POLITECNICO DI TORINO

Master's Degree Program in Mathematical Engineering

Master's Thesis

Reconstructing Historical Rating Distributions for Long-Run PD Calibration: An Intesa Sanpaolo Case Study

Supervisor
prof.ssa Patrizia Semeraro
Co-Supervisor
Francesco Grande

Candidate
Vittoria Drago

Academic Year 2024-2025

Contents

List of Tables 5

List of Figures 6

1 Credit Risk and the Basel Regulatory Framework 11
- 1.1 Credit Risk: Definition and Challenges in Credit Risk Analytics 11
  - 1.1.1 Definition and Key Concepts 11
  - 1.1.2 From Theory to Practice 13
- 1.2 The Basel Framework 14
- 1.3 Risk Parameters and the IRB Framework 16
- 1.4 Lifecycle of a Credit Exposure 17

2 Risk Differentiation and Quantification in PD Models 19
- 2.1 Development and Calibration of PD Models 19
- 2.2 Risk Differentiation 20
  - 2.2.1 Data Collection and Estimation Sample Construction 20
  - 2.2.2 Risk Driver Selection 21
  - 2.2.3 Model Selection 22
  - 2.2.4 Rating Scale Definition 22
- 2.3 Risk Quantification 23
  - 2.3.1 Risk Calibration 23

3 Qualitative and Quantitative Criteria for Rating System Validation 27
- 3.1 Qualitative Validation Criteria for Rating Systems 27
- 3.2 Quantitative Validation Criteria for Rating Systems 28
  - 3.2.1 Contingency tables 28
  - 3.2.2 Receiver Operating Characteristic (ROC) curve 29
  - 3.2.3 Cumulative Accuracy Profile (CAP) 31

4 Effects of Data Inflation on Accuracy Ratio 35
- 4.1 Inflated Accuracy Ratio Formulations 37

5 Confidence Intervals and Statistical Tests for the Accuracy Ratio 43
- 5.1 Theoretical Background for Confidence Intervals 44
- 5.2 Accuracy Ratio Comparison 47
  - 5.2.1 Kolmogorov–Smirnov test 47
  - 5.2.2 T-test for Accuracy Ratio Comparison 48
  - 5.2.3 Bootstrapping on the Accuracy Ratio (AR) 49
  - 5.2.4 Confidence Intervals for the Accuracy Ratio Using DeLong's Method 49

6 Case Study: Long-Run Calibration with Missing Data 51
6.1 Introduction to the Case Study . . . 51
6.2 Description of Available Data . . . 51
6.3 Identification of Missing Data . . . 53
6.4 Initial Data Analysis . . . 53
6.4.1 Default Rate Trends . . . 53
6.4.2 Portfolio Evolution . . . 55
6.5 Challenges and Methodological Blocks . . . 56

7 Reconstruction of Default Rates 57
7.1 Theoretical Background on Regression Analysis . . . 57
7.1.1 Application in Credit Risk: Modeling Default Rates . . . 58
7.2 Reconstruction of Historical Default Rates . . . 60
7.2.1 Tabular Evidence . . . 60
7.2.2 Graphical Evidence . . . 60

8 Markov Chain Models for Rating Transitions 65
8.1 Theoretical background . . . 65
8.1.1 Introduction . . . 65
8.1.2 Communication as an Equivalence Relation . . . 66
8.1.3 Irreducibility and Periodicity in Markov Chains . . . 67
8.1.4 Recurrence and Transience . . . 68
8.1.5 Forward and Backward Transition Matrices . . . 68
8.2 Estimation Framework . . . 69
8.2.1 Matrix Estimation . . . 69
8.2.2 Analysis of Estimated Forward Matrices . . . 73

9 Reconstruction of Missing Data through Dynamic Equations and Migration Matrices 75
9.1 Estimation of Non-Observed Distributions via Dynamic Equation . . . 76
9.2 Reconstructed Rating Distribution . . . 79
9.2.1 Fully Averaged Inputs Approach . . . 79
9.2.2 Regression-Driven Defaults with Averaged Flows Approach . . . 81
9.2.3 Hybrid Approach . . . 83
9.2.4 Comparative Analysis of Rating Evolution (2008–2014) . . . 85
9.2.5 Implied Default Rate Analysis (2008–2014) . . . 87

10 Long-Run PD Calibration Methodology 89
10.0.1 Linear Scaling . . . 89
10.0.2 Ordinary Least Squares (OLS) Regression . . . 90
10.1 Results of Linear Scaling Calibration . . . 90
10.1.1 Reconstruction of Observations via Migration Matrix and Dynamic Equations . . . 91
10.1.2 Estimation of Default Rates Using Regression . . . 94
10.1.3 Results of Direct Linear Scaling Applied to the 2015–2022 Sample . 95

10.1.4 Comparison of Absolute Differences Between Pre- and Post-Calibration 96

11 Accuracy Ratio after Calibration: Results and Statistical Assessment 99
11.1 Accuracy Ratio Results 100
11.2 AR Analysis for Rating 13 Using Statistical Tools 103
11.2.1 Bootstrapping on the Accuracy Ratio (AR) Confidence Intervals for the Accuracy Ratio Using DeLong's 104
11.2.2 Confidence Intervals for the Accuracy Ratio Using DeLong's 105

List of Tables

3.1 Contingency table of credit risk model classification 28
4.1 Rating Grade Data 40
6.1 Annual Data: Performing, Default, Total (TOT), and Default Rate (DR) . 52
7.1 Default rates (DR) per rating and per year reconstructed using the regression-based approach. 60
7.2 Average reconstructed DRs per rating grade for the periods 2008–2014 and 2015–2022 62
7.3 Average Portfolio DR 62
9.1 Normalized distribution of Ratings for Fully Averaged Inputs Approach (2008–2014) 79
9.2 Normalized distribution of Ratings for Regression-Driven Defaults with Averaged Flows (2008–2014) 81
9.3 Normalized distribution of Ratings for Hybrid Method(2008–2014) 83
10.1 Differences between Pre- and Post-Calibration PDs for Each Approach (%) 96
10.2 ρ of each method 97
11.1 Comparative overview of AR results across all methods for ratings 7–13 . . 100
11.2 Differences between AR_{15-22} and pre-calibration values (and DLS) for ratings 7–13. 101
11.3 Differences between pre and post calibration values for ratings 7–13. 102
11.4 Accuracy Rates (AR) for rating 13 across all methods. 103
11.5 Confidence intervals (CI) for the different models. 104
11.6 Percentage of overlap between model AUCs using DeLong's method. 104
11.7 AUC values with DeLong confidence intervals for different models. 105
11.8 Percentage of overlap between model AUCs using DeLong's method. 105
11.9 Summary of AR/AUC comparisons, indicating the method used (Bootstrapping or DeLong) and whether the confidence intervals overlap (✓) or not (×). 106

List of Figures

2.1 Overview of the full risk parameter estimation process European Banking Authority [2017] . . . 25
3.1 Examples of ROC curve Sironi and Resti [2007] . . . 30
3.2 Examples of CAP curve Sironi and Resti [2007] . . . 32
4.1 $ARinflated(x)$ for $x[0,1]$ . . . 41
4.2 $ARinflated(x)$ for $x \in \mathbb{R}$ . . . 41
4.3 Inflation of Default Counts Only-$ARinflated(x)$ . . . 42
6.1 Default Rate (DR) over the years (2008-2014). . . . 53
6.2 Default Rate (DR) over the years (2015-2022). . . . 53
6.3 Annual distribution of performing and defaulted counterparties in the observed period (2015–2022)." . . . 55
6.4 Annual distribution of defaulted counterparties in the observed period (2015–2022). 55
6.5 Annual distribution of performing counterparties in the observed period (2015–2022). . . . 56
7.1 Reconstructed DR dynamics over time. . . . 61
7.2 Reconstructed DR profiles by rating grade (2008-2014). . . . 63
7.3 Reconstructed DR profiles by rating grade (2015-2022). . . . 63
8.1 Year-over-Year Estimated Forward Migration Matrices . . . 71
8.2 Year-over-Year Estimated Backward Migration Matrices . . . 72
8.3 Year-over-Year Percentage Variation for Percentages of Deteriorations and Improvements . . . 73
8.4 Trend of Stable Counterparties Percentage Over Time. . . . 74
8.5 Stability Probability Trend Over Time . . . 74
9.1 Evolution of Ratings over Years - Fully Averaged Inputs Approach . . . 80
9.2 Evolution of Ratings over Years - Regression-Driven Defaults with Averaged Flows Approach . . . 82
9.3 Evolution of Ratings over Years - Hybrid Method . . . 84
9.4 Evolution of normalized distributions per rating (2008–2014), comparison across the three approaches. . . . 85
9.5 Implied Default Rates per year (2008–2014) computed for Approaches 1, 2, and 3, compared with observed portfolio DRs. . . . 87
10.1 Histogram of default rates per grade with PD curve. . . . 91
10.2 Pre-calibration default rates and post-calibration PDs. . . . 91
10.3 Histogram of default rates per grade with PD curve. . . . 92

10.4 Pre-calibration default rates and post-calibration PDs. 92
10.5 Histogram of default rates per grade with PD curve. 93
10.6 Pre-calibration default rates and post-calibration PDs. 93
10.7 Histogram of default rates per grade with PD curve. 94
10.8 Pre-calibration default rates and post-calibration PDs. 94
10.9 Histogram of default rates per grade with PD curve. 95
10.10Pre-calibration default rates and post-calibration PDs. 95
11.1 Absolute differences between AR(15-22) and Pre-calibration values 101
11.2 Absolute differences between Pre and Post calibration 102

Introduction

Credit risk is a fundamental dimension in banking, representing the possibility that an unexpected deterioration in a counterparty's creditworthiness may lead to losses on the associated exposure Sironi and Resti [2007].
It encompasses both **default risk**, which reflects the actual failure to meet financial obligations, and **migration risk**, which refers to changes in credit quality without formal default. Modern credit risk management goes beyond simple binary outcomes, using probability distributions to capture the full spectrum of potential credit events.

Accurate measurement and management of credit risk are essential not only for the stability of individual banks but also for the resilience of the financial system. Regulatory frameworks, particularly the Basel Accords, provide structured guidance on how banks should quantify and mitigate these risks. Basel I introduced minimum capital requirements based on fixed risk weights; Basel II refined the approach by incorporating the **Internal Ratings-Based (IRB)** methodology; Basel III further strengthened capital quality and introduced additional buffers to enhance systemic resilience Baesens et al. [2016].

Within the IRB framework, the **Probability of Default (PD)** plays a central role. PD models estimate the likelihood that a counterparty will default within a specified horizon, providing the basis for internal ratings, risk-weighted asset calculations, and regulatory capital requirements. PD model calibration ensures that estimated probabilities are aligned with historically observed default frequencies and reflect both portfolio characteristics and economic conditions European Banking Authority [2017], European Central Bank [2025].

A key challenge in credit risk modeling arises when historical data are **incomplete**. Interruptions in data collection, portfolio changes, or missing time series create gaps that complicate both model calibration and validation. Accurately reconstructing these missing distributions is therefore essential to **estimate long-term PD reliably** while **preserving the Accuracy Ratio (AR) before and after calibration**, in line with supervisory expectations.

The central aim of this thesis is thus twofold: on one hand, to estimate the long-term PD for a retail mortgage portfolio, even in the absence of internal rating data for the 2008–2014 period; on the other hand, to ensure that PD calibration does not compromise the model's discriminatory power, as measured by the AR, maintaining consistency between estimated PDs and observed default rates.

To achieve these objectives, the thesis integrates multiple sources of information. On

one hand, **detailed loan-level data** from 2015 to 2022 provide a direct basis for calculating PDs for each rating grade. On the other hand, **aggregated portfolio-level data** from 2008 to 2014, published by the Bank of Italy, provide the historical perspective necessary to reconstruct missing default dynamics. These sources are combined using statistical and dynamic approaches, including regression techniques and **Markov chain models**, to estimate the missing distributions and create a coherent dataset across the entire period.

This reconstruction approach preserves essential data characteristics, such as the **monotonicity of PDs across rating grades** and the **stability of ratings over time**, ensuring that long-term PD estimates are reliable and that the model's discriminatory power, measured by the AR, remains consistent **both pre- and post-calibration**.

The thesis is structured in eleven chapters. The first chapters provide a theoretical and regulatory foundation, introducing credit risk, Basel regulatory frameworks, and the IRB approach. Subsequent chapters focus on PD model development, calibration, and validation. The central chapters present a **case study on long-term calibration with missing data**, exploring reconstruction methods and comparing alternative approaches. The final chapters apply calibration techniques to assess differences between pre- and post-calibration PD values and summarize the results, demonstrating that the proposed methods allow for **estimating long-term PD without distorting the Accuracy Ratio**.

Chapter 1

Credit Risk and the Basel Regulatory Framework

The content of this chapter is based on a review of academic and regulatory literature. In particular, the following texts have been used as primary references:

- Sironi and Resti [2007], *Credit Risk Management*
- Baesens et al. [2016] *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*

1.1 Credit Risk: Definition and Challenges in Credit Risk Analytics

This first chapter aims to introduce the fundamental concepts of credit risk and to frame them within the international regulatory framework established by the Basel Committee. After providing a general definition of credit risk and its main components, the chapter examines how these risks have been addressed through the successive Basel Accords and identifies the key parameters that play a central role in the assessment and management of credit exposures. This theoretical and regulatory foundation serves as the groundwork for the subsequent chapters, which will focus on the development and calibration of Probability of Default models and on the empirical case study conducted on Intesa Sanpaolo.

1.1.1 Definition and Key Concepts

Credit risk can be defined as *"the possibility that an unexpected change in a counterparty's creditworthiness may generate a corresponding unexpected change in the market value of*

the associated credit exposure[1] ". This concept incorporates three fundamental aspects:

- The dual nature of risk: default and credit migration
- The inherently unpredictable character of credit events
- The scope and measurement of credit exposure

Credit risk involves not only the possibility that a borrower may become insolvent but also the potential worsening of their creditworthiness. In this sense, it comprises both **default risk**, referring to actual failure to meet financial obligations, and **migration risk**, which arises from downgrades in credit ratings without a formal default.

Therefore, rather than relying on a simple binary model that distinguishes only between default and non-default, it is more appropriate to model credit risk using probability distributions that represent the entire spectrum of potential outcomes. In this framework, default appears as the tail event, and the overall risk profile is more accurately captured.

Another key consideration is the fact that changes in a counterparty's rating can occur unexpectedly. Although banks attempt to forecast financial deterioration at the time of lending—often incorporating it into interest rates and loan terms—the real risk lies in deviations from these expectations. Unexpected negative developments, even if theoretically foreseeable, may not have been accounted for in advance, and thus represent true credit risk.

The final component concerns the definition of exposure. Credit risk is not limited to traditional on-balance-sheet lending (such as loans or debt securities), but also includes off-balance-sheet items like derivatives, unsettled securities trades, and currency transactions. These exposures may still be subject to counterparty risk even if no loan has formally been extended.

It is worth noting that the original definition references the *market value* of exposures, which introduces two practical challenges:

- Financial institutions often record exposures at historical cost, while proper risk measurement requires assessing them at their economic or market value. Under IFRS 9, assets are measured based on their cash flow characteristics and business model.
- Many credit assets are illiquid and do not trade in active markets, making it necessary to estimate their fair value using internal valuation models.

While the theoretical understanding of credit risk is well-established, practical implementation presents several significant challenges.

[1] *Sironi and Resti [2007]*

1.1.2 From Theory to Practice

Commercial banks are generally large institutions whose core activity is financial intermediation. They obtain funding through a combination of deposit collection, wholesale financing, and equity capital, which they then use primarily for lending—by far the principal source of credit risk.

Their loan portfolios are generally heavily concentrated in areas such as residential mortgages, commercial real estate financing, and lending to small and medium-sized enterprises (SMEs), frequently secured by real assets owned by the borrowers. Exposure to the real estate sector thus plays a significant role in their credit risk profile. Mortgage structures range widely and may include prime and subprime loans, reverse mortgages, home equity lines of credit (HELOCs), interest-only products, and loans with fixed, variable, or hybrid interest rates.

Beyond real estate lending, banks also extend credit through consumer loans—such as car loans, student loans, and credit card debt—as well as corporate loans. While loans to large corporations are still offered, many of these firms also rely on capital markets to meet funding needs, via equity issuance or bond placements.

Credit risk also arises from other exposures, including fixed-income instruments (such as sovereign, corporate, or bank bonds), investments in securitized products, contingent claims like credit lines and guarantees, and derivative instruments—especially over-the-counter (OTC) contracts and credit derivatives.

The global financial crisis (GFC) of 2007–2009 placed credit risk at the center of international concern. In response, regulators implemented sweeping reforms that reshaped the risk landscape. Among the most significant changes were:

- **Basel III framework**: Enhanced regulatory standards were introduced to strengthen both the quality and quantity of bank capital, introduce leverage and liquidity requirements, and provide more robust impact assessments. We will delve deeper into Basel III in a dedicated section.
- **Stress testing**: Regulatory bodies now require financial institutions to perform annual stress tests on their risk models to assess their resilience under adverse macroeconomic conditions.
- **Harmonization of regulations**: Efforts have been made to promote greater consistency in risk management rules across institutions and jurisdictions.
- **Recovery of capital markets**: Particularly within the private securitization sector, which had contracted sharply in the aftermath of the crisis.
- **Increased transparency**: Enhanced reporting standards, such as centralized trade repositories and access to granular, loan-level data, have improved the informational landscape for credit risk analysts.

- **Boosting competition and efficiency**: More accurate credit risk measurement contributes to a healthier banking sector by promoting operational efficiency and fairer competition.

Risk modeling techniques have undergone significant development in the years since the crisis. Early approaches tended to be theoretical and were often based on qualitative/expert-based assumptions that failed to reflect economic fluctuations. Contemporary models are grounded in empirical data, incorporating long-term historical observations, including periods of severe stress like the GFC.

Modern credit risk models account for macroeconomic influences and are capable of tracking the full life cycle of financial products—from origination through repayment, default, or maturity—while dynamically adapting to changing economic conditions. These models also apply sophisticated techniques, including Bayesian inference and nonparametric approaches, which help capture both observable factors and latent risks.

Nevertheless, despite these improvements, it is essential to recognize that no model is without limitations. They are built upon assumptions and past data possibly incomplete and typically explain only a portion of the observed outcomes. This underscores the ongoing need for refinement and innovation in the field. Enhancing the precision and robustness of credit risk models will remain a central challenge for the foreseeable future.

1.2 The Basel Framework

Managing credit risk effectively also requires compliance with a global regulatory framework. Among the most important of these is the Basel Framework, established by the Basel Committee on Banking Supervision and then trasformed into effective regulation for EU institutions by means of the CRR. The Basel framework is the full set of standards for the international banking system. It focus on the requirement for banks to maintain enough capital reserves to meet their obligations and absorb unexpected losses.

Banks collect inflows of funds from a variety of sources and allocate these resources across different types of investments. Among these, lending remains one of the core functions of banking activity.

To ensure financial stability and protect both themselves and their depositors, banks must be adequately protected from the risks they assume—particularly on the asset side of their balance sheet. The failure of a bank due to insolvency is a scenario to be avoided, and the risks embedded in a bank's assets should be counterbalanced by sufficient and appropriate liabilities.

A sound capital structure plays a crucial role in risk mitigation. A bank with solid capital reserves, especially in the form of equity, is better equipped to absorb unexpected losses. Consequently, there must be a clear relationship between the level of risk undertaken and the amount of capital held. This relationship is determined through a two-step

process: first, the bank quantifies its risk exposure using specific risk metrics; then, these values are applied in a regulatory formula that determines the minimum level of capital required to cover such risks.

Regulatory frameworks distinguish among different types of regulatory capital, each with varying levels of quality and loss-absorption capacity:

- **Tier 1 Capital:** Considered the highest-quality capital, it includes instruments such as common equity, preferred shares, and retained earnings.
- **Tier 2 Capital:** Comprises supplementary elements like subordinated debt, revaluation reserves, undisclosed reserves, and general loan-loss provisions. Although useful, it offers lower loss capacity compared to Tier 1.
- **Tier 3 Capital:** Introduced under the Basel II framework to cover market risks, it was composed of short-term subordinated debt. However, Tier 3 capital has been phased out under the more stringent Basel III regulations.

Basel I

The Basel I Accord marked a major step in international banking regulation by introducing a capital adequacy framework focused primarily on credit risk. It established the concept of the **Cooke ratio** which compares a bank's regulatory capital to its risk-weighted assets. Under this framework, banks were required to maintain a minimum capital ratio of 8%, meaning that their available capital had to cover at least 8% of the total risk-weighted asset exposure.

It introduced fixed risk weights dependent on the exposure class. For cash exposures, the risk weight was 0 percent, for mortgages 50 percent, and for other commercial exposures 100 percent.

As an example, consider a mortgage of $100. Applying the risk weight of 50 percent, the risk-weighted assets (RWA) then become $50. This is the risk number we referred to earlier. Since the regulatory minimum capital is 8 percent of the risk-weighted assets our $100 mortgage should be financed by least $4 of equity to cover potential credit losses.

Basel II

It was introduced to address the limitations of the Basel I framework and is based on a three-pillar structure:

- **Pillar 1 – Minimum Capital Requirements:** This pillar establishes capital requirements for three categories of risk:
  - *Credit Risk*, which arises from the possibility of a counterparty failing to meet its obligations.
  - *Market Risk*, which results from adverse movements in market variables such as interest rates, exchange rates, and equity prices.

  - *Operational Risk*, defined as the risk of loss resulting from inadequate or failed internal processes, people, systems, or from external events .

  To model credit risk, Basel II allows for three approaches of increasing complexity: the *Standardised Approach*, the *Foundation Internal Ratings-Based (IRB) Approach*, and the *Advanced IRB Approach*.

- **Pillar 2 – Supervisory Review Process:** This pillar emphasizes the role of supervisory authorities in evaluating banks' internal processes for assessing capital adequacy. A key element is the *Internal Capital Adequacy Assessment Process (ICAAP)*, which ensures that banks identify, measure, and manage all relevant risks beyond the minimum capital requirements.

- **Pillar 3 – Market Discipline:** This pillar promotes transparency by requiring banks to disclose qualitative and quantitative information about their risk exposures and management practices. The objective is to strengthen market discipline by enabling stakeholders to assess a bank's risk profile and governance, thereby supporting confidence and potentially lowering funding costs.

Basel III

Basel III emerged as a regulatory response to the Global Financial Crisis (GFC) to strengthen the banking sector's resilience. It places greater emphasis on tangible equity capital, recognizing it as the most effective buffer against unexpected losses. The framework enhances the quality of capital by eliminating Tier 3 capital—considered insufficiently robust—and reinforcing the role of Tier 1 capital, composed mainly of common equity and retained earnings.

A key innovation is the introduction of a non-risk-based leverage ratio (minimum 3%) to act as a safeguard against model risk, supplementing the traditional risk-weighted approach. This ratio includes off-balance-sheet exposures and derivatives, offering a more comprehensive view of total exposure.

Basel III introduces two new buffers:

- a capital conservation buffer of 2.5% of RWA, to be met with common equity;

- a countercyclical capital buffer, ranging from 0 to 2.5% of RWA, designed to build up capital in good times to be used in downturns.

Additionally, Basel III raises the minimum Tier 1 capital ratio from 4% to 6%, and the Common Equity Tier 1 (CET1) ratio from 2% to 4.5%. For systemically important banks, Basel III requires an extra capital surcharge to mitigate the risks their failure could pose to the broader financial system.

1.3 Risk Parameters and the IRB Framework

Following the introduction of enhanced capital requirements and risk buffers under the Basel framework, it becomes essential to understand how banks assess and quantify credit

risk internally. This leads us to the Internal Ratings-Based (IRB) approach—a more refined, risk-sensitive methodology introduced to enhance credit risk evaluation under Basel regulations.

At the core of the IRB approach are four fundamental risk parameters:

- *Probability of Default (PD):* the likelihood that a borrower will default over a one-year horizon.
- *Loss Given Default (LGD):* the proportion of exposure lost in the event of a default, relative to the outstanding amount.
- *Exposure at Default (EAD):* the total value the bank is exposed to when the borrower defaults.
- *Expected Loss (EL):* the product of the previous three parameters:

$$EL = PD \cdot LGD \cdot EAD$$

The IRB framework includes two variants:

- **Foundation IRB (F-IRB):** Banks estimate the PD internally (approved by supervisors and validated through robust internal procedures), while LGD and EAD values are standardized and set by regulators.
- **Advanced IRB (A-IRB):** Banks can internally estimate all three components—PD, LGD, and EAD—subject to supervisory approval and robust internal validation.

It is important to note that the Foundation IRB is generally not applicable to retail exposures. For such cases, banks must opt for either the Standardised Approach or the Advanced IRB. Once the relevant risk parameters are defined, Basel's regulatory formulas are used to compute the required capital.

1.4 Lifecycle of a Credit Exposure

After discussing the Basel framework and the IRB approach, it is important to understand how credit risk unfolds over time in practice. To this end, we now turn to the lifecycle of a credit exposure, which outlines the sequence of stages in the relationship between a financial intermediary and a borrower—from the initial origination to the eventual closure of the exposure. Each phase carries specific risk implications and requires distinct analytical approaches.

The lifecycle of a credit exposure can be articulated through the following stages:

1. **Origination:** This initial phase encompasses the assessment, approval, and disbursement of credit. At this stage, the contractual terms are defined, the borrower's creditworthiness is evaluated, and any collateral arrangements are established.

2. **Monitoring and Observation:** Once the credit has been granted, the exposure enters a continuous monitoring phase. During this stage, the borrower's behavior is evaluated in terms of payment discipline, financial performance, and compliance with contractual obligations.

3. **Default Risk Assessment:** Throughout the monitoring phase, the lender regularly assesses the likelihood that the borrower may face financial distress potentially leading to a default. This risk is formally quantified through the estimation of the *Probability of Default* (PD).

4. **Default Event:** If the borrower exhibits persistent financial difficulties (e.g., overdue payments exceeding 90 days) or is deemed "unlikely to pay" according to regulatory standards, the exposure is classified as defaulted. This transition represents a critical escalation in risk.

   Article 178 of Capital Requirements Regulation (CRR)- No 575/2013 defines default[2]

5. **Post-Default and Recovery:** Once an exposure is classified as defaulted, the lender initiates recovery processes. These may include collateral repossession, legal actions, debt restructuring, or sale of the exposure. During this stage, the actual loss is registered and used for models estimating the LGD, representing the share of the exposure expected not to be recovered.

6. **Closure or Return to Performing Status:** The credit exposure lifecycle concludes either through full repayment, write-off, recovery, or legal resolution. In some cases, if the borrower's financial condition improves significantly, the exposure may be reclassified as performing.

Having outlined the theoretical and regulatory foundations of credit risk, the next chapter moves to the practical construction of Probability of Default (PD) models, focusing on the processes of risk differentiation and quantification.

---

[2]:
"A default shall be considered to have occurred with regard to a particular obligor when either or both of the following have taken place:
(a) the institution considers that the obligor 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 realising security;
(b) the obligor is more than 90 days past due on any material credit obligation to the institution, the parent undertaking or any of its subsidiaries.
In the case of retail exposures, institutions may apply the definition of default laid down in points (a) and (b) at the level of an individual credit facility rather than in relation to the total obligations of a borrower."

Chapter 2

Risk Differentiation and Quantification in PD Models

The content of this chapter is based on a review of academic and regulatory literature. In particular, the following texts have been used as primary references:

- European Banking Authority [2017] *Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures*
- European Central Bank [2025] *ECB guide to internal models*
- Best practices for PD modelling as observed within ISP activities.

2.1 Development and Calibration of PD Models

The estimation of Probability of Default (PD) and other risk parameters generally follows a two-phase process, which, while conceptually distinct, are closely interlinked. These phases are:

1. **Risk Differentiation (Model Development)**
   This phase focuses on developing models that povides a risk ordering and allows to classify credit exposures into homogeneous risk segments or rating grades, based on their underlying credit quality. The objective is to discriminate between borrowers with different levels of credit risk.

   Internal rating systems or scoring models are typically employed to assign exposures to risk categories, using a combination of quantitative (e.g., financial ratios, repayment history) and qualitative (e.g., industry outlook, management quality) information. The output is a risk ranking or grade that reflects relative riskiness among borrowers.

2. **Risk Quantification (Calibration)**
   Once exposures are adequately differentiated, the next step involves quantifying credit risk through the calibration of key risk parameters. This phase assigns numerical values to each risk grade or segment, translating a label-based relative risk differentiation into quantitative measures.

   Calibration aims to estimate parameters that reflect long-term averages and, where applicable, stressed conditions (e.g., downturn LGD). This step ensures that model outputs are not only discriminatory but also aligned with observed historical loss experience and regulatory requirements.

Both phases rely on the preparation and use of datasets, which may partially overlap. However, it is important to note that the data used for model development and the one used for calibration may differ in terms of scope, sample period, or granularity, depending on the specific objective and requirement of each phase. In this case, a lack of sufficient representativeness is not, by itself, a valid reason to exclude the data from the calculation.

We now focus on the processes of Risk Differentiation and Risk Quantification within the framework of Probability of Default (PD) modeling.

The PD model development comprises five main steps: the first four pertain to Risk Differentiation, while the final step concerns Risk Quantification.

2.2 Risk Differentiation

2.2.1 Data Collection and Estimation Sample Construction

Effective data collection is fundamental to the development of reliable PD models. It involves gathering and aggregating data with the goal of ensuring its quality and consistency, forming the basis for estimation and validation samples.

The data collection process must ensure:

- High-quality and reliable data;
- Homogeneity and representativeness of counterparties in the estimation sample;
- Consistency in the definition of default over time;
- Availability[1] of sufficiently long time series.

Data used in PD model development typically falls into the following categories:

- **Personal / Entity Data:**
  Corporate: sector, legal form, size;
  Retail: age, occupation, residence.

[1] Availability is subject to the type of counterparty

- **Financial / Economic Data:**
  Corporate: capital structure, liquidity, profitability, current account balance;
  Retail: income, debt level, savings, repayment capacity, current account balance.

- **Qualitative Data:**
  Corporate: management quality, group affiliation, strategic policies;
  Retail: employment stability, household composition, significant life events.

- **Behavioral Data:**
  Corporate: payment history, credit utilization, account activity;
  Retail: installment punctuality, transaction patterns, past defaults.

Each type of data may lead to different sample structures and modeling outcomes. Therefore, the model design process should evaluate the relative importance of each data source in assessing the borrower's creditworthiness.

Handling of Missing Data Missing data must be treated carefully, taking into account its nature and associated risk. Appropriate imputation techniques or exclusions are applied depending on the type and the impact of the missing information on model performance and reliability.

2.2.2 Risk Driver Selection

The selection of risk drivers represents a foundational step in the construction of a credit rating model, aiming to reduce reliance on expert judgment by automating the identification of the most predictive variables. This process begins with the construction of a long list of candidate variables, drawn from the dataset underlying each specific model module (a specific component of the rating model that analyzes a defined set of homogeneous variables and produces a partial score representing the risk associated with those characteristics).
Beyond their statistical relevance, risk drivers must also exhibit a clear and economically sound relationship with default risk. In other words, the expected economic sense of each variable should be validated and documented. Ensuring this consistency avoids spurious correlations, strengthens the interpretability of the model, and aligns the statistical framework with established principles of credit risk analysis.

Prior to analysis, the variables undergo appropriate transformations based on their data type:

- **Nominal (non-ordered categorical) variables** should be aggregated through dedicated algorithms that optimize predefined criteria—such as maximizing the Accuracy Ratio (AR), ensuring monotonicity of default rates, and preserving predictive consistency.

- **Continuous variables** are typically transformed within the modeling framework to ensure comparability and standardization.

Subsequently, categorical and discrete variables are encoded into numeric formats by assigning risk-consistent scores, facilitating their inclusion in multivariate statistical procedures.

A univariate analysis is conducted to assess the individual predictive power of each candidate variable with respect to the default event. This step enables the reduction of the long list to a short list of variables with the highest explanatory power, evaluated through metrics such as:

- Direction and strength of the Accuracy Ratio (AR)
- Module-specific statistical indicators.To ensure robustness, the selected variables are grouped into homogeneous informational areas based on their nature and economic meaning.

Within and across these areas, two types of validation tests are performed:

- **Performance stability analysis**, comparing indicators across different time frames
- **Distribution stability**, assessed via statistical tests (e.g PSI).

Finally, a correlation analysis is carried out to eliminate the risk of multicollinearity. Through a structured algorithmic approach, the most representative variables are selected among correlated candidates, both within and across informational areas.

2.2.3 Model Selection

Following the definition of the short list and the standardization of the associated variables, the model specification is finalized. A multivariate procedure is applied to the training sample to identify the optimal combination of predictors. Candidate models are evaluated using both statistical and economic criteria, including:

- In-Sample, Out-of-Sample, Out-of-Time discriminatory power (measured via the Accuracy Ratio), Absence of overfitting
- Significance and interpretability of coefficients
- Economic consistency and robustness

Post-selection, the model undergoes a battery of ex-post validation tests to assess its sensitivity to the selected time series and to ensure long-term stability across economic cycles.

2.2.4 Rating Scale Definition

The final score, resulting from the integration of model modules and post-notch components, is discretized through a clustering process if the discrete PD modelling approach is used. A clustering algorithm such as **Ward's minimum-variance method** is employed to minimize intra-cluster variance and maximize inter-cluster distance. The optimal number of clusters (i.e., rating grades) is determined based on portfolio characteristics, such as:

- Overall population structure
- Default rate distribution and monotonicity
- Class concentration and homogeneity

The final rating scale is designed to optimize key properties such as monotonicity, discriminatory power, granularity, stability, and model usability.

2.3 Risk Quantification

2.3.1 Risk Calibration

The risk calibration phase translates the model's ordinal grading into quantitative measures of default probability. This process begins with the estimation of the portfolio's Long-Run Average Default Rate (LRAvDR), established through a robust framework that ensures statistical and economic representativeness as required by the regulation. The framework includes:

- A minimum default observation period of five years and a period sufficiently long to include max and min of a cycle.
- A macroeconomic validation to ensure that the selected time series adequately represents a balanced mix of stressed and non-stressed conditions.

Once the optimal period of LRAvDR has been determined, the portfolio LRAvDR is also computed at the level of rating grades with dedicated approaches (where needed). The result of this outputs is the **Baseline Probability of Default (PD)** for each grade.

Incorporation of Adjustments and Margins of Conservatism (MoC)

In compliance with regulatory guidance, institutions must identify any deficiencies in the data, methodology, or modeling environment that could introduce bias or increase estimation uncertainty beyond standard statistical error. These deficiencies are classified into two distinct, non-overlapping categories:

- **Category A – Data and Methodology Deficiencies:** Includes incomplete or low-quality historical data, limitations in the model's design, or incorrect implementation of the definition of default. For example, missing observations of risk drivers or inaccurate default identification must be treated under this category.
- **Category B – External or Process-Related Uncertainties:** Encompasses changes in the operational environment that impact the predictive stability of the model. These include shifts in underwriting policies, macroeconomic volatility, or legal/regulatory changes—such as amendments to bankruptcy laws or collection procedures—that may affect default or loss behavior, excluding changes specifically targeting internal model regulation or default definitions.

- **Category C – General Estimation Error:** In addition, institutions should quantify a general estimation error, beyond Categories A and B, and present it transparently in a separate category. This accounts for unavoidable statistical uncertainty inherent in the estimation process, even when no specific deficiencies or external process-related uncertainties are identified.

As a general rule, all identified deficiencies must be addressed through appropriate **quantitative adjustments and/or the application of Margins of Conservatism (MoC)**. These MoCs can be implemented at both the portfolio and rating-grade level, depending on the nature and materiality of the deficiency.

The final **Regulatory PDs** are obtained by summing the Baseline PDs and the relevant MoCs, thereby ensuring compliance with prudential requirements and safeguarding against model risk.

The estimation of risk parameters involves more than just model development and calibration; it is embedded within a comprehensive regulatory and operational framework, as summarized in Figure 2.1. The subsequent stages—independent validation, supervisory approval, system implementation, application of risk parameters, and ongoing review—are critical to ensuring the reliability, compliance, and effective use of the models within the institution. These steps guarantee that the risk parameters are operationally integrated and continuously monitored to reflect evolving portfolio and regulatory conditions.

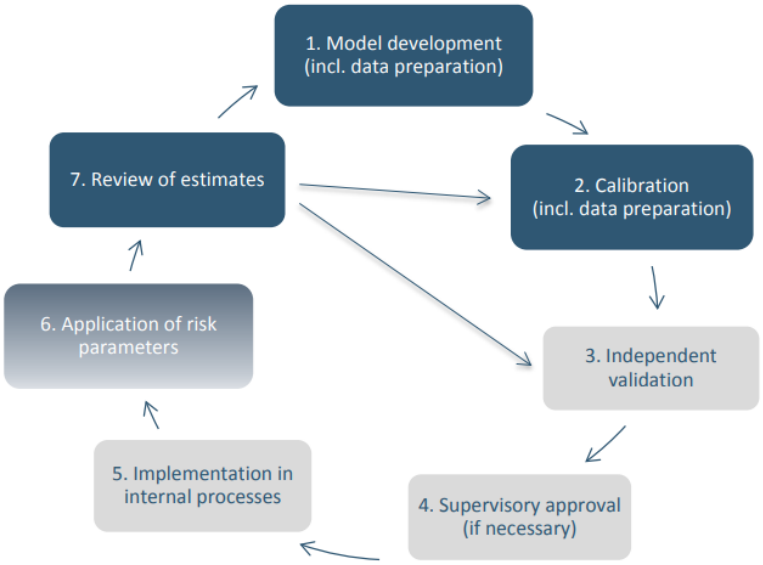

Figure 2.1: Overview of the full risk parameter estimation process European Banking Authority [2017]

The following chapter outlines the validation framework adopted to assess the performance and adequacy of the resulting rating system.

Chapter 3

Qualitative and Quantitative Criteria for Rating System Validation

The content of this chapter is based on a review of academic literature. In particular, the following text has been used as primary reference:

- Sironi and Resti, *Credit Risk Management* Sironi and Resti [2007]

3.1 Qualitative Validation Criteria for Rating Systems

As introduced in the previous chapter, it is essential to assess the adequacy and reliability of a rating system through a structured validation process.

This assessment can be performed by verifying whether the following criteria are satisfied:

- **Monotonicity:** Default rates should increase as the rating worsens, reflecting the expected risk hierarchy across rating grades.
- **Stability of Default Rates:** Default rates within each rating grade should remain stable over time, indicating that the rating system maintains predictive accuracy across different periods.
- **Rating Grade Persistence:** A sufficiently high proportion of exposures should remain in the same rating grade from one year to the next, suggesting stability in borrower credit quality and in the system's assessment (TTCness of the model).

- **Migration Smoothness:** Migration rates toward adjacent rating grade should be more frequent than those toward distant ones, reflecting gradual changes in credit quality rather than abrupt shifts.
- **Early Risk Detection:** Most borrowers who eventually default should have been assigned to high-risk rating grade well before the default event.

In addition to these rules, some quantitative criteria have been proposed to verify the correctness of rating assignments.

3.2 Quantitative Validation Criteria for Rating Systems

3.2.1 Contingency tables

The first method is based on contingency tables: matrices that compare the forecasts of a model with the actual outcomes observed later. Each table is divided into four quadrants, indicating:

- the number **N1** of counterparties correctly rated as "performing" by the model;
- the number **N2** of counterparties incorrectly rated as performing, corresponding to the number of *Type I errors*;
- the number **N3** of counterparties incorrectly rated as high-risk, corresponding to the number of *Type II errors*;
- the number **N4** of counterparties correctly rated as high-risk.

Rating by model	**Performing**	**Defaulting**	**Total**
Low-risk ("pass")	Correct valuation (N_1)	Type II errors (N_3)	$N_1 + N_3$
High-risk ("fail")	Type I errors (N_2)	Correct evaluations (N_4)	$N_2 + N_4$
Total	$N_1 + N_2$	$N_3 + N_4$	$N_1 + N_2 + N_3 + N_4$

Table 3.1: Contingency table of credit risk model classification

Using these values, it is possible to compute several performance indicators. The most important are:

- **Sensitivity:** the percentage of correctly identified defaulting counterparties:

$$\frac{N_4}{N_2 + N_4}$$

- **Specificity:** the percentage of correctly identified performing counterparties:

$$\frac{N_1}{N_1 + N_3}$$

- α **Error rate:** the percentage of defaulting counterparties incorrectly classified as performing:

$$E_\alpha = \frac{N_2}{N_2 + N_4}$$

- β **Error rate:** the percentage of performing counterparties incorrectly classified as defaulting:

$$E_\beta = \frac{N_3}{N_1 + N_3}$$

- **Hit Rate:** the percentage of correctly classified counterparties:

$$\text{HitRate} = \frac{N_1 + N_4}{N_1 + N_2 + N_3 + N_4} = \frac{N_1 + N_4}{N}$$

The quality of a rating model should be assessed through the joint analysis of the Type I (E_α) and Type II (E_β) error rates. These error levels are critically influenced by the choice of the cut-off threshold used to distinguish between performing and defaulting counterparties. In general, a more conservative cut-off value leads to an increase in Type II errors (false negatives) and a reduction in Type I errors (false positives).
Therefore, a robust evaluation of model performance requires examining its sensitivity to variations in the cut-off threshold. This allows the analyst to understand how classification accuracy shifts under different decision rules.

3.2.2 Receiver Operating Characteristic (ROC) curve

Another method for model validation is the Receiver Operating Characteristic (ROC) curve. This graphical representation illustrates the trade-off between the False Alarm and the Sensitivity across all possible cut-off values. Specifically, for each threshold k, a high-performing rating model will exhibit a rapid increase in sensitivity (H_k) with minimal increases in false positives. In other words, as k increases, the model should be able to correctly identify defaulting borrowers (abnormal firms) without mistakenly classifying a significant number of performing borrowers as defaulters.
The figure 3.1 shows an example of ROC curve which express the tradeoff of Type I errors and type II errors. This figure shows two theoretical ROC curves as benchmarks:

- The first one is the curve of a "perfect model", for which a value of k exists that allows 100% (H_k) of defaulted companies to be classified correctly, without making any one mistake ($F_k = 0$)
- The second curve is that of a wholly "naive" model,which lacks any real ability to separate healthy from defaulted companies.

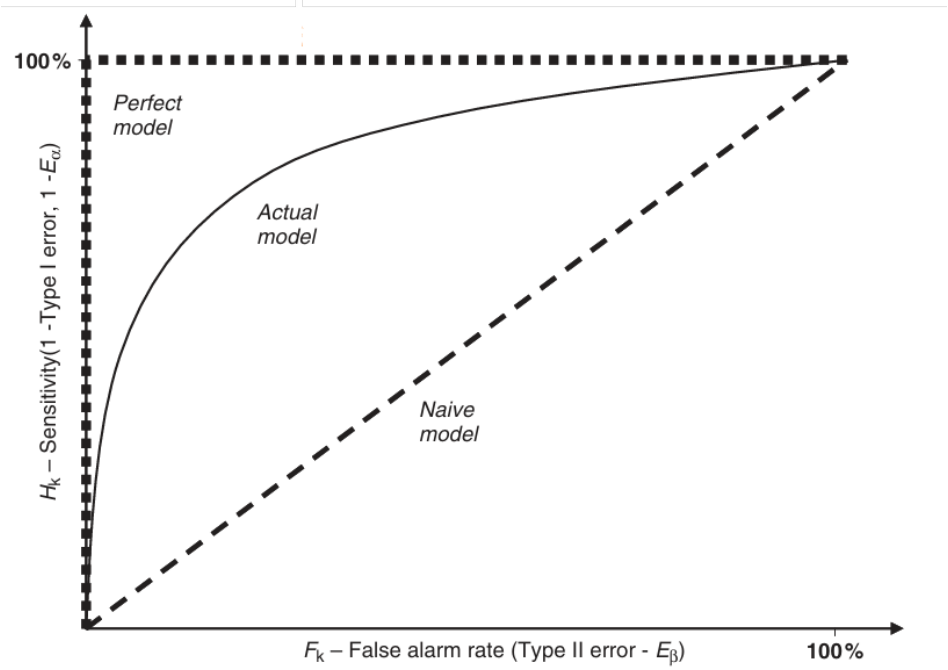

Figure 3.1: Examples of ROC curve Sironi and Resti [2007]

3.2.3 Cumulative Accuracy Profile (CAP)

Another useful way to assess the performance of a rating model is through the Gini curve, also known as the Cumulative Accuracy Profile (CAP).
Let's imagine we have a sample of N companies, each assigned a score by the model. We start by selecting the companies with the worst scores, one by one, and for each step $S = 1, 2, \ldots, N$, we record two things: the number of companies considered so far (on the horizontal axis), and how many among them actually defaulted, which we call $D(S)$ (on the vertical axis).
This process helps us visualize how well the model ranks companies in terms of risk. To interpret the curve, we compare it to three reference cases:

- **An ideal model:** this model perfectly ranks all defaulters at the top. So, for the first $N_2 + N_4$ companies (the actual defaulters), $D(S) = S$. After that, the curve flattens, because there are no more defaults to detect.
- **A naïve model:** this one has no predictive ability. Defaults are randomly distributed, so the number of defaults grows proportionally with S: in other words, $D(S) = p \cdot S$, where $p = \frac{N_2+N_4}{N}$ is the default rate in the sample.
- **A real-world model:** its curve typically falls somewhere between the ideal and naïve cases. The closer it gets to the ideal curve, the more accurate the model is at distinguishing risky companies from safe ones.

A typical CAP curve is therefore the one in 3.2

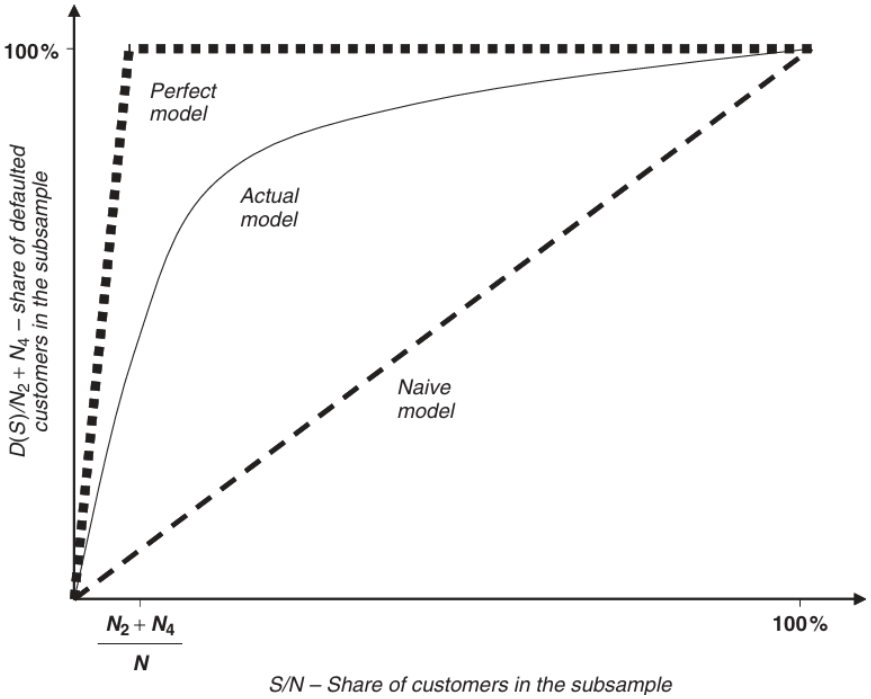

Figure 3.2: Examples of CAP curve Sironi and Resti [2007]

A widely used performance indicator for a rating system is the Gini ratio, also known as the Accuracy Ratio (AR). This measure plays a central role in evaluating the discriminatory power of a model and will be the main focus of the next chapter, where we explore how it is affected by artificial inflation of the underlying numbers.

Chapter 4

Effects of Data Inflation on Accuracy Ratio

In this section, we define the key variables used in the calculation of the Accuracy Ratio (AR) and its inflated versions. The objective is to understand how the artificial increase in the number of defaulted counterparties affects the calculation of the AR. In the context of credit risk, the application of a limitation, an adjustment, or a MoC leads to an artificial increase in the default rate for each rating, which may result in a distortion of the AR compared to its initial value. In the course of this chapter, we will examine under which circumstances this occurs.
The Accuracy Ratio is a metric commonly used to evaluate the performance of classification models, particularly in fields such as credit scoring and fraud detection. It assesses the model's ability to differentiate between distinct classes—such as defaulters and non-defaulters—by comparing its performance to that of a perfect classifier. A higher AR indicates stronger discriminatory power, meaning the model is more effective at correctly separating the two groups.

In order to calculate the Accuracy Ratio (AR) and explore how it may be affected by inflating either defaults or non-defaults, it is essential to first establish a clear set of definitions and notation.
Consider a credit rating system that assigns counterparties to n distinct rating grades.
For each rating grade i, we denote:

- def_i: the number of counterparties that migrated to default within the one-year observation window in rating grade i
- $bonis_i$: the number of non-defaulted (i.e., performing) counterparties in rating grade i

The total number of defaulted counterparties across all grades is denoted by:

$$def = \sum_{i=1}^{n} def_i$$

Similarly, the total number of performing counterparties is:

$$bonis = \sum_{i=1}^{n} bonis_i$$

For each rating grade i, the total number of observed counterparties is:

$$obs_i = def_i + bonis_i$$

and the overall number of observed counterparties is:

$$obs = def + bonis$$

To normalize the distribution of defaults and non-defaults across the rating scale, we define:

$$\pi_i = \frac{def_i}{def} \quad \text{and} \quad \omega_i = \frac{bonis_i}{bonis}$$

where π_i represents the share of defaulted counterparties in grade i, and ω_i the share of performing counterparties in the same grade.

We also introduce the cumulative hit rate, defined as:

$$\text{HitRate}_{i-1} = \mathbf{P}(risk < \text{rating grade } i \mid \text{default})$$

which gives the cumulative proportion of defaults in rating grades worse than i. We assume that the dataframe is sorted such that rating grades are ordered from best to worst credit quality. That is, the default rate (DR) is increasing with increasing rating number:

- Rating 1 has the lowest default rate (best obligors),
- Rating n has the highest default rate (worst obligors).

With this notation, the baseline formula for the Accuracy Ratio (AR) is expressed as:

$$AR = \sum_{i=1}^{n} \pi_i \cdot \omega_i + 2 \sum_{i=1}^{n-1} \omega_i \cdot \text{HitRate}_{i+1} - 1 \tag{4.1}$$

This formula quantifies the discriminatory power of the rating model by comparing the model's ability to distinguish between defaulting and non-defaulting counterparties. A higher AR value indicates a model with greater ability to separate riskier from safer obligors.

4.1 Inflated Accuracy Ratio Formulations

We discuss how the Accuracy Ratio (AR) behaves when the number of defaults is artificially increased while the number of performing counterparties remains unchanged.

Specifically, we examine the scenario in which the default counts in each rating grade are scaled by a common inflation factor of $(1 + x)$, where $x > 0$.

Such an adjustment may arise for several reasons, including data misalignment between the LRAvDR reference period and availability of historical data to reprocess ratings: the calibration target (and period) may not match the calibration sample with computed grades , post-estimation adjustments applied for conservatism or scenario analysis, or regulatory-imposed limitations
The purpose of this analysis is to evaluate the effect of such inflation on the AR calculation, and to understand whether, and to what extent, the model's discriminatory power is impacted by this transformation.
Let us denote the inflated number of defaults in rating grade i as

$$def_i^{(x)} = (1 + x) \cdot def_i,$$

which implies a total inflated number of defaults

$$def^{(x)} = (1 + x) \cdot def$$

.
The proportion of defaults in each grade remains unchanged, since:

$$\pi_i^{(x)} = \frac{def_i^{(x)}}{def^{(x)}} = \frac{(1 + x) \cdot def_i}{(1 + x) \cdot def} = \frac{def_i}{def} = \pi_i$$

Despite the increase in the absolute number of defaults, the relative distribution across rating grades is preserved. Using this, we can write the inflated version of the AR formula as:

$$\begin{aligned} AR_{\text{inflated}} &= \sum_{i=1}^{n} \frac{def_i(1+x)}{def(1+x)} \cdot \frac{bonis_i}{bonis} + 2\sum_{i=1}^{n-1} \frac{bonis_i}{bonis} \left(\sum_{j=1}^{i+1} \frac{def_j(1+x)}{def(1+x)} \right) - 1 \\ &= \sum_{i=1}^{n} \frac{def_i}{def} \cdot \frac{bonis_i}{bonis} + 2\sum_{i=1}^{n-1} \frac{bonis_i}{bonis} \left(\sum_{j=1}^{i+1} \frac{def_j}{def} \right) - 1 \\ &= AR_{\text{baseline}} \end{aligned}$$

This derivation shows that when default counts are uniformly inflated across all rating grades, the Accuracy Ratio remains unchanged. The scaling factor $(1 + x)$ cancels out in both the point-wise default proportions and the cumulative hit rate, leaving the discriminatory power of the model—as measured by the AR—unaffected.

We now consider a more complex scenario in which the number of defaulted counterparties is inflated—just as in the previous case—but, in contrast, the number of performing (non-defaulted) counterparties is adjusted downward to maintain the original total number of observed counterparties in each rating grade.
In this case, the number of non-defaulted counterparties is adjusted as:

$$\text{adjusted\_bonis}_i = obs_i - def_i(1+x)$$

This adjustment ensures that the total number of counterparties in each grade remains unchanged:

$$obs_i = def_i(1+x) + \text{adjusted\_bonis}_i$$

In this framework, the distribution of performing counterparties is no longer preserved, and the proportions ω_i must be recalculated based on the adjusted non-default counts. The advantage is that a constant scaling factor is applied across all rating grades.
The formula for the inflated AR becomes:

$$\begin{aligned}
AR_{\text{inflated}} &= \sum_{i=1}^{n} \frac{def_i(1+x)}{def(1+x)} \cdot \frac{adjusted_bonis_i}{adjusted_bonis} + 2\sum_{i=1}^{n-1} \frac{adjusted_bonis_i}{adjusted_bonis} \left(\sum_{j=1}^{i+1} \frac{def_j(1+x)}{def(1+x)} \right) - 1 \\
&= \sum_{i=1}^{n} \frac{def_i}{def} \cdot \frac{obs_i - def_i(1+x)}{obs - def(1+x)} + 2\sum_{i=1}^{n-1} \frac{obs_i - def_i(1+x)}{obs - def(1+x)} \left(\sum_{j=1}^{i+1} \frac{def_j}{def} \right) - 1 \\
&= \sum_{i=1}^{n} \pi_i \cdot \omega_{i,\text{ inflated}} + 2\sum_{i=1}^{n-1} \omega_{i,\text{ inflated}} \cdot \text{HitRate}_{i+1} - 1 \\
&= \sum_{i=1}^{n} \omega_{i,\text{ inflated}} \cdot (\pi_i + 2 \cdot \text{HitRate}_{i+1}) - 1
\end{aligned}$$

$$\text{where} \quad \text{HitRate}_{n+1} = 0$$

We can define the following two constants, that do not depend on the inflation factor x:

$$\begin{aligned}
C_1 &= \sum_{i=1}^{n} obs_i \cdot (\pi_i + 2 \cdot \text{HitRate}_{i+1}) \\
C_2 &= -\sum_{i=1}^{n} def_i \cdot (\pi_i + 2 \cdot \text{HitRate}_{i+1})
\end{aligned}$$

These constants capture weighted sums over the portfolio, combining the observed counts and default distributions with the cumulative hit rates.
Using these, the Accuracy Ratio after inflating the default counts by $(1+x)$ and adjusting the number of non-defaulted counterparties accordingly can be compactly expressed as:

$$AR_{\text{inflated}} = \frac{C_1 + C_2(1+x)}{obs - def \cdot (1+x)} - 1$$

The two formulations of the Accuracy Ratio (AR) reflect distinct approaches to modeling distortions caused by inflation in observed data:

1. Baseline AR. This formulation uses the observed proportions of defaults (π_i) and non-defaults (ω_i), combined with cumulative Hit Rates, to provide an unbiased measure of the model's discriminatory power.

2. Inflated AR. In the inflated AR, the number of defaults is artificially increased by a factor of $(1 + x)$. Two scenarios arise:

- When only the default counts are inflated while the non-default counts remain unchanged, the overall AR remains unchanged:

$$AR_{\text{inflated}} = AR_{\text{baseline}}.$$

  This occurs because the inflation factor cancels out when computing the relevant proportions.

- When the non-default counts are also recalculated to reflect the inflated defaults, the structure of the AR changes to:

$$AR_{\text{inflated}} = \frac{C_1 + C_2(1 + x)}{obs - def \cdot (1 + x)}.$$

  This rational expression shows that the AR depends explicitly on the inflation parameter x, capturing how inflated observations impact the performance measure.

An interesting result emerging from the analysis is that artificially inflating the number of defaults—while keeping the population of non-defaulters constant—does not change the Accuracy Ratio (AR). This stability is noteworthy in the context of credit risk modeling, as it suggests that the AR remains robust even under stressed default scenarios.
From a practical perspective, this finding implies that certain model adjustments—such as the application of Margin of Conservatism (MoC), overrides, or rating limitations—can be implemented without impacting the AR.
However, when the non-defaults are adjusted accordingly, the AR becomes sensitive to the inflation factor.

Mathematically, the function

$$AR_{\text{inflated}}(x) = \frac{C_1 + C_2(1 + x)}{obs - def \cdot (1 + x)}$$

is a **hyperbola** in terms of the inflation parameter x. This implies a nonlinear and asymptotic relationship between AR and x.

Nevertheless, in practical credit risk applications, inflation factors x are typically small (usually $x < 1$). Within this limited range, the hyperbolic shape is not pronounced, and the AR change function and it may appear approximately linear. This explains why empirical observations often suggest a near-linear relationship despite the underlying nonlinear structure.

This analysis underscores the importance of carefully considering how regulatory constraints, data misalignment, or adjustments may distort the underlying data.
Such distortions can bias the Accuracy Ratio, particularly when the number of non-default counterparties (bonis) is recalculated, as shown. To avoid introducing bias into the AR, it is generally preferable to apply inflation solely to the default counts.

Visual Illustration. To better understand the behavior of the inflated Accuracy Ratio as a function of the inflation parameter x, we provide graphical illustrations. We begin by presenting an illustrative case based on a dataset with 7 rating grades, shown in Table 4.1.

Table 4.1: Rating Grade Data

Rating	obs_i	def_i	DR	$bonis_i$	ω_i	π_i	$HitRate$
1	5000	25	0.51%	4975	6.95%	1.32%	100.0%
2	10000	90	0.91%	9910	13.84%	4.75%	98.7%
3	20000	240	1.21%	19760	27.60%	12.66%	93.9%
4	17000	300	1.78%	16700	23.32%	15.83%	81.3%
5	11000	340	3.12%	10660	14.89%	17.94%	65.4%
6	6000	400	6.73%	5600	7.82%	21.11%	47.5%
7	4500	500	11.22%	4000	5.59%	26.39%	26.4%

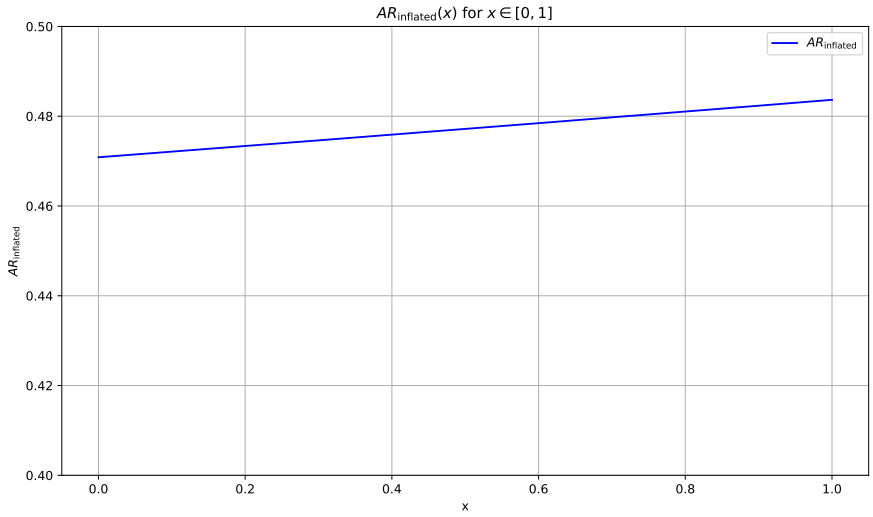

Figure 4.1: $ARinflated(x)$ for $x[0, 1]$

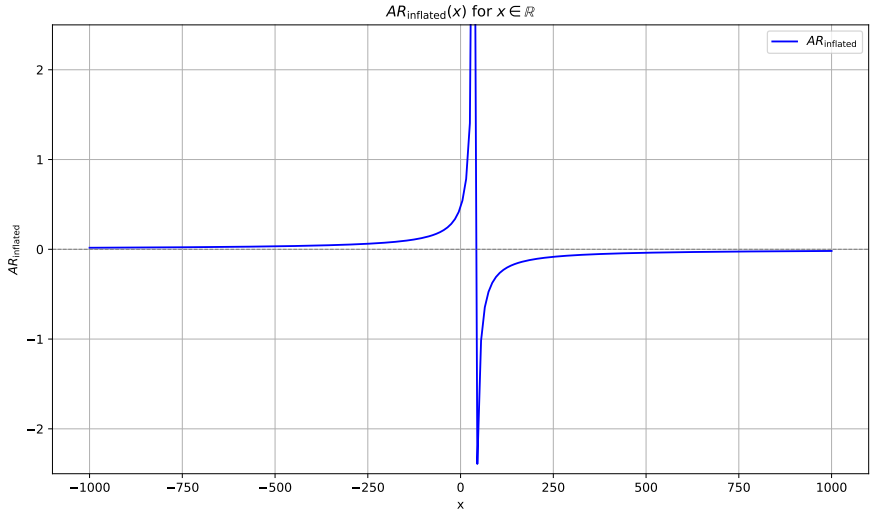

Figure 4.2: $ARinflated(x)$ for $x \in \mathbb{R}$

Figure 4.1 illustrates the plot of $AR_{\text{inflated}}(x)$ for $x \in [0, 1]$, where a near-linear behavior can be observed within this range. In some cases, this approximate linearity may extend to values $x > 1$; however, as x increases further, the curve tends to exhibit a hyperbolic shape, as shown in **Figure 4.2**. In both cases, the performing population has also been updated. By contrast, in **Figure 4.3**, we observe that the accuracy remains constant as x varies, when only the defaulters are inflated while keeping the performing population unchanged.

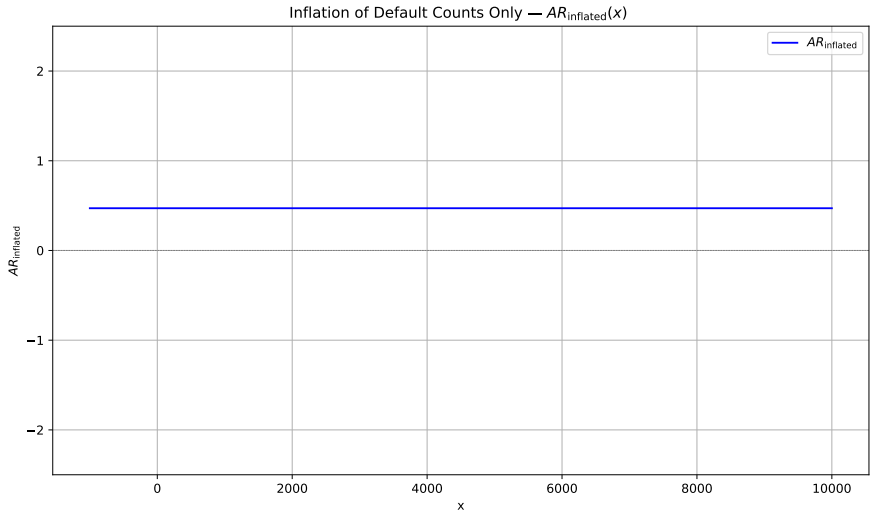

Figure 4.3: Inflation of Default Counts Only-$ARinflated(x)$

Chapter 5

Confidence Intervals and Statistical Tests for the Accuracy Ratio

In line with the European Banking Authority (EBA) guidelines, maintaining the discriminatory power of a rating model represents a fundamental requirement.
This principle is explicitly stated in the latest EGIM guidelines (par. 133(b)), in particular footnote 73, which highlights that:

> "The use of a calibration sample for the purposes of obtaining the scores/raw PDs at grade level or the use of a calibration methodology where the discriminatory power implied by the PDs at grade level is not consistent with the discriminatory power implied by the observed average of the one-year default rates at grade level from the sample corresponding to the period of the LRA default rate may result in PD estimates at grade level which do not reflect the LRA default rate at grade level."

The above extract underlines the importance of ensuring consistency between the observed discriminatory power and the implied default rates at grade level, especially in the context of calibration. Indeed, inappropriate calibration practices may distort the relationship between predicted and observed default frequencies, thus undermining both statistical validity and regulatory acceptability.
In practice, when calibrating a model, the LRA default rate is considered at the segment level, but PDs at grade level may not fully reflect it if the calibration sample is not representative. Bias can occur, for example, if adverse years or riskier grades are over-represented.

A key concept in this context is the **implied discriminatory power**, which measures how well the PDs assigned by the model differentiate between high- and low-risk exposures. Misaligned calibration can weaken this discriminatory power, so ensuring

consistency between the implied and observed discriminatory power is essential for both statistical validity and regulatory compliance.

Against this background, same statistical methods are presented to verify whether the difference between two Accuracy Ratios (e.g., before and after calibration) is statistically significant or not.

5.1 Theoretical Background for Confidence Intervals

This chapter presents the theoretical framework used to derive confidence intervals and variance estimators for the Accuracy Ratio (AR). Given that the AR is a linear transformation of the Area Under the ROC Curve (AUC), much of the statistical theory developed for the AUC can be directly applied to the AR. Specifically, the relationship is given by

$$AR = 2 \cdot AUC - 1, \tag{5.1}$$

which shows that any result concerning the AUC can be straightforwardly translated to the AR.

The AUC represents a measure of the possibility that a randomly selected defaulted counterparty is assigned a worse (i.e., higher-risk) score than a randomly selected non-defaulted counterparty.

The chapter focuses in particular on the estimation of the variance of the AR, which is fundamental for:

- constructing confidence intervals around AR estimates;
- performing hypothesis tests to assess whether differences in AR across models or scenarios are statistically significant;
- evaluating the robustness of model performance under stressed or adjusted conditions.

The main reference for the theoretical treatment is the paper: *"Confidence Intervals for the Area Under the Receiver Operating Characteristic Curve in the Presence of Ignorable Missing Data"*, which offers a detailed non-parametric approach to variance estimation for the AUC.
Thanks to the identity in 5.1 all the variance and confidence interval results derived for the AUC can be applied to the AR through a simple transformation. This connection allows for a consistent and theoretically grounded approach to analyzing model discrimination through AR-based metrics.

Bamber initially proposed a variance estimator of the unbiased estimator of the AUC, defined as:

$$\hat{AUC} = \frac{\sum_{i=1}^{n_Y} \sum_{j=1}^{n_X} \left[\mathbb{I}(y_i > x_j) + \frac{1}{2}\mathbb{I}(y_i = x_j)\right]}{n_Y \cdot n_X}$$

where x_j and y_i are the test scores of the j-th individual in the non-default group and the i-th individual in the default group, respectively, with $j = 1, \ldots, n_X$ and $i = 1, \ldots, n_Y$. Using the equivalence between the Mann–Whitney statistic U and the AUC θ, estimator variance of $\hat{AUC}$ is given by:

$$\begin{aligned} \hat{VAR}(\hat{AUC}) = \frac{1}{4(n_X - 1)(n_Y - 1)} \Big[& p(X \neq Y) + (n_X - 1) b_{XXY} \\ & + (n_Y - 1) b_{YYX} - 4(n_X + n_Y - 1) \left(\hat{\theta} - \tfrac{1}{2}\right)^2 \Big] \end{aligned} \tag{5.2}$$

where:

$$b_{YYX} = \frac{1}{n_X(n_X - 1)n_Y} \sum_{i=1}^{n_Y} \left(u_{\cdot i}\left(u_{\cdot i} - 1\right) + v_{\cdot i}\left(v_{\cdot i} - 1\right) - 2u_{\cdot i}\, v_{\cdot i}\right)$$

$$b_{YYX} = \frac{1}{n_Y(n_Y - 1)n_X} \sum_{j=1}^{n_X} \left(u_{\cdot j}\left(u_{\cdot j} - 1\right) + v_{\cdot j}\left(v_{\cdot j} - 1\right) - 2u_{\cdot j}\, v_{\cdot j}\right)$$

$$v_{\cdot j} = \sum_{i=1}^{n_Y} \left[\mathbb{I}(y_i > x_j) + \frac{1}{2}\mathbb{I}(y_i = x_j)\right], \quad u_{\cdot j} = n_Y - v_{\cdot j}$$

$$v_{i\cdot} = \sum_{j=1}^{n_X} \left[\mathbb{I}(y_i > x_j) + \frac{1}{2}\mathbb{I}(y_i = x_j)\right], \quad u_{i\cdot} = n_X - v_{i\cdot}$$

Here, X_1, X_2, Y_1, Y_2 are randomly sampled independently without replacement from X and Y, respectively.
The quantities b_{XXY} and b_{YYX} are unbiased estimators of:

$$B_{XXY} = P(X_1, X_2 < Y) + P(Y < X_1, X_2) - P(X_1 < Y < X_2) - P(X_2 < Y < X_1)$$

$$B_{YYX} = P(Y_1, Y_2 < X) + P(X < Y_1, Y_2) - P(Y_1 < X < Y_2) - P(Y_2 < X < Y_1)$$

Another shortcoming of the variance formula in Hanley & McNeil (1982) is that it assumes the underlying score or biomarker is sufficiently continuous, meaning the scores assigned to defaulted and performing counterparties do not have any ties—that is, no defaulted and performing counterparties share the same score.
Then, the revised variance estimator is given by:

$$\hat{VAR}(\hat{AUC}) = \frac{1}{(n_X - 1)(n_Y - 1)}\theta(1-\theta) - \frac{1}{4}p(Y = X) + (n_Y - 1)(Q_1 - \theta^2) + \frac{n_X - 1}{(n_X n_Y)^2}(Q_2 - \theta^2)$$

where:

$$Q_1 = \frac{\sum_{j=1}^{n_X} \left[\sum_{i=1}^{n_Y} \left(\mathbb{I}(y_i > x_j) + \frac{1}{2}\mathbb{I}(y_i = x_j) \right) \right]^2}{(n_X n_Y)^2}$$

is an estimator of:

$$Q_1 = P(Y_1, Y_2 > X) + \frac{1}{2}P(Y_1 > Y_2 = X) + \frac{1}{2}P(Y_2 > Y_1 = X) + \frac{1}{4}P(Y_1 = Y_2 = X)$$

which represents the probability that the score of two randomly chosen defaulted counterparties (possibly the same) is greater than or equal to the score of a randomly chosen performing counterparty, with tied scores counted as half.

Similarly:

$$Q_2 = \sum_{i=1}^{n_Y} \sum_{j=1}^{n_X} \left(\mathbb{I}(y_i > x_j) + \frac{1}{2}\mathbb{I}(y_i = x_j) \right)^2$$

is an estimator of:

$$Q_2 = P(Y > X_1, X_2) + \frac{1}{2}P(Y = X_1 > X_2) + \frac{1}{2}P(Y = X_2 > X_1) + \frac{1}{4}P(Y = X_1 = X_2)$$

Hanley & McNeil (1982) further simplified the variance estimator under the assumption that X and Y are exponentially distributed. In this case, the variance becomes:

$$\begin{aligned} \hat{VAR}(\hat{AUC}) = \frac{1}{(n_X - 1)(n_Y - 1)} \Big[& \hat{AUC}(1 - \hat{AUC}) \; - \tfrac{1}{4}p(Y = X) \\ & + \frac{n_Y - 1}{(n_X n_Y)^2} \left(\hat{AUC}(2 - \hat{AUC}) - \hat{AUC}^2 \right) + \frac{n_X - 1}{(n_X n_Y)^2} \left(\frac{2\hat{AUC}^2}{1 + \hat{AUC}} - \hat{AUC}^2 \right) \Big] \end{aligned}$$

where we have:

$$\hat{Q}_1 = \frac{\hat{AUC}}{(2 - \hat{AUC})}, \quad \hat{Q}_2 = \frac{2\hat{AUC}^2}{1 + \hat{AUC}}$$

Considering the relationship between Accuracy Ratio (AR) and AUC, we can derive the standard deviation of AR, denoted as σ_{AR}, using the properties of the standard deviation. Specifically, we have:

$$\sigma_{\hat{AR}} = 2 \cdot \sigma_{\hat{AUC}} \qquad \hat{VAR}(\hat{AR}) = 4 \cdot \hat{VAR}(\hat{AUC})$$

5.2 Accuracy Ratio Comparison

The theoretical framework for the standard deviation of the Accuracy Ratio, developed in the previous sections, is used to implement statistical tests aimed at assessing whether two ARs differ significantly. This estimate of variability is crucial for conducting hypothesis tests and constructing confidence intervals, allowing for a rigorous evaluation of model performance.
To this end, we apply the following statistical methods:

- **Kolmogorov–Smirnov test**, which compares the empirical distributions of the two samples;
- **Two-sample t-test**, which tests for differences between the means of the two AR distributions.

In addition to these tests, we employ a **bootstrap** procedure to approximate the sampling distribution of the Accuracy Ratio and to construct non-parametric confidence intervals. Furthermore, we build **confidence intervals** based on the variance estimated using *DeLong's method*, offering a robust parametric approach.

5.2.1 Kolmogorov–Smirnov test

The Kolmogorov–Smirnov (KS) test is a non-parametric test used to determine whether two samples come from the same continuous distribution. It compares the empirical cumulative distribution functions (ECDFs) of two samples and quantifies the maximum absolute difference between them.
Given two empirical distributions $F_n(x)$ and $G_m(x)$, corresponding to two samples of sizes n and m, the two-sample KS test statistic is defined as:

$$D_{n,m} = \sup_x |F_n(x) - G_m(x)|$$

Large values of $D_{n,m}$ indicate a significant difference between the two distributions, leading to the rejection of the null hypothesis. Conversely, small values of $D_{n,m}$ suggest no significant difference. A small p-value (typically $p < 0.05$) suggests rejecting the null hypothesis that the two samples are drawn from the same distribution.
In our context, the KS test is used to evaluate the robustness of the Accuracy Ratio (AR) under default inflation scenarios. Specifically, we simulate the effect of increasing or decreasing the number of defaults and observe how this inflation impacts the distribution of AR.

Given the two levels of Accuracy Ratio—baseline and inflated—we model each as a normal distribution using the respective means and standard deviations. These distributions are then compared using the two-sample Kolmogorov–Smirnov (KS) test, which assesses whether the difference in AR under stress is statistically significant with respect to the baseline.

If the p-value is greater than 0.05, we fail to reject the null hypothesis, meaning the AR distribution under inflation is statistically indistinguishable from the baseline.

Assuming a normal distribution often results in a test that is not very sensitive to small changes in the Accuracy Ratio (AR). This is especially true in credit risk contexts, where we observe an inflation factor $x < 1$.
Because the AR (which corresponds to the mean of the normal distribution) changes only slightly, and the standard deviation changes accordingly, the test tends to treat the two AR values as essentially the same, reducing its ability to detect meaningful differences.
To overcome this limitation, one can generate random samples without assuming any specific distribution shape, while still preserving the calculated mean and standard deviation.

5.2.2 T-test for Accuracy Ratio Comparison

The t-test is a parametric statistical test used to assess whether the means of two independent samples differ significantly. It assumes that the samples are approximately normally distributed and that their variances are equal. The independent two-sample t-test evaluates the null hypothesis:

$$H_0 : \mu_1 = \mu_2,$$

where μ_1 and μ_2 are the population means of the two groups.
The test statistic is calculated as:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\text{SE}}, \quad \text{with} \quad \text{SE} = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)},$$

where s_p^2 is the pooled sample variance, given by:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

and s_1^2, s_2^2 are the sample variances of the two groups. The resulting t-statistic is compared to the t-distribution with $n_1 + n_2 - 2$ degrees of freedom to compute the p-value.
In our framework, the t-test is used to evaluate whether the baseline Accuracy Ratio (AR) differs significantly from an AR calculated under a default-inflation scenario.
If the p-value is greater than 0.05, we fail to reject the null hypothesis, suggesting that the inflated AR is not significantly different from the original.

Interpretation of KS Test and t-Test in Credit Risk Modeling Application

Both the t-test and the KS test can be used to explore the robustness of the Accuracy Ratio. Specifically, they allow us to:

- Identify the largest inflation factor $x_{\max}$ such that the AR under inflation is not significantly different from the baseline AR.

- Test whether two AR values—such as those observed before and after a model calibration—are statistically distinguishable.

It is important to note that while the t-test focuses on comparing the means of two distributions, the KS test evaluates their entire cumulative distributions. Hence, using both tests in tandem can provide complementary insights into AR stability under different types of perturbations.

5.2.3 Bootstrapping on the Accuracy Ratio (AR)

Bootstrapping is a non-parametric technique that allows for the estimation of the variability of the Accuracy Ratio (AR) by repeatedly simulating samples obtained through resampling with replacement from the original dataset.
In practice, an empirical distribution of the AR is constructed by recalculating it hundreds or thousands of times on different samples, all randomly drawn (with replacement) from the set of evaluated borrowers. From this distribution, it is possible to derive:

- A confidence interval for the AR (e.g., at the 95% level);
- An assessment of the statistical significance of any differences between two models: if the confidence intervals of the two ARs do not overlap, one can conclude that the difference is statistically significant.

5.2.4 Confidence Intervals for the Accuracy Ratio Using DeLong's Method

DeLong's test is a non-parametric statistical method used to compare the AUCs of two correlated ROC curves. It computes the asymptotic variance of the Area Under the Curve (AUC) based on U-statistics, which rely on the differences in predicted scores between defaulters and non-defaulters. This variance can be directly applied to derive a confidence interval for the AUC, and by extension, for the Accuracy Ratio (AR).
Using this estimate, a confidence interval for the AR can be constructed under the assumption of asymptotic normality:

$$CI_{AR}^{(1-\alpha)} = \Big[AR - z_{\alpha/2} \cdot \sigma_{AR}, \quad AR + z_{\alpha/2} \cdot \sigma_{AR}\Big]$$

where:

- AR is the observed Accuracy Ratio;
- σ_{AR} is the standard deviation estimated via DeLong's method;
- $z_{\alpha/2}$ is the critical value from the standard normal distribution corresponding to the desired confidence level.

This approach offers a theoretically grounded alternative to bootstrapping and is particularly advantageous from a computational standpoint, as it is significantly faster.

Both *bootstrapping* and *De Longe* techniques are highly sensitive to even small variations in the Accuracy Ratio (AR). For this reason, it is crucial to ensure that the AR—and, consequently, the model's discriminatory power—is preserved after any post-model adjustments, the application of Margin of Conservatism (MoC), or other overlays.
In addition,these methods were employed to calculate the percentage of overlap between the confidence intervals of the AR estimates.

Chapter 6

Case Study: Long-Run Calibration with Missing Data

6.1 Introduction to the Case Study

In this chapter, we present a case study on the **estimation of long-run probability of default for a retail mortgage portfolio at Intesa Sanpaolo**. The primary challenge lies in the lack of rating data for the period 2008–2014, which hinders the direct computation of long-run PDs by rating grade.

Moreover, the data used in this study do not reflect the true figures of the Intesa Sanpaolo retail portfolio, and the number of ratings does not correspond to the number of ratings actually assigned by the bank. Nevertheless, the results have been tested and validated, and they remain fully applicable when using the actual portfolio data.

6.2 Description of Available Data

The dataset comprises two distinct blocks:

- **2015–2022:** granular loan-level data, containing the following fields:
  - **Counterparty ID:** an alphanumeric code uniquely identifying each counterparty;
  - **Assigned Internal Rating:** a risk grade assigned to the counterparty based on the internal PD model, ranging from 1 to k, where k is the maximum number of internal rating grades;
  - **Reference Year:** the year corresponding to the observation date of the data;
  - **Default Flag:** a binary indicator equal to 1 if the counterparty is classified as in default (based on the applicable default definition), and 0 otherwise.
- **2008–2014:** only aggregated data are available:

- Number of performing and defaulted counterparties
- Overall portfolio default rates

Note that the default rates refer to data published by the Bank of Italy.

Year	**Performing**	**Default**	**TOT**	**DR**
2008	944,256	18,151	962,407	1.89%
2009	940,110	20,548	960,658	2.14%
2010	995,197	17,747	1,012,944	1.75%
2011	1,282,060	18,705	1,300,765	1.44%
2012	1,118,398	19,330	1,137,728	1.70%
2013	946,238	15,836	962,074	1.65%
2014	954,676	14,233	968,909	1.47%
2015	653,253	9,699	662,952	1.46%
2016	696,368	8,672	705,040	1.23%
2017	937,292	9,879	947,171	1.04%
2018	1,076,078	9,948	1,086,026	0.92%
2019	1,080,503	9,285	1,089,788	0.85%
2020	729,915	6,510	736,425	0.88%
2021	408,367	3,717	412,084	0.90%
2022	667,848	4,011	671,859	0.60%

Table 6.1: Annual Data: Performing, Default, Total (TOT), and Default Rate (DR)

6.3 Identification of Missing Data

The absence internal detailed data for rating computation between 2008 and 2014 implies that:

- Yearly default rates by rating grade are unavailable
- The distribution of counterparties across rating grades (rating mix) is missing

These elements are essential for the calibration of the model, as they provide the necessary inputs to accurately estimate the risk parameters, in particular the probability of default (PD) by rating grade. We will therefore focus on reconstructing these quantities and subsequently assess the discriminatory power of the resulting model.

6.4 Initial Data Analysis

6.4.1 Default Rate Trends

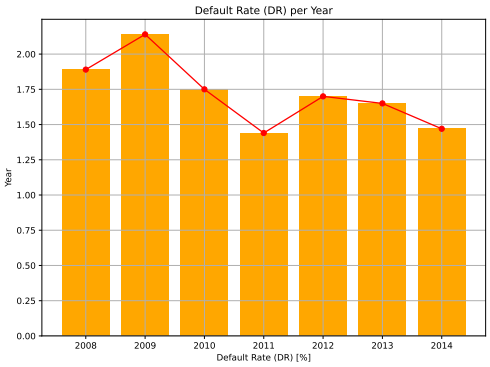

Figure 6.1: Default Rate (DR) over the years (2008-2014).

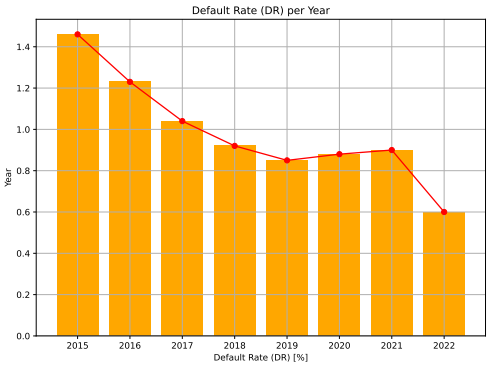

Figure 6.2: Default Rate (DR) over the years (2015-2022).

The figures 6.1 and 6.2 reveals a predominantly downward trend, which can be divided into three distinct phases.

In the aftermath of the global financial crisis, default rates were relatively high, reaching 1.89% in 2008 and peaking at 2.14% in 2009. During the subsequent period (2010–2014), a gradual reduction took place, with a trough of 1.44% in 2011, followed by a modest rebound to 1.70% in 2012, and eventual stabilization around 1.5–1.6%.

From 2015 onwards, the portfolio experienced a marked improvement, with default rates declining steadily from 1.46% in 2015 to 1.04% in 2017. This downward trajectory continued in 2018–2019, when values fell further to 0.92% and 0.85% respectively, representing the lowest levels observed up to that point.

Contrary to initial expectations, the outbreak of the COVID-19 pandemic in 2020 did not lead to a significant deterioration in credit quality. Default rates remained contained at 0.88% in 2020 and 0.90% in 2021.
This resilience can plausibly be attributed to extensive government support measures, which effectively mitigated the immediate impact of the crisis on borrower performance.
Finally, in 2022 the default rate declined further to 0.60%, the lowest value in the entire sample. This result confirms the strengthening of the portfolio's creditworthiness and highlights the structural improvements achieved over the last decade.

6.4.2 Portfolio Evolution

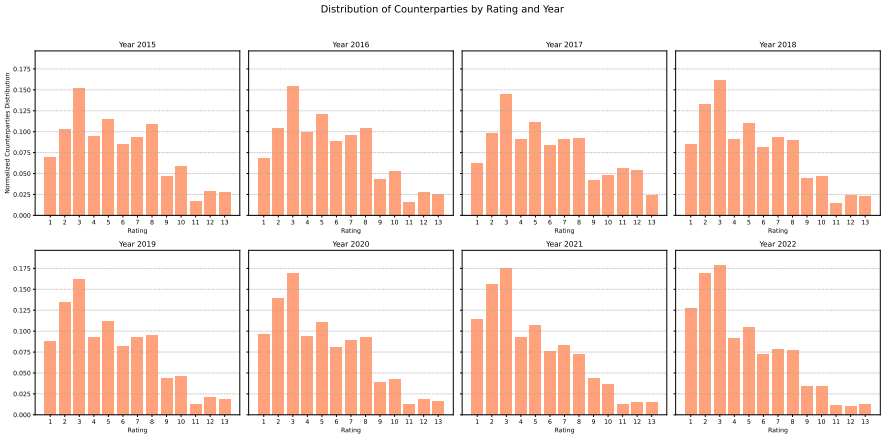

Figure 6.3: Annual distribution of performing and defaulted counterparties in the observed period (2015–2022).”

Figure 6.3 reports the overall distribution of counterparties across rating grades in the observed period (2015–2022). The portfolio composition is dominated by the central rating grade (approximately between 3 and 8), while the extreme rating, both very high quality (1–2) and very low quality (12–13), remain relatively underpopulated.

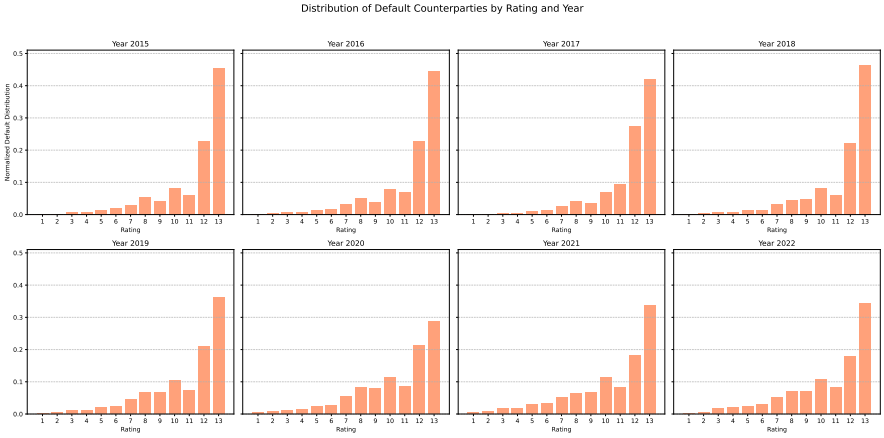

Figure 6.4: Annual distribution of defaulted counterparties in the observed period (2015–2022).

Figure 6.4 illustrates the distribution of *defaulted* counterparties. Here the picture changes substantially: defaults are concentrated almost entirely in the lowest rating grades, especially between 11 and 13. This concentration confirms the discriminative power of the rating system, as weaker grades capture the majority of credit events.

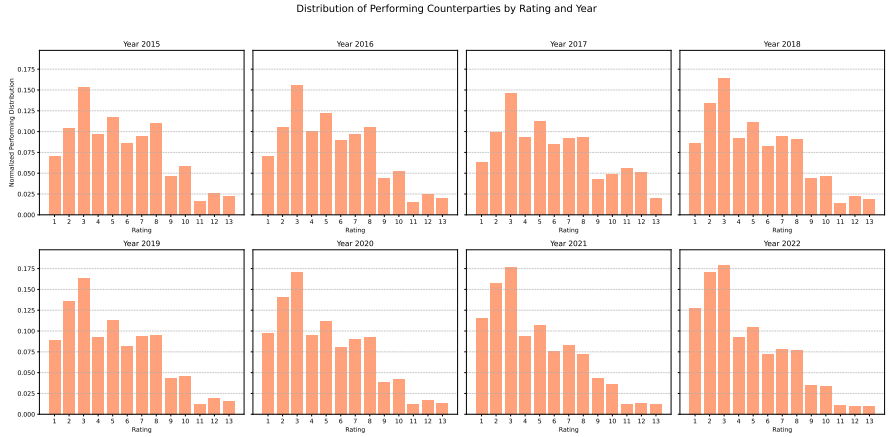

Figure 6.5: Annual distribution of performing counterparties in the observed period (2015–2022).

Figure 6.5 shows the distribution of *performing* counterparties. As expected, the pattern largely mirrors the overall distribution, since the majority of the portfolio is composed of performing exposures. However, the right-hand tail (grades 11–13) is thinner compared to the overall distribution. This is consistent with the fact that exposures in the weakest rating grades are more likely to migrate towards default and therefore are under-represented among performing counterparties.

In summary, the joint analysis of the three distributions highlights a consistent and intuitive pattern: the bulk of the portfolio lies in the intermediate rating grades and the default events concentrate in the weakest ratings.

6.5 Challenges and Methodological Blocks

Based on the available data, we identify four key methodological components:

- **Default Rate Reconstruction (Chapter 7)**: Estimation of rating-level distributions and default rates for the period 2008–2014.
- **Migration Matrices (Chapter 8)**: Use of migration matrices to model rating transitions.
- **Dynamic Equation (Chapter 9)**: Application of a dynamic equation to estimate rating observations for the missing years.
- **Long-Run PD Calibration (Chapter 10)**: Calibration of point-in-time probability of default (PD) estimates to long-term default rates, and assessment of model performance using the Accuracy Ratio (AR).

Each of these steps is detailed in the following chapters.

Chapter 7

Reconstruction of Default Rates

7.1 Theoretical Background on Regression Analysis

Regression analysis is a fundamental statistical tool used to investigate and quantify the relationship between a dependent variable and one or more independent variables. In its simplest form, the method assumes a linear relationship between variables, enabling the prediction of outcomes based on explanatory factors.
In general, regression models take the form:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

where Y is the dependent variable, $X_1, \ldots, X_k$ are the independent variables, $\beta_0, \ldots, \beta_k$ are the model parameters, and ε is the random error term.
In the case of linear regression with one explanatory variable, the parameters β_0 (intercept) and β_1 (slope) can be estimated using the Ordinary Least Squares **(OLS)** method. The formulas are:

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

These estimators minimize the sum of squared residuals:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Once the parameters are estimated, we assess whether the observed relationships are statistically significant. This is typically done through **hypothesis testing** on the regression coefficients.

For the slope β_1, we test the null hypothesis:

$$H_0 : \beta_1 = 0$$

against the alternative:

$$H_1 : \beta_1 \neq 0$$

To do this, we compute the t-statistic:

$$t = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)}$$

where $\text{SE}(\hat{\beta}_1)$ is the standard error of the slope estimate.
The **p-value** associated with this test indicates the probability of observing such an extreme value of the test statistic under the null hypothesis. A small p-value (commonly < 0.05) suggests that the null hypothesis can be rejected, implying that the slope β_1 is statistically significant.

- If p-value < 0.05, the effect is considered statistically significant.
- If p-value ≥ 0.05, there is insufficient evidence to conclude that the predictor has a real effect on the outcome.

This evaluation is crucial for interpreting the regression results and for determining whether the model captures meaningful relationships in the data.

7.1.1 Application in Credit Risk: Modeling Default Rates

In credit risk modeling, regression can be employed to estimate the default rates of individual rating classes during periods where such granular data are missing. However, since default rates are bounded between 0 and 1, applying traditional linear models directly to them can lead to biased or ill-behaved predictions. To address this issue, a transformation such as the log-odds (logit) function is applied to map the default rates onto the entire real line, making linear modeling more appropriate.

The logit function transforms a probability $p \in (0, 1)$ into a value on the real line:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

To overcome the missing information in the years between 2008 and 2014, we estimate default rates (DR) for each rating grade using a regression-based methodology. This approach leverages the available data from 2015 to 2022 to infer class-level default dynamics and reconstruct missing default rates per grade The use of regression is motivated by the fact that default rates display a predominantly decreasing trend over time, making it possible to capture and extrapolate the underlying structural dynamics rather than relying on static averages.

The regression-based methodology is therefore adopted as the reference framework for the backward reconstruction of default rates, ensuring consistency across observed and non-observed periods.
The workflow proceeds as follows:

- Annual portfolio-level default rates from 2008 to 2022 are converted into quarterly estimates via linear interpolation.
  Given the annual default rates $DR_{\text{PTF},Y}$ for year Y, the quarterly rates $DR_{\text{PTF},Q}$ for each quarter $Q \in \{1, 2, 3, 4\}$ are obtained by:

$$DR_{\text{PTF},Q}(Y) = DR_{\text{PTF},Y} + \frac{Q-1}{4} \cdot (DR_{\text{PTF},Y+1} - DR_{\text{PTF},Y}) \quad \text{for } Y = 2008, \ldots, 2021$$

  For the last year:

$$DR_{\text{PTF},Q}(2022) = DR_{\text{PTF},2022} \quad \forall Q$$

- Similarly, annual default rates by rating grade $DR_{i,Y}$ (when available) for the period 2015–2022 are also interpolated into quarterly values using the same linear rule:

$$DR_{i,Q}(Y) = DR_{i,Y} + \frac{Q-1}{4} \cdot (DR_{i,Y+1} - DR_{i,Y}) \quad \text{for } Y = 2015, \ldots, 2021$$

  And for 2022:

$$DR_{i,Q}(2022) = DR_{i,2022} \quad \forall Q$$

  This results in a consistent quarterly panel for both the portfolio and each rating grade, enabling regression and historical reconstruction:

$$DR_{\text{PTF},Q}(\text{2008–2022}), \quad DR_{i,Q}(\text{2015–2022})$$

- For each rating grade i, the following regression is run over the period 2015–2022:

$$\log\left(\frac{1 - DR_{i,Q}}{DR_{i,Q}}\right) = \alpha_i + \beta_i \log\left(\frac{1 - DR_{\text{PTF},Q}}{DR_{\text{PTF},Q}}\right) + \varepsilon_{i,Q}$$

  where α_i and β_i are parameters to be estimated and $\varepsilon_{i,Q}$ is the residual term.

- Using the estimated coefficients $\hat{\alpha}_i$ and $\hat{\beta}_i$, the predicted default rates for each rating grade i in the missing period 2008–2014 are computed by applying the regression to the quarterly portfolio default rates of that period:

$$\widehat{DR}_{i,Q}(\text{2008–2014}) = \frac{1}{1 + \exp\left(-\hat{\alpha}_i - \hat{\beta}_i \log\left(\frac{1 - DR_{\text{PTF},Q}(\text{2008–2014})}{DR_{\text{PTF},Q}(\text{2008–2014})}\right)\right)}$$

- In cases where β_i is not statistically significant, the predicted default rate for class i is obtained by scaling the portfolio default rate for 2008–2014 using the ratio observed in 2015:

$$\widehat{DR}_{i,Q}(\text{2008–2014}) = DR_{\text{PTF},Q}(\text{2008–2014}) \times \frac{DR_i(2015)}{DR_{\text{PTF}}(2015)}$$

7.2 Reconstruction of Historical Default Rates

In this section, we present the reconstructed series of default rates (DR) obtained using a regression-based methodology. The main objective is to fill the missing information for the period 2008–2014 and to construct a consistent dataset suitable for subsequent analyses. The results are presented both in tabular and graphical form to highlight different aspects of the reconstruction.

7.2.1 Tabular Evidence

DR	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
1	0,03%	0,03%	0,02%	0,02%	0,02%	0,02%	0,02%	0,02%	0,03%	0,01%	0,01%	0,02%	0,04%	0,05%	0,02%
2	0,05%	0,06%	0,05%	0,04%	0,05%	0,05%	0,04%	0,04%	0,05%	0,02%	0,03%	0,02%	0,05%	0,07%	0,03%
3	0,10%	0,11%	0,09%	0,07%	0,09%	0,09%	0,08%	0,08%	0,06%	0,04%	0,04%	0,06%	0,08%	0,13%	0,07%
4	0,17%	0,19%	0,16%	0,13%	0,15%	0,15%	0,13%	0,13%	0,12%	0,07%	0,07%	0,11%	0,16%	0,22%	0,14%
5	0,26%	0,29%	0,24%	0,19%	0,23%	0,22%	0,20%	0,20%	0,15%	0,10%	0,11%	0,15%	0,21%	0,32%	0,15%
6	0,44%	0,50%	0,41%	0,34%	0,40%	0,38%	0,34%	0,34%	0,24%	0,19%	0,16%	0,23%	0,32%	0,42%	0,24%
7	0,61%	0,69%	0,57%	0,47%	0,55%	0,53%	0,48%	0,47%	0,42%	0,31%	0,31%	0,41%	0,48%	0,51%	0,34%
8	0,70%	0,73%	0,68%	0,63%	0,67%	0,66%	0,64%	0,67%	0,58%	0,46%	0,44%	0,56%	0,63%	0,62%	0,41%
9	1,13%	1,19%	1,10%	1,03%	1,09%	1,08%	1,04%	1,06%	0,92%	0,78%	0,87%	1,11%	1,01%	0,76%	0,63%
10	1,57%	1,64%	1,53%	1,44%	1,52%	1,50%	1,45%	1,43%	1,34%	1,17%	1,38%	1,64%	1,34%	0,90%	0,97%
11	4,50%	5,11%	4,18%	3,43%	4,06%	3,93%	3,50%	3,49%	3,83%	1,38%	3,23%	4,18%	3,51%	1,88%	2,28%
12	8,78%	9,40%	8,42%	7,55%	8,28%	8,14%	7,64%	8,15%	7,42%	4,18%	7,30%	7,55%	5,71%	3,79%	5,38%
13	22,11%	24,36%	20,86%	17,79%	20,36%	19,85%	18,11%	17,25%	16,12%	14,94%	16,38%	14,30%	9,37%	7,46%	9,54%

Table 7.1: Default rates (DR) per rating and per year reconstructed using the regression-based approach.

The table confirms that the reconstructed values not only fill the historical gaps but also preserve the relative differences across rating grades. In particular, the higher DRs observed in 2008–2010 reflect the financial crisis, while the subsequent decline mirrors the gradual normalization of credit conditions. Another key feature is the stability of distances between rating classes, supporting the credibility of the reconstruction.

7.2.2 Graphical Evidence

Figure 7.1 provides a synthetic overview of the time series. The reconstructed DRs exhibit a generally decreasing trend over time. During the missing period (2008–2014), DRs are consistently above the levels recorded after 2015, reflecting a structurally higher risk environment. The reconstruction captures the decreasing component of credit risk, with higher levels during stressed years and lower values in normal periods.

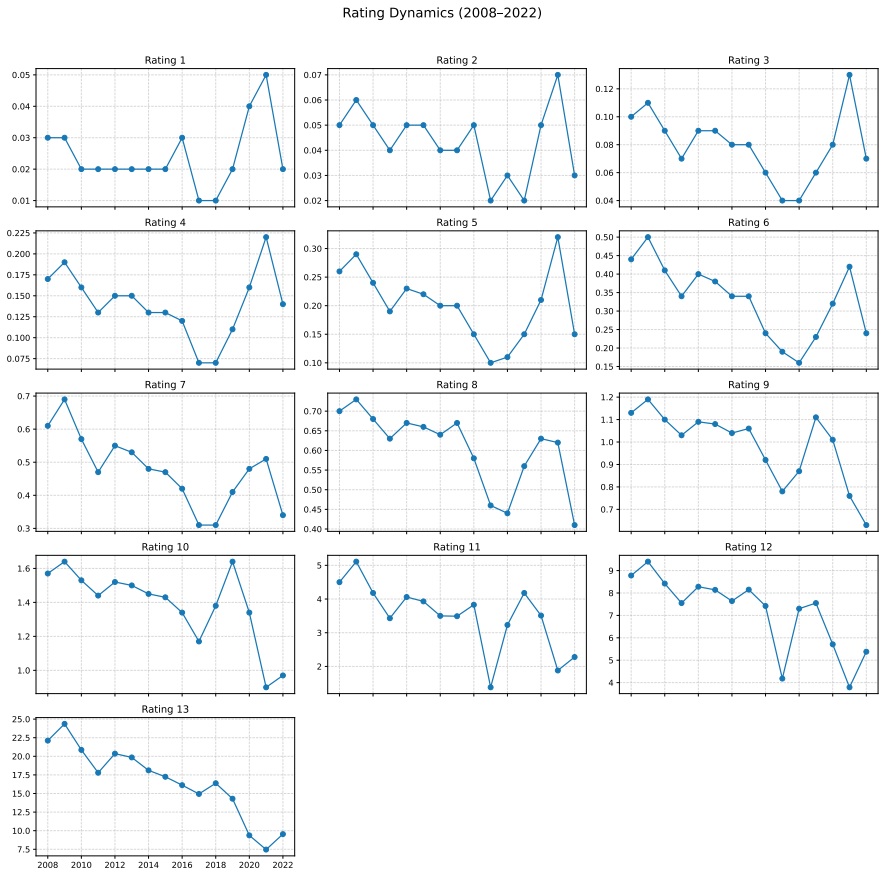

Figure 7.1: Reconstructed DR dynamics over time.

The reconstructed series also allows for an assessment of average DRs across two distinct periods: 2008–2014 and 2015–2022. Table 7.2 summarizes the mean DR per rating grade for these periods, together with the corresponding portfolio-level default rates 7.3. The results show that average DRs in the earlier period were systematically higher, reflecting a riskier credit environment, whereas the later period shows a general stabilization of credit risk, consistent with post-crisis normalization.

Rating	Avg DR 2008–2014	Avg DR 2015–2022
1	0,02%	0,02%
2	0,05%	0,04%
3	0,09%	0,07%
4	0,15%	0,13%
5	0,23%	0,17%
6	0,40%	0,27%
7	0,56%	0,41%
8	0,67%	0,55%
9	1,10%	0,89%
10	1,52%	1,27%
11	4,10%	2,97%
12	8,32%	6,19%
13	20,49%	13,17 %

Table 7.2: Average reconstructed DRs per rating grade for the periods 2008–2014 and 2015–2022

	Average Portfolio DR 2008–2014	Average Portfolio DR 2015–2022
Portfolio	1,72%	0,99%

Table 7.3: Average Portfolio DR

Finally, Figures 7.2 and 7.3 illustrates the class-level DRs. The regression-based method preserves monotonicity across ratings, a crucial property for model consistency. Beyond monotonicity, the figure highlights how different grades respond differently to changes in portfolio-level risk, capturing the heterogeneity of credit quality. This confirms that the reconstructed dataset provides a reliable foundation for further validation exercises and risk quantification.

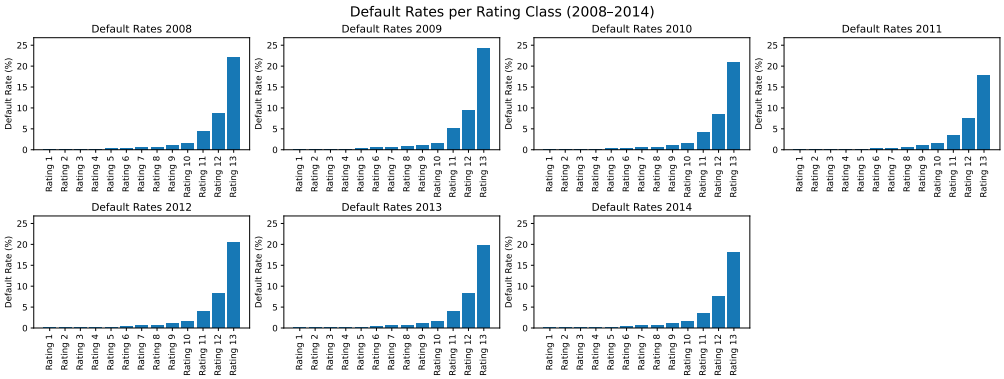

Figure 7.2: Reconstructed DR profiles by rating grade (2008-2014).

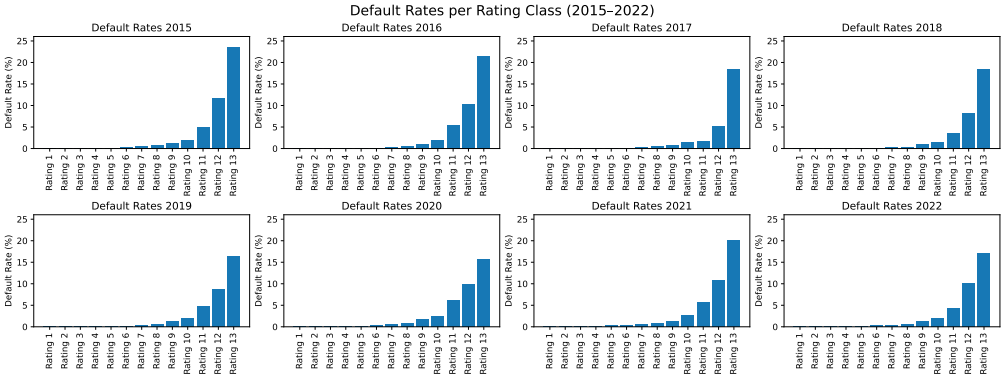

Figure 7.3: Reconstructed DR profiles by rating grade (2015-2022).

Chapter 8

Markov Chain Models for Rating Transitions

The content of this chapter is based on a review of academic literature. In particular, the following text has been used as primary reference:

- Sheldon M.Ross [1995]

In this chapter, we introduce both the theoretical background and the estimation methods for Markov chains, as well as the analysis of the migration matrices, which are fundamental for reconstructing the distribution of counterparties. These concepts and techniques will be applied and further explored in the following chapters.

8.1 Theoretical background

8.1.1 Introduction

Let $\{X_n\}_{n\geq 0}$ be a discrete-time stochastic process (Markov chain) taking values in the state space $\mathcal{S} = \{0, 1, 2, \dots\}$, which may be finite or countably infinite. The defining property of a Markov chain is that the conditional distribution of the next state depends only on the current state, not on the full history. This is known as the **Markov property**, formally:

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j \mid X_n = i) = p_{ij}$$

for all states $i_0, i_1, \dots, i_{n-1}, i, j \in \mathcal{S}$ and for all $n \geq 0$. We define the one-step transition probabilities as:

$$p_{ij} = \mathbb{P}(X_{t+1} = j \mid X_t = i)$$

and organize them into a matrix $P = [p_{ij}]$, known as the **transition probability matrix**:

$$P = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Each row of the matrix sums to 1, since the process must transition to some state:

$$p_{ij} \geq 0 \quad \text{and} \quad \sum_{j=0}^{\infty} p_{ij} = 1 \quad \text{for all } i \in \mathcal{S}$$

We define the n-**step transition probabilities** as:

$$p_{ij}^{(n)} = \mathbb{P}(X_{t+n} = j \mid X_t = i)$$

These describe the probability that the process, starting in state i, is in state j after n steps.
Let $P^{(n)} = [p_{ij}^{(n)}]$ be the matrix of n-step transition probabilities.
The **Chapman-Kolmogorov equations** state that:

$$P^{(n+m)} = P^{(n)} \cdot P^{(m)} \quad \text{for all } n, m \geq 0$$

This recursive relation implies that we can compute the n-step transition matrix as the n-th power of the one-step transition matrix:

$$P^{(n)} = P \cdot P^{(n-1)} = P^n$$

Hence, to obtain the transition probabilities after n steps, it suffices to raise the one-step transition matrix P to the power n using matrix multiplication.

To describe the relationships between states in a Markov chain, we consider the following fundamental concepts:

- A state j is said to be *accessible* from state i if there exists an integer $n \geq 0$ such that $p_{ij}^{(n)} > 0$.
- Two states i and j are said to *communicate* if each is accessible from the other. In that case, we write $i \leftrightarrow j$.

8.1.2 Communication as an Equivalence Relation

Communication between states in a Markov chain is an *equivalence relation*. That is, the relation satisfies the following properties:

1. **Reflexivity:** Every state communicates with itself:

$$i \leftrightarrow i$$

2. **Symmetry:** If state i communicates with state j, then state j communicates with state i:

$$i \leftrightarrow j \quad \Rightarrow \quad j \leftrightarrow i$$

3. **Transitivity:** If state i communicates with state j, and state j communicates with state k, then state i communicates with state k:

$$i \leftrightarrow j \text{ and } j \leftrightarrow k \quad \Rightarrow \quad i \leftrightarrow k$$

Therefore, the set of states of a Markov chain can be partitioned into *communication classes*, i.e., subsets of states where each pair of states communicates with each other.

8.1.3 Irreducibility and Periodicity in Markov Chains

A Markov chain is said to be **irreducible** if all states belong to a single communication class. That is, every state is accessible from every other state:

$$\forall i, j \quad \exists\, n \geq 0 \quad \text{such that} \quad P_{ij}^{(n)} > 0$$

In this case, we say that the chain is *irreducible*, since no subset of states is isolated from the rest.
Let state i be a state in a Markov chain. The **period** $d(i)$ of state i is defined as:

$$d(i) = \gcd\left\{n \geq 1 \ : \ P_{ii}^{(n)} > 0\right\}$$

In words, $d(i)$ is the greatest common divisor of the set of time steps n at which it is possible to return to state i.

- If $d(i) = 1$, then state i is said to be **aperiodic**.
- If $P_{ii}^{(n)} = 0$ for all $n > 0$, then the period is defined to be infinite: $d(i) = \infty$.

Periodicity is a *class property*. That is, if states i and j communicate (i.e., $i \leftrightarrow j$), then:

$$d(i) = d(j)$$

This means that all states in the same communication class share the same period.

8.1.4 Recurrence and Transience

For any two states i and j, let $f_{ij}^{(n)}$ denote the probability that, starting from state i, the first transition into state j occurs exactly at time n. Formally, we define:

$$f_{ij}^{(n)} = \mathbb{P}(X_n = j, X_1 \neq j, X_2 \neq j, \ldots, X_{n-1} \neq j \mid X_0 = i)$$

Let:

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}$$

Then f_{ij} is the probability that, starting in state i, the process will eventually reach state j at some time $n \geq 1$.

- If $f_{ij} > 0$, then state j is said to be **accessible** from state i.
- A state j is said to be **recurrent** (or persistent) if:

  $$f_{jj} = 1$$

  That is, starting from j, the process is guaranteed to return to j eventually.
- Otherwise, if $f_{jj} < 1$, then state j is said to be **transient**.
- A state j is said to be **recurrent** if the expected number of visits to state j, starting from j, is infinite. Formally:

  $$\text{State } j \text{ is recurrent} \iff \mathbb{E}[\text{Number of visits to } j \mid X_0 = j] = \infty$$

This last result has two important implications:

1. A transient state is visited only a finite number of times.
2. In a **finite-state** Markov chain, not all states can be transient.

This conclusion leads to the following fundamental result:

If state i communicates with state j, and i is recurrent, then j is also recurrent. That is, recurrence is a class property.

8.1.5 Forward and Backward Transition Matrices

It is especially important for our application n credit risk modeling to distinguish between two types of transition matrices: **forward** and **backward** transition matrices.
The **forward** transition matrix describes the probability of moving *from* a given rating at time t *to* another rating at time $t + 1$. It is defined as:

$$P^{\text{fwd}} = [\mathbb{P}(X_{t+1} = j \mid X_t = i)]_{i,j}$$

This is the standard Markov transition matrix introduced earlier. Each row represents the conditional distribution of future states given the current state.
The **backward** transition matrix captures the probability that a rating state at time $t+1$ originated from a specific state at time t. Formally, it is defined as:

$$P^{\text{bwd}} = [\mathbb{P}(X_t = i \mid X_{t+1} = j)]_{i,j}$$

Using Bayes' theorem, it can be computed from the forward transition matrix and the marginal distribution of ratings at time t as follows:

$$\mathbb{P}(X_t = i \mid X_{t+1} = j) = \frac{\mathbb{P}(X_{t+1} = j \mid X_t = i) \cdot \mathbb{P}(X_t = i)}{\mathbb{P}(X_{t+1} = j)}$$

Interpretation and Use Cases

- **Forward matrices** (P^{fwd}) are commonly used for simulating future rating migrations and computing multi-period default probabilities. However, they suffer from a limitation: their inverses generally do not represent valid probability matrices, as they may violate probability constraints (e.g., negative entries or rows not summing to one). In particular, to estimate the distribution π_t at time t from a known distribution π_{t+1} at time $t+1$, one would need:

  $$\pi_t = \left(P^{\text{fwd}}\right)^{-1} \pi_{t+1}$$

  which is not guaranteed to yield a valid probability vector.

- **Backward matrices** (P^{bwd}) are particularly relevant for inferring historical rating dynamics and offer a solution to this issue. They allow backward propagation without needing to invert the forward matrix:

  $$\pi_t = P^{\text{bwd}}\, \pi_{t+1}$$

In our application, estimating the distribution of performing and defaulted counterparties at an earlier time requires going backward.
If a forward matrix is used, one would need to invert it, which is problematic due to lack of probabilistic consistency. By contrast, using the backward matrix allows us to compute directly π_t

8.2 Estimation Framework

8.2.1 Matrix Estimation

The transition probabilities were estimated using the Maximum Likelihood Estimation (MLE) method.
For the forward transition matrix, the probability of moving from rating i to rating j is estimated as:

$$\hat{p}_{ij}^{\text{fwd}} = \frac{N_{ij}}{N_i}$$

where:

- N_{ij} is the number of observed transitions from rating i to rating j.
- N_i is the total number of observed transitions originating from rating i.

This represents the probability that an obligor currently in rating i migrates to rating j after one period.

For the backward transition matrix, which models the probability of a previous rating given the current rating, the probability of having been in rating i at time t given rating j at time $t+1$ is estimated as:

$$\hat{p}_{ij}^{\text{bwd}} = \frac{N_{ij}}{N_j^{(+)}}$$

where:

- N_{ij} is the same count of observed transitions from rating i to rating j.
- $N_j^{(+)}$ is the total number of observed transitions arriving at rating j.

This represents the probability that an obligor observed in rating j at time $t+1$ came from rating i at time t.
The resulting one-year forward and backward transition matrices, estimated using the Maximum Likelihood method described above, are reported for each year from 2015 to 2022. These matrices summarize the observed annual rating migrations and serve as the empirical foundation for the subsequent dynamic analysis.

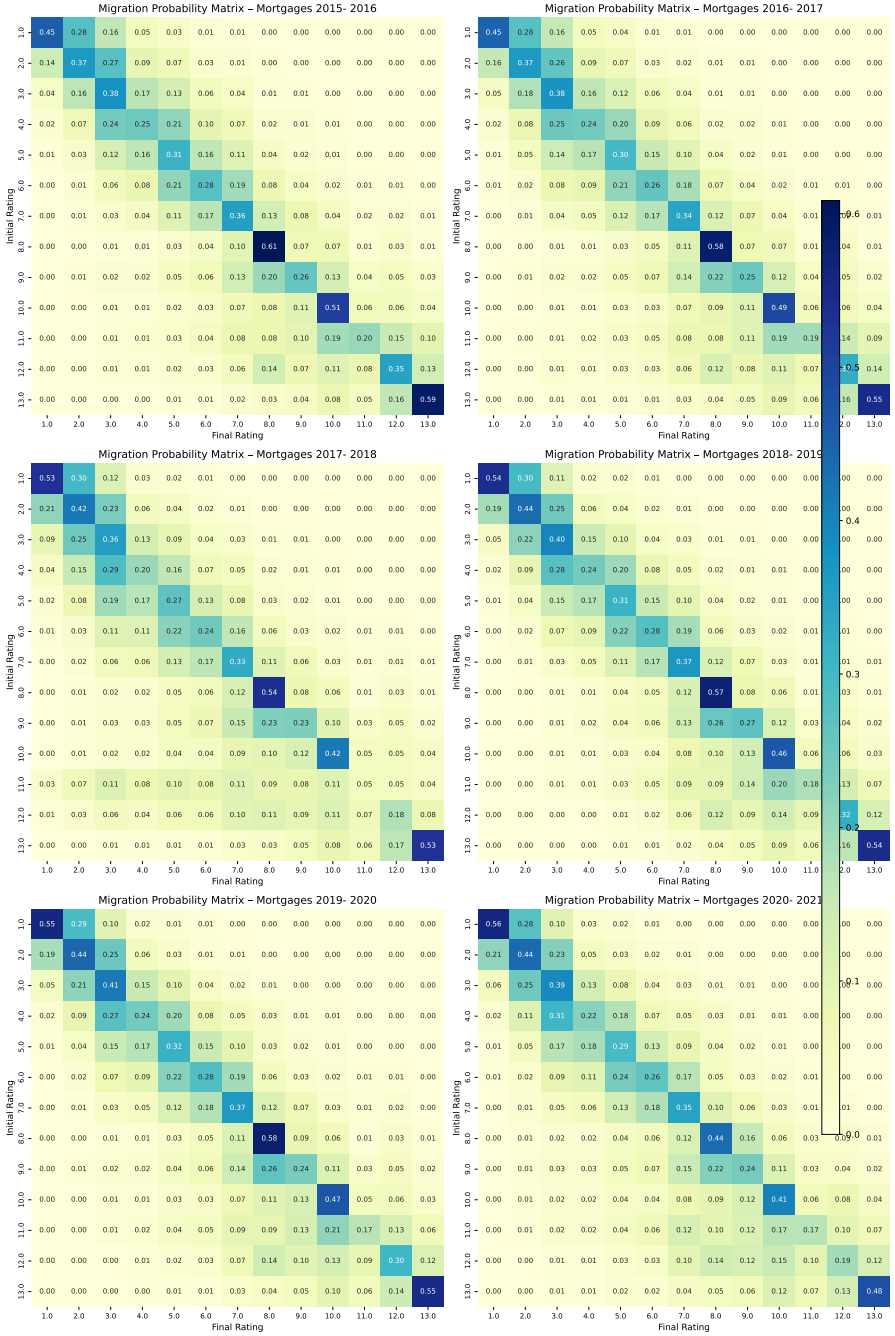

Figure 8.1: Year-over-Year Estimated Forward Migration Matrices

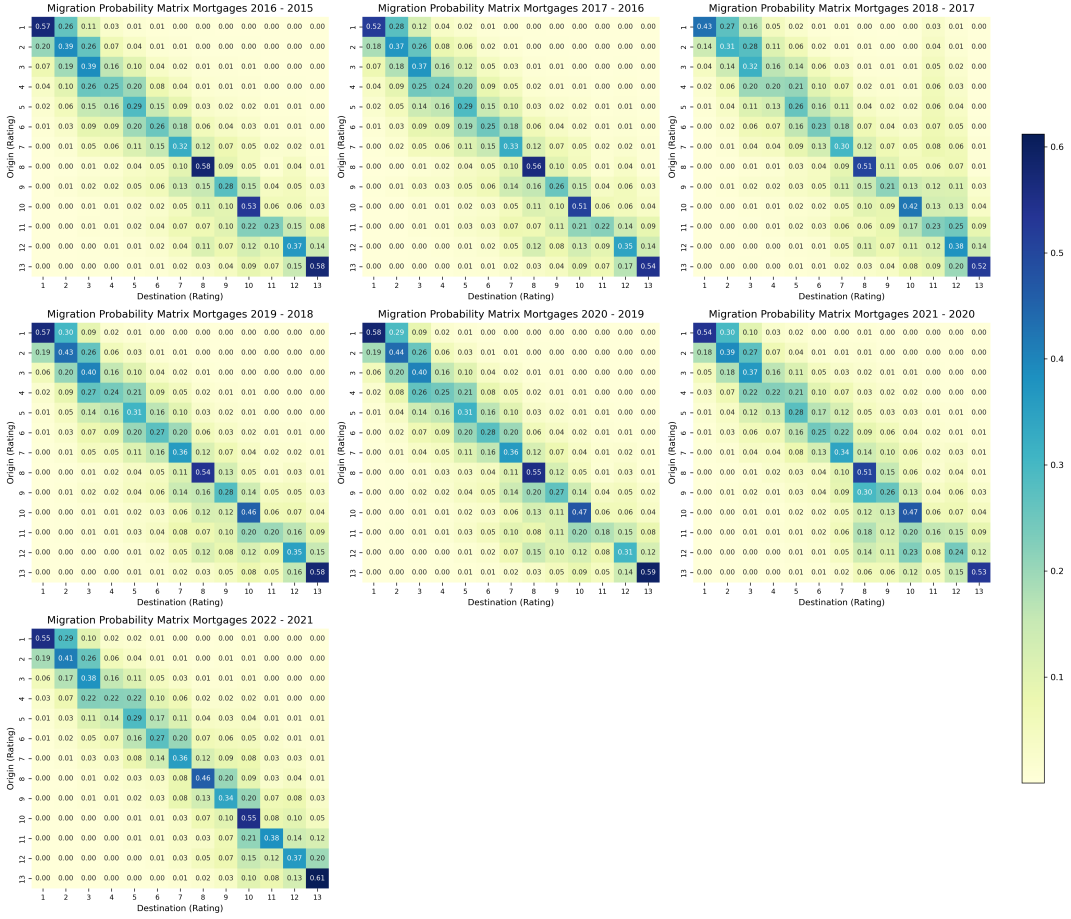

Figure 8.2: Year-over-Year Estimated Backward Migration Matrices

The heatmaps of the migration matrices show that, across all years, the highest probability for each initial rating is concentrated along the main diagonal. This reflects significant credit stability: most borrowers tend to maintain the same rating level over time. Movements outside the diagonal indicate that:

- transitions to **adjacent grades** are the most common changes, representing moderate risk variations;
- **severe downgrades** are relatively rare but highlight cases of material credit deterioration;
- **upgrades** are present but less frequent, confirming a generally prudent portfolio profile.

Next, we present the trends in stability percentage, stability probability, and the percentages of deteriorations and improvements.

8.2.2 Analysis of Estimated Forward Matrices

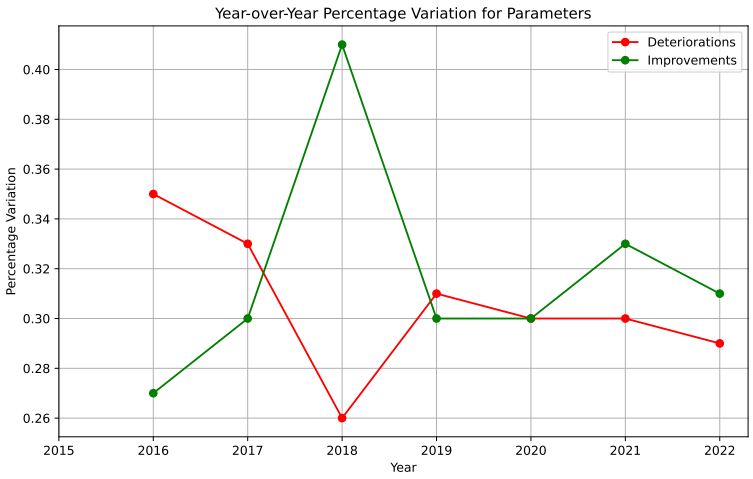

Figure 8.3: Year-over-Year Percentage Variation for Percentages of Deteriorations and Improvements

In the figure 8.3 we can observe the trend of the percentages of **deteriorations** (number of counterparties that downgrade over the total number of counterparties) and **improvements** (number of counterparties that upgrade their rating) over the years.

The deterioration curve reaches its minimum in 2016, when only 27% of counterparties downgraded. It then increases to a peak of 40% in 2018, decreases in 2019, and rises again in 2022.

Conversely, the improvements curve reaches its minimum in 2018, with 26%, then rises in 2019, before decreasing again until 2022.
Overall, the two graphs are almost specular, highlighting the opposing dynamics between deteriorations and improvements.

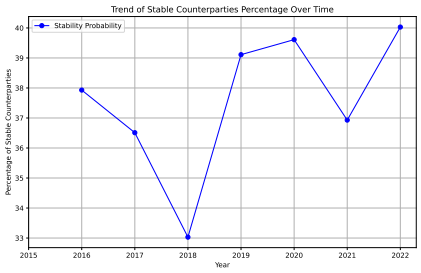

Figure 8.4: Trend of Stable Counterparties Percentage Over Time.

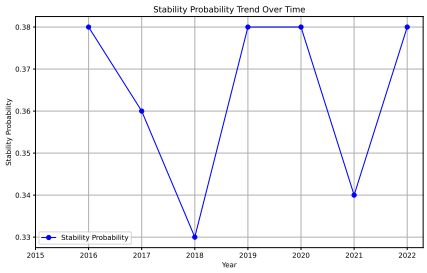

Figure 8.5: Stability Probability Trend Over Time

In Figures 8.4 and 8.5, we report the trends of the percentage of counterparties that do not change their rating from one year to the next and the probability of counterparties maintaining their rating. Both indicators fluctuate within a narrow range, with the percentage ranging from 33% to 40% and the probability from 33% to 38%.

They reach their minimum in 2018 and their maximum in 2022, with another local minimum observed in 2021.

These patterns are consistent with those shown in Figure 8.3: periods of lower stability (2018 and 2021) coincide with higher deterioration rates and fewer improvements, while periods of higher stability (2016 and 2022) correspond to fewer downgrades and more upgrades. This confirms the strong interdependence between stability, improvements, and deteriorations.

Chapter 9

Reconstruction of Missing Data through Dynamic Equations and Migration Matrices

In order to reconstruct the distribution of observations in the missing years, it is necessary to understand the underlying dynamics of the system — specifically, how counterparties move into and out of the mortgage portfolio over time. To formally capture this behavior, we introduce the dynamic equation governing the evolution of the rating distribution over time.

To model the yearly dynamics of the mortgage portfolio, we define several vectors that capture the distribution and transitions of counterparties across rating classes. Each vector is of dimension $1 \times k$, where k is the number of rating classes (e.g., 1, 2, 3, 4, ..., 13). These rating classes remain consistent over time and define the state space of our Markov framework.

The following vectors are used:

- Obs_t: Vector representing the number of counterparties present in the portfolio at time t, distributed across rating classes.
- New_t: Vector representing the number of counterparties entering the portfolio between $t-1$ and t, for instance due to new mortgage originations.
- Exit_t: Vector capturing the number of counterparties leaving the portfolio between $t-1$ and t, due to events such as full loan repayment, refinancing, or portfolio sale.
- Default_t: Vector indicating the number of counterparties transitioning into the default state at time t.
- CuredDefault_t: Vector representing the number of counterparties previously in default who returned to performing states between $t-1$ and t.

All the above vectors have been computed directly from the available dataset for the years 2015 to 2022. They form the empirical basis for reconstructing and understanding the system dynamics during the observed period.

9.1 Estimation of Non-Observed Distributions via Dynamic Equation

In the following, we present the dynamic system equations under both forward and backward formulations.
The **Forward Formulation** leads to

$$\hat{\boldsymbol{Obs}}_t = (\boldsymbol{Obs}_{t+1} - \text{Cured Default}_{t+1} - \text{New}_{t+1}) \cdot \left(P^{\text{fwd}}_{\text{t,t+1}}\right)^{-1} + \text{default}_{t+1} + \text{exit}_{t+1}$$

where $P^{\text{fwd}}_{\text{t,t+1}}$ denotes the forward transition matrix between years t and $t+1$.

The **Backward Formulation** leads to

$$\hat{\boldsymbol{Obs}}_t = (\boldsymbol{Obs}_{t+1} - \text{Cured Default}_{t+1} - \text{New}_{t+1}) \cdot P^{\text{bwd}}_{\text{t,t+1}} + \text{default}_{t+1} + \text{exit}_{t+1}$$

where $P^{\text{bwd}}_{\text{t,t+1}}$ denotes the backward transition matrix between years t and $t+1$.

We have tested both dynamic formulations over the period from 2015 to 2022 and successfully reconstructed the observed distributions with zero reconstruction error.
Based on these results, we choose to adopt the backward formulation for extrapolating the missing distributions from 2008 to 2014. This choice allows us to avoid the numerical instability associated with the inversion of forward transition matrices.
However, to apply this approach to the unobserved period, certain assumptions must be introduced.

To enable the backward reconstruction of rating distributions for the period 2008–2014, we rely on the following set of assumptions:

- **Constant Structure of Entries and Exits**: The inflow of new exposures and the outflow due to portfolio exits are assumed to follow a stable pattern over time. Specifically, for each rating class and each unobserved year, the number of new entries and exits is assumed to be equal to the average observed over the period 2015–2022. It is worth noting that entries and exits may add complexity to the reconstruction process and, if needed, may require additional assumptions or access to more granular historical data.
- **Stable Performing Behavior**: Transition probabilities among performing rating classes are assumed to remain approximately constant, and are set equal to their average values observed during the 2015–2022 period.
- **Default Reconstruction**: The number of exposures transitioning into default was estimated using a regression-based methodology.

- **Consistent Default Definition**: The definition of default is assumed to remain unchanged throughout the historical period under analysis, allowing for comparability across years.

Based on these assumptions, we tested three alternative approaches for estimating the yearly rating distributions during the period 2008–2014:

1. **Fully Averaged Inputs**: Using the average values from 2015–2022 for all components — new entries, exits, cured defaults, and defaults.

$$\hat{\boldsymbol{Obs}}_t = (\boldsymbol{Obs}_{t+1} - \text{Cured Default}_{\text{avg}} - \text{New}_{\text{avg}}) \cdot P_{\text{avg}}^{\text{bwd}} + \text{Default}_{\text{avg}} + \text{Exit}_{\text{avg}}$$

2. **Regression-Driven Defaults with Averaged Flows**: Using regression-estimated default values while keeping new entries, exits, and cured defaults fixed at their average 2015–2022 values.

$$\hat{\boldsymbol{Obs}}_t = (\boldsymbol{Obs}_{t+1} - \text{Cured Default}_{\text{avg}} - \text{New}_{\text{avg}}) \cdot P_{\text{avg}}^{\text{bwd}} + \text{RegressionDefault}_{t+1} + \text{Exit}_{\text{avg}}$$

3. **Hybrid Method**: Using the average values from 2015–2022 for new entries and exits, while estimating cured defaults by applying the average cured default rate (2015–2022) to the regression-estimated default counts for each year.

$$\hat{\boldsymbol{Obs}}_t = (\boldsymbol{Obs}_{t+1} - (\text{Cured Default Rate}_{\text{avg}} \times \text{RegressionDefault}_{t+1}) - \text{New}_{\text{avg}}) \cdot P_{\text{avg}}^{\text{bwd}} + \text{RegressionDefault}_{t+1} + \text{Exit}_{\text{avg}}$$

where:

- $\boldsymbol{Obs}_{t+1}$ is the observed rating distribution vector at time $t+1$.
- For each rating grade $i = 1, \ldots, k$, where k is the number of rating classes:

$$\text{New}_{i,\text{avg}} = \frac{1}{7} \sum_{t=2016}^{2022} \text{New}_{i,t}$$

$$\text{Exit}_{i,\text{avg}} = \frac{1}{7} \sum_{t=2016}^{2022} \text{Exit}_{i,t},$$

$$\text{Cured Default}_{i,\text{avg}} = \frac{1}{7} \sum_{t=2016}^{2022} \text{Cured Default}_{i,t}$$

  represent the average vectors computed over the years 2015–2022.

- $P_{\text{avg}}^{\text{bwd}}$ is the average backward transition matrix computed over 2015–2022.

- $\text{RegressionDefault}_{t+1}$ is the estimated default vector at time $t+1$ obtained via regression.
- Cured Default Rate_{avg} is a vector of average cured default rates computed over 2015–2022. In Method 3, the term

$$\text{Cured Default Rate}_{\text{avg}} \times \text{RegressionDefault}_{t+1}$$

denotes the element-wise (Hadamard) product between the two vectors.

9.2 Reconstructed Rating Distribution

This section presents the reconstructed distribution of counterparties across ratings obtained using the three methodological approaches described in the previous section.

9.2.1 Fully Averaged Inputs Approach

Table 9.1: Normalized distribution of Ratings for Fully Averaged Inputs Approach (2008–2014)

Year	1	2	3	4	5	6	7	8	9	10	11	12	13
2008	5.66%	8.87%	12.20%	7.64%	9.77%	7.80%	9.73%	11.51%	6.53%	8.42%	3.31%	4.63%	3.93%
2009	5.71%	8.91%	12.24%	7.66%	9.77%	7.79%	9.70%	11.46%	6.52%	8.39%	3.32%	4.63%	3.90%
2010	5.78%	8.95%	12.29%	7.68%	9.79%	7.78%	9.67%	11.40%	6.50%	8.33%	3.35%	4.62%	3.85%
2011	5.77%	9.01%	12.41%	7.76%	9.87%	7.84%	9.70%	11.36%	6.45%	8.23%	3.28%	4.53%	3.79%
2012	5.76%	9.08%	12.53%	7.84%	9.97%	7.89%	9.72%	11.31%	6.38%	8.14%	3.20%	4.43%	3.74%
2013	5.76%	9.14%	12.67%	7.93%	10.07%	7.93%	9.71%	11.27%	6.31%	8.06%	3.14%	4.33%	3.70%
2014	5.72%	9.17%	12.86%	8.08%	10.24%	7.99%	9.65%	11.29%	6.15%	8.07%	2.92%	4.15%	3.71%

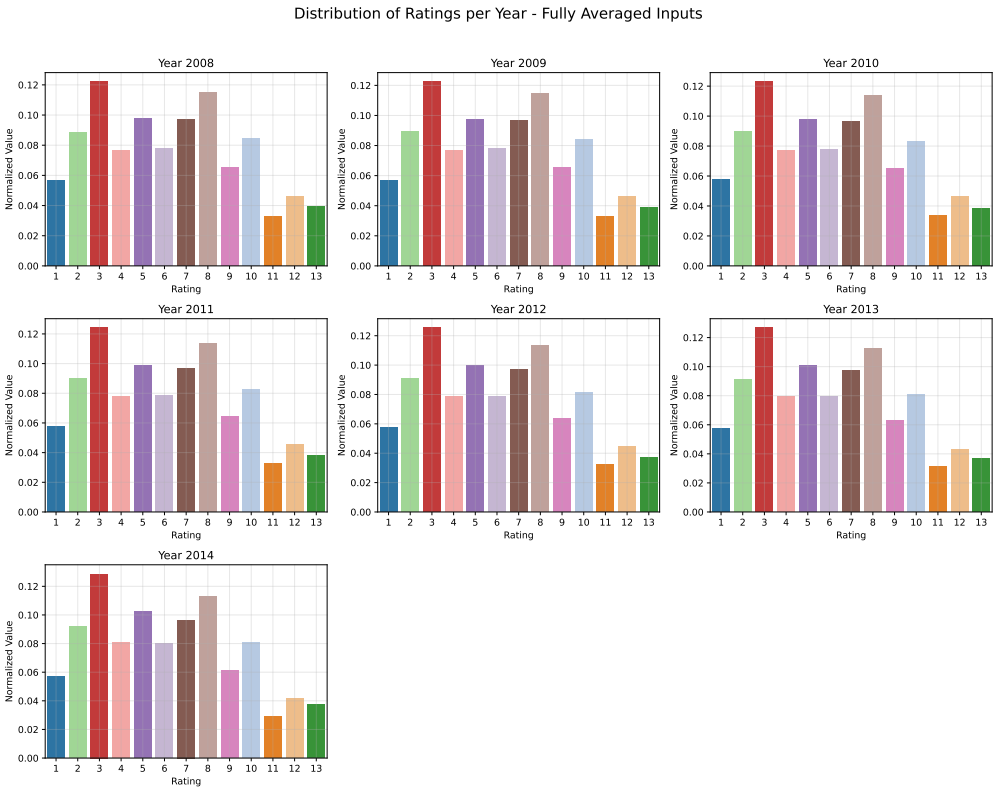

Figure 9.1: Evolution of Ratings over Years - Fully Averaged Inputs Approach

The distribution exhibits a characteristic **bell-shaped pattern**, with central ratings (3–8) consistently exhibiting the highest proportions. Ratings at the extremes (1–2 and 11–13) are comparatively underrepresented, reflecting fewer counterparties in these categories. Over the period 2008–2014, there is a modest upward trend in mid-to-high ratings (particularly 5–8) accompanied by slight declines in the lowest and highest ratings. The overall distribution remains stable, highlighting that this approach fails to capture the year-to-year variations in the distributions of defaults and performing exposures.

9.2.2 Regression-Driven Defaults with Averaged Flows Approach

Table 9.2: Normalized distribution of Ratings for Regression-Driven Defaults with Averaged Flows (2008–2014)

Year	1	2	3	4	5	6	7	8	9	10	11	12	13
2008	5.31%	8.37%	11.58%	7.29%	9.37%	7.53%	9.50%	11.40%	6.56%	8.64%	3.54%	5.25%	5.66%
2009	5.44%	8.51%	11.73%	7.37%	9.44%	7.57%	9.51%	11.37%	6.54%	8.58%	3.50%	5.15%	5.28%
2010	5.56%	8.64%	11.88%	7.45%	9.51%	7.60%	9.51%	11.30%	6.51%	8.47%	3.49%	5.07%	5.00%
2011	5.61%	8.78%	12.10%	7.57%	9.66%	7.69%	9.56%	11.25%	6.43%	8.30%	3.39%	4.89%	4.77%
2012	5.66%	8.93%	12.34%	7.72%	9.83%	7.79%	9.62%	11.22%	6.36%	8.16%	3.27%	4.67%	4.42%
2013	5.71%	9.06%	12.56%	7.86%	9.99%	7.88%	9.64%	11.21%	6.29%	8.06%	3.16%	4.47%	4.11%
2014	5.70%	9.15%	12.83%	8.07%	10.22%	7.98%	9.62%	11.27%	6.13%	8.06%	2.91%	4.20%	3.86%

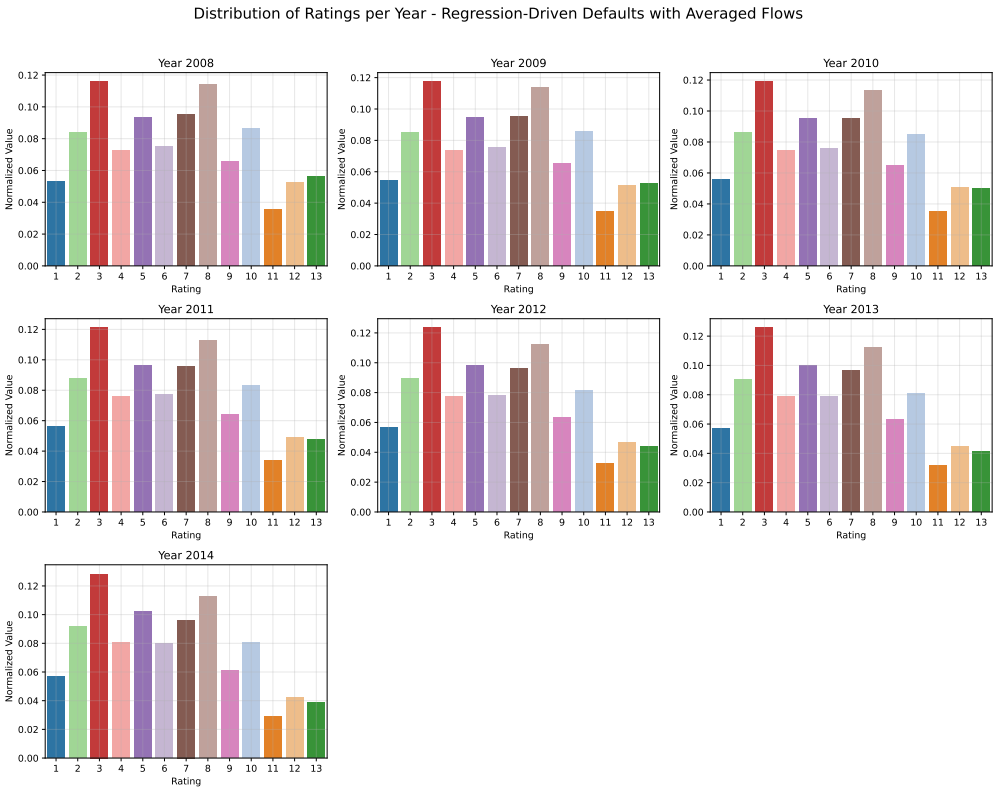

Figure 9.2: Evolution of Ratings over Years - Regression-Driven Defaults with Averaged Flows Approach

The reconstructed distribution for this approach similarly demonstrates a **bell-shaped pattern**, with the highest proportions concentrated in central ratings (3–8). Extreme ratings remain comparatively lower; however, the uppermost rating (13) shows slightly higher representation than in the Fully Averaged Inputs Approach. Across the years 2008–2014, mid-to-high ratings (5–8) display a gradual increase, while the lowest and extreme high ratings are relatively stable or exhibit slight decreases. **Compared to the Fully Averaged Inputs Approach, the distribution here exhibits slightly greater year-to-year variation, suggesting that the method introduces more dynamism in the rating assignments.** The overall shape is consistently maintained over time, indicating a stable emphasis on central ratings despite the observed variations.

9.2.3 Hybrid Approach

Table 9.3: Normalized distribution of Ratings for Hybrid Method(2008–2014)

Year	1	2	3	4	5	6	7	8	9	10	11	12	13
2008	5.37%	8.46%	11.69%	7.35%	9.44%	7.58%	9.54%	11.42%	6.55%	8.59%	3.50%	5.15%	5.36%
2009	5.49%	8.58%	11.82%	7.42%	9.50%	7.61%	9.54%	11.38%	6.53%	8.53%	3.47%	5.08%	5.05%
2010	5.60%	8.69%	11.95%	7.49%	9.56%	7.63%	9.53%	11.31%	6.50%	8.43%	3.46%	5.01%	4.82%
2011	5.64%	8.82%	12.15%	7.60%	9.70%	7.72%	9.58%	11.26%	6.43%	8.28%	3.37%	4.84%	4.62%
2012	5.68%	8.96%	12.37%	7.74%	9.85%	7.81%	9.64%	11.23%	6.36%	8.15%	3.26%	4.64%	4.32%
2013	5.72%	9.07%	12.58%	7.87%	10.00%	7.89%	9.65%	11.22%	6.29%	8.05%	3.16%	4.45%	4.06%
2014	5.71%	9.15%	12.84%	8.07%	10.22%	7.98%	9.63%	11.27%	6.13%	8.06%	2.90%	4.19%	3.84%

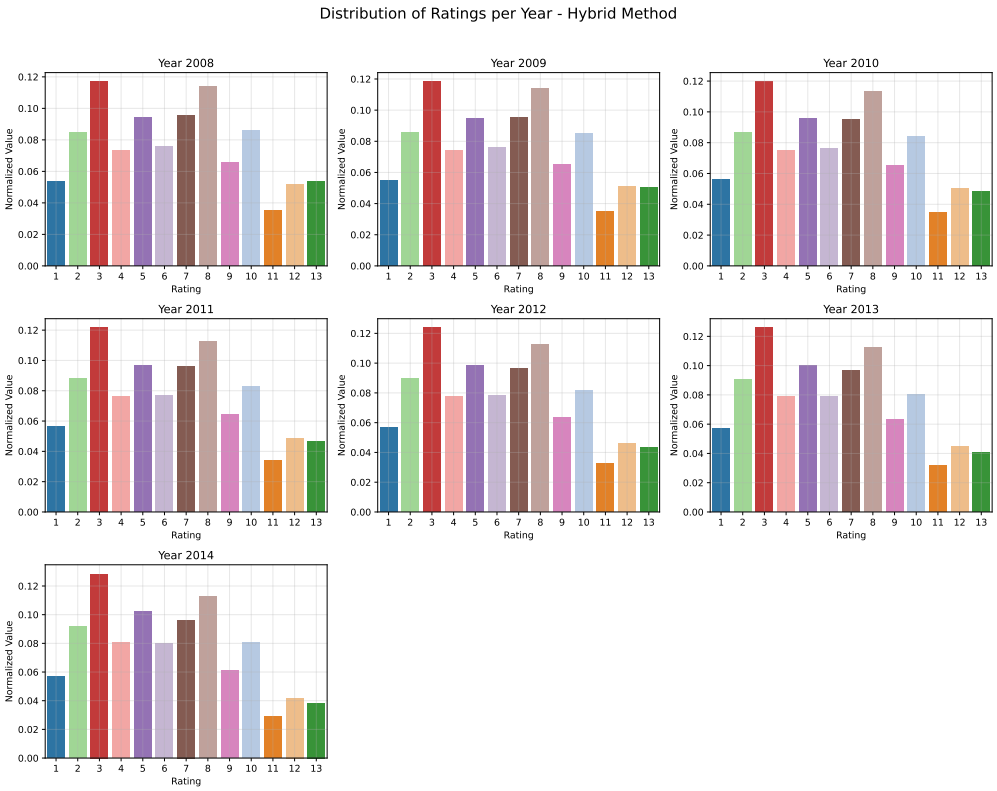

Figure 9.3: Evolution of Ratings over Years - Hybrid Method

The distribution obtained via the Hybrid Method retains a **bell-shaped profile**, with central ratings (3–8) exhibiting the highest frequencies. Extreme ratings (1–2 and 11–13) are consistently lower, though the upper extreme (13) initially presents slightly greater representation relative to the other approaches. From 2008 to 2014, mid-to-high ratings (5–8) gradually increase, while low and extreme high ratings show minor declines or remain stable. **Notably, compared with the Fully Averaged Inputs Approach, the Hybrid Method displays more pronounced year-to-year fluctuations in the distributions, indicating a higher degree of variability introduced by this approach.** Overall, the distribution is maintained across years, reflecting the method's consistent emphasis on central ratings with moderate temporal shifts.

9.2.4 Comparative Analysis of Rating Evolution (2008–2014)

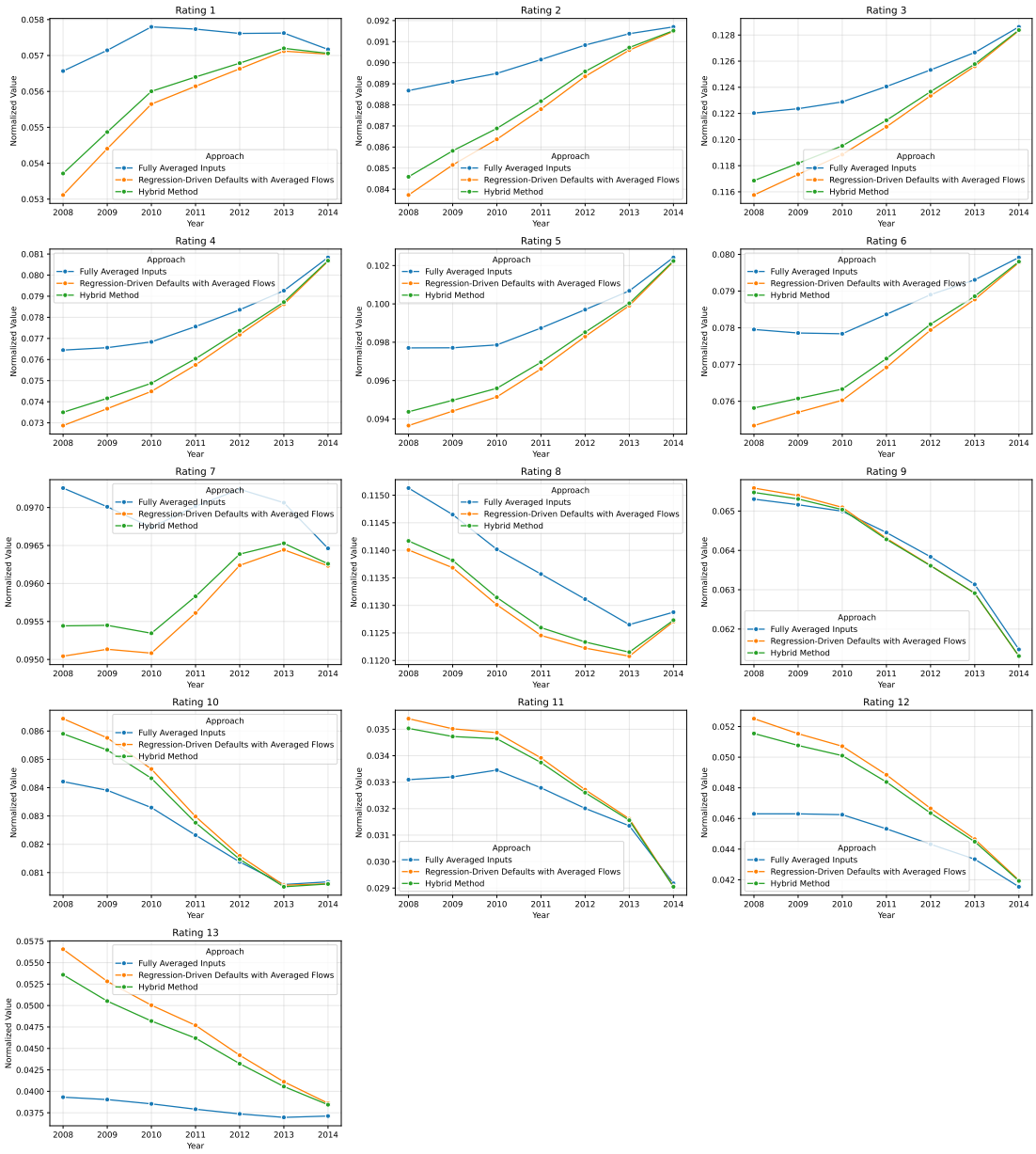

Figure 9.4: Evolution of normalized distributions per rating (2008–2014), comparison across the three approaches.

Figure 9.4 presents the evolution of normalized rating distributions across the period 2008–2014 for all three approaches. Several key observations can be drawn:

- Across all approaches, the distributions maintain a **bell-shaped profile**, with central ratings (3–8) consistently showing the highest frequencies. This indicates a persistent emphasis on mid-level ratings across methodologies.
- For the lower-risk ratings (1–7), the distribution remains relatively **stable** under the Fully Averaged Inputs Approach, while the Regression-Driven Defaults and Hybrid Approaches exhibit slightly **greater year-to-year variability**, reflecting a more dynamic reallocation of counterparties within these categories.
- In the higher-risk ratings (8–13), Approaches 2 and 3 display a modest downward trend, whereas the Fully Averaged Inputs Approach shows only minimal variation. This occurs because, by relying solely on averaged vectors, the method tends to "smooth out" temporal fluctuations, thus limiting its ability to capture distributional shifts over time.
- Overall, annual changes remain **gradual**, with no abrupt variations across methods. The main trend indicates subtle yet consistent shifts toward mid-to-high ratings under Approaches 2 and 3, while Approach 1 stays essentially constant.
- These findings suggest that, although all methods preserve the central tendency of the distributions, Approaches 2 and 3 introduce slightly more dynamics, potentially capturing finer fluctuations in counterparties' risk profiles.
- **Calibration implication:** as will be shown in the following section, the calibration procedure will require the introduction of a smaller *inflation factor* for Approaches 2 and 3, in order to properly adjust the sample and account for the greater variability observed.

In summary, the Fully Averaged Inputs Approach produces highly stable distributions, as averaging tends to smooth out year-to-year fluctuations. By contrast, the Regression-Driven Defaults and Hybrid Approaches reveal slightly stronger dynamics, with less stability in lower ratings and a smoother downward adjustment in higher ratings. Consequently, in calibration, Approaches 2 and 3 will require a smaller inflation factor to balance the sample.

9.2.5 Implied Default Rate Analysis (2008–2014)

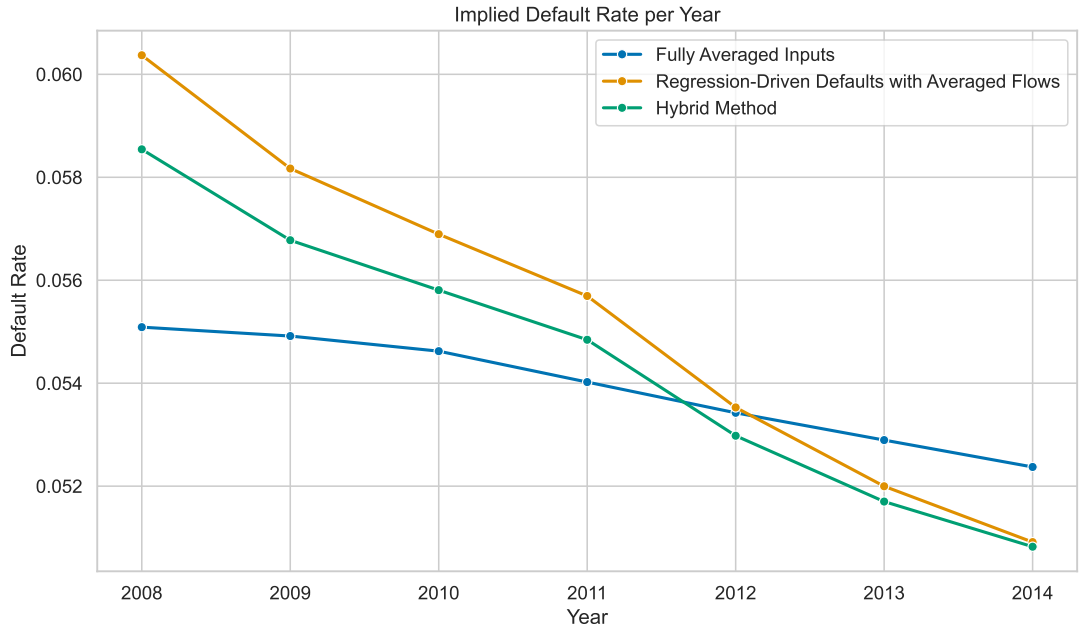

Figure 9.5: Implied Default Rates per year (2008–2014) computed for Approaches 1, 2, and 3, compared with observed portfolio DRs.

Figure 9.5 illustrates the evolution of the implied default rates (DRs) for the three approaches across the period 2008–2014, alongside the actual observed portfolio DRs.

As can be seen, the Fully Averaged Inputs Approach tends to smooth out year-to-year variations and, as a result, does not fully capture the real trend of the observed default rates. This limitation reflects the inherent averaging of the method, which reduces responsiveness to fluctuations in the portfolio. In contrast, both the Regression-Driven Defaults and the Hybrid Approach closely follow the observed DRs, effectively capturing the annual changes. Overall, the comparison indicates that these last two approaches provide a similar trend to the observed default rates, offering a more accurate and responsive estimation of portfolio risk over time.
In conclusion, the implied default rate analysis highlighting that the last two approaches are better suited to reflect dynamic changes in risk profiles over time.

Chapter 10

Long-Run PD Calibration Methodology

In Section 2.3.1 we introduced two calibration methods: **Linear Scaling** and **Ordinary Least Squares (OLS)** regression. In this chapter, we provide a detailed explanation of these procedures, outlining their underlying rationale, implementation steps, and role in translating the Long-Run Average Default Rate (LRAvDR) into grade-level Baseline Probabilities of Default (PDs). In addition, we implement the **Linear Scaling** procedure and report in this section the results obtained from its application.

10.0.1 Linear Scaling

During the linear scaling procedure, the estimated average default rate of the portfolio is computed as:

$$\hat{DR} = \sum_{i=1}^{n} \bar{DR}_i \cdot f_i,$$

where n denotes the total number of grades in the rating scale, $\bar{DR}_i$ is the estimated average default rate of grade i over the reference period (2008–2022, selected as the calibration sample), and $f_i = \frac{obs_i}{obs_{tot}}$ represents the relative frequency of estimated observations in grade i.
The realized average default rate of the portfolio, denoted by $\bar{DR}_{port}$, is then used to compute the scaling factor:

$$\rho = \frac{\bar{DR}_{port}}{\hat{DR}}.$$

Finally, the grade-level probabilities of default are obtained as:

$$PD_i = \bar{DR}_i \cdot \rho.$$

10.0.2 Ordinary Least Squares (OLS) Regression

Differently from the Linear Scaling procedure, the OLS method requires solving the following optimization problem:

$$\min_{\{PD_i\}} \text{ Total Error} \quad \text{subject to} \quad PD_i \leq \hat{DR}_i, \quad \bar{PD} = DR_{port},$$

where the total error is defined as

$$\text{Total Error} = \sum_{i=1}^{n} \left(\ln(DR_i) - \ln(PD_i)\right)^2,$$

and the portfolio-average probability of default is given by

$$\bar{PD} = \sum_{i=1}^{n} PD_i \cdot f_i,$$

with n denoting the number of rating grades, DR_i the observed default rate for grade i, PD_i the calibrated probability of default for grade i, and f_i the relative frequency of observations in grade i within the portfolio.

10.1 Results of Linear Scaling Calibration

In this section, we present the results obtained by applying the Linear Scaling procedure to the estimated default rates derived using the three different methods. Additionally, we compare our method against:

- the approach that only estimates default rates without adjusting performing exposures
- the direct linear scaling.

For each method, the results are visualized using histograms representing the observed default rates per rating grade, accompanied by the corresponding probability of default (PD) curve obtained through the Linear Scaling procedure. Alongside each plot, we include a table reporting the pre-calibration default rates (DR_i) and the post-calibration probabilities of default (PD_i). This layout allows for a clear comparison of the default rate distributions and the calibrated PDs across the different estimation methods.

10.1.1 Reconstruction of Observations via Migration Matrix and Dynamic Equations

This subsection presents the results obtained when the observation distribution is reconstructed using migration matrices and dynamic equations, while default rates are estimated via regression (outside Fully Averaged Inputs Approach). Both the probabilities of default and the distribution of performing exposures across rating grades are considered.

Fully Averaged Inputs Approach

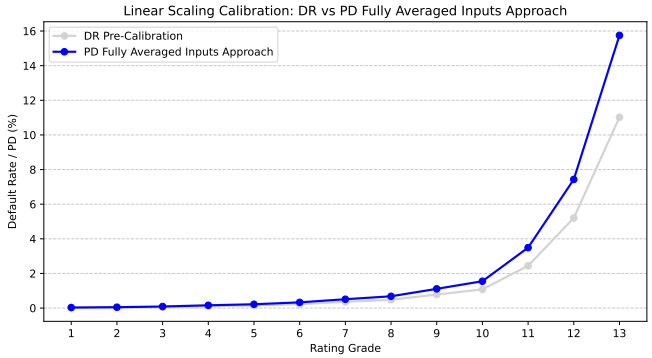

Figure 10.1: Histogram of default rates per grade with PD curve.

Grade	$DR_i(\%)$	$PD_i(\%)$
1	0.02	0.03
2	0.04	0.05
3	0.06	0.09
4	0.11	0.16
5	0.15	0.22
6	0.23	0.33
7	0.36	0.51
8	0.48	0.68
9	0.78	1.11
10	1.08	1.55
11	2.44	3.49
12	5.20	7.43
13	11.02	15.75

Figure 10.2: Pre-calibration default rates and post-calibration PDs.

The Fully Averaged Inputs Approach shows that post-calibration PDs tend to increase progressively across rating grades, with larger deviations in the upper tail. This suggests a systematic overestimation of risk when assuming averaged inputs over time.

Regression-Driven Defaults with Averaged Flows Approach

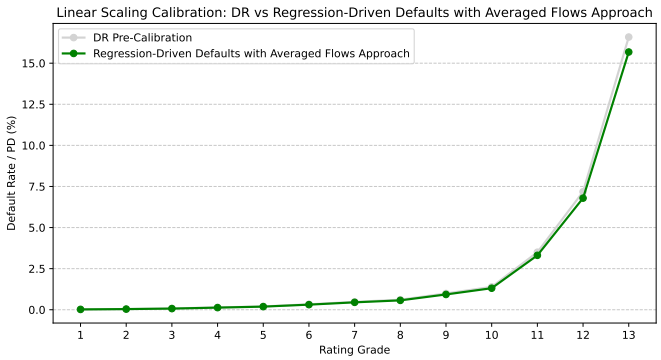

Figure 10.3: Histogram of default rates per grade with PD curve.

Grade	$DR_i(\%)$	$PD_i(\%)$
1	0.02	0.02
2	0.04	0.04
3	0.08	0.07
4	0.14	0.13
5	0.20	0.19
6	0.33	0.31
7	0.48	0.45
8	0.61	0.57
9	0.99	0.93
10	1.39	1.31
11	3.50	3.31
12	7.18	6.79
13	16.59	15.68

Figure 10.4: Pre-calibration default rates and post-calibration PDs.

Here, post-calibration PDs are closely aligned with observed DRs across almost all grades, with only marginal deviations. This indicates that regression-driven estimation combined with averaged flows provides a stable calibration outcome.

Hybrid Approach

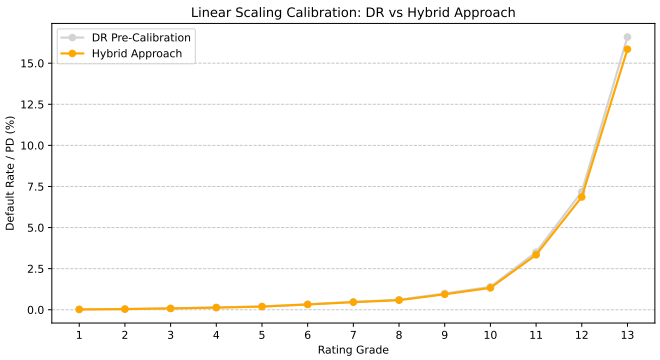

Figure 10.5: Histogram of default rates per grade with PD curve.

Grade	$DR_i(\%)$	$PD_i(\%)$
1	0.02	0.02
2	0.04	0.04
3	0.08	0.07
4	0.14	0.13
5	0.20	0.19
6	0.33	0.31
7	0.48	0.45
8	0.61	0.57
9	0.99	0.93
10	1.39	1.31
11	3.50	3.31
12	7.18	6.79
13	16.59	15.68

Figure 10.6: Pre-calibration default rates and post-calibration PDs.

The Hybrid Approach also yields PDs very close to the observed DRs, showing only limited deviations across grades. This balance suggests that combining regression and flow-based information helps to preserve consistency in the calibration.

10.1.2 Estimation of Default Rates Using Regression

Here we present the results obtained using regression to estimate default rates, assuming that the distribution of performing exposures in the years 2015–2022 remains stable while the default rates $\bar{D}_i$ are averaged over the entire period 2008–2022.

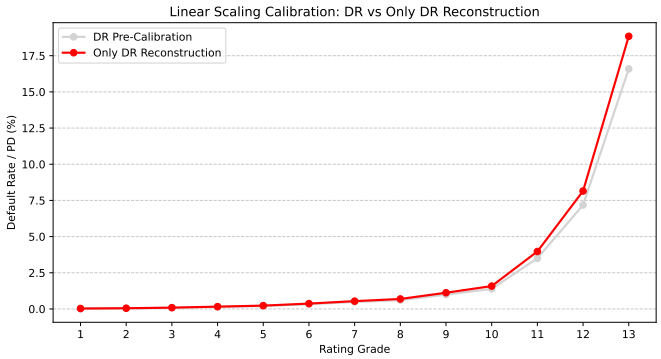

Figure 10.7: Histogram of default rates per grade with PD curve.

Grade	$DR_i(\%)$	$PD_i(\%)$
1	0.02	0.03
2	0.04	0.05
3	0.08	0.09
4	0.14	0.16
5	0.20	0.23
6	0.33	0.37
7	0.48	0.54
8	0.61	0.69
9	0.99	1.12
10	1.39	1.58
11	3.50	3.97
12	7.18	8.15
13	16.59	18.84

Figure 10.8: Pre-calibration default rates and post-calibration PDs.

Regression-based estimation leads to post-calibration PDs that are generally higher than the observed default rates, especially in the upper grades. This reflects the impact of keeping the distribution of performing exposures stable while reconstructing only the default rates.

10.1.3 Results of Direct Linear Scaling Applied to the 2015–2022 Sample

In this subsection, we present the results obtained by applying the Linear Scaling procedure directly to the 2015–2022 sample, under the assumption that the distribution of performing exposures during this period remains stable. In this scenario, the default rates (DR_i) are calculated based solely on the observed defaults within the 2015–2022 period.

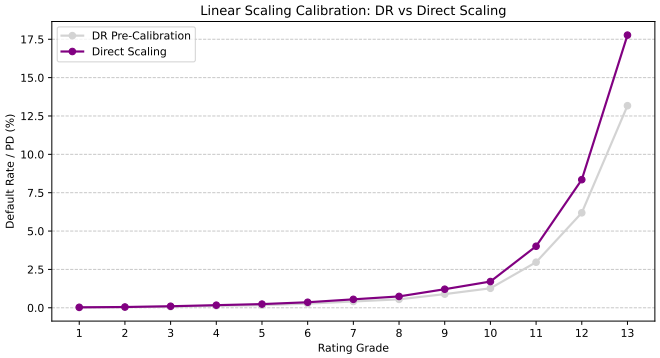

Figure 10.9: Histogram of default rates per grade with PD curve.

Grade	DR_i(%)	PD_i(%)
1	0.02	0.03
2	0.04	0.05
3	0.07	0.10
4	0.13	0.17
5	0.17	0.24
6	0.27	0.36
7	0.41	0.55
8	0.55	0.74
9	0.89	1.21
10	1.27	1.71
11	2.97	4.01
12	6.19	8.35
13	13.17	17.77

Figure 10.10: Pre-calibration default rates and post-calibration PDs.

Direct Linear Scaling produces systematically higher PDs than observed DRs, with increasing distortions in higher grades. This is due to the direct application of scaling without reconstructing either default rates or exposure distributions.

10.1.4 Comparison of Absolute Differences Between Pre- and Post-Calibration

Table 10.1: Differences between Pre- and Post-Calibration PDs for Each Approach (%)

Grade	Fully Averaged Inputs	Regression-Driven Defaults & Avg Flows	Hybrid	Regression-Driven Defaults	Direct LS
1	0.01	0.00	0.00	0.00	0.01
2	0.02	0.00	0.00	0.01	0.01
3	0.03	0.00	0.00	0.01	0.02
4	0.05	-0.01	-0.01	0.02	0.04
5	0.07	-0.01	-0.01	0.03	0.06
6	0.10	-0.02	-0.01	0.04	0.09
7	0.15	-0.03	-0.02	0.06	0.14
8	0.20	-0.03	-0.03	0.08	0.19
9	0.33	-0.05	-0.04	0.13	0.31
10	0.46	-0.08	-0.06	0.19	0.44
11	1.05	-0.19	-0.16	0.47	1.04
12	2.23	-0.39	-0.32	0.97	2.16
13	4.73	-0.91	-0.74	2.25	4.60

Table 10.1 summarizes the differences between the pre-calibration default rates (DR_i) and the post-calibration probabilities of default (PD_i) for each rating grade across the various approaches. The differences highlight how methodological assumptions influence the calibration results.

The last two approaches show the largest distortions in the post-calibration PDs.
In the Regression-Driven Default Approach, PD_i values increase moderately, with deviations ranging from 0.00% for the lowest grades to 2.25% for the highest grade. This comparison reflects that it is not sufficient to only reconstruct default rates (DR_i) without considering the heterogeneity in portfolio risk across periods. When the riskiness of a given sample (e.g., 2015–2022) differs significantly from the longer calibration horizon (2008–2022), approaches that only rely on DR_i reconstruction tend to produce biased results.
Similarly, Direct Scaling produces deviations up to 4.60% in the highest grade, due to applying scaling directly without reconstructing either default rates or exposure distributions.

The first approach, which assumes that the distribution of defaults remains stable over the years while only the distribution of performing exposures changes, consistently leads to an overestimation of risk. In fact, we observe a difference greater than one notch in the riskiness of the rating profile compared to the pre-calibration distribution.

In contrast, the second and the third approaches exhibit the smallest deviations across all

grades, with maximum differences of 0.74% and 0.91% respectively. This demonstrates that reconstructing both default rates and exposure distributions, without assuming stability of the performing exposures, produces the most balanced and realistic PD curves.

These observations confirm the key point: accurate estimation of default rates and proper reconstruction of the exposure distribution are essential to adequately capture the inherent credit risk.

Table 10.2: ρ of each method

	Fully Averaged Inputs	Regression-Driven Defaults & Avg Flows	Hybrid	Regression-Driven Defaults	Direct LS
ρ	142.94%	94.53%	95.56%	113.56%	134.90 %

Table 10.2 reports the ρ coefficients for each method. Values close to 100% (as in the second and in the third, with 94.53% and 95.56% respectively) indicate good alignment with the target distribution, while higher values (e.g., 142.94% for Fully Averaged Inputs Approach and 134.90% for Direct LS) confirm the overestimation of risk observed in Table 10.1.

Chapter 11

Accuracy Ratio after Calibration: Results and Statistical Assessment

The empirical results presented in the last chapter highlight that accurate risk estimation requires more than a simple recalibration of default rates (DR_i). As shown, approaches that reconstruct both the DR_i and the exposure distribution achieve minimal deviations and produce probability-of-default curves that are more balanced and consistent with the estimated DR_i.
In contrast, methods that only adjust the DR_i or methods based on overly simplistic assumptions regarding default exposures, or exclusively use the available sample to recalibrate over a period with missing data, generate significant calibration bias, particularly in the tail risk (see Table 10.1 and 10.2).

As discussed in the chapter on AR, such deviations in default rates directly translate into a deviation of AR: the larger the misalignment of PD_i relative to DR_i, the greater the misalignment in the AR compared to the baseline level.

In line with the *EBA Guidelines*, maintaining the discriminatory power of a rating model is an essential requirement, since significant changes can compromise its predictive reliability and lead to cases of non-compliance with model validation standards (Chapter 5).
This confirms the importance of ensuring consistency between the observed discriminatory power and the implied default rates at grade level, especially during the calibration process.

In this chapter, we present the Accuracy Ratio (AR) resulting from the model after calibration under all approaches.
We will further illustrate the results obtained using statistical methods presented in Section 5.2.

11.1 Accuracy Ratio Results

Using the formula in Equation (4.1), we computed the Accuracy Ratio (AR) for each model.
First, we calculate the AR of the target model, i.e., the sample of mortgages with data from 2015 to 2022, denoted as AR_{15-22}.
Next, we compute the AR for the various models obtained using the methods presented in Chapter 9.

Finally, we evaluate the AR after calibration for each method to assess how calibration impacts the model's discriminatory power and the consistency of the PD curves across rating grades.

For clarity, we denote the AR values as follows:

- AR_{FAI_pre}, AR_{FAI_post}: Fully Averaged Inputs approach
- $AR_{RDD\_AF\_pre}$, $AR_{RDD\_AF\_post}$: Regression-Driven Defaults & Averaged Flows approach
- AR_{HYB_pre}, AR_{HYB_post}: Hybrid approach
- AR_{RDD_pre}, AR_{RDD_post}: Regression-Driven Defaults (only DR) approach
- AR_{DLS} : Direct Linear Scaling approach

These AR calculations were performed considering a variable number of rating grades in the rating scale, ranging from 7 to 13, to evaluate the sensitivity of the results to different granularities of the rating system.
The suffixes `pre` and `post` indicate values computed before and after calibration, respectively.

Rating	AR_{15-22}	AR_{DLS}	AR_{RDD_pre}	AR_{RDD_post}	AR_{FAI_pre}	AR_{FAI_post}	$AR_{RDD\_AF\_pre}$	$AR_{RDD\_AF\_post}$	AR_{HYB_pre}	AR_{HYB_post}
7	82.87	83.10	82.90	83.01	82.00	82.31	83.16	83.07	83.12	83.04
8	83.14	83.34	83.23	83.35	82.10	82.42	83.29	83.19	83.01	82.95
9	83.30	83.50	83.42	83.53	82.22	82.53	83.42	83.33	83.39	83.30
10	83.39	83.59	83.53	83.64	82.11	82.41	83.49	83.40	83.46	83.37
11	83.53	83.73	83.67	83.78	74.02	74.29	83.64	83.54	83.60	83.52
12	83.54	83.39	83.51	83.65	82.18	82.51	83.47	83.40	83.43	83.37
13	83.62	83.81	83.57	83.89	82.47	82.78	83.68	83.58	83.64	83.55

Table 11.1: Comparative overview of AR results across all methods for ratings 7–13

The table reports the results obtained by applying different calibration approaches. When applying direct linear scaling (DLS) to the sample of the period 15–22 in order to calibrate on the period 08–22, the AR increases for ratings 7 to 11, while it decreases for higher ratings. If only default rates (DR) are estimated, the AR systematically increases after calibration. A similar behavior is observed for the Fully Averaged Inputs (FAI) method, whereas for the other two methods the AR generally decreases after calibration.

In line with the EBA Guidelines, the key aspect is not whether the AR is slightly higher or lower after calibration, but rather that the difference between pre- and post-calibration AR values remains limited in absolute terms. To highlight this point, we report below a graph comparing the absolute differences between pre- and post-calibration AR values across methods.

Rating	$AR_{DLS} - AR_{15-22}$	$AR_{RDD\_pre} - AR_{15-22}$	$AR_{FAI\_pre} - AR_{15-22}$	$AR_{RDD\_AF\_pre} - AR_{15-22}$	$AR_{HYB\_pre} - AR_{15-22}$
7	0.23	0.03	-0.87	0.29	0.25
8	0.20	0.09	-1.04	0.15	-0.13
9	0.20	0.12	-1.08	0.12	0.09
10	0.20	0.14	-1.28	0.10	0.07
11	0.20	0.14	-9.51	0.11	0.07
12	- 0.15	-0.03	-1.36	-0.07	-0.11
13	0.19	-0.05	-1.15	0.06	0.02

Table 11.2: Differences between AR_{15-22} and pre-calibration values (and DLS) for ratings 7–13.

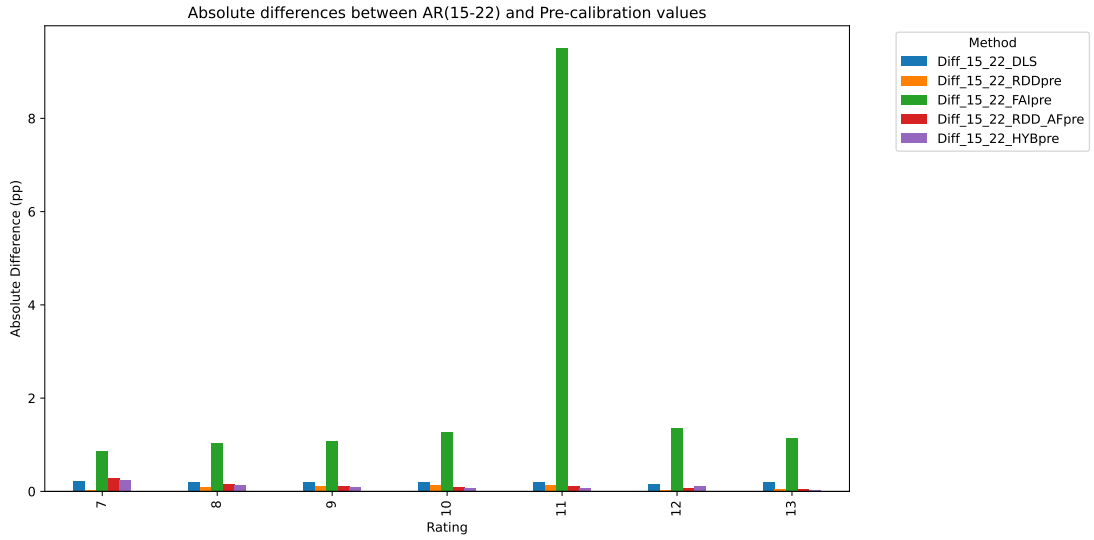

Figure 11.1: Absolute differences between AR(15-22) and Pre-calibration values

Observing Figure 11.1 and Table 11.2, we can note that estimating the distributions using all the averaged flows (FAI) produces the largest absolute differences in AR. The other methods show much smaller differences, generally below 0.30%. Among these, the largest differences are observed for the Direct Linear Scaling (DLS) method.
In the previous graph we looked at the absolute differences with respect to the AR of the 2015–2022 sample. Here, we focus on the differences between pre- and post-calibration

Rating	$AR_{RDD\_post} - AR_{RDD\_pre}$	$AR_{FAI\_post} - AR_{FAI\_pre}$	$AR_{RDD\_AF\_post} - AR_{RDD\_AF\_pre}$	$AR_{HYB\_post} - AR_{HYB\_pre}$
7	0.11	0.31	-0.09	-0.08
8	0.12	0.32	-0.10	-0.06
9	0.11	0.31	-0.09	-0.09
10	0.11	0.30	-0.09	-0.09
11	0.11	0.27	-0.10	-0.08
12	0.14	0.33	-0.07	-0.06
13	0.32	0.31	-0.10	-0.09

Table 11.3: Differences between pre and post calibration values for ratings 7–13.

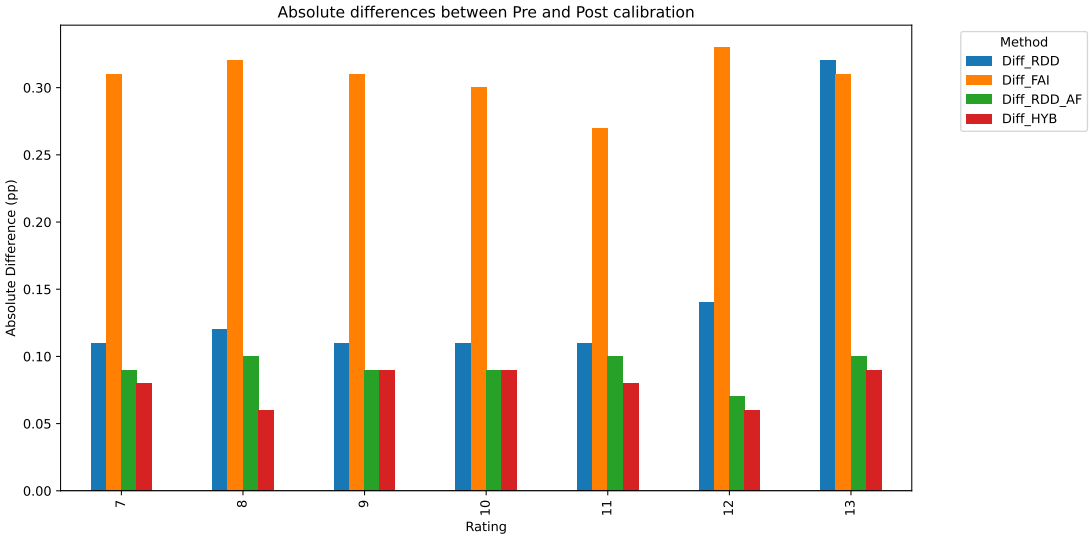

Figure 11.2: Absolute differences between Pre and Post calibration

values within the reconstructed samples, because from now on our reference is no longer the original 2015–2022 sample, but the reconstructed sample obtained using the various calibration approaches. From the results, we can observe that the FAI method continues to produce the largest absolute differences, followed by the RDD method, while the Regression-Driven Defaults with Averaged Flows Approach and the Hybrid approach show the smallest differences.

This observation highlights that, to comply with the EBA Guidelines, it is not sufficient to estimate only the default rates of the missing data in the calibration sample; it is also crucial to estimate the distribution of defaults and performing observations across the ratings.

In this context, all differences for the Regression-Driven Defaults with Averaged Flows and Hybrid methods are negative. Although this indicates that the AR decreases after calibration, it is not problematic in our credit risk application. Indeed, a lower AR implies that if the estimated data produce a lower long-run DR, the higher-risk classes will be populated correctly, or conversely, a negative delta is preferable because it prevents

over-populating the riskier classes. Thus, from a credit risk perspective, this behavior is actually desirable.

11.2 AR Analysis for Rating 13 Using Statistical Tools

In this section, we present the results of applying the statistical tools specifically to observations with rating 13.
We focus on the **Regression-Driven Defaults with Averaged Flows Approach**. Although the same statistical tools could be applied to other approaches, for simplicity and clarity we illustrate the results only for RDD_AF in this section.
The results are compared using the statistical tools introduced in5.2, highlighting the performance of the RDD_AF method relative to both the Direct Linear Scaling (DLS) and the standard RDD approaches.
For reference, we report in the following table all the Accuracy Rate (AR) measures for rating 13.

Method	**AR (%)**
AR_{15-22}	83.62
AR_{DLS}	83.81
RDD_pre	83.57
RDD_post	83.89
FAI_pre	82.47
FAI_post	82.78
RDD_AF_pre	83.68
RDD_AF_post	83.58
HYB_pre	83.64
HYB_post	83.55

Table 11.4: Accuracy Rates (AR) for rating 13 across all methods.

Model	CI Lower (%)	CI Upper (%)
AR_{15-22}	83.43	83.81
AR_{RDD_pre}	83.45	83.68
$AUC_{RDD\_AF\_pre}$	83.59	83.79
AR_{RDD_post}	83.78	84.00
$AUC_{RDD\_AF\_post}$	83.47	83.66
AR_{DLS}	83.72	83.91

Table 11.5: Confidence intervals (CI) for the different models.

11.2.1 Bootstrapping on the Accuracy Ratio (AR) Confidence Intervals for the Accuracy Ratio Using DeLong's

The tables 11.4 and 11.5 report the lower and upper bounds of the confidence intervals for each AR measure, obtained through bootstrapping with 1000 samples. We observe that, when considering only the DR estimation, the two intervals do not overlap. In contrast, applying linear scaling leads to a percentage overlap of 19.38%. The highest overlap, however, is achieved by the *Regression-Driven Defaults with Averaged Flows (RDD-AF)* approach.

Comparison	Percent Overlap (%)
$AR_{RDD\_AF\_pre}$ - $AR_{RDD\_AF\_post}$	20
AUC_{15-22} - AR_{DLS}	19.38
AUC_{RDD_pre} - AUC_{RDD_post}	0

Table 11.6: Percentage of overlap between model AUCs using DeLong's method.

11.2.2 Confidence Intervals for the Accuracy Ratio Using DeLong's

Model	AUC (%)	CI Lower (%)	CI Upper (%)
AUC_{15-22}	78.19	56.15	56.61
$AUC_{RDD\_AF\_pre}$	91.79	83.45	83.70
AUC_{RDD_pre}	91.94	83.76	84.01
AR_{DLS}	91.91	83.69	83.94

Table 11.7: AUC values with DeLong confidence intervals for different models.

Comparison	Percent Overlap (%)
$AUC_{15-22} - AUC_{RDD\_AF\_post}$	0
$AUC_{15-22} - AUC_{RDD\_pre}$	0
$AUC_{15-22} - AR_{DLS}$	0
$AUC_{RDD\_AF\_pre} - AUC_{RDD\_AF\_post}$	41.66
$AUC_{RDD\_pre} - AUC_{RDD\_post}$	0

Table 11.8: Percentage of overlap between model AUCs using DeLong's method.

Observing Tables 11.7 and 11.8, we note that all calibrated models achieve significantly higher AUC values compared to the AR_{15-22}.
The confidence intervals, computed using DeLong's method, are relatively narrow for all LS-based models, suggesting stability in the AUC estimates. Regarding the percentage overlap, a moderate overlap (41.66%) is observed between RDD_AF_post and RDD_AF_post, indicating some similarity in their performance.

To summarize, Table 11.9 reports all AR and AUC comparison results. For each pair of measures, we indicate the method used (Bootstrapping or DeLong) and whether their confidence intervals overlap () or not (). This provides a clear overview of which models show statistically significant differences and which exhibit overlapping performance.

AR Comparison	**Method**	**Overlap**
$AR_{RDD\_AF\_pre}$ - $AR_{RDD\_AF\_post}$	Bootstrapping	✓
AUC_{15-22} - AR_{DLS}	Bootstrapping	✓
AUC_{RDD_pre} - AUC_{RDD_post}	Bootstrapping	×
AUC_{15-22} - $AUC_{RDD\_AF\_post}$	DeLong	×
AUC_{15-22} - AUC_{RDD_pre}	DeLong	×
AUC_{15-22} - AR_{DLS}	DeLong	×
$AUC_{RDD\_AF\_pre}$ - $AUC_{RDD\_AF\_post}$	DeLong	✓
AUC_{RDD_pre} - AUC_{RDD_post}	DeLong	×

Table 11.9: Summary of AR/AUC comparisons, indicating the method used (Bootstrapping or DeLong) and whether the confidence intervals overlap (✓) or not (×).

As we can see, only the RDD-AF method passes the tests, showing overlapping confidence intervals with both pre- and post-calibration AR/AUC measures. All other methods exhibit non-overlapping intervals, indicating statistically significant differences.

Conclusions and Future Work

This thesis addressed a critical challenge in credit risk management: estimating long-term Probabilities of Default (PDs) for a retail mortgage portfolio in the presence of incomplete historical data. By combining detailed loan-level data from recent years with aggregated historical information, and by employing both statistical and dynamic reconstruction techniques, it was possible to generate a coherent dataset spanning the entire period of interest. This allowed for the calibration of PD models that are consistent with observed default rates while preserving the discriminatory power of the models, as measured by the Accuracy Ratio (AR).

The results demonstrate that reconstructing both default rates and the underlying distribution of counterparties significantly improves the consistency of the Accuracy Ratio compared to alternative pre- and post-calibration methods. This confirms the importance of carefully reconstructing historical dynamics, using migration matrices and dynamic equations, to maintain the reliability and regulatory compliance of PD models. The study thus provides a practical approach to bridging gaps in historical data, addressing a common issue faced by banks and supervisors alike.

Despite these achievements, several avenues remain open for further improvement. The assumptions made in this work—such as constant inflows and outflows of exposures, stable behavior of performing rating classes, regression-based default reconstruction, and a fixed definition of default—simplify the modeling process but limit the ability to capture more complex, real-world dynamics. Future research could refine these aspects by incorporating cyclical or economic variations in entries and exits, introducing macroeconomic or counterparty-specific factors into transition probabilities, employing advanced stochastic methods for default reconstruction, or considering evolving regulatory definitions of default.

Pursuing these extensions would allow the development of more flexible and adaptive models, capable of responding to structural changes in the portfolio and dynamic counterparty behaviors, while still maintaining robust predictive performance. In this sense, the thesis represents both a concrete solution to a practical problem and a foundation for further exploration, offering insights into the long-term calibration of PDs and the reconstruction of missing historical information in credit risk models.

Acknowledgments

Al termine di questo lavoro desidero ringraziare sentitamente la Prof.ssa Patrizia Semeraro per la disponibilità, il supporto e soprattutto per avermi dato l'opportunità di svolgere il tirocinio e la tesi presso l'Ufficio Credit Risk Accelerator di Intesa Sanpaolo.
È stata un'esperienza preziosa, che mi ha permesso di crescere molto sia dal punto di vista personale che professionale, facendomi scoprire da vicino il mondo del rischio di credito. Ringrazio tutto l'Ufficio per l'accoglienza, la disponibilità e il supporto dimostrati giorno dopo giorno.
Desidero inoltre ringraziare in particolare il dott. Francesco Grande e il dott. Nicolò Cugno per i preziosi consigli, l'aiuto offerto e l'importante contributo che ha reso possibile la realizzazione di questa tesi.
È stata per me un'esperienza professionale e personale di grande valore, che porterò con me nel mio percorso futuro.

Bibliography

Bart Baesens, Daniel Rösch, and Harald Scheule. *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*. Wiley, 2016. ISBN 9781119143987. doi: 10.1002/9781119449560.

European Banking Authority. *Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures*. November 2017.

European Central Bank. *ECB Guide to Internal Models*. July 2025.

Sheldon M.Ross. *Stochastic Processes*. Wiley, 1995.

Giulio Sironi and Andrea Resti. *Risk Management and Shareholders' Value in Banking*. Wiley, 2007.