LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MASTER

ACTUARIAL SCIENCE

MASTER´S FINAL WORK

PROJECT

BACKTESTING OF A CREDIT SCORING SYSTEM UNDER THE CURRENT REGULATORY FRAMEWORK

MARIA ANA E CASTELLO-BRANCO DOS SANTOS

JUNE-2017

LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MASTER

ACTUARIAL SCIENCE

MASTER´S FINAL WORK

PROJECT

BACKTESTING OF A CREDIT SCORING SYSTEM UNDER THE CURRENT REGULATORY FRAMEWORK

MARIA ANA E CASTELLO-BRANCO DOS SANTOS

SUPERVISION:

MARTA SOFIA SILVESTRE SANTOS
ONOFRE ALVES SIMÕES

JUNE - 2017

If a bank is too big to fail, it is too big to exist.
Bernie Sanders (2016)

Acknowledgements

Em primeiro lugar, gostaria de agradecer ao Professor Onofre Simões pela imediata disponibilidade para me orientar neste trabalho.
Não posso deixar de fazer um agradecimento à Marta Santos pela sua disponibilidade desde o meu primeiro dia de estágio. Um especial obrigada por se disponibilizar a partilhar comigo um pouco do seu imenso conhecimento e experiência e por ser um exemplo indiscutível de dedicação e profissionalismo. O seu contributo para este trabalho é imensurável, bem como o foi a oportunidade de trabalhar ao seu lado.
Um especial obrigada a todos os colegas de trabalho por terem contribuído, cada um à sua maneira, para o meu crescimento pessoal e profissional. Um agradecimento especial ao João Braga pela partilha da sua experiência; e à Carla Fernandes por todo o carinho e incentivo.
Pelo apoio determinante para a conclusão desta etapa e por todos os conselhos que me fizeram manter focada no objetivo final deste projecto, um muito obrigada à Dr.ª Madalena Gomes.
Aos meus amigos que sempre me apoiaram mesmo quando não tive total disponibilidade para estar presente. Um especial obrigada à Rute por todo o apoio, dentro e fora do contexto académico, desde que entrámos na faculdade e que não exclui o seu precioso contributo neste trabalho.
Ao Rui, pela paciência e apoio incondicional que foram indispensáveis para a conclusão deste projeto. Obrigada pelo incentivo e bom humor nos momentos mais críticos e por me relembrar sempre de que somos uma equipa.
Por último, e decerto não menos importante, quero deixar um especial agradecimento aos meus Pais e irmão por serem os meus maiores exemplos de persistência e resiliência, quer a nível pessoal, académico e profissional. À minha sobrinha por ser uma fonte de inspiração para ser e fazer melhor.

Abstract

The contemporary financial crises, especially the one arisen in 2007, have proved the frailty of the overall financial system, even for some institutions that seemed perfectly solvent and "too-big-to-fail". Since the implementation of the current regulatory framework within the global financial system, banks are allowed to rely in a system using their own estimates for credit risk parameters as inputs for the calculation of risk weights and capital requirements. Consequently, in order to assure the stability and soundness of credit institutions, the need for a robust validation system to ensure accuracy and consistency of internal rating systems is greater than ever before.

Although several studies on validation processes already exist, a deeper understanding and agreement on this subject is required, namely in what concerns the accuracy assessment of internal estimates for credit risk parameters, in order to achieve capital requirements stability. Calibration of default probabilities represents one of the quantitative validation procedures underlying the exercise of backtesting that must be performed on a regular basis.

The present text discusses the probability of default (PD) calibration process using a scoring model to illustrate the assessment of the predictive power of these internal estimates in a residential mortgage portfolio. To overcome the challenge of developing an adequate validation scheme in compliance with the current regulatory framework, this project project keeps in mind the legislation from Basel Committee on Banking Supervision (BCBS) and European Banking Authority (EBA), some relevant studies developed on this subject and those that are consider to be the best practices of credit risk management. In particular, following the most recent EBA's guidelines to ensure the stability of internal ratings systems, the purposed PD backtesting and scoring calibration procedures are based on the long-run average of one-year default rates.

Keywords: Credit Risk, Capital Requirements, Risk Weights, IRB Approach, Backtesting, Calibration, Credit Scoring Models, LRDR.

Resumo

As crises financeiras contemporâneas, nomeadamente a de 2007, vieram provar a fragilidade do sistema financeiro, mesmo para algumas instituições que pareciam perfeitamente solventes e *"too-big-to-fail"*. Desde a implementação do atual acordo de supervisão financeira internacional, os bancos podem usar as suas estimativas internas de avaliação de risco de crédito como base para o cálculo dos ponderadores de risco e requisitos de capital. Consequentemente, com vista a assegurar a estabilidade e solvabilidade das instituições de crédito, torna-se crescente a necessidade de um sistema de validação robusto, para garantir a consistência e precisão dos sistemas de notação interna.

Existem vários estudos sobre o processo de validação de estimativas internas. No entanto, aprofundamento e acordo nesta matéria são ainda insuficientes, nomeadamente no que diz respeito à avaliação da precisão das estimativas internas para os parâmetros de risco de crédito, com o objectivo de atingir a estabilidade dos requisitos de capital. A calibração das probabilidades de incumprimento representa um dos procedimentos de validação quantitativa inerentes ao exercício de *backtesting*.

Neste trabalho, será explorado o processo de calibração das probabilidades de incumprimento recorrendo a um modelo de *scoring* para exemplificar como é feita a avaliação da capacidade preditiva destas estimativas internas numa carteira de Crédito à Habitação. Para superar o desafio de desenvolver um sistema de validação adequado, o presente projeto tem em consideração o atual e amplo quadro regulatório proveniente do Comité de Basileia para a Supervisão Bancária (BCBS) e da Autoridade Bancária Europeia (EBA), alguns artigos relevantes nesta matéria e aquelas que são consideradas as melhores práticas de gestão do risco de crédito. Em particular, seguindo as mais recentes orientações da EBA para assegurar a estabilidade dos sistemas de notação interna, o backtesting e a calibração ao modelo de probabilidades de incumprimento propostos são baseados na média de longo-prazo de taxas de incumprimento a um ano.

Palavras-chave: Risco de Crédito, Requisitos de Capital, Ponderadores de Risco, Sistemas de Notação Interna, *Backtesting*, Calibração, Modelos de Scoring, LRDR.

Contents

List of Figures

List of Tables

Abbreviations and Acronyms

A-IRB Advanced Internal-Ratings Based

BCBS Basel Committee on Baking Supervision

BIS Bank for International Settlements

CLT Central Limit Theorem

CRD Capital Requirements Directive

CRR Capital Requeriments Regulation

DR Default Rates

EAD Exposure At Default

EBA European Banking Authority

ECB European Central Bank

ESRB European Systemic Risk Board

EU European Union

F-IRB Foundation Internal-Ratings Based

FSB Financial Stability Board

IMF International Monetary Fund

IRRS Internal Risk Rating Systems

LGD Loss Given Default

LRDR Long Run Default Rate

LRPD Long Run Probability of Default

MCR Minimum Capital Requirements

PD Probability of Default

PIT Point-in-Time

RWA Risk-Weighted Assets

SREP Supervisory Review and Evaluation Process

TTC Through-the-Cycle

UL Unexpected Loss

Chapter 1

Introduction

1.1 Risk and The Basel Accords

Both banks and insurance companies deal and manage risk on a daily basis as a matter of their business field. As a consequence of this, their activity must depend on accurate internal risk models. In order to prevent events with greater impact in the financial environment, supervisors and remaining regulatory competent authorities make sure that insurance companies properly allocate their capital and that banks do the same, as well as they *"establish more stable, long-term sources of funding". – cf.* Deutsche Bank Research (2011). The Bank for International Settlements (BIS) establishes guidelines and strategies for banks, insurance companies and pension funds to follow in their businesses. These guidelines allow the different institutions to rely on their own internal models and assumptions, as an alternative to Standard Approaches, to determine their capital requirements. Although referring to different business models and different regulatory measures, the financial supervision for both banks and insurance companies is made through a risk-based approach and implies thorough reporting to supervisory authorities.

As a matter of fact, the concern to regulate the financial system as a whole and increase a set of best practices in financial and investment institutions from banks (Basel Accords) to insurance companies (Solvency Projects) is not new. Even before the sub-prime crisis and subsequent disruption of the global financial system these institutions were already subject to tight supervision and obliged to apply the regulatory standards in force within their governance. Notwithstanding, some experts point out the previous regulatory framework as a contributor to the recent financial crisis – see § 2.3 of European Actuarial Consultative Group (2013). There are also relevant differences that must be taken into account as referred in European Actuarial Consultative Group (2013). First, banks have to deal mainly with capital and liquidity while insurance companies consider capital and risk. Also, banks and insurance companies are not evolving at the same pace – *cf.* p. 3.

Banks are a fundamental driver of the global economy. Their actions and business models represent risks not just for each institution individually, but also for the entire financial system – the so-called systemic risk. Systemic risk was defined all together by BIS, the International Monetary Fund (IMF) and the Financial Stability Board (FSB), the most relevant international regulators, as *"the disruption to flow of financial services that (i) is caused by an impairment of all or parts of the financial system; and (ii) has the potential to inflict serious negative*

consequences for the real economy" – cf. IMF, BIS and FSB (2009). Also, the European Central Bank (ECB) refers to systemic risk as something that can potentially threat *"the stability of or confidence in the financial system" – cf. Glossary of ECB*.

In other words, systemic risk consists in the possibility that an event in an institution could cause an adverse scenario, instability or even collapse of an entire system or economy due to spillover effects – see, for example, Bandt and Hartmann (2000). Indeed, such side effects were found the main contributors for the most recent financial crisis, triggered in 2007. The repercussions of this crisis spread all over the globe, specially in Europe that faces now a sovereign debt crisis.

The Basel Committee on Banking Supervision (BCBS), founded in 1974 and hosted by BIS, developed and introduced in the banking system an international regulatory framework, the Basel Accords, with the intention of strengthening the resilience and improving the supervision of the international banking system. Basel Accords represent, since 1988, a continuous process in pursuing a global sound financial system through the establishment of a consistent and convergent regulatory framework among different countries and institutions, as well as the definition of a set of common guidelines of assessment and adequacy of capital. The primary basis of this regulatory framework was the set of minimum capital requirements (MCR) for banks, but its extent goes far beyond that.

Basel I was the first of the three Accords that BCBS developed in cooperation with other international entities responsible for banking supervision. Basel I was mainly focused in dealing efficiently with credit risk, known as the major risk faced by a credit institution, and minimizing it through a set of standard international guidelines to be applied in the supervised international banks. These banks were then obliged to allocate an amount of MCR corresponding to a minimum of 8% of the total capital.

After the financial crises have questioned some of the risk management practices carried out by credit institutions, as well as some of the regulator's guidelines supported in Basel I a key concern of the international supervisory community was to highlight the importance of implementing a set of best practices to promote the stability of the overall financial system in order to prevent and mitigate its systemic risk, and not just consider institutions individually. This laid the groundwork to a new accord aiming to *"encompass all banking risk within a new comprehensive adequacy framework"* – see Shimko and Went (2010), p. 122.

Basel II, set forth in 2004, took into account some of the measures put into action to solve problems that steamed from Basel I. Also, the development of Basel II followed the intentions in EU to reconcile its financial markets in order to allow the equal competition between institutions of its member states. The new framework enhanced the effort to develop and implement banks' internal risk models, as will be explained later in this chapter. Also, and besides credit risk and market risk, Basel II further covered operational risk when determining the banks' MCR.

Like its predecessor, Basel III provides common recommendations and guidelines concerning the management of the various risks in banking, such as credit, market, operational and, additionally, liquidity risks.

Even though there are already three versions of the Basel Accords, the most relevant nowadays are the second and the third, Basel Committee on Banking Supervision (BCBS) (2006)

Basel II and Basel Committee on Banking Supervision (BCBS) (2010), Basel III. In the European Union (EU), the legislation under European Parliament (2013a) (Capital Requirements Directive – CRD) and European Parliament (2013b) (Capital Requirements Regulation – CRR) are the current basis in use to strengthen and harmonize the regulation and supervision of banks, in order to contribute to the global financial stability. These documents, jointly with the Regulatory Technical Standards (RTS) and Implementing Technical Standards (ITS), constitute the Single Rulebook and are the transposition of Basel Accords to EU legislation on prudential requirements and capital adequacy for credit and financial institutions.

In order to step supervision further, ECB also conceived the Supervisory Review and Evaluation Process (SREP) used to guide the supervisory review of significant and less significant credit institutions. In order to support this process, European Banking Authority (EBA) prepared a document with a set of relevant guidelines on common procedures and methodologies for credit institutions risk supervision that enabled the participation and contribution with questions from the credit institutions – *cf.* EBA (2014).

The banking supervision in Europe falls under the responsibility of ECB along with the respective national authorities in the member states. EBA and the European Systemic Risk Board (ESRB) also play important roles in this matter. An effort has been made by EU, BCBS and credit institutions to obtain a greater involvement in the construction of a consultation and cooperation platform. These cooperation efforts support the objective set by the Basel Committee, and shared by the European Central Bank, of a global convergent regulatory framework in the financial system. One of their main concerns is the fulfilment of minimum capital requirements, the first pillar of Basel Accords.

The purpose of establishing minimum regulatory capital under Pillar I is to ensure that financial institutions have enough capital to secure obligations and risks that may arise from unexpected losses (UL). Notwithstanding, the requirements from regulators go beyond since they want banks to also consider the internal capital adequacy self-assessment undertaken in Pillar II and, in this way, challenge them to hold more capital than the minimum regulatory capital. As a consequence, banks had to redefine their strategies, to be able to develop internal estimates to assess, manage and mitigate the risk, so that an improvement in their governance and risk management practices would be achieved. Ten years have passed since the introduction of the IRB concept and EBA considers that time has come to review and improve the implementation of this approach, since banks rely now more than ever on mathematical models that need to be accurate in order to reduce the model risk, i.e., the potential losses that might arise from relying in wrong model outputs to business decision-making.

1.2 The Two Approaches to Assess Credit Risk

Within Pillar I of Basel Accords, institutions are allowed to choose between two broad methodologies to assess their credit risk and the respective minimum capital requirements for own funds: the Standardized Approach and the already mentioned Internal Ratings-Based Approach. The main difference between the two methodologies holds in the fact that the Standardized Approach assumes predefined risk weight parameters, applied equally in all institutions that are subject to this approach, while banking institutions that follow the IRB approach (*vide* Section 2.4) are

allowed to use their own estimates for the key parameters of credit risk, not only as inputs to calculate MCR, but also as relevant knowledge to conduct their management activity. As one can expect, this represents a more sensitive and adequate approach to assess credit risk within an organization. Using internal estimates to assess credit risk within the institution's governance will permit the optimization of the bank's own funds and the possibility to reduce the capital requirements, therefore allowing for the distribution of larger dividends to shareholders.

Although the IRB Approach is broadly recognized as a valid risk sensitive way of measuring capital requirements, it compromises the comparability of capital requirements and risk estimates among banks, due to the high degree of flexibility underlying its framework, therefore representing a significant weakness and inconsistency for the supervisory process – see the Executive Summary of EBA/DP/2015/01 (2015). Also, the variations observed in capital requirements over the business cycle – pro-cyclicality – and their impact on credit behaviors and on the overall financial sector, have been a source of worry for supervisors – see Article 502 of European Parliament (2013b). More recently, efforts have been made to analyse the factors that are affecting the comparability between institutions and to manage new regulatory validation techniques and practices. Specially, the factors concerning the definition of default and the assumptions under the calibration of credit risk parameters used by each credit institution.

Internal Risk Rating Systems (IRRS) are based in the assumption of banks using their own data as inputs to develop empirical models to determine the regulatory minimum capital requirements for credit risk. Such systems must be able to estimate the credit risk parameters, Probability of Default (PD), Loss Given Default (LGD) and Exposure at Default (EAD), as will be explained in Section 2.5. According to Ozdemir and Miu (2006) p. , the design of IRRS is *"the very first step of implementing Basel II under the IRB Approach"*, p. 1.

If credit institutions have enough and relevant historical information from their clients' credit behaviour – whether quantitative or qualitative – they will be able to categorize obligors within different risk levels and assign them with the respective risk rating according to the credit risk they represent for the institution. The credit rating, or credit scoring as will be explained, perform as a rank so that every obligor within the same risk level represent the same risk for the institution, ie, they are assigned with the same PD, LGD or EAD – as Figure 1.1 shows.

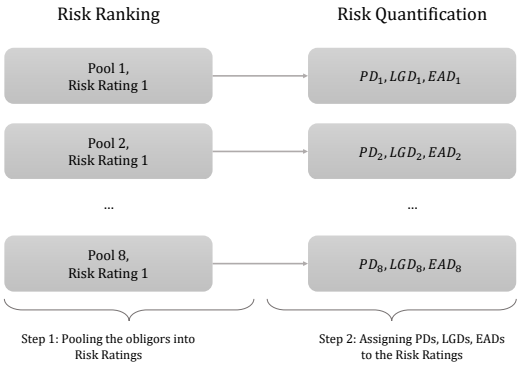

Figure 1.1: Risk Ranking

Adapted from: Ozdemir and Miu (2006)

From the dependence on mathematical models, and since *"more emphasis is now given to the problem of validating a rating system, rather than how to design a rating system" – cf.* Englemann and Rauhmeier (2006), p. 244 – regulators and banks face several challenges in the validation of the IRRS that generate the outputs used to calculate the regulatory capital for credit risk. Some of the most critical challenges for credit institutions arise from: (i) the lack of relevant length and quality information from default experience (when compared to external rating agencies); (ii) the assumptions banks rely on to mitigate the pro-cyclicality on regulatory capital requirements; (iii) the definitions of default and non-performing loans (NPL) applied by the credit institutions, which should be consistent with the definitions set by the regulators – *cf.* default definition in Basel Committee on Banking Supervision (BCBS) (2006).

Regulators believe that IRB Approach is the right step towards a stronger risk management and sounder financial system. Besides, this approach represents a mandatory requisite since banks have to prove and report to the competent authorities that they are assessing their internal risk estimates based in adequate and accurate models, which confirms the rising importance of the validation of credit risk models among institutions' risk management and the implementation of IRB Approach.

1.3 The Aim of the Project

The present project is itself a result of the increasing regulatory framework around the financial system. In particular, it addresses the explicit need to develop accurate validation techniques within the IRB Approach, required by the regulators to enable the comparability between credit institutions and the mitigation of pro-cyclicality in their capital requirements. In fact, and although there is a lot of regulation and guidelines set by the international supervisory community, as well as extensive bibliography concerning risk management, there is still place for broader studies and practical research on how to develop these models adequately. A general agreement on the appropriate validation process for banks' internal risk assessment within the IRB Approach is still to be found.

Considering the most recent guidelines from the ECB and EBA, this work aims to describe and illustrate the empirical procedure of validating a scoring system in what relates to the calibration of the scoring scale. The practical relevance of it urges, on one hand, from the challenging task that the construction and development of these models is, since they are required to meet the necessary conditions to be approved by the supervisor and just then can be implemented internally. An adequate validation methodology entails a complex and demanding process in terms of cost, time, information systems, data quality and human capital. On the other hand, the problem comes from the highly competitive environment of economy and markets, leading to the fact that those credit institutions already using the IRB Approach to generate their own estimates for assessing credir risk are not willing to disclose their knowledge and practices to the overall financial system, since they benefit from an important competitive advantage.

Following the above thoughts, and although it would be interesting to explore the supervisory institutions perspective, an option had to be made to explore the role of credit institutions in the assessment of their capital adequacy. Also, and despite the procedures behind the implementation of IRB models, the focus will be on the methodologies behind the validation process.

The present work intends to be aligned with the regulatory exigencies and guidelines set by regulators along with the support given by some relevant books and studies already existing in these area. Specifically, there is the intention to explore the specificities and assumptions behind the statistical validation and the calibration process of internal estimates for credit risk assessment, since it represents the major risk faced by credit institutions.

Due to the difficulty to analyse more than one credit risk parameter given the existing dimension constraints, this work will focus in the comparison of the predicted probabilities of default with the realized default rates – PD backtesting – and the validation of their accuracy – PD calibration. The procedures behind the backtesting were applied in the context of retail banking. Examples of retail banking products are credit cards, real estate or mortgage and consumer loans. This type of portfolios have the characteristic of high granularity which allows banks to benefit from economies of scale by homogenizing the decision and monitoring processes. Also, the historical data used as basis for scoring models is generally homogeneous and allows the implementation of uniform methodologies within internal estimation. The choice of using a residential mortgage portfolio to illustrate this procedure is due to the availability of the relevant length and quality of information, as required by the supervisor, allowing for the appropriate and accurate processing of data in contrast to other available portfolios. The objective is to propose a practical and adequate statistical validation methodology, compliant with the existing regulatory framework.

1.4 Literature Review

As stated in Basel Committee on Banking Supervision (BCBS) (2005b), p. 1, validation is *"a fundamental aspect of the IRB approach, so much that the Accord Implementation Group has established a Subgroup to examine issues related to validation"* and vast literature can be found on this subject and others concerning credit risk management, regulatory capital allocation, implementation of the IRB Approach, credit scoring models, validation techniques, backtesting and calibration of probabilities of default, including the thorough legislation and studies set up by the international supervisory community.

Common legislation and guidelines are found within Basel Committee on Banking Supervision (BCBS) (2006) and Basel Committee on Banking Supervision (BCBS) (2010). Besides these, other important document – Basel Committee on Banking Supervision (BCBS) (2005a) – was problished by the Subgroup on Validation, justified by the *"need to develop means for validating the systems used to generate the parameters that serve as inputs to the IRB approach to credit risk"*, p. 4. Within this subgroup, and in the context of rating systems, *the term "validation" encompasses a range of processes and activities that contribute to an assessment of whether ratings adequately differentiate risk, and whether estimates of risk components appropriately characterize the relevant aspects of risk"* – *cf.* Basel Committee on Banking Supervision (BCBS) (2005b), p.2.

Basel Committee on Banking Supervision (BCBS) (2005a) gives a compilation of studies on the validation of internal rating systems, to be taken into account by both national supervisors and banks during the implementation of IRB Approach. Specifically in what concerns PD validation, this document makes a distintion between the validation of the discriminatory power of

a rating system and the PD calibration. Also, it develops the assumptions under the classification of rating systems, namely the differences between the so-called Risk Rating Philosophies – point-in-time (PIT) and through-the-cycle (TTC) – and under historical default experience. The authors further point out the scarcity of data as a major obstacle to PD backtesting, expressing their opinion that the application of statistical tests alone is a limited way to adequately validate an IRRS, benchmarking being a complementary instrument.

Within EU regulation, an important reference is, EBA/RTS/2016/03 (2016), which represents a major contribution to the description of the assessment methodology of the compliance of an institution with the requirements needed for the implementation of IRB Approach. It is *"considered an integral part of the efforts of the EBA to ensure consistency in model outputs and comparability of risk weighed exposure amounts"*, p. 3. More precisely, supported by CRR and CRD IV, this document aims to ensure consistency for IRB minimum requirements such as (i) validation, (ii) definition of default and (iii) own funds requirements calculation. Recently, ECB concluded a Guide for the Targeted Review of Internal Models – European Central Bank (2017b) – set as one of the SSM supervisory priorities for 2017 in order to enforce the harmonisation of internal-ratings based approach requirements among credit institutions in EU.

Some relevant bibliography and papers were also considered during the development of this work. Shimko and Went (2010), for example, give a helpful clarification on credit risk management assumptions and practices, including the regulatory view of capital requirements. For banks that are more specialized in traditional lending and credit, the authors state as basic concepts for measuring credit risk the probability of default, the recovery rate, the exposure at default, the expected loss, the loss given default and the unexpected loss, see pp. 13-16. Some valuable considerations about retail credit products, as it is the case of residential mortgage loans, credit scoring and the regulatory perspective of credit risk can also be found in this book.

For the implementation of Basel guidelines on estimation and validation of credit risk parameters, three helpful referrences are Englemann and Rauhmeier (2006), Ozdemir and Miu (2006) and Löeffler and Posch (2007). Englemann and Rauhmeier (2006) detail scoring models for retail exposures, and the statistical approaches to PD validation, both from the perspectives of banks and supervision. They highlight that correct recognition of defaults is crucial to the process, *cf.* pp. 289-290, and include relevant insights about PIT and TTC, the different assumptions underlying and the usual statistical tests for PD validation.

Ozdemir and Miu (2006) give a contribution in internal risk rating systems validation, namely by suggesting a specific overview within PD backtesting and Löeffler and Posch (2007) as a practical manual, useful for the estimation of credit scores with the *logit* function and for the validation of rating systems and credit portfolio models. Also for credit scoring modelling see Thomas et al. (2002). Concerning specifically credit scoring model validation and PD backtesting, Tasche (2006), Wu (2008), and Castermans et al. (2010).

The validation process have been addressed by some authors, who generally divide the backtesting in the validation of discriminatory power and the validation of the accuracy of PD estimates (calibration) – see, for example, Basel Committee on Banking Supervision (BCBS) (2005a) or Ozdemir and Miu (2006). Other researchers additionally include the validation of the model stability as part of the backtesting analysis – see Maarse (2012) and Švec (2012). Authors tend

to use different concepts and assumptions for PIT PD and TTC PD, which is particularly unhelpful in the validation process, since institutions prefer PIT PD for pricing, but are required by the regulators to have their capital requirements stable and based on the TTC assumptions.

Although there are no unique definitions in either literature or legislation, two aspects were found to be relevant and to benefit from generalized agreement. First, the importance of adequate data quality covering a sufficiently large observation period, particularly important to prove the accuracy of estimations and perform further calibration. Second, the demand of qualitative review and expert judgement components to complement the statistical validation techniques in credit risk management.

For the international supervisory community, the comparability of methodologies and estimates among credit institutions is fundamental for an effective implementation of the procedure within the overall financial system. This is why it is so important for European credit institutions to use in their internal governance definitions which are compliant with the ones in European Parliament (2013b). In light of this, this work strictly follows the guidelines on banking supervision provided by EU competent authorities and BCBS.

1.5 Organization of the Text

The text is divided in five chapters. Following the Introduction, Chapter 2 gives a more complete review of concepts related with credit risk and the existing regulatory framework about credit risk management – the Basel Accords. The main concepts concerning the characteristics of credit scoring models and their estimation, as well as the validation process of internal ratings-based systems are also addressed in the second chapter.

Chapter 3 is focused on the discussion of the validation of internal ratings and scoring systems, giving special emphasis to the assessment of the accuracy of PD and the various assumptions underlying the process. The discussion of the Risk Rating Philosophies and the statistical techniques commonly used in the calibration process will also be covered in this chapter.

The process of backtesting and calibration tests is illustrated through an empirical analysis, in Chapter 4. The sample of default experience within a real residential mortgage portfolio that served as input for this application is described and validated, as well as some other available parameters. The way the process was conducted through its different phases is also described in this chapter.

Some conclusions and empirical results are outlined in Chapter 5. Additionally, a few improvement proposals, the constraints and limitations that have been experienced, and some final thoughts and suggestions for further research close the text.

To ensure the confidentiality of the data base, the original information from the default portfolio was modified using a linear transformation. All the statistical results were obtained using MS Excel.

Chapter 2

Credit Risk

All institutions, independently from the nature of their activity, face risks. Risk represents the possibility or threat that an event could cause a damage or an adverse occurrence, wherever it arises from external or internal circumstances. This concept is subsequently connected to the concept of uncertainty. In fact, uncertainty stems from the lack of full information and the quality and reliability of it. For financial institutions, risk is usually referred as the possibility to incur in a loss, or in a lower outcome than expected, due to some uncertain event.

Given the risks that banking has to deal with, risk management is a fundamental part of its governance. According to Shimko and Went (2010), p. 167, *"risk management is a structured approach to monitoring, measuring, and managing exposures to reduce the potential impact of an uncertain event happening"*. An effective risk management allows to mitigate or smooth the impact of the various risks a financial institution runs into.

In order to be effective, banks must be able to identify both the drivers and the controls connected to risk. In risk terminology, drivers are the causes that lead to risk and increase the uncertainty of adverse events, while controls represent the tools that allow the institution to mitigate the results the adverse events can cause. Due to the nature of their activity, banks are mostly exposed to credit risk and a crucial aspect for their survival is its adequate assessment and measurement.

In *Glossary of ECB*, credit risk is defined as *"the risk that a counterpart will not settle the full value of an obligation – neither when it becomes due, nor at any time thereafter. Credit risk includes replacement cost risk, principal risk and risk of the settlement bank failing"*. Gathering this concept with the one of risk management, credit risk management can be defined as the process of monitoring and measuring the exposures to reduce potential losses due to credit risk events, such as default.

Although risk events are uncertain, it is possible to estimate the likelihood of their occurrence and measure their expected impact in the institution's outcome. After the recent global financial crisis, measuring credit risk represents one of the hardest challenges for banks since it became the largest risk for credit institutions. Simultaneously, it is a continuous and thorough task for supervisory authorities in the process of regulating and monitoring banks with the aim of ensuring their solvency and the soundness of the global markets. Due to its complexity, the management of credit risk is a difficult process, especially for those institutions that choose to use internal models and estimates to assess their credit risk within IRB Approach.

2.1 Pillar I – Capital Requirements for Credit Risk

Regulatory capital is the minimum amount of capital that regulators consider to be necessary for banks to hold, in order to guarantee their solvency and absorb potential losses. To avoid past mistakes, regulators set that banks must group its assets by risk category in a way that the amount of capital requirements match with the risk level of each asset. As one can expect, assets that are more likely to default, and that have higher potential losses in case of default, must be considered with higher risk weighting. In this sense, and according with *Financial Times Lexicon*, risk-weighted assets (RWA) are considered *"by adjusting each asset class for risk in order to assess the bank exposure to potential losses"*. Residential mortgage exposures, for example, are less risky since they are associated with property as collateral, as in the case of residential mortgage exposures.

The value of RWA *"equals the sum of various financial assets multiplied by their respective risk-weights and off-balance sheet items weighted for their credit risk, according to the regulatory requirements outlined by banking regulators and supervisors"*, Shimko and Went (2010), p. 167. Under Pillar I, RWA for capital charges must be calculated by one of the two different approaches already mentioned, and described with more detail in Section 2.4. In Appendix A, the formulas that credit institutions that have already implemented the IRB Approach must use to compute RWA and capital requirements for retail exposures are given. These formulas can be found, for this and other exposure categories, both in Basel Committee on Banking Supervision (BCBS) (2006) and European Parliament (2013b).

2.2 Default and Default Rate

Credit risk is closely related with the concept of default. Generally, default events relate to a failure to complete a transfer of funds or securities in accordance with the terms and rules of the system in question – *cf. Glossary of ECB*. This means that default is the failure to pay on the due date or the break of the agreement when the obligor is not able to meet his/her legal obligations. So, the concept of credit risk also represents the possibility of deterioration of credit quality from the counterparts, even if it actually does not imply the default, but increases the probability of it.

Both quantitative and qualitative conditions must be used to identify a default. Credit institutions must define within their governance a definition for default consistent with the one given by Basel Committee on Banking Supervision (BCBS) (2006), § 452-457. In light of these, European Commission published a compliant definition for default in Regulation (EU) no. 575/2013, article 178: *"a default should be considered to have occurred with regard to a particular obliger when either or both of the following events have taken place:*

1. *The institution considers that the obliger is unlikely to pay its credit obligations to the institution, the parent undertaking or any of its subsidiaries in full, without recourse by the institution to actions such as realizing security;*

2. *The obliger is past due more than 90 days on any material credit obligation to the institution, the parent undertaking or any of its subsidiaries. Competent authorities may replace*

the 90 days with 180 days for exposures secured by residential or small and medium-sized enterprises commercial real estate in the retail exposure class, as well as exposures to public sector entities."

It is important to refer that the same document suggests some elements to be taken into account as indications of unlikeness to pay. These can be seen as the qualitative conditions to describe default, for instance:

1. *The institution puts the credit obligation on non-accrued status* [non-performing];
2. *The institution recognizes a specific credit adjustment resulting from a significant perceived decline in credit quality subsequent to the institution taking on the exposure*;
3. *The institution sells the credit obligation at a material credit related economic loss*;
4. *The institution consents to a distressed restructuring of the credit obligation where this is likely to result in a diminished financial obligation caused by the material forgiveness, or postponement, of principal, interest or, where relevant, fees. This includes, in the case of equity exposures assessed under a PD/LGD approach, distressed restructuring of the equity itself*;
5. *The institution has filed for the obliger's bankruptcy or a similar order in respect of an obliger's credit obligation to the institution, the parent undertaking or any of its subsidiaries*;
6. *The obliger has sought or has been placed in bankruptcy or similar protection where this would avoid or delay repayment of a credit obligation to the institution, the parent undertaking or any of its subsidiaries.*

For the purpose of guidance to banks on the non-performing loans (NPL) referred in 1., and given the recent experience of high levels of NPL in Euro-area member states, ECB published final guidelines on this matter – *cf.* European Central Bank (2017a).

In the specific case of retail exposures, and contrary to what happens in the case of corporate exposures, regulators instruct institutions to apply the definition of default at the level of an individual credit operation rather than in relation to the total obligations of a borrower. This means that, for retail exposures, the default by a borrower on one obligation does not require the bank to treat all the other borrower's obligations as defaulted.

Measuring default consists in computing default rates (DR). In Part I, Article 4 (1), § 78 of European Parliament (2013b) "one-year DR" is described *"as the ratio between the number of defaults occurred during a period that starts from one year prior to a date T and the number of obligors assigned to this pool one year prior to that date"*. So, the formula presented below is consistent with the "one-year default rate" definition provided by the European Parliament:

$$DR = \frac{no.\ of\ obligors\ that\ defaulted\ during\ the\ one\text{-}year\ observation\ period}{no.\ of\ non\text{-}defaulted\ obligors\ at\ the\ beginning\ of\ the\ observation\ period} \tag{2.1}$$

In the table below a simple example illustrates the calculation of DR within a scoring system with eight pools, using Equation 2.1.

Table 2.1: Calculation of Default Rates within a Rating System

Pool	**No. of performing operations**	**No. of defaults**	**DR**
1	65 321	14	0,02 %
2	72 596	42	0,06 %
3	105 069	159	0,15 %
4	133 548	700	0,52 %
5	102 365	1 603	1,57 %
6	67 415	2 149	3,19 %
7	26 983	3 123	11,57 %
8	11 078	2 658	23,99 %

2.3 Ratings and Scores

The assessment of credit risk is an essential issue for both credit institutions and regulators. Among the main actors within the banking activity, one must refer rating agencies as well. Rating agencies first came up early in the twentieth century and their mission was to provide information on credit risk of institutions, denominated their 'credit ratings'. However, rating agencies are not the only ones focusing on risk measurement as banks and other credit institutions also measure their own credit risk.

Credit rating *"means an opinion regarding the creditworthiness of an entity, a debt or financial obligation, debt security, preferred share or other financial instrument, or of an issuer of such a debt or financial obligation, debt security, preferred share or other financial instrument, issued using an established and defined ranking system of rating categories"*, see Article 3, § 1 of European Parliament (2009).

Although both ratings and scores were created to assess borrowers' likelihood of repaying their debts, there are a few differences. Rating agencies usually provide ratings and information about businesses, corporations or governments, while banks' are more focused on the credit risk measurement of individuals. Given this, the term rating is used for corporates and governments, and its grades are usually expressed as a letter, although the scales often vary among rating agencies and banks. On the other hand, the term score is often associated to reflect the creditworthiness of the borrower, often a individual or a small enterprise, and is expressed by a number. The common feature hold in the calculation of ratings and scores given that they both rely on historical information that either rating agencies or banks have from companies, governments or individuals.

In this project, since the application is on a residential mortgage default portfolio, the credit scoring models are of particularly interest, as well as the available past default experience on credit behaviour and personal characteristics of the institutions' individual borrowers (*vide* Section 2.6).

2.4 Methodologies to Assess Credit Risk under Basel II/III

As already explained, in order to compute capital requirements regarding credit risk, BCBS allows banks to choose between two broad methodologies to determine the relevant risk weights for capital allocation: the Standardised Approach and the IRB Approach. Further, the Internal Ratings-Based Approach can itself be separated in the Foundation IRB (F-IRB) and the Advanced IRB (A-IRB). The differences between Standardized and IRB are related with the higher sensitivity and complexity that the latter implies, as showen in Figure 2.1.

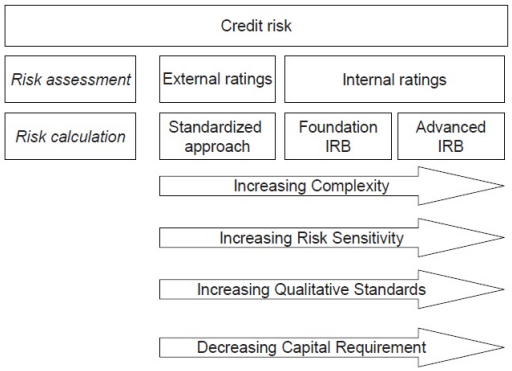

Figure 2.1: Three approaches to manage credit risk under Basel II

Source: Wu (2008)

The calculation of credit risk capital by the Standardized Approach is based broadly in predefined risk weights by the regulator. These risk weights are applied in a standardized manner to the different assets according to the respective asset exposure class, and are the same for all credit institutions using this approach.

In the Capital Requirements Regulation of EU, the risk weights are explained in detail in Section 2. In the specific case of retail exposures secured by mortgages on residential property, the risk-weights must be set respecting the regulatory guidelines in a range of percentages between 35% and 150%.

Within IRB approach, banks are allowed to rely on a system using their own estimates for credit risk parameters as inputs for the calculation of risk weights and capital requirements. According to EBA, the underlying premise for this methodology is that the differences in risk weight of the various exposures effectively reflect the differences in the underlying risk of those exposures, including the structure of the portfolios, the characteristics of the borrowers and credit operations, as well as the internal risk management and collection processes at the institution - see § 7 of EBA/DP/2015/01 (2015).

For the purpose of the calculation of risk weights and capital requirements, credit institutions are allowed to use their own estimations for credit risk parameters as inputs are required to use specified formulas provided by BCBS and EU regulation. Due to the heterogeneity of exposures among loan portfolios of a credit institution (corporate, sovereign, retail, mortgages, etc), banks are required to compute PD estimates for each loan, although they can use different PD models for each class of exposure (examples of exposure classes: sovereign, retail, banks, equity, etc) – see § 395 in Basel Committee on Banking Supervision (BCBS) (2006).

This approach represents a more adequate form for credit risk management within credit institutions governance. The following short table summarizes the main differences between Standardized Approach and IRB Approach, Foundation and Advanced, in what concerns the origin of the estimates for credit risk parameters, i.e., whether these are provided by the supervisor or are estimated by the bank.

Table 2.2: Origin of Risk Parameters under Standard and IRB Approaches

	Standard	**Foundation IRB**	**Advanced IRB**
Probability of Default	Regulatory estimate	Internal estimate	Internal estimate
Loss Given Default	Regulatory estimate	Regulatory estimate	Internal estimate
Exposure at Default	Regulatory estimate	Regulatory estimate	Internal estimate

In Basel II, the minimum conditions needed for IRB Approach approval are described in detail within Part 2 – Chapter III – Credit Risk – The Internal Ratings-Based Approach – Section H (§ 387-537). For retail exposures, banks must use their own estimates of the three credit risk parameters, independently of the approach, F-IRB or A-IRB. The reason for this results from the availability of more complete data within these type of exposures.

2.5 Credit Risk Parameters

Credit risk measurement depends essentially on the three key risk parameters in the table above, used to compute the regulatory capital: (i) Probability of Default (PD), (ii) Loss Given Default (LGD); (iii) Exposure at Default (EAD).

The calculation of risk weights depends on the estimates of the above parameters, as one can understand by observing the formulas provided in Appendix A. As part of an effective risk management system there is the need to develop accurate models to measure the overall exposure of an institution to credit risk. Also, these three parameters are the inputs to compute the amount of EL, already referred in Chapter 1, by simply multiplying them.

Although many definitions exist, PD can be generally defined as the measure of the likelihood that a borrower will default on his obligations over a given time horizon - usually one year. PD per rating grade - or pool, when referring to retail exposures - represents the average ratio of borrowers that default among the rating grade, in one-year period, Wu (2008). Default probabilities are estimated using statistical models. Scoring models are often used to estimate PD since they are recognized as providing accurate estimations for it, as will be explained in the following section.

In general terms, LGD represents the loss severity of exposure at default, i.e., the percentage of likely loss in the eventuality of the borrower's default given the amount that is at risk. This measure considers costs of the associated recovery and the mitigation effect (e.g. collaterals such as in residential mortgage loans that are secured by property).

EAD consists in the amount that the bank can lose if the borrower defaults on his/her obligations, ie, it is the amount in currency that the bank has exposed to the borrower at the time of default.

It is important to ensure the completeness and representativeness of the data to guarantee the reliability of the developed models used to estimate these parameters. In particular, banks must ensure that a sufficiently long period of historic data is used to calculate internal estimates and that it complies with the definition of default provided by the regulators. The complexity within the development, implementation and use of internal models in a bank is the reason why both validation and calibration are so important.

Of the three listed parameters, PD will be the only one developed in this work since it is the one for which *"estimation and validation methodologies are significantly more advanced than those for LGD and EAD"* – see Chapter 2 of Basel Committee on Banking Supervision (BCBS) (2005a). The objective of internal rating systems is to group obligors into homogeneous risk categories and order them correctly in terms of their PD, i.e. the credit risk they represent. Given this, the more accurate the PD estimates are the more reliable a credit model will be and so the better a bank will predict its credit risk.

An adequate identification and counting of defaults in a portfolio is a fundamental step for banks' credit risk management and for the validation of their internal risk rating system. This is so because past defaults (provided by DR) must be regularly compared with the estimations of expected future defaults. Again, the correct count of defaults must be made in compliance with the definition provided by the regulators.

2.6 Scoring Models

In order to assess credit risk, statistical models are used in banks to classify borrowers according to the risk they represent and to predict the overall risk exposure within the various credit portfolios. Statistical rating systems demand the previous investigation of exogenous variables able to adequately explain and forecast the *"possible deterioration"* of the creditworthiness of a borrower – *cf.* Englemann and Rauhmeier (2006). The explanatory variables must include relevant borrower's characteristics and macroeconomic indicators obtained from the available historical information of the credit institution, in order to build accurate forecasting models. With these models, banks aim to predict the future status of borrowers whose performance on meeting obligations is still unknown, i.e., the borrowers' likelihood to default or not default during the next year.

In order to predict the likelihood of default of the borrowers within a portfolio, parametric models are developed such as credit scoring models. These models provide broad advantages when managing credit risk – credit approval, portfolio monitoring and credit risk quantification. They rely on statistical models and not on the traditional lenders' "gut feel" – see Thomas et al. (2002) or Wu (2008).

Credit scoring models are the banking standard for the rating of retail portfolios, as is the case of residential mortgage loans – *vide* Englemann and Rauhmeier (2006). Such portfolios have the characteristic of high granularity which allows banks to benefit from economies of scale by homogenizing their decision and monitoring processes.

When assessing the borrower's creditworthiness among scoring models the output is expressed in terms of a number on a continuous scale – the score. Usually, higher scores correspond to lower default probabilities. Using historical data and statistical methodologies, score models

will be mapped within default probabilities. This allows the segmentation of borrowers into homogeneous risk groups and separation between the "good" from the "bad" borrowers. In this way, banks will assign a single PD for clients in the same class of risk, i.e., with the same risk grade.

Standard scoring models linearly combine the factors (both quantitative and qualitative) that can affect a borrower's default probability, such as the salary, professional status, age, and other characteristics concerning his/her credit behaviour history. So, according to Löeffler and Posch (2007), let these factors be represented by x and let b represent their weights. Then, the score for scoring instance i, and assuming K factors, can be represented as follows:

$$Score_i = b_1 x_{i1} + b_2 x_{i2} + ... + b_K x_{ik} = \mathbf{b'x}_i. \tag{2.2}$$

2.6.1 Estimating Scoring Models with *Logit*

There is a range of statistic methodologies to build a scoring model, such as *logit* models, neural networks and decision trees. Since it is necessary to link scores to the correspondent default probabilities, this can be done by associating these probabilities as a function of scores. The score function must be constrained to the interval from 0 to 1, so that each possible score has a correspondent default probability. This requirement can be satisfied using the logistic cumulative distribution function,

$$\Lambda(z) = \frac{e^z}{1 + e^z} \tag{2.3}$$

Logit models represent an econometric technique intended to analyse dummy dependent variables and it is the most used technique in credit scoring, see Löeffler and Posch (2007). In this way, PD can be linked with the score value using the following formula:

$$Probability(Default_i) = \Lambda(Score_i) = \frac{e^{\mathbf{b'x_i}}}{1 + e^{\mathbf{b'x_i}}} = \frac{1}{1 + e^{-\mathbf{b'x_i}}}. \tag{2.4}$$

The usual process is to collect all the relevant x factors and then estimate the respective b weights using the maximum likelihood function. For convenience, in the present work, higher pools correspond to higher PD, although the opposite reasoning is also possible.

Chapter 3

Calibration of a PD Scoring System

In order to be compliant with the regulatory standards, banks must perform annual reviews of risk parameter estimates computed internally and used for allocation of capital requirements. The calibration of credit risk parameters under the internal validation of rating systems is one of its complex processes. Despite the existing guidelines and some papers related to this matter, by both regulators – such as Basel Committee on Banking Supervision (BCBS) (2005a), European Parliament (2013b) or European Central Bank (2017b) – and other authors – Englemann and Rauhmeier (2006), Ozdemir and Miu (2006) or Tasche (2006) – closer guidance for credit institutions on the subject of internal validation is still missing. Nevertheless, one thing is clear: rating and scoring systems are the core for the implementation of IRB Approach, affecting directly the calculation of regulatory capital requirements, so validation must be performed in order to ensure the good performance of these internal systems and their predictive power.

3.1 Internal Validation

In light of Basel Accords, banks are required to implement a *"robust system to validate accuracy and consistency of rating systems, processes, and the estimation of all relevant risk components"*. As so, the need for reliable validation methodologies in order to prove to the supervisors that credit risk is being assessed adequately within a credit institution is irrefutable. The important concern in this process is to assure that internal estimates are sufficiently predictive and reliable to assess the risk. Banks are required to, in addition, to adequately report this process proving a comprehensive validation, at least annually.

Notwithstanding the relevance given recently to model validation, there is still lacking a thorough guidance for an adequate and commonly accepted validation framework in banking inndustry. Still, Basel Committee on Banking Supervision (BCBS) (2005a) set six principles summarizing what banks must take into account within their IRRS validation process:

1. *The bank has primary responsibility for validation*;
2. *Validation is fundamentally about assessing the predictive ability of a bank's risk estimates and use of ratings in credit processes*;
3. *Validation is an iterative process.*

4. *There is no single validation method*;

5. *Validation should encompass both quantitative and qualitative elements*;

6. *Validation processes and outcomes should be subject to independent review.*

A comprehensive description of each one is provided in Basel Committee on Banking Supervision (BCBS) (2005b).

In Ozdemir and Miu (2006), the authors state that the validation of a rating system has the objective to assess and discriminate accurately credit risk. The concept of validation entails the various procedures included in the assessment of whether a rating system performs well and ensures accuracy, consistency and a certain level of conservatism, to be served as input for the calculation of capital requirements. Internal validation aims to ensure the quality of the rating systems as demanded by the relevant requirements. It starts from the beginning of model development to the posterior model monitoring and outcomes periodic review. Additionally, it embraces the task of ascertaining whether each particular risk parameter estimate differentiates appropriately the credit risk faced by the institution, see for example Wu (2008). It is relevant to highlight the importance of fully access to relevant and quality risk databases to perform the validation of any rating system – *cf.* European Central Bank (2017b), p 19 – "Evaluation of input data" and pp 34-38 – "Data quality".

One of the aspects that hinders this process holds in the requirements for validation of internal estimates in Basel II. Banks must demonstrate that the quantitative testing methods and other validation methods do not vary systematically with the credit cycle. Also, when validating their internal estimates and assessing the performance of their own rating systems, banks must be sure that the analysis is based on long data series covering a range of economic conditions and one or more complete credit cycles – see § 502 of Basel II. Moreover, the same document requires banks to have well-articulated internal standards for situations where deviations in realized PD become significant enough to question the validity of the systems, with respect to their ability to capture of credit cycles and similar systematic variability in default experiences, see § 504, Basel II. As a consequence, banks must be forward-looking and use the available historical data and empirical evidence to improve their validation techniques and their forecasts. As has been said through this text, an accurate, consistent and reliable rating system discriminates risk in an effective way, assigning better ratings to borrowers who represent lower risk (and calibrating it) correctly quantifying the risk.

The above considerations are mostly consequence of regulatory MCR. In fact, banks must be aware that supervisors are not keen to let capital requirements vary over time, but to be stable through all the credit cycle. This leads to the risk rating philosophies and to the concepts of point-in-time, through-the-cycle and long-run average default rates (LRDR), as will be further explain in this chapter.

3.2 Backtesting and Calibration

One particular purpose of this work is to illustrate quantitative validation, although qualitative validation and expert judgement are also relevant and mandatory for an effective validation

framework. As can be observed in Figure 3.1, regulators suggest that there are two complementary quantitative ways to validate banks' IRRS systems: backtesting and benchmarking.

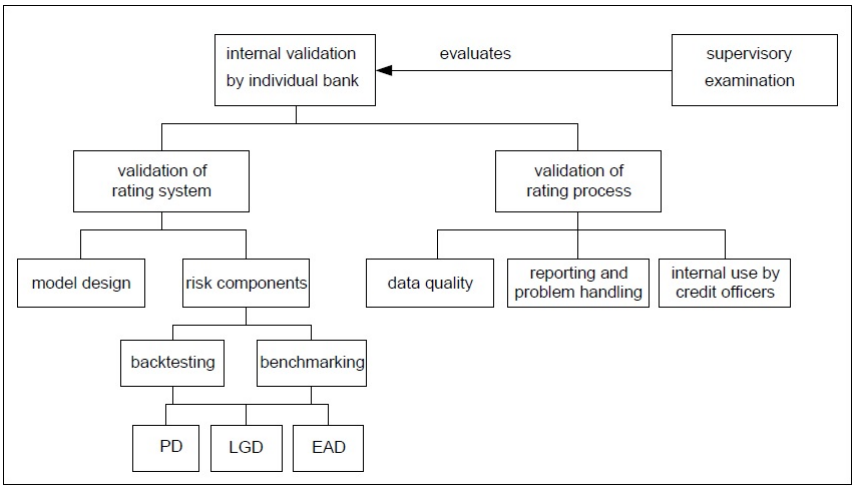

Figure 3.1: Internal validation components

Source: Basel Committee on Banking Supervision (BCBS) (2005a)

To better understand the figure above, and according to Ozdemir and Miu (2006), the validation process can be separated in the following three important dimensions: (i) *validation of concept, methodology and assumptions*; (ii) *verification of replicability and examination of override and exceptions monitoring, key performance indicators, data integrity and the use test* – for the latter *cf.* Article 289 in European Parliament (2013b); (iii) *backtesting, benchmarking and annual health check*. This project mostly covers the last dimension, i.e., the analysis of the outcome of internal rating systems and whether they predict future default behaviour as accurately as possible.

Following Ozdemir and Miu (2006), backtesting is the assessment of the performance of the risk rating system build on its historical data by the comparison between predicted and realized outcomes. Default probabilities must be compared with default rates, in order to assess whether the deviation between the estimates and the observed rates is relevant to call into question the accuracy of the model outcomes. According to Castermans et al. (2010) and Tasche (2006), backtesting evaluates the predictive power, the discriminatory power and the stability of the risk model supported by statistical tests.

Benchmarking is the assessment of the performance and consistency of a risk rating system relative to the comparable risk rating systems, usually done by comparing internal risk measurements with external risk measurement provided by rating agencies. Some limitations can be pointed out concerning benchmarking. In fact, significant difficulties may arise when comparing, for example, internal and external estimates for default probabilities, since banks need to ensure that the definition of default and time horizon used by both parts is the same. It is also important to understand whether PD estimates are through-the-cycle or point-in-time, as the two rating systems need to be assigned to the same risk rating philosophy, in order to be comparable.

In Englemann and Rauhmeier (2006) calibration is defined as the assignment of a PD to a

certain rating or scoring grade. Also Tasche (2006) refers to the concept of calibration of rating systems and score variables as the issue holding with how accurate the estimates of the default probabilities are given the score. From the supervisory point of view, calibration also holds with the validation of the accuracy of the PD quantification - see Basel Committee on Banking Supervision (BCBS) (2005a).

As a regulatory guidance, the sample used for the calibration of PD estimates must cover at least five years of relevant data or a margin of conservatism must be applied if a period less than five years is not available. Banks must assess if the comparison of the most recent information of defaults with the dataset used for PD estimation is leading to material differences, in particular if including the most recent data is leading to a significant difference on long-run average default rate. In order for this to be possible, institutions must define what they consider as a likely range of variability on default rates incorporating "good" and "bad" years of the credit cycle.

3.2.1 Rating Philosophy and Credit Cycle

EBA/CP/2016/21 (2016) makes clear that the rating philosophy must be considered for back-testing purposes. It is easy to understand that the credit behaviour will change over the credit cycle. In other words, it is expected that during a recession period the number of defaults rises, while in an expansion period the number of defaults may descrease. This constraint of under- or over-estimation of PD must be solved in order for the internal ratings system to be sufficiently accurate to be used in the calculation of the regulatory capital within the credit institution.

Usually, where rating systems are designed to be more sensitive to economic environment – PIT – the assignment of PD among different grades will have to be corrected frequently due to significant migration rates, while the observed default within each pool will continue stable. As a matter of fact, if an institution opts for a PIT rating system, it will have to increase the borrowers' PD when facing a recession period and reduce them when facing an expansion period. On the other hand, for less sensitive rating systems – TTC – the default rates observed yearly within each pool will follow the cycle average default rate. Institutions must choose the philosophy they consider most convenient to their business strategy as long as it is applied in a consistent way over time. Also, they must take it into account for both the PD assignment by pools and for posterior outcome analysis.

An important reference is made to this subject in Ozdemir and Miu (2006), stating that banks can only use most recent information and recalibration of the obligors within the different pools at the time of review. What shows that *"a "pure" PIT IRRS is not achievable in practice even if it is the financial institution's intention to attain such a philosophy. It is in fact a hybrid of the philosophies, somewhere in between "pure" PIT and "pure" TTC"*. To better understand this extent, the figure below is presented where the thick curve represents the PIT PD, the horizontal line represents the TTC PD and the broken line between the two represents a hybrid PD.

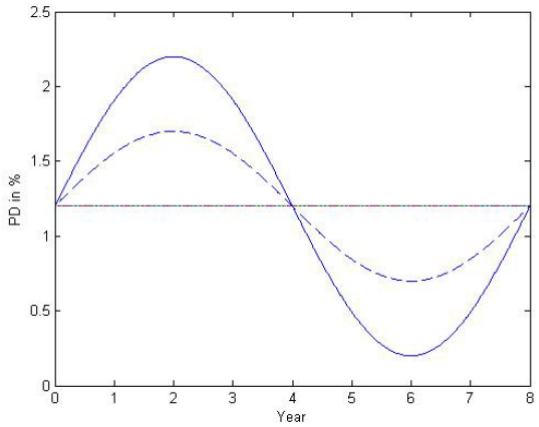

Figure 3.2: Point-In-Time, Through-The-Cycle and Hybrid PD Curves

Source: Gobeljic (2012)

Credit and financial institutions usually use PIT PD since they better take into account the current economic conditions. The difference between PIT and TTC PD must be also made from the perspective of the horizon period they represent. PIT ratings represent the assessment of the creditworthiness of a borrower over a relatively short horizon, usually one year. The other way around, TTC ratings reflect a longer horizon of time, trying to eliminate the credit cycle influences and economic conditions, making them more stable over the cycle than PIT ratings. Assumptions on the basis for risk rating philosophies and credit cycle definition must be thoroughly analysed and defined in order to attain an adequate and comparable validation of default probabilities.

Economic cycle, credit cycle or business cycle are often used when referring to the historical observation period that financial institutions must take into account within their governance and risk management. Also, validation methodologies *"take account of the way business cycles and the related systematic variability in default experience are considered in the internal ratings and risk parameters, especially regarding PD estimation"* – *cf.* EBA/RTS/2016/03 (2016). Also, it is difficult to find an explicit definition of what can be considered a complete credit cycle, since it may vary from institution to institution. One can generally define complete economic cycle as the period between two upturn peaks of economic expansion. However, and as described in EBA/CP/2016/21 (2016), institutions set various requirements to be able to consider a completed credit cycle. A general accepted requisite is that a complete credit cycle must include both good and bad years. Additionally, supervisors refer that banks should consider historical observation periods with a likely range of variability and, in particular, should include a downturn period.

Some authors define other assumptions for risk rating philosophies, for example Tasche (2006), where PIT PD is referred to as conditioned since their estimates are depending on economic variables and to the current state of the credit cycle. With the opposite reasoning, TTC PD are seen as unconditional PD estimates since they incorporate the information of an entire credit cycle and not specific variables for specific periods.

In some studies, such as Miu and Ozdemir (2007), various authors adressed the concept of long-run probability of default (LRPD) and long-run default rate (LRDR) associated to the

regulatory requirement that PD assignment of IRRS should reflect long-run default experience. Also, EBA has been making some considerations regarding LRPD estimation after observing that the differences in the PD calibration come both from the choice of data and the calibration methods used. So, in order to ensure stability of the rating system, EBA suggests that the long run average of one-year default rates should be the basis for the PD calibration – see EBA/DP/2015/01 (2015) and EBA/CP/2016/21 (2016).

3.2.2 Long-Run Average Default Rate

In order to provide clarification on the requirements, EBA states in EBA/RTS/2016/03 (2016) that PD estimates must be based on the long-run average of yearly default rates (LRDR) with the purpose of stability over time and also to avoid the cyclicality of MCRs. Recall that DR are computed as described in Equation 2.1. To perform model calibration, institutions must consider observations from the most recent five years available. If these observations are representative of a likely range of variability of default rates, as already referred, the long-run average of default rates should be colculated as the observed average of the one-year default rates in that period – *cf.* EBA/CP/2016/21 (2016), p. 49, as follows:

$$LRDR = \frac{1}{N}\sum_{i=1}^{N} DR(i), \tag{3.1}$$

where $i = 1, ..., N$ denotes the number of years considered for the average and $DR(1), ..., DR(N)$ are calculated yearly with respect to a certain reference date.

The usual procedure in credit institutions is to estimate PIT PD based on their own historical observed default experience and then calibrate the rating and scoring systems through LRDR or, as often called, through a central tendency – see Bonini and Caivano (2014). In the long run, one can expect that LRDR tends to converge to the long-run average PD. Institutions still need to guarantee the comparability of rating assignments and the posterior observed default frequency, as well as the adequacy of the LRDR to be used for the system's calibration before applying the calibration tests.

Following the thoughts of Ozdemir and Miu (2006) on the impossibility to achieve either "pure" PIT PD and "pure" TTC PD, and knowing that scoring scales are previously adjusted to LRDR reflecting long-run default experience, banks must ascertain if the central tendency underlying is still adequate before measuring the differences between PD scale and observed DR. Institutions must demonstrate that any tests chosen are appropriate.

In fact, calibration starts immediately after the rating system is developed, in order to perform a preliminary validation and ensure that the system meets the minimum requirements to be applied in the bank management and governance. This first model calibration is made to include the available long-run information on default experience that the bank owns. For this reason, the scoring scale that results from the model is preliminarily calibrated to the LRDR and benefits from a PD scale that already includes the information from a relevant historical period of observation.

In the annual review and monitoring exercise, the model is analysed again in order to assess if it remains with a good predictive power. Given that the scale in force already reflects a TTC

philosophy (as required for own funds allocation), the comparability between the PD assigned to this scale and the most recent yearly default events – and then underlying a PIT philosophy – is compromised. Thus, and similarly to the adjustment applied to the rating system at the development, institutions must perform the same correction to the recent observed DR that will be subject to the backtesting exercise. Otherwise, institutions would be comparing predictions and observations that are not measured in the same scale or neither are proportional. Since the present work is focus on the monitored exercise, one is interested that the systems reflects a philosophy rather TTC aiming to obtain a rating or scoring scale in which PD assignment in each risk pool underlays stability over the economic cycle.

For the annual health check, institutions must guarantee that the LRDR is still adequate to serve as adjustment to the new information on default events, observed in the portfolio from the beginning of observation period until the end of the posterior 12 months. There is not a specific supervisor recommendation on the best methodology to apply in this adjustment. Notwithstanding, competent authorities refer that the methodology chosen must be statistically appropriate and duly documented for posterior assessment of compliance with the regulatory framework in force.

In this project, and since this technique is often used for the preliminary calibration of scoring systems, the Bayes' Rule was applied to perform the adjustment of the most recent available one-year DR. In fact, Bayes' Rule is often adopted to achieve revised or posterior probabilities when institutions hold additional information – *cf.* McClave et al. (2014).

Analogously, and used as a common practice by the credit institution that provided the data, the following equation is applied in order to obtain the DR adjusted to the LRDR of the most recent five-year period, denoted as DR_{adjust}.

$$DR_{adjust} = \frac{DR_i \times \frac{LRDR}{DR_N}}{DR_i \times \frac{LRDR}{DR_N} + (1 - DR_i) \times \frac{1-LRDR}{1-DR_N}}, \tag{3.2}$$

with $i = 1, ..., K$ representing the risk pools of the scoring system and DR_N represents the most recent one-year DR available.

To measure if the most recent LRDR is still adequate to the credit cycle in the "new" year, a percentile confidence interval was constructed using the Bootstrap technique. For Bootstrap technique the available yearly default information was replicated to achieve a lower and a upper bound for LRDR. If the most recent LRDR (possible to compute) falls inside the confidence interval then central tendency is still adequate and institutions can use Equation 3.2 to adjust the observed DR and further apply the calibration statistical tests. Otherwise, the LRDR is no longer adequate which suggests that the scoring scale requires revision even before the statistical tests are conducted. Nevertheless, these tests can still be performed in order to assess the magnitude of the differences between estimates and observed rates.

3.3 Calibration Statistical Tests

The calibration process is based on the application of specific statistical techniques that allow to conclude if the differences between the expected PD and the *a posteriori* observed default

rates are significant and if the recalibration of the rating system is necessary.

In this matter, supervisors want the estimates to be more conservative, i.e. not too low, since they are being used to determine regulatory capital requirements. It is important to highlight that for a correct and reliable validation process statistical techniques do not exempt the complementary expert judgement and alternative qualitative approaches, such as the assessment of the performance of the variables introduced in the credit operation process.

When performing the calibration analysis, some statistical tests are more adequate for single-period rating systems and to assess each rating grade individually, while others are more adequate for single-period rating systems but for an aggregate analysis taking into account the total of risk pools together. Moreover, there are some tests more suitable for multiple-period rating systems. For the PIT system, conditional tests such as the binomial test and the Hosmer-Lemeshow test must be applied. For the rather TTC systems, unconditional tests such as the Normal Approximation to Binomial test must be undertaken. The following will be described according to Basel Committee on Banking Supervision (BCBS) (2005a).

3.3.1 Mean Square Error and *Spiegelhalter* Test

Mean Square Error (MSE), is the squared difference between PD estimates and observed DR averaged across the different obligors, and can be represented as follows:

$$MSE = \frac{\sum_{i=1}^{K} n_i \times ((DR_i) - (PD_i))^2}{N-1}, \tag{3.3}$$

where N is the total number of operations in the portfolio and n_k is the number of operations in each risk pool k, with $i = 1, ..., K$ risk pools.

As can be expected, MSE presents a small value when the PD estimate assigned to default events is high and PD estimate assigned to non-default events is low. This means that a low MSE suggests a good rating system.

Spiegelhalter Test is based on MSE since it allows to verify if the computed MSE significantly differs from its expected value. Assuming the independence of the default events in each risk pool and among the different risk pools, the hypotheses underlying this test are:

$$H_0 : PD_k = DR_k \text{ against } H_1 : PD_k \neq DR_k, \text{ for each risk pool } k.$$

Under the null hypothesis, the MSE has, respectively, an expected value and variance of:

$$E[MSE] = \frac{1}{N} \sum_{i=1}^{K} n_i PD_i (1 - PD_i) \tag{3.4}$$

$$Var[MSE] = \frac{1}{N} \sum_{i=1}^{K} n_i (1 - 2PD_i)^2 PD_i (1 - PD_i) \tag{3.5}$$

Using the Central Limit Theorem (CLT), it is possible to show that under H_0 the test statistic, Z_S, that must be used to perform the usual steps of a test decision follows a Standard Gaussian Distribution and is given by:

$$Z_S = \frac{MSE - E[MSE]}{\sqrt{Var[MSE]}} \sim N(0, 1). \tag{3.6}$$

3.3.2 Binomial Test

As previously referred, Binomial Test can be performed for each risk pool individually and for a single-time period. When choosing a fixed risk grade to perform this test, the independence of default events within that grade must be assumed. The hypotheses tested are then:

H_0: the PD estimate for the risk pool is conservative enough *versus*
H_1: PD estimate for the risk pool is underestimated.

So, basically, the hypotheses underlying this test represent the possibility of the observed DR in risk pool i to be less than or equal to the assigned PD estimate against the unilateral hypotheses.

Assuming a critical value of k^* where:

$$k* = min\left\{k : \sum_{i=k}^{n} \begin{pmatrix} n \\ i \end{pmatrix} PD^i(1-PD)^{n-i} \leq (1-\alpha)\right\}, \tag{3.7}$$

one can expect that, if the number of observed default within risk pool i is greater then or equal to k^*, the null hypothesis will be rejected for a confidence level α.

A limitation in performing this test holds with the assumption of independence assumed for the default events, since it is not empirically realistic, but performing Binomial Test assuming the correlation between defaults would make the mathematical approach more difficult. Still, assuming the independence, the Binomial Test is considered by some experts as the most powerfull among all fixed level tests – *cf.* Basel Committee on Banking Supervision (BCBS) (2005a).

3.3.3 Normal Approximation to the Binomial Test

An approximation can be applied to the previous test. This simplification is obtained asymptotically applying the CLT and assuming that, as the number of obligors increases, our previously Binomial Distribution approaches a Normal Distribution. The Normal Approximation to the Binomial Test is particularly useful since it measures the stability over time in each risk grade.

In order to guarantee that this approximation is asymptotically valid, institutions must confer that the following conditions are met:

$n_i PD_i(1-PD_i) > 9$ or, less restrictively, $n_i PD_i > 5$ and $(1-PD_i) > 5$.

Then the following test of hypothesis can be performed:

H_0: PD estimate for risk pool i is correct *versus*
H_1: PD estimate for risk pool i is underestimated.

The following formula gives the confidence interval for this test that can be performed for each risk pool in the scoring system under study,

$$DR_i \approx PD_i \pm \Phi^{-1}(\alpha)\sqrt{\frac{PD_i(1-PD_i)}{n_i}}, \tag{3.8}$$

where DR_i and PD_i represent respectively the observed DR and the PD estimate for risk pool i, n_k represents the number of obligors in each risk pool i and Φ^{-1} represents the inverse Standard Gaussian cumulative function.

3.3.4 Hosmer-Lemeshow Test

The Hosmer-Lemeshow Test performs an aggregate analysis considering the different risk grades simultaneously for a single-period validation. Also, it tests the goodness-of-fit for models derived from *logit* regression, as is the case of the scoring model studied in this work. To perform this test, two assumptions underlay: estimated PD and observed DR are identically distributed and all default events, either within each risk pool or between all risk pools, are independent. The hypotheses tested are:

H_0: all the PD estimates are correct *versus*
H_1: at least one PD estimate is not correct.

Then our test-statistic is defined as follows:

$$HL = \sum_{i=1}^{K} \frac{(n_i PD_i - d_i)^2}{n_i PD_i(1 - PD_i)}, \tag{3.9}$$

where n_i represents the number of obligors in risk pool i, PD_i is the PD estimate in risk pool i and d_i is the number of observed defaults in risk pool i, with $i = 1, ..., K$ different risk pools.

Keeping the assumption of independence (within the same risk pool and between different risk pools), and assuming the CLT when $n_i \xrightarrow{\infty}$ at the same time for all i, $HL \xrightarrow{d} \chi^2_{(k-2)}$ distribution if all PD estimates are the true PD. This is the most accurate test-statistic for institutions using sub-samples. Then the p-value of this test can be used as a measure in favour of the null hypotheses, i.e., a p-value closer to zero suggests a bad estimation of PD. To accomplish the described test, it is advisable to guarantee that $n_k \times PD_k > 5$ and that exists at least one obligor in default in each risk pool. Like the Binomial Test, Hosmer-Lemeshow Test may provide an underestimation of Type I error, i.e., the likelihood of rejecting erroneously the hypothesis of an adequate PD estimate.

Chapter 4

Empirical Analysis

The credit scoring model supporting this part of the project, to illustrate the process of back-testing and further calibration tests of an IRB system, is the one used to estimate the default probabilities, from a certain reference date and over the next 12 months, for borrowers within a residential mortgage portfolio. Data was provided by an anonymous institution willing to meet the minimum requirements to be able to use F-IRB Approach in its risk management, having to prove compliant validation techniques for the internal estimates.

Since the present work is not focused in the development of the model, there is no need to to describe it in detail. It is enough to say that it was developed taking historical data from mortgage loans covering the period from 1999 to 2004 and includes the usual variables in this type of models.

The statistical validation methodology that will be described took into account ten years of historical default experience within the banks' of residential mortgage loans portfolio. The objective is to demonstrate a backtesting methodology, compliant with the current supervisory guidelines, to assess whether the bank relies on accurate and consistent PD estimates for the portfolio in study. The project gives special attention to the recommendations to take into account the systematic variability in default experience – see Article 12 (f) of EBA/RTS/2016/03 (2016). The original data was transformed in order to ensure the confidentiality of the information and all the calculations were performed using MS Excel. Still, results remain valid, coherent and meaningful.

4.1 Data Description

The available data set contains the yearly default information from years 2005 to 2014 for a residential mortgage portfolio, see Table 4.1. For the correct count of defaults, banks have to choose a reference date, $t = T$, and measure the occurrence of defaults during the next 12 months, obtaining the total number of defaults in the portfolio at the end of the observation period, $t = T + 1$. The yearly DRs were computed using Equation 2.1.

Table 4.1: Yearly Default Rates

Year	No. of operations	No. of defaults	DR
2005	570 745	8 414	1,4743 %
2006	590 186	10 596	1,7954 %
2007	597 352	16 873	2,8246 %
2008	600 990	11 116	1,8496 %
2009	614 077	12 029	1,9588 %
2010	620 149	6 906	1,1136 %
2011	620 639	8 130	1,3099 %
2012	603 127	10 273	1,7033 %
2013	575 898	7 416	1,2878 %
2014	559 115	8 325	1,4890 %

The data used for validation covers a historical observation period longer than the recommended (of at least five years). As provided in EBA/RTS/2016/03 (2016), the period length for backtesting is also relevant and representative of a likely range of variability of default rates. For these reasons the available data is adequate for the exercise of statistical validation.

The credit scoring model under consideration classifies the loans in eight numeric different pools of risk, where 1 has the lowest PD associated (representing the better level of risk) and 8 has the higher PD associated (representing the worst level of risk). Figure 4.1 shows the portfolio's credit scoring.

Risk Bucket	PD
1	0,01 %
2	0,05 %
3	0,10 %
4	0,35 %
5	1,18 %
6	2,52 %
7	6,71 %
8	15,73 %

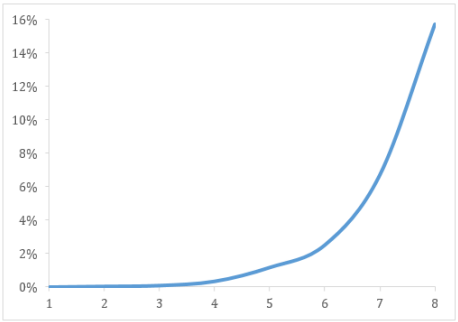

Figure 4.1: Portfolio's Credit Scoring

4.2 Backtesting

As mentioned before, in the process of backtesting the objective is to compare *ex ante* estimation of default probabilities from the credit scoring model with the *ex post* realized default rates. This exercise can only be done at the end of each observation period. Since the estimates from the credit scoring model are PIT, i.e., are contingent on the fluctuations of the economic cycle, the first step in this process is to study if the model stays stable, independently from the current point in the cycle, and if the tendency of the long run default experience stays inside a reliable confidence interval. In order to allow the comparability when performing the calibration tests, the observed DR information must be adjusted to the most recent LRDR available.

4.2.1 Long-Run Default Rate Adequacy

Before applying the usual calibration tests, a fundamental step is to study the adequacy of the long run default rates. For this purpose, and since the supervisors require a minimum length of five years as well for PD validation, the formula for LRDR given in Equation 3.1 was used to compute the simple averages of the observed default rates from five consecutive years. Therefore, six long-run default rates were obtained, which will be referred to as LRDR_j, where $j = 1, 2, ..., 6$. Note that $j = 1$ refers to the period comprising the years from 2005 to 2009, $j = 2$ refers to the period between 2006 and 2010, and so on until period between years 2010 and 2014. The values obtained are in Table 4.2.

Table 4.2: Long Run Averages of One-year Default Rates

Period j	**Observation Period**$_j$	**LRDR**$_j$
1	2005-2009	1,980047 %
2	2006-2010	1,907919 %
3	2007-2011	1,810838 %
4	2008-2012	1,586579 %
5	2009-2013	1,474201 %
6	2010-2014	1,380311 %

Plotting the values of the default rates in Table 4.2, the variability and fluctuation of default rates can be observed, apparently showing a "complete credit cycle", as provided in the previous chapter.

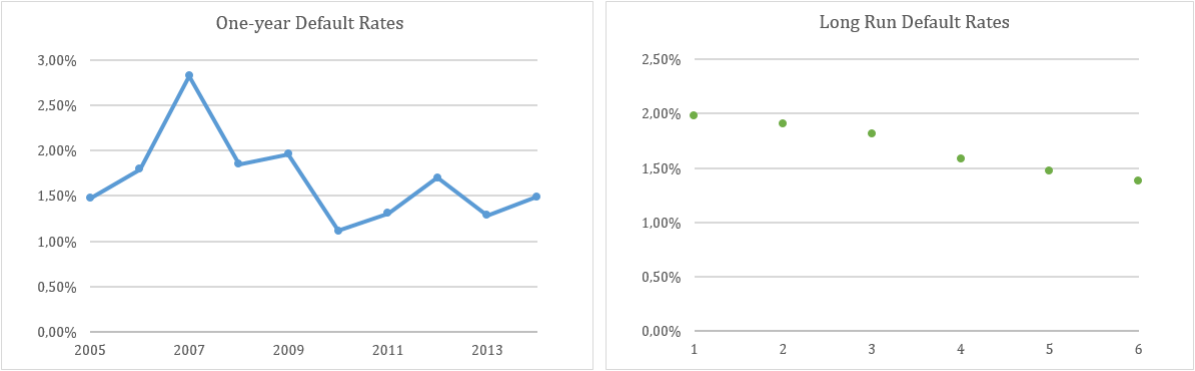

Figure 4.2: Observed Default Rates: Yearly and Long Run Averages

Given the availability of a sample range covering 10 years of default experience of the portfolio, it was found convenient to use the first 9 years' information to infer about the default in year 2014, as if this information regarding the last year wasn't already known, and then compare with the effective observed value. This inference was made using the Bootstrap technique, already described in Chapter 3, to construct a confidence interval for the Long-Run Default Rate corresponding to the period from 2010 to 2014. For this purpose, the observed yearly default rates from 2005 to 2013, provided in Table 4.1, were replicated to simulate 1000 samples of five observations each (one for each year of the 6_{th} period), 2010-2014. Then, the simple averages for each one of the 1000 samples of observations, as in Equation 3.1, were computed.

The bootstrap technique was conducted separately using two different probability distributions: the Discrete Uniform Distribution and the Continuous Uniform Distribution. The percentile confidence intervals obtained for the two cases, at a level of confidence of 95%, are the following.

Table 4.3: Confindence Interval Bounds for $LRDR_6$

Distribution	**Lower Bound**	**Upper Bound**
Discrete	1,333945 %	2,186382 %
Continuous	1,552976 %	2,417755 %

Observing the real value for $LRDR_6$, 1,380311%, one can conclude that this value falls inside the confidence interval obtained using the Discrete Uniform Distribution but the same does not happen with the Continuous Uniform Distribution. Given the limitations of the present work and for convenience, the adequacy of the $LRDR_6$ is assumed valid to use it to perform the calibration tests, considering the Discrete Uniform Distribution.

The default behaviour may have changed when compared to the previous five-year periods. This implies that the adjustments that must be applied to each pooled DR may change these in a significant way. Finally, the effects over the scoring PD scale in force my also be significant. To identify and assess the magnitude of such effects, some statistical calibration tests were conducted directly on the scoring scale.

For a rigorous measure of the accuracy of a model it is important to ensure a consistent comparability between estimates and empirical results. This purpose is compromised since the scoring scale is adjusted to the long-run default experience but default rates are only reflecting the yearly experience. To adjust the observed DR for 2014, Equation 3.2 was used. Only after this adjustment it is possible to compare the values for PD estimates and observed DRs and ensure that the assessment of the accuracy is adequate.

Before presenting the results from the calibration tests, it is important to display the information available for the eight risk pools in the residential mortgage loans portfolio under analysis, where the default events are assumed to be mutually independent.

Table 4.4: Default Rates for 2014 after Adjustment to $LRDR_6$

Risk Pool	**PD Estimate**	**No. of operations**	**No. of defaults**	**DR**	**DR_{adjust}**
1	0,01 %	67 389	17	0,03 %	0,02 %
2	0,05 %	73 488	36	0,05 %	0,05 %
3	0,10 %	96 598	150	0,15 %	0,15 %
4	0,35 %	125 168	588	0,47 %	0,45 %
5	1,18 %	90 280	1 405	1,56 %	1,49 %
6	2,52 %	58 554	2 270	3,88 %	3,71 %
7	6,71 %	26 342	2 738	10,39 %	9,97 %
8	15,73 %	10 952	2 204	20,12 %	19,38 %

Table 4.5: Portfolio Averages

Average PD Estimate	1,20%
Average DR	1,71%
Average DR_{adjust}	1,64%

As expected, the DR_{adjust} displays slightly lower values than the yearly DR. This is so because the formula for DR_{adjust} takes into account the $LRDR_6$ which, as one can see in Figure 4.2, shows a decreasing tendency for long-run default rates.

4.2.2 Calibration Tests

After confirming the adequacy of LRDR for the most recent period available and concluding about its validity for the adjustment of the default rates of 2014, it is important to assess the accuracy of each PD associated to each risk pool, i.e., assess whether the magnitude of the differences between the estimated PDs and the observed DRs after adjustment are significant. In this way, the bank will know if there is need to adjust the current distribution of the portfolio and respective scoring scale. In this process, the calibration tests described in Chapter 3 are applied to the scoring scale in Table 4.4.

Briefly comparing columns for PD Estimate and DR it is easy to observe that there are differences in most of the risk pools. For example, in risk pools 7 and 8, the PD estimates may no longer be valid, since the gap between them and the corresponding observed DR_{adjust} appears to be significant. In fact, observing Figure 4.3, is visible that for higher levels of risk the model is underestimating the likelihood of default. Also, comparing the average PD estimate (1,20%) with the average default rate after adjustment to $LRDR_6$ (1,64%) it is possible to conclude that the model is underestimating the risk of default.

The main question that the present empirical analysis of calibration wants to answer is: is this a significant difference? In order to answer it, a range of tests were performed.

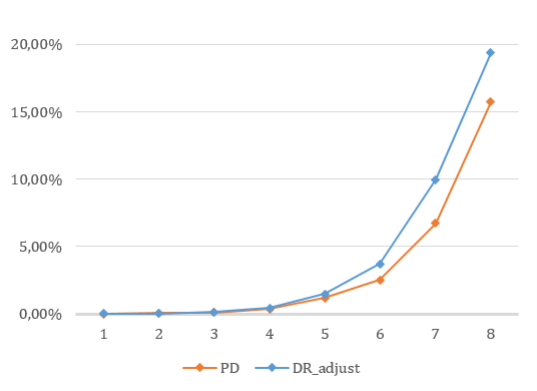

Figure 4.3: PD Estimate vs. DR after adjustment

4.2.2.1 Mean Square Error and *Spiegelhalter* Test

The MSE can be helpful when measuring the distance between estimated PDs and observed DRs. For the model in study a MSE of, approximately, 0,0001 was obtained indicating *a priori*

non significant differences.

Performing the *Spiegelhalter* Test, that has in its basis the MSE, it is possible to assess if the difference provided by the observed MSE is significantly different from the expected. The expected value and variance of MSE, $E[MSE]$ and $Var[MSE]$, were computed.

Table 4.6: Values for *Spiegelhalter* Test

$E[MSE]$	$Var[MSE]$	Z_S
0,011240275	0,008726235	-0,119316366

As $Z_S \sim N(0,1)$, the p-value obtained is ≈0,905 and the null hypothesis, under which the scoring model displays an adjusted granularity and adequate estimation of the risk, considering the eight different levels, is not rejected.

4.2.2.2 Binomial Test

In order to test the null hypothesis "the estimated PD for the risk pool is not underestimated" against the alternative hypothesis "the estimated PD of the rating pool is underestimated", for a significance level of α= 1%, the Binomial Test, as described in Section 3.3.2, was applied.

Table 4.7: Binomial Test Results

Risk Pool	***k*-statistic**	**Decision**
1	0,06 %	Reject H0
2	57,10 %	Do not Reject H0
3	0,00 %	Reject H0
4	0,00 %	Reject H0
5	0,00 %	Reject H0
6	0,00 %	Reject H0
7	0,00 %	Reject H0
8	0,00 %	Reject H0

The Binomial Test only rejects the null hypothesis when the observed default rate exceeds the default probability estimate. The test is performed this way since the regulators are particularly focused on confirming that the capital requirements calculated with these estimates are not underestimating the risk the bank is facing. Observing Table 4.7, the test only fails to reject the PD estimate for the second risk pool, ie, that is the only estimate that might be accurate, given the corresponding observed default rate. For all the other pools the PD estimates are evidenced as underestimating the effective risk. A deeper evaluation of the scoring scale accuracy is necessary.

4.2.2.3 Normal Approximation to the Binomial Test

Given the limitations regarding the binomial distribution, referred in 3.3.4, the assumption that it approaches asymptotically to the normal distribution can be used to construct confidence intervals at 99% for each risk level PD estimate. The results are in the following table.

Table 4.8: Confidence Intervals at 99%

Risk Pool	Lower Bound	Upper Bound	DR_{adjust}	Decision
1	0,00 %	0,02 %	0,02 %	Not Calibrated
2	0,03 %	0,07 %	0,05 %	Calibrated
3	0,08 %	0,12 %	0,15 %	Not Calibrated
4	0,31 %	0,39 %	0,45 %	Not Calibrated
5	1,10 %	1,26 %	1,49 %	Not Calibrated
6	2,37 %	2,67 %	3,71 %	Not Calibrated
7	6,35 %	7,07 %	9,97 %	Not Calibrated
8	14,92 %	16,54 %	19,38 %	Not Calibrated

Observing Table 4.8, there is only one risk pool that falls inside the obtained confidence interval for the probability of default, the second one. All the other pools are underestimating the risk of default, as already concluded from the Binomial Test.

If a closer look is given at the average values, the scoring model is underestimating the default events. Constructing a confidence interval at 99% identically to what was previously done, one can verify that the average observed DR after adjustment falls outside the bounds defined for the average PD, see Table 4.9.

Finally, either observing each risk pool individually or the global scoring scale in average, and similarly to the conclusions reached with the Binomial Test, there is evidence to consider that the scoring scale needs recalibration.

Table 4.9: Confidence Interval at 99% for the Average PD

	Values
Average PD Estimate	1,20%
Average Observed DR after Adjustment	1,64%
Lower Bound	1,17%
Upper Bound	1,24%

4.2.2.4 Hosmer-Lemeshow Test

Contrary to the previous tests, the Hosmer-Lemeshow test the different scoring levels simultaneously. Testing the hypothesis that the PD estimates are correct against the hypothesis that at least one PD estimate is not correct, this test will reject the null hypothesis for a value of HL greater than $\chi^2_{(k-2)}$, where HL is the test statistic and k is the number of risk pools in the model.

Table 4.10: Hosmer-Lemeshow

***t*-statistic**	$\chi^2_{(6)}$
1375,74	16,81

Comparing the two values provided in the table above, the null hypothesis is rejected so at

least one PD estimate is not correctly adjusted.

4.2.3 Conclusions of Empirical Analysis

Regarding to the calibration of the scoring system presented in this chapter, is shown that several statistical tests can be performed in order to assess its accuracy and predictive ability.

Either using Mean Square Error or *Spiegelhalter* Test, a global analysis of the scoring system is performed. In both, the conclusion is the same: the model is estimating the risk adequately and there is no need to recalibrate the scoring scale in force. In other words, the observed default events do not differ significantly from the estimated PDs.

Contrarily to MSE and *Spiegelhalter* Test, the other tests came up with the conclusion that the model is underestimating the risk, either in a individually risk pool analysis or in a global analysis. It can be noted for the tests that allow for a risk pool by risk pool analysis – Binomial Test and Normal Approximation to Binomial Test – that the second risk pool is still calibrated. Although, the relevant conclusion is that the scoring scale requires review because the difference between estimated PDs and observed DRs after the adjustment to long run experience are statistically significant. Remember that the regulator want the rating or scoring systems not to underestimate the risk faced by the bank.

In fact, this outcome could be expected due to the points already raised in Section 4.2.2 and the negative result obtained when subtracting the average PD estimate from the average DR_{adjust}, see Table 4.1, disclosing an underestimation of the risk and consequent need to review the scoring scale.

The review process of the scoring scale might include redistributing the PD estimates in the different risk pools and recomputing the PD for each risk pool, based in the available default history, namely the most recent.

Chapter 5

Conclusions

Basel II and III give strong emphasis to model validation by supervisors and the need for banking industry to comply with their requirements. Therefore, the goal of this project consisted on proposing a statistical validation methodology for the annual regular assessment of the calibration of PD estimates banks must do, aligned with the current regulatory framework. The purpose of this work is to allow banks to evaluate their risk estimates in a more sensitive way in order to subsequently have a more precise allocation of capital. Extensive regulation can be found on this subject although often complex and not all clear for banking industry. Besides Basel Accords and EU Single Rulebook, this project took into consideration the most recent guidelines of the ECB's targeted review of internal models, European Central Bank (2017b), and EBA's PD estimation EBA/CP/2016/21 (2016).

The aim of IRB Approach is to encourage financial institutions to have a better risk management and to ensure risk sensitivity when calculating capital requirements assuring their efficiency and stability. The process of backtesting within the various portfolios is fundamental to ensure the accuracy of the overall credit risk measurement system in credit institutions. From that stems the need of a regular inspection and validation in order to measure the performance and adequacy of internal rating systems and assess whether it is necessary to revise or recalibrate them. The present work focused particularly on the assessment of the predictive ability of a scoring model, even though it would be interesting to assess the discriminatory power and model stability, as well.

Notwithstanding the fact that the presented exercise is model based and technically statistical, qualitative analysis and expert judgement are not expendable in the achievement of an adequate and reliable validation analysis of internal rating systems. In fact, underlying validation, data quality and internal and external standards must also be considered. Also, it is important to understand that the scope only relates to the diagnosis of eventual deficiencies of the developed models. The posterior correction required for the model is not covered in this project.

An available data set containing the yearly default information from 2005 to 2014 for a residential mortgage for individuals portfolio was used to illustrate the process of backtesting and further calibration of an IRB system. Following the existing bibliography and legislation, the first step of the analysis was to check the adequacy of the long-run average default rate for the most recent five-year period available, LRDR_6. A resampling using the available yearly DR

was made through a Bootstrap technique in order to construct percentile confidence intervals and observe if LRDR_6 was still adequate for the credit cycle. Given the constraints of the present work and for convenience to perform the calibration tests, the adequacy was assumed valid.

Before performing the statistical tests, and to ensure the comparability and consistency between estimates and empirical results, an adjustment rate as described in Chapter 3 was applied to the yearly default information. For the performance of the calibration tests, as suggested by Basel Committee on Banking Supervision (BCBS) (2005a) and Englemann and Rauhmeier (2006), the aim was to assess if the difference between estimates and observed values was significant to question the accuracy of the scoring scale. The conclusion was that, in general, the scoring scale in study was not adequate and was not accurately estimating the risk faced by the institution, i.e., the default probabilities predicted by the scoring model didn't match the observed default rates with the degree of accuracy they should. This evaluation derives from both individual (risk pool by risk pool) and global analysis of the outcomes produced by the system. This diagnose could be expected since the result obtained from the difference between the average PD estimate and the average DR_{adjust} disclosed an underestimation of the risk and subsequent need to review the assignment of the PD to the scoring model.

The analytical part of this project was simplified to make calculations easier due to the work constraints, as well as to the confidentiality agreement with the credit institution. Nonetheless, an effort was made to clarify the statistical validation process and the tools and assumptions used for it.

The project doesn't cover the posterior correction of the model, although this might include redistributing the PD estimates in the scoring scale and recomputing the PD for each risk pool considering the most recent default history. Also, and given the limitations of any statistical performing tests, a margin of conservatism and expert judgement by credit risk management units must be considered when validating internal models.

Some relevant topics need further research as the way to deal with insufficient data for model development and validation which is compounded by the short awareness on the importance of data integrity. Also, there is still room for research progress in the validation of LGD and EAD since these represent more complex and less developed topics in this field.

Bibliography

Bandt, O. d. and P. Hartmann (2000). "Systemic Risk: A Survey". *ECB Working Paper Series*.

Basel Committee on Banking Supervision (BCBS) (2005a). "Studies on the validation of internal rating systems. Working paper no. 14". *Basel Committee on Banking Supervision, Basel*.

Basel Committee on Banking Supervision (BCBS) (2005b). "Update on work of the Accord Implementation Group related to validation under the Basel II Framework".

Basel Committee on Banking Supervision (BCBS) (2006). "Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework – Comprehensive Version". *Basel Committee on Banking Supervision, Basel*.

Basel Committee on Banking Supervision (BCBS) (2010). "Basel III: A global regulatory framework for more resilient banks and banking systems". *Basel Committee on Banking Supervision, Basel*.

Bonini, S. and G. Caivano (2014). "Probability of Default: A Modern Calibration Approach". *Mathematical and Statistical Methods for Actuarial Sciences and Finance*. Springer, pp. 41–44.

Castermans, G., D. Martens, T. Van Gestel, B. Hamers, and B. Baesens (2010). "An overview and framework for PD backtesting and benchmarking". *Journal of the Operational research society* 61(3), pp. 359–373.

Deutsche Bank Research (2011). "Solvency II and Basel III".

EBA (2014). "Guidelines on common procedures and methodologies for the Supervisory Review and Evaluation Process (SREP)".

EBA/CP/2016/21 (2016). "Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures". *European Banking Authority*.

EBA/DP/2015/01 (2015). "Discussion Paper: Future of IRB Approach". *European Banking Authority*.

EBA/RTS/2016/03 (2016). "Final Draft Regulatory Technical Standards on the specification of the IRB assessment methodogy". *European Banking Authority*.

Englemann, B. and R. Rauhmeier (2006). *The Basel II risk parameters: Estimation, Validation and Stress Testing*. Springer Science & Business Media.

European Actuarial Consultative Group (2013). "Comparison of the Regulatory Approach in Insurance and Banking in the Context of Solvency II".

European Central Bank. *Glossary of ECB*. `https://www.ecb.europa.eu/home/glossary/html/index.en.html`.

European Central Bank (2017a). "Guidance to banks on non-performing loans".

European Central Bank (2017b). "Guide for the Targeted Review of Internal Models (TRIM)".

European Parliament (2009). "Regulation (EC) No 1060/2009". *Council of 16 September 2009 on credit rating agencies.*

European Parliament (2013a). "Directive 2013/36/EU". *Council of 26 June 2013 on prudential requirements for credit institutions and investment firms.*

European Parliament (2013b). "Regulation (EU) No 575/2013". *Council of 26 June 2013 on prudential requirements for credit institutions and investment firms.*

Gobeljic, P. (2012). "Classification of Probability of Default and Rating Philosophies".

Guillaume Hingel. *Financial Times Lexicon.* `http://lexicon.ft.com/Term?term=risk_weighted-assets`.

IMF, BIS and FSB (2009). "Guidance to Assess the Systemic Importance of Financial Institutions, Markets and Instruments: Inital Considerations".

Löeffler, G. and M. P. N. Posch (2007). *Credit risk modeling using Excel and VBA*. John Wiley & Sons.

Maarse, B. (2012). *Backtesting framework for PD, EAD and LGD.* University of Twente.

McClave, J. T., P. G. Benson, and T. Sincich (2014). *Statistics for business and economics.* Pearson Essex.

Miu, P. and B. Ozdemir (2007). "Estimating and validating long-run probability of default with respect to Basel II requirements".

Ozdemir, B. and P. Miu (2006). *Basel II implementation: a guide to developing and validating a compliant, internal risk rating system.* McGraw Hill Professional.

Shimko, D. and P. Went (2010). *Credit Risk Management.* Global Association of Risk Professionals GARP.

Švec, M. (2012). *PD backtest empirical study on credit retail portfolio.* Hentede.

Tasche, D. (2006). "Validation of internal rating systems and PD estimates". *The analytics of risk model validation* 28, pp. 169–196.

Thomas, L. C., D. B. Edelman, and J. N. Crook (2002). *Credit scoring and its applications.* Siam.

Wu, X. (2008). *Credit Scoring Model Validation.* Master Thesis, University of AMTERDAM, 2009.

Appendix A

Risk-Weigthed Assets for Retail Exposures

A.1 Residential Mortgage Exposures

$$\text{Correlation (R)} = 0.15$$

A.2 Qualifying Revolving Retail Exposures

$$\text{Correlation (R)} = 0.04$$

A.3 Other Retail Exposures

$$\text{Correlation (R)} = 0.03 \times \left[\frac{1-e^{-35\times PD}}{1-e^{-35}} + \frac{1-(1-e^{-35\times PD})}{1-e^{-35}}\right]$$

A.4 Common for all the above

$$\text{Capital requirement (K)} = \text{LGD} \times N\left[\frac{N^{-1}(\text{PD})}{\sqrt{1-\text{R}}} + \sqrt{\frac{\text{R}}{1-\text{R}}}N^{-1}(0.999)\right]$$

$$\text{Risk-weighted Assets (RWA)} = \text{K} \times 12.5 \times \text{EAD}$$

$$\text{Regulatory Capital for Credit Risk} = 8\% \times \text{RWA}$$

Where:

- PD and LGD are measured as decimals and EAD is measured as currency (€);
- N(.) and N^{-1}(.) are, respectively, the Gaussian and the Inverse Gaussian cumulative distribution functions;
- For a defaulted exposure, EL = EL_{BE} and K = $\max\{0, \text{LGD} - \text{EL}_{\text{BE}}\}$, EL_{BE} is the bank's best estimate of Expected Loss.

Source: § 327-330 of Basel Committee on Banking Supervision (BCBS) (2006).