UNIVERSITY OF SOUTHAMPTON

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES

School of Management

Credit Risk Models for Mortgage Loan Loss Given Default

by

Mindy Leow

Thesis for the degree of Doctor of Philosophy

August 2010

To my lgz.

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES
SCHOOL OF MANAGEMENT

Doctor of Philosophy

CREDIT RISK MODELS FOR MORTGAGE LOAN LOSS GIVEN DEFAULT

by Mindy Leow

Arguably, the credit risk models reported in the literature for the retail lending sector have so far been less developed than those for the corporate sector, mainly due to the lack of publicly available data. Having been given access to a dataset on defaulted mortgages kindly provided by a major UK bank, this work first investigates the Loss Given Default (LGD) of mortgage loans with the development of two separate component models, the Probability of Repossession (given default) Model and the Haircut (given repossession) Model. They are then combined into an expected loss percentage. Performance-wise, this two-stage LGD model is shown to do better than a single-stage LGD model (which directly models LGD from loan and collateral characteristics), as it achieves a better R-square value, and it more accurately matches the distribution of observed LGD. We next investigate the possibility of including macroeconomic variables into either or both component models to improve LGD prediction. Indicators relating to net lending, gross domestic product, national default rates and interest rates are considered and the interest rate is found to be most beneficial to both component models. Finally, we develop a competing risk survival analysis model to predict the time taken for a defaulted mortgage loan to reach some outcome (i.e. repossession or non-repossession). This allows for a more accurate prediction of (discounted) loss as these periods could vary from months to years depending on the health of the economy. Besides loan- or collateral-related characteristics, we incorporate a time-dependent macroeconomic variable based on the house price index (HPI) to investigate its impact on repossession risk. We find that observations of different loan-to-value ratios at default and different security type are affected differently by the economy. This model is then used for stress test purposes by applying a Monte Carlo simulation, and by varying the HPI forecast, to get different loss distributions for different economic outlooks.

TABLE OF CONTENTS

LIST OF TABLES VIII

LIST OF FIGURES X

DECLARATION OF AUTHORSHIP XIII

ACKNOWLEDGEMENTS XIV

LIST OF ABBREVIATIONS XV

CHAPTER 1. INTRODUCTION 1

1.1 BASEL COMMITTEE AND BASEL II REGULATIONS 2
1.2 MORTGAGE LOANS IN THE UK 6
1.3 OVERVIEW OF THE THREE RESEARCH PROJECTS 9

CHAPTER 2. DATA 13

2.1 TIME PERIODS AND GEOGRAPHY 13
2.2 MULTIPLE DEFAULTS 14
2.3 DATE OF DEFAULT 15
2.4 TIME ON BOOKS 16
2.5 REPOSSESSION 17
2.6 VALUATION OF SECURITY AT DEFAULT AND HAIRCUT 18
2.7 TYPE OF SECURITY 19
2.8 TRAINING AND TEST SET SPLITS 19
2.9 LOSS GIVEN DEFAULT 19

CHAPTER 3. PREDICTING LOSS GIVEN DEFAULT (LGD) FOR RESIDENTIAL MORTGAGE LOANS: A TWO-STAGE MODEL 22

3.1 LITERATURE REVIEW 22
3.1.1 Risk models for residential mortgage lending 22
3.1.2 Single vs. two-stage LGD models 23
3.2 RESEARCH OBJECTIVES 24

3.3 THE PROBABILITY OF REPOSSESSION MODEL 25
3.3.1 Modelling methodology 25
3.3.2 Model variations 26
3.3.3 Performance measures 26
3.3.4 Model results 27
3.4 THE HAIRCUT MODEL 28
3.4.1 Modelling methodology 29
3.4.2 Model variations 30
3.4.3 Performance measures 31
3.4.4 Model results 31
3.4.5 Haircut standard deviation modelling 34
3.5 LOSS GIVEN DEFAULT MODEL 36
3.5.1 Modelling methodology 36
3.5.2 Alternative single-stage model 39
3.5.3 Model performance 39
3.6 CONCLUSIONS 43

CHAPTER 4. THE ECONOMY AND RETAIL LGD 44

4.1 LITERATURE REVIEW 44
4.1.1 Overview of corporate credit risk models 44
4.1.2 The economy, PD and LGD 45
4.1.3 The retail sector: LGD and the economy 47
4.2 RESEARCH OBJECTIVES 48
4.3 MACROECONOMIC VARIABLES 50
4.4 MODELLING METHODOLOGY 55
4.4.1 Specific details for mortgage loans model 56
4.4.2 Specific details for personal loans model 56
4.5 MORTGAGE LOAN LGD RESULTS 57
4.5.1 Probability of Repossession Model 57
4.5.2 Haircut Model 59
4.5.3 Two-stage LGD model 62
4.6 UNSECURED PERSONAL LOANS LGD MODEL RESULTS 67
4.7 CONCLUSIONS 71

CHAPTER 5. COMPETING RISKS SURVIVAL MODEL WITH SIMULATED LOSS DISTRIBUTIONS 73

5.1 LITERATURE REVIEW 74
5.1.1 Survival analysis in retail credit models 74
5.1.2 Monte Carlo and stress testing 76
5.2 RESEARCH OBJECTIVES 78
5.3 DATA PREPARATION 80
5.3.1 Data 80
5.3.2 Defining events 81
5.3.3 Variable selection and pre-processing 81
5.4 COMPETING RISKS SURVIVAL MODEL 83
5.4.1 Survival models 83
5.4.2 Competing risks 85
5.4.3 Survival model for repossession 85
5.4.4 Survival model for closure 89
5.5 MONTE CARLO SIMULATION 91
5.5.1 Framework 92
5.5.2 Stress testing 95
5.5.3 Simulation results 98
5.5.3.1 Distribution of total loss 98
5.5.3.2 Number of repossessions 102
5.5.3.3 Distribution of LGD 104
5.6 CONCLUSIONS 110

CHAPTER 6. CONCLUSIONS AND FURTHER RESEARCH 114

6.1 REVIEW OF CHAPTERS AND MAJOR CONCLUSIONS 114
6.2 ISSUES FOR FURTHER RESEARCH 118

LIST OF REFERENCES 120

APPENDIX A. PARAMETER ESTIMATES FOR TWO-STAGE AND SINGLE-STAGE MORTGAGE LGD MODEL 126

APPENDIX B. PLOTS OF MACROECONOMIC VARIABLES 134

APPENDIX C. PARAMETER ESTIMATES FROM MACROECONOMIC MODELS 139

APPENDIX D: RESULTS OF LGD MACROECONOMIC MODELS WITH TIME LAGS AND LEADS.....143

APPENDIX E. TABLE OF CONTRIBUTIONS TO MARGINAL RISK FOR REPOSSESSION AND CLOSURE.....149

APPENDIX F. PARAMETER ESTIMATES FOR SURVIVAL MODELS 151

APPENDIX G. GRAPHS OF HOUSE PRICE GROWTH FOR REGIONS IN THE UK.....153

LIST OF TABLES

TABLE 1.1: RISK COMPONENTS TO BE DEVELOPED ACCORDING TO BASEL II APPROACHES 3
TABLE 3.1: REPOSSESSION MODELS PERFORMANCE STATISTICS 27
TABLE 3.2: PARAMETER ESTIMATE SIGNS FOR PROBABILITY OF REPOSSESSION MODEL *R2* 28
TABLE 3.3: HAIRCUT MODEL PERFORMANCE STATISTICS 32
TABLE 3.4: PARAMETER ESTIMATE SIGNS OF HAIRCUT MODEL *H1* 33
TABLE 3.5: HAIRCUT STANDARD DEVIATION MODEL PERFORMANCE STATISTICS 35
TABLE 3.6: PERFORMANCE MEASURES OF TWO-STAGE AND SINGLE-STAGE LGD MODELS 40
TABLE 4.1: MACROECONOMIC VARIABLES CONSIDERED IN THE ANALYSIS 53
TABLE 4.2: PERFORMANCE OF PROBABILITY OF REPOSSESSION MODEL (TEST SETS) WITH MACROECONOMIC VARIABLES 58
TABLE 4.3: PERFORMANCE OF HAIRCUT MODEL (TEST SETS) WITH MACROECONOMIC VARIABLES 60
TABLE 4.4: PERFORMANCE STATISTICS OF MORTGAGE LOAN LGD MODELS (TEST SETS) 62
TABLE 4.5: RESULTS OF PERSONAL LOANS LGD BASE AND MACROECONOMIC MODELS (TEST SETS) 68
TABLE 5.1: MIGRATION MATRIX OF DLTV AND STRESSED DLTV (IN PERCENTAGES) (FOR TEST SET) 97
TABLE 5.2: MEAN LGD VALUES FOR VALIDATION AND STRESSED SIMULATION, COMPARED AGAINST OBSERVED LGD, SEGMENTED BY TYPE OF SECURITY, FOR COHORT OF 1991 107
TABLE 5.3: MEAN LGD VALUES FOR VALIDATION AND STRESSED SIMULATION, COMPARED AGAINST OBSERVED LGD, SEGMENTED BY DLTV BANDS, FOR COHORT OF 1991 109
TABLE A1: PARAMETER ESTIMATES FOR PROBABILITY OF REPOSSESSION MODEL *R0* . 126
TABLE A2: PARAMETER ESTIMATES FOR PROBABILITY OF REPOSSESSION MODEL *R1* . 126
TABLE A3: PARAMETER ESTIMATES FOR PROBABILITY OF REPOSSESSION MODEL *R2* . 127
TABLE A4: PARAMETER ESTIMATES FOR HAIRCUT MODEL *H1* 127
TABLE A5: PARAMETER ESTIMATES FOR HAIRCUT MODEL *H2* 128

TABLE A6: PARAMETER ESTIMATES FOR HAIRCUT STANDARD DEVIATION MODEL ... 130
TABLE A7: PARAMETER ESTIMATES FOR SINGLE-STAGE LGD MODEL 130
TABLE A8: PARAMETER ESTIMATES FOR PROBABILITY OF REPOSSESSION MODEL FROM ROBUSTNESS TEST (WHERE ONLY FIRST INSTANCE OF DEFAULT WAS INCLUDED IN MODELLING) 131
TABLE A9: PARAMETER ESTIMATES FOR HAIRCUT MODEL FROM ROBUSTNESS TEST (WHERE ONLY FIRST INSTANCE OF DEFAULT WAS INCLUDED IN MODELLING) 132
TABLE C1: PARAMETER ESTIMATES AND P-VALUES OF PROBABILITY OF REPOSSESSION MACROECONOMIC MODEL 139
TABLE C2: PARAMETER ESTIMATES, P-VALUES AND VIFS OF HAIRCUT MACROECONOMIC MODEL 139
TABLE C3: PARAMETER ESTIMATES, P-VALUES AND VIFS OF PERSONAL LOANS MACROECONOMIC LGD MODEL 141
TABLE D1: PERFORMANCE OF PROBABILITY OF REPOSSESSION MODEL (TEST SETS) WITH MACROECONOMIC VARIABLES WITH SIX-MONTH LAG 143
TABLE D2: PERFORMANCE OF PROBABILITY OF REPOSSESSION MODEL (TEST SETS) WITH MACROECONOMIC VARIABLES AT SIX-MONTH LEAD 144
TABLE D3: PERFORMANCE OF HAIRCUT MODEL (TEST SETS) WITH MACROECONOMIC VARIABLES WITH SIX-MONTH LAG 145
TABLE D4: PERFORMANCE OF HAIRCUT MODEL (TEST SETS) WITH MACROECONOMIC VARIABLES AT SIX-MONTH LEAD 146
TABLE D5: PERFORMANCE OF PERSONAL LOAN LGD MODEL (TEST SETS) WITH MACROECONOMIC VARIABLES WITH SIX-MONTH LAG 147
TABLE D6: PERFORMANCE OF PERSONAL LOAN LGD MODEL (TEST SETS) WITH MACROECONOMIC VARIABLES WITH SIX-MONTH LEAD 148
TABLE E1: CONTRIBUTIONS TO CALCULATION OF MARGINAL RISK FOR REPOSSESSION AND/OR CLOSURE 149
TABLE F1: PARAMETER ESTIMATES FOR REPOSSESSION SURVIVAL MODEL 151
TABLE F2: PARAMETER ESTIMATES FOR CLOSURE SURVIVAL MODEL 152

LIST OF FIGURES

FIGURE 1.1: TOTAL CONSUMER CREDIT: UK NET LENDING, IN STERLING MILLIONS 6
FIGURE 1.2: BANKS UK NET LENDING SECURED ON DWELLINGS (RESIDENTIAL PROPERTIES), IN STERLING MILLIONS 7
FIGURE 1.3: HALIFAX HOUSE PRICE INDEX FOR THE UK 8
FIGURE 2.1: ESTIMATION OF DEFAULT DATE AND CALCULATION OF VARIABLE TIME ON BOOK 15
FIGURE 2.2: MEAN TIME ON BOOK OVER TIME WITH REFERENCE TO YEAR OF DEFAULT 16
FIGURE 2.3: DEFAULTS OVER TIME WITH REFERENCE TO LENGTH OF TIME TO REPOSSESSION 17
FIGURE 3.1: DISTRIBUTION OF HAIRCUT. SOLID CURVE REFERENCES THE NORMAL DISTRIBUTION 29
FIGURE 3.2: RELATIONSHIP BETWEEN HAIRCUT AND (RANKED) VALUATION OF SECURITY AT DEFAULT TO AVERAGE PROPERTY VALUATION IN THE REGION RATIO 30
FIGURE 3.3: RELATIONSHIP BETWEEN HAIRCUT AND (RANKED) LTV AT TIME OF LOAN APPLICATION 32
FIGURE 3.4: PREDICTION PERFORMANCE OF HAIRCUT MODEL (FOR TEST SET) 34
FIGURE 3.5: MEAN HAIRCUT STANDARD DEVIATION BY TIME ON BOOK BINS 35
FIGURE 3.6: DISTRIBUTION OF OBSERVED LGD (EMPIRICAL), PREDICTED LGD FROM TWO-STAGE HAIRCUT POINT ESTIMATE MODEL (HC PT. EST), TWO-STAGE EXPECTED SHORTFALL MODEL (E.SHORTFALL), SINGLE-STAGE MODEL (SINGLE STAGE) 40
FIGURE 3.7: SCATTERPLOT OF PREDICTED AND ACTUAL LGD IN LGD BANDS 42
FIGURE 4.1: OBSERVED BANK REPOSSESSION RATE AND MEAN LGD ACROSS DEFAULT YEARS FOR MORTGAGE LOANS DATASET 49
FIGURE 4.2: MEAN LGD OVER DEFAULT YEARS FOR PERSONAL LOANS DATASET 50
FIGURE 4.3: MEAN OBSERVED AND PREDICTED MORTGAGE LOAN LGD OVER EACH YEAR OF DEFAULT FOR BASE AND MACROECONOMIC LGD MODELS (TEST SETS) 63
FIGURE 4.4: MEAN DIFFERENCE BETWEEN OBSERVED AND PREDICTED MORTGAGE LOAN LGD, FOR BASE AND MACROECONOMIC LGD MODELS (TEST SETS) 64
FIGURE 4.5: DISTRIBUTION OF OBSERVED LGD (EMPIRICAL), PREDICTED LGD FROM TWO-STAGE EXPECTED SHORTFALL BASE MODEL (E.SHORTFALL), TWO-STAGE EXPECTED SHORTFALL MACROECONOMIC MODEL (MACRO) 65

FIGURE 4.6: SCATTERPLOT OF (BASE AND MACROECONOMIC) PREDICTED AND ACTUAL LGD IN LGD BANDS 66
FIGURE 4.7: MEAN OBSERVED AND PREDICTED PERSONAL LOAN LGD OVER EACH YEAR OF DEFAULT FOR BASE AND MACROECONOMIC MODELS (TEST SETS) 69
FIGURE 4.8: MEAN DIFFERENCE BETWEEN OBSERVED AND PREDICTED PERSONAL LOAN LGD, FOR BASE AND MACROECONOMIC LGD MODELS (TEST SETS) 70
FIGURE 5.1: REPOSSESSION MARGINAL RISK CONTRIBUTIONS FOR DIFFERENT TYPES OF SECURITY ACCORDING TO DLTV BANDS 87
FIGURE 5.2: BASELINE HAZARD RATES FOR REPOSSESSION SURVIVAL MODEL 89
FIGURE 5.3: CLOSURE MARGINAL RISK CONTRIBUTIONS FOR DIFFERENT TYPES OF SECURITY ACCORDING TO DLTV BANDS 90
FIGURE 5.4: BASELINE HAZARD RATES FOR CLOSURE SURVIVAL MODEL 91
FIGURE 5.5: OBSERVED AND STRESSED HPIG FOR REGION OF GREATER LONDON 96
FIGURE 5.6: DISTRIBUTION OF DLTV AND STRESSED DLTV (FOR TEST SET) 97
FIGURE 5.7: DISTRIBUTION OF PREDICTED TOTAL LOSS ACROSS 1,000 SIMULATION RUNS FOR VALIDATION SIMULATION FOR COHORT OF LOANS THAT DEFAULT IN 1995 99
FIGURE 5.8: DISTRIBUTION OF PREDICTED TOTAL LOSS ACROSS 1,000 SIMULATION RUNS FOR VALIDATION SIMULATION FOR COHORT OF LOANS THAT DEFAULT IN 1991 100
FIGURE 5.9: COMPARATIVE DISTRIBUTION OF PREDICTED TOTAL LOSS ACROSS 1,000 SIMULATION RUNS FOR COHORT OF LOANS THAT DEFAULT IN 1995; FOR VALIDATION (TOP PANEL) AND STRESSED (BOTTOM PANEL) SIMULATION 101
FIGURE 5.10: COMPARATIVE DISTRIBUTION OF PREDICTED TOTAL LOSS ACROSS 1,000 SIMULATION RUNS FOR COHORT OF LOANS THAT DEFAULT IN 1991; FOR VALIDATION (TOP PANEL) AND STRESSED (BOTTOM PANEL) SIMULATION 102
FIGURE 5.11: NUMBER OF OBSERVED REPOSSESSIONS (SOLID) IN THE MONTHS AFTER DEFAULT AND AVERAGE (OVER 1,000 RUNS) PREDICTED NUMBER OF REPOSSESSIONS FOR VALIDATION (LINE) AND STRESSED (CIRCLES) SIMULATIONS, FOR 1995 COHORT 103
FIGURE 5.12: NUMBER OF OBSERVED REPOSSESSIONS (SOLID) IN THE MONTHS AFTER DEFAULT AND AVERAGE (OVER 1,000 RUNS) PREDICTED NUMBER OF REPOSSESSIONS FOR VALIDATION (LINE) AND STRESSED (CIRCLES) SIMULATIONS, FOR 1991 COHORT 104

FIGURE 5.13: DISTRIBUTION OF MEDIAN PREDICTED LGD FOR VALIDATION (LEFT PANEL) AND STRESSED SIMULATIONS (RIGHT PANEL) FOR 1991 COHORT, SEGMENTED BY TYPE OF SECURITY 105
FIGURE 5.14: DISTRIBUTION OF 95TH PERCENTILE PREDICTED LGD FOR VALIDATION (LEFT PANEL) AND STRESSED SIMULATIONS (RIGHT PANEL) FOR 1991 COHORT, SEGMENTED BY TYPE OF SECURITY 106
FIGURE 5.15: DISTRIBUTION OF MEDIAN PREDICTED LGD FOR VALIDATION (LEFT PANEL) AND STRESSED SIMULATIONS (RIGHT PANEL) FOR 1991 COHORT, SEGMENTED BY DLTV BANDS 108
FIGURE 5.16: DISTRIBUTION OF 95TH PERCENTILE PREDICTED LGD FOR VALIDATION (LEFT PANEL) AND STRESSED SIMULATIONS (RIGHT PANEL) FOR 1991 COHORT, SEGMENTED BY DLTV BANDS 110
FIGURE B1: NET LENDING CHANGE; ONS; 1983 – 2003 134
FIGURE B2: DISPOSABLE INCOME GROWTH; ONS; 1983 – 2003 134
FIGURE B3: GDP GROWTH; ONS; 1983 – 2003 135
FIGURE B4: PURCHASING POWER OF POUND GROWTH; ONS; 1983 – 2003 135
FIGURE B5: UNEMPLOYMENT RATE; ONS; 1983 – 2003 136
FIGURE B6: SAVING RATIO; ONS; 1983 – 2003 136
FIGURE B7: INTEREST RATES; BOE; 1983 – 2003 137
FIGURE B8: NET LENDING GROWTH FOR DWELLINGS; ONS; 1983 – 2003 137
FIGURE B9: HPI YEAR ON YEAR QUARTERLY GROWTH; HALIFAX; 1983 – 2003; DIFFERENTIATED BY REGION 138
FIGURE B10: HPI YEAR ON YEAR MONTHLY GROWTH; HALIFAX; 1983 – 200 138
FIGURE G1: OBSERVED AND STRESSED HPIG FOR UK (NORTH) 153
FIGURE G2: OBSERVED AND STRESSED HPIG FOR UK (YORKSHIRE AND HUMBLESIDE) 153
FIGURE G3: OBSERVED AND STRESSED HPIG FOR UK (NORTH WEST) 154
FIGURE G4: OBSERVED AND STRESSED HPIG FOR UK (EAST MIDLANDS) 154
FIGURE G5: OBSERVED AND STRESSED HPIG FOR UK (WEST MIDLANDS) 155
FIGURE G6: OBSERVED AND STRESSED HPIG FOR UK (EAST ANGLIA) 155
FIGURE G7: OBSERVED AND STRESSED HPIG FOR UK (WALES) 156
FIGURE G8: OBSERVED AND STRESSED HPIG FOR UK (SOUTH WEST) 156
FIGURE G9: OBSERVED AND STRESSED HPIG FOR UK (SOUTH EAST) 157
FIGURE G10: OBSERVED AND STRESSED HPIG FOR UK (NORTHERN IRELAND) 157
FIGURE G11: OBSERVED AND STRESSED HPIG FOR UK (SCOTLAND) 158

DECLARATION OF AUTHORSHIP

I, Mindy Leow

declare that the thesis entitled

Credit Risk Models for Mortgage Loan Loss Given Default

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work are under review as: Leow and Mues (2010), Predicting Loss Given Default for Residential Mortgage Loans: A Two-Stage Model and Empirical Evidence for UK Bank Data, with the International Journal of Forecasting

Signed: ……..

Date:……

ACKNOWLEDGEMENTS

This thesis would not be possible without the people around me, my heartfelt thanks to everyone who has helped me in one way or another throughout my time at the University of Southampton, UK.

Special thanks go to my supervisors: Dr Christophe Mues, for his invaluable guidance, his patience and his encouragement which pulled me through my (many) panic attacks; and Prof Lyn Thomas, for his kindness, his expert knowledge and time. I am also grateful to the University of Southampton and the ORSAS for giving me the financial support and hence the opportunity to pursue my postgraduate studies; as well as the UK bank(s) who provided the data, making this work possible.

I could not have survived this without friends, my family away from home: to friends from CCF – there are too many of you to name! – thank you for the makan sessions, the mostly impromptu but crazy mj & wii sessions and all that laughter, as well as prayers and reminders to lose neither faith nor hope; to LL & YF, our food adventures, random midnight conversations and sleepovers will be memories I shall always hold dear; to my fellow research students, those who have "seen the light" and those who are still working towards it – Jacq, Kurt, Meko, Pundita, Iain, Ed, Grant, Wendy, Eric, Katie, Sara – thank you for your stimulating discussions and listening to my thoughts, questions and especially to my whining.

I would also like to thank my parents, for their unwavering support, encouragement and MSN conversations; to the Ng sisters and the best brother in the world, for providing J&Z videos and reading through my thesis just because I said pretty please; to my girlies, my constant and tireless cheerleaders, enough said.

Above all, I wish to thank God, without whom none of these would have any meaning; and Nic, for loving me on this journey of life, for pampering me with home-cooked food of the highest standard and so very little of the washing-up, and for showing me happiness and contentness I would otherwise have never known about.

LIST OF ABBREVIATIONS

BOE	Bank of England
CML	Council of Mortgage Lending
DLTV	Loan to value at default
EAD	Exposure at default
EL	Expected Loss
FSA	Financial Services Authority
HC	Haircut
HCSD	Haircut standard deviation
HPI	House price index
HPIG	House price index growth
IRB approach	Internal ratings-based approach
LGD	Loss given default
LTV	Loan to value
OLS	Ordinary least squares
ONS	Office of National Statistics
PD	Probability of default
RR	Recovery rate (1 – LGD)
SME	Small- and medium-sized entity
TOB	Time on book
UL	Unexpected loss
VaR	Value at risk

CHAPTER 1. INTRODUCTION

The term credit risk refers to the risks associated with lending as financial institutions run the risk of losses if debtors default on their loans. In order to absorb some of these losses, banks hold certain levels of capital – the more they hold, the less risky it is that they would encounter bankruptcy, but it also means that they would make less profit. How much a financial institution holds not only depends on their risk appetite, as the Basel II accord regulates the minimum amount of capital it should hold. In order to comply with these regulations, an institution is required to estimate a number of components that make up credit risk losses. They are the Probability of Default (PD), Exposure at Default (EAD) and Loss Given Default (LGD). In this thesis, a series of novel approaches for LGD estimation are developed and validated, thereby focusing mostly on the LGD of mortgage loans.

This first chapter provides a brief introduction to the Basel II accord, its history and implications, and the credit risk components it defines. Also, further background information is included about the consumer lending portfolios that are the main focus of this thesis, i.e. UK mortgage loans. At the end of the chapter, the three main research parts of this work are introduced and motivated, and the main research contributions are identified. Subsequently, this thesis will explore the credit risk models associated with residential mortgage lending in three parts. First, a loss model is developed for the prediction of mortgage loan LGD, which, in the second part, is further extended by including macroeconomic variables. In the third part, mortgage loan loss is reinvestigated using a different modelling technique, viz. survival analysis, which will provide a novel framework to stress test the estimates. Chapter 2 will detail the data used in the analysis. The three main research parts are then documented in Chapters 3 to 5, respectively. Chapter 6 concludes with a summary of the main results and issues for further work.

1.1 Basel Committee and Basel II Regulations

The Basel Committee on Banking Supervision was established by central bank governors of 10 major industrial countries at the end of 1974, with the aim of providing a set of broad supervisory standards or guidelines for regulatory bodies to adapt for banking infrastructures in different countries. Although the committee does not carry any formal or legal powers, it has the endorsement of governments to steer banking practises in individual countries towards two main objectives: that no foreign banking institution escapes supervision; and that this regulation should be adequate (Basel Committee on Banking Supervision (2001)). In 1988, a set of recommendations were introduced, which regulated minimum capital requirements of financial institutions in an attempt to decrease the chance of bank insolvency. This is commonly known as the Basel Capital Accord (also referred to as "Basel I"), and was subsequently adopted across many non-member countries by banks that were active in international banking. Under Basel I, the different assets of a financial institution are categorised into five different risk classes and assigned a respective risk weight. Capital ratio, defined as the ratio of available capital to weighted risk assets, should then be at least 8%.

In June 1999, the committee set out to define a second Capital Accord, commonly known as Basel II, mainly because different kinds of loans that would have different risk levels were not adequately differentiated and acknowledged; for example, loans that were backed by security (also known as collateral) should be considered less risky than those that are not. According to recent surveys by the Financial Stability Institute, 95 countries are expected to implement some form of the Basel framework by 2015, which would translate to more than 5,000 financial institutions controlling about 75% of banking assets (Financial Stability Institute (2004), Financial Stability Institute (2006)). This widespread implementation of the Basel II Accord highlights the impact it will make on the stability of the financial sector and hence the world economy. It was unfortunate timing that the implementation of Basel II was carried out by international banks as from January 2008 just as the world economy destabilized.

The Basel II Accord consists of three pillars: the Minimum Capital Requirement, the Supervisory Review Process, and a part relating to Market Discipline and Public Disclosure. Pillar I encompasses three main types of risk. They are credit risk, defined to be risks associated with bank lending; operational risk, defined to be risk arising from the internal operations or staff of the financial company; and trading book issues (known as market risk), defined to be risks related to financial instruments or assets held for the purpose of hedging. Nestled within Pillar I of the new Basel II capital framework, there are guidelines and rules on the minimum amount of capital that financial institutions are required to hold for their estimated exposure to credit risk, market risk and operational risk. According to the most advanced approach of these (cf. infra), the minimum capital required by financial institutions to account for their exposure to credit risk is to be calculated for each section of their credit risk portfolios, via three components: the Probability of Default (PD), i.e. the probability of default of a debtor in the following 12 months; Exposure at Default (EAD), the outstanding amount at default; and the Loss Given Default (LGD), the proportion of the remaining loan that the bank would be unable to recover. Expected Loss (EL) is then derived to be the product of the three, for example, a £50,000 loan that has a 2% chance of going into default and a 10% LGD would have an expected loss of 0.02 x 0.1 x 50,000 = £100.

Banks can choose to implement either one of two approaches, the Standardized Approach or the Internal Ratings Based (IRB) Approach.

Table 1.1: Risk components to be developed according to Basel II Approaches

	PD	EAD	LGD
Standardised Approach	-	-	-
Foundation IRB Approach	Internal	Supervisory	Supervisory
Advanced IRB Approach	Internal	Internal	Internal

The Standardized Approach, essentially an extension of the first Basel Accord, does not require any internal estimation of PD or LGD parameters. It divides bank liabilities into sections and dictates that the capital to be set aside should be a direct percentage of the value of exposures. For example, most

retail exposures are risk weighted at 6% of the loan, while residential mortgages are at 2.8%. The IRB approach is further split into two and can be implemented using either the Foundation IRB Approach or the Advanced IRB Approach. Under both IRB approaches, the bank portfolio is split into five asset classes, which are (a) corporate, in which five further sub-classes are defined (project finance, object finance, commodities finance, income-producing real estate, and high volatility commercial real estate); (b) sovereign; (c) bank; (d) retail, which is further divided into three sub-classes (exposures secured by residential properties, qualifying revolving retail exposures, all other retail exposures); and (e) equity. In these asset classes, (internal or external) estimates must be provided for each of the three credit risk components, viz. PD, EAD and LGD. The estimates from these risk models are then inputs to a series of risk-weight functions defined in the Basel II document. Under the Advanced IRB Approach, financial institutions are required to develop their own models for the estimation of all three components for each section of their credit risk portfolios, subject to an estimate floor value (for example, the minimum PD is set to be 0.03%), whereas only the PD component is required for non-retail under the Foundation IRB Approach (see Table 1.1).

Expected losses, which should be calculated via the three credit risk components, are alternatively known as provisions, because they should be incorporated into the pricing of the loan. Also part of the first pillar is the calculation of Unexpected Loss (UL), for which regulatory capital is required, which is a function of PD, downturn LGD, EAD, and a correlation parameter among loans. This calculation is based on a Value at Risk (VaR) Model to estimate unexpected loss based on a 99.9% confidence level; i.e., assuming all model assumptions were correct, there is a 0.1% chance, or a once in a 1,000 years occurrence, that the amount of capital set aside would not cover the unexpected losses, which would in theory cause the financial institution to enter insolvency.

The second pillar of Basel II, the Supervisory Review Process, aims to highlight not only the responsibility of regulatory bodies in terms of supervision and transparency, but also the responsibility of the banks themselves in terms of the development of an internal system of risk

assessment. This would include any aspects of risk not covered by Pillar I, as well as having a risk management structure in place, for example for when to backtest and, if necessary, update the model, or what measures to undertake when the bank takes on unexpected losses. The third pillar of Basel II aims to introduce a standard set of terms for disclosure that all participating financial institutions would have to adhere to. More details for the second and third pillars can be found in the online Basel II documentation and will not be covered here.

The Basel II accord also addresses the need and importance of stress testing, defined to be "a risk management tool used to evaluate the potential impact on a firm of a specific event and/or movement in a set of financial variables"[1]. This would straddle across both Pillars 1 (IRB only) and 2. On the level of Pillar 1, it is necessary to ensure the respective internal risk models developed are able to adequately prepare banks for any unexpected events in the economy, whilst on the broader level of Pillar 2, banks should also be aware of correlations amongst different parts of their portfolios, and have a course of action in place to react to such events quickly and systematically. Stress testing will be again discussed in a later chapter.

In addition, the recent credit crisis of 2008 has drawn much attention to the area of credit risk modelling. Not only is there an increased awareness that these credit risk models are what underpins the amount of regulatory capital banks are required to hold, the quality of these models (as well as the role of regulatory bodies) are also questioned. Also, even though only some portfolios of loans were affected (in particular, residential mortgages and the structured products into which they were packaged), the resulting lack of confidence by the public or the markets could drive a bank to the brink of bankruptcy, which could then affect the wider economy swiftly. Therefore, the importance of building and validating robust and accurate risk models has become ever more apparent in recent years.

[1] Section 2.1 of Stress Testing at Major Financial Institutions: Survey Results and Practice (Committee on the Global Financial System (2005)).

1.2 Mortgage loans in the UK

As described in the section above, the portfolios of a financial institution can be broadly divided into five classes, including corporate and retail exposures. Traditionally, corporate exposures have had much more attention paid to them in the literature, which can be partly explained by the greater availability of (public) data and because the financial health or status of the debtor companies can be directly inferred from share and bond prices traded on the market. However, this is not the case for retail exposures. Retail exposures are further divided into three sub-classes, viz. exposures secured by residential properties, qualifying revolving retail exposures (e.g. credit cards) and all other retail exposures (e.g. Small and Medium Entities (SME) retail). Each of these is given different risk-weight functions, based on their PD and LGD estimates. The work here focuses mainly on residential mortgage loans.

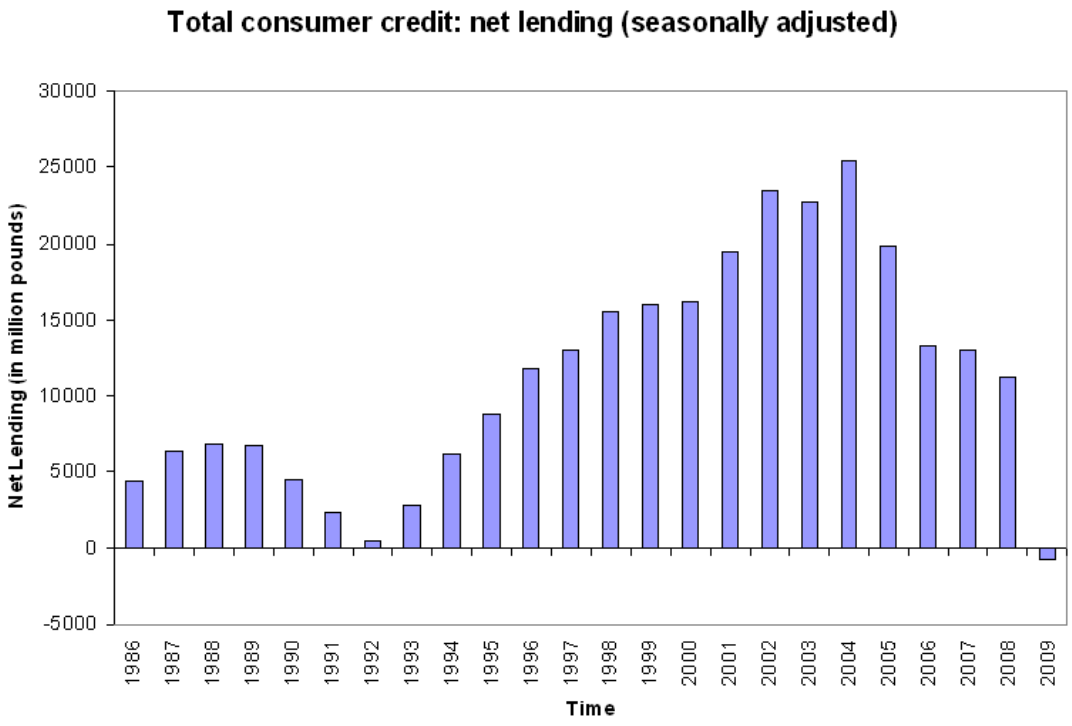

Figure 1.1: Total consumer credit: UK net lending, in sterling millions, seasonally adjusted, for years 1986 to 2009, source from ONS and BOE.

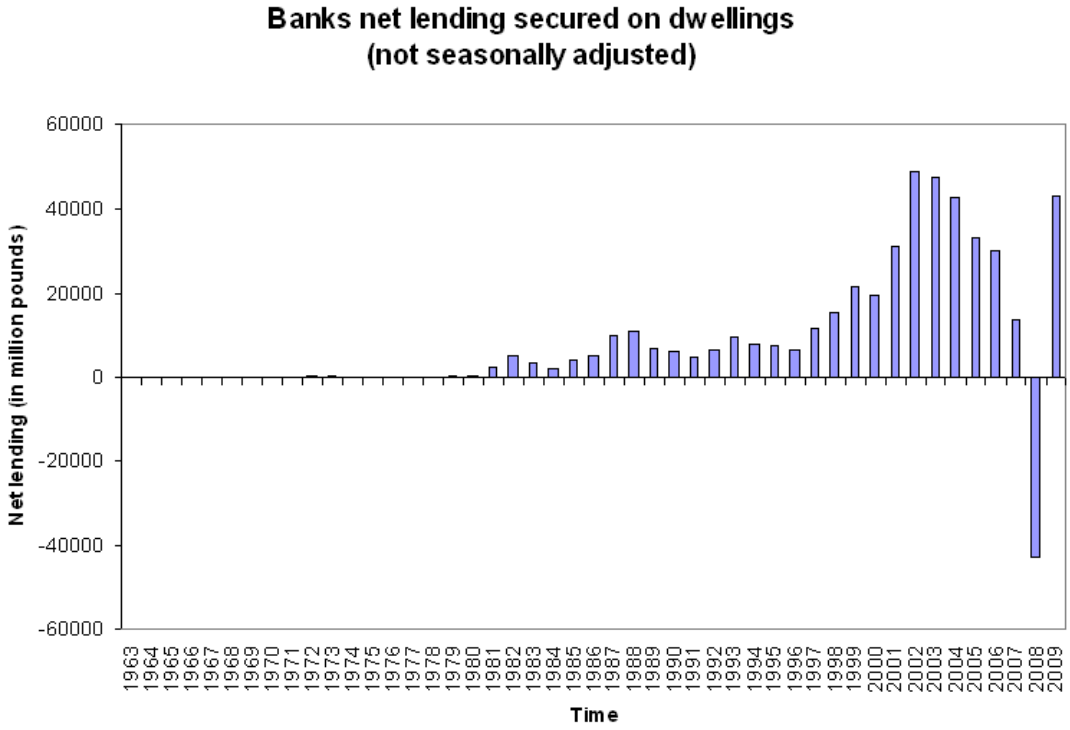

Figure 1.2: Banks UK net lending secured on dwellings (residential properties), in sterling millions, for years 1963 to 2009, source from ONS

In the UK, until the recent downturn, lending to consumers has seen steady growth since the 1990s, as shown in Figure 1.1. The drops in net lending[2] observed on the graph correspond to the two economic downturns experienced by the UK in 1991 and 2008. However, we also see that net lending levels recovered quickly in the mid 1990s and saw a growth even greater than before, which again reinforces the growing importance of retail lending and indicates how vital it is to have a good and robust credit model tailored for each sub-class of retail exposures. According to the Council for Mortgage Lending (CML), there are about 11.4 million mortgages worth over 1.2 trillion pounds as of 2010[3]. Figure 1.2 sees how banks' net lending secured on dwellings (residential properties) has grown since 1963. After the economic downturn in 1991, net lending in mortgages increased from 4.7 billion pounds to 48.9 billions pounds in 2002. This exponential rise is probably due to a combination of consumers having easier access to loans

[2] Net lending is the flow of gross lending less the flow of repayments.
[3] Council of Mortgage Lending website.
URL: http://www.cml.org.uk/cml/media/press/2674.

and because UK house prices doubled in the 2000s (see Figure 1.3). The Basel Committee has also recognised this, and close inspection of the amount of savings (in terms of capital) a bank can make by adopting the advanced IRB approach (as compared against the old Accord) shows that the bulk of it is in the mortgage portfolio (see Table 8 of Quantitative Impact Study 5 in Basel Committee on Banking Supervision (2006)).

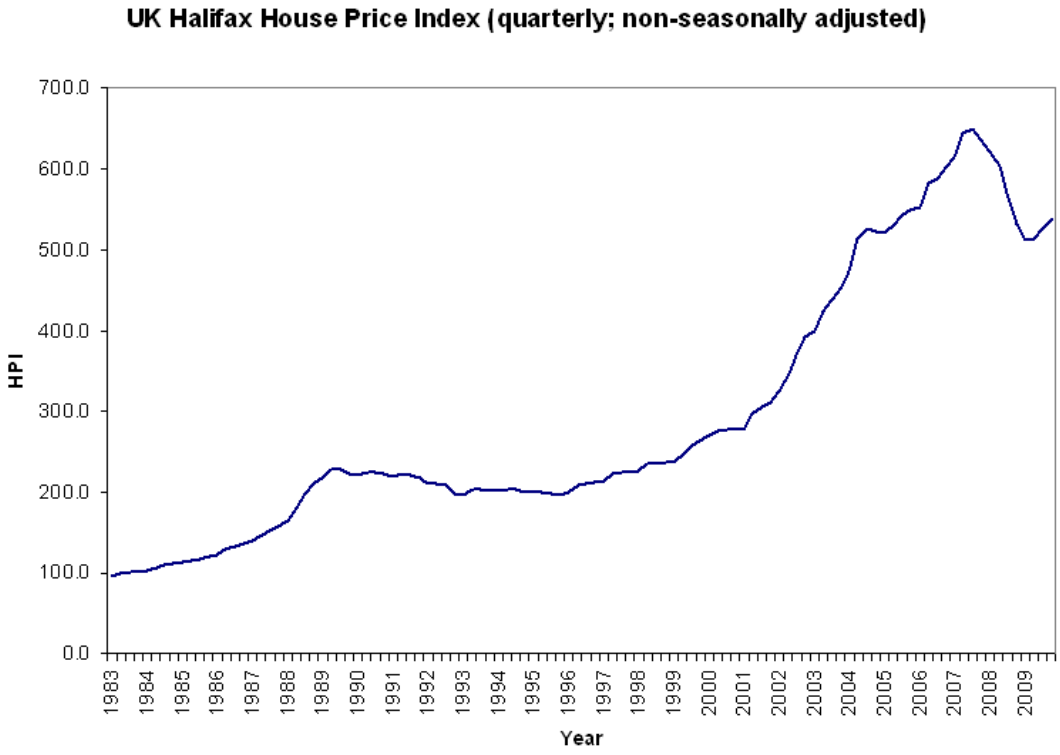

Figure 1.3: Halifax house price index for the UK, quarterly, non-seasonally adjusted, for years 1983 to 2009, source from Lloyds Banking Group

In the United Kingdom, as in the US, the local Basel II regulation specifies that a mortgage loan exposure is in default if the debtor has missed payments for 180 consecutive days (The Financial Services Authority (2009), BIPRU 4.3.56 and 4.6.20; Federal Register (2007)). In the case of residential mortgages, loss does not necessarily occur when an account goes into default if the property does not undergo repossession (and also in the case where sale proceeds from the property are able to cover the outstanding loan and other related costs).

When a loan goes into default, financial institutions could contact the debtor for a re-evaluation of the loan whereby the debtor would have to pay a slightly higher interest rate on the remaining loan but have lower and more manageable monthly repayment amounts; or banks could decide to sell the loan to a separate company which works specifically towards collection of repayments from defaulted loans; or, because every mortgage loan has a physical security (also known as collateral), i.e. a house or flat, the property could be repossessed (i.e. enter foreclosure) and sold by the bank to cover losses. In this case, there are two possible outcomes: either the sale of the property is able to cover the outstanding loan (and the costs associated with default and repossession), and any excess the bank has made in repossessing and selling the property must be returned, so this scenario would result in a loss rate of zero; the alternate outcome is that the sale proceeds are less than the outstanding balance and there is a loss. In the case where the bank does not repossess the property, or was not able to sell it after repossession, the loss rate is again often assumed to be zero (if no further data is available). Note that the distribution of LGD in the event of repossession is thus capped at one end. Another thing to note is that when a property undergoes repossession, one cannot expect it to fetch a price equal to its market value due to the circumstances. The repossessed property is usually sold at a price lower than its market value and this difference is referred to as the haircut (defined here to be the ratio of forced sale price over its current market valuation)[4]. More detail is given in the next section, as well as in later chapters.

1.3 Overview of the three research projects

Having had an overview of the importance and widespread implementation of the Basel II Accord and an understanding of the increasing worth of the retail portfolio, we formulate a number of research questions that this thesis addresses.

[4] Haircut can either be defined as $\frac{\text{forced sale price}}{\text{current market valuation}}$ or $\frac{\text{current market valuation - forced sale price}}{\text{current market valuation}}$

In the first part, we investigate, using a large set of recovery data of residential mortgage defaults from a major UK bank, the LGD of mortgage loans according to the repossession procedure that defaulted mortgage loans would undergo. The aim of LGD modelling in the context of residential mortgage lending is to accurately estimate the loss as a proportion of the outstanding loan, if the loan were to go into default. A two-stage approach is proposed here: in order to differentiate between accounts that would be repossessed and those that would not, a Probability of Repossession Model is developed; and to investigate the level of haircut each repossessed property would be expected to undergo, a Haircut Model is developed. In those models, a common indicator is the loan-to-value ratio, which is the ratio of the amount of loan over the current valuation of the property, which can be updated at various points of time throughout the loan period to reflect the amount of loan the debtor has paid back and any appreciation in the value of the property. In the first part of the thesis, we therefore also investigate the role of the loan-to-value ratio and which is more suitable for use within each component model (LTV at origination, henceforth referred to as LTV; or LTV at default, henceforth referred to as DLTV), as well as find out what is the most appropriate method of combining the component models to get an expected loss percentage. We note however that repossession decisions and hence LGD are partly related to a bank's policy decisions (as well as external legislative factors); hence any statistical model will inherently have certain limitations as far as predictive power or forward-lookingness is concerned.

In the second part of the thesis, we subsequently examine the role of macroeconomic variables in retail LGD by testing the inclusion of a range of macroeconomic variables in two different retail LGD models: the two-stage mortgage LGD model developed in the first part and an unsecured personal loans LGD model, built on a second available dataset. There has been ample evidence in the corporate sector that some macroeconomic variables are able to improve predictions of PD and LGD, which is to be expected because a poor economic environment would affect all debtors, as well as magnify any adverse effects due to the empirical correlation found between PD and LGD. We are inclined to expect similar results in the retail sector and indicators relating to e.g. net lending, gross domestic product and interest rates are considered.

In the third and final part, we re-visit the development of an LGD model for residential mortgage loans. In the previous parts, LGD models were developed using a combination of logistic and linear regression models. These regression techniques have been the popular choice and are typically able to produce decent predictions of LGD, but a repossession model based on logistic regression is only able to predict whether an event (here: repossession) will happen but not when. Also, because traditional regression models are only able to take into account static variables, time-dependent macroeconomic variables (i.e. macroeconomic variables that change with time) cannot be fully incorporated. Survival models, on the other hand, are able to incorporate time-dependent variables and produce estimates for the likelihood of an event happening at each (say monthly) time step. A competing risks survival model is to be developed in order to reflect that a defaulted mortgage account could either enter repossession or be closed, both of which would give very different loss estimates. Eventually, this model shall be used for the purpose of stress testing by applying a Monte Carlo simulation, and by varying the macroeconomic variables, to get different loss distributions for different economic outlooks.

In this sense, we see that all three parts of the thesis explore the development of LGD models for residential mortgage loans. Both regression and survival analysis methods are covered, and the role of macroeconomic variables in the prediction of LGD is explored. We recognise the importance of stress testing and try to incorporate a suitable framework for it during our analysis and modelling.

Hence, this work aims to contribute to the existing academic literature on credit risk modelling, especially the still relatively under-explored area of LGD models for retail exposures. More specifically, the major contributions of this thesis are:

- The development of a Probability of Repossession Model that is shown to perform significantly better than a model with just the commonly used DLTV as the explanatory variable;
- The detailed description and empirical validation of a two-stage model consisting of two component models, a Probability of Repossession Model and a Haircut Model (which consists of the Haircut Model itself

and the Haircut Standard Deviation Model) to produce estimates for LGD, and its comparison against a single-stage model that it was shown to outperform;

- Our empirical investigation based on two real-life datasets of the extent to which macroeconomic variables could be beneficial for LGD models in retail lending, the results of which suggest limited scope for improvement in predictive performance, particularly so for unsecured personal loans;
- The development of a novel approach based on survival analysis models with time-dependent variables to better estimate the time taken for defaulted mortgage loan accounts to go from default to some event (repossession or otherwise), which can be used alongside the former two-stage model to more accurately estimate discounted loss;
    - It was found that the time-dependent variable (HPIG) gives valuable insight on how drivers of risk are different for different types of securities of different DLTV bands in different economic climates
- The development and validation of an appropriate framework using Monte Carlo simulation through which mortgage loan LGD stress testing can be carried out and which is shown to produce intuitive typical and downturn loss distributions.

CHAPTER 2. DATA

The dataset used in this study is supplied by a major UK Bank, with observations coming from all parts of the UK, including Scotland, Wales and Northern Ireland. There are more than 140,000 observations and 93 variables in the original dataset, all of which are on defaulted mortgage loans, with each account being identified by a unique account number. About 30 percent of the accounts in the dataset undergo repossession, and time between default and repossession varies from a couple of months to several years. After pre-processing we retain about 120,000 observations.

Under the Basel II framework, financial institutions are required to forecast default over a 12-month horizon and resulting losses from a given time point (referred to here as "observation time"). As such, LGD models developed should not contain information that only becomes available at time of default. However, due to limitations in the dataset, in which information on the state of the account in the months leading up to default (e.g. outstanding balance at observation time) are unavailable, we use approximate default time instead of observation time. When applying this model at a given time point, a forward-looking adjustment could then be applied to convert the current value of that variable, for example, outstanding balance, to an estimate at time of default. Default-time variables for which no reasonable projection is available are removed.

2.1 Time periods and geography

This dataset consists of observations that encompass a fairly long time frame. 90% of observations have a start year (also known as loan origination) ranging from 1983 to 2002, and 99% of observations originate between the years 1970 to 2002. In the end, all loans predating 1983 were removed because of the unavailability of house price index data for these older loans. Note that this sample does not encompass observations from the recent economic downturn.

All observations in this dataset enter a state of default between the years 1988 to 2002. For properties that undergo repossession, together with their subsequent outcome of being sold or otherwise, we only have information about these dates if repossession occurred in year 2003 or earlier. Depending on which LGD model is being developed, we use different periods of time. In the first two parts of the thesis where LGD models are developed using a combination of logistic and linear regression models, only loans that default between the years of 1988 and 2001 are used, because we have to allow at least a two year outcome window for repossession to happen, if any. Logistic regression is only able to differentiate between whether the account experienced the (repossession) event in some fixed time frame or not, so an outcome window that is too short might wrongly handle observations that might experience the event after this time period. In the third part where survival analysis was used in LGD model development, we were able to use all available data, i.e. accounts that defaulted even in 2002, because survival models are able to handle censoring, where the more recent observations that did not (yet) experience any event in the first t months after default are differentiated from those for which we have observed an actual outcome.

In terms of geographical information, although exact postcodes of properties are not known, we have information about the region that the security is in. The UK is divided into a number of sections, namely: North, Yorkshire and Humberside, North West, East Midlands, West Midlands, East Anglia, Wales, South West, South East, Greater London, Northern Ireland, Scotland. These districts are standard and common, with Housing Indices using the same divisions.

2.2 Multiple defaults

Some accounts have repeated observations, which means that some customers were oscillating between keeping up with their normal repayments and going into default, whereby each default is recorded as a separate observation of the characteristics of the loan at that time. The UK Basel II regulations state that the financial institution should return an exposure to non-default status in the case of recovery, and record another default should

the same exposure subsequently go into default again (Financial Services Authority (2009), BIPRU 4.3.71). For the LGD models detailed in the first two parts of the thesis, we include all instances of default in our analysis, and record each default that is not the final instance of default as having zero LGD.

2.3 Date of default

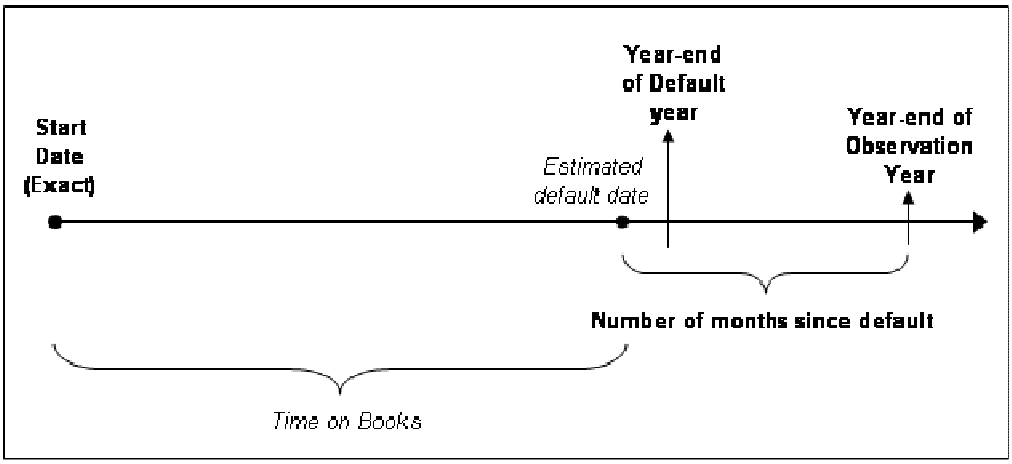

Figure 2.1: Estimation of default date and calculation of variable time on book. Variables in bold are given in the dataset, variables in italics are calculated from other variables.

In this dataset, we are given only the year of default. However, we have also another year variable which we believe to be an indication of an observation year post default, and a variable containing the number of months between the estimated[5] date of default and the year-end of this observation year. By subtracting number of months in default from the year-end of observation year, as illustrated in Figure 2.1, we are able to extract a more specific (estimated) date of default. As a check, we compare the year from estimated default date with the year of default as given in the dataset, and find that both values coincide for all observations. From this estimated default date, we are also able to calculate time on books (see next sub-section).

[5] Date of default was estimated by the bank using the arrears status and amount of cumulated arrears at the end of each year for each account. However, we are not explicitly given this date in the original dataset.

2.4 Time on books

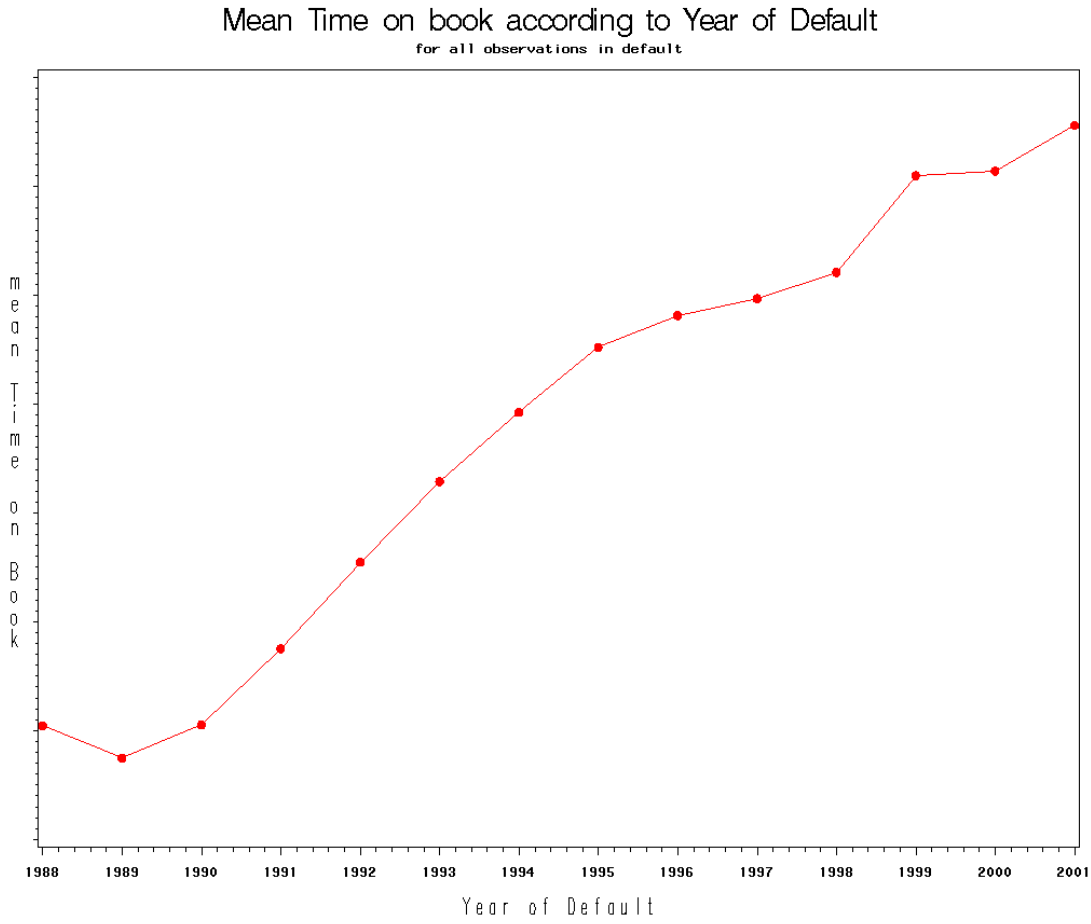

Figure 2.2: Mean time on book over time with reference to year of default. Due to a data confidentiality agreement with the data provider, the scale for the vertical axis has been omitted in some of the reported figures.

Time on Book is calculated to be the time between the start date of the loan and the approximate date of default (see Section 2.3). The variable time on book exhibits an obvious increasing trend over time (cf. Figure 2.2) which might be partly due to the composition of the dataset. In the dataset, we have defaults between years 1988 and 2002, which just about coincides with the start of that economic downturn in the UK in the early nineties. We observe that the mean time on book for observations that default during the economic downturn is significantly lower than the mean time on book for observations that default in normal economic times.

2.5 Repossession

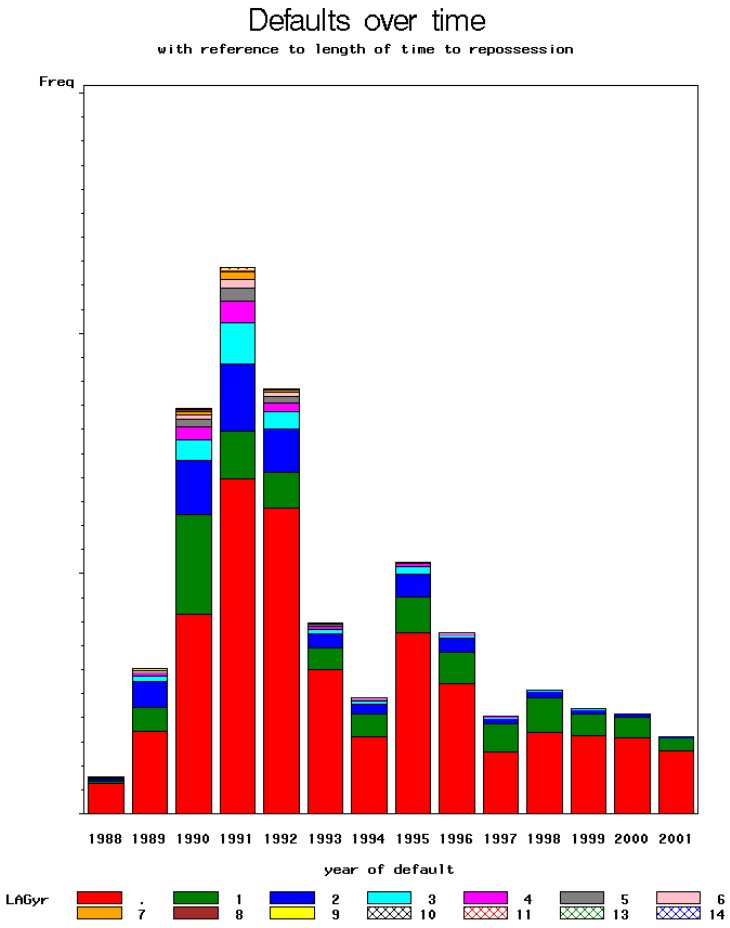

Figure 2.3: Defaults over time with reference to length of time to repossession. The red parts represent observations that were not repossessed, and the rest of the colours correspond to the number of years it took for repossession to occur.

About 35% of observations in this dataset undergo repossession. Properties that are repossessed are usually repossessed within a couple of years of the account going into default – about 75% of observations that get repossessed are repossessed within 24 months after default. The histogram in Figure 2.3 gives a graphical representation of the subsequent outcomes for those observations that went into default in a given year. The different colours in each vertical bar represents whether they were subsequently repossessed, and if so, whether they were repossessed within the 12-month intervals after default, up to 14 years. We see that most defaulted accounts do not undergo repossession (represented by the red parts); but of those that do, the majority is repossessed within two years of time of default (the green and

blue parts representing repossession within the first and second 12-month intervals after default, respectively). Also, we observe a large increase in number of repossessions between the years of 1990 and 1992, which corresponds to the UK economic downturn of the early nineties as shown in historical economic data.

2.6 Valuation of security at default and haircut

At the time of the loan application, information about the market value of the property is obtained. As reassessing its value would be a costly exercise, no new market value assessment tends to be undertaken thereafter and valuation of property at various points of the loan can be obtained by updating the initial property value using the publicly available Halifax House Price Index [6] (all houses, all buyers, non-seasonally adjusted, quarterly, regional). The valuation of security at default is calculated according to Equation 2.1.

$$\text{Valuation of security}_{\text{def}} = \frac{\text{HPI}_{\text{def yr, qtr, region}}}{\text{HPI}_{\text{start yr, qtr, region}}} \times \text{Valuation of security}_{\text{start}} \tag{2.1}$$

Using this valuation of security at default, other variables are then updated. One is the ratio of valuation of property over the average property value in the region, which gives an indication of the quality of the property relative to the rest of the properties in the same area; another is loan to value at default (DLTV), which is defined to be the ratio of the outstanding loan amount at default over the valuation of security at default. The calculation of haircut, which will be the target variable in our second component model, also depends on valuation of security at default. It is only applicable for observations with a valid forced sale price, i.e. where the security has undergone repossession and sale, and is defined in Equation 2.2.

[6] Available at: http://www.lloydsbankinggroup.com/media1/research/halifax_hpi.asp

$$\text{Haircut} = \frac{\text{forced sale price}}{\text{valuation of property}_{\text{default}}} \tag{2.2}$$

For example, a property estimated to be valued at £1,000,000 but repossessed and sold at £700,000 would have a haircut of $\frac{£700{,}000}{£1{,}000{,}000} = 0.7$.

2.7 Type of security

There are five types of securities in this dataset – flats, terraced, semi-detached, detached, and others. However, because there are only a small number of observations under others, they are combined with flats.

2.8 Training and test set splits

To obtain unbiased performance estimates of model performance, we make sure to set aside an independent test dataset. We develop all component models (i.e. the Probability of Repossession Model, the Haircut Model, the Haircut Standard Deviation Model, the macroeconomic LGD models and the survival models) on a training set before applying the models onto a separate test set that was not involved in the development of the model itself, to gauge the performance of the model and to ensure there is no over-fitting. To do so, we split the cleaned dataset into two-third and one-third sub-samples, keeping the proportion of repossessions the same in both sets (i.e. stratified by repossession cases). This training and test set split remains the same throughout all three parts.

2.9 Loss given default

Only recoveries from forced sales are included in this dataset; hence loss is defined to be the outstanding balance less recovery amount from the sale of the repossessed property.

When a loan goes into default and the property is subsequently repossessed by the bank and sold, legal, administrative and holding costs are incurred. As this process might take a couple of years to complete, revenues and costs have to be discounted to present value in the calculation of LGD, and should include any compounded interest incurred on the outstanding balance of the loan. In our analysis, because we are not provided with information about the legal and administrative costs associated with each loan, the definition of LGD is simplified to exclude both the extra costs incurred and the interest lost. Two slightly different definitions of LGD are used for different parts of this work.

In Chapters 3 and 4, where regression models for LGD are developed, we are only able to predict if repossession, and hence any loss, happens, but not when these losses occur. Hence, LGD is as defined in Equation 2.3, the ratio of the final (nominal) loss from the defaulted loan over the outstanding loan balance at (year end of) default, and where loss is defined to be the difference between outstanding loan at default and the forced sale amount, if the property was sold at a price that is lower than the outstanding loan at default (i.e. outstanding loan at default > forced sale amount). If the property was able to fetch an amount greater than or equal to the outstanding loan at default, then loss is defined to be zero. If the property was not repossessed, or repossessed but not sold, loss is also assumed to be zero, in the absence of any additional information. With loss defined to be zero, LGD is of course also 0.

$$\mathrm{LGD}_{\mathrm{nominal}} = \max\left\{0, \frac{\mathrm{loan\,balance}_{\mathrm{default}} - \mathrm{forced\,sale\,amount}}{\mathrm{loan\,balance}_{\mathrm{default}}}\right\} \tag{2.3}$$

In Chapter 5 however, where we use survival analysis to get an estimate of when repossession will happen (in the case of default and repossession), we are able to form an estimate for the time required from time of default to time of sale, so here we choose to model discounted (rather than nominal) loss, as defined by Equation 2.4.

$$\text{LGD}_{\text{discounted}} = \max\left\{0, \frac{\dfrac{\text{loan balance}_{\text{default}} - \text{forced sale amount}}{(1+d)^{k}}}{\text{loan balance}_{\text{default}}}\right\} \tag{2.4}$$

Where $d = \text{discount rate}$ and $k = \dfrac{\text{number of months between default and sale}}{12}$.

CHAPTER 3. PREDICTING LOSS GIVEN DEFAULT (LGD) FOR RESIDENTIAL MORTGAGE LOANS: A TWO-STAGE MODEL

In this chapter, the LGD of mortgage loans is modelled using a combination of logistic and linear regression. A Probability of Repossession Model and a Haircut Model (comprising of a Haircut Model and a Haircut Standard Deviation Model) are developed and then combined into an expected loss percentage[7].

3.1 Literature review

Much of the work on prediction of LGD, and to some extent PD, proposed in the literature pertains to the corporate sector (see Schuermann (2004), Gupton and Stein (2002), Jarrow (2001), Truck, Harpaintner and Rachev (2005), Altman et al. (2005)), which can be partly explained by the greater availability of (public) data and because the financial health or status of the debtor companies can be directly inferred from share and bond prices traded on the market. However, this is not the case in the retail sector, which partly explains why the LGD models are not as developed as those pertaining to corporate loans.

3.1.1 Risk models for residential mortgage lending

Despite the lack of publicly available data, particularly on individual loans, there are still a number of interesting studies on credit risk models for mortgage lending that use in-house data from lenders. However, the majority of these have in the past focused on the prediction of default risk, as comprehensively detailed by Quercia and Stegman (1992). One of the earliest papers on mortgage default risk is by von Furstenberg (1969) who

[7] This chapter is based on a paper currently under review at the International Journal of Forecasting.

found that characteristics of a mortgage loan can be used to predict whether default will occur. These include LTV (at origination), term of mortgage, and age and income of the debtor. Following that, Campbell and Dietrich (1983) further expanded on the analysis by investigating the impact of macroeconomic variables on mortgage default risk. They found that LTV is indeed a significant factor, and that the economy, especially local unemployment rates, does affect default rates. This is confirmed more recently by Calem and LaCour-Little (2004), who looked at estimating both default probability and recovery on defaulted loans from the Office of Federal Housing Enterprise Oversight (OFHEO). Of interest was how they estimated recovery by employing spline regression to accommodate the non-linear relationships that were observed between both loan-to-value ratios (LTV and DLTV) and recovery, which achieved an R-square of 0.25.

Similarly to Calem and LaCour-Little (2004), Qi and Yang (2009) also modelled loss directly using characteristics of defaulted loans, using data from private mortgage insurance companies, in particular on accounts with high loan-to-value ratios that have gone into default. In their analysis, they were able to achieve high values of R-square (around 0.6) which could be attributed to their being able to revalue properties at time of default (near-perfect expert-based information that would not normally be available to lenders on all loans; hence one would not be able to use it in the context of Basel II which requires the estimation of LGD models that are to be applied to all loans, not just defaulted loans).

3.1.2 Single vs. two-stage LGD models

Whereas the former models estimate LGD directly and will thus be referred to as “single-stage" models, the idea of using a so-called “two-stage" model is to incorporate two component models, the Probability of Repossession Model and the Haircut Model, into the LGD modelling. Initially, the Probability of Repossession Model is used to predict the likelihood of a defaulted mortgage account undergoing repossession. It is sometimes thought that the probability of repossession is mainly dependent on one variable, viz. a loan-to-value ratio, hence some probability of repossession models currently in use only consist of this single variable (Lucas (2006)). This is then followed

by a second model which estimates the amount of discount the sale price of the repossessed property would undergo. The Haircut Model predicts the difference between the forced sale price and the market valuation of the repossessed property. These two models are then combined to get an estimate for loss, given that a mortgage loan would go into default. An example study involving the two-stage model is that of Somers and Whittaker (2007), who, although they did not detail the development of their Probability of Repossession Model, acknowledged the methodology for the estimation of mortgage loan LGD. In their paper, they focus on the consistent discount (haircut) in sale price observed in the case of repossessed properties and because they observe a non-normal distribution of haircut, they propose the use of quantile regression in the estimation of predicted haircut. Another paper that investigates the variability that the value of collateral undergoes is by Jokivuolle and Peura (2003). Although their work was on default and recovery of corporate loans, they highlight the correlation between the value of the collateral and recovery.

In summary, despite the increased importance of LGD models in consumer lending and the need to estimate residential mortgage loan default losses at the individual loan level, still relatively few papers have been published in this area apart from the ones mentioned above.

3.2 Research objectives

From the literature review, we observe that the few papers which looked at mortgage loss did so either by directly modelling LGD ("single-stage" models) using economic variables and characteristics of loans that were in default or did not look at both components of a two-stage model, i.e. haircut as well as repossession. This might be due to their analysis being carried out on a sample of loans which had undergone default and subsequent repossession, and thus removed the need to differentiate between accounts that would undergo repossession from those that would not. We note also that there was little consideration for possible correlation between explanatory variables.

Hence, the two main objectives of this chapter are as follows. Firstly, we intend to evaluate the added value of a Probability of Repossession Model with more than just one variable (loan-to-value ratio). Secondly, using real-life UK bank data, we would also like to empirically validate the approach of using two component models, the Probability of Repossession Model and the Haircut Model, to create a model that produces estimates for LGD. We develop the two component models before combining them by weighting conditional loss estimates against their estimated outcome probabilities.

3.3 The Probability of Repossession Model

Our first model component will provide us with an estimate for the probability of repossession given that a loan goes into default. Observations that undergo repossession within the observation period of the dataset are defined to be repossessions. About 35% of the observations in this dataset are repossessed.

3.3.1 Modelling methodology

We first identify a set of variables that are eligible for inclusion in the Repossession Model. Variables that cannot be used are removed, including those which contain information that is only known at time of default and for which no reasonably precise estimate can be produced based on their value at observation time (e.g. arrears at default, the number of months between observation time and default), or those that have too many missing values[8], are related to housing or insurance schemes that are no longer relevant, or where the computation is simply not known. We also then check the correlation coefficient between pairs of remaining variables, and find that none are greater than | 0.6 |. Using these, a logistic regression is then fitted onto the repossession training set and a backward selection method based on the Wald test is used to keep only the most significant variables (p-value of at most 0.01). We then check that the signs of each parameter estimate

[8] The variables in the original dataset were either consistently present or missing. In the case where variables do have missing observations, at least 30% of the observations are not available.

behave logically, and that parameter estimates of groups within categorical variables do not contradict with intuition.

3.3.2 Model variations

Using the methodology above, we obtain a Probability of Repossession Model *R1*, with four significant variables: LTV (at origination), a binary indicator for whether this account has had a previous default, time on book in years and type of security, i.e. detached, semi-detached, terraced, flat or others. In a second model, we replace LTV and time on book with DLTV, referred to as Probability of Repossession Model *R2*. Including all three variables (LTV, DLTV and time on books) in a single model would cause counter-intuitive parameter estimate signs. Another simpler repossession model fitted on the same data, against which we will compare our model, is Model *R0*. The latter model only has a single explanatory variable, DLTV, which is often the main driver in models used by the retail banking industry.

3.3.3 Performance measures

Performance measures applied here are accuracy rate, sensitivity, specificity, and the Area Under the ROC Curve (AUC).

In order to assess the accuracy rate (i.e. total number of correctly predicted observations as a proportion of total number of observations), sensitivity (i.e. number of observations correctly predicted to be events – in this context: repossessions – as a proportion of total number of actual events) and specificity (i.e. number of observations correctly predicted to be non-events – in this context: non-repossessions – as a proportion of total number of actual non-events) of each logistic regression model, we have to define a cut-off value for which only observations with a probability higher than the cut-off are predicted to undergo repossession. How the cut-off is defined affects the performance measures above, as it affects how many observations shall be predicted to be repossessions or non-repossessions. For our dataset, we choose the cut-off value such that the sample proportions of actual and predicted repossessions are equal. However, we note that the exact value

selected here is unimportant in the estimation of LGD itself as the method later used to estimate LGD does not require selecting a cut-off.

The Receiver Operating Characteristic (ROC) curve is a 2-dimensional plot of sensitivity and 1 – specificity values for all possible cut-off values. It passes through points (0,0), i.e. all observations are classified as events, and (1,1), i.e. all observations are classified as non-events. A straight line through (0,0) and (1,1) represents a model that randomly classifies observations as either events or non-events. Thus, the more the ROC curve approaches point (0,1), the better the model is in terms of discerning observations into either category. As the ROC curve is independent of the cut-off threshold, the area under the curve (AUC) gives an unbiased assessment of the effectiveness of the model in terms of classifying observations.

We also use the DeLong, DeLong and Clarke-Pearson test (DeLong, DeLong and Clarke-Pearson (1988)) to assess whether there are any significant differences between the AUC of different models.

3.3.4 Model results

Table 3.1: Repossession models performance statistics

Model	AUC	Cut-off	Specificity	Sensitivity	Accuracy
R1 Test Set (LTV, time on books, Security, Previous default)	0.727	0.435	57.449	75.688	69.186
R2 Test Set (DLTV, Security, Previous default)	0.743	0.432	59.398	76.203	70.213
R0 Test Set (DLTV)	0.737	0.436	58.626	76.008	69.812
DeLong et al. p-value, *R1* vs. *R0*	<0.001				
DeLong et al. p-value, *R2* vs. *R0*	<0.001				

Applying the DeLong, DeLong and Clarke-Pearson test we find that the AUC value for model *R2* is significantly better than that for *R0* (cf. Table 3.1). Model *R2* is selected for further inclusion in our two-stage model. Table 3.2 gives the direction of parameter estimates used in the Probability of Repossession Model *R2*, together with a possible explanation. The parameter estimate values and p-values of all repossession model variations can be found in Appendix A, Tables A1, A2 and A3.

Table 3.2: Parameter estimate signs for Probability of Repossession Model *R2*

Variable	Relation to probability of repossession (given default)	Explanation
DLTV	+	If a large proportion of loan is tied up in security, likelihood of repossession increases
Previous default	+	Probability of repossession increases if account has been in default before
Security	+	Lower-range properties such as flats (compared to higher-range properties such as detached) are more likely to be repossessed in the case of default

3.4 The Haircut Model

The Haircut Model is only applicable to observations that have undergone the repossession and forced sale process, where haircut is defined to be the ratio of forced sale price to valuation of security at default. Therefore, securities that were not repossessed, or repossessed but not sold do not have a haircut value, and are thus excluded from the development of the Haircut Model.

An OLS model is also developed to explicitly model haircut standard deviation, as a function of time on books, as suggested by Lucas (2006).

The distribution of haircut is shown in Figure 3.1 with the solid curve referencing the normal distribution. Statistics from the Kolmogorov-Smirnov and Anderson-Darling Tests (Peng (2004)) suggest non-normality with p-values of <0.01 and <0.005 respectively, but for the purposes of the prediction of LGD, we approximate haircut by a normal distribution.

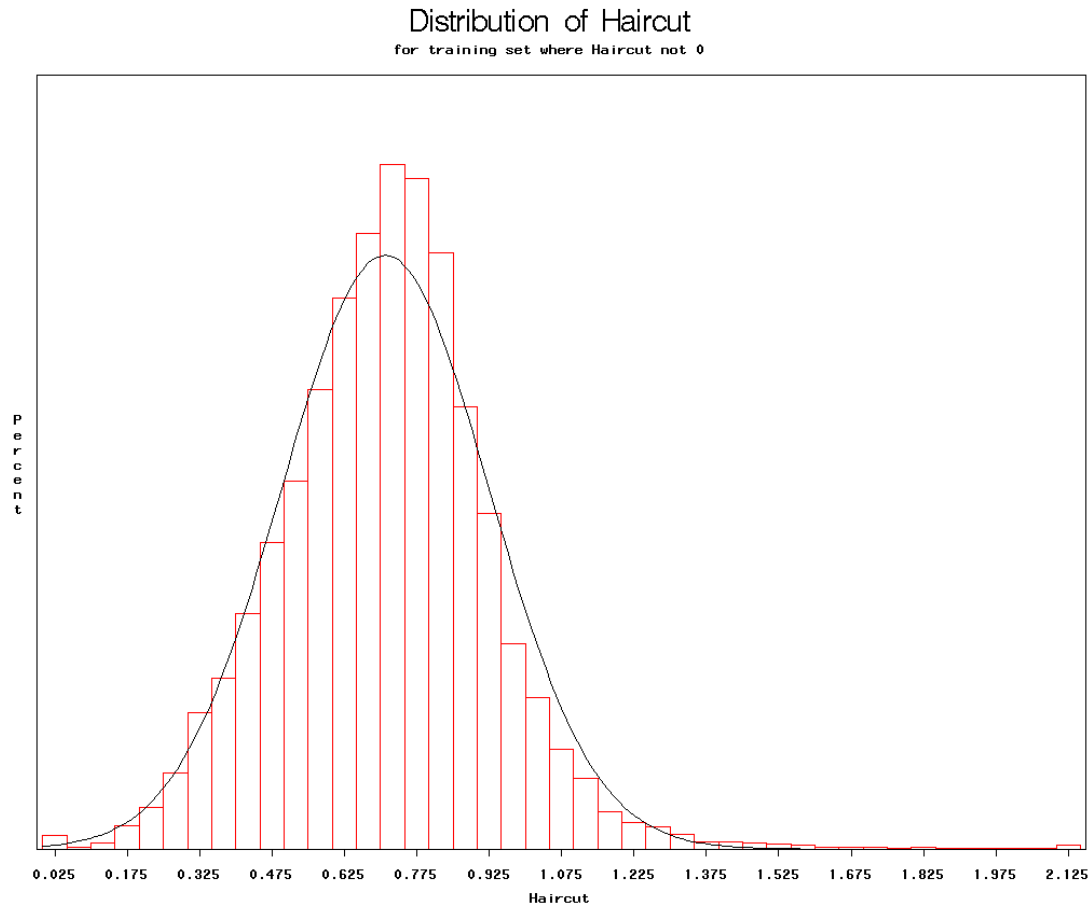

Figure 3.1: Distribution of haircut. Solid curve references the normal distribution.

3.4.1 Modelling methodology

The top and bottom 0.05 percent of observations (26 cases) for haircut are truncated before we establish the set of eligible variables to be considered in the development of an OLS linear regression model for the Haircut Model. We also check the relationship between variables and haircut. In particular, the valuation of security at default to average property valuation in the region ratio displays high non-linearity (cf. Figure 3.2) and is binned into 6 groups for model development. Backward stepwise regression is used to

remove insignificant variables and individual parameter estimate signs are checked for intuitiveness. We also check for intuition within categorical variables, and examine the Variance Inflation Factors (VIF)[9].

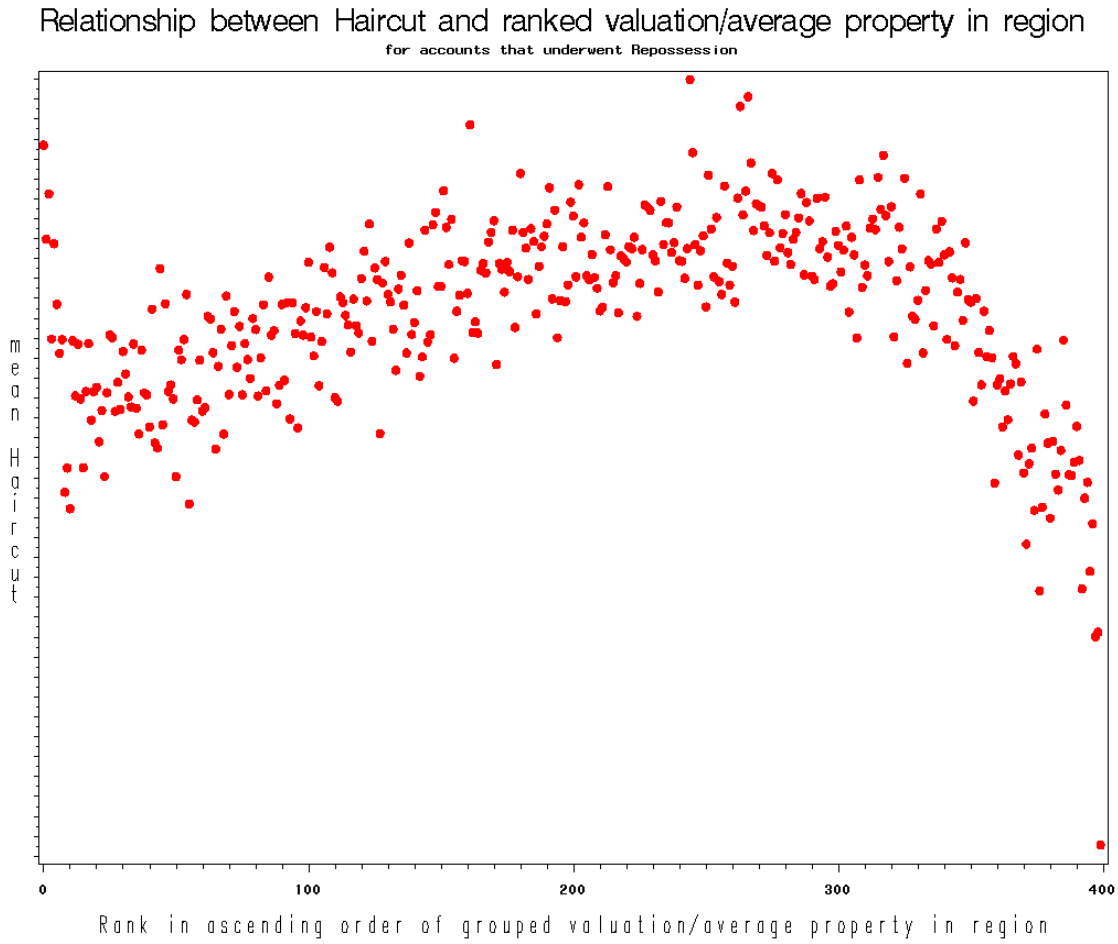

Figure 3.2: Relationship between haircut and (ranked) valuation of security at default to average property valuation in the region ratio

3.4.2 Model variations

Using the methodology above, we obtain a Haircut Model *H1*, with seven significant variables: LTV (at origination), a binary indicator for whether this account has had a previous default, time on book in years, ratio of valuation of property to average in that region (binned), type of security, i.e. detached,

[9] If variables within the model are highly correlated with each other, it would be reflected in a high value of VIF. Any value above 10 would imply severe collinearity amongst variables while values less than 2 would mean that variables are almost independent (Fernandez (2007)).

semi-detached, terraced, flat or other, age group of property and region. In a second model, we replace LTV and time on book with DLTV, referred to as Haircut Model *H2*; note that, as previously, including all three variables (LTV, DLTV and time on books) in a single model would cause counter-intuitive parameter estimate signs. Comparative performance measures for the two models are reported in the following section.

3.4.3 Performance measures

The performance measures considered here are the R-square value (see Equation 3.1), Mean Squared Error (MSE) and Mean Absolute Error (MAE). R-square is calculated as follows:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{3.1}$$

where $\hat{y}$ is the predicted target value and $\bar{y}$ is the mean observed target value.

To create a graphical representation of the results, we also present a binned scatterplot of predicted haircut value bands against actual haircut values, where predicted haircut values are put into ascending order and binned into equal-frequency value bands; the mean actual haircut value is then compared against the mean predicted haircut value in each haircut band.

3.4.4 Model results

First, we note that all parameters for all models have low VIF values, the only ones above 2 belonging to geographical indicators (see Appendix A, Tables A4 and A5). In the Haircut Model, the combination of LTV and time on books seems to be able to capture the information carried in DLTV because, as it is observed from Table 3.3, Model *H1* gives the better performance. This could

be because LTV gives an indication of the (initial) quality of the customer whereas values of DLTV could be due to changes in house prices since the purchase of the property. Based on this, Model *H1* is selected as the Haircut Model to be used in the LGD estimation.

Table 3.3: Haircut model performance statistics

Model	MSE	MAE	R^2
H1, Test Set	0.039	0.147	0.143
H2, Test Set	0.039	0.148	0.131

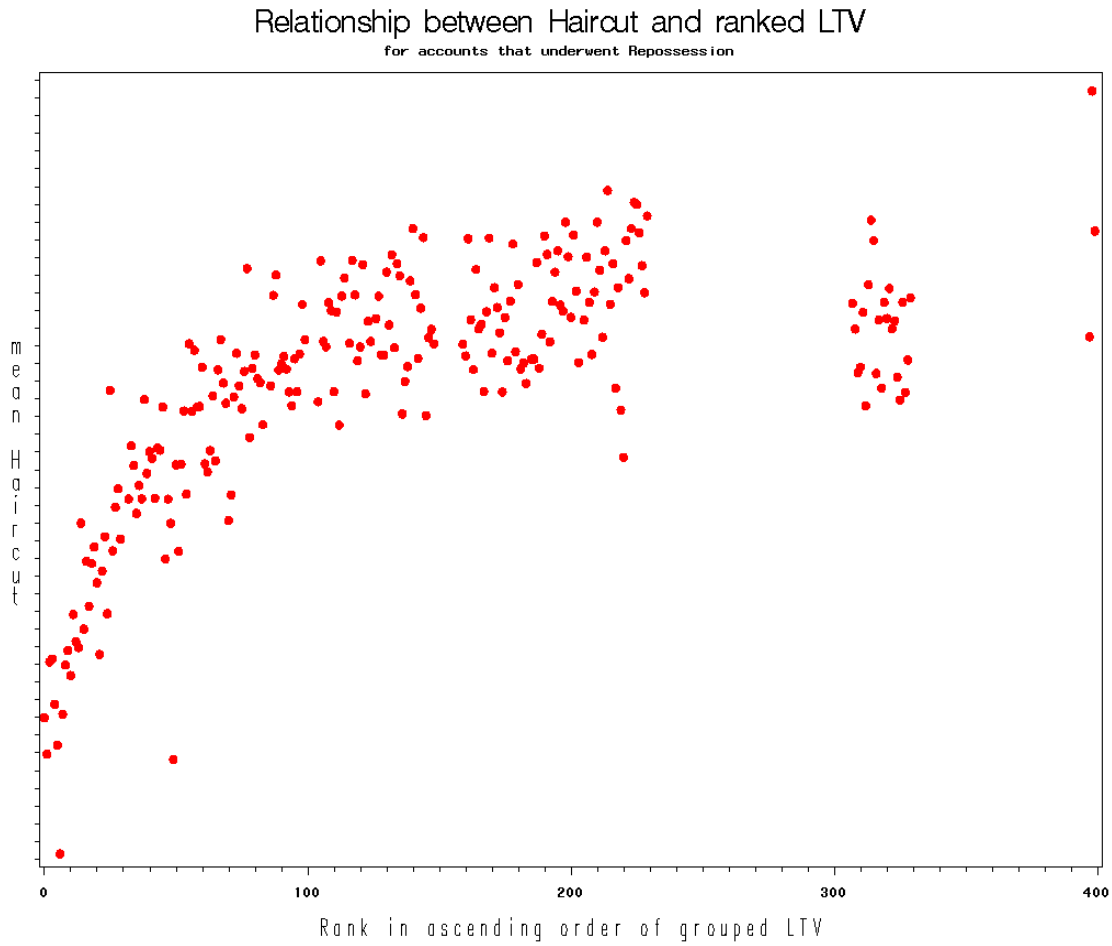

Figure 3.3: Relationship between haircut and (ranked) LTV at time of loan application[10]

[10] The gaps observed between ranked LTV groups 250 to 300, and again at 350 to 400, are due to the large number of observations that have very similar LTV values, which has caused a number of observations to have tied ranks.

Table 3.4: Parameter estimate signs of Haircut Model *H1*

Variable	Relation to haircut	Explanation
LTV	+	Refer to Figure 3.3 and explanation in Section 3.4.4
Ratio of valuation of security at default to average property valuation in that region, binned	+/-	Medium-end properties (relative to the region the property is in) have higher haircut than lower-end properties, but higher-end properties tend to have lowest haircut (cf. Figure 3.2 in Section 3.4.1)
Previous default	+	Haircut is higher for accounts that have previously defaulted
TOB (Time on book in years)	+	Older loans imply greater uncertainty and error in estimation of value of security at default, so higher haircut is possible
Security	+	Haircut tends to be higher for higher-end property types (e.g. detached)
Age group of property (oldest to newest)	+	Haircut tends to be higher for newer properties
Region	N/A	Haircut differs across regions

Table 3.4 details parameter estimate signs. From it, we see that a greater LTV at start implies a higher haircut (i.e. a higher forced sale price). This would mean that the larger the loan a debtor took at time of application in relation to property value, the higher the forced sale price of the security would be in the event of a default and repossession. At first, it might seem as though this parameter estimate sign might be confused due to the number of variables in the Haircut Model, or due to some hidden correlation between variables. In order to rule out this possibility, we look at the relationship between LTV at start and haircut. From Figure 3.3, we observe that there indeed appears to be a positive relationship between haircut and LTV. An explanation for this might be found in policy decisions taken by the bank. For loans with high LTV, due to the large amount the bank has committed towards the property, when the account does go into default and subsequent

repossession, the bank may be reluctant to let the repossessed property go unless it is able to fetch a price close to the current property valuation.

To further validate the model, we also include in Figure 3.4 a scatterplot of mean (grouped) predicted and actual haircut. From it, we observe that our model produces unbiased estimates of haircut. Parameter estimates of all models can be found in Appendix A, Table A4 and A5.

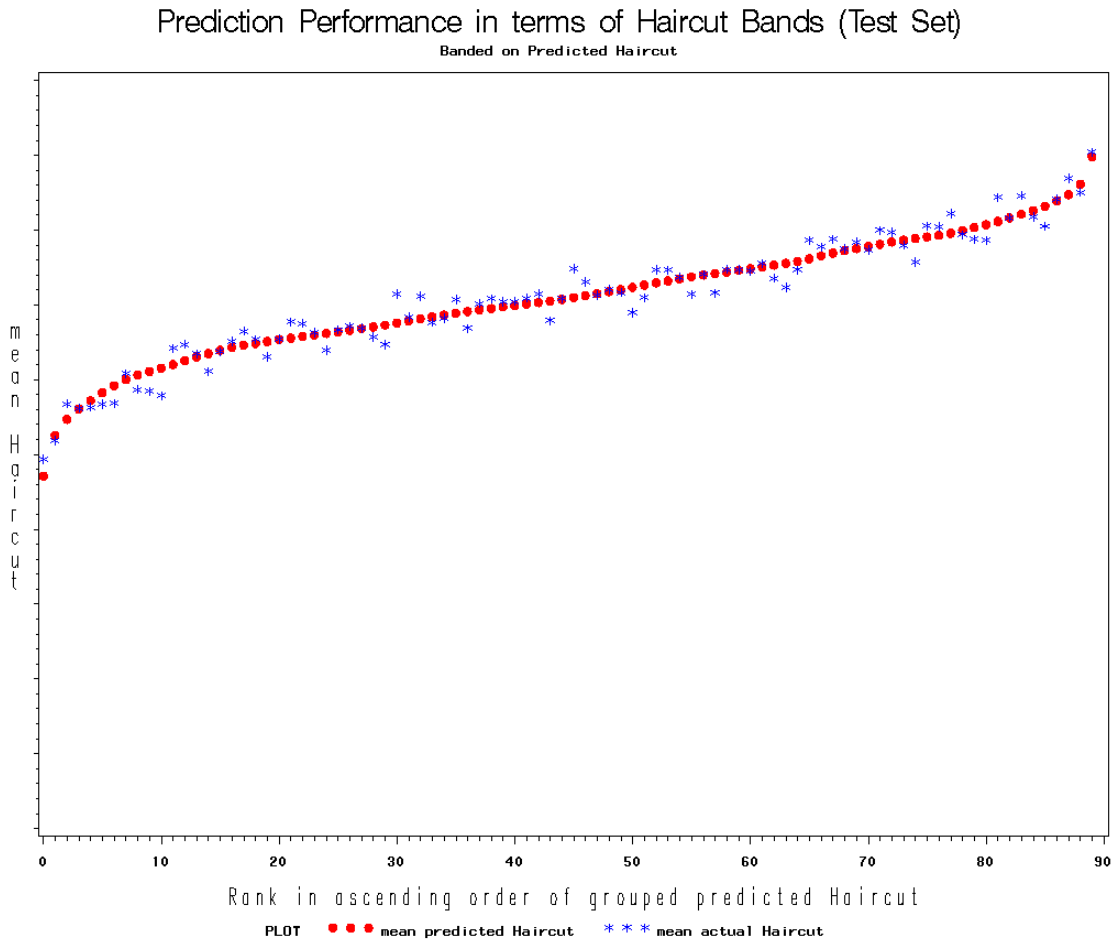

Figure 3.4: Prediction performance of haircut model (for test set). The stars represent the mean actual haircut and the solid dots represent the mean predicted haircut, in each haircut band.

3.4.5 Haircut standard deviation modelling

To be able to produce an expected value for LGD (see later, Section 3.5.1), we will not only require a point estimate for haircut but also a model component for haircut variability. Further inspection reveals that the

standard deviation of haircut increases with longer time on books (cf. Figure 3.5), which can be expected because the valuation of a property is usually updated using publicly available house price indices (instead of commissioning a new valuation process), and the longer an account has been on the books, the greater the uncertainty and error in the estimation of current valuation of property, which will affect the error in the prediction of haircut as well.

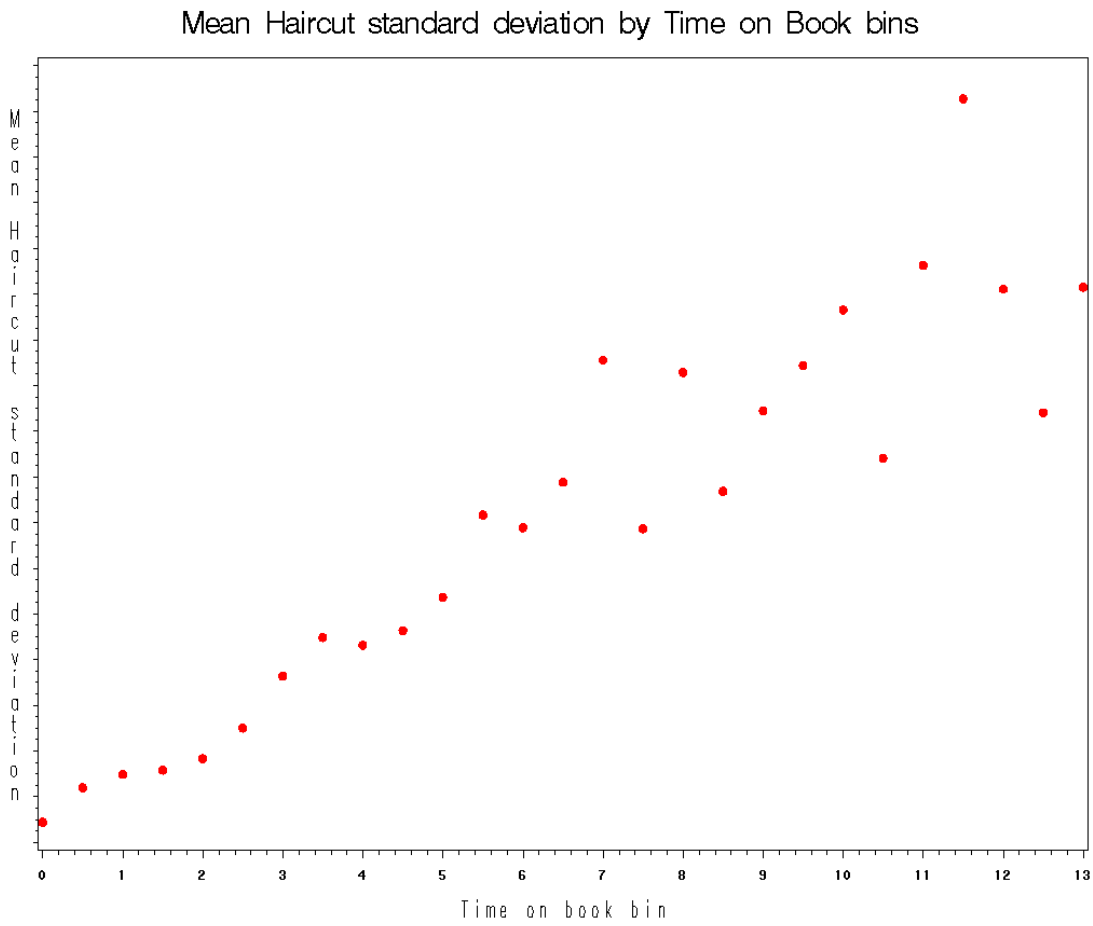

Figure 3.5: Mean haircut standard deviation by time on book bins

Table 3.5: Haircut Standard Deviation Model performance statistics

Model	MSE	MAE	R^2
Training Set	0.0001	0.0046	0.9315
Test Set	0.0002	0.0105	0.8304

As suggested by Lucas (2006), to model this relationship, a simple OLS model was fitted that estimates the standard deviation for different time on

books bins[11]. Time on books is binned into equal-length intervals of 6 months, and standard deviation of haircut is calculated for each group based on the mean haircut in that group. This model will later on be used to calculate the expected values for LGD (cf. Section 3.5.1). Performance statistics for this auxiliary model are detailed in Table 3.5; parameter estimates can be found in Appendix A, Table A6.

3.5 Loss Given Default Model

After having estimated the Probability of Repossession Model, the Haircut Model and the Haircut Standard Deviation Model, we now combine these models to get an estimate for Loss Given Default. Here we illustrate two ways of combining the component models, report their respective LGD predictions, and advocate use of the more conservative approach producing an expected value for LGD that takes into account haircut variability. We also compare these results against the single-stage model predictions and performance statistics.

3.5.1 Modelling methodology

A first approach referred to in this chapter as the "haircut point estimate" approach would be to keep the probabilities derived under the Probability of Repossession Model and apply the Haircut Model onto all observations. The latter would give all observations a predicted haircut value in the event of repossession. Using this predicted value of haircut, predicted sale price and predicted loss (outstanding balance at default less sale proceeds), if any, can be calculated. We then find predicted LGD by multiplying the probability of

[11] Alternatively, because standard deviation of haircut is different for different groups of observations, the weighted least squares regression method was considered to adjust for heteroscedascity in the OLS model developed in Section 6.4. Two different weights were experimented with – the error term variance of each observation (from running an OLS model for haircut) and time on books. Both models produced similar parameter estimates to the selected haircut model, which suggests that the OLS model was able to produce robust parameter estimates even though the homoscedasticity assumption was violated. Also, because both models did not explicitly model and produce standard deviation of haircut, which is required in the calculation of expected LGD, a separate OLS model for standard deviation is necessary.

repossession with this predicted loss if repossession happens. Although this method does produce some estimate for LGD, regardless of whether the observation is predicted to enter repossession or not, it uses only a single value of haircut (although it is the most probable value). However, if the true haircut happens to be lower than predicted, sale proceeds would be overestimated, which would mean that a loss could still be incurred (provided that haircut falls below DLTV). This is an illustration of how misleading LGD predictions could be produced if the component models were not combined appropriately. Hence, to produce a true expected value for LGD, one should also take into account the distribution to the haircut estimate, and the associated effect on loss in its left tail.

Hence, the second and more conservative approach, suggested e.g. by Lucas (2006) and referred to here as the "expected shortfall" approach, also takes into account the probabilities of other values of haircut occurring, and the different levels of loss associated with these different levels. To do so, we first apply the Probability of Repossession Model to get an estimate of probability of repossession given that an account goes into default. We then apply the Haircut Model onto the same dataset to get an estimate for haircut, $\hat{H}_j$, for each observation *j*, regardless of whether the security is likely to be repossessed. A minimum value of zero is set for predicted haircut, as there is no meaning to a negative haircut. The Haircut Standard Deviation Model is then applied onto each observation *j* to get a predicted haircut standard deviation, σ_j, depending on its value of time on books (see Section 3.4.5). From these predicted values, we approximate the distribution of each predicted haircut by a normal distribution, $h_j \sim N\left(\hat{H}_j, \sigma^2\right)$.

For simplicity, the subscript *j*, which represents individual observations, will be dropped from here on.

As long as the haircut (sale amount as a ratio of valuation of property at default) is greater than DLTV (outstanding balance at default as a ratio of valuation of property at default), and ignoring any additional administrative and repossession-associated costs, the proceeds from the sale would be able

to cover the outstanding balance on the loan, i.e. there would be no shortfall. Hence, the expected shortfall expressed as a proportion of the indexed valuation of property is:

$$E(\text{shortfall percent} \mid \text{repossession}) = \int_{-\infty}^{\text{DLTV}} p(h)(\text{DLTV} - h)dh \tag{3.2}$$

where *p(.)* denotes the probability density function of the distribution for *h*.

To convert the latter into a standard normal distribution, we let:

$$z = \frac{h - \hat{H}}{\sigma} \sim N(0,1);\ D = \frac{\text{DLTV} - \hat{H}}{\sigma}$$

Hence, expected shortfall can easily be derived as follows:

$$\begin{aligned} E(\text{shortfall percent} \mid \text{repossession}) &= \int_{-\infty}^{D} p(z)(D - z)\sigma dz \\ &= \left(\sigma D \int_{-\infty}^{D} p(z)dz\right) - \left(\sigma \int_{-\infty}^{D} p(z)zdz\right) \\ &= \sigma D\,\text{CDF}_Z(D) - \sigma(-\text{PDF}_Z(D)) \end{aligned} \tag{3.3}$$

where $\text{CDF}_Z(D)$ and $\text{PDF}_Z(D)$ denote the cumulative distribution function and probability density function of the standard normal distribution, respectively.

Expected loss given repossession is then obtained from the probability of non-repossession and the expected shortfall calculated for the repossession scenario (cf. Equation 3.4 below). The probability of an account undergoing repossession given that it has gone into default is multiplied against the expected LGD the account would incur in the event of repossession. We also multiply the probability of an account not going into repossession against the expected LGD for non-repossessions (denoted by *c*). We can use the

average observed LGD for actual non-repossessions as the expected conditional LGD for non-repossessions.

$$E(\text{loss} \mid \text{default}) = \begin{bmatrix} E(\text{shortfall percent} \mid \text{repossession}) \\ \times \text{indexed valuation} \\ \times \text{P}(\text{repossession} \mid \text{default}) \end{bmatrix} + \left[c \times (1 - \text{P}(\text{repossession} \mid \text{default}))\right] \tag{3.4}$$

where c is the loss associated with non-repossessions (assumed to be 0 in the absence of additional information).

Finally, we obtain predicted LGD by taking the ratio $E(\text{loss} \mid \text{default})$ to (estimated) outstanding balance at default.

3.5.2 Alternative single-stage model

To be able to compare this two-stage model, we also developed a simple single-stage model using the same data. A backward stepwise selection on the same set of eligible variables used earlier in the two-stage model building was applied, and resulting model parameter estimates are added in Appendix A, Table A7. However, it is noted that whatever the results of the single-stage model, because it directly predicts LGD based on loan and collateral characteristics, it does not provide the same insight into the two different drivers (i.e. repossession risk and sale price haircut) of mortgage loss, and as such does not provide as rich a framework for stress testing.

The performance measures of this single-stage model are then compared against those of the preferred two-stage model developed in the previous section (i.e. using the expected shortfall approach), as well as the two-stage model that would result from the so-called "haircut point estimate" approach.

3.5.3 Model performance

Using the same performance measures as those used for the Haircut Model, we compare the MSE, MAE and R-square (given in Equation 3.1) values of our

two-stage and the single-stage models (cf. Table 3.6). It is observed that both two-stage model variations achieve a substantially better R-square of just under 0.27 (compared to 0.233 for the single-stage model) on the LGD Test set, which is competitive to other LGD models currently used in the industry.

Table 3.6: Performance measures of two-stage and single-stage LGD Models

Method, Dataset	MSE	MAE	R^2
Single stage, Test set	0.026	0.121	0.233
Two-stage (haircut point estimate), Test set	0.025	0.108	0.268
Two-stage (expected shortfall), Test set	0.025	0.101	0.266

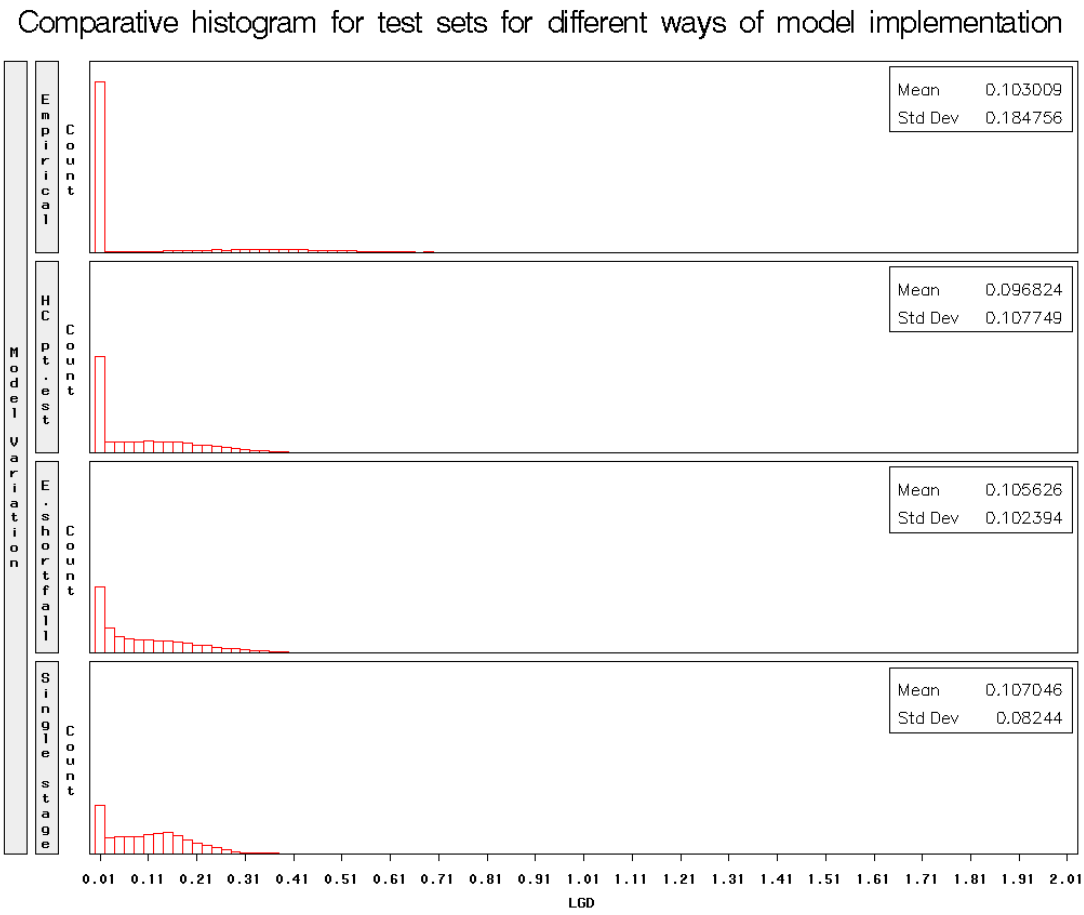

Figure 3.6: Distribution of observed LGD (Empirical), predicted LGD from two-stage haircut point estimate model (HC pt. est), two-stage expected shortfall model (E.shortfall), single-stage model (single stage) (from top to bottom). These are results from test sets.

The distributions of predicted LGD and actual LGD for all LGD models (test sets) are shown in Figure 3.6. In the original empirical distribution of LGD (see top section of Figure 3.6), there is a large peak near 0 (where losses were zero either because there was no repossession, or because the sale of the house was able to cover the remaining loan amount). Firstly, we observe that the single-stage model (shown in the bottom section of Figure 3.6) is unable to produce the peak near 0. Moreover, note that the two-stage model using the haircut point estimate seemingly is the model that most closely reproduces the empirical distribution of LGD, as it is able to bring out the peak near 0. Although the R-square value achieved by the two two-stage LGD model variations are very close (see Table 3.6), their LGD distributions are quite different. However, the haircut point estimate approach does so at the expense of underestimating the average loss (cf. mean predicted LGD from haircut point estimate method being lower than mean actual LGD). Unlike the former approach, the expected shortfall method takes a more conservative approach in its estimation of LGD, which takes into account the haircut distribution and its effect on expected loss based on probabilities of different haircut values occurring. This will make a difference especially for observations that would be predicted to have low or zero LGD under the haircut point estimate method because these very accounts are now assigned at least some expected loss amount, hence moving observations out of the peak and into the low LGD bins.

To further verify to what extent these various models are able to produce unbiased estimates at an LGD loan pool level, we create a graphical representation of the results (from the test sets). We look at a binned scatterplot of predicted LGD value bands against actual LGD values, where predicted LGD values are put into ascending order and binned into equal-frequency value bands. For each method we used in the calculation of LGD, we plot the mean actual LGD value against the mean predicted LGD value (for that LGD band) onto a single graph, included in Figure 3.7. Observe that both of the two-stage models are able to consistently estimate LGD fairly closely, whereas the single-stage model either overestimates or underestimates LGD (i.e. estimates are further removed from the diagonal). Furthermore, the expected shortfall approach is shown to produce the more reliable estimates in the lower-LGD regions, outperforming the haircut point

estimate approach in the lower-left part of the graph, where the haircut point estimate approach indeed underestimates risk.

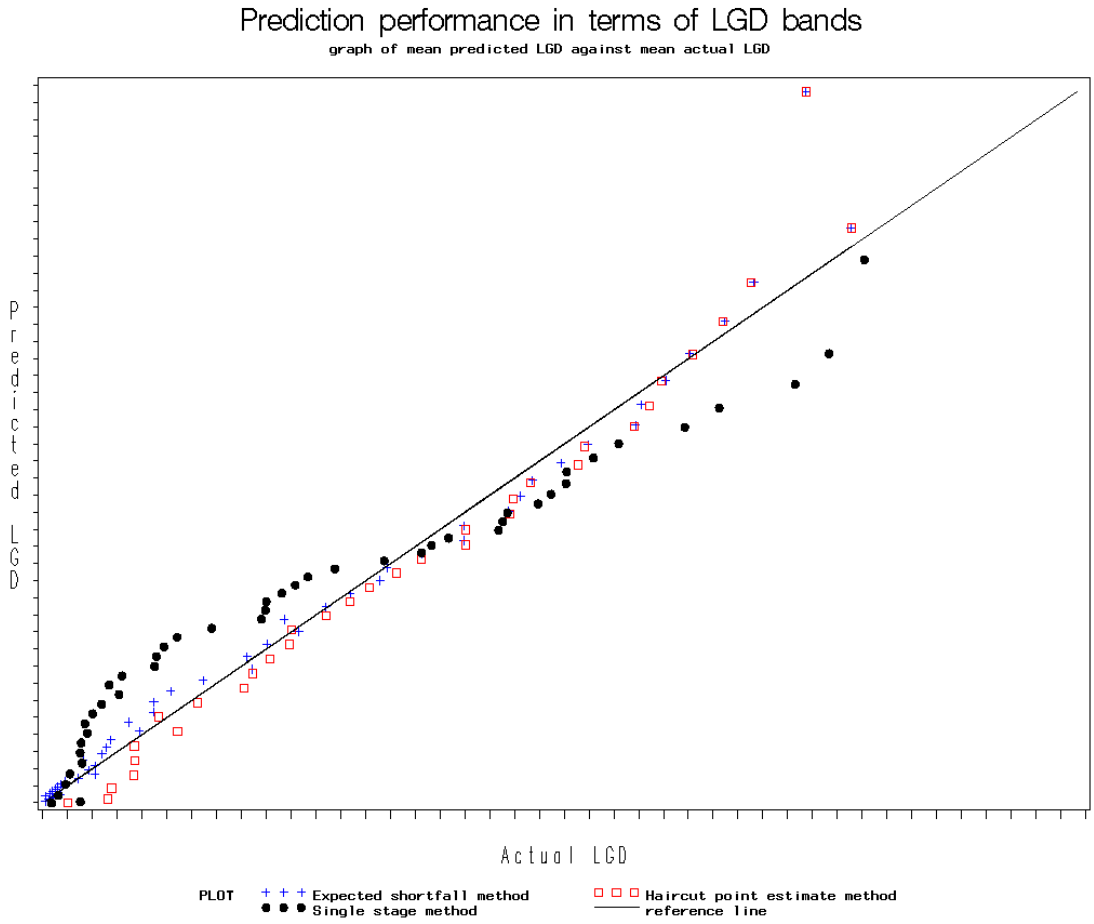

Figure 3.7: Scatterplot of predicted and actual LGD in LGD bands. The solid diagonal line represents perfect prediction, the squares represent predictions from the two-stage haircut point estimate method, the pluses represent predictions from the two-stage expected shortfall method and the solid dots represent predictions from the single-stage method. These are results from test sets.

Finally, in order to check robustness of our two-stage LGD model, we have also experimented with re-estimating the two component models this time including only the first instance of default for customers with multiple defaults (i.e. not all instances of default included for observations with multiple defaults). For both component models, we obtained the same parameter estimate signs, and parameter estimates were similar in size (see Tables A8 and A9 in Appendix A).

3.6 Conclusions

In this chapter, we developed and validated a number of models to estimate the LGD of mortgage loans using a large set of recovery data of residential mortgage defaults from a major UK bank. The objectives of this chapter were two-fold. Firstly, we aimed to evaluate the added value of a Probability of Repossession Model with more than just LTV at default as its explanatory variable. We have developed a Probability of Repossession Model with three variables, and showed that it is significantly better than a model with only the commonly used DLTV.

Secondly, we wanted to validate the approach of using two component models, a Probability of Repossession Model and a Haircut Model, which consists of the Haircut Model itself and the Haircut Standard Deviation Model, to create a model that produces estimates for LGD. Here, two methods are explained, both of which will produce a value of predicted LGD for every default observation because the Haircut Model, which gives a predicted sale amount and predicted shortfall, shall be applied to all observations regardless of its probability of repossession. However, we then show how the first method, which uses only the haircut point estimate would end up underestimating LGD predictions. The second and preferred method (expected shortfall) derives expected loss from an estimated normal haircut distribution having the predicted haircut from the Haircut Model as the mean, and the standard deviation obtained from the Haircut Standard Deviation Model.

For comparison purposes, we also developed a single-stage model. This model produced a lower R-square value, and was also unable to fully emulate the actual distribution of LGD.

CHAPTER 4. THE ECONOMY AND RETAIL LGD

In the previous chapter, a two-stage modelling approach for mortgage loan Loss Given Default (LGD) was outlined, and it was empirically shown that the LGD model worked well on real-life data. The following literature review highlights the increasing significance of the macroeconomy on both Probability of Default (PD) models and LGD models in the corporate sector, so we are inclined to expect similar results for risk models in the retail sector. Together with another retail loan dataset on personal loans that is also available, we use the mortgage loan dataset, and their respective LGD models, and extend the modelling to include relevant macroeconomic variables that might improve the predictions of LGD.

4.1 Literature review

In the research literature, credit risk models estimating PD and/or LGD, or its complement recovery rate (RR) tend to be more developed for corporate lending, partly due to the availability and ease of access of public data. The majority of credit models before and during the nineties assumed LGD to be independent of PD, estimating LGD using the average of historically observed LGD, with variability of losses depending only on PD. Since then, LGD models have been recalibrated to include explanatory variables like seniority of loan on the capital structure, the industry the firm is in, the collateral committed towards the loan, and in recent years, some variable representing the state of the economy at the time of default Schuermann (2004). In this literature review, we discuss a number of papers supporting this, before moving on to that part of the literature that specifically documents similar developments within the retail sector.

4.1.1 Overview of corporate credit risk models

In the earliest credit risk models (Black and Scholes (1973), Merton (1974)), because all components of credit risk including both PD and LGD were assumed to be a function of risk and asset elements of a firm, PD and LGD

were implicitly assumed to be correlated. However, perhaps because the Merton model was not as predictive as expected (Jones, Mason and Rosenfeld (1984)), credit models started to treat PD and LGD as independent variables, and it was not until the late 1990s that the relationship between PD and LGD was closely examined again. Besides credit models treating PD and RR as separate and independent variables, some modelled RR to be the average of historically observed losses (Asarnow and Edwards (1995)), which might be further segmented on industry or seniority (Altman and Kishore (1996)), or type of loan (Hamilton, Gupton and Berthault (2001)). More detail about how credit risk models evolved through the years can be found in Altman (2006).

The introduction of Basel II regulations in the late 1990s shifted credit models towards Credit Value at Risk (VaR) models which aim at predicting distributions of loss in order to estimate expected and unexpected losses. These models, including JP Morgan's CreditMetrics[12] and CreditVaR[13], KMV[14]'s Expected Default Frequency (EDF), Credit Suisse Financial Products' CreditRisk+[15] and McKinsey's CreditPortfolioView[16], were implemented in the industry with RR either being treated as a deterministic variable or constant (Crouhy, Galai and Mark (2000)). Also, only CreditPortfolioView acknowledges the effect the economy might have on default risk and risk grade migration beyond that of interest rates (Crouhy, Galai and Mark (2000)).

4.1.2 The economy, PD and LGD

From 2000 onward, corporate credit models not only started to pay attention to the relationship that RR might have with PD but also how they might have some form of dependence on the economy. Firstly, in Frye (2000a) and Frye (2000b), it was observed that recoveries differ according to when in the economic cycle the default happened, and he deduced that the PD and value of collateral (and hence recovery) are both dependent on the economy, and that they move in opposite directions (for example, an economic downturn

[12] CreditMetrics is a trademark of JP Morgan.
[13] CreditVaR is CIBC's proprietary credit value at risk model.
[14] KMV is a trademark of KMV Corporation.
[15] CreditRisk+ is a trademark of Credit Suisse Financial Products (CSFP), described in CSFP's publication (Credit Suisse (1997)).
[16] CreditPortfolioView is a risk measurement model developed by Wilson (1998).

would increase default frequencies and reduce the amount recovered from repossession of collateral). Jarrow (2001) comes to similar conclusions and introduces a recovery model incorporating debt and equity. A number of studies including Altman, Resti and Sironi (2001) and Hu and Perraudin (2002) have since confirmed the negative relationship between PD and LGD.

Gupton and Stein (2001) observed that RR falls when frequency of defaults in the market increases and have since expanded and developed their work into the LGD model we now know as LossCalc. According to Gupton and Stein (2002) and Gupton and Stein (2005), predictive variables of recovery are type of collateral, type of debt and seniority, type of firm, and industry and macroeconomic conditions (geographic location and distance to default aggregated at regional and industry level). From these variables, they are able to develop a robust and predictive LGD model which can be applied to a number of defaulted instruments, and which is consistently able to outperform LGD models that rely on historical averages. The study by Acharya, Bharath and Srinivasan (2003) on defaulted securities finds similar significant determinants of RR, and that whether the industry is in distress at time of default is a robust and important indicator of recovery, which emphasizes the relevance of the macroeconomy in the analysis of recovery rates.

However, it is the conclusions of Altman et al. (2005) and Pesaran et al. (2006) that confirm what has always been suspected. From their benchmarking study, Altman et al. (2005) conclude that while no single macroeconomic variable in recovery models is adequate, recovery models do benefit statistically from the inclusion of a variable which represents the macroeconomy. Pesaran et al. (2006) show that economic movements anywhere in the world do have an effect on individual firms' portfolio losses, but that these are not necessarily proportional to the macroeconomic changes themselves. More recently, Figlewski, Frydman and Liang (2008) undertook an extensive study in which they considered a broad range of macroeconomic variables in the prediction of hazard rates for different credit events. They found that macroeconomic variables do appear statistically significant at times (in their default intensity models), but that they might also be unstable and unreliable when used for certain models (i.e. most of

the risk grade migration models), thus re-emphasizing the complexity of incorporating macroeconomic indicators and the need for more work to be carried out in this area.

4.1.3 The retail sector: LGD and the economy

In Chapter 3, a literature review was included on LGD models for mortgage lending, which showed that the reported LGD models for retail lending are not as developed as those in the corporate sector, mainly due to the unavailability of public data. This is also reported in Allen, DeLong and Saunders (2004). We note a number of issues here.

Firstly, the retail sector work that has been reported on mortgage lending focuses on probability of default (e.g. von Furstenberg (1969), Avery et al. (1996) and Wong et al. (2004)). Some acknowledge the volatility of loss, but do so only alongside default models.

Pennington-Cross (2003), using data from Fannie Mae and Freddie Mac, finds that mortgage loan defaults are affected, amongst other loan characteristics, by economic conditions, but that the effect differs depending on whether the mortgage loan is prime or subprime. Recovery rates are predicted using a simulation over time, but the main explanatory variables only consist of DLTV (binned) and its interaction with the loan being prime or subprime. Calem and LaCour-Little (2004) find DLTV to be one of the significant variables of LGD, and achieve an R-square of 0.25. Neither of these LGD models included any macroeconomic variable. Also, although models like those of Campbell and Dietrich (1983) and Pennington-Cross (2003) include a macroeconomic effect in their development of mortgage default models, there has not been a definitive study investigating the impact of the economy on mortgage lending credit models for LGD.

Secondly, there are not many fully developed and validated credit models in the retail sector documented in the academic literature, even fewer for those specifically pertaining to mortgage lending. Jokivuolle and Peura (2003) studied the connection between the variability of LGD and uncertainty in collateral value, and suggested that LGD is correlated to likelihood of default

and some other macroeconomic-related variables. Although the study was in reference to corporate debt, mortgage loans too have collateral put up with the loan, whose value changes over time. Hui et al. (2006) did an empirical study on a sample of mortgage loans in Hong Kong, and came to similar conclusions, i.e. that there is a correlation between collateral value and PD. In addition, they also show that amongst other variables, LTV and volatility of collateral value should play a part in the assessment of the risk of a retail portfolio.

Finally, where LGD models for other types of retail loans are concerned, Bellotti and Crook (2009b) carried out a similar analysis in which they incorporated macroeconomic variables into an LGD model for credit card loans. They find that macroeconomic variables do play a significant role in reducing the difference between predicted and observed LGD.

4.2 Research objectives

As shown in the literature review of corporate risk models, the macroeconomy does play a part in recovery rates, but this potential relation has not been expanded much upon in the work on retail credit risk models, particularly for residential mortgage and personal loans.

In this chapter, we aim to investigate the impact the economy has on the predictive power of retail lending credit models for LGD. Empirically, we have reason to believe that the economy affects mortgage recovery. Figure 4.1 is a graph of the rate of repossessions[17] observed in the mortgage loan dataset. It shows a substantial increase for defaults that occurred during the economic downturn that the UK experienced in the early 1990s, with mean observed LGD showing an obvious increase as well. Note that the increase in LGD seems dampened because of the non-repossessions that will be posting

[17] Due to data restrictions, we define rate of repossession in year *i* to be the proportion of total number of loans that went into default in year *i* and repossessed by 2003 (latest observation in dataset) to total number of defaults that took place in year *i*.

a zero loss[18]. Subsequently, in Figure 4.2, we show the mean LGD of each default year of the personal loans dataset, from which an (albeit less pronounced) increase in LGD corresponding to the economic downturn is again observed.

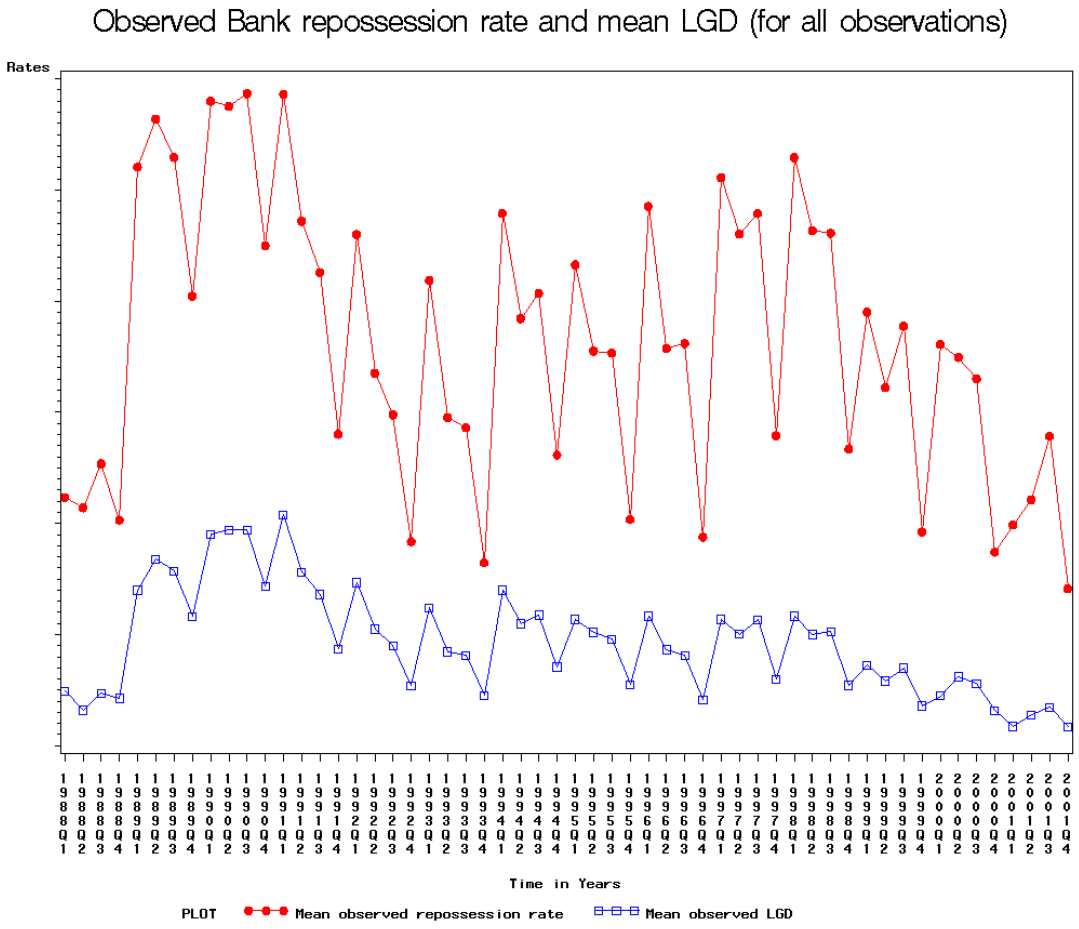

Figure 4.1[19]: Observed bank repossession rate and mean LGD across default years for mortgage loans dataset. The solid dots represent the bank's mean observed repossession rate and the squares represent the bank's mean observed LGD

There is some evidence that loss rates are affected by the economic conditions, but some of these effects have already been incorporated in the models. So the question now becomes: by adding variables related to the economy to the two LGD models and comparing their predictions of losses before and after the inclusion of macroeconomic variables, we hope to

[18] It is common to assume that there is zero loss (i.e. LGD is 0) if there is no repossession and/or sale of the repossessed property.
[19] Due to confidentiality issues, the scale of the vertical axis is omitted.

investigate if, for mortgage LGD, the current two-stage LGD model is already able to cope with these economic fluctuations, or if macroeconomic variables are able to improve the loss prediction further. Similarly, for personal loans LGD, we investigate whether the loan-related variables that were taken at different points of the loan term used in the OLS LGD model are already able to reflect changes in the economy.

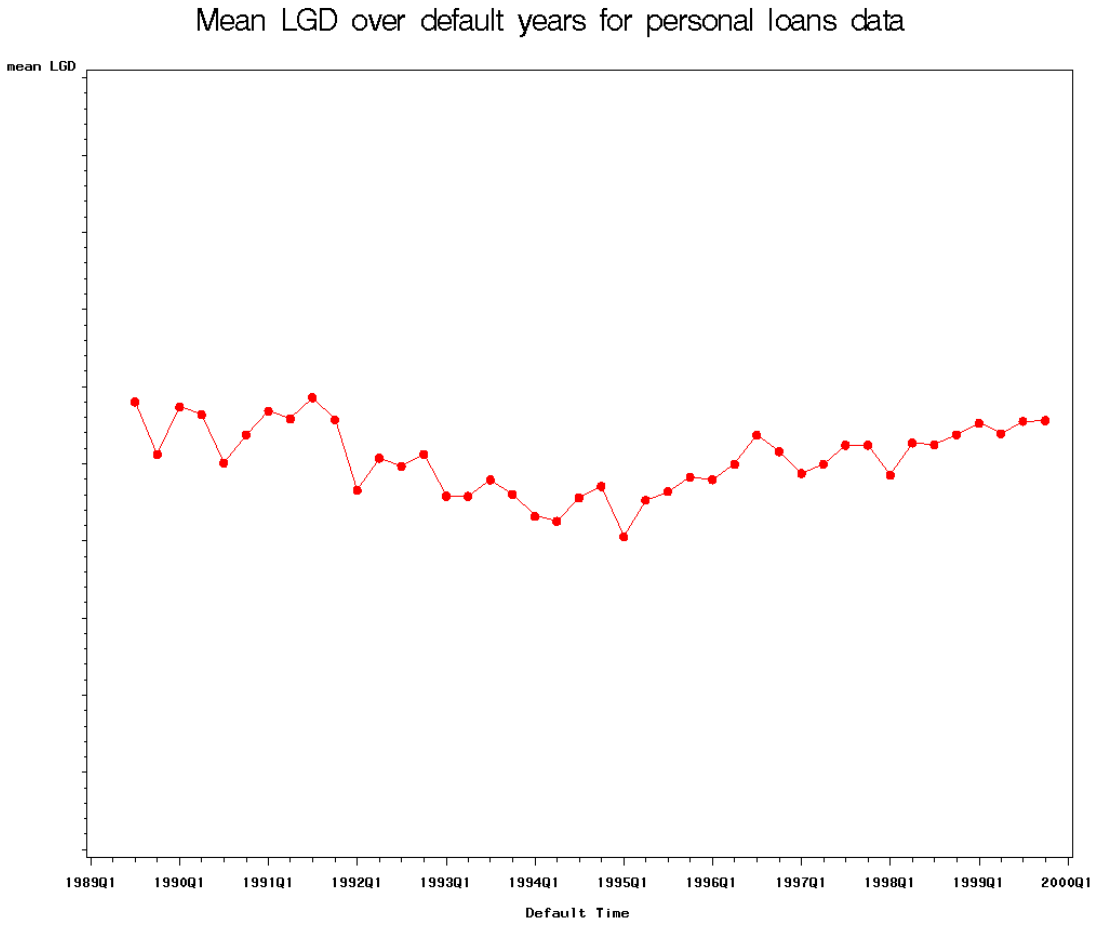

Figure 4.2: Mean LGD over default years for personal loans dataset.

4.3 Macroeconomic variables

Data of two different retail loans are used here, residential mortgage loans and unsecured personal loans, where observations in both datasets are of accounts that have already gone into default.

The UK experienced a major economic recession in the early 1990s, after which most aspects of the UK economy have experienced a steady and

significant growth. The datasets available here record defaults that happened between the late 1980s and the early 2000s, so they encompass these years. The next major recession happened in 2008, and although macroeconomic variables for this period are available and would be very interesting to include in our analysis, we unfortunately are unable to gain access to the corresponding bank data; hence we suggest the possibility of a future follow-up study.

The macroeconomic variables considered in this study are listed in Table 4.1 (graphs are attached in Appendix B). Lucas (2003) suggested the use of a few macroeconomic variables, i.e. change in interest rates, unemployment rates, asset prices and wage inflation, in a stress testing context, and this selection was expanded upon. All variables were sourced from the Office for National Statistics (ONS), and supplemented where necessary with data from the Council for Mortgage Lending (CML), Bank of England (BoE), and Halifax. With the exception of the variables related to the Halifax Housing Price Index, the rest of the selected variables have data available from the year 1976 through to 2007, which also explains the selection of BOE interest rates over the more directly relevant mortgage interest rates (which is only available from 1993). We select variables that are seasonally adjusted unless they are unavailable or not consistent with information from the source. Also, if the macroeconomic variables are expressed in terms of a monetary amount, we use growth rates such that the real change is reflected. It is inevitable that most of the macroeconomic variables are correlated with each other. However, removing the inappropriate variable at an early stage would mean the potential exclusion of valuable information. Hence, we employ stepwise selection, as explained later in Section 4.4.

For most of the macroeconomic variables, it is difficult to make any prior expectation of how they would specifically affect the two types of LGD (i.e. mortgage loan LGD and personal loan LGD), but we try to motivate the selection of each variable in Table 4.1. Things are made more complicated in the case of mortgage loans because its LGD is the product of two component models. Macroeconomic variables taken at loan origination could give an indication of the quality of customers, which could affect their probability of being repossessed should they go into default. For example, if net lending

growth at loan origination is high, it could mean that it was relatively easy for customers to secure their mortgage loan (hence lower-rated customers would also be accepted), which means that probability of repossession could increase should they later go into default. On the other hand, if net lending growth at default is high, it could mean that debtors would find it easier to obtain loans from other sources in order to keep up with repayments, and probability of repossession would decrease. If the HPIG at loan origination is high, the value of the property purchased might be inflated. In the event of default, because banks have invested a larger amount, they might be reluctant to sell the property unless they get a good price for it (i.e. haircut is high). If, on the other hand, there is high HPIG at default, the debtor is defaulting at a period of housing market growth, so banks are likely to be able to sell the property at a higher haircut.

Table 4.1: Macroeconomic variables considered in the analysis; variables are taken at two time points (where available), at start and default time.

Macroeconomic Variable	Source	Time Unit	Definition	Motivation
Net Lending Growth	ONS	Quarterly	Total consumer credit, net lending, seasonally adjusted, year on year quarterly change	This is an indication of how easy it is for consumers to get extra funds, and might give an indication of how difficult it is to secure a loan. Generally expect an increase to lead to a drop in LGD.
Disposable Income Growth	ONS	Quarterly	Real households' disposable income per head, seasonally adjusted, (constant 2003 prices), year on year quarterly change	This gives an indication of post-tax income, which might be a more accurate indicator of wealth than income itself. Generally expect an increase to lead to a drop in LGD.
GDP Growth	ONS	Quarterly	Gross Domestic Product, seasonally adjusted, (constant 2003 prices), year on year quarterly change	A common indicator of overall economic growth. Generally expect an increase to lead to a drop in LGD.
Purchasing Power Growth	ONS	Annually	Internal purchasing power of the pound (based on Retail Prices Index), not seasonally adjusted, (constant 2003	This gives an indication of wealth and inflation. Generally expect an increase to lead to a drop in LGD.

			prices)	
Net lending Growth on Dwellings	ONS	Quarterly	Banks net lending secured on dwellings, not seasonally adjusted, year on year quarterly change	This is an indication of how easy it is for consumers to secure a mortgage loan. Generally expect an increase to lead to a drop in LGD.
Unemployment Rate	ONS	Monthly	Unemployment rate, UK, All aged 16 and over, percentage, seasonally adjusted	An economic indicator for the job market. Generally expect an increase to lead to an increase in LGD as well.
Saving Ratio	ONS	Quarterly	Household saving ratio, seasonally adjusted	A reflection of level of saving to level of disposable household resources. Generally expect an increase to lead to a drop in LGD.
Interest Rate	BOE	Monthly	Bank of England interest rate, mean over the month	A pseudo indicator of cost of a mortgage loan. Generally expect an increase to lead to an increase in LGD.
House Price Index Growth (for mortgage loan dataset)	Halifax	Quarterly	All houses, all buyers, non seasonally adjusted, regional, year on year quarterly change	An economic indicator on the housing market. Generally expect an increase to lead to a drop in LGD.
House Price Index Growth (for personal loan dataset)	Halifax	Monthly	All houses, all buyers, non seasonally adjusted, year on year monthly change	An economic indicator on the housing market. Generally expect an increase to lead to a drop in LGD.

4.4 Modelling methodology

The LGD models for the mortgage and personal loans datasets have already been developed in the previous chapter, and other work, respectively, and are summarised next.

We refer to the LGD models already developed on each dataset as the base models. As far as possible, these models do not have any macroeconomic variables. After assigning the relevant macroeconomic indicators to each observation in each dataset, we find that correlations between the significant variables of each LGD model and the macroeconomic variables are no greater than | 0.6 |.

Because of the correlations amongst the macroeconomic variables themselves, there is limited value in including anything more than one macroeconomic variable (Campbell and Dietrich (1983) incorporate only local unemployment rates in the development of their mortgage loan default model; Hui et al. (2006) incorporates only the monthly private domestic price index in Hong Kong in their measurement for provision for risk under Basel II regulations). In order to assess the relevance and significance of each macroeconomic variable for each LGD model, we include the macroeconomic variables one at a time into a series of extended models. From each such resulting model, we extract the following pieces of information:

- Any variable in the extended or base model that becomes insignificant (has p-value greater than 0.01) [20], which might also be the macroeconomic variable itself; if there are any, the macroeconomic variable will not be considered any further.
- Any variable in the extended or base model with high Variance Inflation Factor[21,22] (only applicable to OLS models, i.e. the mortgage

[20] This excludes categorical variables as long as at least one group within the variable is significant.
[21] Refer to Chapter 3, Section 4.
[22] According to Fernandez (2007), VIF above 10 would imply severe multi-collinearity, while VIF less than 2 means that variables are almost independent.

loans haircut model and personal loans LGD model); if there are any, the macroeconomic variable will not be considered any further[23].

- The performance statistics of the model for both test and training sets (ROC for the mortgage loan probability of repossession model and R-square for the mortgage loans haircut model and personal loans LGD model)

4.4.1 Specific details for mortgage loans model

The data used in the development of the mortgage loan LGD model is supplied by a major UK bank, and is described in Chapter 2. Due to how the default date was recorded in this dataset, we chose to focus only on two time intervals per loan, i.e. the year the loan started and the year it went into default. Each observation will have a set of macroeconomic variables corresponding to that of its origination year, which would give an indication of the economic climate at the time the loan was approved, as well as act as some differentiating factor between lending practices at various points of the economic cycle as suggested in Breeden, Thomas and McDonald (2008). Secondly, the time of default would give an indication of the current state of economy. The inclusion of time lags and leads to the macroeconomic variables are also investigated.

4.4.2 Specific details for personal loans model

This data is supplied by a major UK bank. There are about 50,000 observations in the original dataset, all of which are on defaulted personal loans, where each observation describes an existing credit account and the debtor(s). These personal loans are unsecured, so there is no physical security that the bank could repossess in the case of a default. Observations with a significant number of missing values and LGD outliers were removed during pre-processing, and about 48,000 observations were retained. We have about 45 variables describing customer information collected at time of loan application, as well as information collected during the loan term and at

[23] This criterion is relaxed for categorical variables due to the correlation between some of the groups within the categorical variables that were originally present.

default. Accounts default between the years of 1989 and 1999. More details about this dataset can be found in Matuszyk, Mues and Thomas (2010).

In this dataset, only default date information is readily known, so the only time point of interest here would be default time (instead of both start and default time points of the loan). Again, lags and leads were investigated.

4.5 Mortgage loan LGD results

The development of the two-stage LGD model (base model for mortgage loans LGD) is described in Chapter 3. This consists of two component models: a probability of repossession model, which gives the probability of a defaulted loan going into repossession; and a haircut model, consisting of two further sub-models, one which estimates haircut itself and the other that produces the standard deviation of predicted haircut. During that analysis, we found that including the following three variables, LTV, DLTV and time on books, in any single model would cause counter-intuitive parameter estimate signs. As such, the suitability of incorporating either LTV and time on books, or DLTV, was investigated for each of the component models and one of either options was selected based on which gave the best overall performance measure (ROC for the Probability of Repossession Model; R-square for the Haircut Model).

4.5.1 Probability of Repossession Model

The final variables in the Probability of Repossession Model are DLTV, type of security and an indicator for whether the account has gone into default before. We note that this means that the model is not totally free of economic effects. The fact that DLTV is already one of the explanatory variables in the base model means that an updated value of the House Price Index (HPI) was already included in its estimation (see Chapter 3, Section 3). Still, we want to investigate if the inclusion of macroeconomic variables is able to improve LGD prediction any further. Hence, macroeconomic variables will be included, independently and separately, to the Probability of Repossession Model. The resulting test-set ROC values (only if variables

remain significant at the 99% level of confidence) are detailed in Table 4.2. Most of the macroeconomic variables taken from the start of the loan turn out to be insignificant and the variable that gives the best improvement in ROC is the interest rate at default (parameter estimates attached in Appendix C, Table C1), which is found to increase test set ROC from 0.743 to 0.758.

Table 4.2: Performance of Probability of Repossession Model (test sets) with macroeconomic variables

PROBABILITY OF REPOSSESSION MODEL			
Model	Additional Variable	Model Estimate	ROC (Test)
Base			0.743
Macroeconomic variables at origination			
Model	Additional Variable	Model Estimate	ROC (Test)
Base (DLTV) + MV 1	Net Lending Growth at origination	+	insignificant
Base (DLTV) + MV 2	Disposable Income Growth at origination	+	insignificant
Base (DLTV) + MV 3	GDP Growth at origination	+	insignificant
Base (DLTV) + MV 4	Purchasing Power Growth at origination	-	0.743
Base (DLTV) + MV 5	Net Lending Growth for Dwellings at origination	-	insignificant
Base (DLTV) + MV 6	Unemployment Rate at origination	-	0.743
Base (DLTV) + MV 7	Saving Ratio at origination	+	insignificant
Base (DLTV) + MV 8	Interest Rate at origination	+	insignificant
Base (DLTV) + MV 9	House Price Index Growth at origination	-	0.744
Macroeconomic variables at default			
Model	Additional Variable	Model Estimate	ROC (Test)
Base (DLTV) + MV 10	Net Lending Growth at default	-	0.744

Base (DLTV) + MV 11	Disposable Income Growth at default	-	0.745
Base (DLTV) + MV 12	GDP Growth at default	+	0.755
Base (DLTV) + MV 13	Purchasing Power Growth at default	-	0.757
Base (DLTV) + MV 14	Net Lending Growth for Dwellings at default	-	0.747
Base (DLTV) + MV 15	Unemployment Rate at default	-	0.754
Base (DLTV) + MV 16	Saving Ratio at default	-	0.752
Base (DLTV) + MV 17	Interest Rate at default	+	0.758
Base (DLTV) + MV 18	House Price Index Growth at default	+	insignificant

Interest rate at default is found to be positively related to probability of repossession. This is probably because an increase in interest rate would imply that borrowing has become more expensive (for the borrower), leading to increased repossession. However, we see that the macroeconomic variables on the whole do not particularly improve the ROC value, perhaps because HPI (which is the main macroeconomic indicator in the housing industry) is already embedded in the calculation of DLTV.

Because the results seem to suggest that macroeconomic variables at the time of origination are not as useful as those at time of default, we experiment only with a six-month lag and lead time from default. ROC values for the test set corresponding to these variations are slightly higher at 0.760 compared to 0.758 achieved by interest rates at default (see Tables D1 and D2 in Appendix D). Because the improvement is small, it was decided to use the macroeconomic variable corresponding to the time of default. Besides, time leads are less desirable because the goal here is prediction.

4.5.2 Haircut Model

The Haircut Model is made up of two sub-models: one which predicts the haircut, and the other which estimates the standard deviation of haircut

(details in Chapter 3, Section 4). The Haircut Standard Deviation Model is left unchanged because of the reasonably high R-square achieved (0.93 for the training set; 0.83 for the test set). Hence, it would seem that any contribution macroeconomic variables can give to predicted haircut would be in the prediction of haircut itself. The final variables in this sub-model for haircut are LTV, time on books, the ratio of current valuation of property to the average property value in that region, an indicator for whether the account has gone into default before, type of security, the age band of the property and region. Again, the HPI is to some extent already factored into the model through the calculations for the current valuation of property (updated from the original property value using the HPI at the start and default time of the loan) which is then divided by the average property value in that region to get average valuation ratio for the region (see Chapter 3, Section 4).

Table 4.3: Performance of Haircut Model (test sets) with macroeconomic variables

HAIRCUT MODEL			
Model	Additional Variable	Model Estimate	R^2 (Test)
Base (LTV)			0.143
Macroeconomic variables at origination			
Model	Additional Variable	Model Estimate	R^2 (Test)
Base (LTV) + MV 1	Net Lending Growth at origination	-	0.143
Base (LTV) + MV 2	Disposable Income Growth at origination	-	insignificant
Base (LTV) + MV 3	GDP Growth at origination	-	0.146
Base (LTV) + MV 4	Purchasing Power Growth at origination	+	0.146
Base (LTV) + MV 5	Net Lending Growth for Dwellings at origination	-	0.145
Base (LTV) + MV 6	Unemployment Rate at origination	-	0.147

Base (LTV) + MV 7	Saving Ratio at origination	+	insignificant
Base (LTV) + MV 8	Interest Rate at origination	-	0.149
Base (LTV) + MV 9	House Price Index Growth at origination	+	0.145
	Macroeconomic variables at default		
Model	Additional Variable	Model Estimate	R^2 (Test)
Base (LTV) + MV 10	Net Lending Growth at default	+	insignificant
Base (LTV) + MV 11	Disposable Income Growth at default	+	0.144
Base (LTV) + MV 12	GDP Growth at default	-	0.149
Base (LTV) + MV 13	Purchasing Power Growth at default	+	0.167
Base (LTV) + MV 14	Net Lending Growth for Dwellings at default	+	0.148
Base (LTV) + MV 15	Unemployment Rate at default	+	0.143
Base (LTV) + MV 16	Saving Ratio at default	+	0.144
Base (LTV) + MV 17	Interest Rate at default	-	0.170
Base (LTV) + MV 18	House Price Index Growth at default	+	0.148

Similarly to Section 4.5.1, we try adding macroeconomic variables independently and separately to the Haircut Model. The resulting models' R-square values (only if variables remain significant at the 99% level of confidence) are as summarised in Table 4.3. Although more macroeconomic variables taken at the start of the loan turn out to be significant, they do not substantially improve the R-square. Again, interest rate at time of default contributes most to the haircut model, increasing R-square from 0.143 to 0.170 (parameter estimates in Appendix C, Table C2). An increase in the interest rate at default implies that potential buyers may find it harder to find

an affordable loan; hence, the reduced demand may force banks to release repossessed properties at a less favourable price.

Again, time lags and leads from default were considered here. However, the inclusion of macroeconomic variables corresponding to a six-month lag or lead time from default did not result in the same improvement in R-square achieved by using the interest rate at default (see Tables D3 and D4 in Appendix D).

4.5.3 Two-stage LGD model

From the above two sub-sections, it was found that interest rates at default gave the best improvements to the component models. This variable was then included into the Probability of Repossession Model and the Haircut Model. Together with the standard deviation obtained from the Haircut Standard Deviation Model, estimates for LGD were produced according to the methodology described in Chapter 3, Section 5.1. Table 4.4 displays the performance statistics for the test sets of the two LGD models developed (the original (base) LGD model developed in Chapter 3, and the macroeconomic LGD model developed in this chapter).

Table 4.4: Performance statistics of mortgage loan LGD models (Test Sets)

Method, Dataset	R^2	MSE	MAE
Two-stage base model, Test	0.268	0.025	0.108
Two-stage macroeconomic model, Test	0.303	0.024	0.103

Within each component model, we see an improvement in performance measures, and this eventually translates to an improvement in R-square value in the overall LGD model. We highlight two observations. First, this improvement is noteworthy because, as it was noted earlier, the HPI was already embedded into both the component models. On top of this, the HPI is also used in the calculation of LGD (see Equation 3.4 in Chapter 3, Section 5.1), because the indexed valuation is involved in the calculation for expected shortfall (if default occurs), and this indexed valuation is updated using the original valuation and HPI at the start and default times of the loan.

Second, probably because HPI was already involved in the LGD base model, the variable HPI growth (HPIG), which is the main macroeconomic indicator of the housing market, itself only gave small improvements or appeared insignificant.

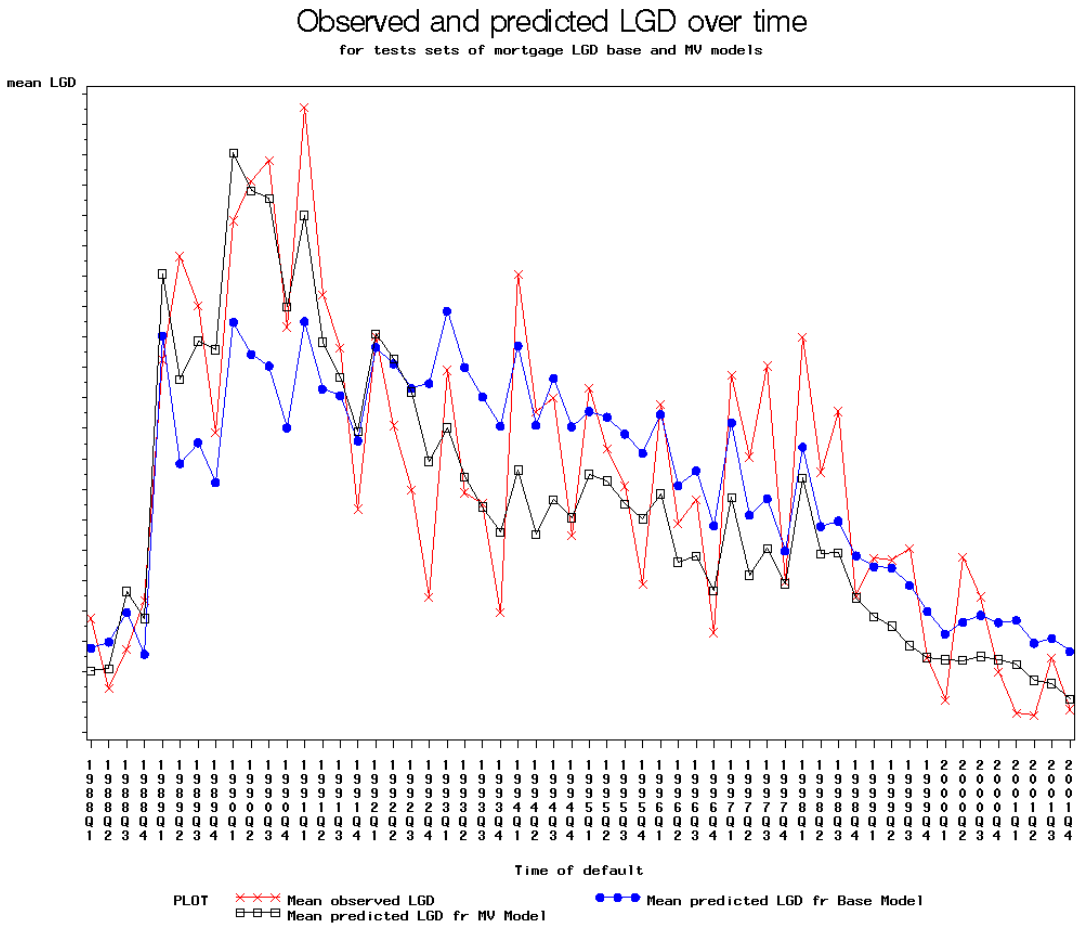

Figure 4.3: Mean observed and predicted mortgage loan LGD over each year of default for base and macroeconomic LGD models (Test sets). The crosses represent mean observed LGD, the solid dots represent mean predicted LGD from the base model and the squares represent mean predicted LGD from the macroeconomic LGD model.

Figure 4.3 puts the performance of the two two-stage LGD models on a single graph, from which we see that both give decent predictions but are unable to fully reflect the highs and lows that the observed LGD goes through in each year. Also, we see that the two models give better predictions for different time periods, which is more obvious in Figure 4.4. During the economic downturn the UK experienced in the early 1990s, the

base LGD model (represented by the solid dots) was underestimating LGD, whilst the macroeconomic LGD model (represented by the squares) was able to give a closer prediction. However, during the rest of the period observed (1993 onwards), the macroeconomic LGD model consistently gives a lower prediction than the base LGD model, i.e. underestimating LGD more frequently and by a further distance than the base LGD model. So, although the inclusion of macroeconomic variables improved model predictions during the period of downturn, it caused the model to underestimate predictions of LGD during periods outside of the downturn.

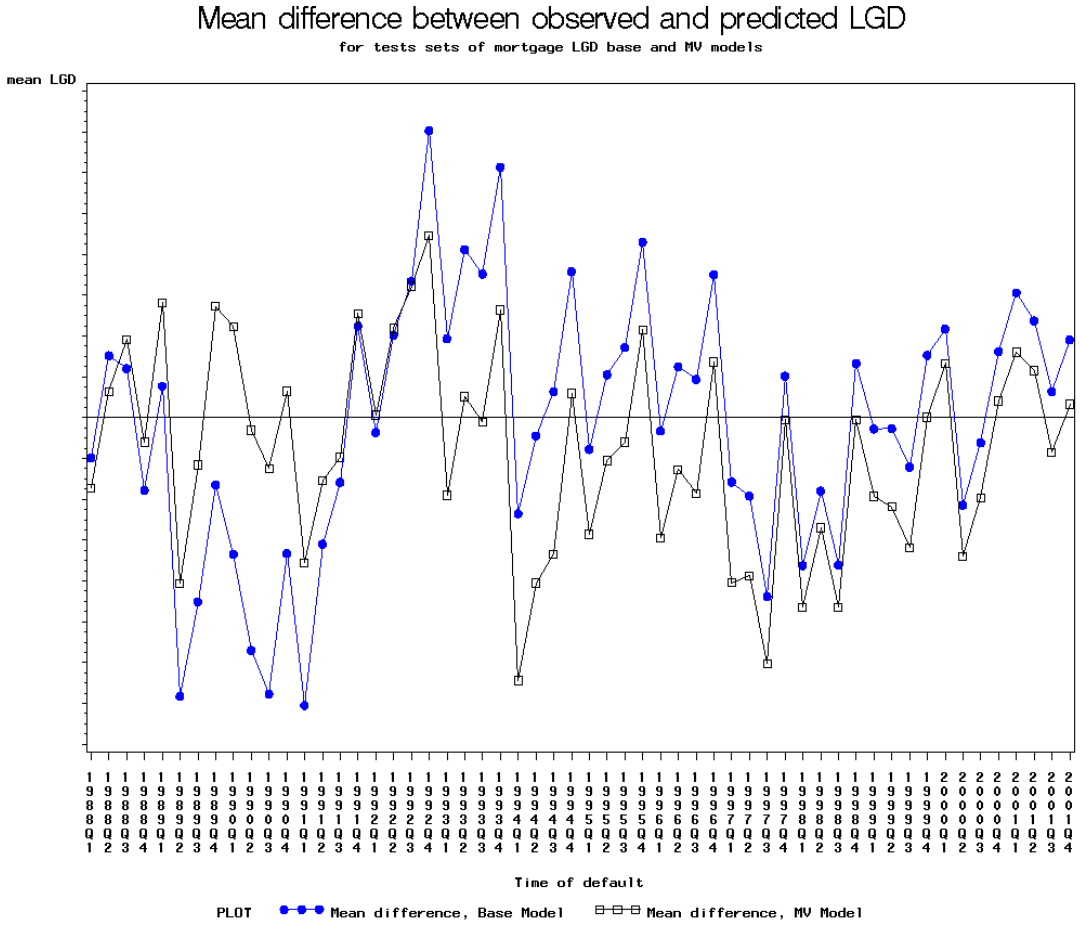

Figure 4.4: Mean difference between observed and predicted mortgage loan LGD, for base and macroeconomic LGD models (Test sets). The solid dots represent mean difference coming from the base model and the squares represent mean difference coming from the macroeconomic LGD model.

Consistent with the performance graphs produced in Chapter 3 (cf. Figures 3.6 and 3.7), we subsequently look at distribution of predicted LGD and prediction performance of the two models in terms of LGD bands. The

former is a graph which compares the distributions of predicted LGD estimated by the base and macroeconomic models (cf. Figure 4.5). The latter is a binned scatterplot of predicted LGD value bands against actual LGD values (cf. Figure 4.6).

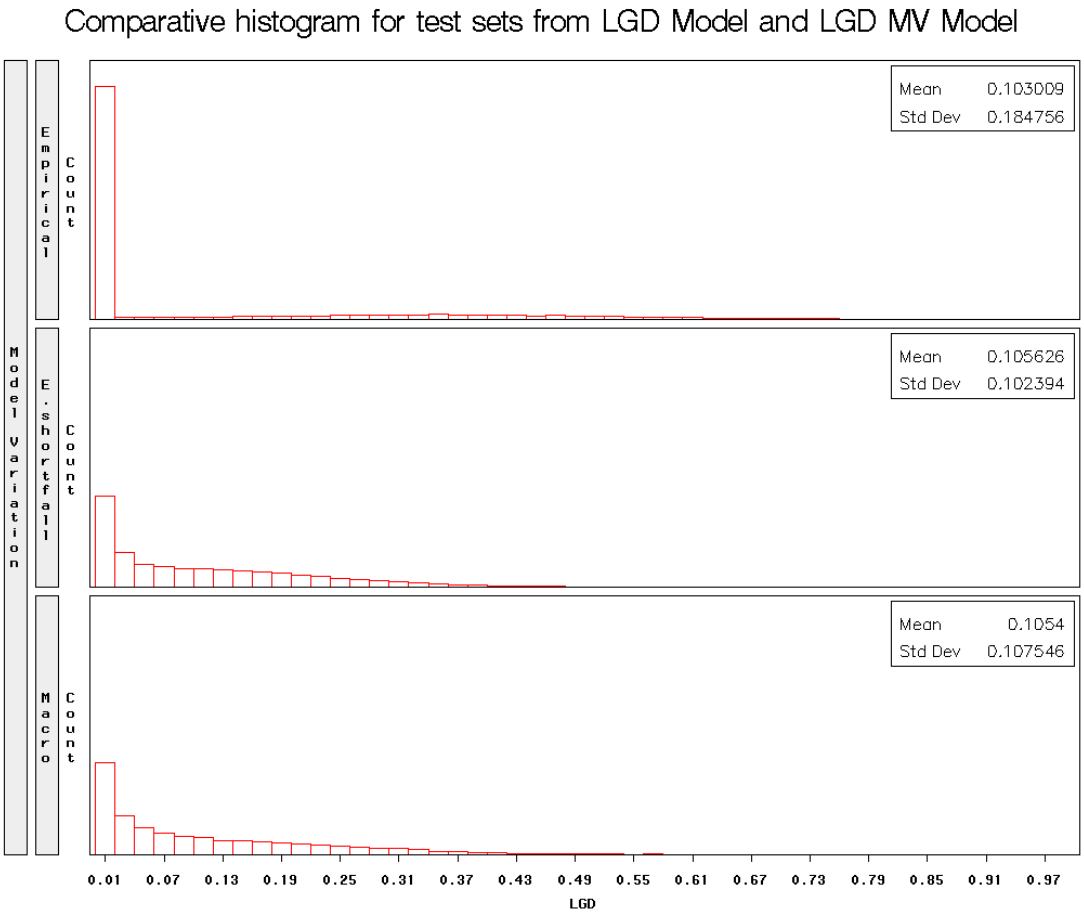

Figure 4.5: Distribution of observed LGD (Empirical), predicted LGD from two-stage expected shortfall base model (E.shortfall), two-stage expected shortfall macroeconomic model (macro) (from top to bottom). These are results from test sets.

As shown in Figure 4.5, recall that in the original empirical distribution of LGD (see top section), there is a large peak near 0 (where losses were zero either because there was no repossession, or because the sale of the house was able to cover the remaining amount). We see that the distributions of predicted LGD from the base model and the macroeconomic model are very similar. Both are able to bring out the peak near 0, but do not match the magnitude. We then examine a binned scatterplot shown in Figure 4.6,

where the mean actual LGD value is plotted against the mean predicted LGD value of the base and macroeconomic models (for that LGD band). Again, the base and macroeconomic models give very similar performances. Towards the larger values of LGD, we see that the macroeconomic model was able to give a closer mean predicted value than the base model. This is in line with our earlier observation that larger LGD values are associated with loans that go into default during downturn periods, and that the macroeconomic model was able to better predict for these accounts. However, in the lower LGD bands, this advantage no longer holds.

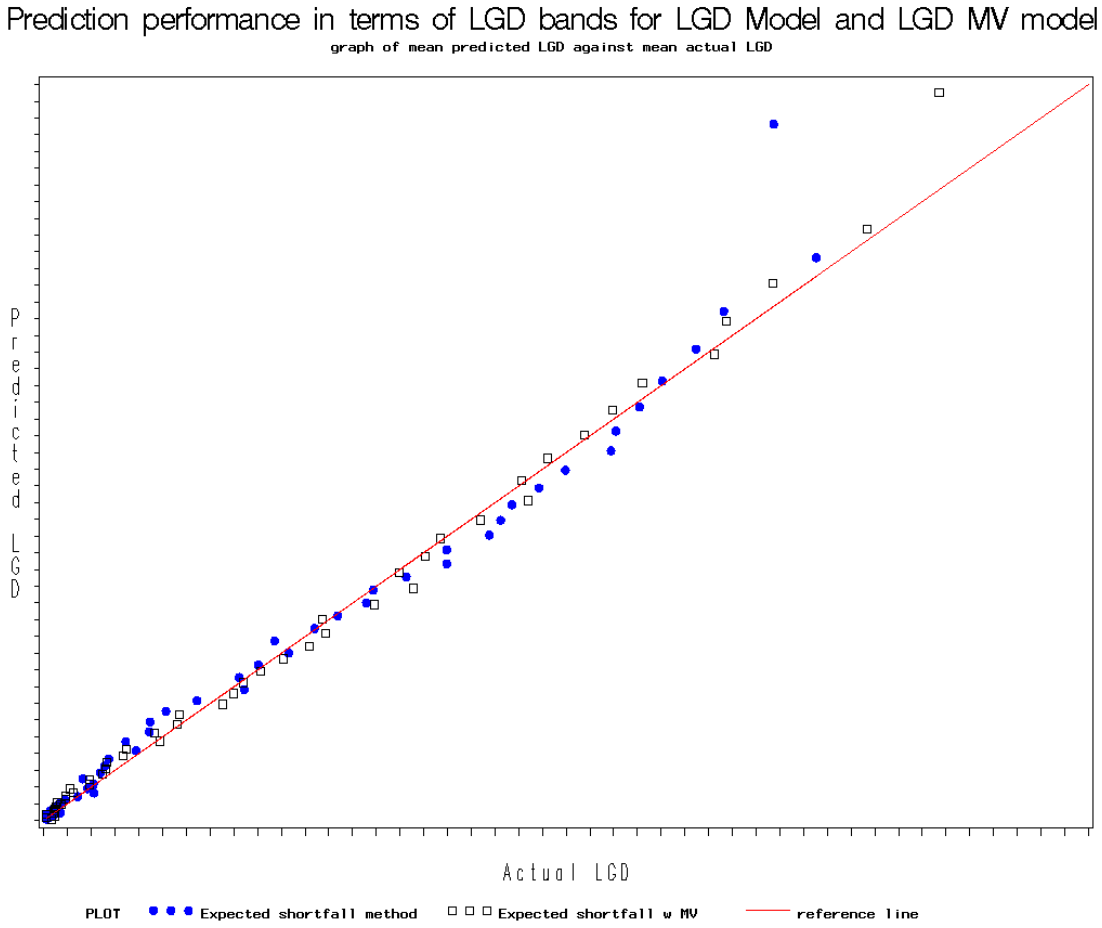

Figure 4.6: Scatterplot of (base and macroeconomic) predicted and actual LGD in LGD bands. The solid diagonal line represents perfect prediction, the solid dots represent predictions from the base LGD model and the squares represent predictions from the macroeconomic LGD model. These are results from test sets.

In summary, although the macroeconomic model was able to produce a higher R-square value, it did not seem to be able to consistently improve

prediction performance in terms of distribution of LGD and mean predictions. This could suggest that macroeconomic variables could be non-linearly related to recovery rates, so for example, a drop in interest rates from 5% to 4% would affect recovery rates differently than if interest rates went from 1.5% to 0.5%. Also, the effect could be different for different types of loans, so the inclusion of interaction terms could be considered in future work.

Bruche and González-Aguado (2010) suggest that default rates would be a good predictor of recovery rates, however, the internal bank default rate was unavailable, and although a general UK default rate was considered, this data is only available starting from 1993. We believe that it is likely the economy could have a more obvious impact on likelihood of default; for example, Campbell and Dietrich (1983) showed a statistically significant relationship between default and the economic situation, in particular local unemployment rates. However, it can be difficult to show a direct relationship between macroeconomic variables and recovery rates.

4.6 Unsecured personal loans LGD model results

An LGD model for the personal loans dataset has already been developed in earlier work by Loterman et al. (2009). One of the models developed was a linear regression model that directly models LGD using characteristics of defaulted loans. The variables used in the LGD model can be roughly divided into two groups – customer-related, which include the application score, employment and residential status of account holder at the start of the loan, as well as information about other loan commitments (for example, whether they have a mortgage, current or personal accounts); or loan-related characteristics such as the loan amount, the term and purpose of the loan, the length of time the loan has been at the bank, as well as whether the loan was ever in arrears and the extent of arrears, if any.

Table 4.5 displays the R-square values of the personal loans LGD base, as well as R-square values with each macroeconomic variable included, as measured at time of default. Considering that this dataset consists of defaulted unsecured personal loans, we expect LGD to be more likely

affected by general indicators of the economy than by industry-specific macroeconomic indicators like the House Price Index. Perhaps surprisingly though, the only macroeconomic variable that is found to be statistically significant is net lending growth at default. A potential explanation for the observed negative relationship could be that, when there is an increase in net lending at the time of default, the debtor might be more successful in securing funds from other sources to try and repay his loan, hence a lower LGD.

Table 4.5: Results of personal loans LGD base and macroeconomic models (test sets)

Model	Additional Variable	Model Estimate	p-value	R^2 (Test)
Base				0.073
PERSONAL LOANS LGD MODEL; Macroeconomic variables at default				
Base + MV 1	Net Lending Growth at default	-	<0.01	0.073
Base + MV 2	Disposable Income Growth at default	-	insignificant	0.073
Base + MV 3	GDP Growth at default	-	insignificant	0.073
Base + MV 4	Purchasing Power Growth at default	-	insignificant	0.073
Base + MV 5	Unemployment Rate at default	-	insignificant	0.073
Base + MV 6	Saving Ratio at default	+	insignificant	0.073
Base + MV 7	Interest Rate at default	+	insignificant	0.073
Base + MV 8	Net Lending Growth for Dwellings at default	+	insignificant	0.073
Base + MV 9	House Price Index Growth at default	+	insignificant	0.073
Base + MV 10	House Price Index at default	+	insignificant	0.073

In our model variations with six-month lags or leads, none of the macroeconomic variables were statistically significant (see Tables D5 and D6 in Appendix D).

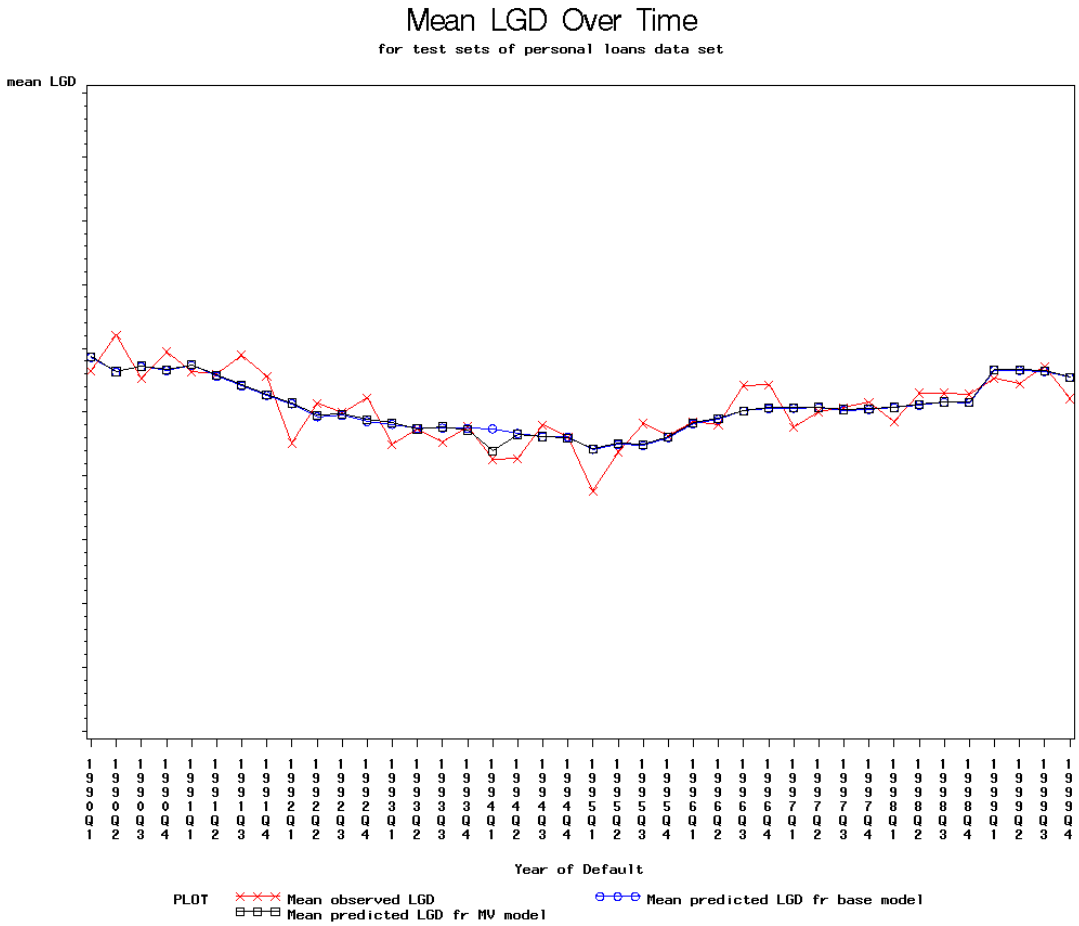

Figure 4.7: Mean observed and predicted personal loan LGD over each year of default for base and macroeconomic Models (Test Sets). The crosses represent the mean observed LGD, the circles represent predictions from the base LGD model and the squares represent predictions from the macroeconomic LGD model.

However, we note that, despite this variable being significant, there is no improvement in R-square value. We also observe in the dataset that the mean LGD is quite flat across the default years (cf. Figure 4.7), and that the mean LGD observed during the downturn years was not that different to non-downturn years. The predictions from the two LGD models are also very similar. It is possible that loans that went into default during downturn years have a longer recovery process but finally achieve similar LGD values to default loans from non-downturn years. However, in this dataset, we only

have information on LGD but not about how long each recovery process took. Figure 4.8 reinforces the observation that the predictions coming from the LGD macroeconomic model are very similar to those from the LGD base model.

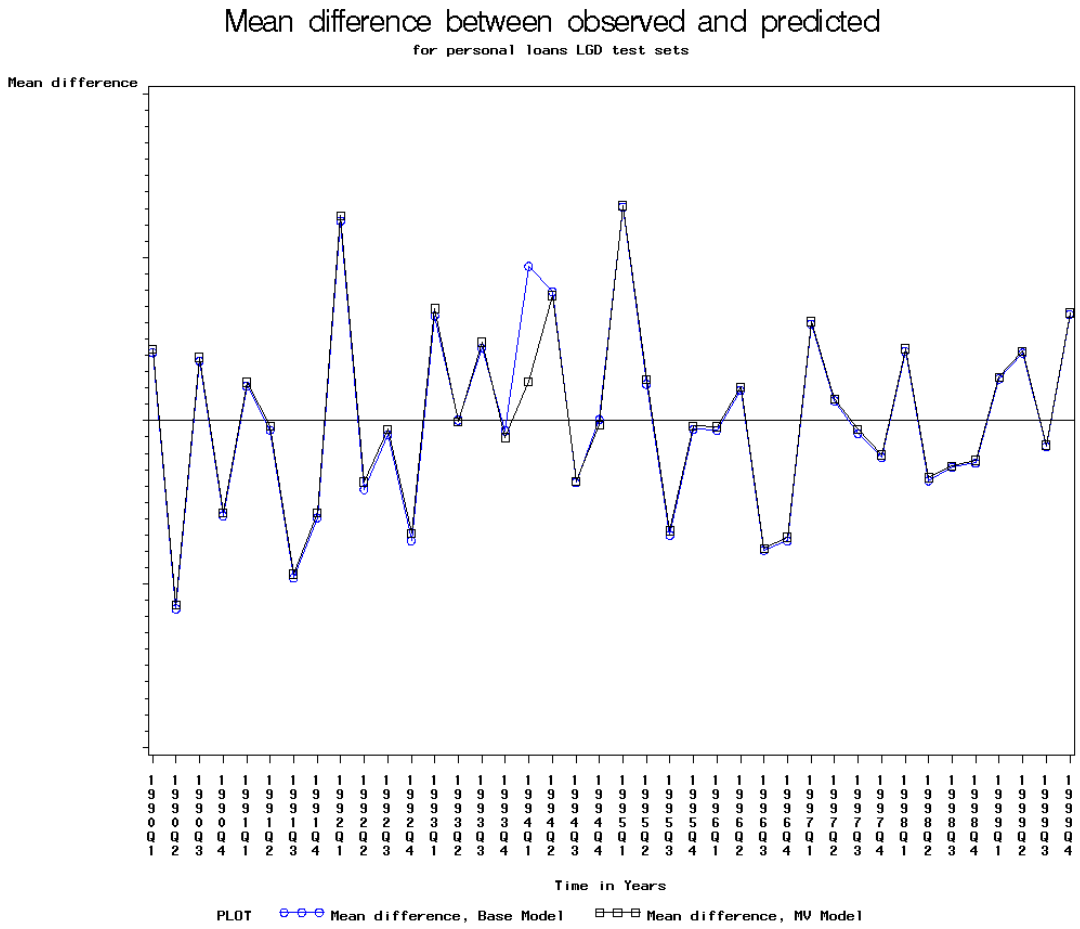

Figure 4.8: Mean difference between observed and predicted personal loan LGD, for base and macroeconomic LGD models (Test sets). The circles represent mean difference coming from the base LGD model and the squares represent mean difference coming from the macroeconomic LGD model.

In summary, although there were some benefits to incorporating interest rate as an additional variable in our mortgage LGD model(s), no such improvement was found for the model built on the personal loans dataset.

4.7 Conclusions

In this chapter, we investigated the inclusion of macroeconomic variables on two different retail loan LGD models, one on residential mortgage loans and the other on unsecured personal loans. Although macroeconomic variables have gained significance in corporate LGD models, they do not seem to have the same level of importance in retail LGD models.

In the case of residential mortgage loan LGD, both the Probability of Repossession Model and the Haircut Model benefit from the inclusion of macroeconomic variables (more specifically, interest rates at default). This is an interesting contribution because it should be noted that the HPI, which is the leading macroeconomic variable in the housing market, is already embedded in the Probability of Repossession Model, the Haircut Model and the calculation of mortgage loan LGD. Combining the component models, we find that although the overall R-square for the macroeconomic LGD model increases, there seems to be no other obvious improvements in the other performance measures. The distribution of predicted LGD produced by the macroeconomic LGD model is similar to that produced by the base LGD model (see Figure 4.5). Mean predicted values of LGD for each quarter are also close, but the macroeconomic model seemed to be able to predict better for higher LGD bands. This implies that LGD predictions from the macroeconomic model are skewed towards the downturn period: the model performs better on observations that went into default during the economic downturn, but was consistently underestimating LGD for non-downturn periods. More work is required here, perhaps to investigate if macroeconomic variables are non-linearly related to recovery rates for mortgage defaults, or if they affect different loans (e.g. loans of different DLTV bands or security type) differently. Possible techniques include binning the macroeconomic variable, or even applying interaction terms.

Secondly, where unsecured personal loan LGD is concerned, all of the macroeconomic variables turn out statistically insignificant, the exception being net lending growth (at default). However, it brings no discernable improvement in prediction of LGD, which seems to suggest that personal

loan LGD seems to be less affected by the economy. This result is similar to that of Bruche and González-Aguado (2010), who find that the impact that macroeconomic variables have on recovery rates are limited.

Ultimately, one can argue that, although there was a modest performance improvement for the mortgage models, neither of the retail loan LGD models investigated here benefited from further incorporating the effects of the economy to the extent that was perhaps expected. However, perhaps macroeconomic variables would be able to contribute more substantially to credit risk retail models in other ways, for example, in survival analysis models, where they are able to accommodate time-dependent variables, and this is investigated in the following chapter.

CHAPTER 5. COMPETING RISKS SURVIVAL MODEL WITH SIMULATED LOSS DISTRIBUTIONS

In previous projects, Loss Given Default (LGD) models for residential mortgage loans were developed using a combination of logistic and linear regression models. These models have produced decent predictions of LGD, and although there were some improvements in the individual component models (i.e. Probability of Repossession Model and Haircut Model) as well as overall R-square value, we see that macroeconomic variables caused model prediction to be skewed towards the downturn years. Also, the regression methods would always be measuring macroeconomic indicators at just one snapshot in time per observation (e.g. at default) whereas these indicators essentially change over the course of the workout. Survival models, on the other hand, are able to take into account these time-dependent variables.

Therefore, in this chapter, we first model mortgage loan LGD by developing a competing risks model for the possible outcomes after default (i.e. repossession or the default case being closed), which will help us better understand what impacts time to repossession. This model will also include a time-dependent macroeconomic variable, i.e. house price index growth. We then apply a Monte Carlo simulation, taking into account this time-dependent covariate and variability of haircut, to assign each loan default a predicted event type and timing of that event. With this time estimate, it is then possible to estimate a discounted loss (whereas previous chapters looked at nominal loss). A loss distribution is then obtained by subjecting a given set of defaulted loans to multiple such runs. Subsequently, it will be shown how to use this framework for stress testing; i.e., by varying the macroeconomic time series used in the simulation, distributions of predicted losses can be obtained under various scenarios.

5.1 Literature review

Survival analysis is a class of statistical methods commonly used to study the distribution of event time; e.g. in the medical domain, survival models typically aim to model the amount of time a patient survives (i.e. time to death). However, it is also useful in other areas where it is of interest to estimate when a certain event is likely to occur; for example, in the case of production, the prediction of machine failure, or in the case of credit risk, the prediction of default. There are a number of papers which have applied survival analysis to credit risk modelling. However, we note that, to our knowledge, most of them are PD models (i.e. the survival models developed predict the time the loan will enter default, based on loan characteristics taken at time of application). The work here differs in that the survival analysis model is developed on data to try and identify accounts that are to be repossessed, as well as predict the length of time the repossession process will take given the macroeconomic conditions.

5.1.1 Survival analysis in retail credit models

Narain (1992) was the first to introduce survival analysis to the area of credit scoring but did not compare it to any alternative methods. Banasik, Crook and Thomas (1999) was one of the first papers to advocate the use of survival models for credit problems, when logistic regression was (and still is) the established methodology for building PD models in retail lending. In it, they acknowledged the exploration of alternative methods like discriminant analysis and partitioning trees (by Rosenberg and Gleit (1994), Hand and Henley (1997) and Thomas (1998)) but concluded that all were inadequate because eventually these were methods that rely on static information recorded during a particular time point of the loan yet expected to predict towards something in the future. Even if the event could be predicted, the time at which the event would take place cannot be predicted and this is a key component in the calculation of loss. Hence, they developed a competing risks survival model using loan characteristics for default vs. early prepayment and showed that survival analysis produces comparative results to a logistic regression model.

The idea of using survival analysis in credit scoring is popular because, if successful, it is able to give an approximation of when a credit event (e.g. default, repossession, early prepayment) is likely to occur, which is key to the estimation of loss. Stepanova and Thomas (2002) outlined the use of survival analysis models in the development of credit scorecards for personal loans and showed how coarse classification of characteristics could be done in the context of survival analysis. They then compared the performance of their survival models with their logistic regression equivalents using ROC values and found survival models to be better able to predict for event occurrence, especially for the first 12 months after the start of the loan. Bellotti and Crook (2009a), working on credit card data, compared the performance of survival models (with and without time-dependent variables) against corresponding regression models, and using different performance measures, came to similar conclusions. Note that these papers focused on the prediction of an event (default) and the comparison of survival and logistic models.

Work has also been done specifically for mortgage loans. Phillips and VanderHoff (2004), using a logit model, investigated the probability of a defaulted loan going into repossession (also known as foreclosure), get cured (where debtors resume payment) or prepay (where debtor pays off the loan in a lump sum). They were interested in which outcome the defaulted loan would eventually experience, with no indication of the extent of losses. So although local economic and housing market conditions were included in their logit model, their objective was to run simulations to investigate how local legislation (e.g. lowered foreclosure costs, or different redemption laws) would affect the outcomes of these defaulted loans. Pennington-Cross (2010) uses a competing risks model to investigate the time period subprime loans spend in repossession. Specifically, these accounts are those that are in default and where repossession is already in process, and which could now experience one of a number of outcomes (cure, partial cure, repayment in full, real estate owned). Again, no loss estimates were produced.

McDonald, Matuszyk and Thomas (2010) applied survival analysis to estimate losses in a mortgage loan portfolio. A competing risks survival model was developed on a dataset of mortgage loans in order to predict for default or

early prepayment, using both loan related characteristics and macroeconomic indicators (which are time-dependent). A Monte Carlo simulation is then applied which, at each run, translates the survival (or hazard) probabilities into predicted dates at which predicted events would occur. Based on these predicted events and their respective dates, any loss is calculated and cash flow is modelled. Although the simulation did account for probability of repossession and haircut, it used only a general estimation of percentages (not detailed in the paper). The methodology adopted in this work is similar to their work – a competing risks survival model to predict two mutually exclusive events followed by a Monte Carlo simulation to translate failure probabilities (i.e. 1 – survival probabilities) to predicted events and event times, but instead of focusing on default prediction, we explicitly model post-default events using survival analysis.

In summary, our work differs from these papers firstly because we focus on the prediction of LGD by specifically developing survival models predicting for repossession or the default case being closed without having to repossess (henceforth referred to as "closure"). Time-dependent macroeconomic variables are incorporated, with the intention of investigating the risk drivers of repossession. The simulation proposed in this work will then investigate the impact of such time-dependent variables, as well as take into account the variability experienced by haircut, on loss distributions.

5.1.2 Monte Carlo and stress testing

It is acknowledged that the number of defaults and losses differ in good economic times and bad (see Frye (2000a), Frye (2000b) , Jarrow (2001), Altman, Resti and Sironi (2001), Hu and Perraudin (2002)). Furthermore, the number of defaults and LGD are correlated such that losses are further magnified during poor economic situations. Similarly, in this dataset, we also see an increased number of repossessions during the economic downturn in the UK in the early 1990s. Also, from the Basel perspective, it is important to differentiate between long-run PDs and LGDs and their downturn counterparts. For PD, a formula derived from a single-factor Value at Risk (VaR) model is provided in the Basel II documentation (Basel Committee on Banking Supervision (2005)) to translate PD to downturn PD,

but the equivalent is not available for LGD. Hence, banks are required to develop their own methods to estimate downturn LGD. Whereas such a downturn LGD estimate is supposed to give an indication of how bad loss rates typically are in an "average" bad year, the aim of stress testing on the other hand is to investigate the impact on the amount of capital needed to cope with specific unlikely but still plausible events or stress scenarios, and is meant as an adjunct to the simple VaR model adopted for Basel II regulatory capital calculations. Although it is not reasonable to expect banks to hold this amount of capital in preparation for such an extreme event, the minimum objective is for banks to be aware of their worst case scenarios, as well as their available options should such a situation ever arise. With the implementation of Basel II, this risk management tool became a prominent part in the development of risk models, with many papers upholding its importance (see Blaschke et al. (2001) and the Financial Services Authority (2005)). However, because it is a relatively new concept, especially in retail credit, there is still uncertainty regarding the most appropriate methodologies of stress testing (see Berkowitz (2000) and Sorge (2004)).

Coleman et al. (2005) carried out a case study on Australian banks and their housing loan portfolios, modelling default and loss rates using a regression model with house price changes as an explanatory variable. Consistent with Sorge (2004), they defined loss to be a function of individual bank variables coming from the components of expected loss (i.e. probability of default, loss given default, mortgage insurance recoveries and resource costs) and macroeconomic shocks. The data was segmented and analysed on a number of variables (e.g. loan to value, type of mortgage, loan age) and various situations were considered in order to stress test the amount of capital required to absorb the losses. They found that a 30% drop in house prices translated into a large increase in both default rates and LGD. Rodriguez and Trucharte (2007) carried out a case study using a dataset of Spanish mortgage loans, in which macroeconomic variables (unemployment rate and the official mortgage interest rate) were included in their logistic regression model towards the prediction of probability of default. This model is made to be time-dependent with the inclusion of the corresponding variable and macroeconomic values of each time, essentially creating a system of models. If, as in this particular case study, some of the mortgage loan characteristics

are time-dependent, this model would require a lot of data which would also need to be consistently available at each time step, which might not always be possible. In this paper, it is not a major issue because simulation was only done on one cohort year (2004). Rodriguez and Trucharte (2007) therefore illustrate the advantage of using simulation to generate loss distributions under different economic situations, including stressed economic situations. These are then compared against the Basel II capital requirements and the authors find that they give adequate protection.

Although both these papers predict losses for mortgage loans, they are based on PD models. It is thus necessary to investigate the relevance of a mortgage LGD model, which should also take into account the impact on losses that time to repossession could have. Although it is assumed that there is no loss if repossession does not occur, once repossession does occur, any losses would have to be discounted appropriately. Also, in order to more accurately estimate the valuation of the property at the time of repossession and sale (because the time between default and repossession could take an average of two years), it is necessary to model the time between default and repossession, if repossession does happen. In addition, it would also be beneficial to develop a mortgage loan LGD model which takes into account both the volatility of haircut (which is a big factor in mortgage LGD) and macroeconomic changes. None of this has been covered in the literature yet.

5.2 Research objectives

Retail LGD models are usually developed using regression models, which forces a deterministic solution onto a dynamic problem. LGD varies depending on when the default occurred, how long the recovery process takes, as well as when the recoveries come back in, which is believed to depend on the economy. This is especially evident in the case of mortgages, because if the bank decides to repossess the property, this repossession process may vary between months and years after default (as was observed in the dataset, see Figure 2.3), and the sale of the property might not be immediate. If, on the other hand, the bank has not repossessed the property, perhaps because the debtor has restarted repayments, the account is said to

be closed (or have entered closure). In this case, even though there may not be any costs associated with it[24], it is still important to recognise this as an event in itself because closure and repossession are mutually exclusive events.

Firstly, we wish to model the probability of a defaulted mortgage account experiencing some event (either repossession or closure) at a given time after default. In order to appropriately model events (i.e. repossession or closure) that follow the default of a mortgage loan, survival analysis is considered. Not only is it able to produce probabilities of repossession or closure happening in the months after default, it is also able to take into account censoring (observations that did not yet experience any event). The Cox survival model is also able to take into account time-dependent variables, which would allow for the incorporation of macroeconomic variables. By developing this model, we should also be able to develop a better understanding of the different risk drivers of repossessions and closure.

In order to reflect the possibility of either of two events occurring, the individual models are combined into a competing risks survival model, which allows for more than one type of event taking place, whereby the events are mutually exclusive. In other words, if the account experiences repossession, it is no longer susceptible to closure and vice versa.

Secondly, we wish to produce an LGD model within a framework appropriate for stress testing. The output from the survival models give probabilities, for each observation, of that event happening in each month after default has occurred. Using simulation and translating these probabilities into some predicted event and some predicted event time, we would be able to (a) validate the model by comparing actual and predicted events and event times and (b) stress test the model by adjusting the macroeconomic variables within the survival models and, depending on how survival probabilities are affected, investigate how the pattern of repossession, and LGD, changes.

[24] It is widely accepted in industry that loss will only occur if repossession and sale of the property takes place, and if sale proceeds are unable to cover the outstanding balance.

5.3 Data preparation

5.3.1 Data

The data used here is as described in Chapter 2. However, note that if we were using regression methods, observations that entered default towards the end of the sample (2001-2002) would have to be removed from the analysis because of the short outcome window (between default and repossession); failing to do so would otherwise cause a misrepresentation of a seemingly low repossession rate for accounts that default in the early 2000s. Survival analysis on the other hand is able to handle such accounts as censored observations, avoiding any such issue.

In the case of accounts that experienced multiple defaults, only the last instance of default is included here. There are two main reasons for this. The first is due to data limitations. For these accounts, we only have the estimated date of (repeated) default, so there is no information with regards to when the debtors recovered from their previous default. Secondly, taking the last instance of default, which implies that the account is at risk of repossession, would ensure that the model is most conservative. This would also be convenient for simulation purposes, since, if all instances of default of a single account were to be included, each would get estimates for if and when that instance would enter repossession or closure, and it would not make sense if the same account were predicted to be both repossessed and closed (at different times).

Under the Basel II Internal Ratings Based framework, banks have to predict PD for the next 12 months, and their resulting losses, for each portfolio of loans. In tandem with a PD model, the model developed here would be used to estimate the LGD of mortgage loans, which will depend on whether a repossession or closure takes place, as well as the time it takes for the event to happen. In order to mimic the way the model would be used, we further subset the test set by years. Validation will be done on two years: 1995, which represents a typical non-downturn year, and 1991, which is selected to represent a downturn year.

5.3.2 Defining events

In the dataset, we have dates which correspond to the events that each account experienced. An account that has a repossession and sale date is defined to be a repossession; an account that has a close date is defined as a closure. In the case of accounts with no repossession or close date, we define them as censored at the latest observation time in the sample, December 2003. All these are mutually exclusive events; an observation cannot be a closure and repossession at the same time, and if an account gets closed, it will no longer be at risk of repossession. As such, it is necessary to have a competing risks model.

5.3.3 Variable selection and pre-processing

For each observation, besides information that was collected at the time of loan application (e.g. LTV, valuation of property, type of security), we also have some default time information, including an estimated default date from which we calculate a more current valuation for the property, outstanding balance, DLTV and relative standing of the property in its region (ratio of valuation of property at default over the average regional average). Although we do have a lot of information on loan and collateral related characteristics, it was important that the model developed remains relevant and easy to implement, as well as suitable for stress test purposes. Also, from the previous chapters, although we find that repossession risk depends on a few variables – loan-to-value ratios, type of security, time on books and whether default happened before – we also see that it is the loan-to-value ratios that are most significant. Both LTV and DLTV carry important information. LTV gives an indication of risk at the time of loan application and is stable because it is not related to the economy as the loan progresses, whereas DLTV gives an updated indication of debtor equity but will change with time and the economy. Similarly to the repossession model developed in Chapter 3.3, we find that including both LTV and DLTV causes counter-intuitive parameter estimate signs, so it is necessary to select only one. Eventually, we decided to use DLTV as it is as an indication of debtor equity at default.

Another variable included is the type of security. This is a categorical variable that differentiates between detached, semi-detached, terraced properties and flats.

Only three macroeconomic variables were considered here. They are unemployment rates, interest rates, which are both taken from the Office of National Statistics (ONS), and the Halifax House Price Index (HPI). All three are year-on-year growth rates and are available monthly. Eventually, we use only the HPI growth rate (HPIG). The use of growth rates eliminates the use of absolute values and ensures that the real change in HPI is reflected. We have the HPIG for each month following default, so this is a variable that changes with time.

Time lags and leads of up to 12 months were considered here (i.e. HPIG values from 12 month before default to 12 month after default). Due to computational limitations, these variations were applied only on a preliminary repossession survival model, where DLTV was not binned and interaction terms were not yet included. Some of the models, especially where the lead or lag interval considered becomes larger, have the parameter estimate sign for HPIG switched from positive (i.e. generally repossession is more likely to happen in good economic times) to negative (i.e. generally repossession less likely to happen in good economic times). This is not intuitive and it was decided to use the model with neither leads nor lags. However, further work could be done to investigate the application of such time lags, especially after interaction terms are considered, and whether these changes might impact the overall prediction of LGD.

We also observe that the economy affects different types of security differently, as well as observations of different ranges of DLTV. In particular, we find that although the risk of repossession increases with DLTV, once DLTV reaches a certain threshold, this risk seems to remain constant. In order to accurately model this, interaction terms are introduced between DLTV and HPIG, as well as between type of security and HPIG. DLTV is binned into 7 groups, which also aids the interpretation of the parameter estimates of these interaction variables. The bins are created according to boundaries of DLTV values, with smaller bins for larger DLTV values.

5.4 Competing risks survival model

5.4.1 Survival models

Survival analysis aims to model the period of time an observation takes to experience an event. In general terms, if T is defined to be the time period at which death is observed, where T is some non-negative continuous random variable, then the survival function at time t is the probability that the observation survives up to time t, i.e. T is greater than t (see Equation 5.1).

$$S(t) = P(T > t) \tag{5.1}$$

Given that the observation has survived up to time t, then the probability that it will not survive in the next time interval ∂t is $P(t < T \leq t + \partial t \mid T > t)$. Another term of interest is the hazard rate, $\lambda(t)$, which is the risk rate at each unit of time, for observations that have survived up to time t (see Equation 5.2).

$$\lambda(t) = \lim_{\partial t \to 0} \left(\frac{P(t < T \leq t + \partial t \mid T > t)}{\partial t} \right) \tag{5.2}$$

The relationship between hazard rate and survival function is given by Equation 5.3.

$$S(t) = P(T > t) = \exp\left(- \int_0^t \lambda(x) dx \right) \tag{5.3}$$

There are a number of different types of survival models, according to the type of distribution that event time follows. However, it can be difficult to identify an appropriate distribution from the available data, and because each model assumes a certain underlying distribution, applying the wrong

type of model would produce invalid results. A popular class of models is the Cox Proportional Hazards Model (Cox (1972)), because it removes the need for the identification of the underlying distribution of survival times. In this model, the hazard rate, given in Equation 5.4, of an individual depends on a component that is common to all (baseline hazard rate, $\lambda_0(t)$) and its individual characteristics (covariates), $X_1, X_2, ..., X_m$.

$$\lambda(t \mid X) = \lambda_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m) \tag{5.4}$$

Another advantage of the Cox Proportional Hazards Model is that it is able to take into account time-dependent variables. In this case, the individual characteristics of an observation's hazard rate will now consist of two parts, the static characteristics of the loan (*X* covariates), and the time-dependent variables (*Y(t)* covariates), given in Equation 5.5.

$$\lambda(t \mid X, Y(t)) = \lambda_0(t)\exp(\beta_1 X_1 + ... + \beta_m X_m) + (\gamma_1 Y_1(t) + ... + \gamma_q Y_q(t)) \tag{5.5}$$

The Proc PHREG in SAS will estimate the values of the β and γ parameters and produce, for each time step, the baseline hazard rate. Once the baseline hazard rates are produced, hazard rates and survival probabilities for each time step can be calculated subsequently. However, when time-dependent variables are included, the calculations used for baseline hazard rate in this SAS procedure do not hold. It is then necessary to apply the Nelson-Aalen formula (see Andersen (1992) and Chen et al. (2005)) to calculate the time-dependent baseline hazards, according to Equation 5.6.

$$h_0(t) = \frac{d_t}{\sum_{R_t} \exp\left[(\beta_1 X_1 + ... + \beta_m X_m) + (\gamma_1 Y_1(t) + ... + \gamma_q Y_q(t))\right]} \tag{5.6}$$

Where d_t = number of events that occurred in time t and R_t = observations that are at risk of the event at time *t*.

5.4.2 Competing risks

The explanation on survival models so far have been for the case where the observation is assumed to be at risk of experiencing only one type of event, for example, death. However, it is common to want to differentiate between different kinds of events that an observation can experience, for example, when modelling the survival time of a piece of machinery, whether it breaks down due to age or human error would have different risk factors and thus should be modelled separately. This is known as a competing risks model.

In this work, there are two events of interest: repossession and closure. Once the account is closed, it is no longer susceptible to repossession and vice versa. A survival model for each event (repossession and closure) is developed. In the survival model for repossession, all non-repossessions (i.e. closed and censored observations) are assumed to be censored; in the survival model for closure, all non-closure observations (i.e. repossessed and censored observations) are assumed to be censored. Each survival model produces survival and hazard probabilities and a random number generator is used here to produce random numbers and compared against survival probabilities to produce predicted event times (more details in Section 5 later).

The competing risks survival model will then define survival time to be the minimum of the time to repossession and the time to closure, given in Equation 5.7.

$$T = \min\{T_r, T_c\} \tag{5.7}$$

Where T_r = time predicted to repossession and T_c = time predicted to closure.

5.4.3 Survival model for repossession

A survival model is developed here to predict the number of months it takes for an observation to go from default to repossession. Only observations

that have been repossessed are defined to have undergone the event, and observations that were closed or censored are defined to have been censored at the time of their closure or censoring. For example, an account that went into default in February 1992 and was repossessed in January 1995 is said to have experienced the event at 35 months, whereas an account that went into default in February 1992 and closed in January 1995 is said to be censored at 35 months.

Variable selection is as described in Section 5.3.3. The final variables used in this survival model are DLTV (a continuous variable now binned into seven groups), security (a categorical variable), HPIG at time of default, and two sets of interaction terms – HPIG and security, and HPIG and DLTV (binned). Their parameter estimates and p-values are given in Appendix E, Table E1, but because there are interaction terms in this model, interpreting the model is not as straightforward as looking at parameter estimates and interpreting their signs. The marginal effect of the log-risk of repossession for different types of security can be split into two parts: it is affected differently depending on what kind of property it is, and which DLTV band it is in (non-time-dependent); and it will also react differently depending on the changes in the HPIG (time-dependent).

In order to make any sense of the results, we have to separate between the types of security and then look at the overall parameter estimates for different bands of DLTV. This means that the marginal risk of repossession for different types of securities and different DLTV bands have different linear relationships with HPIG, and this is represented by a linear line, $\text{repossession risk}_{\text{security, DLTV group}} = m\text{HPIG} + c$, where m represents the slope of the line, i.e. the sum of the relevant parameter estimates that are HPIG-related (time-dependent variables) and c represents the intercept, i.e. the sum of all other relevant parameter estimates.

For example, an account that has a flat as the security (security0) and a DLTV of 0.64 (groupdtv2) would be calculated to have:

$$m = 0.022 + (-0.002) + (-0.030) = 0.01$$
$$c = -1.194 + 0.321 = -0.873,$$

(5.8)

which will give:

$$\text{repossession risk}_{\text{flat, groupdtv2}} = -0.01 - 0.873 \times \text{HPIG}$$

(5.9)

This calculation is repeated for the rest of the DLTV groupings within the flat category, as well as the other types of securities. Table D1 in Appendix D shows how the variables contribute towards the slope and coefficient during the calculation of risk for the different types of securities.

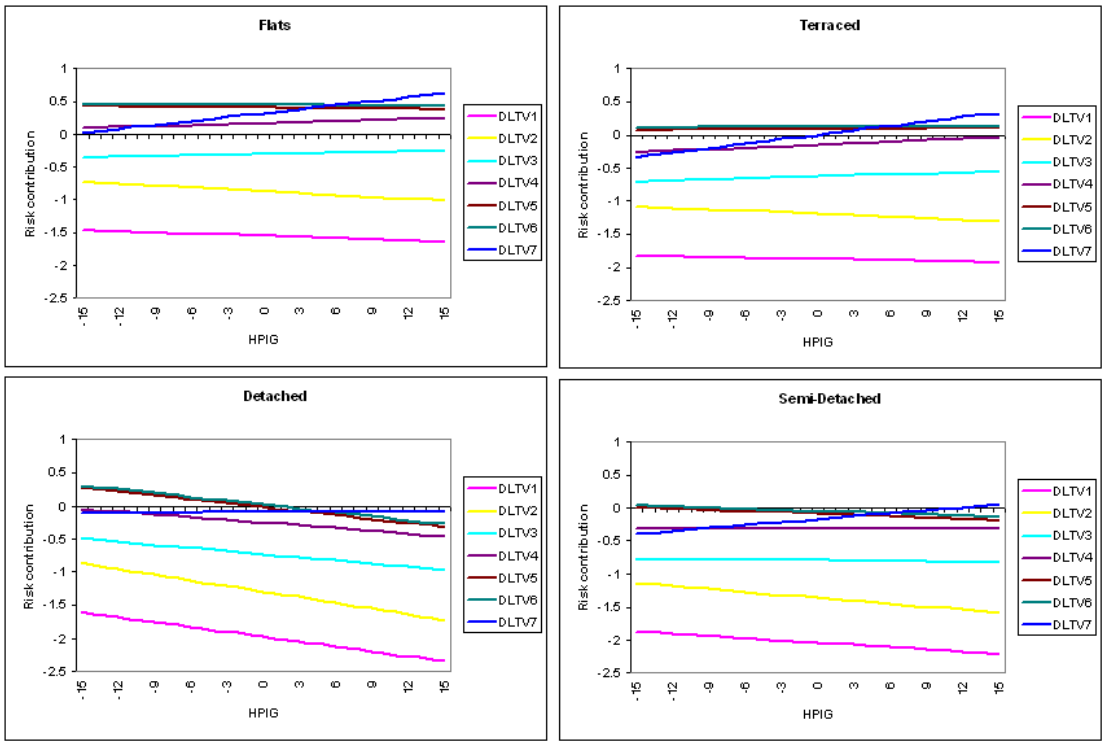

Figure 5.1: Repossession marginal risk contributions for different types of security according to DLTV bands. For comparison purposes, all the horizontal and vertical axes have a common scale.

Figure 5.1 displays four graphs representing the four types of securities, for risk of repossession. On the horizontal axis, we have HPIG which goes from −15% to 15% and each line on the graph represents a DLTV band. From here,

we see that for all types of securities, marginal risk of repossession generally increases with a larger DLTV (the lower DLTV bands are lower on the graphs) but that the increase is smaller for higher DLTV bands. Within each DLTV band, marginal risk of repossession decreases with an improvement in HPIG (a negative slope), and this is most pronounced in the low DLTV bands, with mid-range DLTV bands being rather flat. We observe that flats, terraced and semi-detached properties have lines with similar slopes. The highest DLTV band behaves differently in that the marginal risk of repossession increases with improvement in HPIG (a positive slope). This could be because during poor economic times, banks are more likely to hold on to properties that have high DLTV (since they have more invested in them). This implies that these properties are more likely to be repossessed in good economic times, giving a positive slope. The rest of the DLTV bands have slopes that are negative, i.e. the risk of repossession decreases when the economy is doing well. Detached properties have marginal risk contributions that are different from the rest of the other types of security, in that the slopes are more pronounced, so detached properties are more affected by changes in HPIG.

From the parameter estimates and using the Nelson-Aalen estimator, the baseline hazard rates for each time step for the survival model is calculated and attached in Figure 5.2. Redefining the terms in Equation 5.6, d_t is the number of repossessions that happened in time t, and R_t is the set of observations that are still in the risk set at time t, i.e. are already in default but not yet repossessed, closed or censored. This baseline hazard is common to all observations, regardless of their individual characteristics and from this, we observe that the risk of repossession (hazard) increases sharply immediately after default, and then drops off steadily. This implies that banks generally try to repossess the property within two years after default, after which the risk of repossession gradually bottoms out.

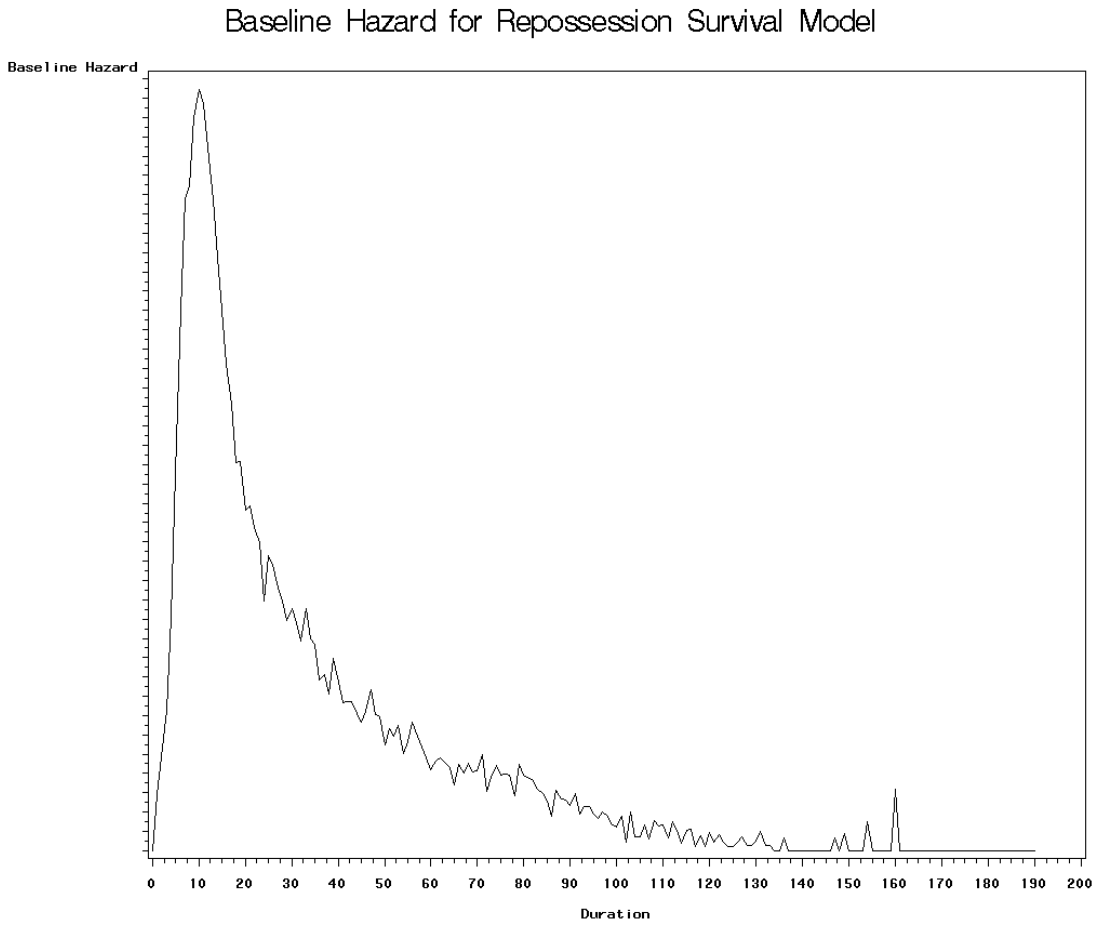

Figure 5.2: Baseline hazard rates for repossession survival model

5.4.4 Survival model for closure

Similar to the repossession survival model described above, a survival model is developed here to predict the number of months it takes for an observation to go from default to close. Only observations that are closed are defined to have undergone the event, and observations that were repossessed or censored are defined to have been censored at the time of their repossession or censoring.

Variable selection is as described in Section 5.3.3. However, the interaction variables between HPIG and security were mostly insignificant and thus dropped. The final variables used in this survival model are DLTV (binned into seven groups), security (a categorical variable), HPIG at time of default, and one set of interaction terms – HPIG versus DLTV (binned). Their parameter estimates and p-values are given in Appendix E, Table E2. Again,

because of the interaction terms, interpreting the parameter estimates is tricky. Similar to the repossession survival model above, we have Figure 5.3, which displays four graphs representing the four types of securities and their marginal risk profile. The rankings, and slopes, of DLTV bands behave similarly across the different types of securities. The positive slope observed for all DLTV bands implies that risk of being closed is higher in good economic times. This could be because debtors are more likely to be successful in securing funds during good economic times, which would allow them to recover from default. Low DLTV bands are also the most likely to be closed without repossession, but this difference is less pronounced in good economic times.

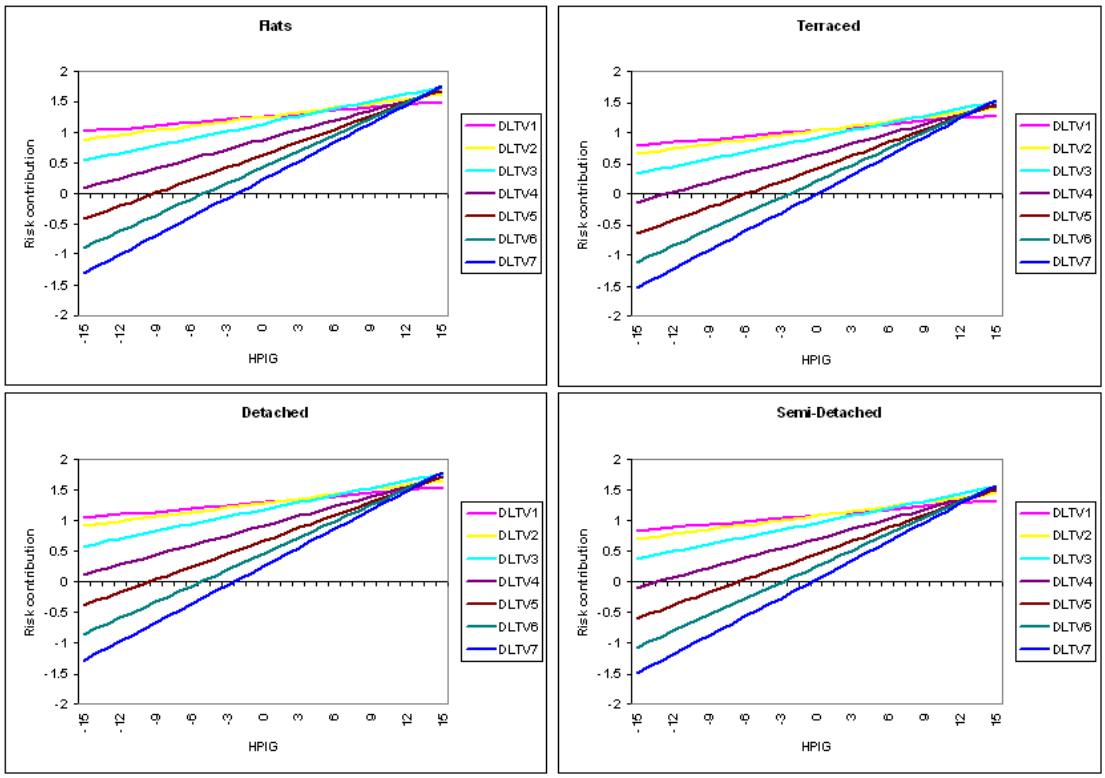

Figure 5.3: Closure marginal risk contributions for different types of security according to DLTV bands. For comparison purposes, all the horizontal and vertical axes have a common scale.

The baseline hazard rates at each time step for the survival model is calculated and attached in Figure 5.4. Redefining the terms in Equation 5.6, d_t is the number of closure observations that happened in time t, and R_t is the set of observations that are still in the risk set at time t, i.e. are already in

default but not yet repossessed, closed or censored. The baseline hazard for closure seems to indicate that the risk of closure is highest in the months following default, and although it decreases with time, does not fall away completely such that this risk of closure remains active (comparative to the earlier post-default months) throughout the observation period we have in the sample. Note that the y-axis scales of Figures 5.2 and 5.4 are not directly comparable.

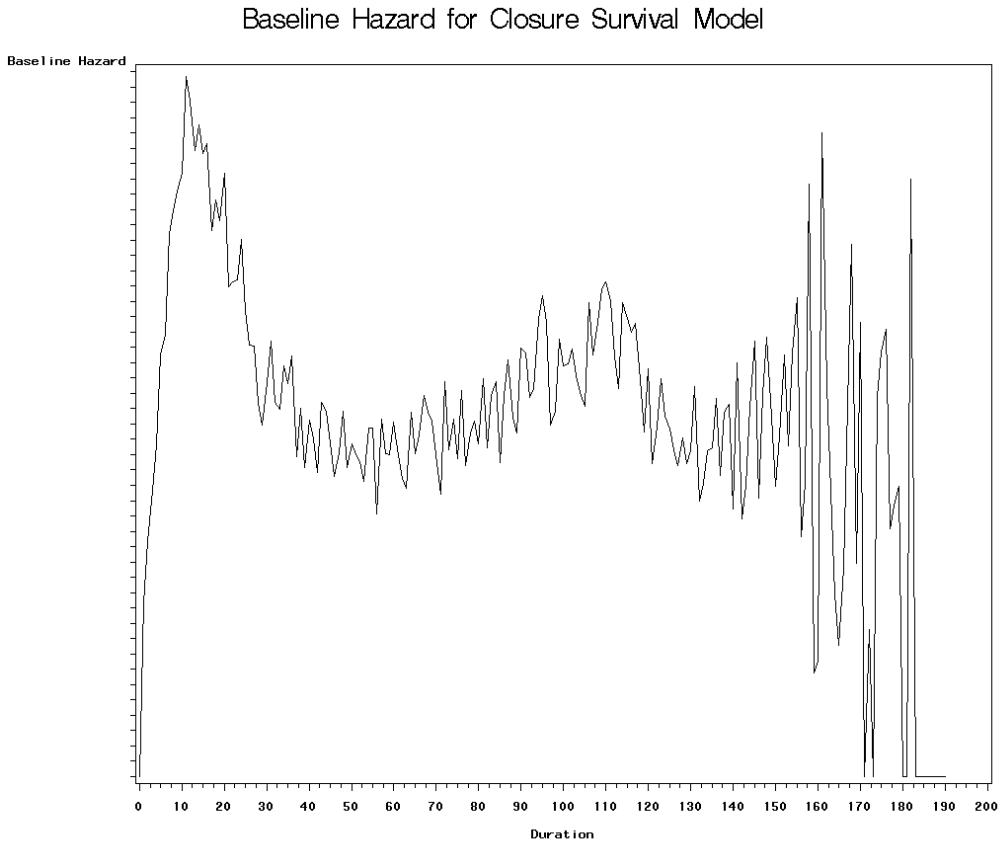

Figure 5.4: Baseline hazard rates for closure survival model

5.5 Monte Carlo simulation

According to the Oxford English Dictionary, Monte Carlo refers to "designating or involving any of various methods of estimating the solution to numerical problems by the random (or pseudorandom) sampling of numbers with some chosen frequency distribution". It is useful, especially in this context of risk and estimation of losses, because it could be used to simulate uncertainty repeatedly yet independently. From the survival models,

we have probabilities of survival to repossession and/or closure for each time step after default, but unless some deterministic time period is determined from these probabilities, they remain just that. With the application of a Monte Carlo simulation, these survival probabilities can be converted into events from which not only the validation of the model can be done, but also, the framework necessary for stress test purposes can be created.

5.5.1 Framework

Each observation *j* is put through each survival model and survival probabilities for repossession, denoted by $s_{r,j}(t)$, and closure, $s_{c,j}(t)$, are produced for each time step *t* (the number months after default). It is then necessary to find conditional survival probabilities during the simulation, i.e. conditional on having survived up to time $t-1$. Using Bayes Theorem, conditional survival probabilities for repossession, $S^*_{r,j}(t)$, and closure, $S^*_{c,j}(t)$, are then calculated for each time *t*, according to Equation 5.10.

$$\begin{aligned} S^*_{r,j}(t) &= P(T_r > t \mid T_r > t-1) \\ &= \frac{P(T_r > t)}{P(T_r > t-1)} \\ &= \frac{s_{r,j}(t)}{s_{r,j}(t-1)} \end{aligned}$$

$$\begin{aligned} S^*_{c,j}(t) &= P(T_c > t \mid T_c > t-1) \\ &= \frac{P(T_c > t)}{P(T_c > t-1)} \\ &= \frac{s_{c,j}(t)}{s_{c,j}(t-1)} \end{aligned}$$

(5.10)

The simulation then takes place as follows for each observation *j*, starting from $t = 1$:

- A random number (between 0 and 1), U_t, is generated and compared first against the conditional survival probability of repossession at time t, $S^*_{r,j}(t)$
- If $U_t > S^*_{r,j}(t)$, then the account is said to undergo repossession at time t
- If $U_t \leq S^*_{r,j}(t)$, then another random number, Z_t, is generated and compared against the conditional survival probability of closure at time t, $S^*_{c,j}(t)$
- If $Z_t > S^*_{c,j}(t)$, then the account is said to be closed at time t
- If $Z_t \leq S^*_{c,j}(t)$, then repeat steps (a) to (d) for time $t+1$, until some event happens, or until the end of the selected observation period (for the purpose of this work, a 12-year observation period is assumed)

Following this simulation run, we are able to get predicted events, and predicted event times for each defaulted account from a single run, which allows us to make predictions of loss. This is then repeated 1,000 times, so that we get a distribution of total loss, as shown later.

At each run, for accounts that are predicted to be non-repossessions, loss is assumed to be zero.

For accounts j that are predicted to be repossessions, there is also a predicted repossession time, and loss is calculated as follows. First, the average time taken for properties to go from repossession to sale is calculated for the available data (about 7 months), and we assume, for simplicity, that the predicted sale date of the repossessed property is taken to be 7 months after the predicted repossession date. However, it would be an area of further work to develop a similar survival model to estimate the time between repossession and sale. For this predicted sale date, the Halifax House Price Index (all houses, all buyers, non-seasonally adjusted, quarterly, regional) is extracted. The valuation of the property at time of sale is calculated according to Equation 5.11.

$$\text{valuation of property}_{\text{sale}} = \frac{\text{HPI}_{\text{sale yr, sale qtr, region}}}{\text{HPI}_{\text{start yr, start qtr, region}}} \times \text{original property value} \tag{5.11}$$

Using the Haircut Model (from Chapter 3, Section 4), we find predicted haircut $\hat{H}$ and predicted haircut standard deviation σ for each observation *j*. Although it was necessary to translate the predicted haircut and standard deviation into a standard normal distribution and weight probabilities of haircut with their corresponding losses, it is not necessary to do so here. Because this haircut model is used in a simulation, there can only be one value of haircut (instead of a range of haircut values with their different chances of occurring). However, instead of taking the value of predicted haircut as it is predicted from the haircut model, we introduce variability to the predicted haircut by adding some randomness (from the normal distribution) which depends on its standard deviation, given in Equation 5.12. This new haircut, $\tilde{h}$, is used in the calculation of shortfall and loss.

$$\tilde{h} = \hat{H} + N(0, \sigma) \tag{5.12}$$

Applying haircut $\tilde{h}$ to the valuation of the property at time of sale, we are able to get a prediction of sale price of the property, from which predicted shortfall amount can be calculated. It is possible that the property is sold for a price that can cover the outstanding balance on the loan, and in this case, any extra proceeds would be returned to the debtor and loss is predicted to be zero. If however, there is a loss (i.e. sale proceeds < outstanding balance), then this loss undergoes discounting (see Equation 5.13).

$$\text{predicted loss} = \max\left\{0, \frac{\text{balance}_{\text{default}} - \left(\text{valuation of property}_{\text{sale}} \times \tilde{h}\right)}{(1+d)^k}\right\} \tag{5.13}$$

where $k = \dfrac{\text{number of months between default and predicted sale time}}{12}$.

For a discussion between regulatory supervisors and banks on the appropriate discount rate that can be used, we refer to FSA's Expert Group Paper on Loss Given Default (Financial Services Authority Expert Group (2009)). In our analysis, a discount rate of 5% is used.

This prediction of loss can be easily translated to a prediction of LGD, according to Equation 5.14. In the case of actual loss, any loss is also discounted by the time taken between default and sale in years (cf. Equation 2.4).

$$\text{predicted LGD} = \frac{\text{predicted loss}}{\text{balance}_{\text{default}}} \tag{5.14}$$

5.5.2 Stress testing

The importance of stress testing was mentioned earlier in the text. Helbling and Terrones (2003) looked at historical house prices for 15 countries from the 1970s to early 2000s, identifying peaks and troughs. They find that housing crashes meant that house prices tumbled an average of 30% and took longer to recover (as compared to crashes in the stock market). Here, we stress the model by creating a simple but plausible economic downturn scenario. There is no suggested percentage on which how badly to stress the macroeconomic variable but we observed that the drop in house prices during the credit crisis of 2008 was about double that experienced in 1991.

A stressed HPIG is created as follows:

- For the period of 1984 to 1988: there is no negative growth observed, so the HPIG values are left are they are.
- For the period of 1988 to 1998: we impose the HPIG values observed from July 2008 to December 2009 (aggravating further the existing economic downturn) onto July 1992 to December 1993. Because we are imposing the HPIG values from a different period, and to ensure that there is a smooth joining trend, any negative HPIG immediately surrounding these dates are doubled.

- For the period of 1998 to 2008: any negative HPIG is doubled
- This is done separately for each region in the UK

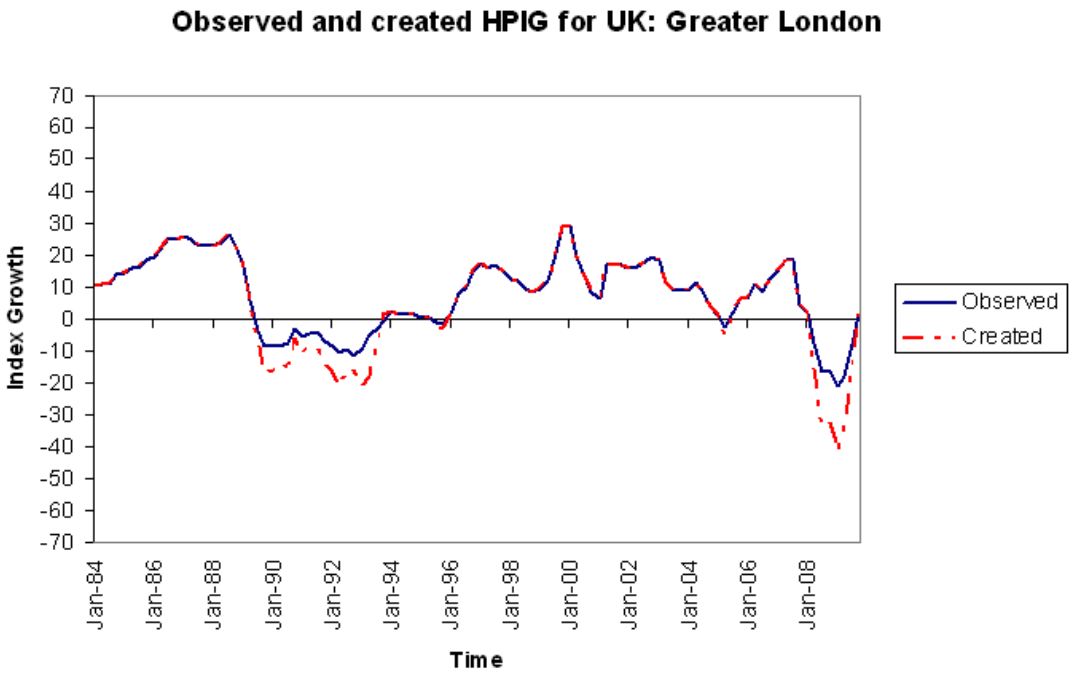

Figure 5.5: Observed and stressed HPIG for region of Greater London. The solid line represents historically observed HPIG and the dotted line represents the created stressed HPIG.

Figure 5.5, for the region of Greater London, shows an example of the extent which HPIG will be stressed. The graphs for the rest of the regions in the UK are in Appendix F.

Table 5.1: Migration matrix of DLTV and stressed DLTV (in percentages) (for test set). The top (bottom) row represents the lowest (highest) DLTV value range, respectively.

	stressed DLTV1	stressed DLTV2	stressed DLTV3	stressed DLTV4	stressed DLTV5	stressed DLTV6	stressed DLTV7
DLTV1	74.73	24.39	0.84	0.05	0.00	0.00	0.00
DLTV2	0.96	60.74	34.81	2.58	0.65	0.18	0.08
DLTV3	0.00	2.26	55.31	23.20	12.62	4.76	1.84
DLTV4	0.00	0.02	5.84	38.45	26.35	18.08	11.25
DLTV5	0.00	0.00	0.43	3.70	34.82	28.20	32.85
DLTV6	0.00	0.00	0.00	0.05	1.36	26.67	71.92
DLTV7	0.00	0.00	0.00	0.00	0.00	0.38	99.62

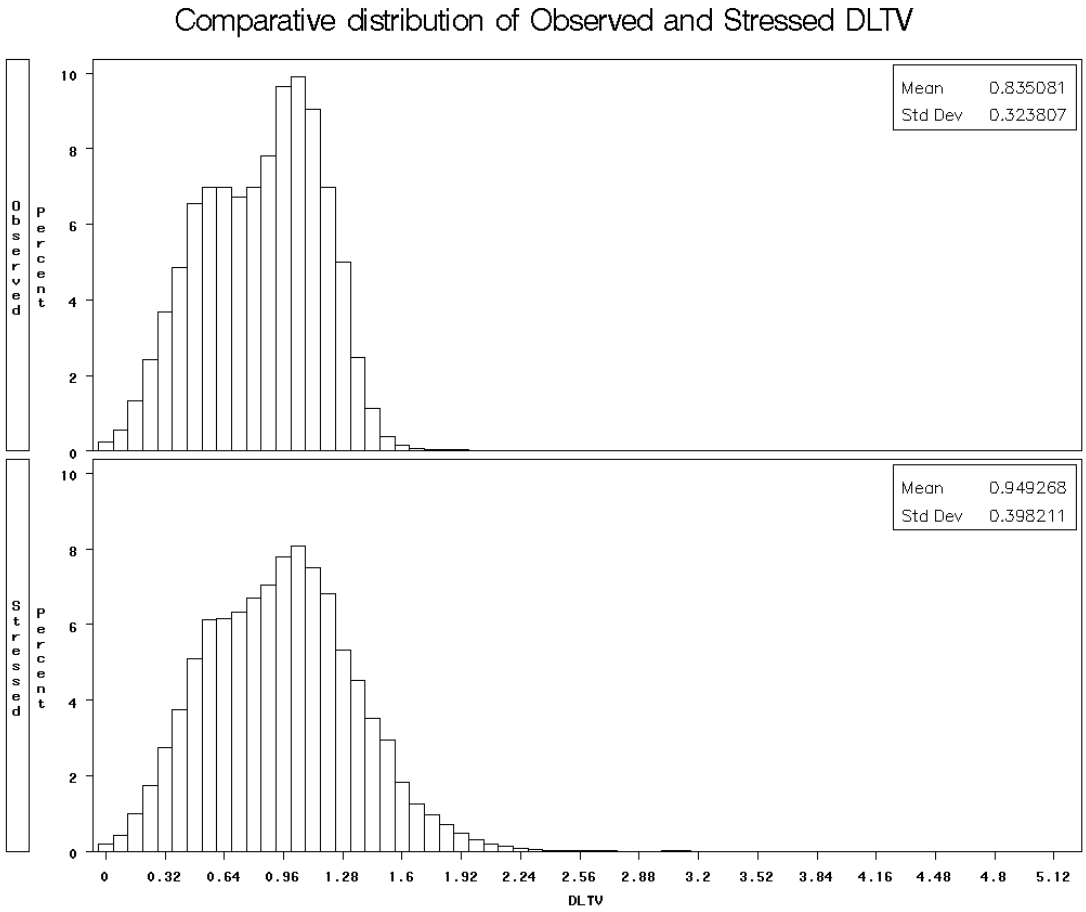

Figure 5.6: Distribution of DLTV and stressed DLTV (for test set). Top panel shows distribution for observed DLTV and bottom panel shows distribution for stressed DLTV.

Because DLTV is a model variable that is calculated using the HPI, we need to recalculate a "stressed" DLTV to take into account the stressed HPIG. The

distribution of DLTV and stressed DLTV is displayed in Figure 5.6, showing that stressed DLTV has a wider spread, and higher mean. Because we used binned DLTV in our models, the differences between DLTV and stressed DLTV are dampened for part of the observations, but we still see a sizeable number of observations moving across DLTV bands (from lower to higher bands), with the higher DLTV bands having fewer accounts staying in the same DLTV band (see Table 5.1). Note that a small number of observations migrate onto a lower (better) DLTV band due to the way the stressed HPIG values were created (because parts of the HPIG from 2008 were imposed onto 1992). These are left as such in the modelling, but can be avoided if preferred, by simply disallowing any such upgrades, thus giving more conservative results. These stressed HPIG and DLTV values are used in the simulation later.

5.5.3 Simulation results

Simulations are carried out for two different time periods. Firstly, we run our analysis on the cohort of loans that defaulted in 1995 (about 3,000 accounts), and then to investigate the performance of the model on a downturn year, the analysis is repeated on the cohort of loans defaulted in 1991 (about 7,000 accounts). Two types of simulations are carried out for each of these default cohorts. The first is the validation simulation, in which the actual observed HPIG time series as given by the Halifax HPI (regional, quarterly) is used. The second is the stressed simulation in which the newly created stressed HPIG values are used (cf. Section 5.5.2). For both loan cohorts, performance measures from the validation and stressed scenarios are produced and compared here.

5.5.3.1 Distribution of total loss

There are a number of indicators we are interested in, the first of which is the distribution of total loss. As described in the earlier section, the predicted loss is calculated for each observation according to its outcome for each run. Loss is totalled for each run, which gives us a distribution of predicted total loss over all runs for 1995 and 1991, shown in Figures 5.7 and 5.8

respectively. The dotted line represents the actual total loss encountered in the validation set for that year, and so is not directly comparable. The model has slightly overestimated losses for both years, which could be due to the model being developed on a group of loans that defaulted in a range of years (during good and bad economic climates), whereas the validation was run on a particular cohort of loans at each time. Also, in the calculation of total losses, a number of additional factors are involved, like the estimation of sale date, the discount rate used, the haircut model and its variability. Finally, although the total losses predicted are in the region of observed total loss, it seems that during good economic times, this overestimation is more pronounced.

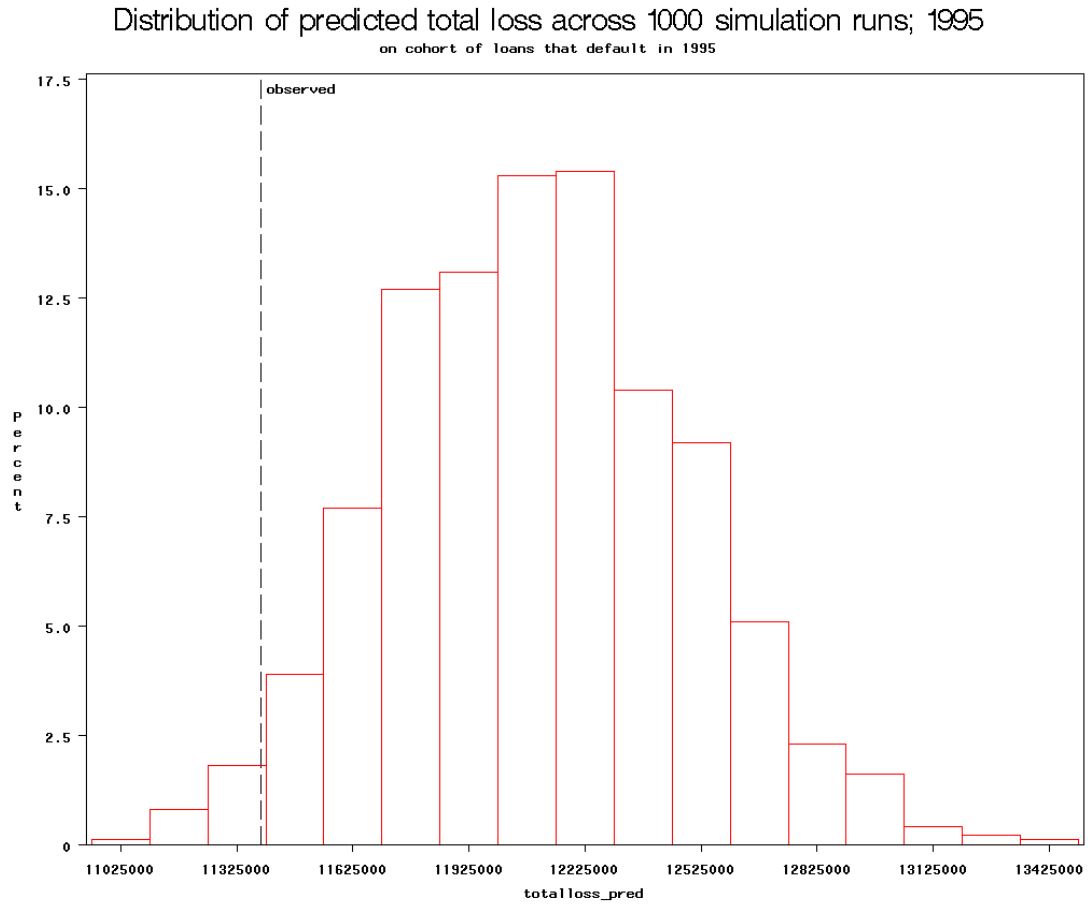

Figure 5.7: Distribution of predicted total loss across 1,000 simulation runs for validation simulation for cohort of loans that default in 1995. The dotted line represents the actual total loss encountered in this cohort.

The simulation is then repeated for the stressed scenario. In order to put the values in perspective, we put the distribution of predicted losses from the validation and stressed simulations onto a single graph with a common scale. These are shown in Figures 5.9 and 5.10, for 1995 and 1991 respectively.

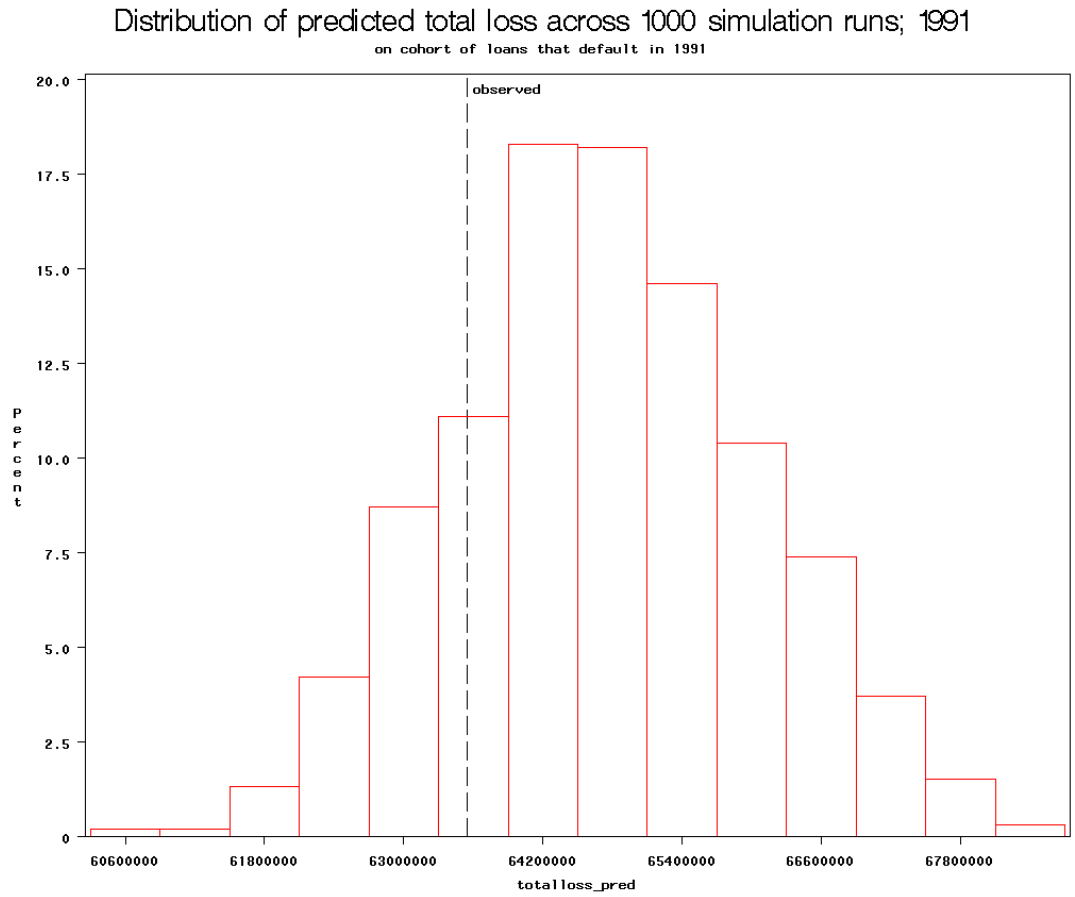

Figure 5.8: Distribution of predicted total loss across 1,000 simulation runs for validation simulation for cohort of loans that default in 1991. The dotted line represents the actual total loss encountered in this cohort.

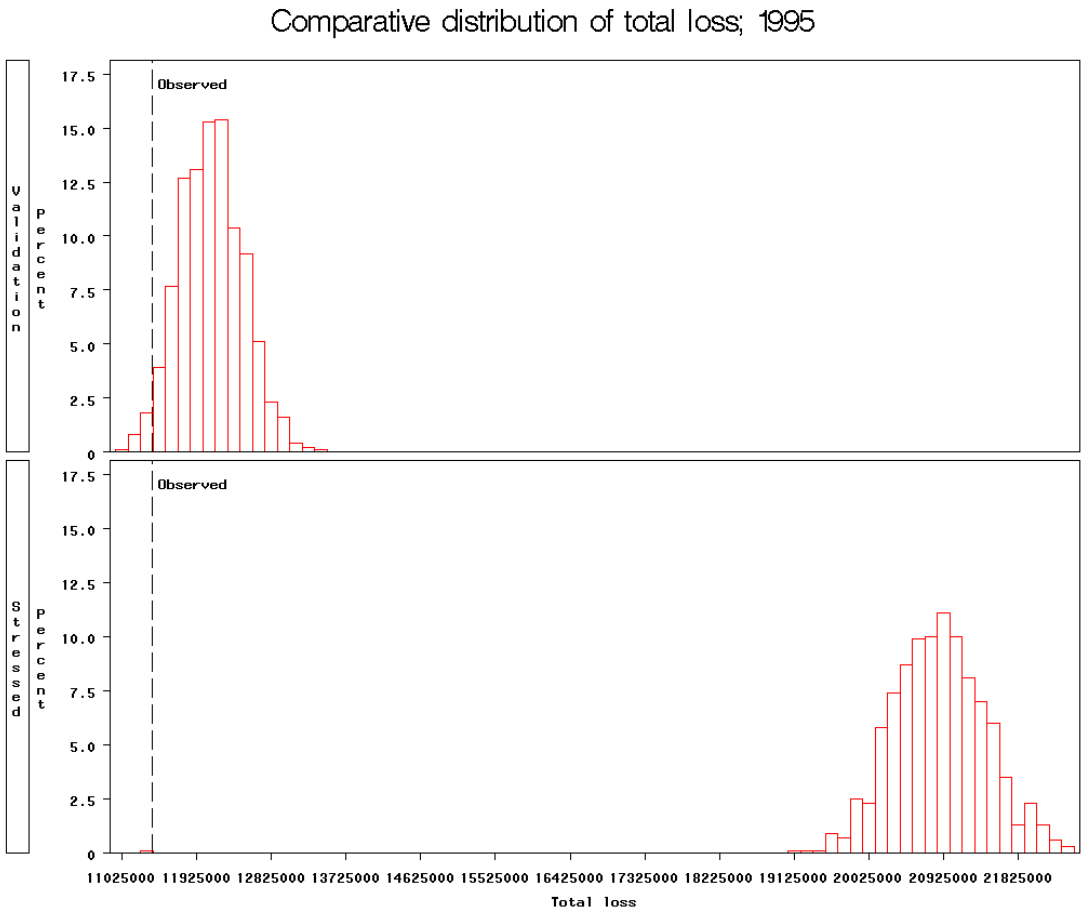

Figure 5.9: Comparative distribution of predicted total loss across 1,000 simulation runs for cohort of loans that default in 1995; for validation (top panel) and stressed (bottom panel) simulation. The dotted line represents the actual observed total loss. Mean of the stressed distribution increased by approximately 75% compared to the validation distribution.

In both cohorts, the model has predicted for much greater losses during the stressed simulation, and the proportion of increase is different in the different cohorts. The model predicts losses to be about 75% more during the stressed simulation in the 1995 cohort, but only 50% more in the 1991 cohort. This is an intuitive result since 1995 is a non-downturn year, so when the economy is stressed, the difference is more obvious. In contrast, 1991 was already a downturn year, so the stress scenario used has affected losses to a lesser extent. Ultimately, it is difficult to assess the plausibility of such stress test results because of data limitations and because there is no way of predicting how poor the economy can get.

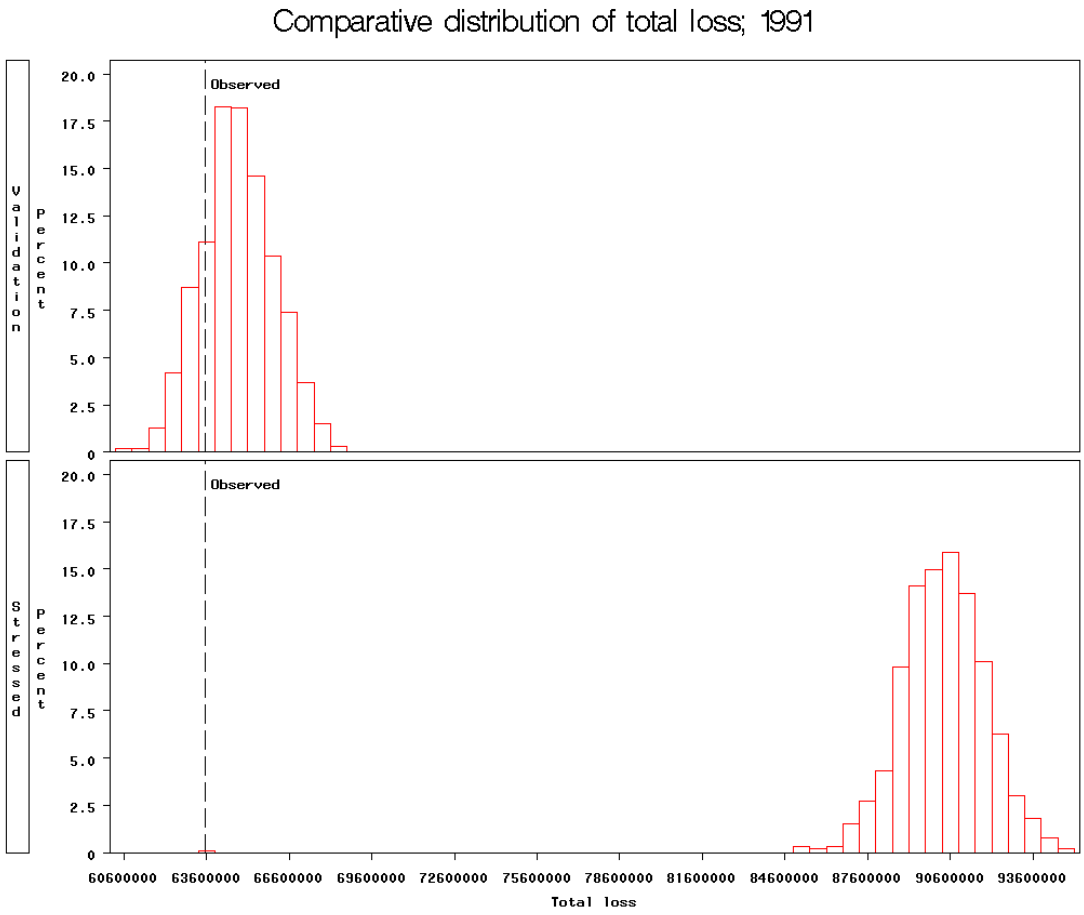

Figure 5.10: Comparative distribution of predicted total loss across 1,000 simulation runs for cohort of loans that default in 1991; for validation (top panel) and stressed (bottom panel) simulation. The dotted line represents the actual observed total loss. Mean of the stressed distribution increased by approximately 50% compared to the validation distribution.

5.5.3.2 Number of repossessions

Another component of interest here is the number of repossessions predicted to take place in each month following default against the number that is observed. The number of observed repossessions in each month following default, together with the average (over 1,000 runs) that was predicted by the validation and stressed simulations respectively is plotted on a single graph. These are in Figures 5.11 and 5.12, for the 1995 and 1991 cohort years respectively.

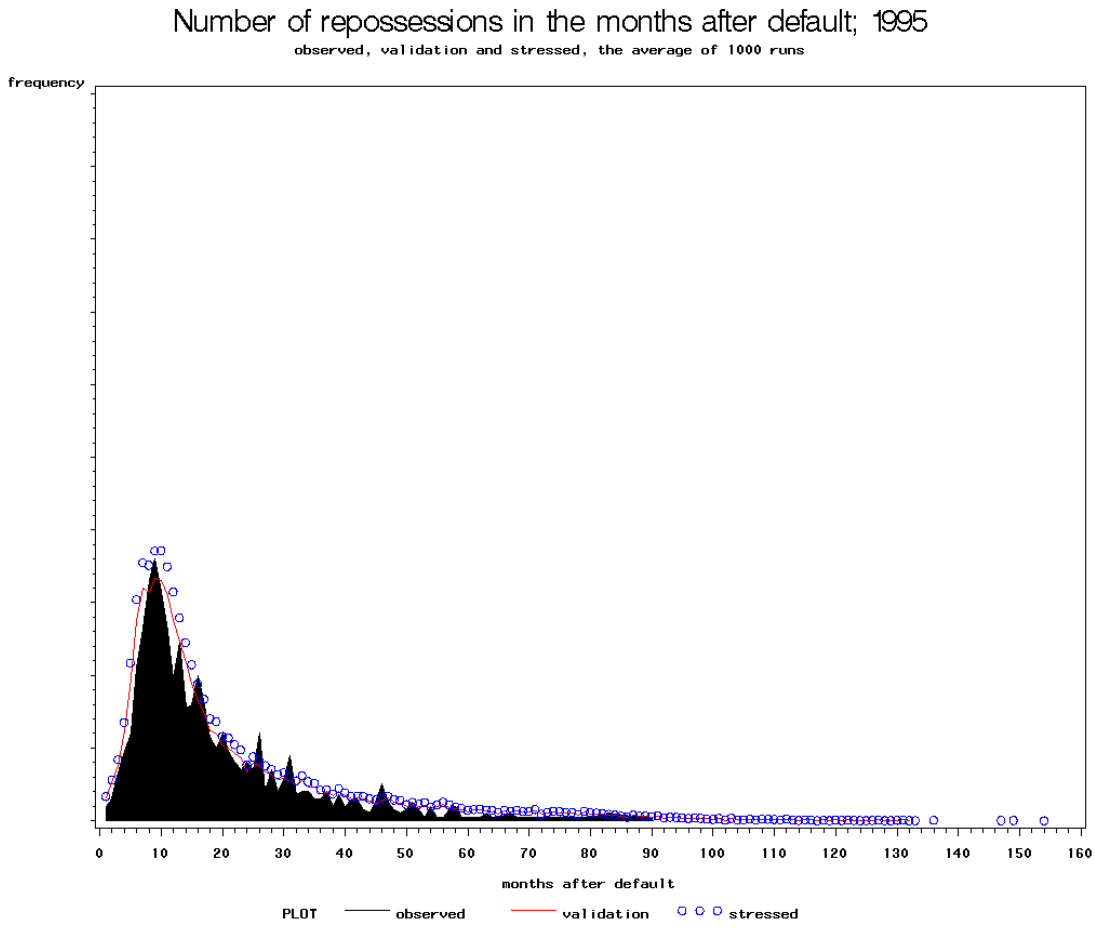

Figure 5.11: Number of observed repossessions (solid) in the months after default and average (over 1,000 runs) predicted number of repossessions for validation (line) and stressed (circles) simulations, for 1995 cohort. To allow for direct comparison, the vertical axis is the same as in Figure 5.12.

A number of observations can be made here. Firstly, the number of repossessions predicted for 1995 in each month following default is close to what that was historically observed in the dataset. 1995 is a fairly typical year and the model is predicting based on what it was trained to do. However, the model overestimates the number of predictions in 1991, so it seems to have identified that 1991 is a downturn year and has predicted for more repossessions, but not at the correct times. At times where the economic outlook is poor, banks are less likely to enter repossession procedures because it would mean that they would be trying to sell these properties in a depressed housing market, which would mean a lower selling price on top of the haircut that would be expected. Secondly, we see that the repossession pattern from the two simulations and the two cohort years are

very similar, which implies that the way the macroeconomic variable (HPIG) was incorporated into the survival model might not be sensitive enough.

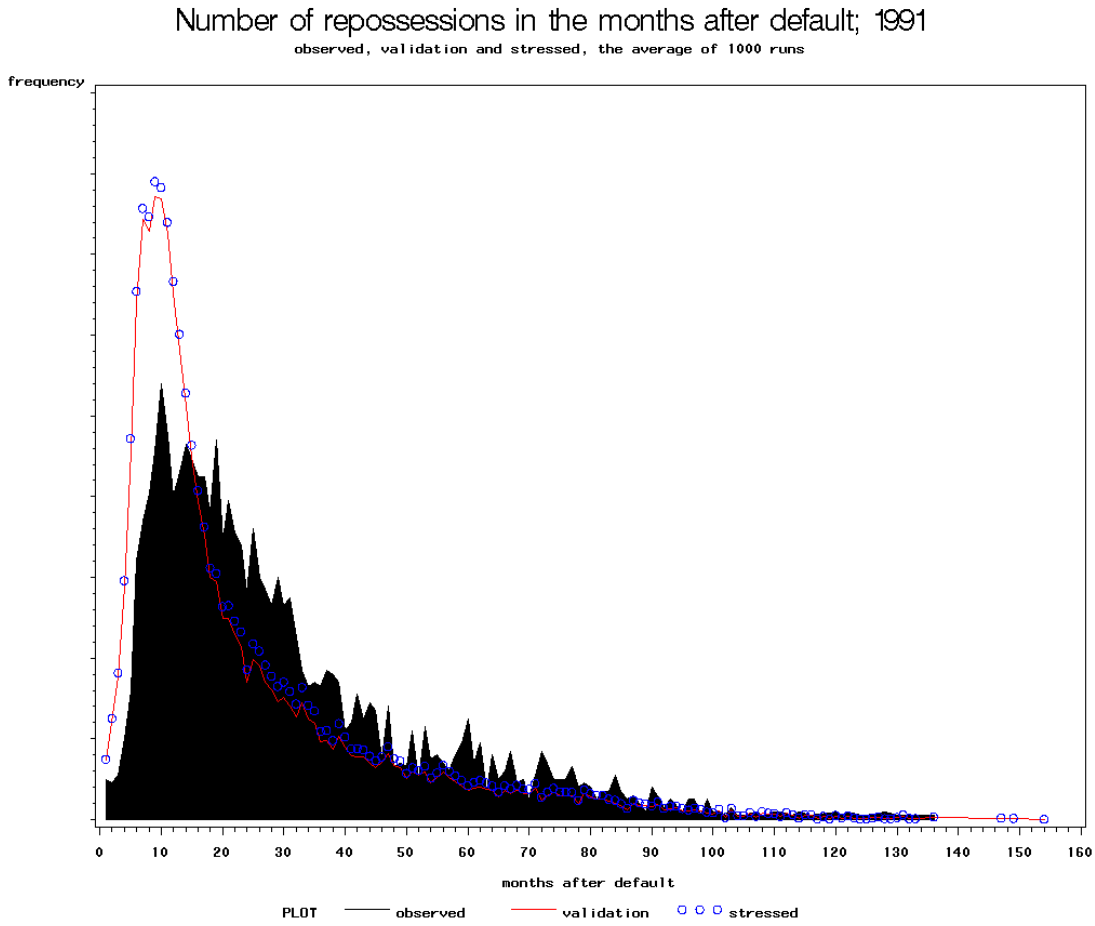

Figure 5.12: Number of observed repossessions (solid) in the months after default and average (over 1,000 runs) predicted number of repossessions for validation (line) and stressed (circles) simulations, for 1991 cohort. To allow for direct comparison, the vertical axis is the same as in Figure 5.11.

5.5.3.3 Distribution of LGD

The main goal of the development of this model is the prediction of LGD, and what are given by the model are LGD estimates for individual accounts within each separate run. Given that the simulation is run 1,000 times, and that the cohort datasets are quite large (about 3,000 observations for 1995 and 7,000 observations for 1991), there is plenty of output data here. The aim would be to use the data to calculate some confidence boundaries, without compromising individual information. In order to do this, for each individual

observation *j*, we look at its 1,000 predictions of LGD, and extract the median, $P50_j$, and 95th percentile, $P95_j$, values. The median is chosen over the mean because the difference between losses due to repossession (which could lead to some loss) and closure (which would mean zero loss) is large and distinct. Also, only a percentage of default loans will end in repossession so a sizeable number of observations will be predicted to be closed (i.e. have zero loss). Taking the mean for a distribution which is very skewed and has a pronounced point density at one end is not very meaningful, so taking the median is a more appropriate measure of centrality.

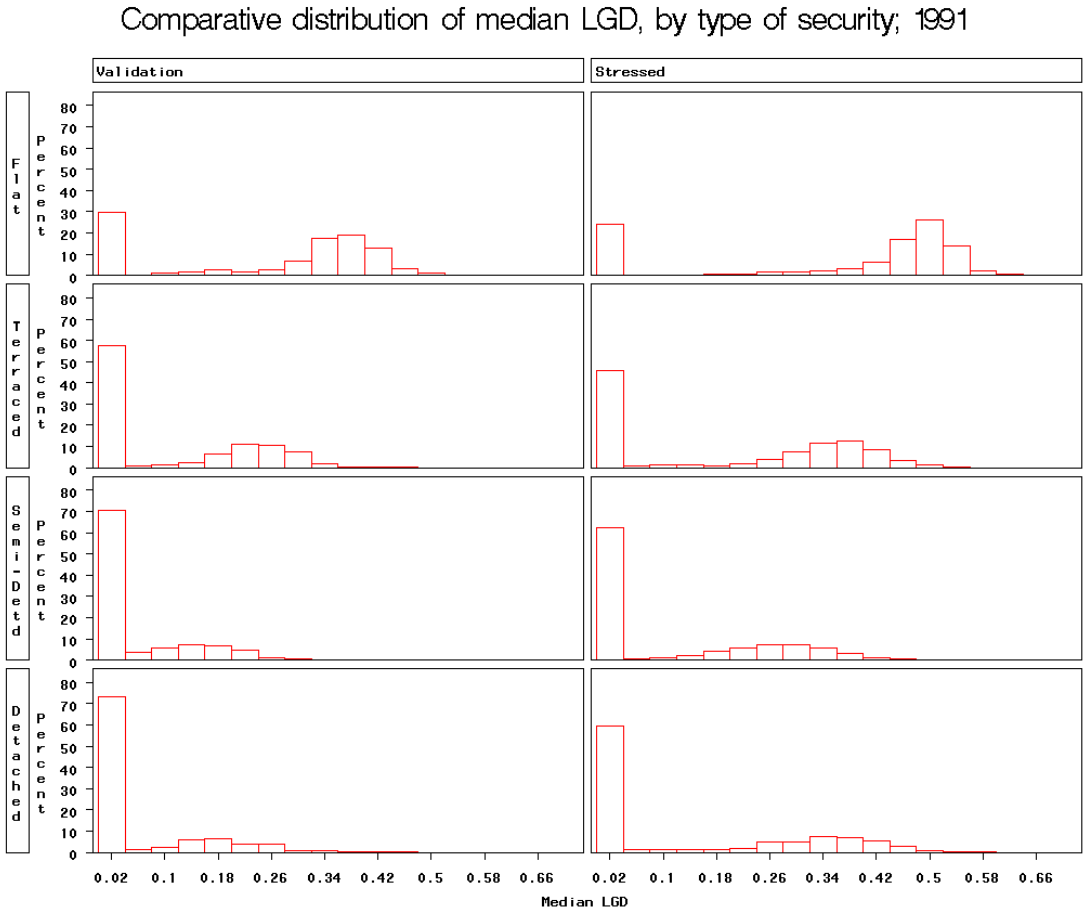

Figure 5.13: Distribution of median predicted LGD for validation (left panel) and stressed simulations (right panel) for 1991 cohort, segmented by type of security

The 5th percentile value is usually a statistic of interest, but is left out here because LGD is constrained to be a non-negative value (a profit or negative LGD would be recorded as zero LGD), and this value is likely to be zero. With the P50 and P95 values of each observation, the distribution of median

predicted LGD and the distribution of the 95th upper boundary predicted LGD can be determined. Note that the 95th percentile LGD is a value compiled by taking the bottom 5th percentile (of 1,000 runs) of LGD of each account, essentially putting together a distribution of LGD from the worst runs. By doing so, we hope to get a distribution of typical LGD (via the P50 distribution) as well as downturn LGD and potential losses (via the P95 distribution).

Because we make similar observations for the two cohorts of loans, only the graphs for 1991 are included here.

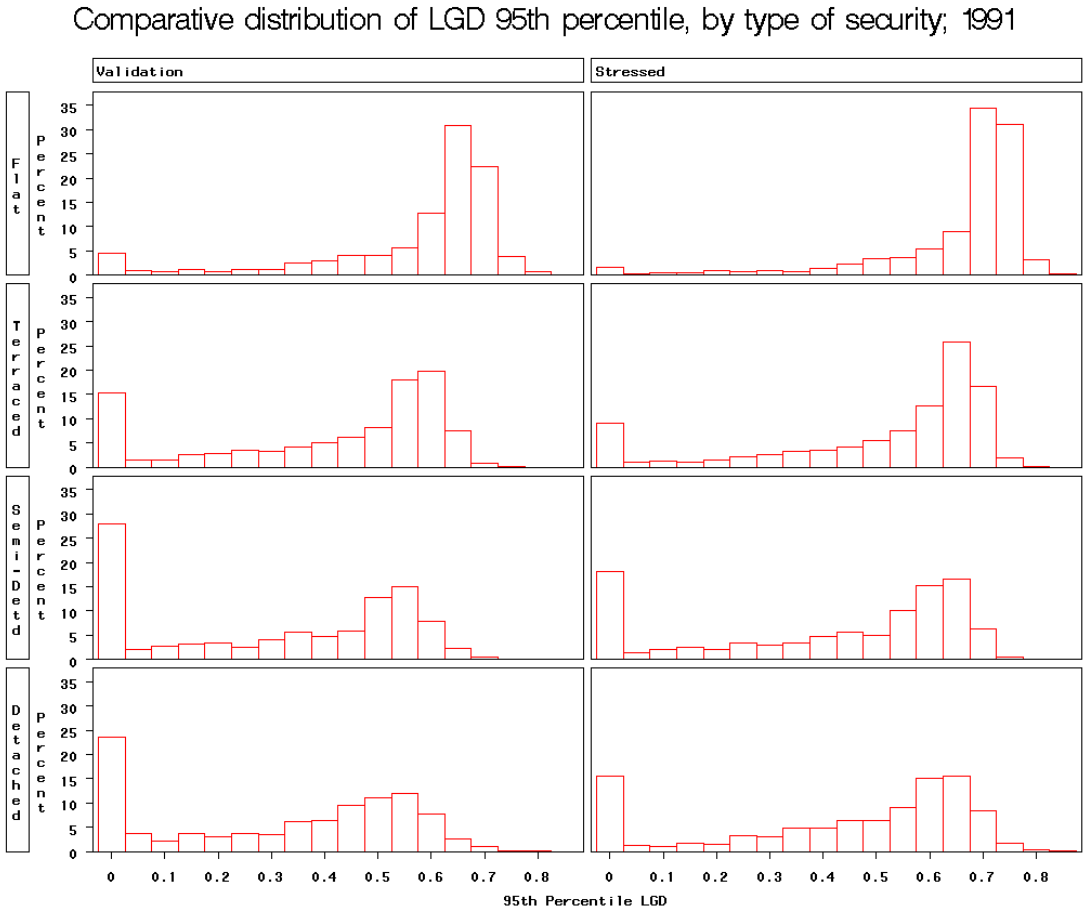

Figure 5.14: Distribution of 95th percentile predicted LGD for validation (left panel) and stressed simulations (right panel) for 1991 cohort, segmented by type of security

Firstly, we examine the distribution of LGD, segmented by type of security. Figures 5.13 and 5.14 give the distribution of 50th percentile predicted LGD

and 95th percentile LGD respectively, for the cohort of 1991, by type of security. Clearly, the distribution of LGD for the stressed simulation is more severe than that produced by the validation simulation. The stressed scenarios have fewer occasions of zero loss and where there is loss, there is higher loss. In general, we still see higher loss rates for lower-range properties, but the LGD of higher-range properties appears to be more affected by poor economic conditions (see Table 5.2).

Table 5.2: Mean LGD values for validation and stressed simulation, compared against observed LGD, segmented by type of security, for cohort of 1991. Top panel gives the mean LGD for the 50th percentile LGD; bottom panel gives the mean LGD for the 95th percentile LGD.

50th Percentile LGD				
Security	Mean Actual	Mean of predicted median from validation simulation	Mean of predicted median from stressed simulation	Increase (from validation to stressed)
Flat	0.238	0.245	0.356	0.110
Terraced	0.121	0.100	0.186	0.086
Semi-Detached	0.090	0.046	0.101	0.056
Detached	0.087	0.050	0.131	0.081
95th Percentile LGD				
Security		Mean of predicted median from validation simulation	Mean of predicted median from stressed simulation	Increase (from validation to stressed)
Flat		0.569	0.655	0.086
Terraced		0.406	0.509	0.102
Semi-Detached		0.309	0.413	0.104
Detached		0.319	0.439	0.119

Table 5.2 gives the mean LGD values (the 50th percentile and 95th percentile) from the validation and stressed simulations, of each security. The observed mean LGD values have also been included as an indication of the sensible values produced.

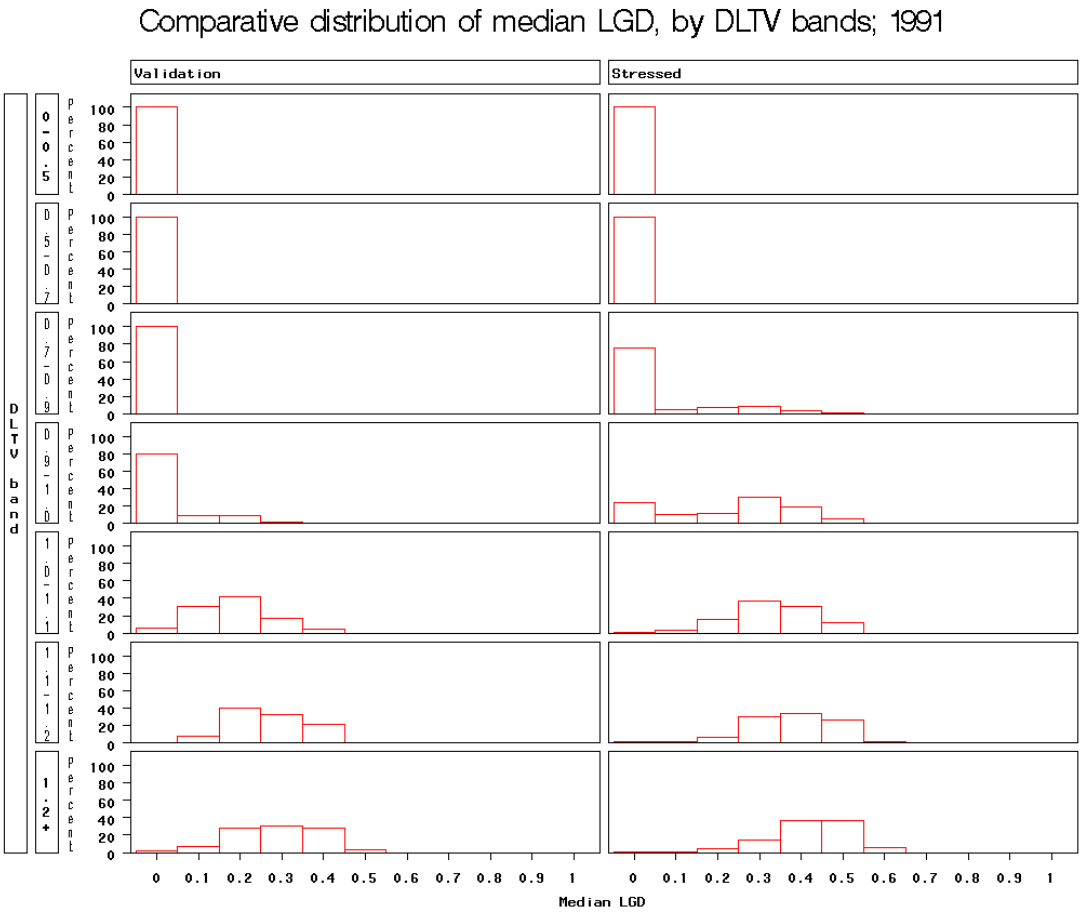

Figure 5.15: Distribution of median predicted LGD for validation (left panel) and stressed simulations (right panel) for 1991 cohort, segmented by DLTV bands

Next, Figures 5.15 and 5.16 give the distribution of 50th and 95th percentile predicted LGD, respectively, now segmented by DLTV bands. We make similar observations here. The distribution of LGD for the stressed simulation is more severe than that produced by the validation simulation. The model once again produces sensible predictions of LGD for the different DLTV bands. Higher loss rates are observed for higher DLTV bands, but, interestingly, the results suggest that the increase in losses during stressed situations could be higher for mid-range DLTV bands (see Table 5.3).

Table 5.3: Mean LGD values for validation and stressed simulation, compared against observed LGD, segmented by DLTV bands, for cohort of 1991. Top panel gives the mean LGD for the 50th percentile LGD; bottom panel gives the mean LGD for the 95th percentile LGD.

50th Percentile LGD					
DLTV Band		Mean Actual	Mean of predicted median from validation simulation	Mean of predicted median from stressed simulation	Percentage Increase (from validation to stressed)
1	DLTV <= 0.5	0.015	0.000	0.000	0.000
2	0.5 < DLTV <= 0.7	0.024	0.000	0.000	0.000
3	0.7 < DLTV <= 0.9	0.066	0.000	0.064	0.064
4	0.9 < DLTV <= 1.0	0.122	0.032	0.227	0.195
5	1.0 < DLTV <= 1.2	0.192	0.184	0.331	0.147
6	1.1 < DLTV <= 1.2	0.246	0.267	0.376	0.109
7	DLTV > 1.2	0.290	0.287	0.420	0.133
95th Percentile LGD					
DLTV Band			Mean of predicted median from validation simulation	Mean of predicted median from stressed simulation	Percentage Increase (from validation to stressed)
1	DLTV <= 0.5		0.011	0.089	0.078
2	0.5 < DLTV <= 0.7		0.200	0.373	0.172
3	0.7 < DLTV <= 0.9		0.422	0.551	0.129
4	0.9 < DLTV <= 1.0		0.524	0.625	0.101
5	1.0 < DLTV <= 1.2		0.569	0.645	0.076
6	1.1 < DLTV <= 1.2		0.600	0.670	0.069
7	DLTV > 1.2		0.625	0.697	0.072

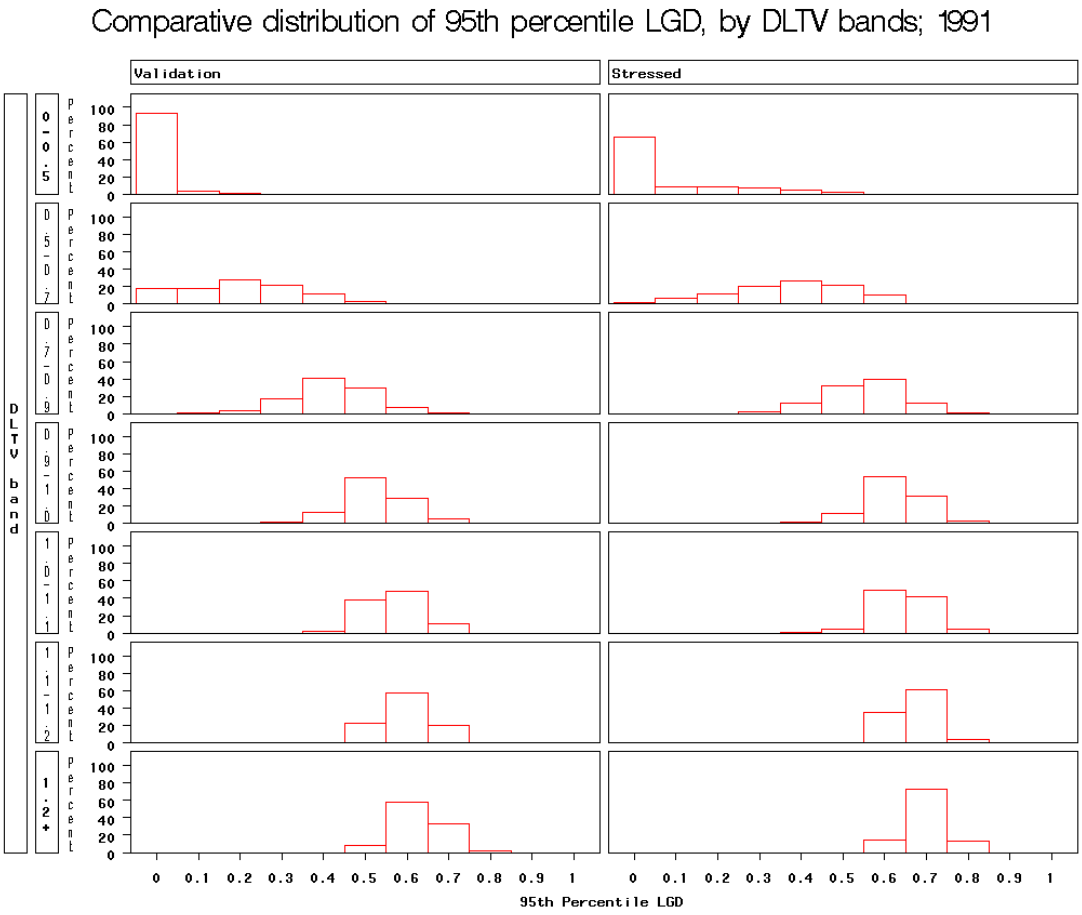

Figure 5.16: Distribution of 95th percentile predicted LGD for validation (left panel) and stressed simulations (right panel) for 1991 cohort, segmented by DLTV bands

5.6 Conclusions

In this chapter, we have set out an approach to model the time taken for defaulted residential mortgage loans to experience some event, either a repossession – which usually leads to loss – or a closure – which is commonly assumed to have zero loss. The motivation for doing so was that the time it takes for a loan to go from default to repossession will affect the amount of losses the bank will suffer. It was expected that time to repossession is affected by some loan characteristics (for example, DLTV), but that it will also depend on the state of the economy. During a downturn, property prices generally decrease which could leave the bank trying to sell the property at reduced prices in order to cover the outstanding balance on the loan, so more losses are expected to occur.

The method of survival analysis was selected for a number of reasons. For one, it is able to model time to event and handle censoring, where accounts simply have not had enough time to experience the event. The defaulted accounts could also experience one of two mutually exclusive events, repossession or closure. Finally, it can accommodate time-dependent variables, which would allow for the inclusion of macroeconomic variables which vary with time.

A number of issues were addressed and covered in this work.

Firstly, using similar loan characteristics and HPIG, two response survival models were developed to estimate the time from default to repossession and time from default to closure, creating a competing risks survival model. We find that survival analysis models could be a good class of models in the analysis and prediction of LGD because they are able to give some perspective on time in terms of when recoveries start arriving after default. They were also able to incorporate the main housing industry macroeconomic variable, the HPI, when regression models were unable to (macroeconomic variables were either insignificant or were not able to bring any obvious impact to prediction of LGD, as shown in Chapter 4). These time-dependent variables give valuable insight on how drivers of risk are different for different types of securities of different DLTV bands in different economic climates.

Secondly, the application of a Monte Carlo simulation allowed for the translation of (conditional) survival probabilities into predicted events and their corresponding predicted event times. This simulation was done on two selected loan cohorts: 1995 representing a non-downturn year, and 1991 representing a downturn year.

Next, the combination of survival analysis models and a Monte Carlo simulation provided an appropriate framework within which stress testing can be carried out. A stressed HPIG scenario was also created and the simulation results from this stressed scenario were compared against what was observed for the original scenario (validation simulation).

Using the output from these simulations, three types of performance measures were produced for each cohort of loans: the distribution of predicted total loss, the pattern of predicted repossession in each month following default, and distributions of predicted LGD.

When looking at prediction of total loss, we find that the model somewhat overestimates losses, which could be due to a number of reasons. The model was developed on a set of loans encompassing the economic booms and busts from the mid 1980s to the early 2000s, but was used to validate only a single cohort of loans at a time. Also, the calculation of total loss involved some external components including the estimation of time between repossession and sale and a haircut model previously developed. However, the macroeconomic variables obviously still have an impact on LGD prediction because the model predicted for higher losses in stressed situations, with a 75% increase in losses for the 1995 cohort and 50% increase in the 1991 cohort. Note that these figures are based solely on accounts that are already in default, which would likely increase further when the potential impact of macroeconomic variables on default rates is also taken into consideration.

Looking at the predicted repossession in each month following default, we found that the model performed fairly well for 1995 but less so for 1991. The model was able to anticipate a higher number of repossessions during a downturn year, but was not able to accurately predict the time at which repossession would happen. During a downturn year, where house prices are depressed and transacted prices of properties are low, banks might be reluctant to start repossession procedures. Because the pattern of repossession observed from the validation and stressed simulations are similar, we suspect that the survival model is still not sensitive enough to changes in HPIG. We believe more work may be required here, especially in the development of a separate model from which downturn LGD can be more reliably predicted.

In terms of LGD prediction, the distributions of the 50th and 95th percentile of predicted LGD were generated. These would give an indication of typical as well as downturn LGD and potential losses. The LGD distributions are

displayed segmented according to type of security and DLTV bands. We find that the validation simulation was able to produce reasonable predictions of LGD, and these predictions stand even when the loans were segmented. In line with expectations, lower-range securities (i.e. flats and terraced properties), as well as higher-range DLTV bands, are observed to experience higher LGD rates.

We believe more work is required in the investigation of time periods between default and repossession, and repossession and sale, as well as the differentiation of accounts that would undergo repossession or otherwise. More work on the impact of risk drivers on risk of repossession and/or closure would be beneficial. It would also be interesting to validate this model using data from the credit crisis of 2008, and check if the level of losses is as predicted. Finally, based on this work, the potential benefits of building separate models for downturn and non-downturn years are highlighted.

CHAPTER 6. CONCLUSIONS AND FURTHER RESEARCH

In this chapter, we give a summary of the major contributions of this thesis, followed by some suggested areas for further research.

6.1 Review of chapters and major conclusions

Using a large set of recovery data of residential mortgage defaults from a major UK bank, a number of LGD models were explored and developed in this thesis.

The first part of this thesis investigated the use of a two-stage approach for modelling the LGD of mortgage loans. The first component model in this approach was a Probability of Repossession Model, which was developed to predict repossession (for accounts that go into default), and the second component model was a Haircut Model, which was developed to investigate the level of discount a repossessed property would be expected to undergo.

A Probability of Repossession Model was developed with more than just the commonly used loan-to-value ratio at default (DLTV), and was shown to be significantly better. In order to account for the variability of haircut, it was necessary to model the distribution of predicted haircut and this was done using two sub-models: one to estimate the mean and the other to estimate the standard deviation. The two component models were then combined into an expected loss percentage.

In this chapter, we found that the prediction of mortgage loan LGD benefited from the differentiation of the two different drivers (i.e. repossession risk and sale price haircut) of mortgage loss. Performance-wise, this two-stage LGD model was shown to do better than a single-stage LGD model (which directly modelled LGD from loan and collateral characteristics), as it achieved

a better R-square value, and it more accurately matched the distribution of observed LGD (see Figures 3.6 and 3.7) .

Since it was suggested in the literature that the macroeconomy could be correlated with recovery rates, we subsequently hypothesized that the inclusion of relevant macroeconomic variables might improve prediction performance of these credit models. Therefore, in the second part, we investigated the inclusion of a number of macroeconomic variables into two different retail LGD models: the first was the mortgage loan dataset described and used in the first part, and the second was an unsecured personal loans dataset. Although there were strong indications from the literature on corporate-sector credit risk that macroeconomic variables are able to improve predictions of probability of default (PD) and loss given default (LGD), this was not consistently seen here (see Figures 4.3 and 4.7). Only a small number of macroeconomic variables turned up statistically significant, and even fewer were able to bring about any positive contribution towards LGD prediction.

In the case of mortgage loan LGD prediction, interest rates gave the best improvements in both component models (i.e. Probability of Repossession Model and Haircut Model). Interestingly, the House Price Index, despite being the leading economic indicator in the housing market, did not, probably because it was already unavoidably included in some of the base models' variables. Overall, the mortgage loan LGD model experienced an increase in overall R-square with the inclusion of macroeconomic variables. However, the distributions of predicted LGD produced by the base and macroeconomic models, and mean predicted values of LGD of each quarter were very similar (see Figures 4.5 and 4.6 respectively). It was also observed that the macroeconomic model was able to produce better estimates for LGD during downturn periods, but ended up underestimating LGD during non-downturn periods (see Figure 4.4).

In the case of unsecured personal loan LGD prediction, the only macroeconomic variable that turned up statistically significant was net lending growth at default, but this did not bring any improvement to overall

R-square of the LGD model, or prediction values (see Figure 4.8). This seemed to indicate that personal loans LGD are less affected by the economy.

In the first two parts, LGD models were developed using a combination of logistic and linear regression models, which are popular methods for credit risk modelling in industry. These models have produced decent predictions of LGD, and although there were some improvements in the individual component models (i.e. Probability of Repossession Model and Haircut Model) as well as overall R-square values, we saw that macroeconomic variables caused model prediction to be skewed towards the downturn years. Also, the regression methods would always be measuring macroeconomic indicators at just one snapshot in time per observation (e.g. at default) whereas these indicators essentially change over the course of the workout. Survival models, on the other hand, are able to take into account these time-dependent variables and produce estimates for the likelihood of an event happening at each time step, which could then predict not just *whether* an event will happen, but also *when* it is most likely to occur. They would also able to account for observations that have not yet experienced the event (i.e. censoring).

In the third and final part of this thesis, we therefore revisited the development of an LGD model for residential mortgage loans using survival models. The defaulted accounts could experience one of two mutually exclusive events, repossession or closure and two response survival models were developed to model the timing of each event. Significant variables for both survival models include loan-to-value at default (DLTV), type of security and the house price index growth rate (HPIG), as well interaction variables between the loan-related variables and the house price index growth rate. It was also observed that the time-dependent variable (i.e. the HPIG) gave valuable insight on how drivers of risk were different for different types of securities of different DLTV bands in different economic climates (see Figures 5.1 and 5.3).

We then went on to show how this model could be used for stress testing. To do so, we applied a Monte Carlo simulation, which not only allowed us to translate the (conditional) survival probabilities into a predicted event and a

predicted time, but also provided an appropriate framework from which stress testing could be carried out. With the predicted event and event time estimate, it was then possible to estimate a discounted loss (whereas the previous two chapters looked at nominal loss).

Two types of simulation were done here, the first using historically observed HPIG values ("validation simulation") and the second using stressed HPIG rates ("stressed simulation") created for this purpose. By taking into account how HPIG would affect not just survival probabilities but also DLTV values, haircut variability and discounting (if applicable), we were able to produce distributions for total losses and LGD. The simulations of 1,000 runs were carried out on two selected default cohorts: 1995 and 1991, each representing a year with a very different economic outlook.

We made a number of observations here: first, that macroeconomic variables did affect survival time and had a substantial impact on potential losses. The model predicted higher losses in stressed situations, with a 75% increase in losses for the 1995 cohort and 50% in the 1991 cohort (see Figures 5.9 and 5.10 respectively). Note that these estimates were based solely on accounts that were already in default; hence, losses would likely increase further when the economy's potential impact on default rates would also be taken into consideration. Secondly, the model developed was also able to give a fairly accurate representation of when repossessions would happen in the months after default (see Figure 5.11). It was also able to predict a higher number of repossessions during a downturn year (see Figure 5.12), but was less successful at getting the timing right. Next, the simulation of losses produced distributions of the 50th and 95th percentile of predicted LGD, which gave an indication of typical as well as downturn LGD and potential losses. The validation simulation produced reasonable predictions of LGD, both when loans were segmented by type of security and DLTV bands. In line with expectations, lower-range securities (i.e. flats and terraced properties), as well as higher-range DLTV bands, were observed to experience higher LGD rates.

6.2 Issues for further research

There are a number of areas for further work.

In the development of the Haircut Model, it was observed that one of the variables, the ratio of valuation of the security at default over the average property value in the region, was not linearly related to haircut. We binned the variable in order to use it as an explanatory variable in the linear regression model, but alternate techniques like spline regression could be investigated.

We also found that different types of loans are affected differently by macroeconomic variables, but more work may be done to investigate if macroeconomic variables are non-linearly related to recovery rates for mortgage defaults. If so, possible techniques include binning the macroeconomic variable or experimenting with interaction terms between macroeconomic economic variables and loan-related ones.

A mortgage loan could also experience more than the two events (i.e. repossession and non-repossession) currently defined here. For the repossessions, further research on the time period between repossession and sale could improve prediction of the entire default to repossession to sale process and improve prediction of LGD further. Given that such data were available, it would also be interesting to further distinguish between different types of outcomes, in the event that there is no repossession (e.g. the default could be cured, a new payment plan could be agreed, or the loan is written off altogether). If so, one could investigate the factors that contribute to either event. This could help banks identify individuals who might need extra help in their loan repayments and prevent the larger losses associated with the writing-off of loans.

We note also that the economic downturn experienced by the UK during the early 1990s is different to that of 2008. For example, during the 1990s, interest rates were high and in the region of 10%, whereas the interest rates

in 2008-09 were less than 2%. If updated data were to become available, it would be useful to re-calibrate and validate the LGD models developed here.

Finally, with respect to stress testing and downturn predictions, in future work, we would like to assess the potential advantages of building a separate LGD model for downturn and non-downturn years. This is especially apparent in mortgage loans because the value of the security, which affects the recovery amount, can be affected by the economy.

In summary, all three parts of this thesis explore the development of LGD models for retail loans, specifically in the case of residential mortgages. Both regression and survival analysis methods are covered, as well as the relevance of macroeconomic variables in the prediction of LGD. The significance and importance of stress testing is acknowledged and a suitable framework is investigated and proposed. It is believed that this work contributes significantly to the existing academic literature by advancing our understanding of LGD models for retail exposures, and mortgage lending in particular.

LIST OF REFERENCES

Acharya, V. V., Bharath, S. T. & Srinivasan, A. (2003). Understanding the Recovery Rates on Defaulted Securities.

Allen, L., Delong, G. & Saunders, A. (2004) Issues in the credit risk modeling of retail markets. *Journal of Banking & Finance,* 28**,** 727-752.

Altman, E. I. (2006). Default Recovery Rates and LGD in Credit Risk Modelling and Practice: An Updated Review of the Literature and Empirical Evidence. This is an updated and expanded review of the original artical by Altman, Resti and Sironi (2006).

Altman, Edward i., Brady, B., Resti, A. & Sironi, A. (2005) The Link between Default and Recovery Rates: Theory, Empirical Evidence, and Implications. *The Journal of Business,* 78**,** 2203-2228.

Altman, E. I. & Kishore, V. M. (1996) Almost Everything You Wanted to Know about Recoveries on Default Bonds. *Financial Analysts Journal,* 52**,** 57-64.

Altman, E. I., Resti, A. & Sironi, A. (2001). Analyzing and Explaning Default Recovery Rates. A report submitted to The International Swaps & Derivatives Association.

Andersen, P. K. (1992) Repeated Assessment of Risk Factors in Survival Analysis. *Statistical Methods in Medical Research,* 1**,** 297-315.

Asarnow, E. & Edwards, D. (1995) Measuring Loss on Defaulted Bank Loans: A 24-Year Study. *Journal of Commercial Lending,* 77**,** 11-23.

Avery, R. B., Bostic, R. W., Calem, P. S. & Canner, G. B. (1996) Credit Risk, Credit Scoring, and the Performance of Home Mortgages. *Federal Reseve Bulletin***,** 621-648.

Banasik, J., Crook, J. N. & Thomas, L. C. (1999) Not if but When will Borrowers Default. *The Journal of the Operational Research Society,* 50**,** 1185-1190.

Basel Committee on Banking Supervision (2001). History of the Basel Committee and its Membership

Basel Committee on Banking Supervision (2005). An Explanatory Note on the Basel II IRB Risk Weight Functions

Basel Committee on Banking Supervision (2006). Results of the Fifth Quantitative Impact Study (QIS 5)

Bellotti, T. & Crook, J. N. (2009a) Credit Scoring with Macroeconomic Variables using Survival Analysis. *Journal of the Operational Research Society,* 60**,** 1699-1707.

Bellotti, T. & Crook, J. N. (2009b). Loss Given Default Models for UK Retail Credit Cards. Credit Research Centre (Edinburgh) Working Paper 09/1

Berkowitz, J. (2000) A Coherent Framework for Stress-Testing. *Journal of Risk,* 2**,** 1-11.

Black, F. & Scholes, M. (1973) The Pricing of Options and Corporate Liabilities. *The Journal of Political Economy,* 81**,** 637-654.

Blaschke, W., Jones, M. T., Majnoni, G. & Peria, S. M. (2001). Stress Testing of Financial Systems: An Overview of Issues, Methodologies, and FSAP Experiences. An International Monetary Fund (IMF) Working Paper WP/01/88

Breeden, J. L., Thomas, L. C. & Mcdonald, J. I. (2008) Stress-Testing Retail Loan Portfolios with Dual-Time Dynamics. *The Journal of Risk Model Validation,* 2**,** 43-62.

Bruche, M. & González-Aguado, C. (2010) Recovery rates, default probabilities, and the credit cycle. *Journal of Banking & Finance,* 34**,** 754-764.

Calem, P. S. & Lacour-Little, M. (2004) Risk-based capital requirements for mortgage loans. *Journal of Banking & Finance,* 28**,** 647-672.

Campbell, T. S. & Dietrich, J. K. (1983) The Determinants of Default on Insured Conventional Residential Mortgage Loans. *The Journal of Finance,* 38**,** 1569-1581.

Chen, L. S., Yen, M. F., Wu, H. M., Liao, C. S., Liou, D. M., Kuo, H. S. & Chen, T. H. H. (2005) Predictive survival model with time-dependent prognostic factors: development of computer-aided SAS Macro program. *Journal of Evaluation in Clinical Practice,* 11**,** 181-193.

Coleman, A., Esho, N., Sellathurai, I. & Thavabalan, N. (2005) Stress Testing Housing Loan Portfolios: A Regulatory Case Study. *Basel Committee on Banking Supervision Conference on Banking and Financial Stability - A Workshop on Applied Banking Research.* Vienna.

Committee on the Global Financial System (2005). Stress Testing at Major Financial Institutions: Survey Results and Practice

Cox, D. R. (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological),* 34**,** 187-220.

Credit Suisse (1997). CreditRisk+: A Credit Risk Management Framework. Credit Suisse Financial Products Publication.

Crouhy, M., Galai, D. & Mark, R. (2000) A comparative analysis of current credit risk models. *Journal of Banking & Finance,* 24**,** 59-117.

Delong, E. R., Delong, D. M. & Clarke-Pearson, D. L. (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics,* 44**,** 837-845.

Federal Register (2007). Risk-Based Capital Standards: Advanced Capital Adequacy Framwork - Basel II; Final Rule

Fernandez, G. C. J. (2007) Effects of Multicollinearity in All Possible Mixed Model Selection. *PharamaSUG Conference (Statistics & Pharmacokinetics).* Denver, Colorado.

Figlewski, S., Frydman, H. & Liang, W. (2008) Modelling the Effect of Macroeconomic Factors on Corporate Default and Credit Rating Transitions. *5th Annual Credit Risk Conference.* NYU Stern School of Business, New York.

Financial Services Authority (2005). Stress Testing. FSA Discussion Paper 05/02

Financial Services Authority (2009). Prudential Sourcebook for Banks, Building Societies and Investment Firms

Financial Services Authority Expert Group (2009). Expert Group Paper on Loss Given Default Other

Financial Stability Institute (2004). Implementation of the New Capital Adequacy Framework in Non-Basel Committee Member Countries Occasional Paper 4

Financial Stability Institute (2006). Implementation of the New Capital Adequacy Framework in Non-Basel Committee Member Countries Occasional Paper 6

Frye, J. (2000a) Collateral Damage. *RISK,* 13.

Frye, J. (2000b) Depressing Recoveries. *RISK,* 13**,** 108-111.

Gupton, G. M. & Stein, R. M. (2001) A Matter of Perspective. *CREDIT,* 2.

Gupton, G. M. & Stein, R. M. (2002). LOSSCALC: Model for Predicting Loss Given Default (LGD). Moody's KMV Research Publication

Gupton, G. M. & Stein, R. M. (2005). LOSSCALC V2: Dynamic Prediction of LGD. Moody's KMV Research Publication

Hamilton, D. T., Gupton, G. M. & Berthault, A. (2001). Default and Recovery Rates of Corporate Bond Issuers: 2000. Moody's KMV Research Publication

Hand, D. J. & Henley, W. E. (1997) Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 160**,** 523-541.

Helbling, T. & Terrones, M. (2003) Real and Financial Effects of Bursting Asset Price Bubbles. *World Economic Outlook,* Chapter 2: When Bubbles Burst**,** 61-76.

Hu, Y. T. & Perraudin, W. (2002). The Dependence of Recovery Rates and Defaults. Unpublished Paper

Hui, C. H., Lo, C. F., Wong, T. C. & Man, P. K. (2006) Measuring provisions for collateralised retail lending. *Journal of Economics and Business,* 58**,** 343-361.

Jarrow, R. (2001) Default Parameter Estimation Using Market Prices. *Financial Analysts Journal,* 57**,** 75-92.

Jokivuolle, E. & Peura, S. (2003) Incorporating Collateral Value Uncertainty in Loss Given Default Estimates and Loan-to-value Ratios. *European Financial Management,* 9**,** 299-314.

Jones, E. P., Mason, S. P. & Rosenfeld, E. (1984) Contingent Claims Analysis of Corporate Capital Structures: An Empirical Investigation. *The Journal of Finance,* 39**,** 611-625.

Loterman, G., Brown, I., Martens, D., Mues, C. & Baesens, B. (2009) Benchmarking State-Of-The-Art Regression Algorithms for Loss Given Default Modelling. *Credit Scoring and Credit Control XI Conference.* Edinburgh, UK.

Lucas, A. (2003) The New Basel Accord - Implications. *Credit Scoring and Credit Control Conference 2003.* Edinburgh, UK.

Lucas, A. (2006) Basel II Problem Solving. *Conference on Basel II & Credit Risk Modelling in Consumer Lending* Southampton, UK.

Matuszyk, A., Mues, C. & Thomas, L. C. (2010) Modelling LGD for Unsecured Personal Loans: Decision Tree Approach. *Journal of the Operational Research Society,* 61**,** 393-398.

Mcdonald, R., Matuszyk, A. & Thomas, L. C. (2010) Application of Survival Analysis to Cash Flow Modelling for Mortgage Products. *OR Insight,* 23**,** 1-14.

Merton, R. C. (1974) On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. *The Journal of Finance,* 29**,** 449-470.

Narain, B. (1992) Survival Analysis and the Credit Granting Decision. IN Thomas, L. C., Crook, J. N. & Edelman, D. B. (Eds.) *Credit Scoring and Credit Control.* Oxford University Press, USA.

Peng, G. (2004) Testing Normality of Data Using SAS. *PharmaSUG.* San Diego, California.

Pennington-Cross, A. (2003). Subprime And Prime Mortgages: Loss Distributions. Office of Federal Housing Enterprise Oversight (OFHEO) Working Paper 03-1

Pennington-Cross, A. (2010) The Duration of Foreclosures in the Subprime Mortgage Market: A Competing Risks Model with Mixing. *The Journal of Real Estate Finance and Economics,* 40**,** 109-129.

Pesaran, M. H., Schuermann, T., Treutler, B.-J. & Weiner, S. M. (2006) Macroeconomic dynamics and credit risk: A global perspective. *Journal of Money Credit and Banking,* 38**,** 1211-1261.

Phillips, R. A. & Vanderhoff, J. H. (2004) The Conditional Probability of Foreclosure: An Empirical Analysis of Conventional Mortgage Loan Defaults. *Real Estate Economics,* 32**,** 571-587.

Qi, M. & Yang, X. (2009) Loss given default of high loan-to-value residential mortgages. *Journal of Banking & Finance,* 33**,** 788-799.

Quercia, R. G. & Stegman, M. A. (1992) Residential Mortgage Default: A Review of the Literature. *Journal of Housing Research,* 3**,** 341-379.

Rodriguez, A. & Trucharte, C. (2007) Loss coverage and stress testing mortgage portfolios: A non-parametric approach. *Journal of Financial Stability,* 3**,** 342-367.

Rosenberg, E. & Gleit, A. (1994) Quantitative Methods in Credit Management: A Survey. *OPERATIONS RESEARCH,* 42**,** 589-613.

Schuermann, T. (2004) What Do We Know About Loss Given Default. IN Shimko, D. (Ed.) *Credit Risk: Models and Management.* 2nd ed., Risk Books.

Somers, M. & Whittaker, J. (2007) Quantile regression for modelling distributions of profit and loss. *European Journal of Operational Research,* 183**,** 1477-1487.

Sorge, M. (2004). Stress-Testing Financial Systems: An Overview of Current Methodologies. Bank for International Settlements (BIS) Working Paper No. 165

Stepanova, M. & Thomas, L. (2002) Survival Analysis Methods for Personal Loan Data. *OPERATIONS RESEARCH,* 50**,** 277-289.

Thomas, L. C. (1998) Methodologies for Classifying Applicants for Credit. IN Hand, D. J. & Jacka, S. D. (Eds.) *Statistics in Finance.* Hodder Arnold H&S.

Truck, S., Harpaintner, S. & Rachev, S. T. (2005). A Note on Forecasting Aggregate Recovery Rates with Macroeconomic Variables. Unpublished paper as part of thesis.

Von Furstenberg, G. M. (1969) Default Risk on FHA-Insured Home Mortgages as a Function of the Terms of Financing: A Quantitative Analysis. *The Journal of Finance,* 24**,** 459-477.

Wilson, T. C. (1998) Portfolio Credit Risk. *FRBNY Economic Policy Review,* 4**,** 71-82.

Wong, J., Fung, L., Fong, T. & Sze, A. (2004). Residential Mortgage Default Risk and the Loan-To-Value Ratio. Hong Kong Monetary Authority Quarterly Bulletin

APPENDIX A. PARAMETER ESTIMATES FOR TWO-STAGE AND SINGLE-STAGE MORTGAGE LGD MODEL

Table A1: Parameter estimates for Probability of Repossession Model *R0*

Variable	Variable explanation	Estimate	StdErr	WaldChiSq	ProbChiSq
Intercept	-	-3.069	0.028	12235.289	<0.01
DLTV	Loan to value at default	2.821	0.029	9449.349	<0.01

Table A2: Parameter estimates for Probability of Repossession Model *R1*

Variable	Variable Explanation	Estimate	StdErr	WaldChiSq	ProbChiSq
Intercept	-	-1.138	0.040	795.605	<0.01
LTV	Loan to value at loan application	2.101	0.040	2809.703	<0.01
TOB	Time on books (in years)	-0.188	0.003	2899.616	<0.01
Previous default	Indicator for previous default	0.102	0.034	8.869	<0.01
security0 (base)	Flat or other	-	-	-	-
security1	Detached	-0.625	0.031	413.989	<0.01
security2	Semi-detached	-0.670	0.024	787.436	<0.01
security3	Terraced	-0.421	0.021	395.497	<0.01

Table A3: Parameter estimates for Probability of Repossession Model *R2*

Variable	Variable Explanation	Estimate	StdErr	WaldChiSq	ProbChiSq
Intercept	-	-2.570	0.034	5769.803	<0.01
DLTV	Loan to value at default	2.679	0.029	8295.648	<0.01
Previous default	Indicator for previous default	-0.471	0.032	211.064	<0.01
security0 (base)	Flat or other	-	-	-	-
security1	Detached	-0.461	0.031	219.425	<0.01
security2	Semi-detached	-0.546	0.024	503.458	<0.01
security3	Terraced	-0.343	0.022	253.470	<0.01

Table A4: Parameter estimates for Haircut Model *H1*

Variable	Variable Explanation	Estimate	StdErr	ProbT	VIF
Intercept	-	0.508	0.009	<0.01	0.000
LTV	Loan to value at loan application	0.243	0.007	<0.01	1.136
TOB	Time on book (in years)	0.005	0.001	<0.01	1.251
VVAratio1 (base)	Value of property / region average <= 0.9	-	-	-	-
VVAratio2	0.9 < Value of property / region average <= 1.2	-0.005	0.004	0.248	1.134
VVAratio3	1.2 < Value of property / region average <= 1.5	-0.059	0.006	<0.01	1.149
VVAratio4	1.5 < Value of property / region average <= 1.8	-0.092	0.008	<0.01	1.127
VVAratio5	1.8 < Value of property / region average <= 2.4	-0.090	0.009	<0.01	1.161
VVAratio6	Value of property / region average > 2.4	-0.138	0.009	<0.01	1.226
Previous default	Indicator for previous default	0.042	0.006	<0.01	1.168

Propage1	Very old property (before 1919)	-0.085	0.003	<0.01	1.273
Propage2	Old property (1919-1945)	-0.032	0.004	<0.01	1.194
Propage3 (base)	Built after 1945	-	-	-	-
security0 (base)	Flat or other	-	-	-	-
security1	Detached	0.165	0.006	<0.01	1.875
security2	Semi-detached	0.129	0.004	<0.01	1.764
security3	Terraced	0.094	0.003	<0.01	1.739
region1	North	-0.112	0.010	<0.01	1.753
region2	Yorkshire & Humberside	-0.095	0.008	<0.01	2.898
region3	North West	-0.099	0.008	<0.01	3.163
region4	East Midlands	-0.100	0.008	<0.01	2.489
region5	West Midlands	-0.065	0.008	<0.01	2.449
region6	East Anglia	-0.067	0.009	<0.01	1.968
region7	Wales	-0.115	0.009	<0.01	2.140
region8	South West	-0.047	0.008	<0.01	3.272
region9	South East	-0.062	0.007	<0.01	6.348
region10	Greater London	-0.010	0.007	0.166	5.214
region11	Northern Ireland	-0.034	0.014	0.017	1.256
region12 (base)	Scotland or others / missing	-	-	-	-

Table A5: Parameter estimates for Haircut Model *H2*

Variable	Variable Explanation	Estimate	StdErr	ProbT	VIF
Intercept	-	0.591	0.008	<0.01	0.000
DLTV	Loan to value at default	0.162	0.005	<0.01	1.175
VVAratio1 (base)	Value of property / region average <= 0.9	-	-	-	-
VVAratio2	0.9 < Value of property / region average <= 1.2	-0.011	0.004	<0.01	1.126
VVAratio3	1.2 < Value of property / region average <= 1.5	-0.069	0.006	<0.01	1.141

VVAratio4	1.5 < Value of property / region average <= 1.8	-0.108	0.008	<0.01	1.116
VVAratio5	1.8 < Value of property / region average <= 2.4	-0.108	0.009	<0.01	1.149
VVAratio6	Value of property / region average > 2.4	-0.158	0.009	<0.01	1.209
Previous default	Indicator for previous default	0.064	0.005	<0.01	1.010
Propage1	Very old property (before 1919)	-0.079	0.003	<0.01	1.261
Propage2	Old property (1919-1945)	-0.030	0.004	<0.01	1.193
Propage3 (base)	Built after 1945	-	-	-	-
security0 (base)	Flat or other	-	-	-	-
security1	Detached	0.162	0.006	<0.01	1.874
security2	Semi-detached	0.126	0.004	<0.01	1.761
security3	Terraced	0.092	0.003	<0.01	1.736
region1	North	-0.109	0.010	<0.01	1.752
region2	Yorkshire & Humberside	-0.094	0.008	<0.01	2.897
region3	North West	-0.098	0.008	<0.01	3.159
region4	East Midlands	-0.112	0.008	<0.01	2.497
region5	West Midlands	-0.076	0.008	<0.01	2.454
region6	East Anglia	-0.102	0.009	<0.01	2.007
region7	Wales	-0.125	0.009	<0.01	2.141
region8	South West	-0.080	0.008	<0.01	3.325
region9	South East	-0.095	0.007	<0.01	6.489
region10	Greater London	-0.042	0.007	<0.01	5.323
region11	Northern Ireland	-0.030	0.014	0.040	1.256
region12 (base)	Scotland or Others / Missing	-	-	-	-

Table A6: Parameter estimates for Haircut Standard Deviation Model

Variable	Variable Explanation	Estimate	StdErr	ProbT
Intercept	-	0.181	<0.001	<0.01
TOB bins	Time on book (in years)	0.010	<0.001	<0.01

Table A7: Parameter estimates for single-stage LGD model

Variable	Variable Explanation	Estimate	StdErr	ProbT	VIF
Intercept	-	-0.093	0.005	<0.01	0.000
DLTV	Loan to value at default	0.230	0.002	<0.01	1.263
secondapp	Second applicant present	-0.003	0.001	0.012	1.105
VVAratio1	Value of property / region average <= 0.9	-0.049	0.004	<0.01	8.976
VVAratio2	0.9 < Value of property / region average <= 1.2	-0.050	0.004	<0.01	5.416
VVAratio3	1.2 < Value of property / region average <= 1.5	-0.035	0.004	<0.01	3.093
VVAratio4	1.5 < Value of property / region average <= 1.8	-0.018	0.005	<0.01	2.148
VVAratio5	1.8 < Value of property / region average <= 2.4	-0.018	0.005	<0.01	2.037
VVAratio6 (base)	Value of property / region average > 2.4	-	-	-	-
Previous default	Indicator for previous default	-0.032	0.002	<0.01	1.018
Propage1	Built before 1919	0.023	0.002	<0.01	1.653
Propage2 (base)	Built between 1919 and 1945	-	-	-	-
Propage3	Built after 1945	-0.010	0.001	<0.01	1.536
Propage4	Age unknown	-0.133	0.014	<0.01	1.017
security0	Flat or other	0.065	0.002	<0.01	1.370
security1	Detached	-0.020	0.002	<0.01	1.628
security2	Semi-detached	-0.013	0.002	<0.01	1.370
security3 (base)	Terraced	-	-	-	-

region0	Others or Missing	0.054	0.013	<0.01	1.054
region1	North	0.041	0.004	<0.01	1.758
region2	Yorkshire & Humberside	0.041	0.003	<0.01	3.011
region3	North West	0.047	0.003	<0.01	2.940
region4	East Midlands	0.052	0.004	<0.01	2.233
region5	West Midlands	0.037	0.004	<0.01	2.364
region6	East Anglia	0.047	0.004	<0.01	1.782
region7	Wales	0.047	0.004	<0.01	2.047
region8	South West	0.038	0.003	<0.01	2.936
region9	South East	0.050	0.003	<0.01	5.244
region10	Greater London	0.030	0.003	<0.01	4.265
region11	Northern Ireland	0.028	0.006	<0.01	1.333
region12 (base)	Scotland	-	-	-	-

Table A8: Parameter estimates for Probability of Repossession Model from robustness test (where only first instance of default was included in modelling)

Variable	Variable Explanation	Estimate	StdErr	WaldChiSq	ProbChiSq
Intercept	-	-2.638	0.036	5520.344	<0.01
dtv_UOS	Loan to value at default	2.757	0.031	7874.367	<0.01
ndef_previously	Indicator for previous default	0.000	-	-	-
security0 (base)	Flat or Others	-	-	-	-
security1	Detached	-0.446	0.032	191.324	<0.01
security2	Semi-detached	-0.547	0.025	469.607	<0.01
security3	Terraced	-0.354	0.022	250.396	<0.01

Table A9: Parameter estimates for Haircut Model from robustness test (where only first instance of default was included in modelling)

Variable	Variable Explanation	Estimate	StdErr	Probt	VIF
Intercept	-	0.524	0.010	<0.01	0.000
ltv	Loan to value at loan application	0.233	0.007	<0.01	1.139
TOB_UOS	Time on book (in years)	0.005	0.001	<0.01	1.081
VVAgroup1 (base)	Value of property / region average <= 0.9	-	-	-	-
VVAgroup2	0.9 < Value of property / region average <= 1.2	-0.003	0.004	0.429	1.138
VVAgroup3	1.2 < Value of property / region average <= 1.5	-0.057	0.006	<0.01	1.154
VVAgroup4	1.5 < Value of property / region average <= 1.8	-0.092	0.008	<0.01	1.131
VVAgroup5	1.8 < Value of property / region average <= 2.4	-0.087	0.009	<0.01	1.167
VVAgroup6	Value of property / region average > 2.4	-0.137	0.009	<0.01	1.230
ndef_previously	Indicator for previous default	0.000	-	-	-
propage (base)	Built after 1945	-	-	-	-
propage_vold	Very old property (before 1919)	-0.084	0.003	<0.01	1.271
propage_old	Old property (1919-1945)	-0.033	0.004	<0.01	1.194
security0 (base)	Flat or Others	-	-	-	-
security1	Detached	0.160	0.006	<0.01	1.891
security2	Semi-detached	0.127	0.004	<0.01	1.762
security3	Terraced	0.093	0.003	<0.01	1.731
region1	North	-0.109	0.010	<0.01	1.768
region2	Yorkshire & Humberside	-0.094	0.008	<0.01	2.926
region3	North West	-0.098	0.008	<0.01	3.201
region4	East Midlands	-0.102	0.008	<0.01	2.532

region5	West Midlands	-0.069	0.009	<0.01	2.491
region6	East Anglia	-0.073	0.009	<0.01	2.007
region7	Wales	-0.115	0.009	<0.01	2.155
region8	South West	-0.054	0.008	<0.01	3.374
region9	South East	-0.069	0.007	<0.01	6.589
region10	Greater London	-0.018	0.007	0.012	5.440
region11	Northern Ireland	-0.041	0.015	<0.01	1.267
region12 (base)	Scotland or Others / Missing	-	-	-	-

APPENDIX B. PLOTS OF MACROECONOMIC VARIABLES

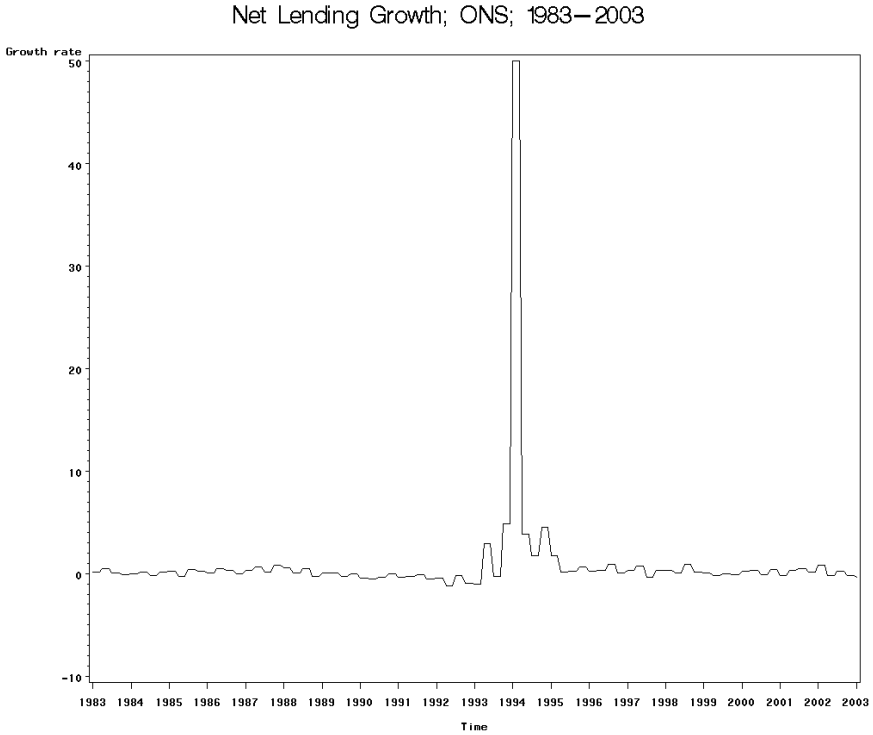

Figure B1: Net Lending Change; ONS; 1983 – 2003

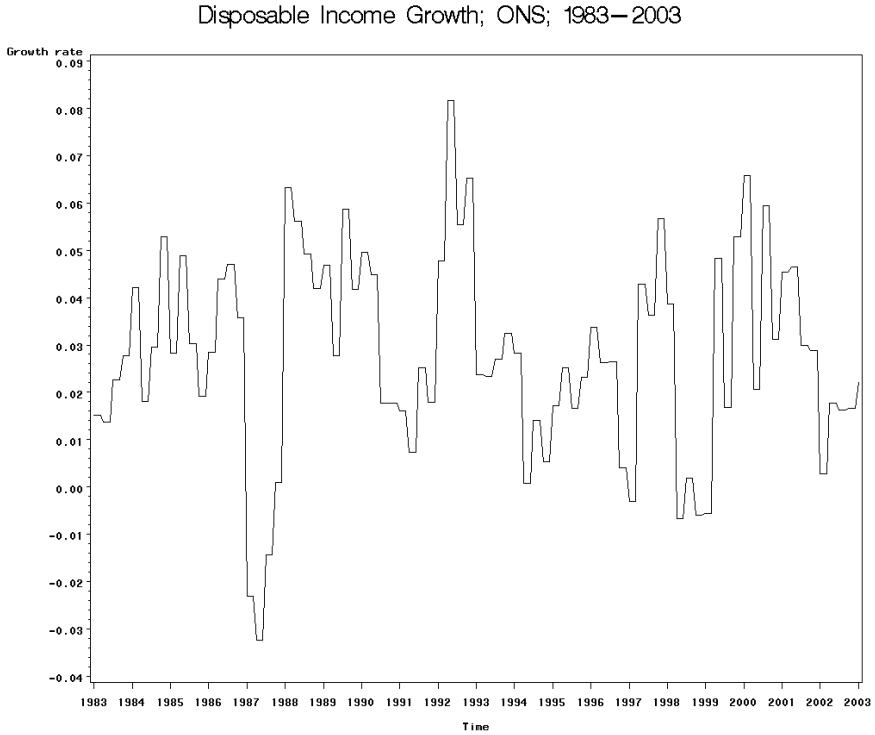

Figure B2: Disposable Income Growth; ONS; 1983 – 2003

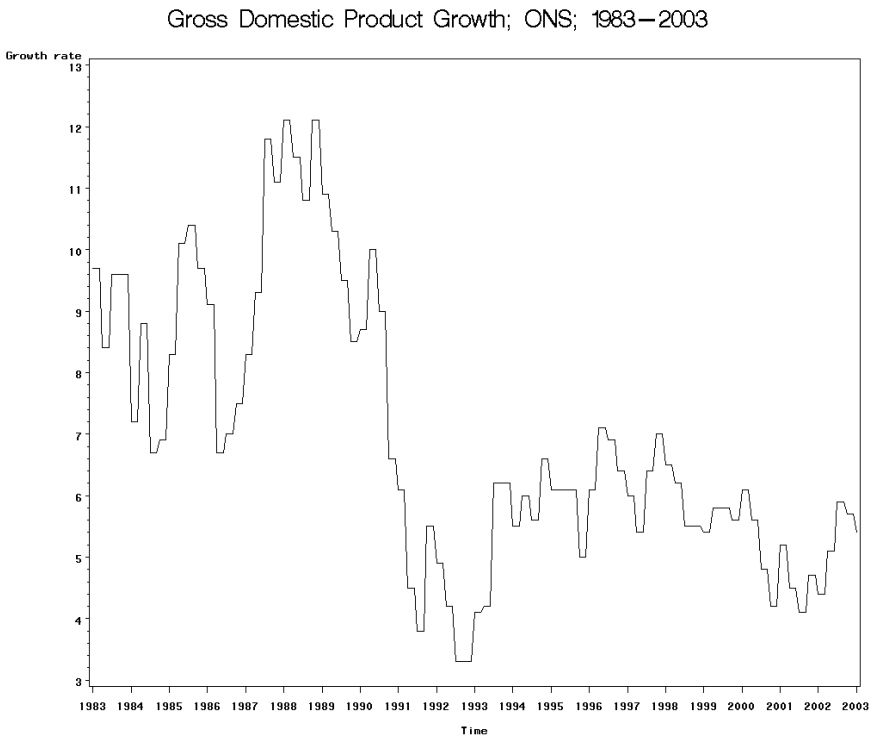

Figure B3: GDP Growth; ONS; 1983 – 2003

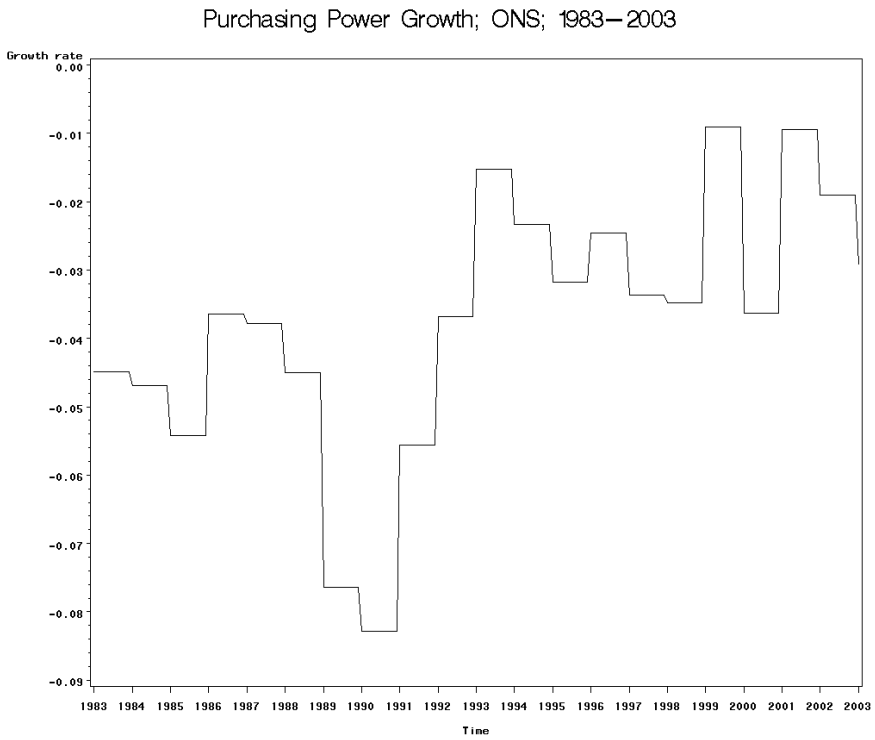

Figure B4: Purchasing Power of Pound Growth; ONS; 1983 – 2003

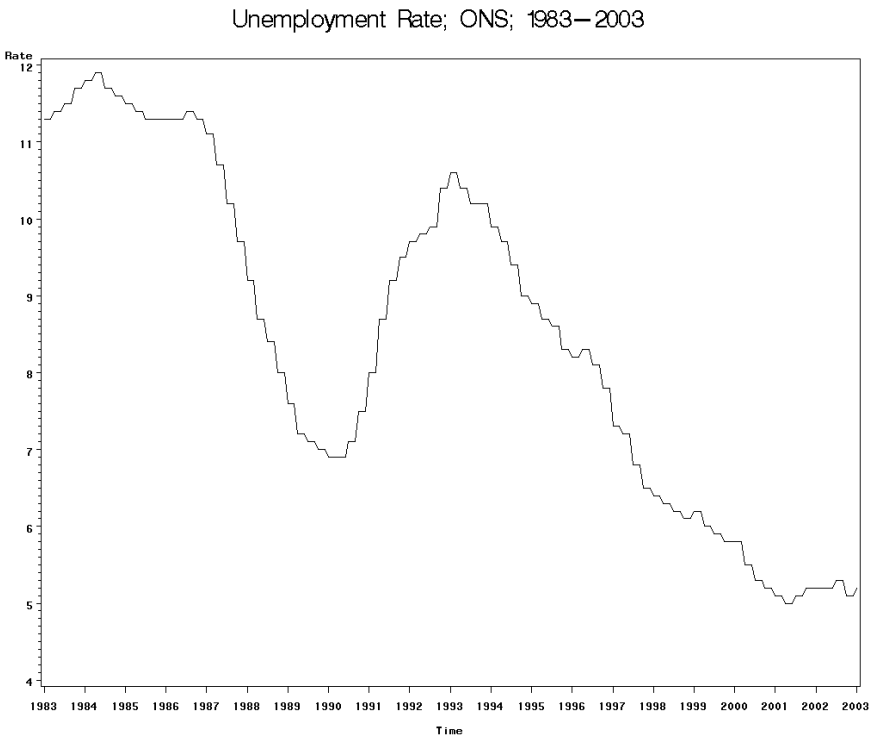

Figure B5: Unemployment rate; ONS; 1983 – 2003

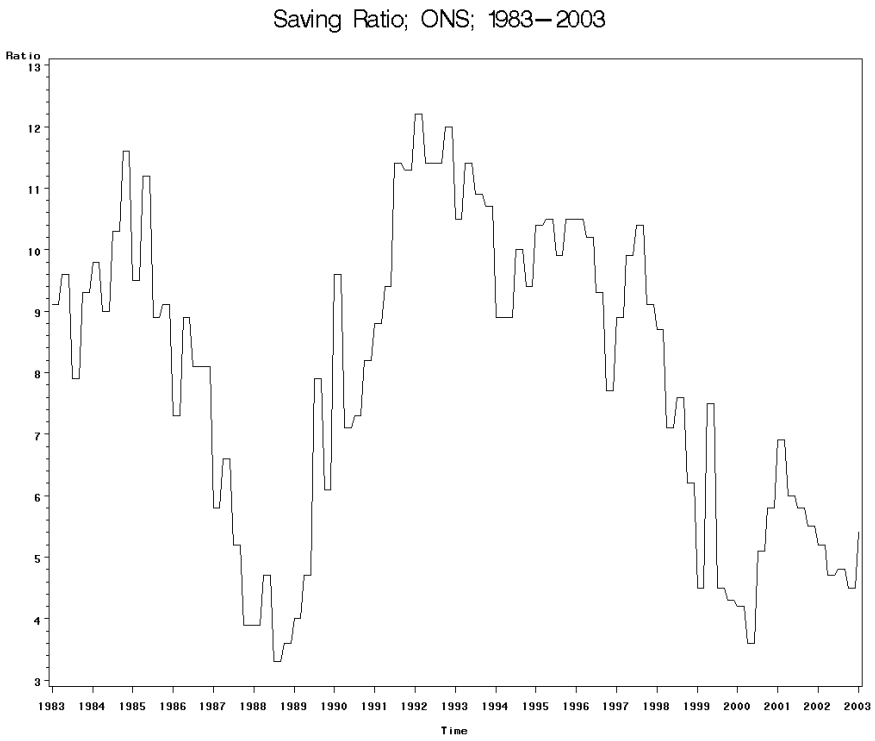

Figure B6: Saving ratio; ONS; 1983 – 2003

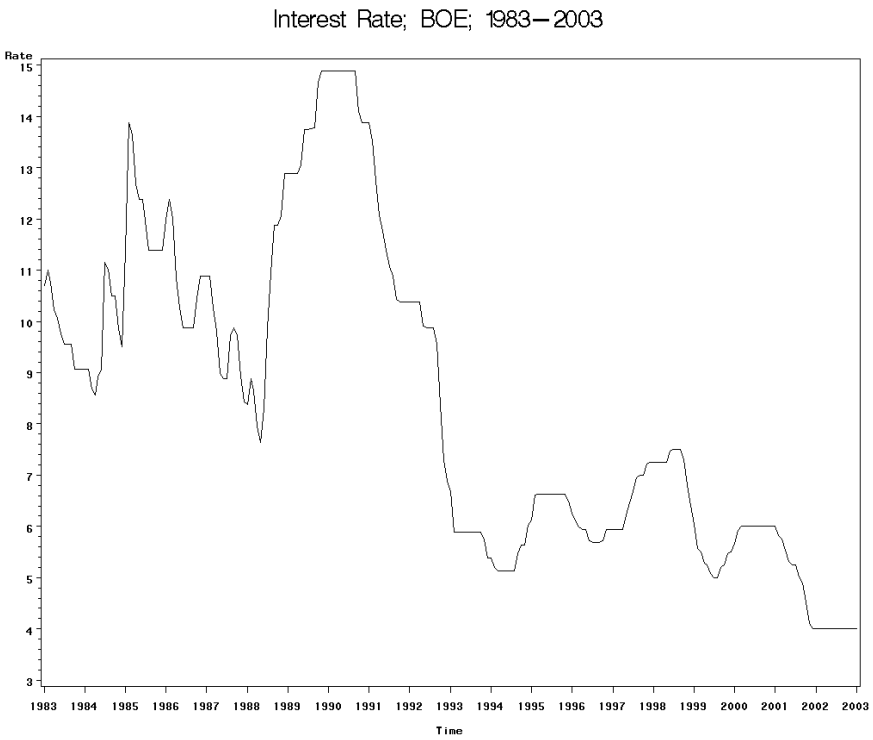

Figure B7: Interest rates; BOE; 1983 – 2003

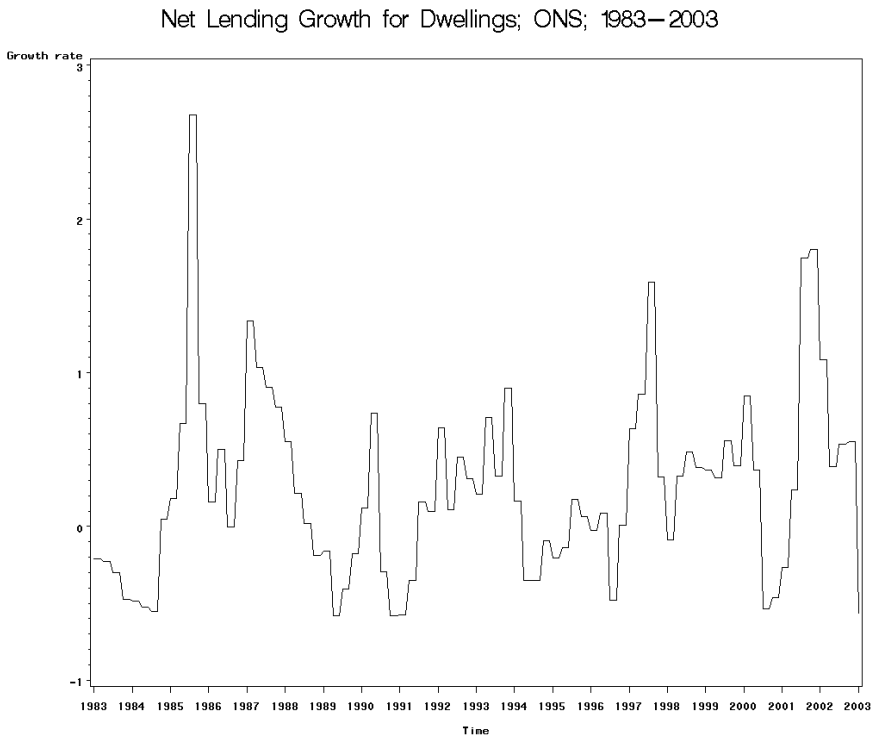

Figure B8: Net lending growth for dwellings; ONS; 1983 – 2003

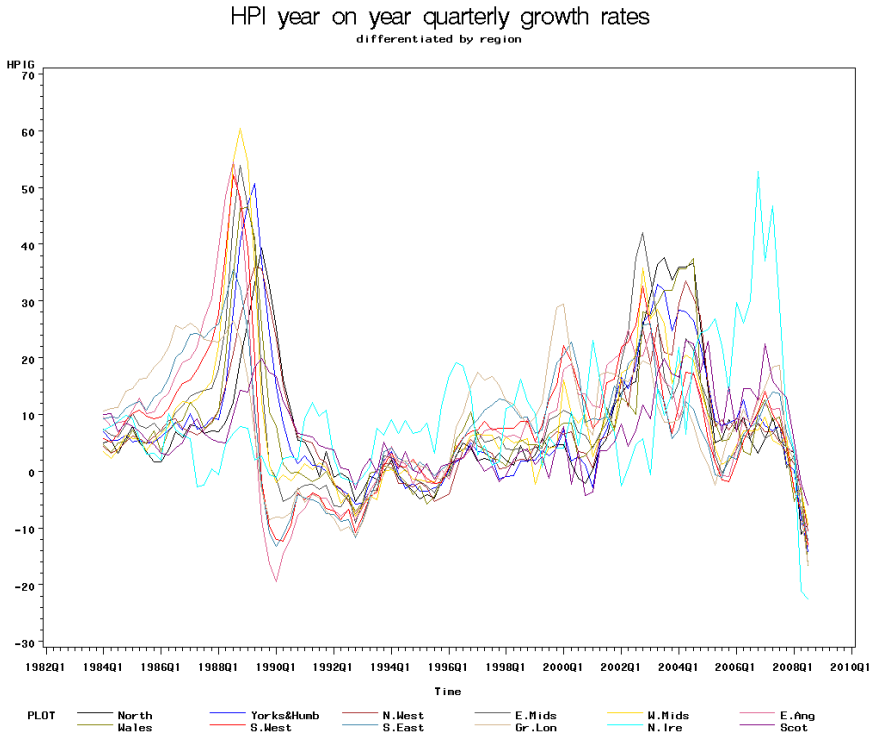

Figure B9: HPI year on year quarterly growth; Halifax; 1983 – 2003; differentiated by region

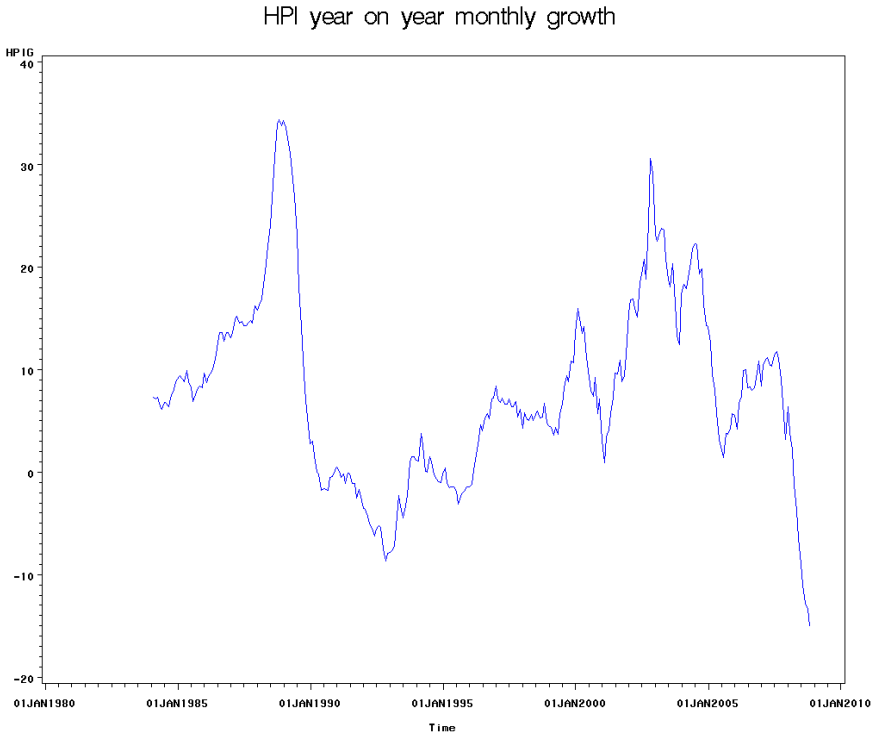

Figure B10: HPI year on year monthly growth; Halifax; 1983 – 200

APPENDIX C. PARAMETER ESTIMATES FROM MACROECONOMIC MODELS

Table C1: Parameter estimates and p-values of Probability of Repossession Macroeconomic Model

Variable	Variable Explanation	Estimate	StdErr	WaldChiSq	ProbChiSq
Intercept	-	-3.620	0.042	7267.330	<0.01
DLTV	Loan to value at default	2.644	0.030	7973.549	<0.01
Previous default	Indicator for previous default	-0.162	0.033	23.682	<0.01
security0 (base)	Flat or other	-	-	-	-
security1	Detached	-0.481	0.032	231.280	<0.01
security2	Semi-detached	-0.532	0.025	461.975	<0.01
security3	Terraced	-0.323	0.022	216.771	<0.01
INTR_DE	Interest rate at default	0.112	0.003	1992.278	<0.01

Table C2: Parameter estimates, p-values and VIFs of Haircut Macroeconomic Model

Variable	Variable Explanation	Estimate	StdErr	ProbT	VIF
Intercept	-	0.674	0.011	<0.01	0.000
LTV	Loan to value at loan application	0.223	0.007	<0.01	1.146
TOB	Time on book (in years)	-0.005	0.001	<0.01	1.711
VVAratio1 (base)	Value of property / region average <= 0.9	-	-	-	-
VVAratio2	0.9 < Value of property / region average <= 1.2	0.006	0.004	0.131	1.142
VVAratio3	1.2 < Value of property / region average <= 1.5	-0.044	0.006	<0.01	1.157

VVAratio4	1.5 < Value of property / region average <= 1.8	-0.071	0.008	<0.01	1.134
VVAratio5	1.8 < Value of property / region average <= 2.4	-0.072	0.009	<0.01	1.166
VVAratio6	Value of property / region average > 2.4	-0.116	0.009	<0.01	1.232
Previous default	Indicator for previous default	0.042	0.006	<0.01	1.168
Propage1	Very old property (before 1919)	-0.078	0.003	<0.01	1.280
Propage2	Old property (1919-1945)	-0.027	0.004	<0.01	1.196
Propage3 (base)	Built after 1945	-	-	-	-
security0 (base)	Flat or other	-	-	-	-
security1	Detached	0.162	0.006	<0.01	1.876
security2	Semi-detached	0.131	0.004	<0.01	1.765
security3	Terraced	0.097	0.003	<0.01	1.741
region1	North	-0.115	0.010	<0.01	1.753
region2	Yorkshire & Humberside	-0.092	0.008	<0.01	2.899
region3	North West	-0.102	0.008	<0.01	3.163
region4	East Midlands	-0.084	0.008	<0.01	2.498
region5	West Midlands	-0.055	0.008	<0.01	2.453
region6	East Anglia	-0.043	0.009	<0.01	1.981
region7	Wales	-0.104	0.009	<0.01	2.143
region8	South West	-0.021	0.007	<0.01	3.311
region9	South East	-0.033	0.007	<0.01	6.459
region10	Greater London	0.021	0.007	<0.01	5.313
region11	Northern Ireland	-0.007	0.014	0.631	1.261
region12 (base)	Scotland or others / missing	-	-	-	-
INTR_DE	Interest rate at default	-0.014	0.000	<0.01	1.663

Table C3: Parameter estimates, p-values and VIFs of personal loans macroeconomic LGD model

Variable	Variable Explanation	Estimate	StdErr	Probt	Variance Inflation
Intercept	-	0.684	0.022	<0.01	0.000
AGE_OF_EXP	Age of exposure (months)	-0.006	0.000	<0.01	8.041
LOAN_AMT_LAST	Amount of loan at opening	0.000	0.000	<0.01	1.508
APP_SCORE_FIRST	Application score of applicant at start of loan	-0.001	0.000	<0.01	1.406
no_mths_arrs_0_12m	Number of months with arrears >0 within the last 12 months	-0.014	0.001	<0.01	2.056
no_mths_arrs_0_ever	Number of months with arrears >0 within the life of the loan	0.006	0.000	<0.01	9.552
TERM_LAST	Term of loan (months)	0.003	0.000	<0.01	1.738
JOINT_APP	Joint applicant present	-0.060	0.005	<0.01	1.145
TADD	Time at address (months)	0.000	0.000	<0.01	1.123
TIME_AT_BANK	Time with Bank (months)	0.000	0.000	<0.01	1.169
RESID_STATUS_FIRST2	Residential status at time of application = Tenant	0.032	0.006	<0.01	1.281
EMPL_STATUS_C1_FIRST4	Employment Status: Self-employed	0.069	0.009	<0.01	1.293
HBS_MORT_HELD_FIRST	Mortgage held	-0.029	0.006	<0.01	1.419

EMPL_STATUS_C1_FIRST3	Employment Status: Private sector	0.021	0.005	<0.01	1.269
PL_HELD_FIRST	Personal loan account held	-0.018	0.005	<0.01	1.168
EMPL_STATUS_C1_FIRST8	Employment Status: Military	-0.157	0.038	<0.01	1.022
MAXIM_HELD_FIRST	Current account held	0.022	0.005	<0.01	1.192
PURP2	Purpose of Loan	-0.019	0.005	<0.01	1.148
EMPL_STATUS_C1_FIRST10	Employment Status: Unemployed without income	0.304	0.089	<0.01	1.005
EMPL_STATUS_C1_FIRST6	Employment Status: Student	0.420	0.134	<0.01	1.002
MARITAL_STATUS_FIRST4	Divorced / Separated	0.020	0.007	<0.01	1.033
SAV_HELD_FIRST	Savings account held	0.012	0.005	<0.01	1.135
NLG_DE	Net lending growth at default	-0.001	0.000	<0.01	1.013

APPENDIX D: RESULTS OF LGD MACROECONOMIC MODELS WITH TIME LAGS AND LEADS

Table D1: Performance of Probability of Repossession Model (test sets) with macroeconomic variables with six-month lag

PROBABILITY OF REPOSSESSION MODEL; Macroeconomic variables with six-month lag			
Model	Additional Variable	Model Estimate	ROC (Test)
Base (DLTV) + MV 1	Net Lending Growth with six-month lag	+	Insignificant
Base (DLTV) + MV 2	Disposable Income with six-month lag	-	0.743
Base (DLTV) + MV 3	GDP Growth with six-month lag	+	0.760
Base (DLTV) + MV 4	Purchasing Power Growth with six-month lag	-	0.757
Base (DLTV) + MV 5	Net Lending Growth for Dwellings with six-month lag	-	0.744
Base (DLTV) + MV 6	Unemployment Rate with six-month lag	-	0.756
Base (DLTV) + MV 7	Saving Ratio with six-month lag	-	0.754
Base (DLTV) + MV 8	Interest Rate with six-month lag	+	0.754
Base (DLTV) + MV 9	House Price Index Growth with six-month lag	+	0.745

Table D2: Performance of Probability of Repossession Model (test sets) with macroeconomic variables at six-month lead

PROBABILITY OF REPOSSESSION MODEL; Macroeconomic variables with six-month lead			
Model	Additional Variable	Model Estimate	ROC (Test)
Base (DLTV) + MV 10	Net Lending Growth with six-month lead	-	0.745
Base (DLTV) + MV 11	Disposable Income Growth with six-month lead	-	0.745
Base (DLTV) + MV 12	GDP Growth with six-month lead	+	0.746
Base (DLTV) + MV 13	Purchasing Power Growth with six-month lead	-	0.760
Base (DLTV) + MV 14	Net Lending Growth for Dwellings with six-month lead	-	0.749
Base (DLTV) + MV 15	Unemployment Rate with six-month lead	-	0.748
Base (DLTV) + MV 16	Saving Ratio with six-month lead	-	0.746
Base (DLTV) + MV 17	Interest Rate with six-month lead	+	0.761
Base (DLTV) + MV 18	House Price Index Growth with six-month lead	+	0.745

Table D3: Performance of Haircut Model (test sets) with macroeconomic variables with six-month lag

HAIRCUT MODEL; Macroeconomic variables with six-month lag			
Model	Additional Variable	Model Estimate	ROC (Test)
Base (DLTV) + MV 1	Net Lending Growth with six-month lag	+	Insignificant
Base (DLTV) + MV 2	Disposable Income with six-month lag	-	Insignificant
Base (DLTV) + MV 3	GDP Growth with six-month lag	+	Time on book becomes insignificant
Base (DLTV) + MV 4	Purchasing Power Growth with six-month lag	-	0.164
Base (DLTV) + MV 5	Net Lending Growth for Dwellings with six-month lag	-	0.144
Base (DLTV) + MV 6	Unemployment Rate with six-month lag	-	0.145
Base (DLTV) + MV 7	Saving Ratio with six-month lag	-	0.147
Base (DLTV) + MV 8	Interest Rate with six-month lag	+	0.167
Base (DLTV) + MV 9	House Price Index Growth with six-month lag	+	0.144

Table D4: Performance of Haircut Model (test sets) with macroeconomic variables at six-month lead

HAIRCUT MODEL; Macroeconomic variables with six-month lead			
Model	Additional Variable	Model Estimate	ROC (Test)
Base (DLTV) + MV 10	Net Lending Growth with six-month lead	-	0.745
Base (DLTV) + MV 11	Disposable Income Growth with six-month lead	-	Insignificant
Base (DLTV) + MV 12	GDP Growth with six-month lead	+	0.746
Base (DLTV) + MV 13	Purchasing Power Growth with six-month lead	-	0.760
Base (DLTV) + MV 14	Net Lending Growth for Dwellings with six-month lead	-	0.749
Base (DLTV) + MV 15	Unemployment Rate with six-month lead	-	Insignificant
Base (DLTV) + MV 16	Saving Ratio with six-month lead	-	0.746
Base (DLTV) + MV 17	Interest Rate with six-month lead	+	0.761
Base (DLTV) + MV 18	House Price Index Growth with six-month lead	+	Time on book becomes insignificant

Table D5: Performance of personal loan LGD Model (test sets) with macroeconomic variables with six-month lag

PERSONAL LOANS LGD MODEL; Macroeconomic variables with six-month lag				
Model	Additional Variable	Model Estimate	p-value	R-square (Test)
Base + MV 1	Net Lending Growth with six-month lag	-	Insignificant	0.073
Base + MV 2	Disposable Income Growth with six-month lag	-	Insignificant	0.073
Base + MV 3	GDP Growth with six-month lag	-	Insignificant	0.073
Base + MV 4	Purchasing Power Growth with six-month lag	-	Insignificant	0.073
Base + MV 5	Unemployment Rate with six-month lag	-	Insignificant	0.073
Base + MV 6	Saving Ratio with six-month lag	-	Insignificant	0.073
Base + MV 7	Interest Rate with six-month lag	+	Insignificant	0.073
Base + MV 8	Net Lending Growth for Dwellings with six-month lag fault	+	Insignificant	0.073
Base + MV 9	House Price Index Growth with six-month lag	+	Insignificant	0.073

Table D6: Performance of personal loan LGD Model (test sets) with macroeconomic variables with six-month lead

PERSONAL LOANS LGD MODEL; Macroeconomic variables with six-month lead				
Model	Additional Variable	Model Estimate	p-value	R-square (Test)
Base + MV 1	Net Lending Growth with six-month lead	+	Insignificant	0.073
Base + MV 2	Disposable Income Growth with six-month lead	+	Insignificant	0.073
Base + MV 3	GDP Growth with six-month lead	-	Insignificant	0.073
Base + MV 4	Purchasing Power Growth with six-month lead	+	Insignificant	0.073
Base + MV 5	Unemployment Rate with six-month lead	-	Insignificant	0.073
Base + MV 6	Saving Ratio with six-month lead	-	Insignificant	0.073
Base + MV 7	Interest Rate with six-month lead	-	Insignificant	0.073
Base + MV 8	Net Lending Growth for Dwellings with six-month lead	+	Insignificant	0.073
Base + MV 9	House Price Index Growth with six-month lead	+	Insignificant	0.073

APPENDIX E. TABLE OF CONTRIBUTIONS TO MARGINAL RISK FOR REPOSSESSION AND CLOSURE

Table E1: Contributions to calculation of marginal risk for repossession and/or closure. In a general form for the equation of a line, *c* represents the coefficient and *m* represents the slope. The equations for marginal risk are different for each type of security and each DLTV band, with marginal risk as the independent variable and HPIG as the explanatory variable. The table below shows how each variable contributes to the calculation of marginal risk according to the type of security and DLTV type.

Variable	Explanation	Groupdtv1				Groupdtv2				Groupdtv3				Groupdtv4				Groupdtv5				Groupdtv6				Groupdtv7			
		F	T	S	D	F	T	S	D	F	T	S	D	F	T	S	D	F	T	S	D	F	T	S	D	F	T	S	D
groupdtv1	DLTV <= 0.5	c	c	c	c																								
groupdtv2	0.5 < DLTV <= 0.7					c	c	c	c																				
groupdtv3	0.7 < DLTV <= 0.9									c	c	c	c																
groupdtv4	0.9 < DLTV <= 1.0													c	c	c	c												
groupdtv5	1.0 < DLTV <= 1.1																	c	c	c	c								
groupdtv6	1.1 < DLTV <= 1.2																					c	c	c	c				
groupdtv7	DLTV > 1.2																									c	c	c	c
security0	Flat	c				c				c				c				c				c				c			
security1	Detached				c				c				c				c				c				c				c
security2	Semi-Detached			c				c				c				c				c				c				c	

security3	Terraced		c				c				c				c				c				c				c		
HPIG0	HPIG (time-dependent)	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m
HPIGsec0	HPIG x Flat	m				m				m				m				m				m				m			
HPIGsec1	HPIG h x Detached				m				m				m				m				m				m				m
HPIGsec2	HPIG x Semi-Detached			m				m				m				m				m				m				m	
HPIGsec3	HPIG x Terraced		m				m				m				m				m				m				m		
HPIGdtv1	HPIG x DLTV (0-0.5)	m	m	m	m																								
HPIGdtv2	HPIG x DLTV (0.5-0.7)					m	m	m	m																				
HPIGdtv3	HPIG x DLTV (0.7-0.9)									m	m	m	m																
HPIGdtv4	HPIG x DLTV (0.9-1.0)													m	m	m	m												
HPIGdtv5	HPIG x DLTV (1.0-1.1)																	m	m	m	m								
HPIGdtv6	HPIG x DLTV (1.1-1.2)																					m	m	m	m				
HPIGdtv7	HPIG x DLTV (1.2 and above)																									m	m	m	m

APPENDIX F. PARAMETER ESTIMATES FOR SURVIVAL MODELS

Table F1: Parameter estimates for repossession survival model

Variable	Explanation	Estimate	ProbChiSq	HazardRatio
groupdtv1	DLTV <= 0.5	-1.876	<0.01	0.153
groupdtv2	0.5 < DLTV <= 0.7	-1.194	<0.01	0.303
groupdtv3	0.7 < DLTV <= 0.9	-0.630	<0.01	0.532
groupdtv4	0.9 < DLTV <= 1.0	-0.154	<0.01	0.858
groupdtv5	1.0 < DLTV <= 1.2	0.082	<0.01	1.086
groupdtv6	1.1 < DLTV <= 1.2	0.121	<0.01	1.129
groupdtv7	DLTV > 1.2	0.000	-	-
security0	Flat	0.321	<0.01	1.379
security1	Detached	-0.097	<0.01	0.907
security2	Semi-Detached	-0.169	<0.01	0.844
security3	Terraced	0.000	-	-
HPIG	HPIG (time-dependent)	0.022	<0.01	1.023
HPIGsec0	HPIG x Flat	-0.002	0.333	0.998
HPIGsec1	HPIG h x Detached	-0.021	<0.01	0.979
HPIGsec2	HPIG x Semi-Detached	-0.007	<0.01	0.993
HPIGsec3	HPIG x Terraced	0.000	-	-
HPIGdtv1	HPIG x DLTV (0-0.5)	-0.026	<0.01	0.974
HPIGdtv2	HPIG x DLTV (0.5-0.7)	-0.030	<0.01	0.971
HPIGdtv3	HPIG x DLTV (0.7-0.9)	-0.017	<0.01	0.983
HPIGdtv4	HPIG x DLTV (0.9-1.0)	-0.015	<0.01	0.985
HPIGdtv5	HPIG x DLTV (1.0-1.1)	-0.022	<0.01	0.979
HPIGdtv6	HPIG x DLTV (1.1-1.2)	-0.021	<0.01	0.979
HPIGdtv7	HPIG x DLTV (1.2 and above)	0.000	-	-

Table F2: Parameter estimates for closure survival model

Variable	Explanation	Estimate	ProbChiSq	HazardRatio
groupdtv1	DLTV <= 0.5	1.043	<0.01	2.837
groupdtv2	0.5 < DLTV <= 0.7	1.034	<0.01	2.813
groupdtv3	0.7 < DLTV <= 0.9	0.924	<0.01	2.519
groupdtv4	0.9 < DLTV <= 1.0	0.662	<0.01	1.939
groupdtv5	1.0 < DLTV <= 1.2	0.414	<0.01	1.513
groupdtv6	1.1 < DLTV <= 1.2	0.209	<0.01	1.233
groupdtv7	DLTV > 1.2	0.000	-	-
security0	Flat	0.219	<0.01	1.245
security1	Detached	0.248	<0.01	1.282
security2	Semi-Detached	0.039	<0.01	1.040
security3	Terraced	0.000	-	-
HPIG	HPIG (time-dependent)	0.102	<0.01	1.108
HPIGdtv1	HPIG x DLTV (0-0.5)	-0.086	<0.01	0.918
HPIGdtv2	HPIG x DLTV (0.5-0.7)	-0.078	<0.01	0.925
HPIGdtv3	HPIG x DLTV (0.7-0.9)	-0.063	<0.01	0.939
HPIGdtv4	HPIG x DLTV (0.9-1.0)	-0.049	<0.01	0.952
HPIGdtv5	HPIG x DLTV (1.0-1.1)	-0.032	<0.01	0.968
HPIGdtv6	HPIG x DLTV (1.1-1.2)	-0.014	<0.01	0.986
HPIGdtv7	HPIG x DLTV (1.2 and above)	0.000	-	-

APPENDIX G. GRAPHS OF HOUSE PRICE GROWTH FOR REGIONS IN THE UK

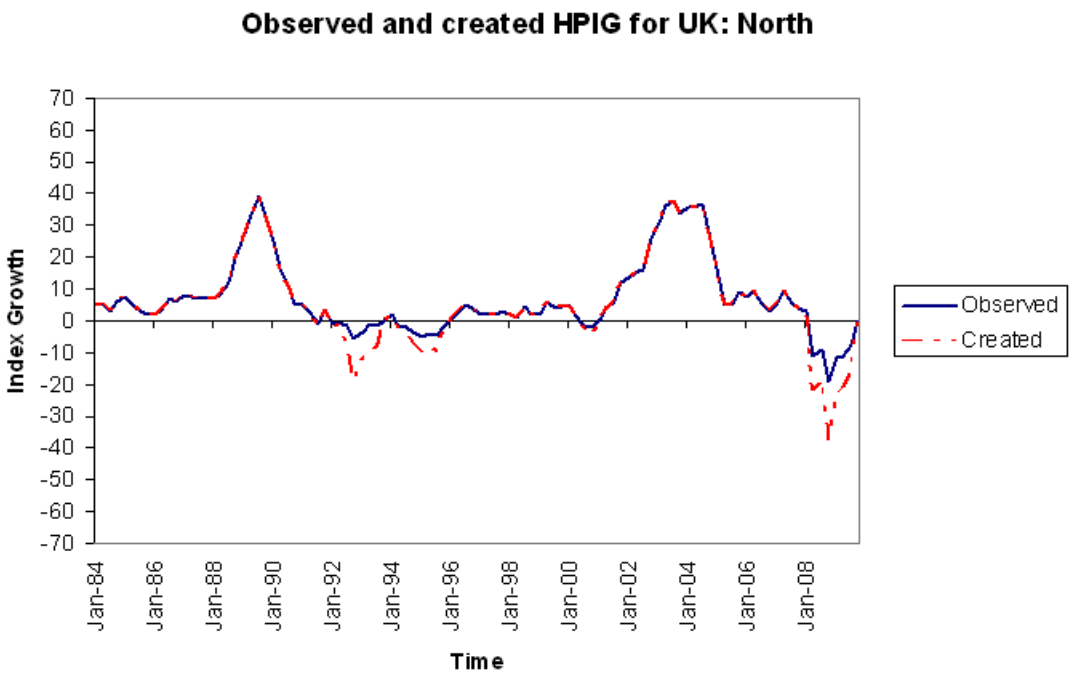

Figure G1: Observed and stressed HPIG for UK (North)

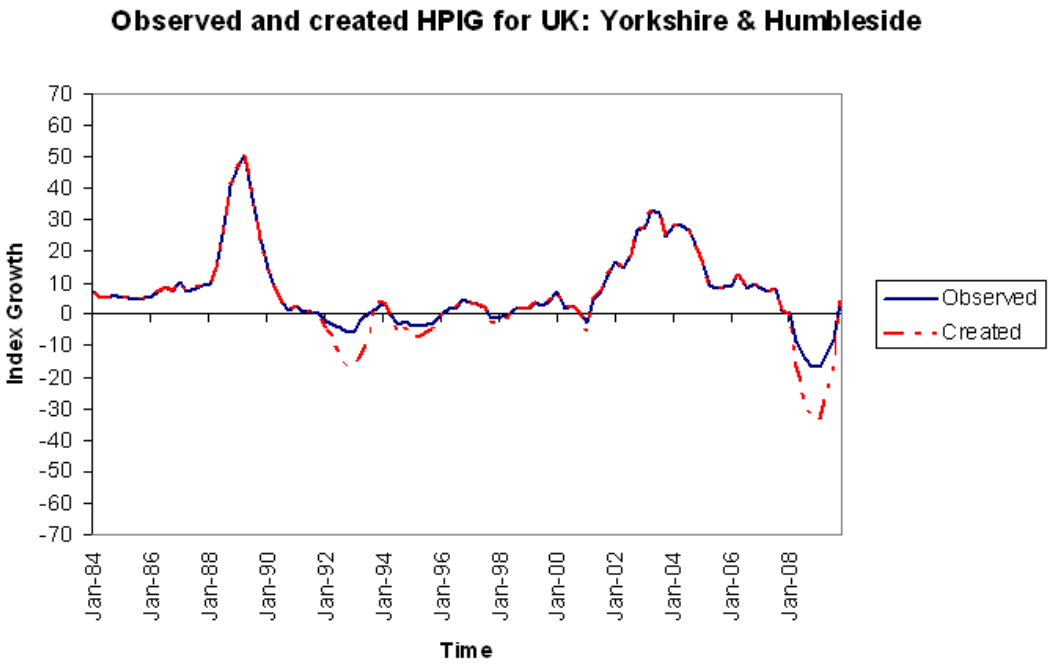

Figure G2: Observed and stressed HPIG for UK (Yorkshire and Humbleside)

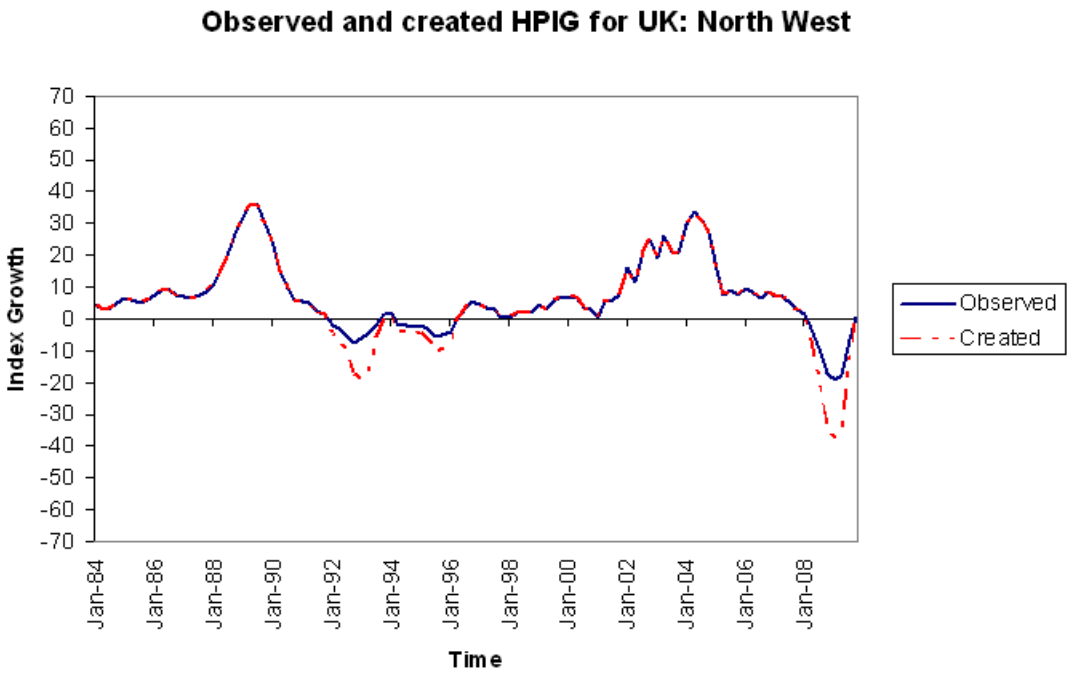

Figure G3: Observed and stressed HPIG for UK (North West)

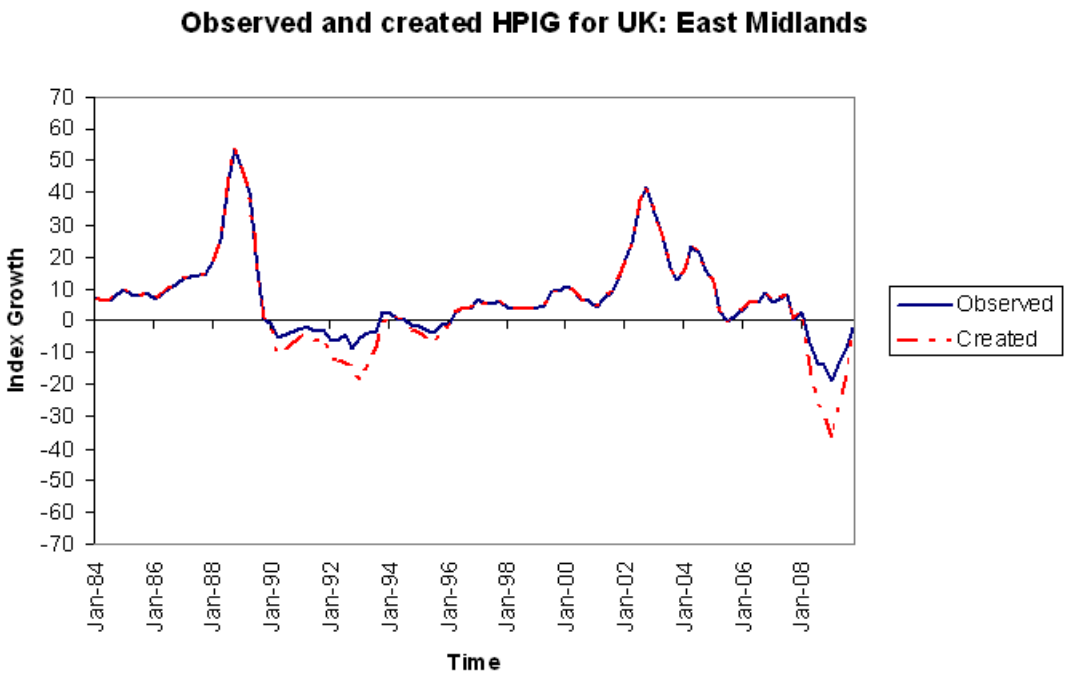

Figure G4: Observed and stressed HPIG for UK (East Midlands)

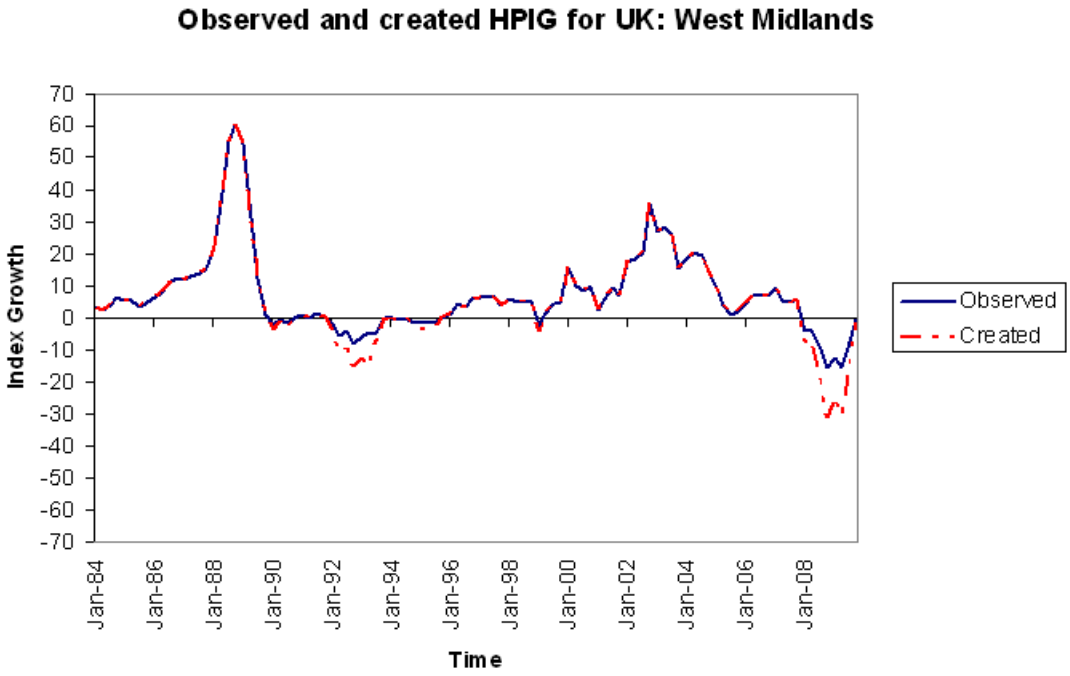

Figure G5: Observed and stressed HPIG for UK (West Midlands)

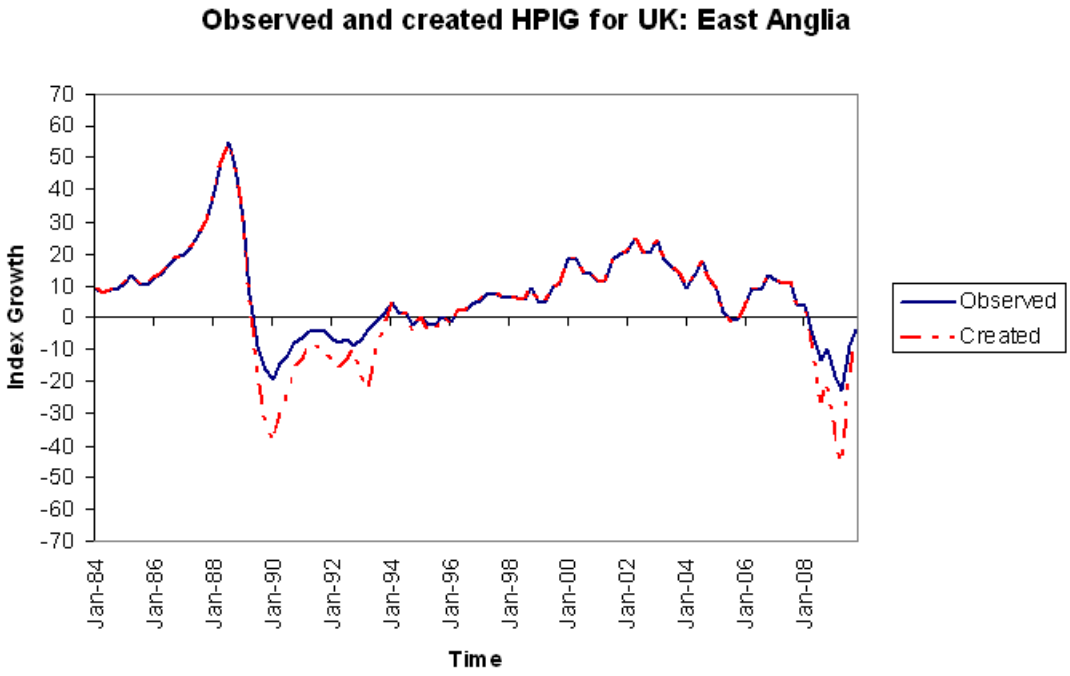

Figure G6: Observed and stressed HPIG for UK (East Anglia)

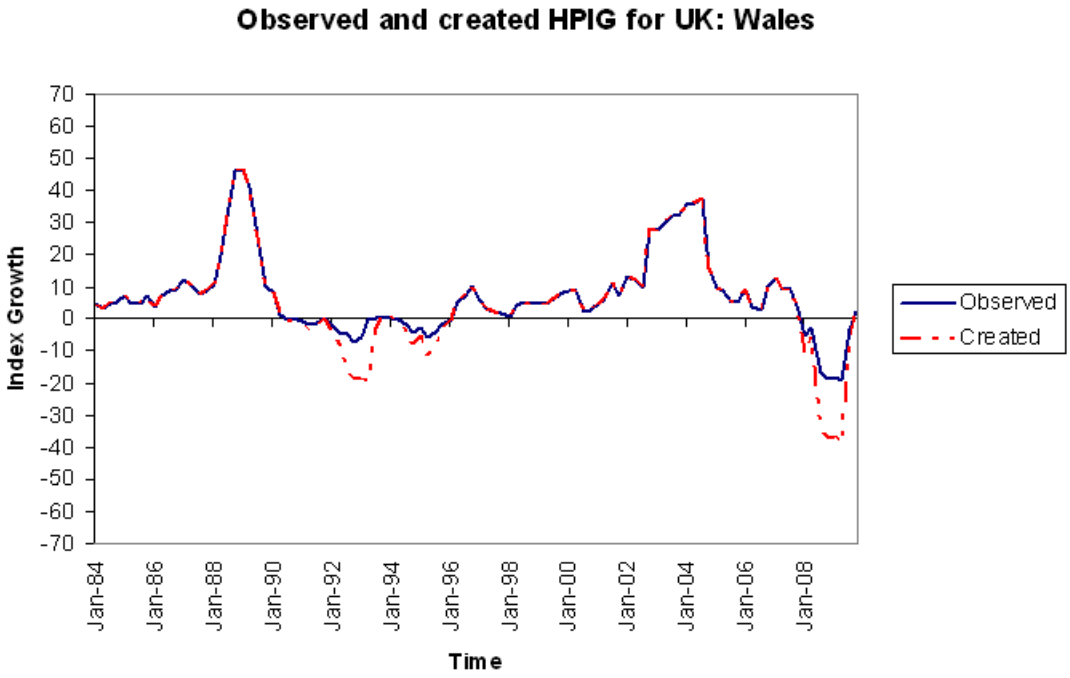

Figure G7: Observed and stressed HPIG for UK (Wales)

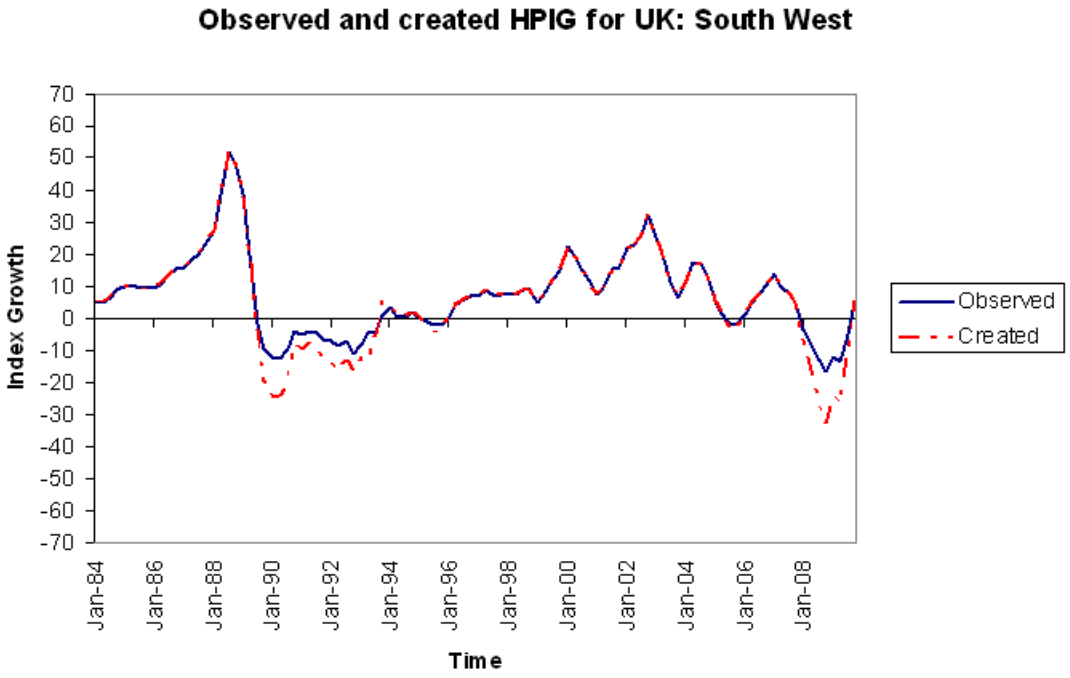

Figure G8: Observed and stressed HPIG for UK (South West)

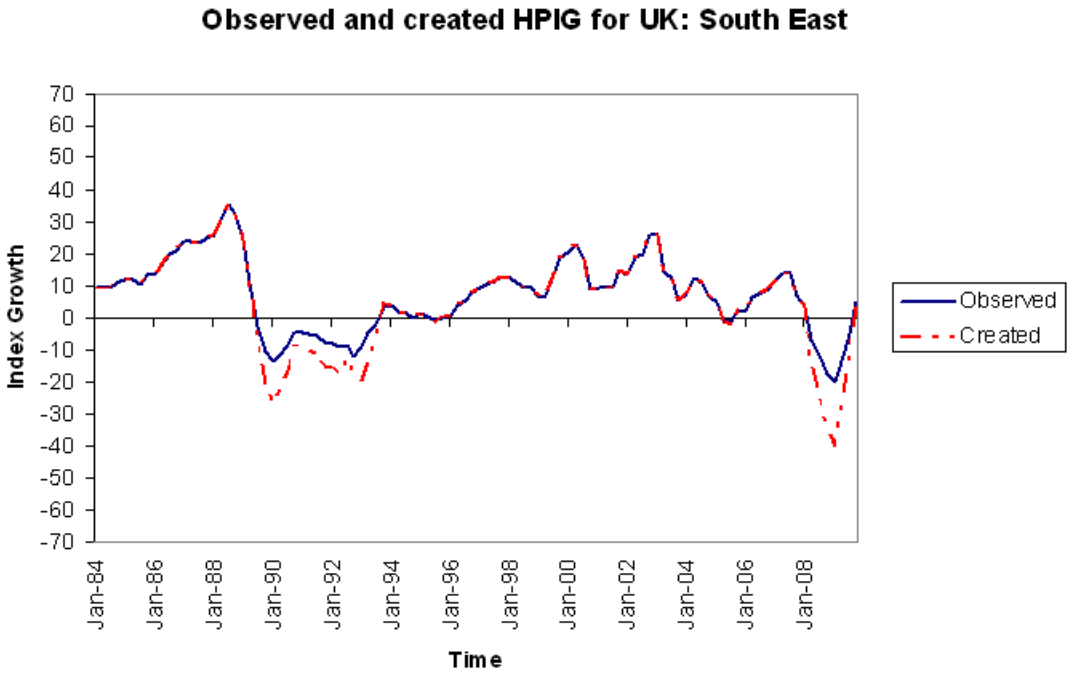

Figure G9: Observed and stressed HPIG for UK (South East)

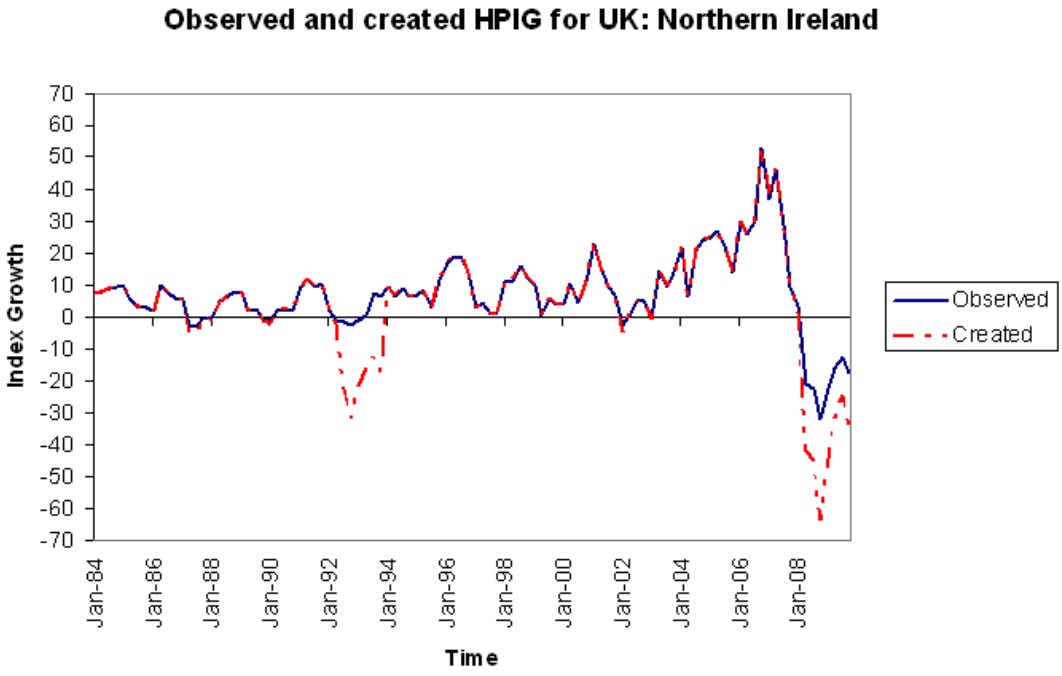

Figure G10: Observed and stressed HPIG for UK (Northern Ireland)

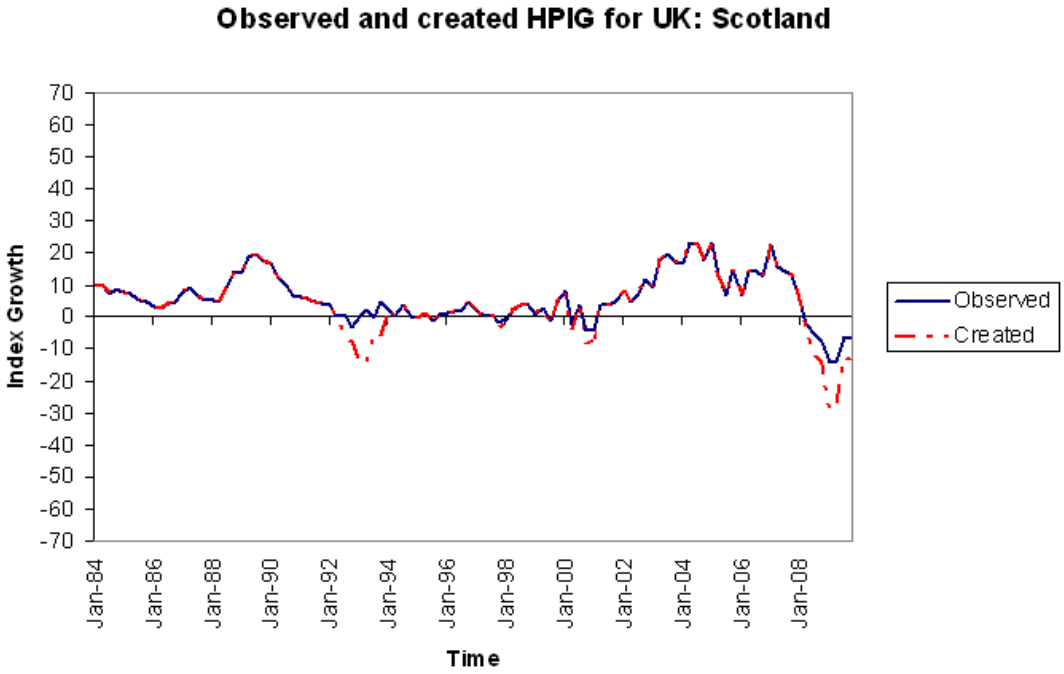

Figure G11: Observed and stressed HPIG for UK (Scotland)