler, Bloomberg Opinion.	174,264
UK Prime Minister	10DowningStreet	10 Downing Street, London	Official page for Prime Minister @BorisJohnson's office, based at 10 Downing Street	5,851,505
Wiley Economics	WileyEconomics	Global	We publish Economics research and books. We'll help you build your career, get published and connect you.	9,260
William E. Spriggs	WSpriggs	Washington, DC	Chief Economist, @AFLCIO and Professor, Dept. of Economics, @HowardU	10,531
William Easterly	bill_easterly	New York City	NYU Economics Professor. Field is Development. I am an expert on the other experts.	125,833
Wolf Richter	wolfofwolfst	San Francisco	Publisher of finance and econ site https://t.co/Xt9NMmBHdg	19,957
Raoul Pal	RaoulGMI	Little Cayman, Cayman Islands	Founder/CEO - Global Macro Investor and Real Vision Group, Business Cycle Economist, Investment Strategist, Economic Historian, Traveller and Rum Drinker..	297,037
Reuters	Reuters	Around the world	Top and breaking news, pictures and videos from Reuters. For more breaking business news, follow @ReutersBiz.	22,796,915
Rishi Sunak	RishiSunak		Member of Parliament, Richmond. Chancellor of the Exchequer.	436,297
Robert P. Murphy	BobMurphyEcon	Illinois	Christian, Austrian economist, and theorist of non-violent social mechanisms. Senior Fellow at Mises Institute. Author of *Choice*. Host of https://t.co/ByuMumY4eb	46,447
Robert Wenzel	WenzelEconomics	San Francisco	Writes at: https://t.co/7qiNAuFk25 Videos: https://t.co/AM06Hfb4HA Author of "The Fed Flunks: My Speech at the NY Fed" https://t.co/SSWDfwSh2y	4,541
Roger E. A. Farmer	farmerrf	London UK and Los Angeles USA.	Roger Farmer's Economic Window. Professor of Economics, Warwick University and Distinguished Professor of Economics UCLA.	14,025

Table 35: Twitter handles table (7/8).

Name	Screen Name	Location	Description	Follower Count
Sandy Darity	SandyDarity	Durham, NC	Now in the fourth print run From Here to Equality!	41,502
Scott Minerd	ScottMinerd	USA	Global Chief Investment Officer of @GuggenheimPtnrs. These are my views on the global economy and financial markets https://t.co/G7sajGQUxG	87,721
Stephen Redding	ReddingEcon	Princeton, NJ	Economist. Tweets and retweets are not endorsements.	2,301
Steve Burns	SJosephBurns	Tennessee USA	I tweet about trading, financial markets, and financial freedom. I also share what I find inspiring & motivating. I am a trader & the founder of https://t.co/gqmZ6wZxcM	307,616
Steve Keen	ProfSteveKeen	Sutton, London	We need a new economics. Help me build it https://t.co/arS8Ordxqy Debunking Economics, Minsky UCL Honorary Research Fellow profstevekeen@moneybutton.com	74,583
Stocktwits	Stocktwits	New York, NY	The world's largest community of investors and traders.	737,035
Sven Henrich	NorthmanTrader	Britannia	Navigating changing markets. Keeping it real. Occasional sarcasm. To subscribe: ttps://t.co/ZyyUuwbyMz...	236,435
The Economist	TheEconomist	London	News and analysis with a global perspective. Subscribe here: https://t.co/SHA0LZG0e2	25,271,206
The Sunday Times	thesundaytimes	London	The best of our journalism. Become a subscriber: https://t.co/Kq4ltERnQC. Contact our customer service team: https://t.co/VIDSflmdIL	447,915
The Wall Street Journal	WSJ	New York, NY	Sign up for our newsletters and email alerts: https://t.co/WFU7oLKkip	18,351,907
The Washington Post	washingtonpost	Washington, DC	Democracy Dies in Darkness.	16,951,889
Thomas Piketty	PikettyLeMonde		Compte officiel de Thomas Piketty, professeur @EHESS_fr & @PSEinfo, co-directeur @WIL_inequality	172,297
Tim Harford	TimHarford	Oxford	Author of "How To Make The World Add Up". Cautionary Tales podcast. Undercover conomist at the FT. BBC More or Less. Views my own, of course.	172,906
Tim Worstall	worstall	Messines, Portugal	Senior Fellow @ASI Editor: @ItsExpunct Usually in @Forbes, @WSJ, @NYtimes, @Telegraph, DCExaminer.	6,287
TIME	TIME		Breaking news and current events from around the globe.	17,742,072
Timothy Taylor	TimothyTTaylor	Minnesota	Journal of Economic Perspectives, Conversable Economist	7,154
Tom Woods	ThomasEWoods	Harmony, FL	New York Times bestselling author, and host of The Tom Woods. Show	96,602

Table 36: Twitter handles table (8/8).

26.5 UK based news - Guardian news

Field name	Description and example	Format
Data provider: The Guardian News		
first_pub_date	First recorded date that the article was published. This is a primary date field, if missing, then web_pub_date is used [YYYY-MM-DD]	Datetime
web_pub_date	Date that article was published on-line [YYYY-MM-DD]	Datetime
section_name	Section name filter ("Sport", "Business", "Markets").	String
head_line	Headlines of the articles as it is featured in the newspaper	String
author	Author's name as it appears in the newspaper	String
paper_page_num	Page of the print newspaper where the article appeared	Integer
word_count	Number of words in the article, character cound available	Integer
body_text	The text in the article.	String
Data provider: New York Times		
pub_date	date that article was published on-line. Format arrives as date, month, day ["2020-05-02"] , or YYYY-MM-DD	datetime
abstract	Short abstract of the article	String
document_type	This field is used to categorise whether it is video, print, audio and it is used to filter content, to print only	String
web_url	Current URL of the article	String
lead_paragraph	Lead paragraph of the article	String
section_name	Each article is list under sections, such as "Sport", "Business", "Markets". This will be used to filter out data which is not required, such as "Sport"	String
subsection_name	Each article is list under sub-sections, such as "Rugby" within a section of "Sports". This may again be used to filter out data which is not required, such as "Sport"	String
word_count	Number of words used in the article	Integer

Table 37: The Guardian & The New York News data API description, keys, field names and data sources.

26.6 Google GDELT data

Field name	Description and example	Format
Data provider: Google GDELT		
start date	Required - The start date for the filter ["2020-05-01"]	YYYY-MM-DD
end date	Required - The end date for the filter ["2020-05-02"]	YYYY-MM-DD .
num records	The number of records to return. Article list mode, up to 250	Integer
keyword	Return articles containing the exact phrase keyword within the article text. ["climate change"]	String
domain	Return articles from the specified domain. Filtering field ["bbc.co.uk", "nytimes.com"]	String
country	Return articles published list of countries, formatted as the FIPS 2 letter country code. Filtering field ["UK", "US"]	String
theme	Return articles that cover one of GDELT's GKG Themes. A full list of themes. Filtering field ["GENERAL HEALTH]	String
repeat	Return articles containing a single word repeated at least a number of times. Use repeat() eg. repeat = repeat(3, "environment")	String
timelinevol	A timeline of the volume of news coverage matching the filters, represented as a percentage of the total news articles monitored.	Integer
timelinevolraw	Similar to timelinevol, but has the actual number of articles and a total rather than a percentage	Integer
timelinelang	Similar to timelinevol but breaks the total articles down by published language. Each language is column in the DataFrame.	String
timelinesourcecountry	Similar to timelinevol but breaks the total articles down by the country they were published in, returns a DataFrame column	String
timelinetone	a timeline of the average tone of the news coverage matching the filters. See GDELTs documentation for more information.	Float

Table 38: GDELT Events data description, keys, field names and data sources.

26.7 Modelling data

26.7.1 Training data - The Guardian

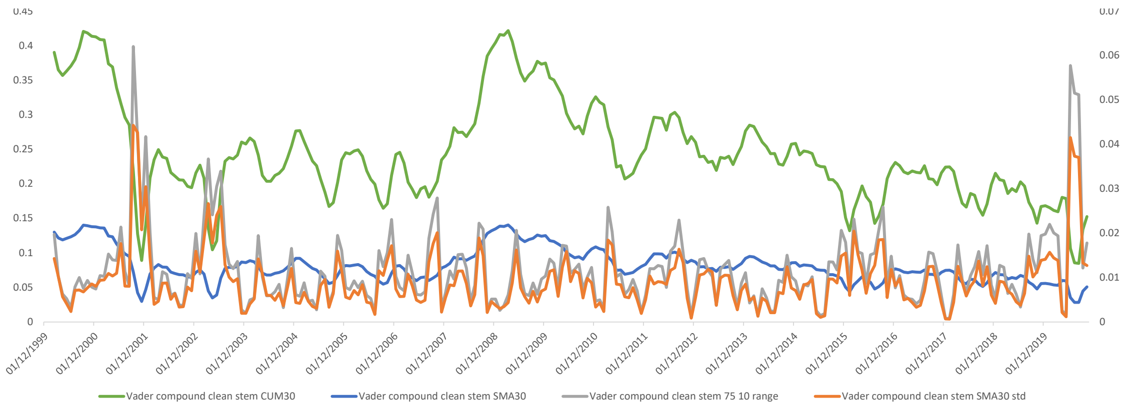


Figure 87: The green line plots the rolling cumulative sentiment value, showing expected increases in the run up to the GFC and a sharp decrease toward the beginning of the pandemic. The standard deviation of the simple moving average (SMA30 day) and the range between the 75th and 10th percentile show a dramatic increase during 2002 and 2003, with an even sharper increase at the beginning of the pandemic.

26.7.2 Training data - New York Times

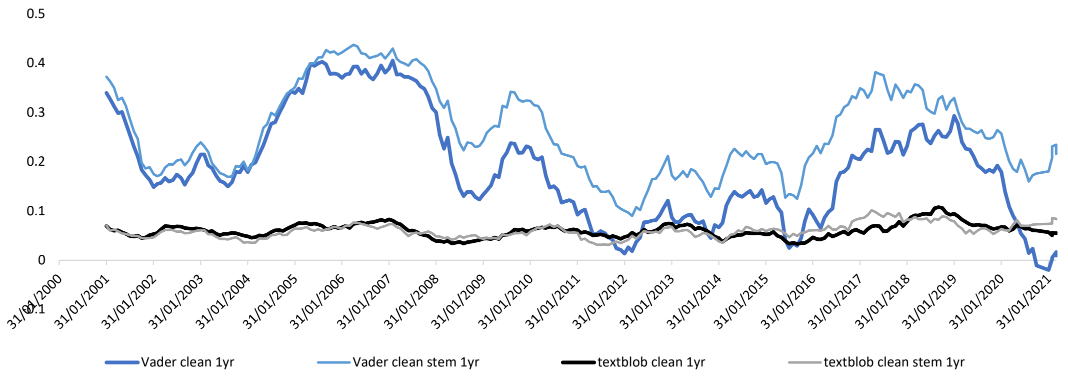


Figure 88: Textblob and VADER, clean plotted and stemmed text plotted for New York Times text. Unlike the same analysis for The Guardian, the stemmed results follow the same trend as the raw text, but it does widen significantly.

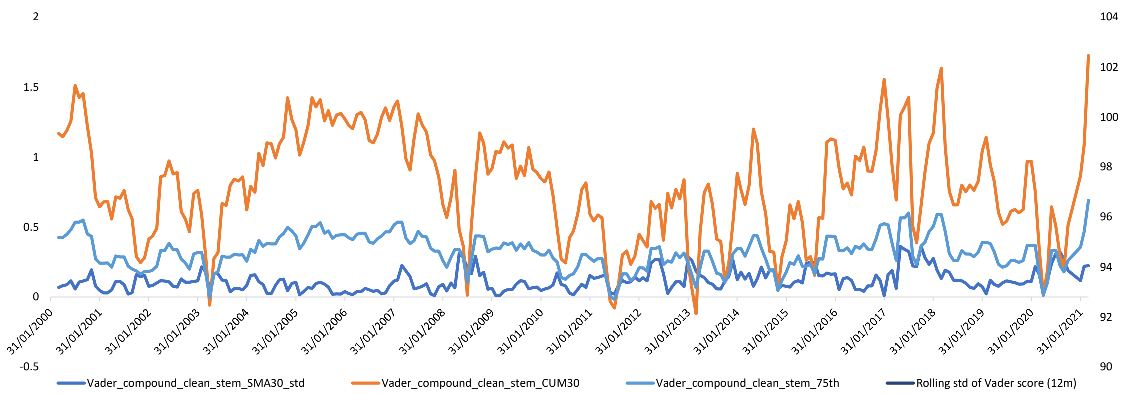


Figure 89: The New York Times cumulative sentiment (CUM30) and simple moving average standard deviation (SMA30 std) are quite a volatile trend, but features shocks around the periods of the 2002 crisis, the GFC. There is a notable increase in sentiment for the CUM30 variable in 2021, which is counter-intuitive.

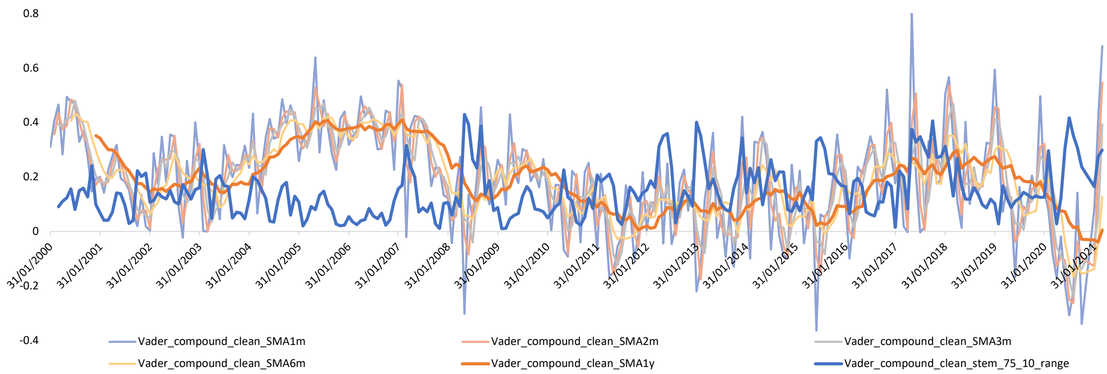


Figure 90: The simple moving average plot of The New York Times show large decreases over the 2002 period, 2008 GFC, and early 2021 pandemic.

26.8 Lexical analysis

New York Times	mean	std	min	25%	50%	75%	max
VADER compound clean	0.090	0.031	-0.021	0.071	0.087	0.105	0.191
VADER compound clean stem	0.080	0.030	-0.151	0.066	0.078	0.092	0.148
Textblob polarity clean	0.079	0.012	0.055	0.071	0.078	0.084	0.214
Textblob polarity clean stem	0.074	0.014	0.053	0.068	0.072	0.078	0.247

Table 39: New York Time lexical summary statistics.

Twitter	mean	std	min	25%	50%	75%	max
VADER compound clean	0.060	0.167	-0.418	-0.025	0.049	0.153	0.422
VADER compound clean stem	0.070	0.149	-0.399	0.017	0.076	0.127	0.494
Textblob polarity clean	0.069	0.068	-0.091	0.038	0.074	0.111	0.268
Textblob polarity clean stem	0.077	0.059	-0.05	0.044	0.075	0.106	0.208

Table 40: Twitter lexical summary statistics.

26.9 Modelling results

26.9.1 LSTM modelling results

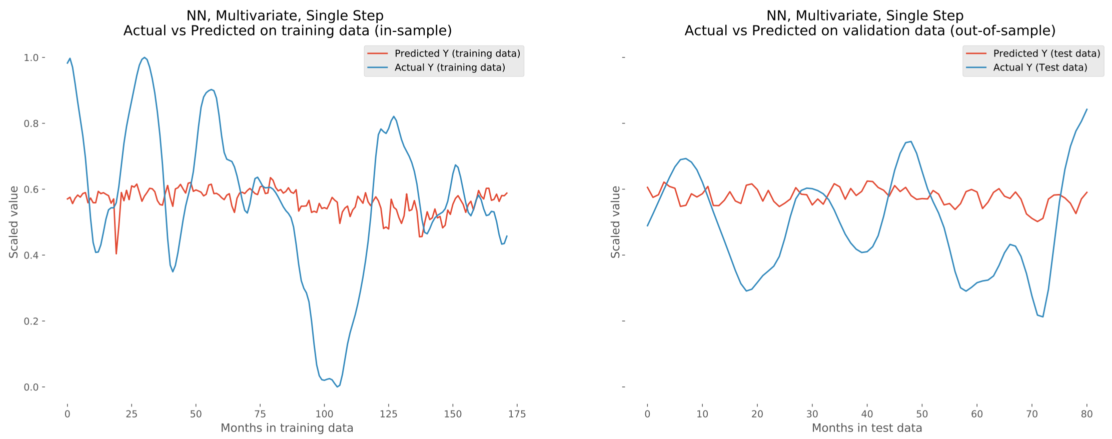


Figure 91: Neural network model with multi-factor and multi-lagged variables (times steps).

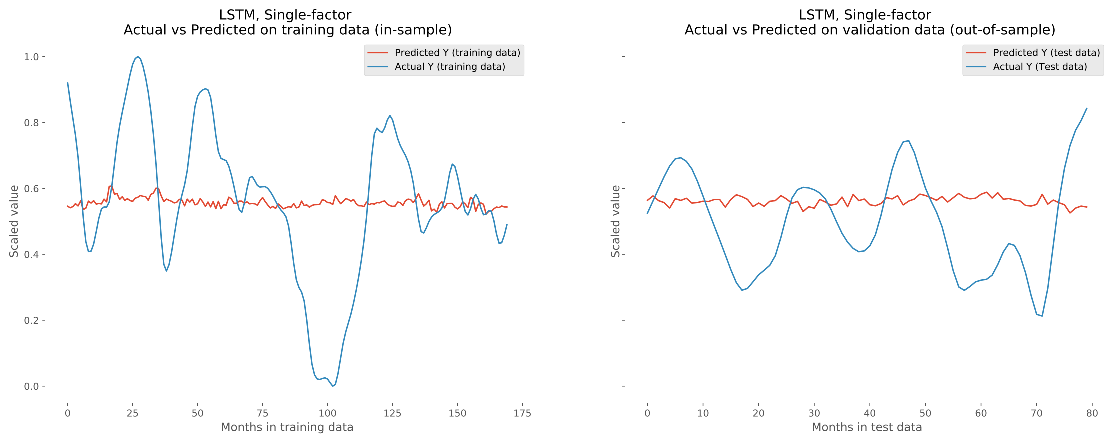


Figure 92: LSTM model results - single factor.

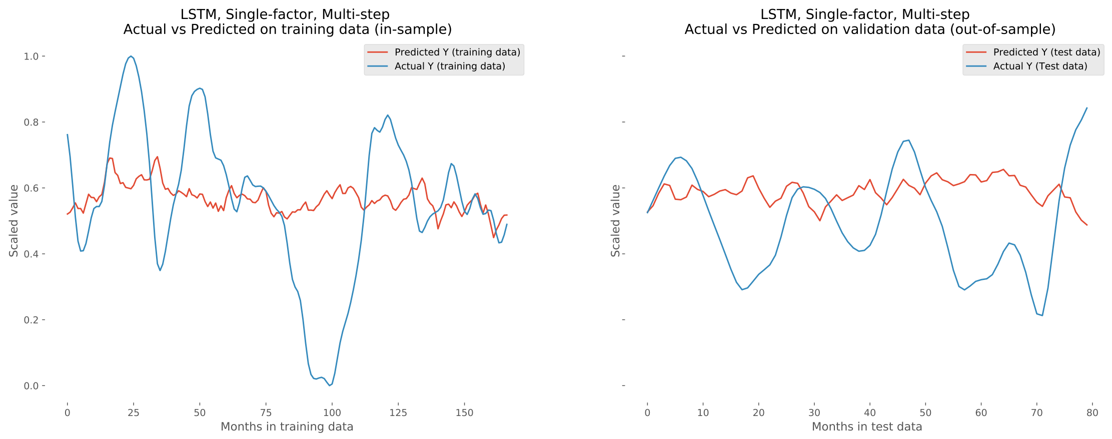


Figure 93: LSTM modelling that includes multi-variate, no lagged variables (times steps).

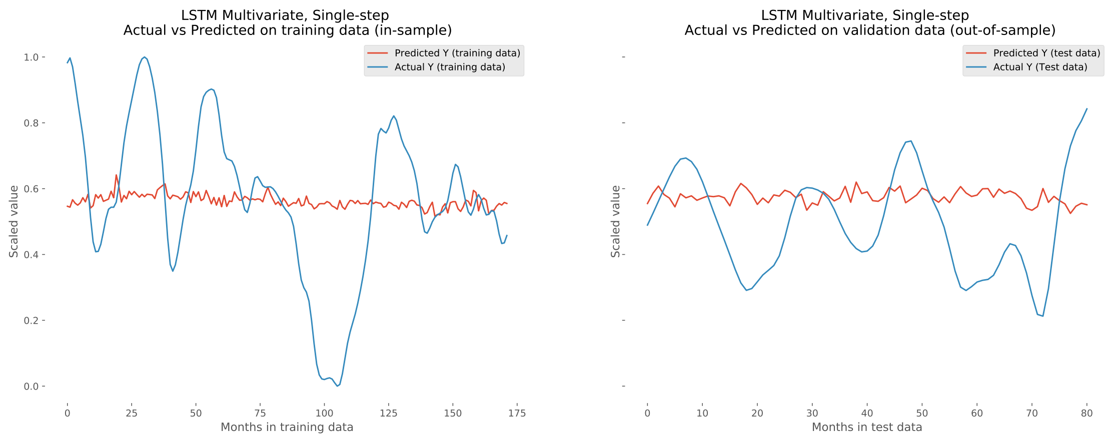


Figure 94: LSTM modelling that includes multi-variate, multi-lagged variables (times steps).

26.9.2 All country results

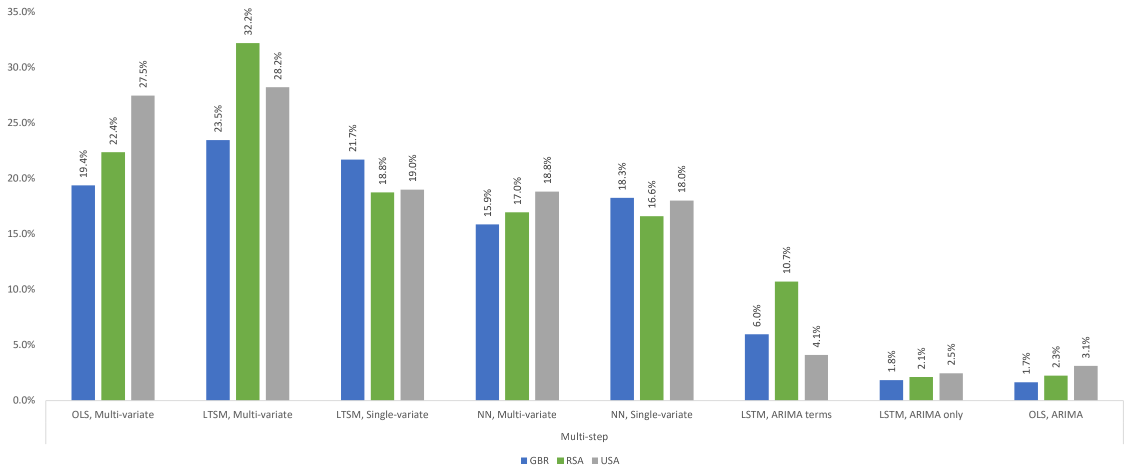


Figure 95: Out-of-sample RMSE test results broken down by factor category, single/multi factor category and region.

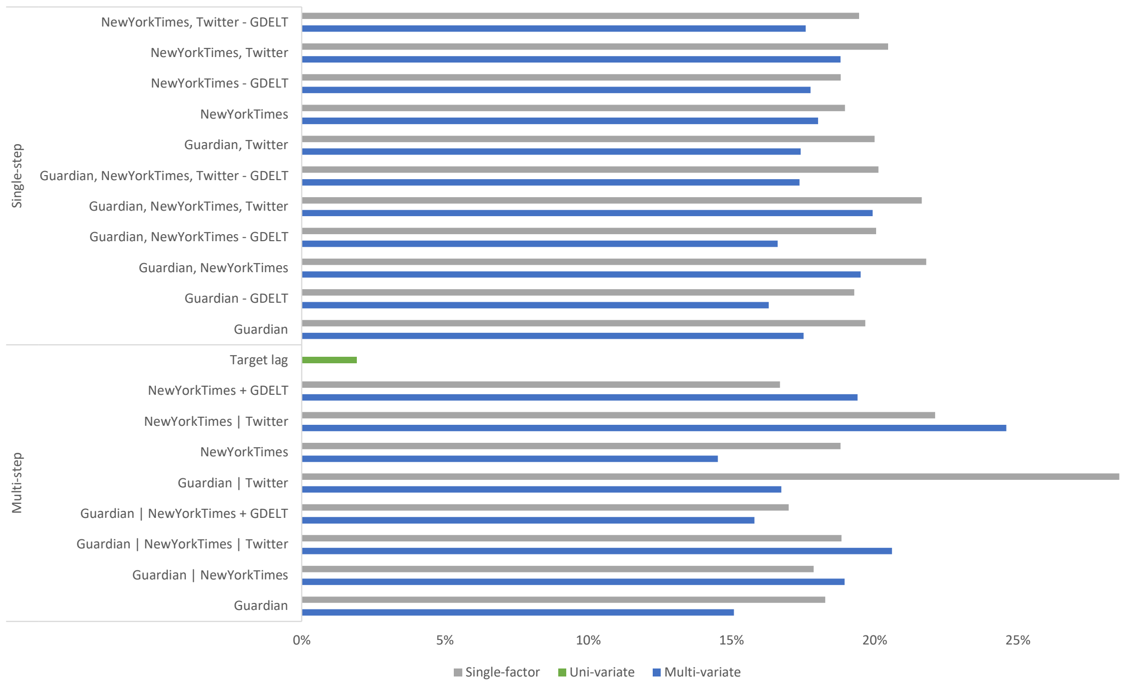


Figure 96: Out-of-sample RMSE test results broken down by data source, single/multi factor category across all regions per country.

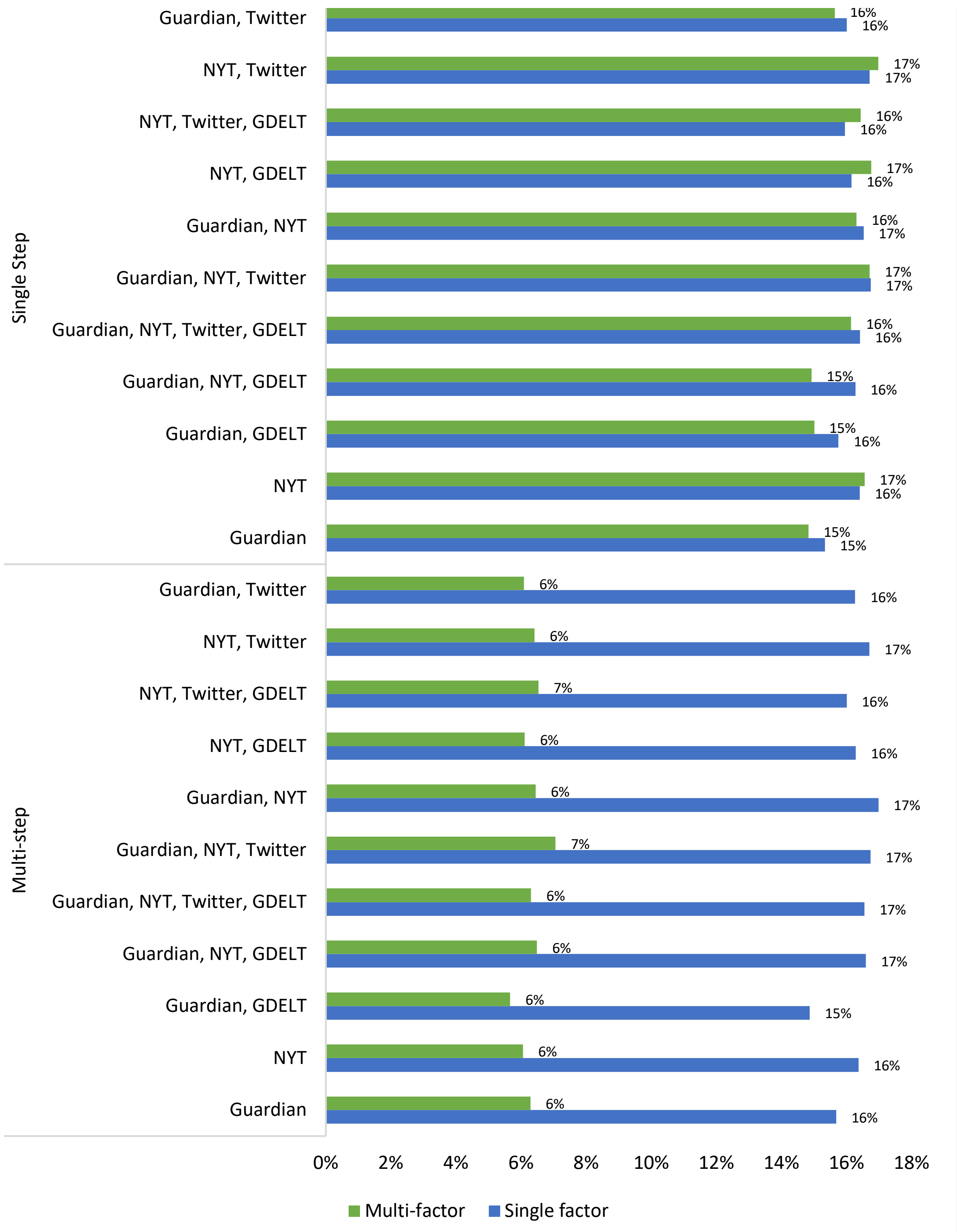


Figure 97: Out-of-sample RMSE test results broken down by data source, single/multi factor category across all regions.

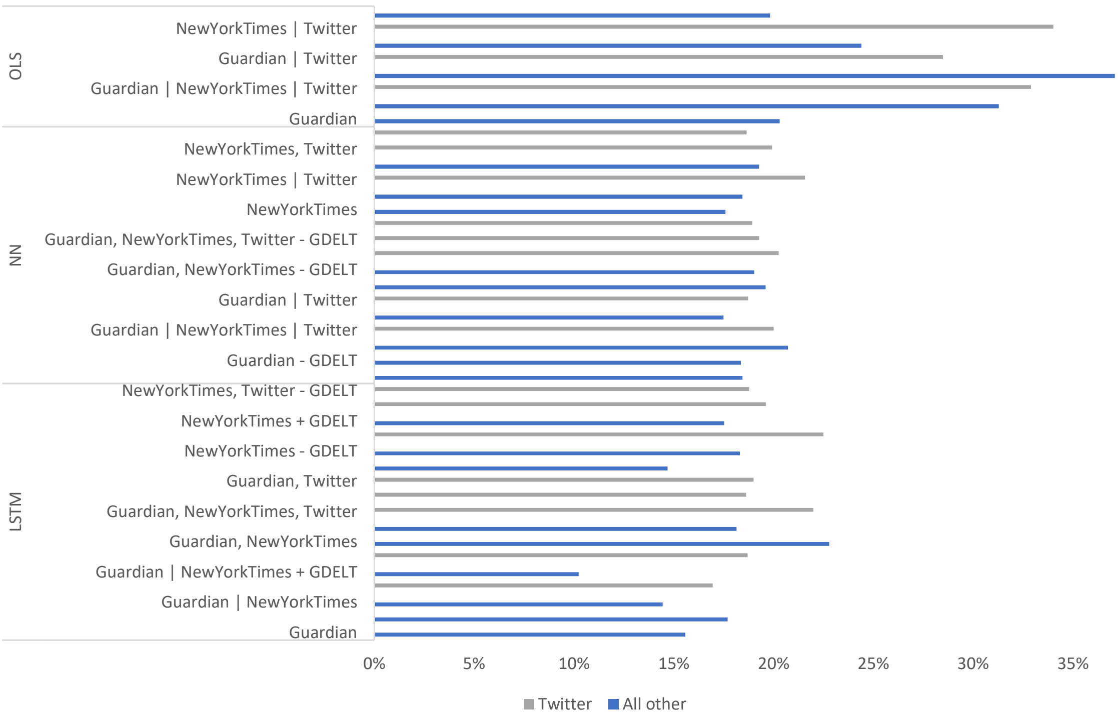


Figure 98: Out-of-sample RMSE test results broken down by data source and region for single/multi factor category split.

26.9.3 UK modelling results

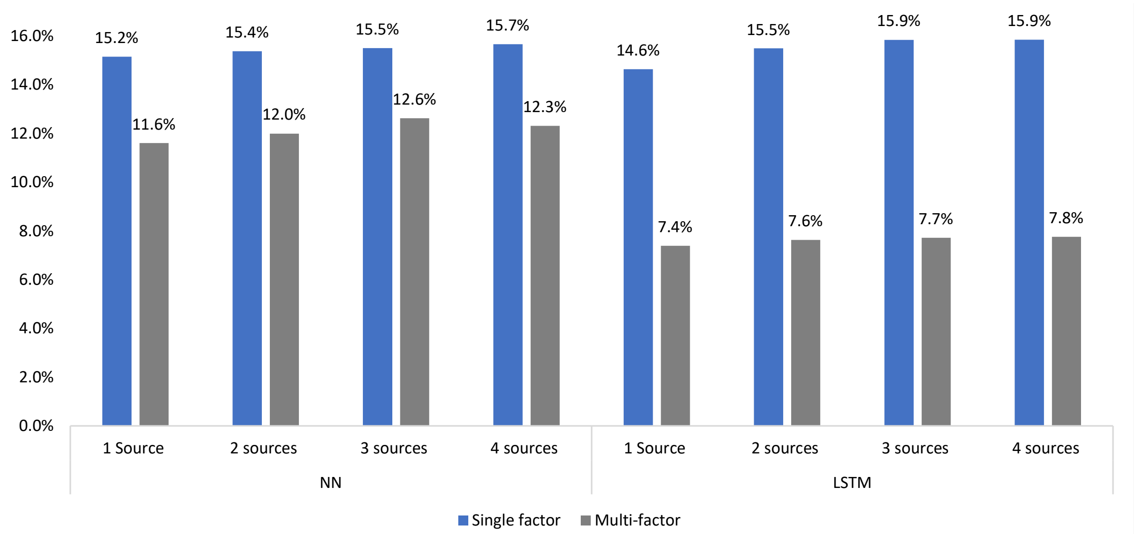


Figure 99: Out-of-sample RMSE test results broken down by model type category and number of data sources.

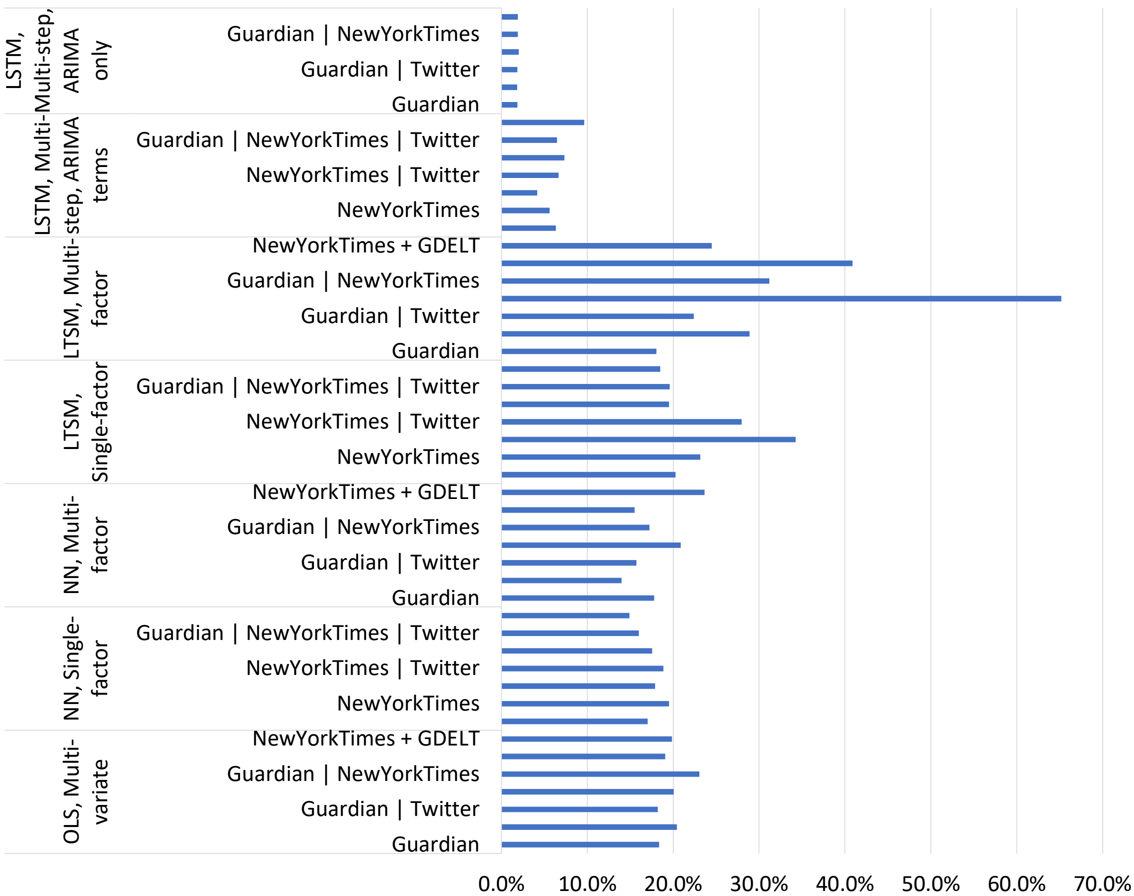


Figure 100: MAE and RMSE test results modelling type category and data sources.

26.9.4 USA modelling results

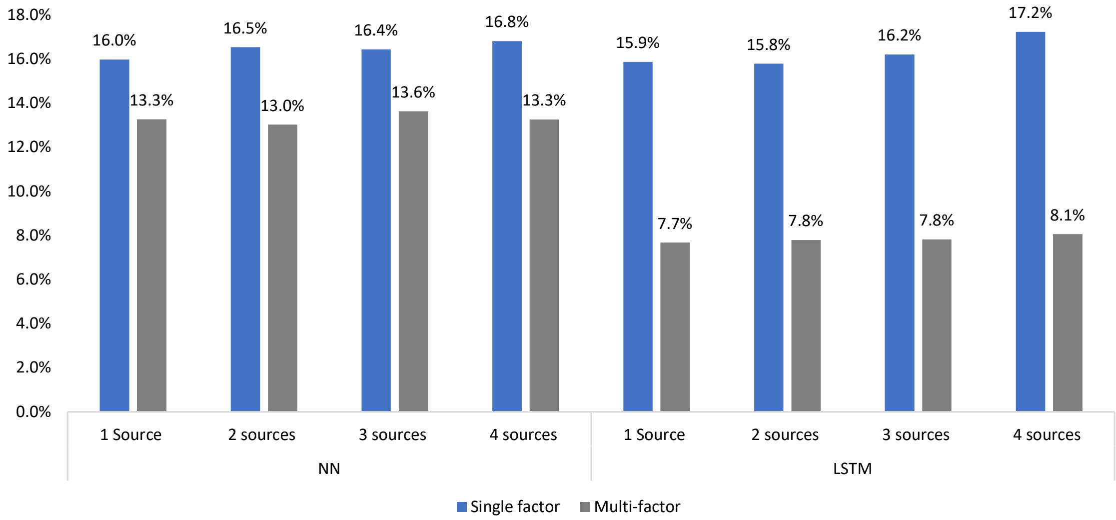


Figure 101: Out-of-sample RMSE test results broken down by model type category and number of data sources.

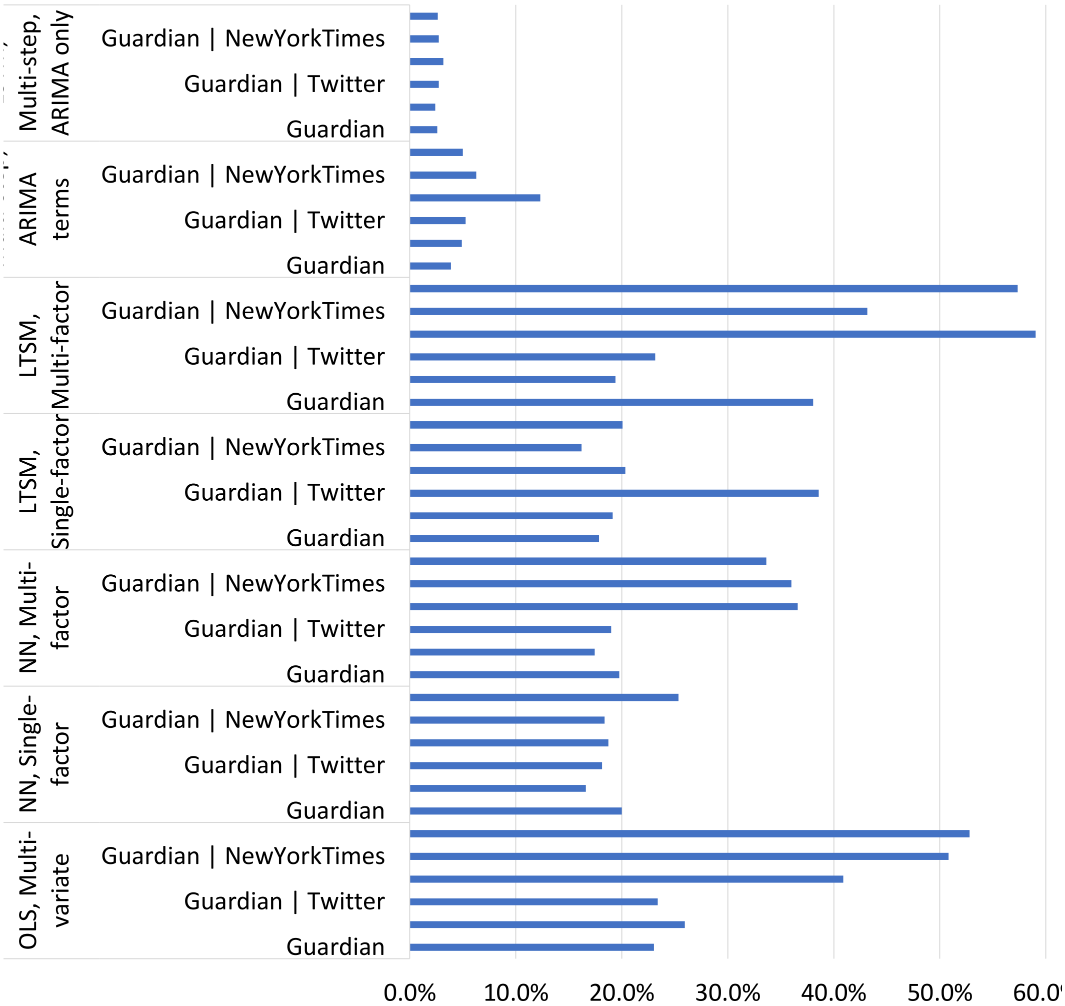


Figure 102: RMSE test results modelling type category and data sources.

26.9.5 RSA modelling results

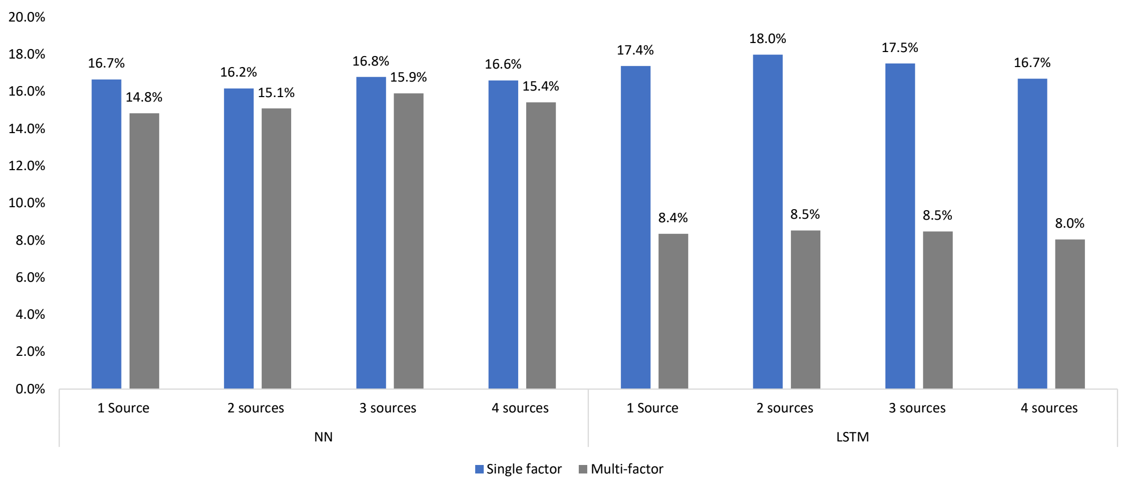


Figure 103: Out-of-sample RMSE test results broken down by model type category and number of data sources.

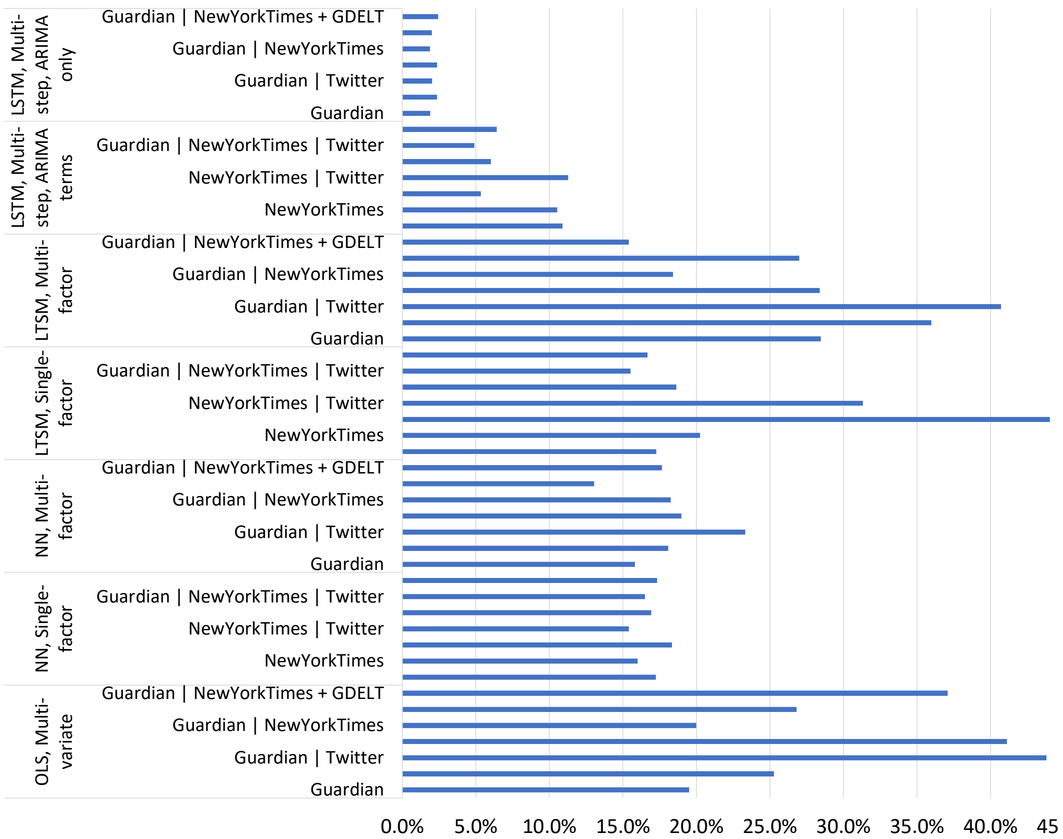


Figure 104: MAE and RMSE test results modelling type category and data sources.

26.9.6 UK sentiment vs credit spread results

From the time-series data I can see a relationship in and around the GFC in 2009 and the period of 2002 to 2004. The relationship is also observed at and around 2016, which reflects the time period of the Brexit vote.

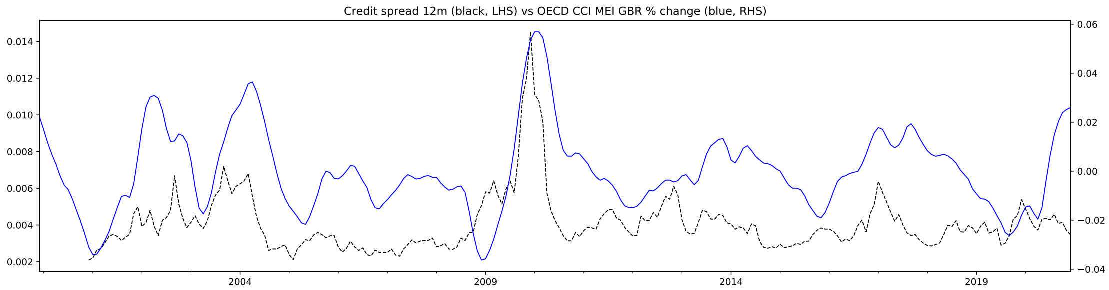


Figure 105: UK OECD BCI index relative to the 12 month UK credit spread 12 month forward.

The Vader, three month rolling window, shows a similar pattern, with drawdowns in sentiment around the shock periods, however, this relationship is not as clear as the Textblob index.

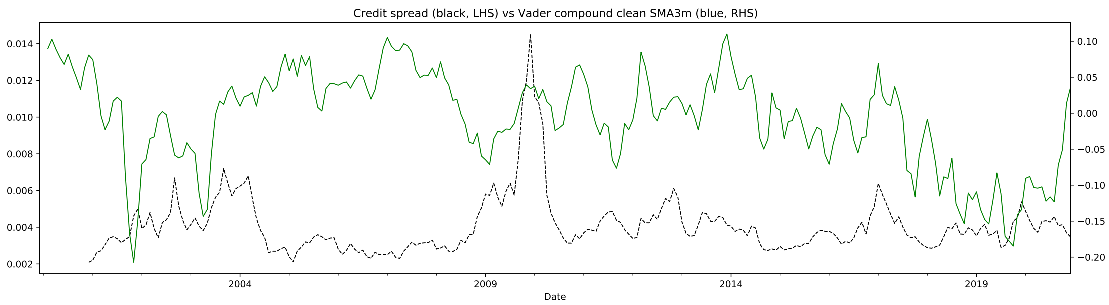


Figure 106: UK Vader based on The Guardian data plotted against the 12 month UK credit spread 12 month forward.

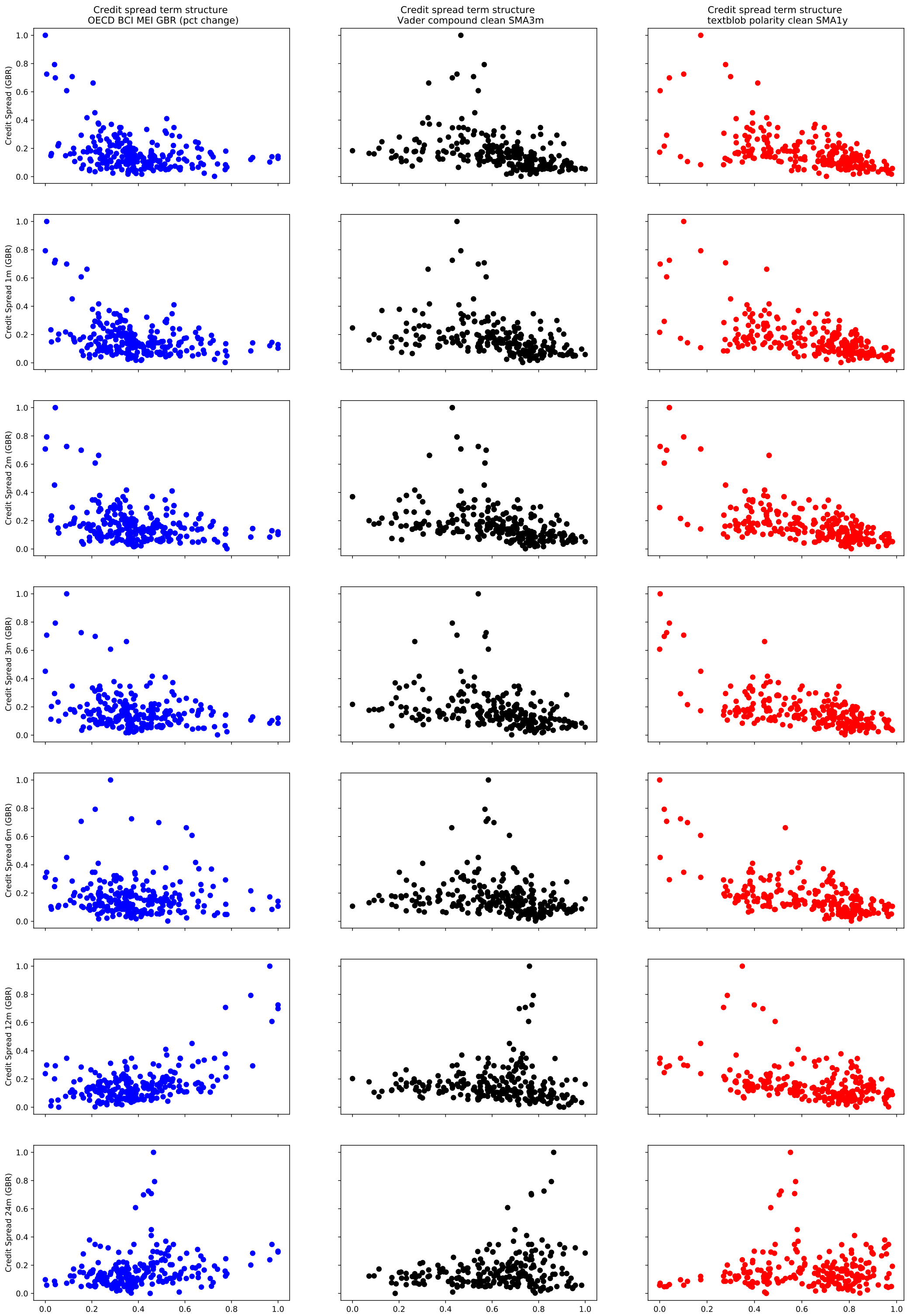


Figure 107: UK Credit spread as it relates to Business Confidence Indicator, Vader and TextBlob.

26.9.7 USA sentiment vs credit spread results

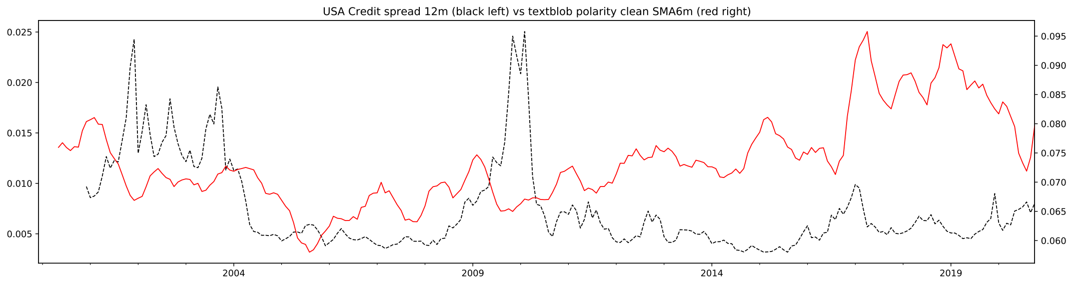


Figure 108: US Credit spread as it relates to Business Confidence Indicator and TextBlob. There are times where the red (textblob) provides a leading indicator for stress events.

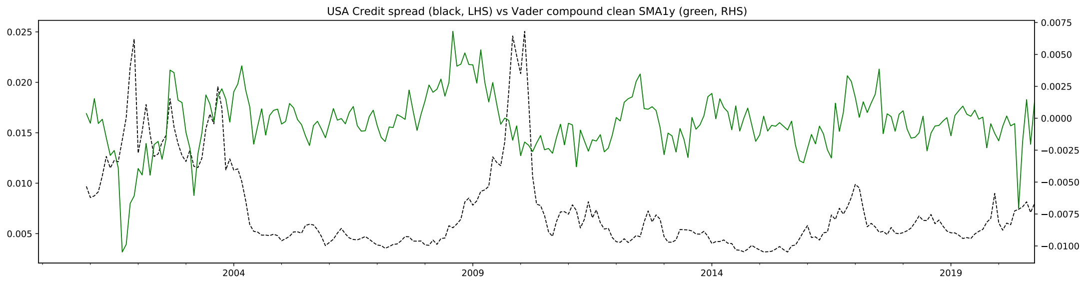


Figure 109: US Credit spread as it relates to Business Confidence Indicator, and simple moving average of the Vader time-series. The green line (Vader SMA) is directionally showing properties, in that the directionality of the green line precedes the credit spread before the GFC, after the GFC too.

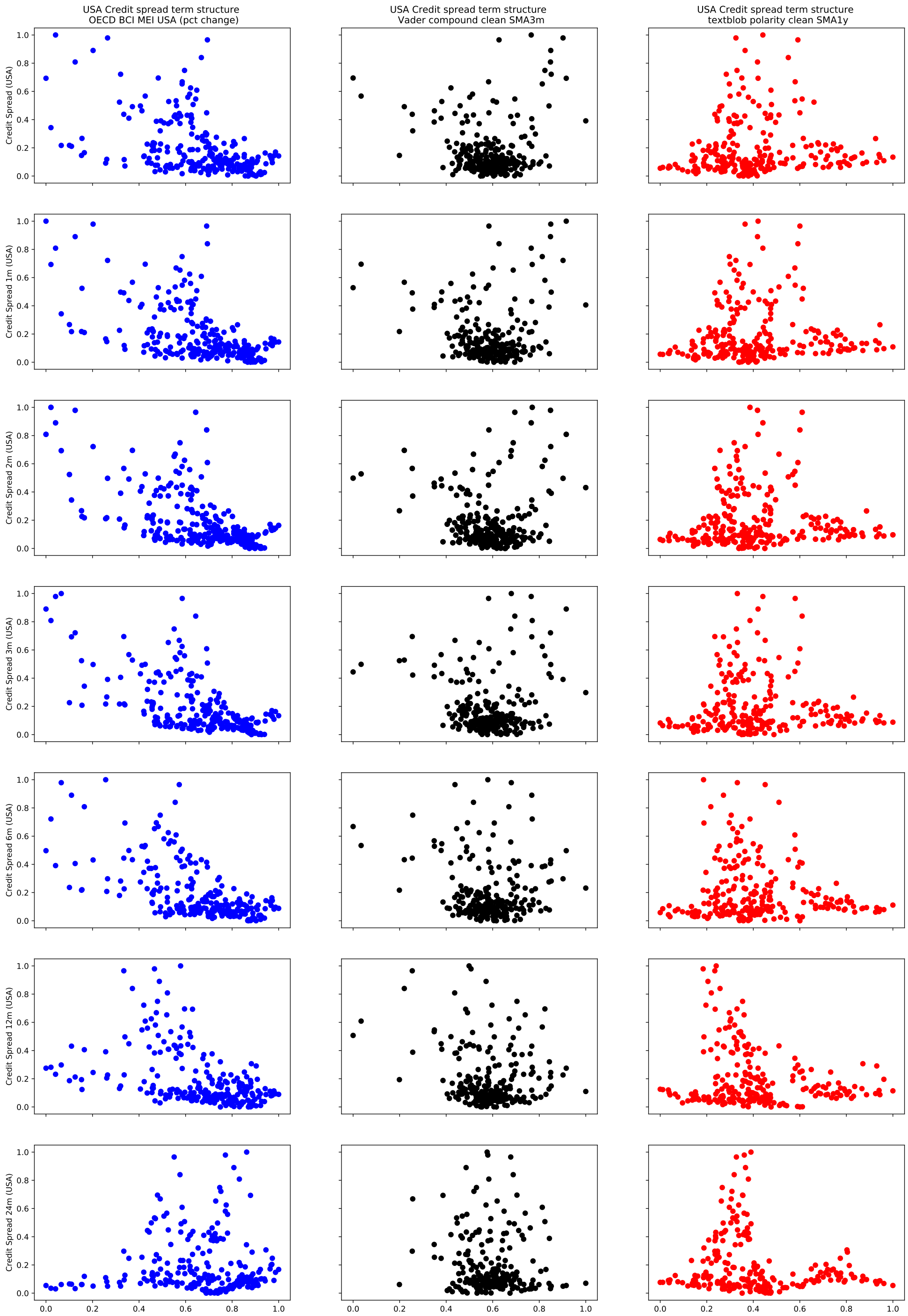


Figure 110: For investigating the correlation between the Business Confidence Indicator, VADER and textblob variables, versus the US credit spread, I make use of a the XY plot, at differing lag time periods and this also helps to review which variabSle shows a more clear signal for credit spread. From this analysis, the signal appears to be more clear for the original OECD variable than it does for the VADER or textblob NLP variables.

26.9.8 RSA sentiment vs credit spread results

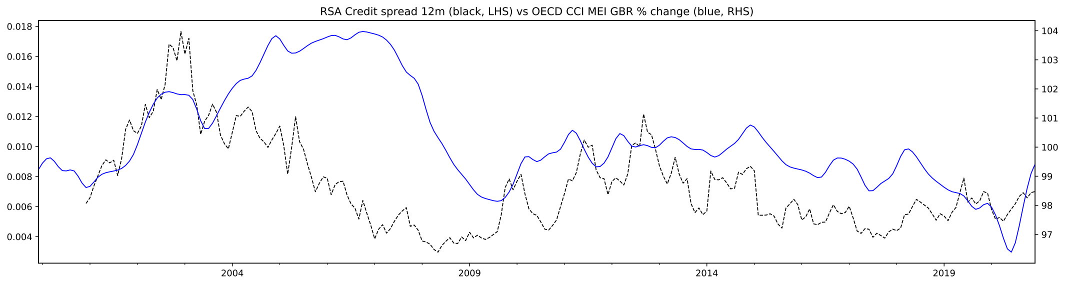


Figure 111: RSA Credit spread as it relates to Business Confidence Indicator, TextBlob polarity. The trend of the textblob time-series does not appear, on visual inspection, to be a leading indicator for the credit spread variable.

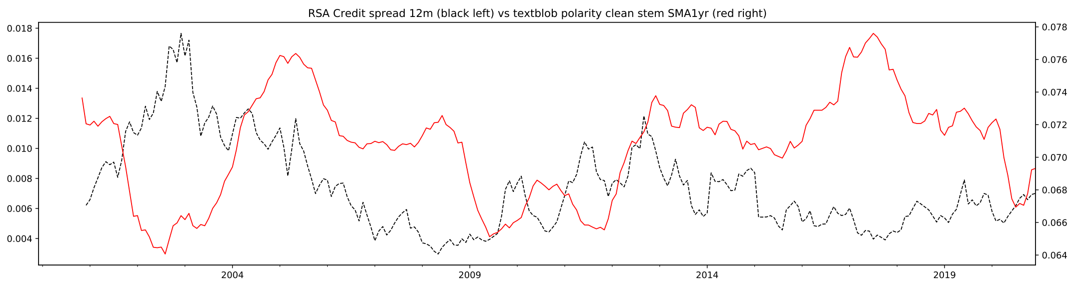


Figure 112: RSA Credit spread as it relates to textblob polarity variable, the trends do not appear to correlate.

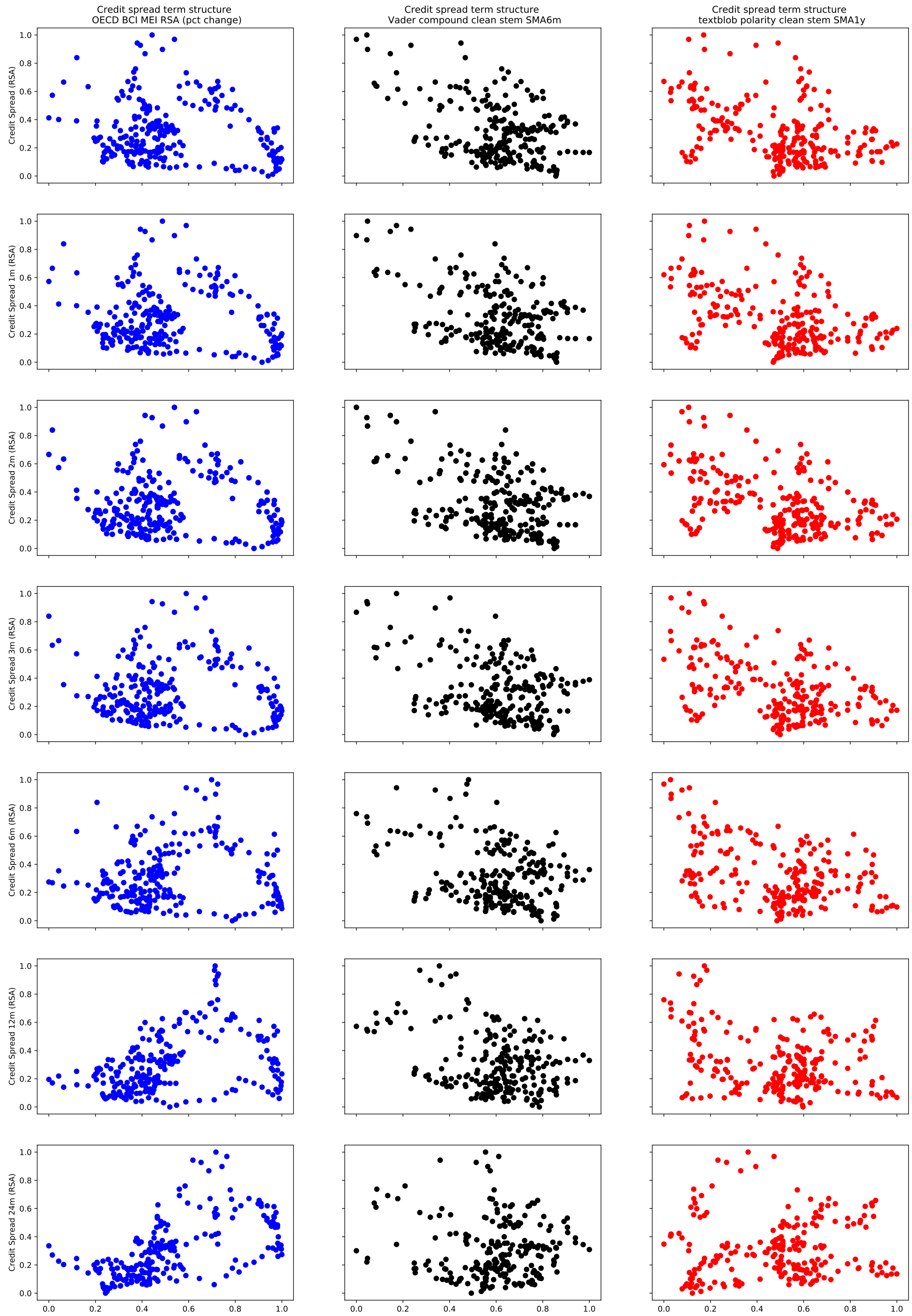


Figure 113: For investigating the correlation between the Business Confidence Indicator, VADER and textblob variables, versus the RSA credit spread, I make use of a the XY plot, at differing lag time periods and this also helps to review which variable shows a more clear signal for credit spread. From this analysis, the signal appears to weaken after the 12 month lag.

27 Economic scenario generator

27.1 Economic scenario assumptions

Region	**UK**		**USA**		**RSA**	
Metric ⟶	Return	Volatility	Return	Volatility	Return	Volatility
Cash & Income						
Cash	1.5%	0.6%	2.0%	0.7%	6.0%	3.0%
Sov Bonds	2.3%	2.2%	2.6%	3.5%	7.5%	4.8%
Inv Bonds	2.6%	4.3%	3.2%	4.5%	8.0%	6.3%
Growth assets						
Equity	6.8%	17.2%	6.0%	18.0%	13.0%	22.0%
Property	6.9%	15.5%	5.4%	16.0%	12.5%	19.5%
Alternatives						
DC Inv	5.8%		5.2%		8.7%	
DC Sub-Inv	7.75%		7.15%		13.0%	
Gold	5.0%	14.6	5.0%	14.7	9.0%	16.5

Table 41: Short term asset class input parameters for CCI=0% (DC = Direct credit / Private debt).

Region	**UK**		**USA**		**RSA**	
Metric $\longrightarrow$	Return	Volatility	Return	Volatility	Return	Volatility
Cash & Income						
Cash	1.5%	0.6%	2.0%	0.7%	6.0%	3.0%
Sov Bonds	2.3%	2.2%	2.6%	3.5%	7.5%	4.8%
Inv Bonds	2.6%	4.3%	3.2%	4.5%	8.0%	6.3%
Growth assets						
Equity	6.8%	17.2%	6.0%	18.0%	13.0%	22.0%
Property	6.9%	15.5%	5.4%	16.0%	12.5%	19.5%
Alternatives						
DC Inv	5.8%		5.2%		8.7%	
DC Sub-Inv	7.75%		7.15%		13.0%	
Gold	5.0%	14.6	5.0%	14.7	9.0%	16.5

Table 42: Short term asset class input parameters for CCI=50%(DC = Direct credit / Private debt).

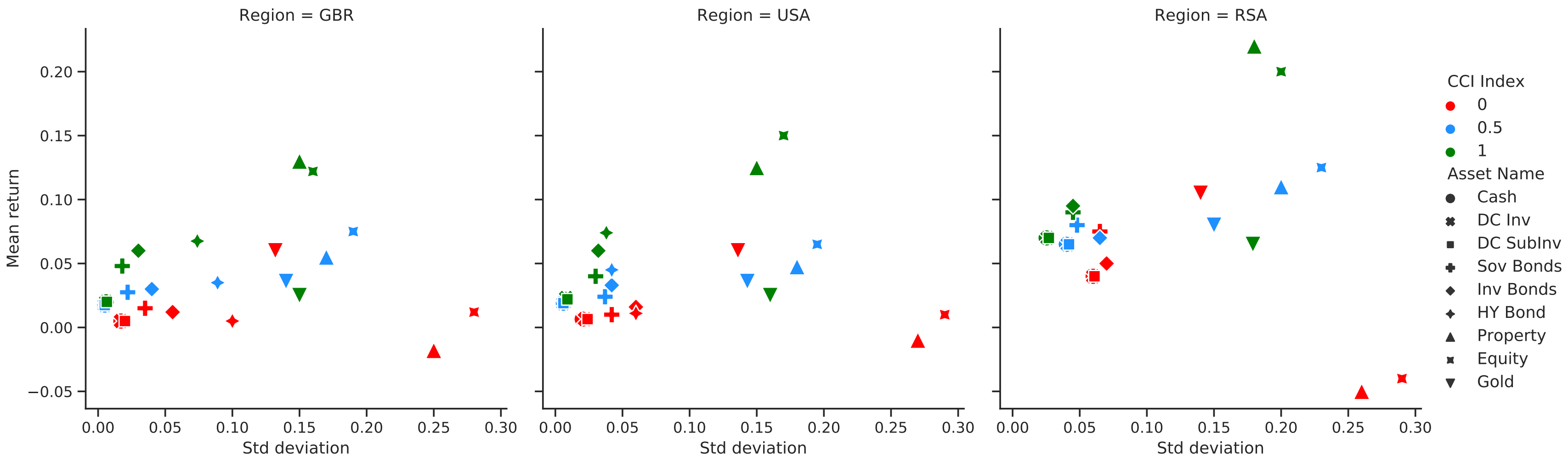


Figure 114: The inputs were divided by region and CCI regime (0%, 50% = E[R], and 100%). The colour in the chart distinguishes the asset category, showing the expected risk and return features of asset classes over an economic cycle. This layout also allows the reader to compare reasonability across regions. Notably, risky assets (green) outperform fixed-income (blue) in terms of return and risk for the boom scenario (CCI = 100%), expected short-term scenario (CCI = 50%), and long-run assumptions (CCI = "LT"). The bust scenario indicates a relatively poor return from growth assets but a decrease across all asset classes. Emerging market returns are expected to be higher risk and higher return, which the Region = "RSA" assumptions reflect.

Region	UK		USA		RSA	
Metric ⟶	Return	Volatility	Return	Volatility	Return	Volatility
Cash & Income						
Cash	1.5%	0.6%	2.0%	0.7%	6.0%	3.0%
Sov Bonds	2.3%	2.2%	2.6%	3.5%	7.5%	4.8%
Inv Bonds	2.6%	4.3%	3.2%	4.5%	8.0%	6.3%
Growth assets						
Equity	6.8%	17.2%	6.0%	18.0%	13.0%	22.0%
Property	6.9%	15.5%	5.4%	16.0%	12.5%	19.5%
Alternatives						
DC Inv	5.8%		5.2%		8.7%	
DC Sub-Inv	7.75%		7.15%		13.0%	
Gold	5.0%	14.6	5.0%	14.7	9.0%	16.5

Table 43: Short term asset class input parameters for CCI=100% (DC = Direct credit / Private debt).

USA Assets	Cash	Sov Bonds	FI corp Inv	FI corp HY	Equity	Property	DC Inv	DC Sub Inv	Gold
Cash	1.00								
Agg Bonds	0.18	1.00							
FI corp Inv	0.04	0.79	1.00						
FI corp HY	-0.08	0.19	0.57	1.00					
Equity	-0.15	0.07	0.27	0.52	1.00				
Property	-0.42	-0.20	-0.10	0.28	0.22	1.00			
DC Inv	0.84	0.17	0.02	0.02	-0.12	-0.20	1.00		
DC Sub Inv	0.54	0.17	0.02	0.02	-0.12	-0.20	0.85	1.00	
Gold	0.10	0.43	0.27	-0.08	-0.02	-0.24	0.30	0.36	1.00

Table 44: US asset correlation assumptions used in the ESG asset simulation.

USA Assets	Cash Cash	Sov Bonds	Corp bonds Inv	Equity	Real estate	DC Inv	DC Sub Inv	Gold
Cash	1.00							
Sov Bonds	0.66	1.00						
Corp bonds Inv	0.42	0.80	1.00					
Large Cap	0.25	0.45	0.55	1.00				
Real Estate	-0.18	0.03	0.05	0.40	1.00			
DC Inv	0.84	0.57	0.40	0.20	-0.12	1.00		
DC Sub Inv	0.84	0.57	0.40	0.20	-0.12	0.85	1.00	
Gold	0.06	0.02	0.15	0.36	-0.27	0.05	0.10	1.00

Table 45: RSA asset correlation assumptions used in the ESG asset simulation.

27.2 Simulation results and visualisations

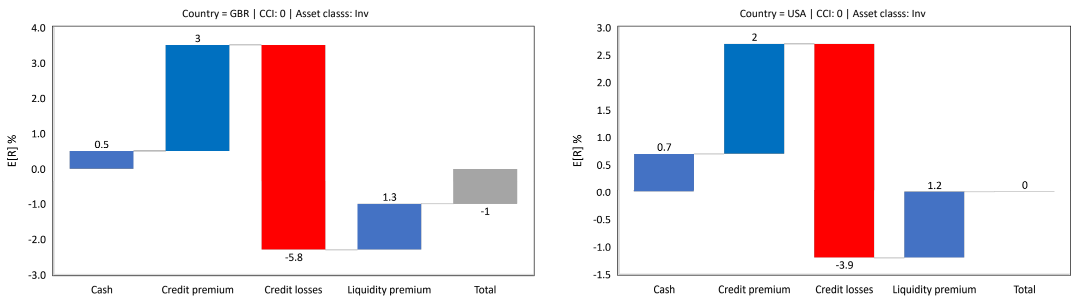


Figure 115: UK and US Investment grade CCI=0%.

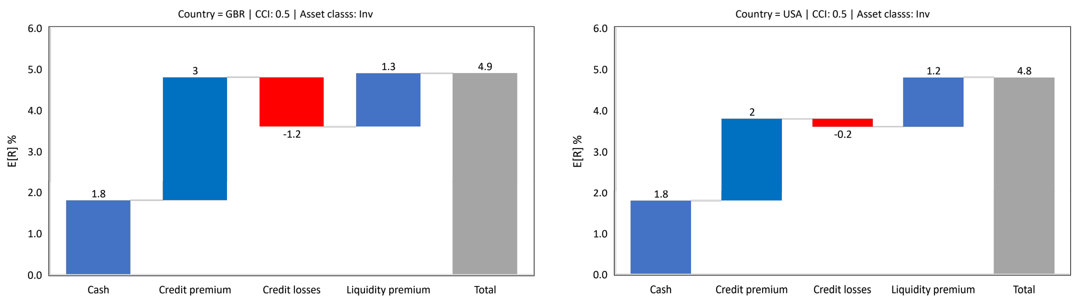


Figure 116: UK and US Investment grade CCI=50%.

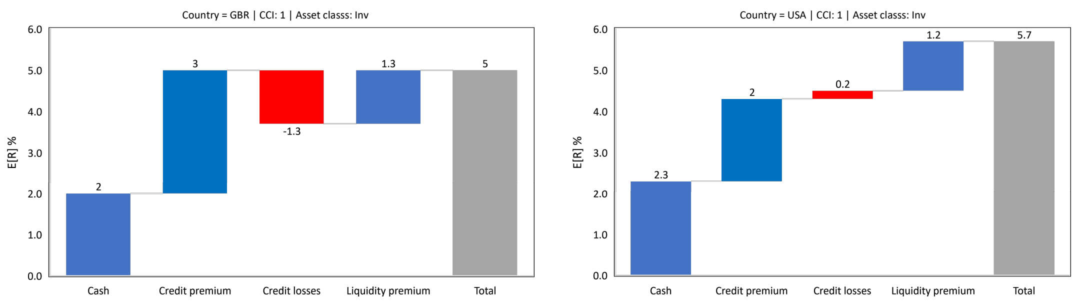


Figure 117: Cost of credit for the sub-investment grade credit block (UK on the left, USA on the right), per CCI deciles, where CCI = 1 is top credit cycle decile and CCI=0 is the lowest CCI decile. Note the significant cost differentiate and upfront planning required for the premium.

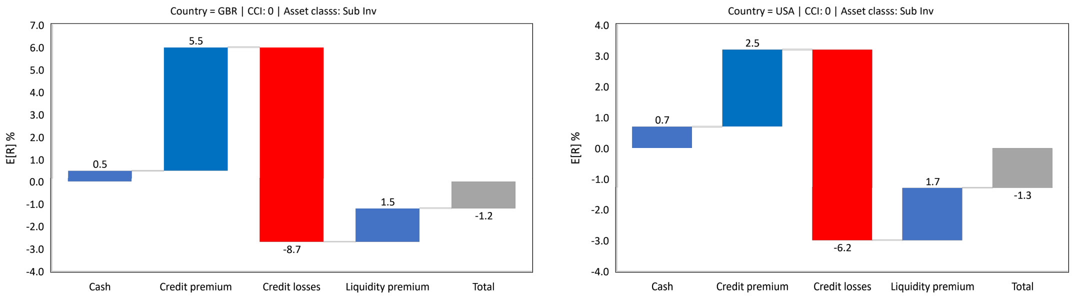


Figure 118: UK and US Sub-Investment grade CCI=0%.

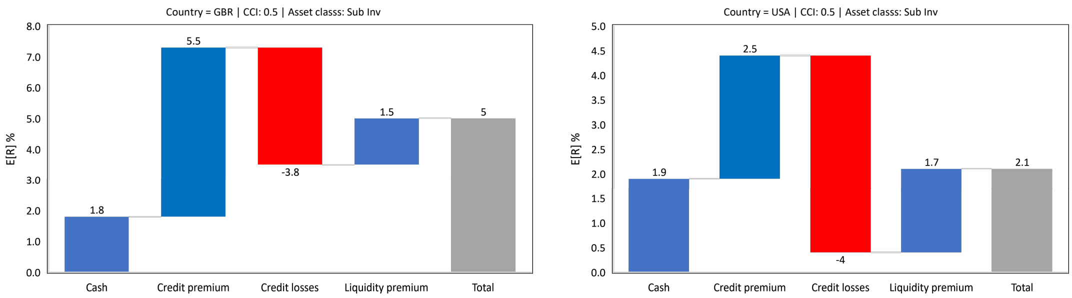


Figure 119: UK and US Sub-Investment grade CCI=50%.

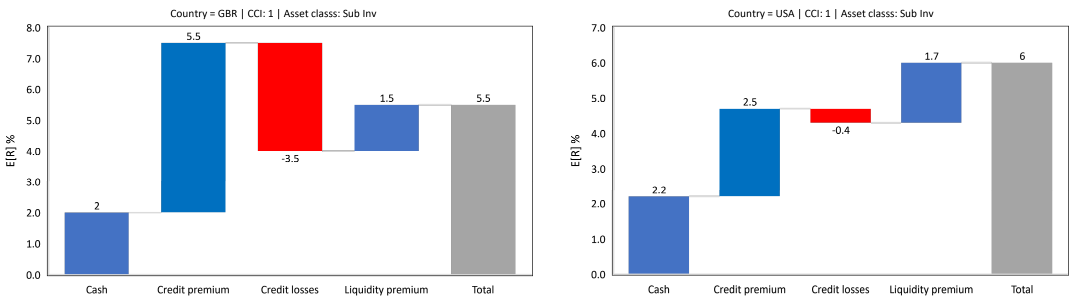


Figure 120: Cost of credit for the sub-investment grade credit block (UK on the left, USA on the right), per CCI deciles, where CCI = 1 is top credit cycle decile and CCI=0 is the lowest CCI decile. Note the significant cost differentiate and upfront planning required for the premium.

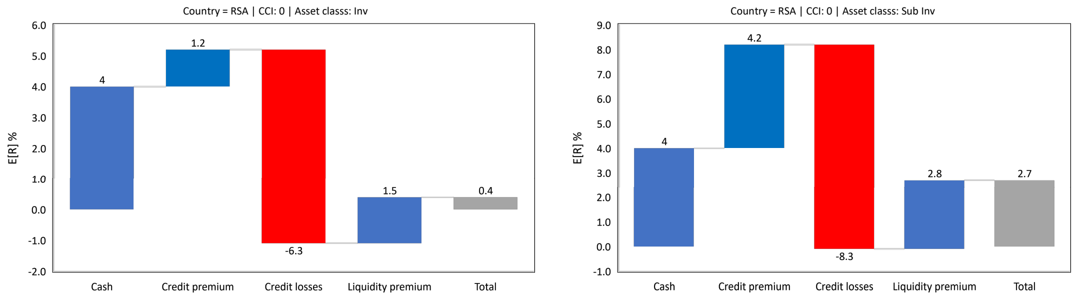


Figure 121: EL waterfall CCI=0% DC Investment grade and Sub-Investment grade RSA.

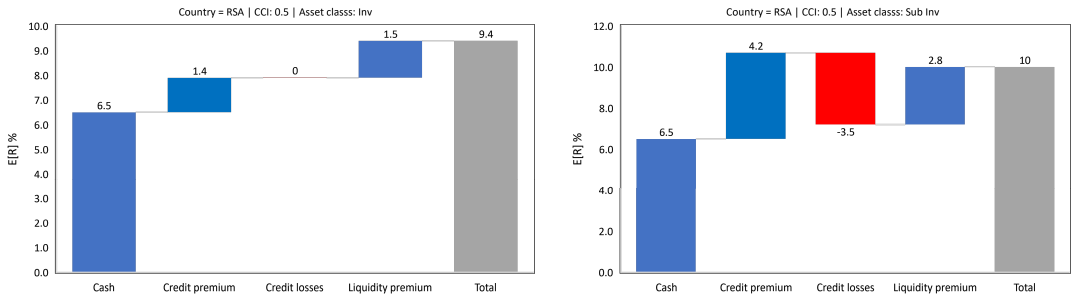


Figure 122: EL waterfall CCI=50% DC Investment grade and Sub-Investment grade RSA.

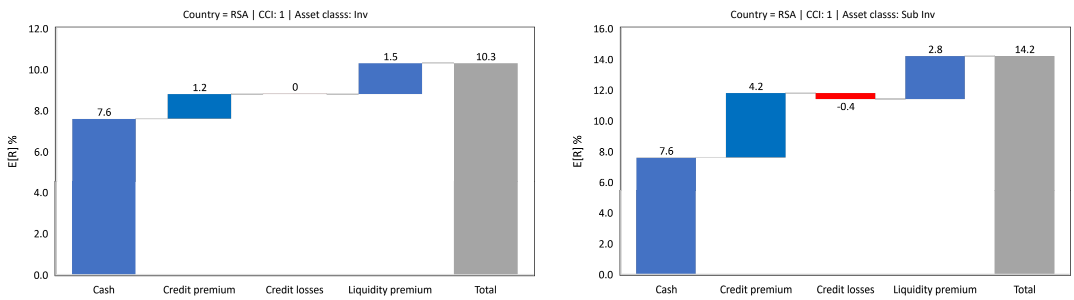


Figure 123: Cost of credit for the RSA region (Investment on the left, Sub-investment on the right), per CCI deciles, where CCI = 1 is top credit cycle decile and CCI=0 is the lowest CCI decile. Note the significant cost differentiate and upfront planning required for the premium.

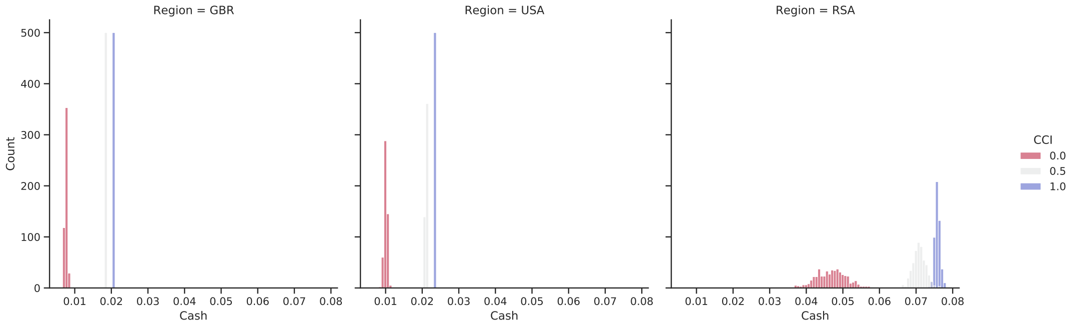


Figure 124: ESG distributions 1yr cash.

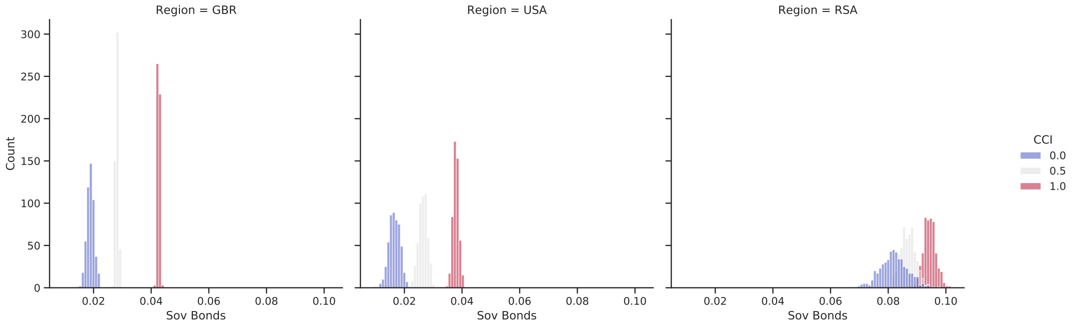


Figure 125: ESG distributions 1yr Sov Bonds.

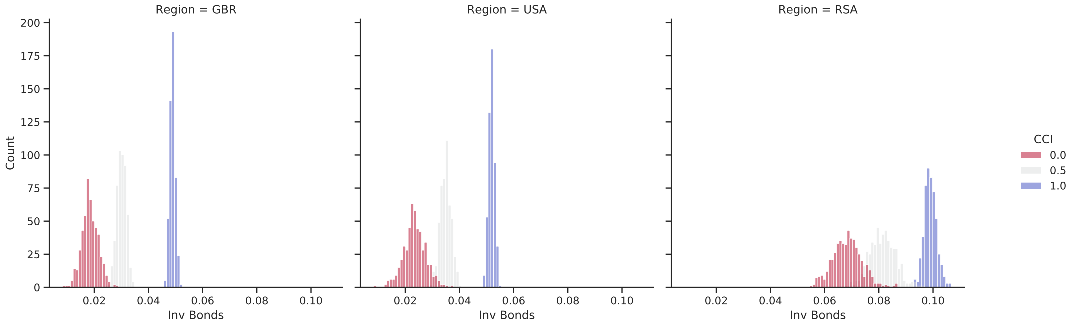


Figure 126: ESG distributions 1yr Inv Bonds.

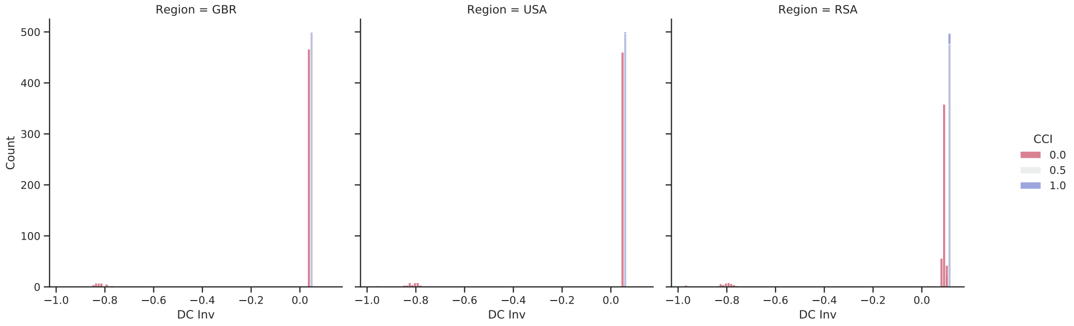


Figure 127: ESG distributions 1yr DC Investment grade.

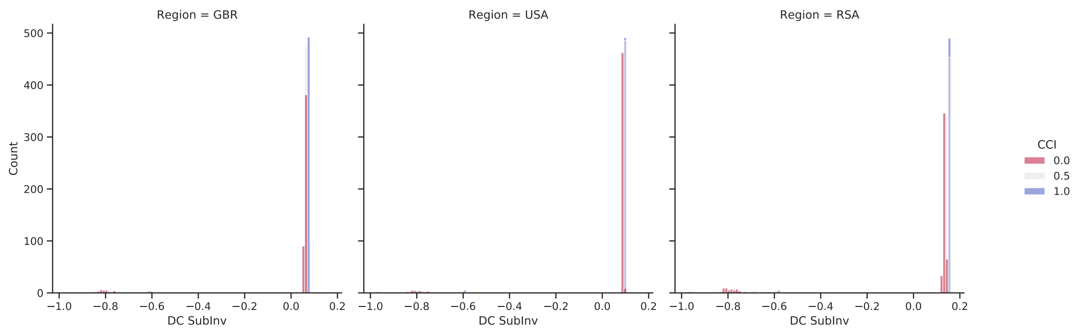


Figure 128: ESG distributions 1yr Sub-Investment grade.

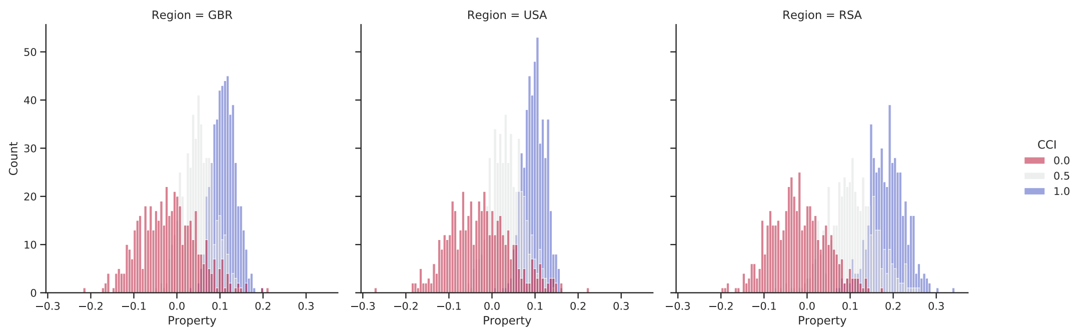


Figure 129: ESG distributions 1yr property.

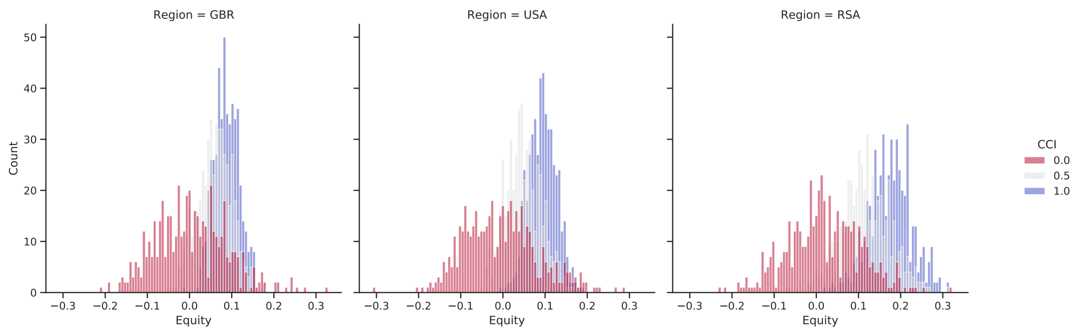


Figure 130: ESG distributions 1yr equity.

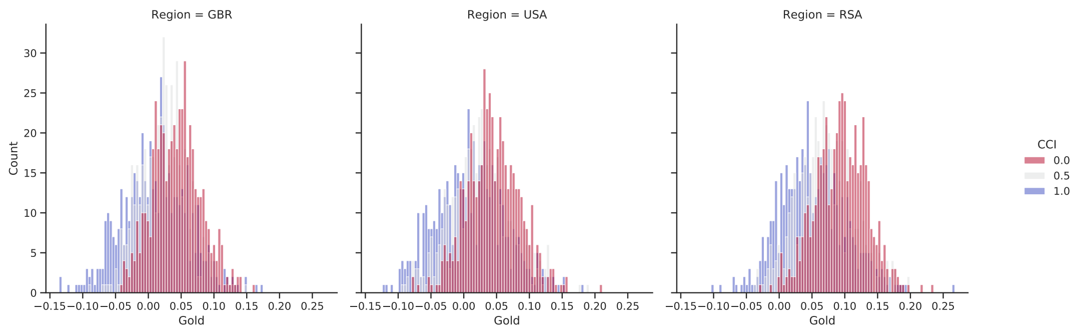
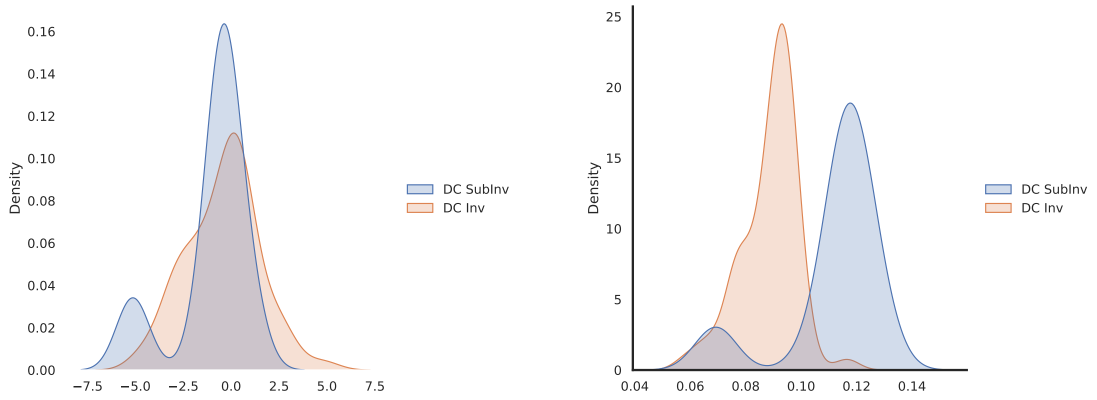


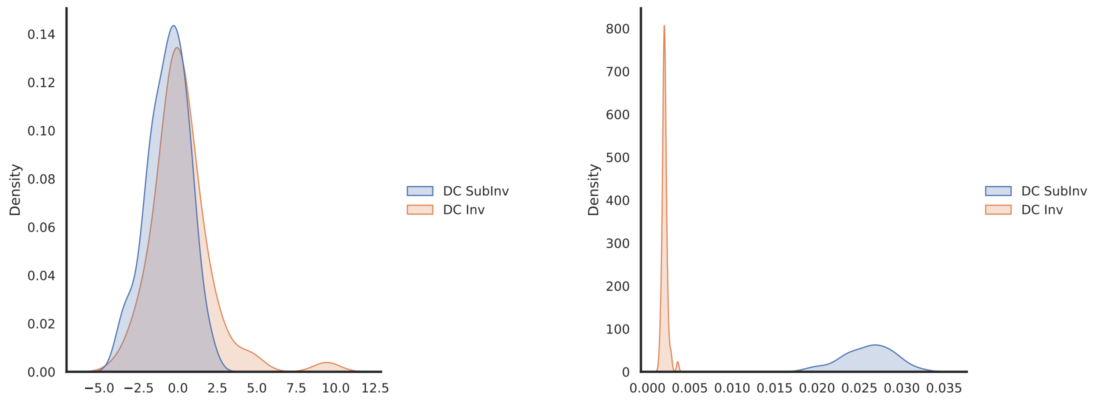
Figure 131: ESG distributions 1yr gold.



(a) Sampled PD distributions.

(b) Scaled PD distributions.

Figure 132: Credit distribution simulation example, note the scaling difference between DC Investment grade and DC Sub-Investment grade.



(a) Sampled PD distributions.

(b) Scaled PD distributions.

Figure 133: Credit distribution simulation example, note the scaling difference between DC Investment grade and DC Sub-Investment grade, significant different in scaled values between Inv and Sub-Investment grade in this specific scenario/sample.

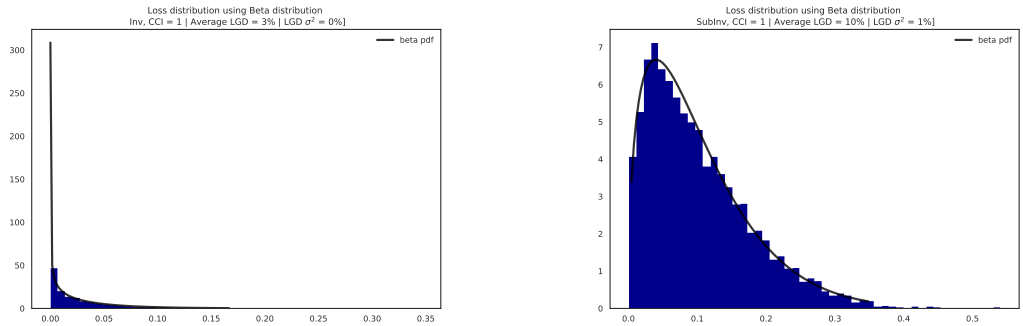


Figure 134: LGD distribution for CCI=100%.

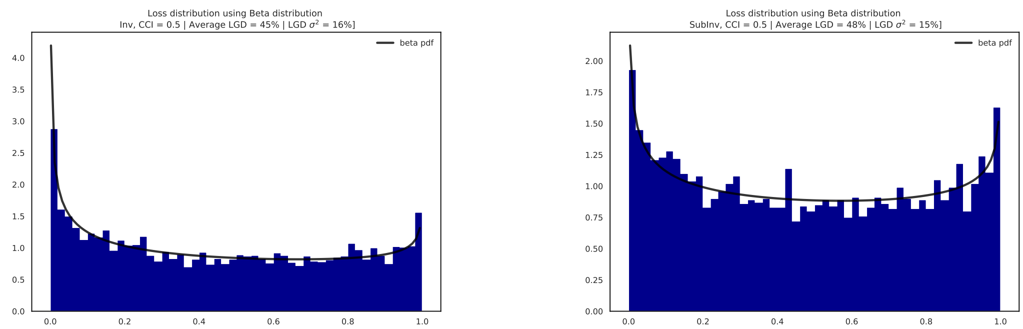


Figure 135: LGD distribution for CCI=50%.

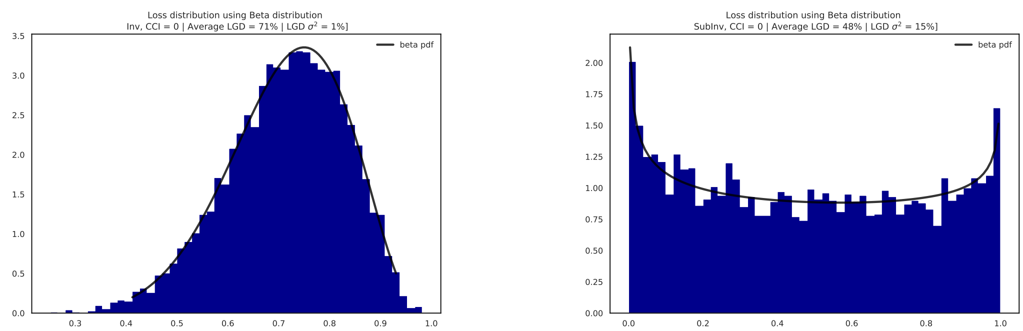


Figure 136: LGD distribution for CCI=0%.

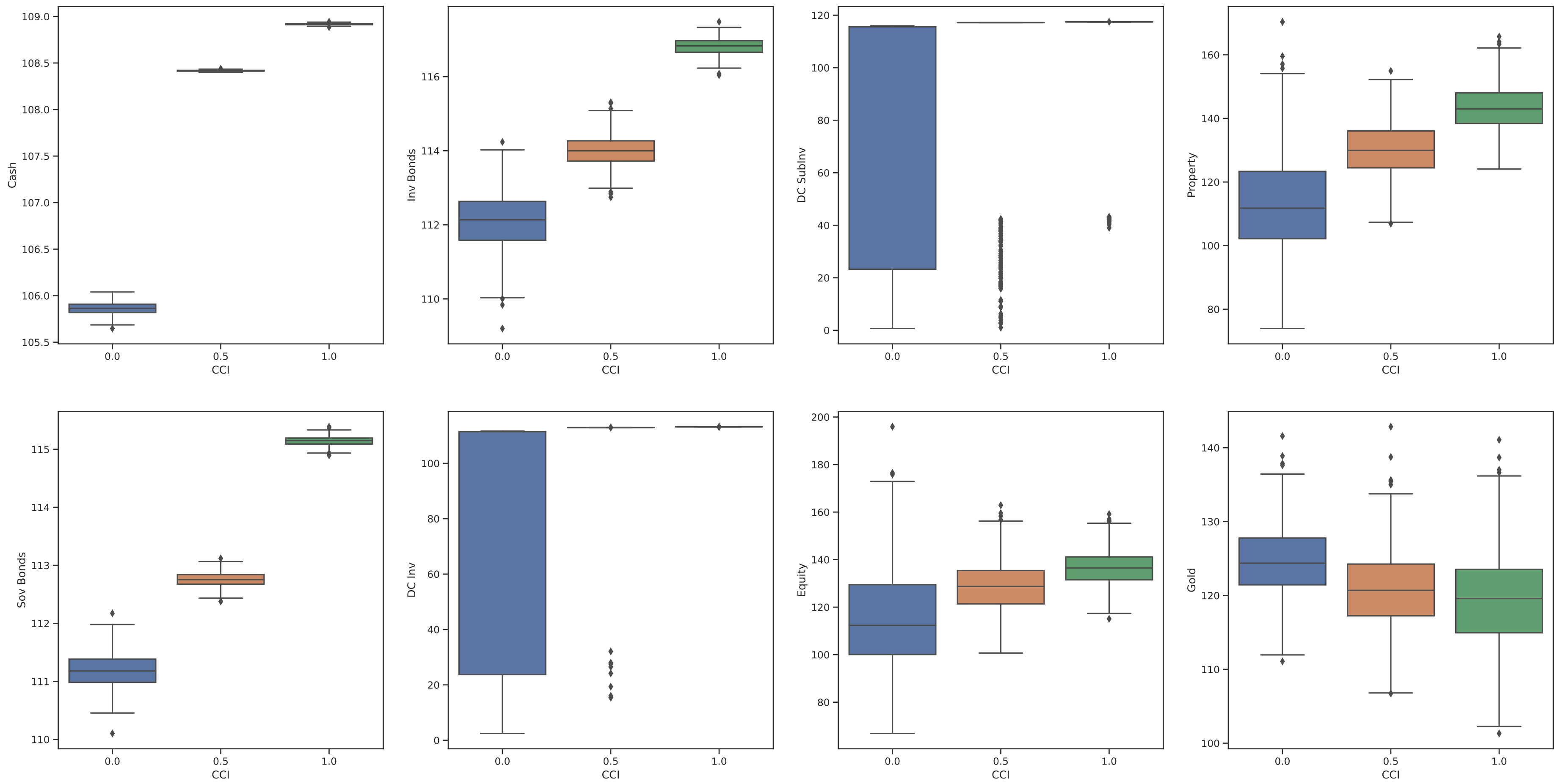


Figure 137: UK Asset class CAGR distributions from one year of investing, split by CCI category.

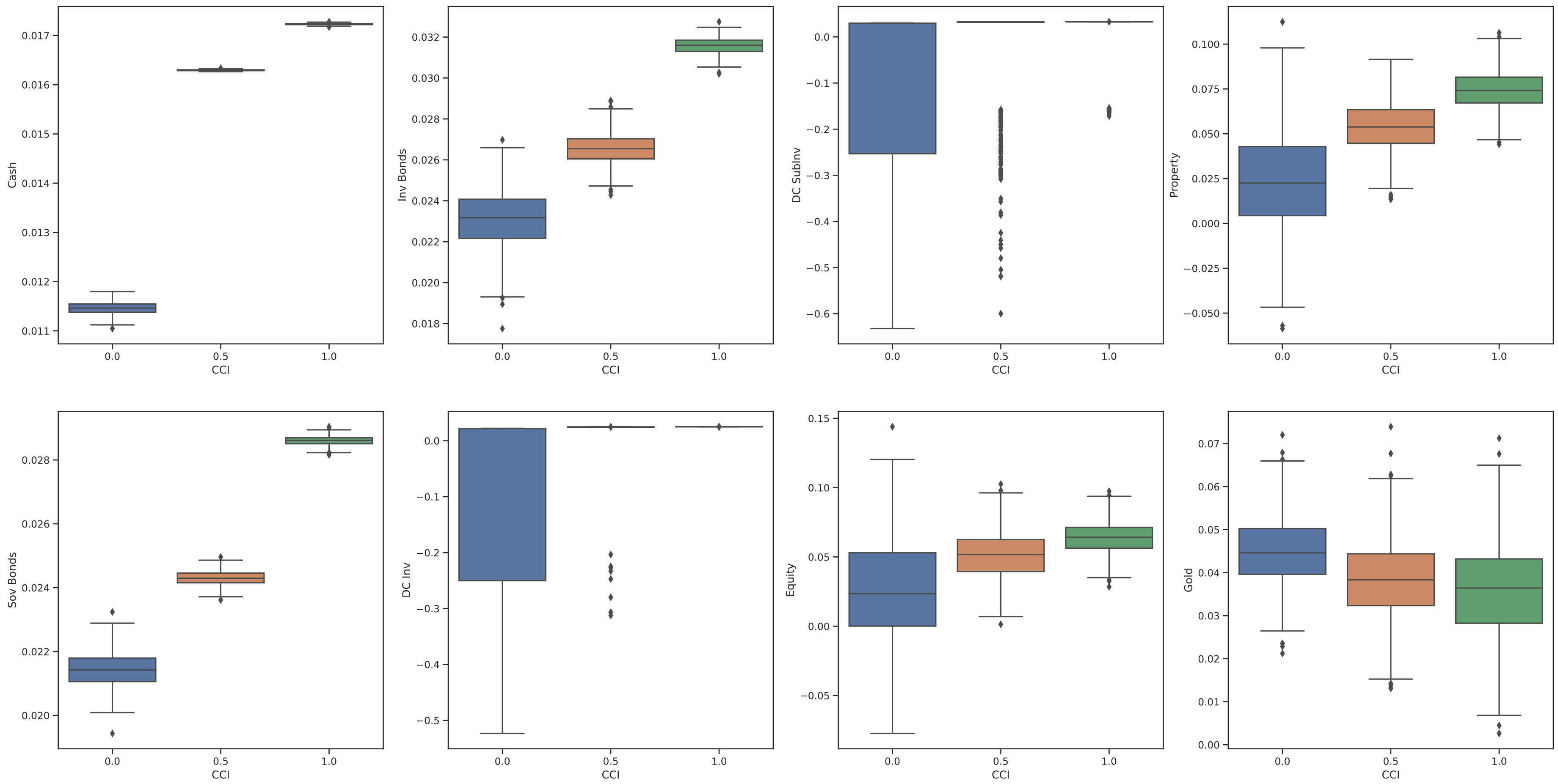


Figure 138: UK Asset class CAGR distributions from five years of investing, split by CCI category.

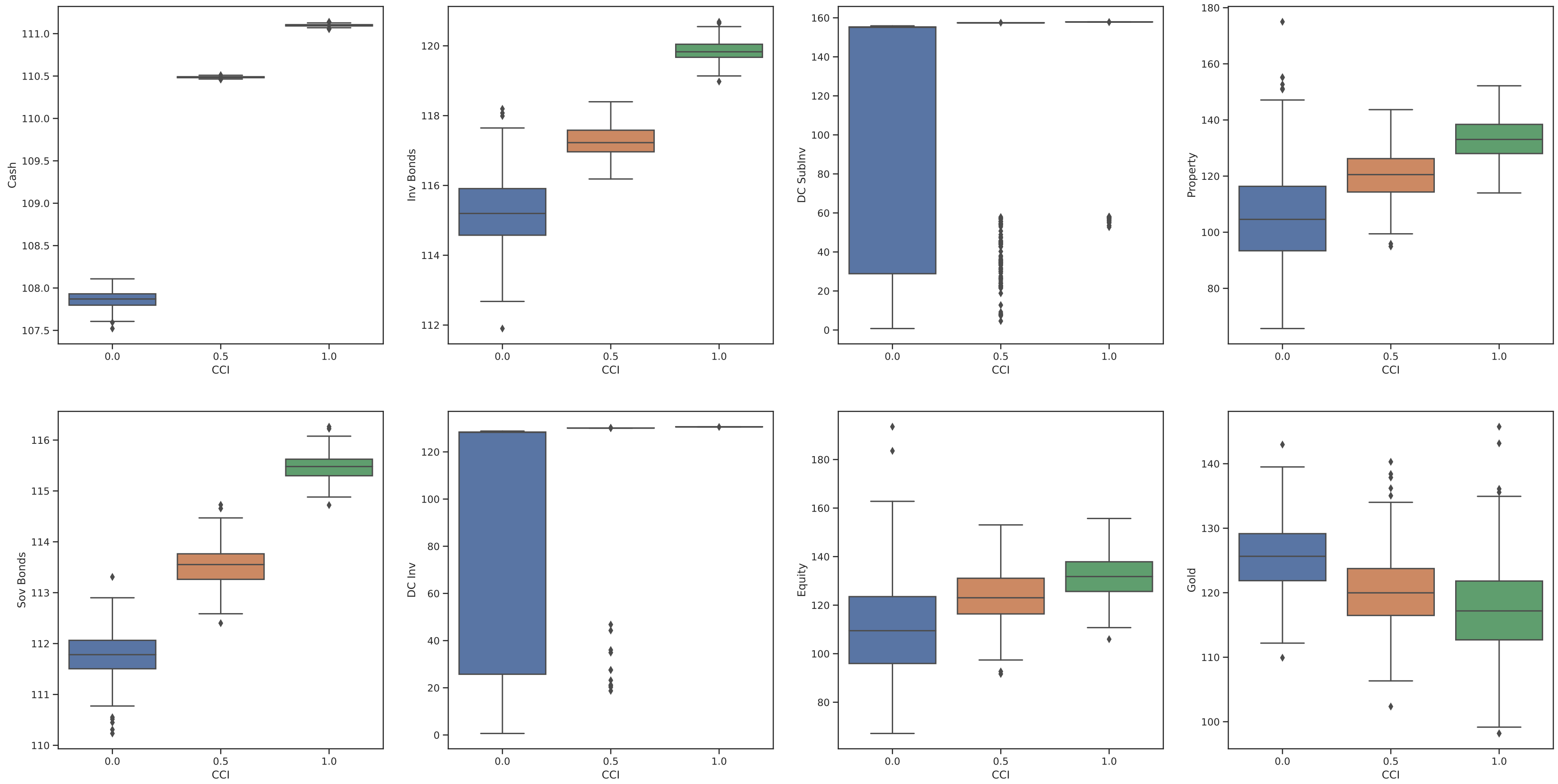


Figure 139: USA Asset class CAGR distributions from one year of investing, split by CCI category.

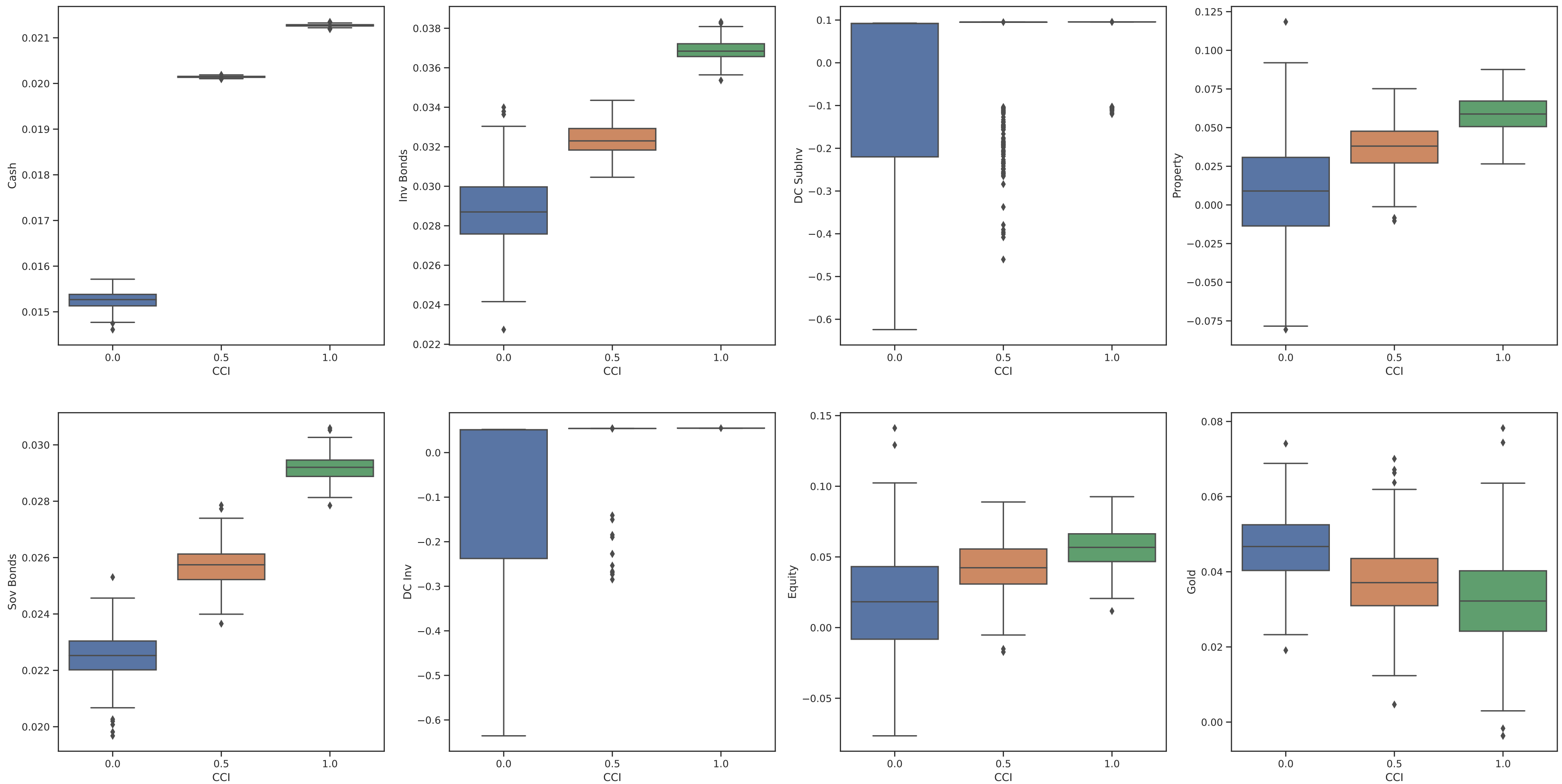


Figure 140: USA Asset class CAGR distributions from five years of investing, split by CCI category.

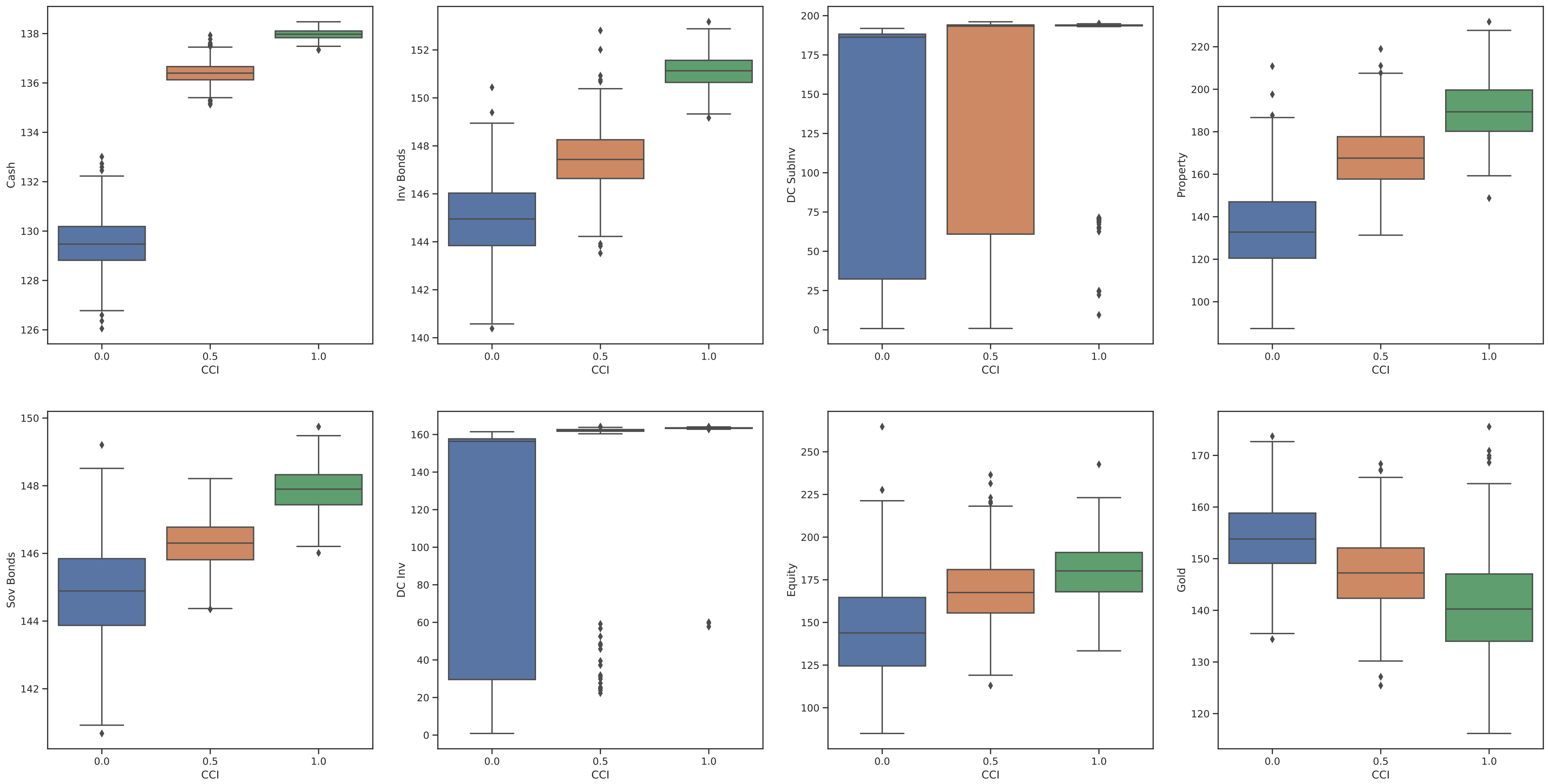


Figure 141: RSA Asset class CAGR distributions from one year of investing, split by CCI category.

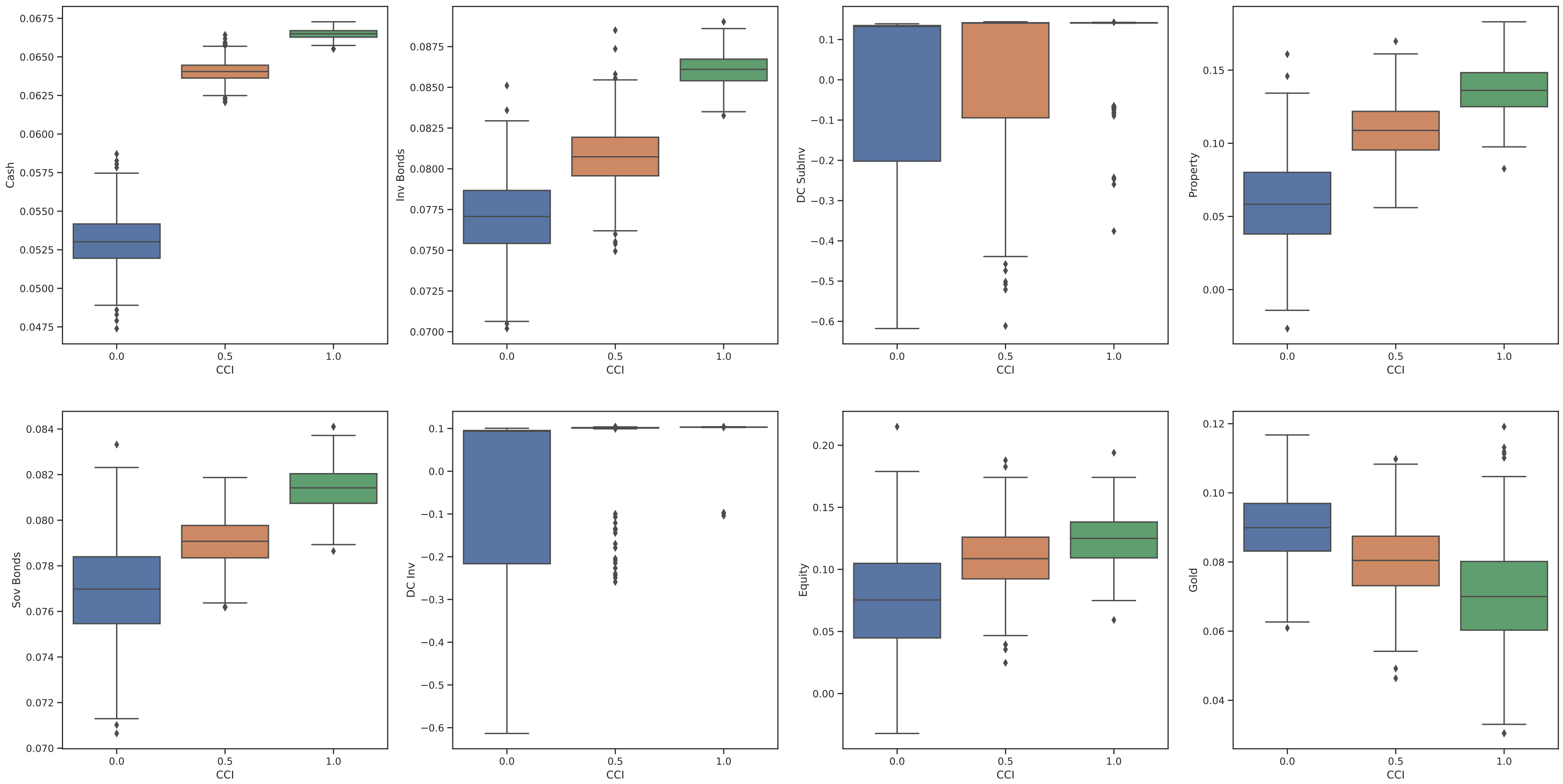


Figure 142: RSA Asset class CAGR distributions from five years of investing, split by CCI category.

28 Building block portfolio returns

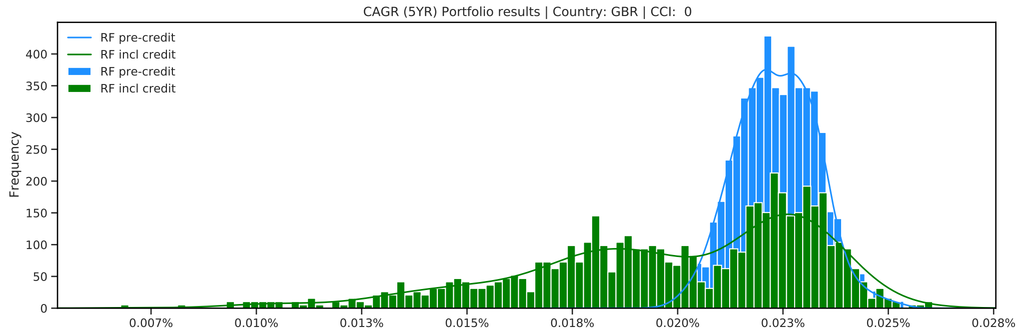


Figure 143: CAGR (1YR) CCI=0%: UK.

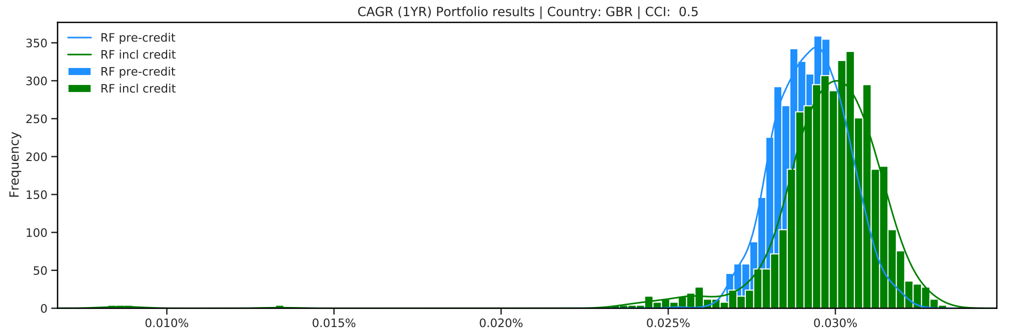


Figure 144: CAGR (1YR) CCI=50%: UK.

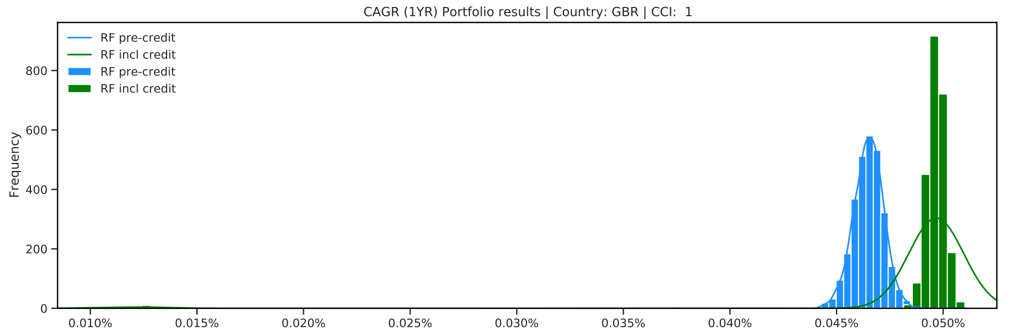


Figure 145: CAGR (1YR) CCI=100%: UK.

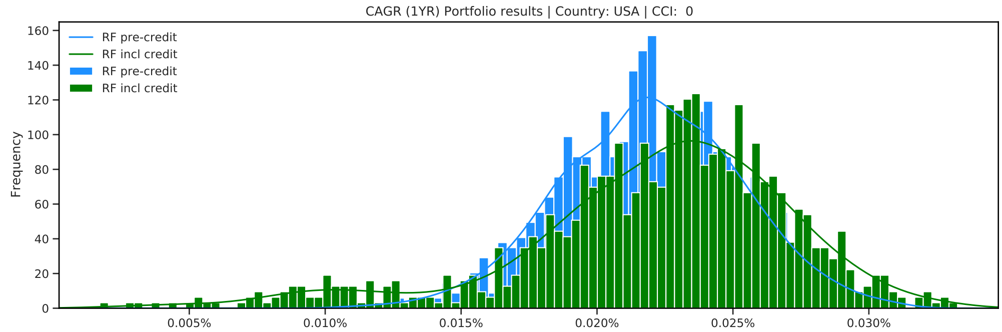


Figure 146: CAGR (1YR) CCI=0%: USA.

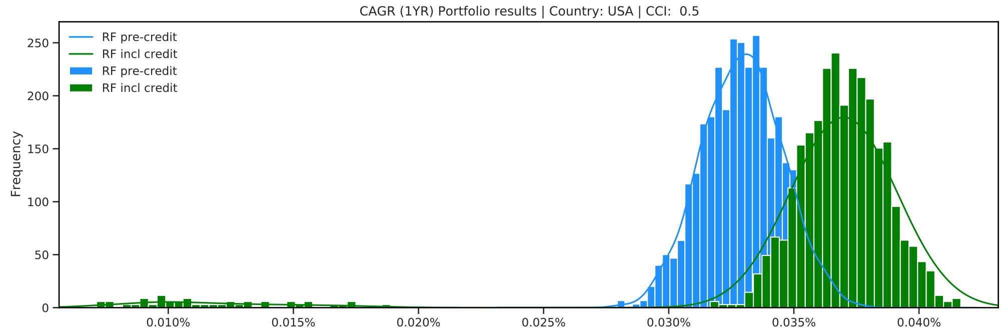


Figure 147: CAGR (1YR) CCI=50%: USA.

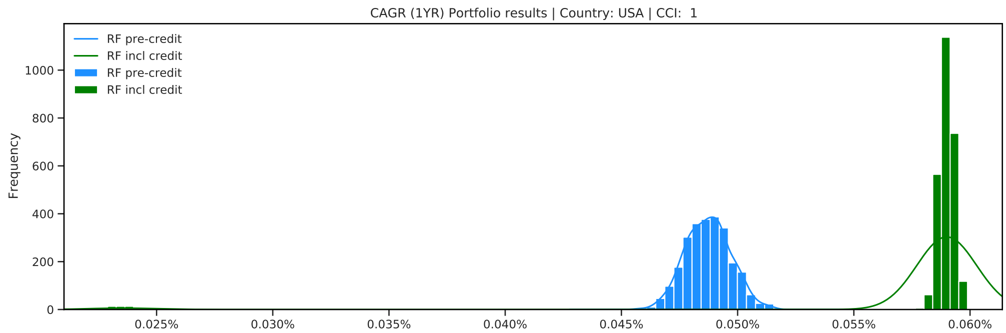


Figure 148: CAGR (1YR) CCI=100%: USA.

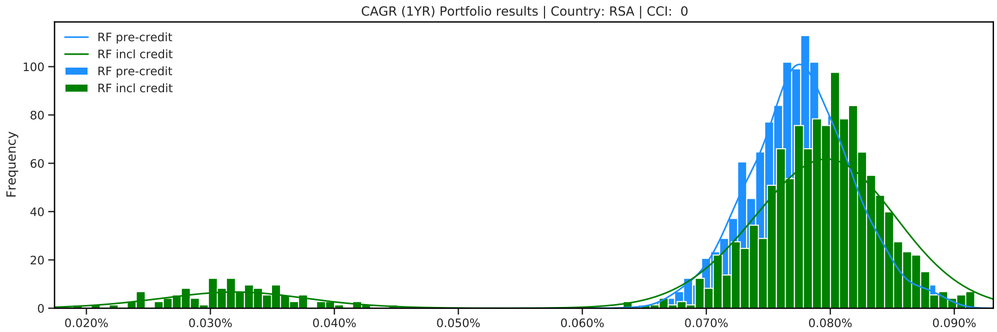


Figure 149: CAGR (1YR) CCI=0%: RSA.

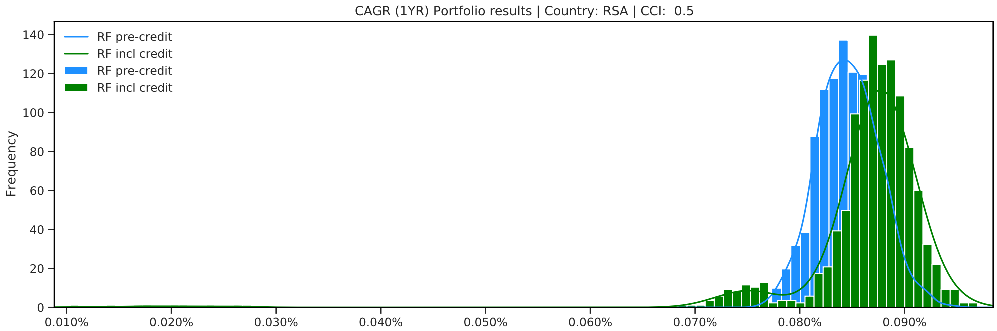


Figure 150: CAGR (1YR) CCI=50%: RSA.

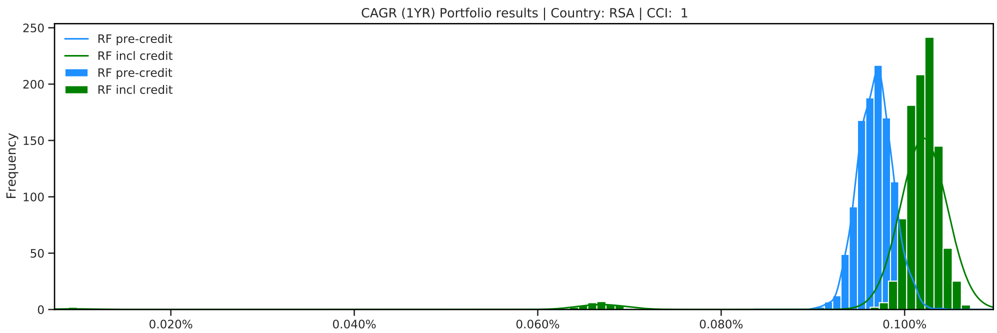


Figure 151: CAGR (1YR) CCI=100%: RSA.

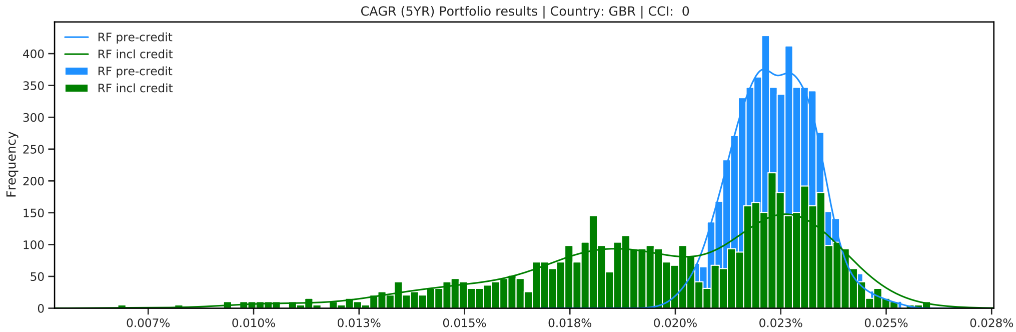


Figure 152: CAGR (5YR) CCI=0%: UK.

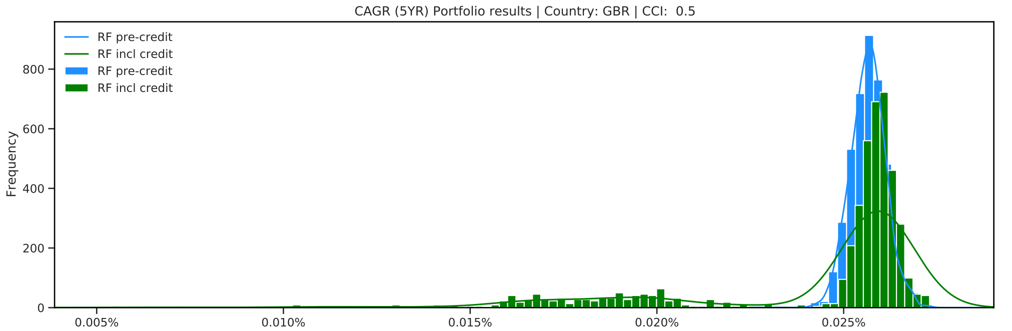


Figure 153: CAGR (5YR) CCI=50%: UK.

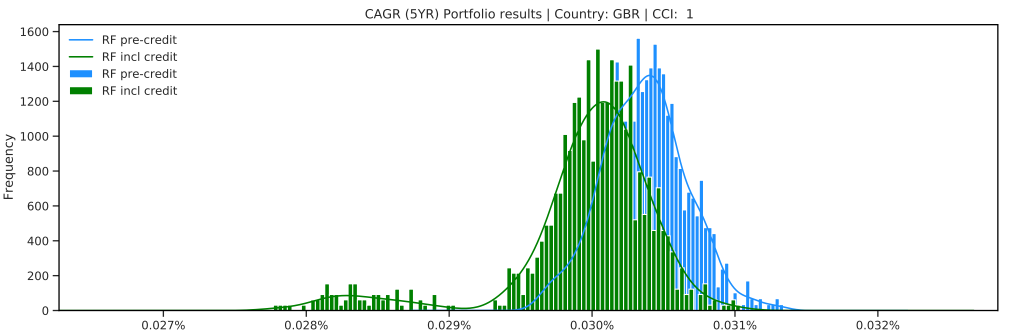


Figure 154: CAGR (5YR) CCI=100%: UK.

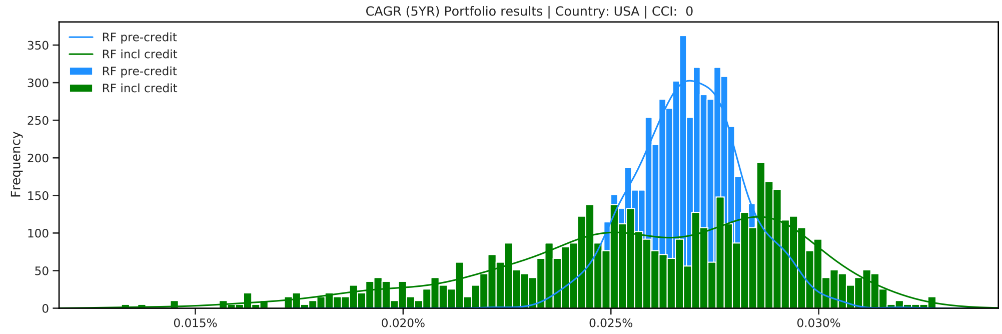


Figure 155: CAGR (5YR) CCI=0%: US.

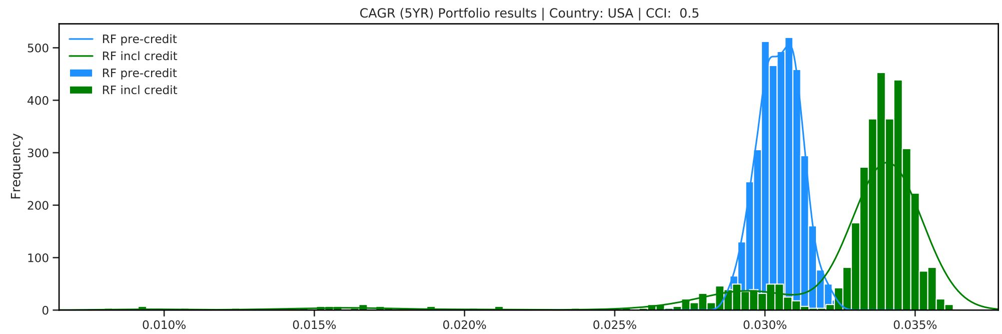


Figure 156: CAGR (5YR) CCI=50%: US.

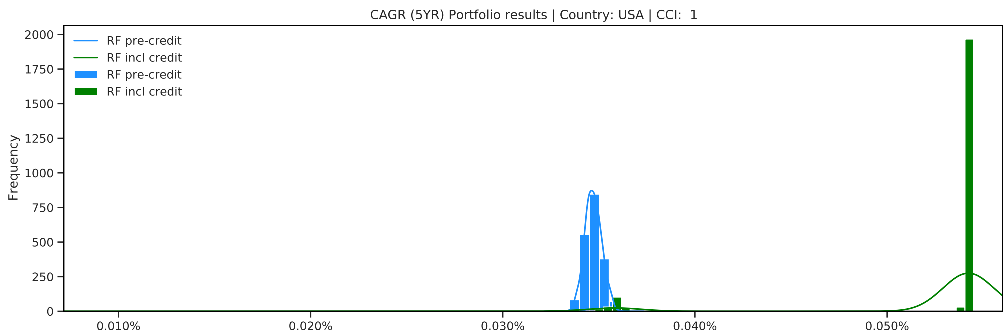


Figure 157: CAGR (5YR) CCI=100%: US.

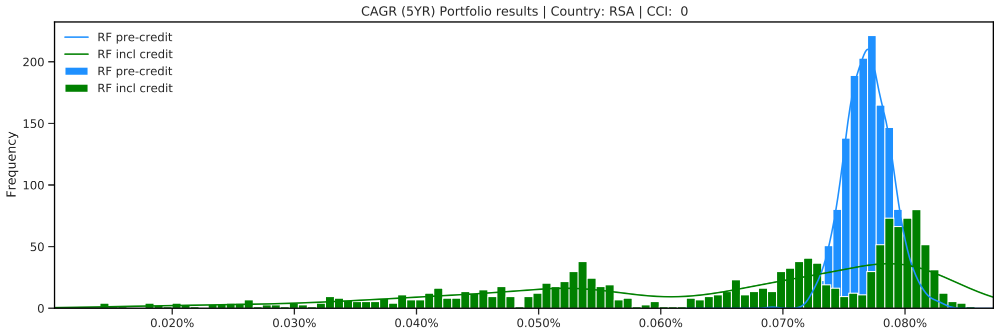


Figure 158: CAGR (5YR) CCI=0%: RSA.

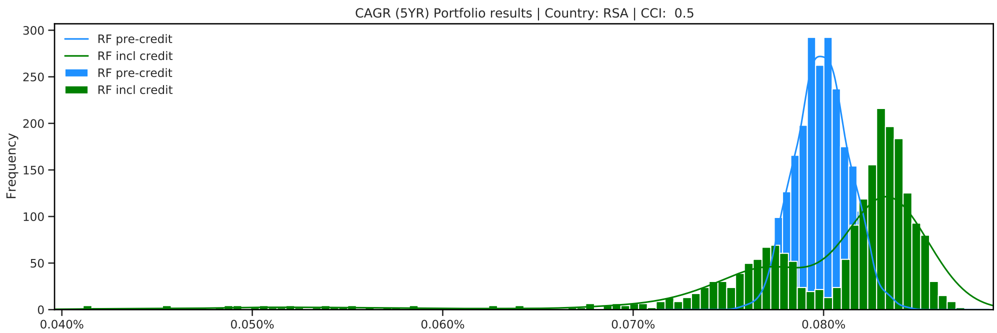


Figure 159: CAGR (5YR) CCI=50%: RSA.

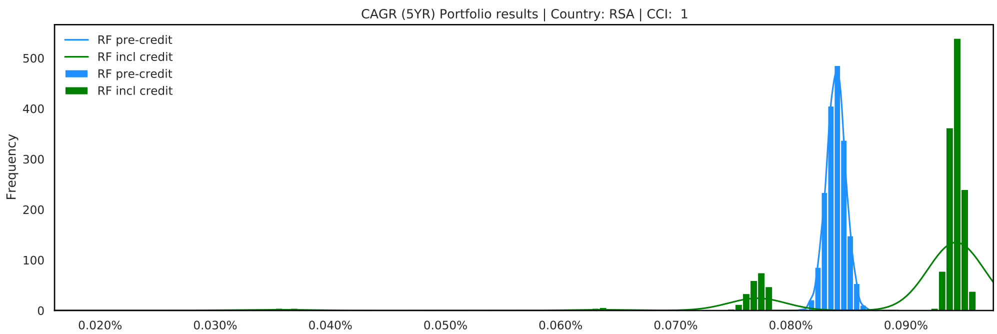


Figure 160: CAGR (5YR) CCI=100%: RSA.

29 Liability parameters

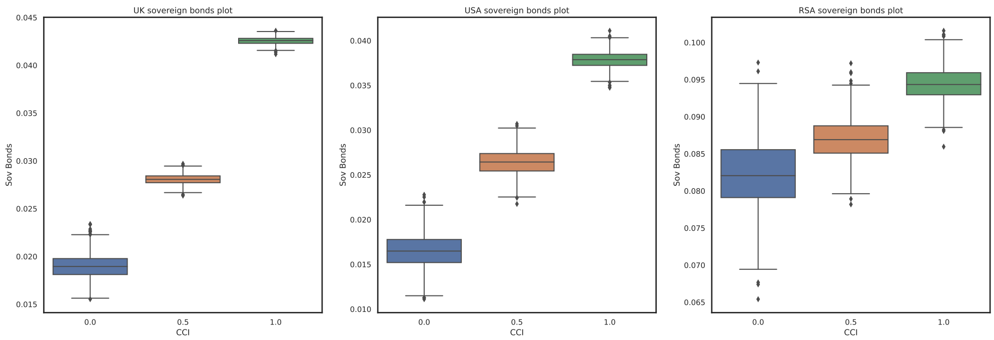


Figure 161: Box plots featuring the distribution of the one year expected rates for Sovereign bonds in each region and boom, bust or normal markets.

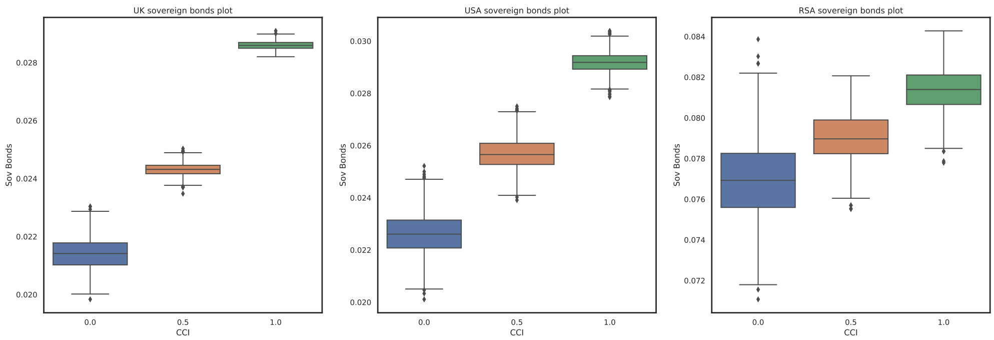


Figure 162: Box plots featuring the distribution of the five year expected rates for Sovereign bonds in each region and boom, bust or normal markets.

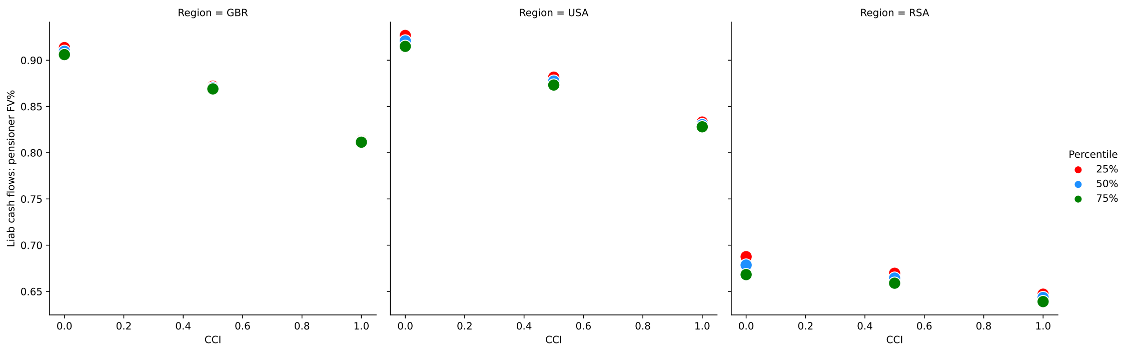


Figure 163: Once the liability cash flows are discounted, using the different scenarios reporting in table 26, the effects to liability values is clear. When interest rate are expected to be higher (CCI=0%), the liability values are lower and visa versa. The chart indicates the proportion of FV the PV represents.

30 Software and Python packages

30.1 Colab and Python code

Please refer to the following Github (https://github.com/sroydenturner/50797891-PhD) for the coded sets. Password on request (50707981@mylife.unisa.ac.za).

30.2 Python packages

```
from datetime import date, timedelta datetime
from sklearn.preprocessing import StandardScaler
from gdeltdoc import GdeltDoc, Filters, near, repeat
from gensim.corpora.dictionary import Dictionary
from gensim.models import CoherenceModel
from gensim.models.coherencemodel
from gensim.models.ldamodel import LdaModel
from gensim.utils import lemmatize
from gensim.utils import simple_preprocess
from sklearn.preprocessing import LabelEncoder
from keras.layers import Dense
from keras.layers import Flatten
from keras.layers import GlobalMaxPooling1D
from keras.layers import LSTM
from keras.layers.convolutional import Conv1D
from keras.layers.convolutional import MaxPooling1D
from keras.layers.core import Activation, Dropout
from keras.layers.embeddings import Embedding
from keras.models import Sequential
from keras.preprocessing.text import one_hot
from keras.preprocessing.text import Tokenizer
from keras_preprocessing.sequence
from math import sqrt
from matplotlib import pyplot
from nltk.corpus import stopwords
nltk.download('punkt')
nltk.download('vader_lexicon')
from numpy import array
from numpy import concatenate
from numpy import zeros
from os import makedirs
from os.path import join, exists
from pandas import concat
import riskfolio.Portfolio as pf
```

```
from sklearn.preprocessing import MinMaxScaler
from dateutil.relativedelta import relativedelta
from time import gmtime, strftime, strptime
from timeit import default_timer as timer
from nltk.sentiment.vader SentimentIntensityAnalyzer
import boto3
import calendar
import csv
import matplotlib.pyplot as plt
import matplotlib.ticker as mtick
import datetime as dt
import datetime as dt
import datetime as dt
import dateutil.parser
import dateutil.parser as parser
import gensim
import gensim.corpora as corpora
import json
import logging
import math
from textblob import TextBlob
import pad_sequences
import nltk
import nltk; nltk.download('stopwords')
import numpy as np
import pandas as pd
import pandas.util.testing as tm
import quandl
import re
import requests
import scipy.interpolate as interpolate
import seaborn as sns
import sidetable
import spacy # lemmatization
```

from pandas import read_csv
from scipy.stats import beta
from scipy.stats import binned_statistic
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import mean_squared_error
from sklearn.metrics import roc_curve, auc
from sklearn.model_selection import train_test_split
from sklearn.model_selection import train_test_split

import statsmodels.api as sm
import statsmodels.formula.api as smf
import tensorflow.keras.layers as layers
import tensorflow.keras.models as models
import time
import timeit
import warnings
import waterfall_chart