BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

The in-house credit assessment system of Banca d'Italia

by Filippo Giovannelli, Alessandra Iannamorelli, Aviram Levy and Marco Orlandi

November 2020

Number 586

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

The in-house credit assessment system of Banca d'Italia

by Filippo Giovannelli, Alessandra Iannamorelli, Aviram Levy and Marco Orlandi

Number 586 – November 2020

ISSN 1972-6627 (print)
ISSN 1972-6643 (online)

Printed by the Printing and Publishing Division of the Bank of Italy

THE IN-HOUSE CREDIT ASSESSMENT SYSTEM OF BANCA D'ITALIA

by Filippo Giovannelli*, Alessandra Iannamorelli*,
Aviram Levy* and Marco Orlandi*

Abstract

Banca d'Italia's In-house Credit Assessment System (ICAS) is one of the sources for the valuation of collateral agreed upon within the Eurosystem's monetary policy framework. It helps to provide liquidity to those Italian banks that cannot rely on an internal model (IRB). Its role has become all the more important in the aftermath of the financial crisis relating to the COVID-19 pandemic of 2020. The paper first outlines the Eurosystem's collateral framework and describes Banca d'Italia's ICAS in terms of architecture and governance. It then presents in detail the underlying statistical model, including the definition of default adopted, and the validation process for the statistical model and for the expert system. The paper concludes by providing data on the amount of collateral pledged with an ICAS rating and on the main features, including the probabilities of default, of the Italian non-financial companies rated by the system.

JEL Classification: G32.
Keywords: collateral framework, credit risk.
DOI: 10.32057/0.QEF.2020.586

Contents

* Bank of Italy, Directorate Financial Risk Management.

Introduction[1]

Since 2012, Banca d'Italia has extended the range of eligible collateral to improve banks' access to monetary policy operations, with the ultimate goal of supporting the provision of bank credit to firms and households. One of the tools deployed for this purpose is an internal model for assessing the creditworthiness of Italy's non-financial companies, to be used in the context of the Eurosystem's collateral framework.

The Eurosystem's internal credit assessment systems (ICASs) play an important role in the conduct of monetary policy in ordinary times. Their contribution becomes even more important following financial crises and economic shocks, such as the COVID-19 pandemic of 2020, as they preserve the transmission mechanism of monetary policy. Euro-area banks have taken advantage of the wide range of measures adopted since March 2020 by the Governing Council of the European Central Bank (ECB) in response to the crisis. Banca d'Italia's ICAS (henceforth BI-ICAS) has allowed Italian banks to fully exploit one of these new measures, namely the easing of the collateral framework: by improving banks' collateral availability, this measure has countered one of the negative effects of the pandemic, the drying up of other sources of bank funding. BI-ICAS is particularly important for smaller banks, specialized in lending to small and medium-sized enterprises that do not have an Internal Based Model (IRB) for credit risk assessment.

In May 2020, the net value of bank loans pledged by Italian counterparties amounted to approximately €113 billion, €15 billion of which were assessed by BI-ICAS. This paper provides a detailed description of BI-ICAS.

ICASs represent one of three sources for the valuation of collateral currently agreed upon by the Eurosystem, the other two being the rating agencies and banks' IRBs.[2]

ICAS systems are also currently in use at the central banks of Austria, France, Germany, Ireland, Portugal, Slovenia and Spain. Since 2013, BI-ICAS has been instrumental in providing an important source of liquidity for Italian banks, whose loans to non-financial companies represent a large portion of banks' assets but are difficult to use as collateral because they are typically non-marketable and unrated.

[1] The authors wish to thank Gianluca Viggiano for his valuable help in data processing and Roberto Venturini for his helpful contribution to the sections on the system's validation. Patrizia Ferrauto, Giuseppe Grande, Franco Panfili, Tommaso Perez, Antonio Scalia and Stefano Siviero provided useful comments and suggestions.

[2] Until 2019, the Eurosystem Credit Assessment Framework included a fourth source of valuation, namely 'rating tools', i.e. statistical tools for measuring credit risk. In 2019, the Eurosystem decided to phase out this source of valuation and the two rating tools currently authorized in Italy will lose their status in March 2021.

The model calculates the borrowing firm's probability of default (PD) over a one-year horizon. The estimates are obtained by means of a statistical model, which draws on both National Credit Register (NCR) reports and balance sheet data; the output of the statistical model is then combined with an 'expert system', wherein the analysts take into account a range of supplementary information, such as sector risk and the quality of corporate governance.

The ICAS of Banca d'Italia generates ICAS ratings on a yearly basis (i.e. PDs based on both the statistical model and the expert system) for about 4,000 firms to which banks have granted potentially eligible loans. In addition, purely statistical PDs are available for around 300,000 non-financial firms. The degree of utilization of ICAS is demand driven, i.e. it relies on the initiative of banks.

This document is organized as follows: Section 1 outlines the Eurosystem framework within which the ICAS operates, Section 2 describes the architecture of the system, Section 3 provides details on the underlying statistical model and on the calibration approach, Section 4 describes the validation process for the statistical model and the expert system, Section 5 presents some data on the amount of collateral pledged with ICAS and on the characteristics and PDs of Italian non-financial companies rated by the system. Section 6 concludes.

1. The Eurosystem credit quality standards for access to Eurosystem refinancing

The eligibility requirements set by the Eurosystem for collateral are mainly aimed at mitigating the financial, legal and operational risks incurred when a national central bank provides liquidity to its bank counterparties. Minimum credit quality requirements are a key element of the eligibility criteria, both for marketable and non-marketable assets.[3]

The Eurosystem Credit Assessment Framework (ECAF) defines the Eurosystem's minimum credit quality requirements as well as the procedures, rules and techniques meant to ensure that the Eurosystem only accepts as collateral assets with high credit standards.[4]

Since the Eurosystem accepts a very broad range of marketable and non-marketable assets as collateral, it has to rely on various sources of credit assessment. The ECAF currently relies on three sources of credit assessment:[5]

- Credit rating agencies, formally defined as 'external credit assessment institutions' (ECAIs; 4 of them are currently authorized)
- In-house credit assessment systems (ICASs) developed and run by NCBs (7 systems are currently operating)
- Counterparties' internal ratings-based (IRB) systems used by banks (around 40 systems are currently authorized)

ECAIs are mainly used for assessing marketable collateral, whereas ICASs and IRB systems are mainly used for non-marketable collateral. In order to enhance its internal credit assessment capabilities, in recent years, the Eurosystem has encouraged the development of ICASs and their number has increased significantly.

The information provided by all these credit assessment systems has to be brought together in a harmonized way. To this end, the ECAF makes the credit ratings from all ECAF-accepted credit assessment systems comparable by mapping each of their rating grades in an appropriate credit quality step (CQS) within the Eurosystem's harmonized rating scale (see Table 1).[6] For instance, Table 1 shows that CQS3 includes triple-Bs: this CQS is equivalent to a probability of default of between 0.1 and 0.4 per cent over a one-year horizon and it currently represents the minimum credit quality requirement for the eligibility of all assets in the general framework.[7]

Once a credit assessment system has been authorized, The Eurosystem carries out due diligence

[3] Eligible non-marketable assets mainly encompass credit claims, which include bank loans (including shares of syndicated loans), some leasing and factoring credit claims, and drawn credit lines.

[4] https://www.ecb.europa.eu/paym/coll/risk/ecaf/html/index.en.html.

[5] See above, footnote.2.

[6] European Central Bank (2015), The financial risk management of the Eurosystem's monetary policy operations.

[7] Specific requirements apply to asset-backed securities (ABS) and retail mortgage-backed debt instruments (RMBDs).

Table 1: The Eurosystem's harmonized rating scale for ECAI

ECAI credit assessment		Credit quality steps (CQS)				
		1	2	3	4	5
Short-Term	DBRS Morningstar		R-1H, R-1M	R-1L, R-2H, R-2M, R2-L, R-3		
	FitchRatings		F1+	F1, F2, F3		
	Moody's		P-1	P-2, P-3		
	S&P's		A-1+, A-1	A-2, A-3		
Long-Term	DBRS Morningstar	AAA/AAH/AA/AAL	AH/A/AL	BBBH/BBB/BBBL	BBH	BB
	FitchRatings	AAA/AA+/AA/AA-	A+/A/A-	BBB+/BBB/BBB-	BB+	BB
	Moody's	Aaa/Aa1/Aa2/Aa3	A1/A2/A3	Baa1/Baa2/Baa3	Ba1	Ba2
	S&P's	AAA/AA+/AA/AA-	A+/A/A-	BBB+/BBB/BBB-	BB+	BB

Source: ECB

every year as part of its Performance monitoring procedure, which is designed in a way that allows the information derived from different sources to be compared. The key tool for the annual ECAF due diligence is known as the ECAF performance monitoring process, which consists of two components: (i) a quantitative statistical component, meant to check whether the system has accurately predicted default rates and, as a result, whether the mapping of the ratings of each credit assessment system in the Eurosystem's harmonized rating scale is still appropriate, and (ii) a qualitative component, which looks at credit assessment processes and methodologies.

Within the ECAF credit assessment systems, ICASs play a special role: on the one hand, they make it possible to reduce reliance on rating agencies, as was recommended by regulators after the great financial crisis. On the other hand they provide banks that do not have IRB systems with the incentives to grant loans to the non-financial sector and thus contribute to improving the transmission mechanism of monetary policy.

Besides the ordinary ECAF framework, a temporary Additional Credit Claims (ACC) framework has been in place since 2013 in order to improve the availability of collateral and overcome the liquidity shortages that followed the sovereign debt crisis of 2012. Within such a framework, banks that are monetary policy counterparties are allowed to pledge a wider range of credit claims as collateral, with lower credit requirements, in particular when those claims are pledged in a pool. The ICAS ratings are also widely used for the assessment of collateral within the ACC framework.

2. Banca d'Italia's In-house Credit Assessment System's architecture

2.1 Aims and governance of the system

As an ECAF-recognized credit assessment source, BI-ICAS provides important support to the conduct of monetary policy. The main aim of BI-ICAS is to allow counterparties in monetary policy operations to use a larger portion of their balance sheet assets as collateral; in addition, BI-ICAS also makes it possible to mobilize credit claims to banks that do not own or intend to use an IRB system.

The wealth of data and information on non-financial companies made available by BI-ICAS is also being used by Banca d'Italia for the purpose of financial stability analysis and occasionally for banking supervision purposes.

As regards disclosure, the assessment produced by BI-ICAS is confidential and is not disclosed to counterparties; however, in order to encourage banks to use the system, the list of eligible obligors of each counterparty is disclosed to the latter.

BI-ICAS is managed by the Financial Risk Management (FRM) Directorate, which is part of the Directorate General for Markets and Payment Systems. Within the Risk Management Directorate, the Credit Risk Assessment Division is in charge of the ICAS activity, which includes not only model development and producing ratings but also coordination and control tasks.

Since 2015, a number of Banca d'Italia's local branches[8] have been actively involved in producing ratings, on the one hand allowing them to better exploit their knowledge of the local economic context and on the other hand to increase the number of BI-ICAS assessments (see Section 5 below). Based on a rotation mechanism, each month, a list of firms to be assessed is allocated to analysts, mostly in local branches; a selected number of firms is assessed in the Central Administration.

2.2 Definition of default

In order to rate an entity, BI-ICAS has to merge reports from several banks. Therefore, a rule is needed for integrating different information and defining a (system-wide) state of default. The definition adopted by Banca d'Italia is based on national regulations. Based on this default definition, a borrower is considered in default if both of the following conditions are met:

[8] At the end of 2019, the following local branches had participated: Turin, Genoa, Milan, Venice, Bolzano, Trieste, Bologna, Florence, Ancona, L'Aquila, Rome, Naples, Bari, Catanzaro and Palermo.

1) the total amount of exposures reported as bad debt, unlikely to pay and past-due exposures by each bank is greater than 5 per cent of the total exposure of the borrower to the whole banking system (the materiality rule) and greater than €500;

2) the condition from the point 1) is verified for three consecutive months (the persistency rule).

Moreover, if a loss is reported by a bank to the National Credit Register (NCR), then no materiality and persistence threshold is applied and the firm is immediately considered as being in default.

The aforementioned definition of system-wide default is used for estimation, calibration and internal validation of BI-ICAS.

2.3 Input data and collection process

*Debtors' 'static' data***.** The system includes static (i.e. snapshot) data information on non-financial firms, such as name, identification number, tax number, legal form, location, sector of economic activity, registered office and so on.

*Credit behaviour data***.** The National Credit Register is a national archive of banks and financial intermediaries' debtors, and AnaCredit is its pan-European equivalent. Banks and other financial intermediaries are required to communicate to the central register a wide set of information about the financial liabilities of individual entities (e.g. companies, public entities and households). In return, banks have access to information on the debt exposure of their borrowers vis-à-vis the whole banking system.[9] The Credit Register is therefore the main data source for credit behaviour information to be used in the system. In particular, the ICAS system default status, compliant with ECAF rules, is determined for each firm in the AnaCredit and in the CR database.

*Financial statement data***.** For financial statement data, ICAS relies on Banca d'Italia's financial statements archive (Sistema informativo economico-finanziario, SIEF) based on data collected from Cerved Group.

Such data are available in two databases (Centrale dei Bilanci, also known as the CEBI, and Cerved). The CEBI collects about 80,000 financial statements per year from Italian medium-large corporations (all of them limited liability companies). These data are collected in part through banks participating in the CEBI program and for the rest through the Official Business Register. Cerved's database covers a very large portion of Italian small and medium-sized corporations:

[9] The NCR collects information about all credit relationships in the Italian banking system with a minimum size of €30,000, including the defaulted ones.

data are provided to Cerved by the National Official Business Register. Financial statement data are reclassified according to the CEBI reclassification accounting scheme, which can be applied to both national GAAP and IFRS financial statements. ICAS uses Cerved reclassified financial statements data as input and, as a consequence, comparability among the financial statements of different companies is safeguarded**.**

SIEF also has a 'Gruppi' (groups) archive, which provides information on the structure and the composition of the main Italian industrial groups. It covers around 500 of the most important groups, referring to listed companies or groups having a consolidated income above €250 million. Groups under foreign control are also included in the database whenever they have a significant Italian shareholder stake. The database is updated every month based on a detailed and regular analysis of annual and semi-annual reports, consolidated financial statements, balance sheet data of the subsidiaries and any other useful information.

Additional and qualitative data. In order to assess management quality, analysts can also rely on data from the Infocamere database, the Official Business Register of Italian Chambers of Commerce. It includes data on 6 million companies and individual firms, covering both company information (standing data, legal events) and corporate governance or staff information (officers, controllers, auditors, top managers).

Within the BI-ICAS production process, additional sources of information can be drawn on by the expert system. Such sources include the main financial news providers together with Banca d'Italia's articles and publications; in addition, the expert system also refers to rating agencies' research, for individual economic sectors and firms, and on the internal rating models (IRBs) of Italian banks.

3. The BI-ICAS process for calculating PDs

For classification purposes, the models for predicting default probabilities can be divided into three main macro-categories: judgmental assessment models, capital market-based models and credit scoring models. The latter category is the most heterogeneous and is generally divided in turn into two different groups, namely traditional credit scoring models and machine learning credit scoring models.[10]

Capital market-based models derive the probability of insolvency for an issuer starting from the listed prices of financial instruments such as shares, bonds and CDS. Market-based model usage has followed the development of financial markets and theoretical studies on the pricing of financial assets.[11] There are two macro-categories of capital market-based models: the reduced form models (Jarrow, Lando and Turnbull (1997); Duffie and Singleton (1999)) and the structural models (Merton (1974); Crosbie and Bohn (2002)).

Credit scoring models are commonly used for the prediction of defaults.[12] The conventional scoring models used by the banking industry (IRB models) can be divided into discriminant analysis models and regression models (linear, logit, probit). Machine learning scoring models relying on more advanced mathematical approaches, such as neural networks and genetic algorithms, have recently been added to the previous two.[13]

Discriminant analysis involves the assumption that independent variables may provide an indication of firms belonging to predefined groups (i.e. defaulted and non-defaulted), while regression models provide the means to obtain reliable estimates of output variables based on a set of predictor (independent) variables that are usually easier to measure. The most common regression model among IRBs is the logit, which relies on logistic regression to determine the conditional probability that a firm is insolvent, based on independent variables such as economic and financial indicators. In general, logistic regression is preferred over linear discriminant analysis, as the initial working hypotheses required are more easily satisfied.[14]

[10] It is worth underlining the difference between rating and scoring. The concept of rating is generally associated with a qualitative assessment in which the analyst's judgment is essential and which is not based exclusively on the processing of quantitative data. The scoring, on the other hand, generally indicates evaluation systems that derive their judgments mainly from an automatic quantitative analysis, therefore based on objective data.
[11] See Hillegeist, Keating, Cram and Lundstedt (2004).
[12] Scoring models can be defined as multivariate models that use the main economic-financial indices of a company as input, obtaining a numerical value, called a score, which represents the probability of default.
[13] A recent paper by Banca d'Italia (Moscatelli et al. (2019)) shows that the use of machine learning techniques can improve the performance of scoring models mainly when the latter only rely on publicly available information (such as non-financial companies' financial statements) whereas the value added of machine learning declines when the scoring models also rely on high quality, non-publicly available data, such as credit behavior indicators based on the Credit Register.
[14] See Ohlson (1980).

The BI-ICAS rating process is based on a two-stage procedure, which combines a statistical module assessment with a judgmental model (the Expert assessment, Figure 1).

Figure 1: BI-ICAS rating process

Stage 1 (Statistical Module) consists of a system of logit models that exploit two sets of variables: indicators derived from the NCR and indicators based on financial statements data.
The first set of indicators aims to describe the relationship of the firm with the banking system (credit behaviour). These variables are derived from information relating to the ratio of drawn to granted credit and the number and average amount of overdrafts.

The second set of variables is derived from financial statement information and aims to represent profitability, financial structure, debt service capacity, asset quality and dynamics and operating risk.

Model parameters are estimated using the observed defaults derived from NCR as a dependent variable. The model determines a one-year default probability and a rating level according to the master scale mapping.[15]

Stage 2 (Expert Assessment) involves the financial analysts' assessment through the use of a wider range of information sources as well as updated data and news. The analysis can either confirm the rating derived from Stage 1 or modify it by notching the master score up or down.

The assessment relies on information derived from financial statements (e.g. profitability, cash generation and financial structure) as well as on other types of information, such as the company's strategic position within the sector of activity, governance and management quality or geographical location. While all the ICASs developed within the Eurosystem are similar in their general characteristics and they all comply with Eurosystem requirements, some differences can be observed, either in the sources of information or in the methodology adopted in order to calculate the rating.

The Banque de France, for example, also collects qualitative information through direct contacts with firms' managers during the assessment process and by means of a monthly monitoring of the

[15] In the context of rating systems, two approaches can be adopted, one that includes cyclical effects and one that does not. The two approaches generate different rating types, commonly known as point-in-time (PIT) and through-the-cycle (TTC). PIT ratings aim at evaluating the current situation of an entity by taking into account both cyclical and permanent effects. In contrast, TTC ratings focus mainly on the permanent component of default risk and are essentially independent from cyclical changes in the entity's creditworthiness.

economic situation.[16] Within its ICAS, Banco d'España has developed different statistical models for non-financial companies and economic groups, based on their stand-alone and consolidated financial statements.[17] The Bundesbank and the Oesterreichische Nationalbank have jointly developed a Common Credit Assessment System (CoCAS)[18], which is used also by other Eurosystem NCBs for their credit assessment procedure.

3.1 The statistical model

3.1.1 Overview

The BI-ICAS statistical component predicts the probability of default (PD) over a one-year horizon for Italian non-financial companies having both an exposure to the banking system of at least €30,000 as reported in the NCR and financial statement data available. The module derives a point-in-time (PIT) PD estimate using data from the NCR, operated by Banca d'Italia, on one hand and from Banca d'Italia's financial statements database (SIEF) on the other hand.

The score obtained from the PD produced by the statistical component is then submitted to the Expert System module, through which financial analysts can either confirm or modify the rating attributed by the statistical module. Probabilities of default are estimated by relating ECAF-compliant default data (see Section 2.2 – for a Default Definition) with firm-level financial ratios and firm-bank level credit data.

The general architecture of the BI-ICAS statistical model consists of two independent components:

1) a *credit behaviour component*, namely a logit regression aimed at modelling data from the NCR;
2) a *financial component*, namely a logit regression based on yearly financial statement data, such as financial structure and profitability indicators.

The two components are then merged through a further logit regression. The potential correlation existing between some financial features and credit relationship variables can generate biased results in the event a variable selection is carried out simultaneously in the two sets of variables. In particular, some crucial financial variables would be overridden by credit variables, also due to their different predictive horizon, resulting in a weaker aggregate model performance. For these reasons, developing a model using financial and non-financial information together may not be the optimal choice. Moreover, the different frequency of data (monthly data for credit relationship

[16] Banque de France (2010) and Banque de France (2015).
[17] Banco d'España (2020).
[18] Deutsche Bundesbank (2015).

information and annual data for balance sheets) could generate additional distortions in the PD estimation.[19]

The components are estimated separately during model development.[20] In line with bank industry practices, such a model structure that keeps the two components of financial statements and credit behaviour separate is also quite common among Italian IRBs.

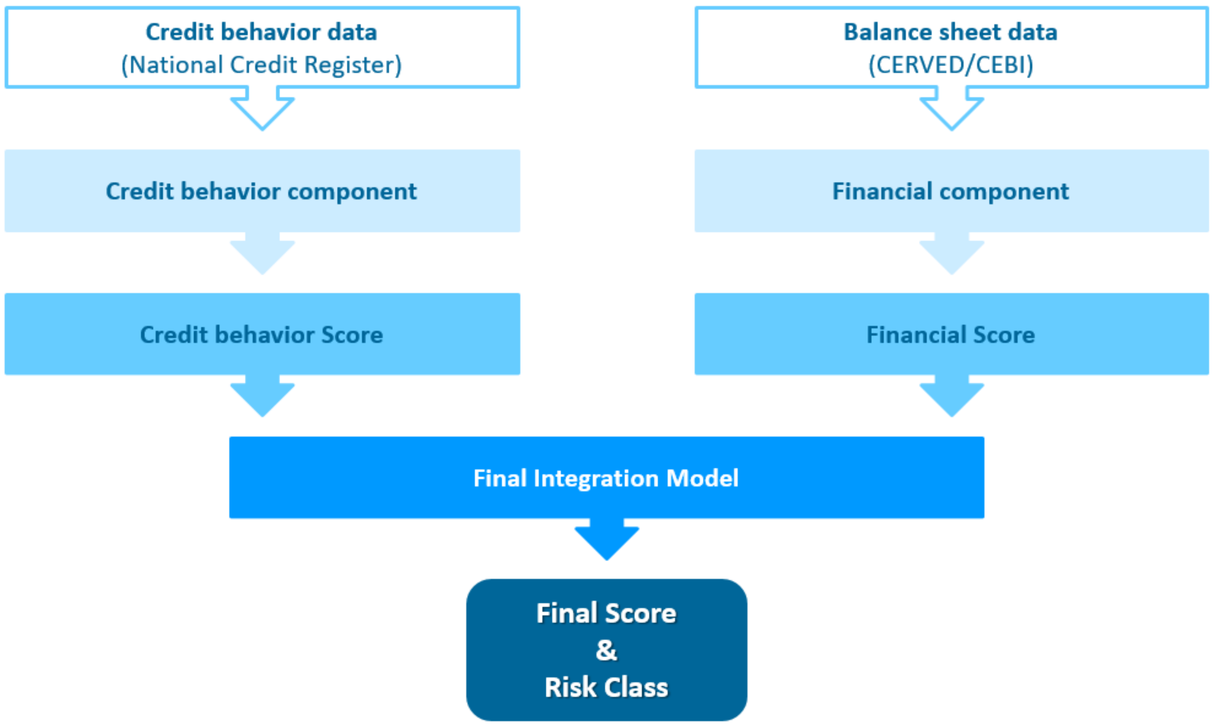
The absence of collinearity, i.e. the absence of a correlation between elementary credit behaviour and financial indicators included in the aggregate logit so that they independently predict the value of the PD, is verified before the integration of the two scores.

In order to map the score into a rating class, the final score is transformed into a PD via the inverse logit function:

$$PD = \frac{1}{(1 + e^{-Score})}$$

In light of the highly heterogeneous geographical distribution of observed default rate levels across the country, the final score is determined by including dummy variables for geographical areas in order to account for the relevant effects on a firm's credit quality. A scheme summarizing the structure of the model is presented in Figure 2.

Figure 2. Statistical model architecture



[19] Giannozzi et alii (2013).
[20] EBA (2017).

Three different models that exploit credit behaviour variables are set up for three different classes of firms, identified based on their size: small firms with a financial exposure below €300,000, medium firms with exposures between €300,000 and €20 million, and large firms with exposures exceeding €20 million.

The financial component instead consists of six different sub-models, in order to account for different classes of firms according to their industry sector. Each sub-model uses indicators tailored to capture credit risk for firms belonging to broad sectors corresponding to the structure of a firm's financial statements as reported in the financial statements database.[21]

Finally, obligors are allocated to different risk classes according to their estimated level of PD. Risk classes were empirically determined, analogously to rating agencies' current practices, according to historical default rates and frequency of cases and then mapped into the Eurosystem Credit Quality Steps.

3.1.2 Calibration approach and data

BI-ICAS is calibrated using a cross-sectional approach, i.e. default data and estimation data samples refer to different (two-year long) time periods.[22] Within a logit-modelling framework, this calibration approach implies that the (in-sample) average firm-level PD equals the default rate recorded in the calibration sample. The model is scheduled to be re-calibrated on a regular basis in order to keep its parameters updated with the latest information about the business and financial cycle phase as embedded into financial ratios and credit indicators.

In order to address the 'rare events' bias,[23] it is usually advisable to rely on downsampling in order to obtain a more balanced sample. In this case, we chose to use a sample containing an equal number of defaulted and non-defaulted firms, hence with a 50 per cent default rate.[24] Moreover, in downsampling, the sample of non-defaulters was chosen via a stratified sampling scheme in order to represent non-defaulters both in terms of their financial exposure to the financial system and in terms of the geographical distribution of Italian non-financial firms.

Credit behaviour and financial models were then rescaled before their integration into the final score. By doing so, the estimated PD reflects the actual default rates prevailing in each sector and credit-size calibration datasets. Figure 3 shows the rescaling step:

---

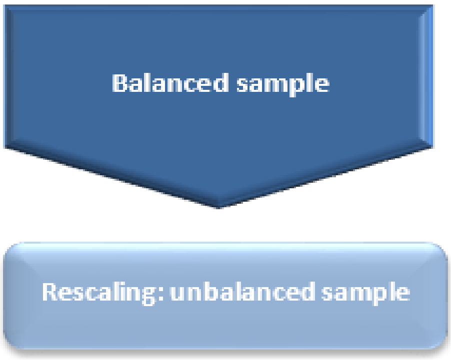
[21] Industrial, Trade, Construction, Services, Real estate and Holdings.

[22] For instance, in the case of 'Release 2017', the default data samples refer to the years 2015 and 2016 while the estimation data samples refer to, respectively, end-2014 and end-2015 for credit data and to 2013 and 2014 for financial data.

[23] The number of defaulted firms in a sample is usually small when compared with non-defaulters, leading to a reduced capacity to estimate significant relationships between firms' specific variables and the event of default.

[24] King and Zeng (2001).

Figure 3. Rescaling



For each sub-sample i:

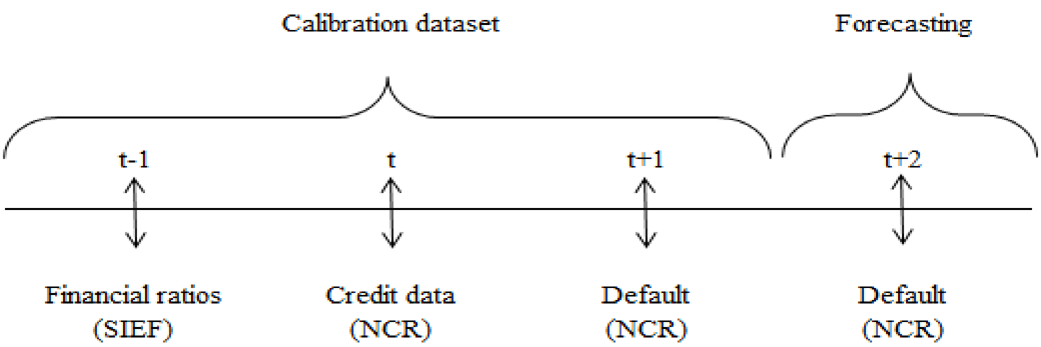
$$Balanced\ Score_i = \alpha + \sum_{j=1}^{n} \beta_{i,j}\ Indicator_{i,j}$$

$$Rescaled\ Score_i = \delta + \widehat{\beta}_i Balanced\ Score_i$$

The two data sources used to calibrate BI-ICAS, namely credit data and financial ratios, are available at a monthly and yearly frequency respectively.

Furthermore, financial ratios are typically available with a time lag of one year, while credit data are available with a monthly lag: the timeline of the data used to calibrate the model, as well as the forecasting horizon where actual PDs are estimated and used for credit risk assessment purposes, account for such a data mismatch, as shown in Figure 4.

Figure 4. Data timeline and forecasting horizon



Regarding the calibration sample, the entire population of Italian non-financial firms with both a credit exposure in the NCR and financial statement data available at the target date was used: this population includes about 250,000 firms for each of the two years of the calibration period, amounting to around 500,000 firms.

3.2 The expert assessment

Overall, the expert assessment aims at revising the statistical assessment through a qualitative analysis based on the key credit risk drivers. The expert assessment is applied to a small fraction of those non-financial firms for which a statistical assessment is available (around 4,000 out of 300,000 firms at the end of 2019). These firms are selected by Banca d'Italia among the borrowers which, according to the National Credit Register data, have received the largest outstanding loans from banks that either do not have an Internal Based Model (IRB) for credit risk assessment or, if they have such a model, they do not intend to use it for other purposes – such as monetary policy – than the supervisory ones.

The expert assessment starts from the 'automatic' PD and proceeds by following different steps of analysis. It involves the analyst's assessment by means of a wider range of information sources as well as recent data and news. The assessment relies in particular on information derived from balance sheets (e.g. profitability, cash generation, debt service coverage ratio and financial equilibrium) as well as on other types of information such as the strategic position in the company's sector of activity, the governance and management quality or the geographical distribution of sales. The analysis can either confirm the rating derived from the statistical stage or modify it by notching the master score up or down.

The analyst's assessment takes into account the different profiles and determines a risk score independently for each profile. Such score will express an opinion on whether the data considered improve, confirm or worsen the risk assessment produced by the statistical module.[25]

The assessment starts with a review of static data, in order to get an initial overview of the firm and its main features (size, sector of activity, age, financial group structure, geographical location and so on). This information does not produce a specific assessment but represents the necessary background driver for the subsequent assessment. Then the automatic rating is assessed including its components (the Credit Register data PD, the financial data PD) as well as the variables that have been used in the respective models, in order to get an initial indication of the position of the company compared with the general population or peer groups (by sector, geographical area or

[25] The score could assume three to four different levels for each profile (positive, neutral, negative or very negative).

risk class).[26] In particular, the analyst compares the signal coming from Credit Register with the one from financial information. Then, the last 4 years of available balance sheet data are taken into account to perform an analysis of the balance sheet ratios and a peer group analysis. Information on the firm is compared with the average and percentile value for the same sector or peer group. In this regard, data, trends and balance-sheet ratios according to four different profiles are considered: a) profitability; b) financial structure and debt sustainability; c) liquidity and cash generation; and d) growth indicators.[27] At this stage, analysts are also required to read and assess the full financial statement of the company.

Furthermore, the information available relating to payment behaviour and access to external financing is considered in order to perform an analysis of financial flexibility. The relevant data are derived from the National Credit Register and concern the relationship between the assessed firm and the banking system.[28] In the next stage, the management's quality and corporate governance are examined, based on the principle that i) a transparent and fair governance structure has a positive impact on company performance and ii) firms with strong governance systems tend to outperform peers.[29]

According to best practices, a more advanced corporate structure fosters internal control and consistency in the management strategy and execution. An analysis of the economic environment, the industrial sector and the geographic location where the company primarily operates is then performed in order to assess the contribution of these factors to the company's credit profile. Then the focus moves to an analysis of the group: if the assessed company is part of a group, as either a parent or a controlled firm, an analysis of the group's influence on its credit profile is performed.

If the group is judged to be a risky one, this can lead to a notching down of the company. A positive contribution to the firm's credit profile can arise from explicit forms of financial guarantee from a stronger parent company or other firms in the group. As a general rule, the group's rating represents a ceiling for the individual rating. Finally, third party opinions and collections of recent news from media, press and internet are reviewed. This potentially includes

[26] For all data, the relative quintile is calculated both on the total population of companies assessed by the statistical module and the subsample composed of all the firms in the same sector, geographical area and size class.

[27] The main balance sheet indicators are included in the statistical module and hence already reflected in the automatic rating. However, at this stage, it is possible to consider a larger range of indicators and to look at balance sheet data in a more comprehensive manner.

[28] NCR information is already used in the statistical module. However, a more detailed inspection of the relationship between the firm and each single bank has been considered useful at this stage. This information is relevant because in the Italian system, firms tend to turn to multiple banks for financing needs. The change in the number of banking relationships and in the use of granted credit lines is considered relevant information for spotting possible tensions in the financing lines and for assessing the availability of financing sources.

[29] The general quality and track record of the management are analysed on the basis of information collected from the Italian Business register database, which includes around 6 million companies and individual firms, covering both company information (standing data, legal events) and corporate governance and staff information (officers, controllers, auditors, top managers). From the same database, information on compliance with best practices in terms of corporate governance can be retrieved.

analyses carried out by other financial institutions or other external sources. In addition, every recent news item that could have an impact on the firm's assessment and that is not yet reflected in other sources is collected. If news that could significantly affect the overall assessment of the firm emerges, it will be taken into due consideration, in some cases leading to a downward revision of the assessment or to its suspension.

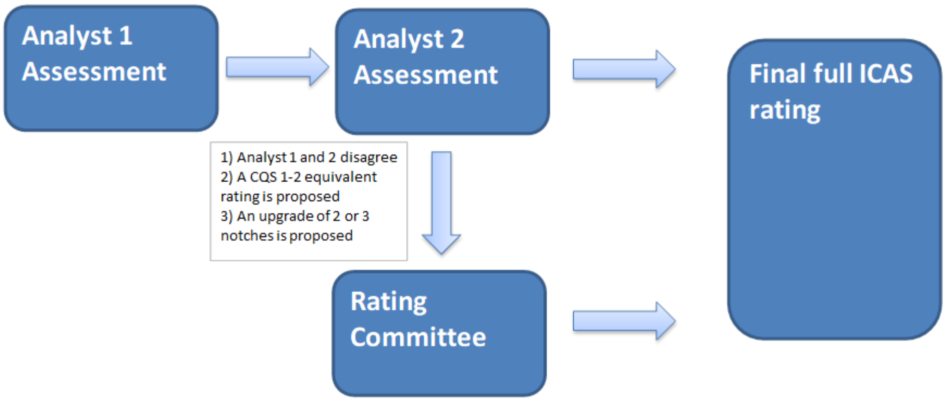
In order to reach a high level of consistency and reliability, the expert assessment is based on a template that guides the analysts through a number of predefined steps associated with each profile, considering specific information and producing an evaluation based on a set of predetermined rules. On top of the analysis template, analysts are allowed to consider whatever other information is deemed relevant and to evaluate it in qualitative terms. For each of the abovementioned steps, an assessment is conducted. Partial scores resulting from each step are weighted and aggregated in order to produce a final grade. A decision matrix translating the final grade into the rating decision (upgrade, confirm or downgrade the automatic rating) provides a non-binding guideline to analysts for taking the final decision. For each firm, an independent assessment by at least two analysts is required.

The final assessment is upper-bounded: analysts can lower the assessment as much as they want, while they can raise the final rating only by one notch. If they want to notch up the automatic assessment by more than one level, they are required to propose the upgrade to the ICAS Rating Committee.

The Rating Committee is composed of senior management representatives of the Directorate "Financial Risks Management" and the analysts involved in the assessment are required to attend the meetings. The Rating Committee normally meets with a monthly frequency. More frequent meetings are held in the event of urgent matters to be decided. For decisions regarding selected straightforward cases, written procedures are adopted.

The last step of the process is a final report containing all the relevant details about the firm's assessments. In the final report, analysts are also required to briefly describe the main motivations of the assessment. Each credit assessment decision has to be properly documented in all its different stages.

Figure 5. Decision process



Once the final rating decision has been taken either by analysts or by the Rating Committee, the final rating is recorded in the ICAS database and normally remains valid for the next 12 months, unless new relevant information becomes available.

4. The model validation

4.1 The role of the validation

The best practices in the field of credit risk assessment recommend the separation of the validation activity from model development and from rating production. This separation guarantees the independence and impartiality of validators. In Banca d'Italia, the responsibility for validation has been assigned to a separate unit since 2014.

The validation activity usually includes both the internal validation and the monitoring process.

The purpose of internal validation is to check that the rating attribution process is carried out following methodologies and procedures consistent with best practices; furthermore, internal validation verifies that the rating model, before being used, is adequately robust and efficient. This latter activity is performed each time a new version of the model is adopted or whenever significant innovations are introduced (initial validation) and it consists in checking the whole structure of the model, both in the statistical component and in the expert system component.

The goal of performance monitoring is to assess the stability of the rating system over time and to verify the predictive ability of the model over the course of the year.

Since BI-ICAS is focused strictly on the PD calculation, the internal validation does not include the examination of the Loss Given Default (LGD) and of the Exposure At Default (EAD).

The validation and monitoring of the PDs can be assessed with, among other tools, benchmarking and backtesting.

Benchmarking techniques compare the PDs calculated by the model under assessment with the ones calculated by other models on the same portfolio of firms. These techniques consist of calculating a statistical distance between the models. The main drawback of this approach is the limited availability of ratings calculated by other models, since the ICAS typically assesses small and medium-sized enterprises (SMEs) that are not assessed by other systems. Furthermore, benchmarking requires strong confidence in the rating system used for the comparison, since it is taken for granted that its PDs are a good reference for this kind of analysis.

Backtesting procedures are based on a comparison between the ratings calculated ex-ante and the number of defaults observed ex-post. The scientific literature provides several statistical tests, which help the validator in understanding how accurate a model is in predicting the PDs correctly. Within the validation framework, two main aspects are typically taken into account: the discriminatory power and the predictive power. The first one measures the ability of the model to distinguish the rated entities according to their future status (defaulted or not defaulted) at a predefined time horizon. The second one compares the number of defaults that actually occurred in a certain rating class with the number of defaults that are predicted by the model. Discriminatory and predictive power analysis are described in more detail in Annex 1.

4.2 Expert system validation

Validating the Expert System is a major challenge, since the goal is to perform a qualitative analysis using quantitative measures. After having calculated the statistical PD, the ICAS system produces a final rating, obtained by combining the results of the quantitative step with a qualitative analysis, performed by the Expert System, which relies on the work of a dedicated team of analysts.

The team has to consider eight different profiles in its assessment. Two profiles are just informative for the analyst and do not require a score. For the other six profiles, the analysts assign a score (‘very bad’, ‘bad’, ‘neutral’, ‘good’). The profiles are: 1) balance sheet ratios and peer group analysis, 2) financial flexibility, 3) quality of the management and corporate governance, 4) industrial sector, geographic location and economic environment, 5) group analysis and 6) third party opinions and other recent information. The goals of the validation of the expert system are:

1. To understand how analysts use the different profiles of the expert system.
2. To assess how the final judgment is influenced by the different profiles of analysis.
3. To explore the influence of the second analyst on the assessment proposed by the first analyst.
4. To study the effect of the expert system on the risk classes attributed by the statistical system.
5. To understand if the assessment of the analyst is able to anticipate the evolution of the risk class of the companies which can be observed one year later.
6. To understand the analysts’ behaviour in assigning the final ratings and the scores to the profiles.
7. To backtest the ICAS ratings.

The purpose of this step of the validation is to understand how active the analysts are in attributing the scores, i.e. which scores among the neutral, positive and negative ones are the most frequently assigned. It can also be of interest to compare the behaviour of the analysts over time, analysing how often and by how much the analysts’ proactiveness changes.

A detailed analysis of the six profiles assessed by the validation of the expert system is presented in Annex 2.

4.3 Banca d'Italia's internal backtesting analysis: results for 2019

As already mentioned in section 4.1, the validation analysis which is carried out by a separate unit of Banca d'Italia includes a backtesting procedure based on a comparison between the probabilities of default estimated ex-ante by BI-ICAS and the actual default rates observed ex-post.

More in detail, the pool of obligors assessed by the system at a certain date is split among different risk classes based on their estimated PD level (the 'static pool'). For instance, one can rely on the risk classes (Credit Quality Steps or CQS) defined by the Eurosystem in its credit assessment framework. Then for each class, the firms that have defaulted in the 12 months following the date of creation of the static pool are counted. The default rate is obtained by calculating the ratio of defaulted firms to total firms and is compared with the expected default rate in each class. If the expected and observed default rates are close enough, the outcome of the test is positive and the model has performed well. In the case of BI-ICAS, backtesting is carried out separately on the Statistical Module and then on the full BI-ICAS (Statistical module and Expert System) in order to verify whether the Expert System is able to improve the Statistical Module's predictive power.

As an illustration, the following are the results of the exercise carried out in early 2020 as part of the monitoring process of the BI-ICAS predictive power with reference to their annual PDs, estimated on 31 October 2018 (for the Statistical Module) and on 1 January 2019 (for the full BI-ICAS). The expected defaults implied in these PDs were compared with the effective defaults observed during the subsequent year.

The results of the BI-ICAS backtesting for 2019 are positive.[30] First, the predictive power analysis for the statistical module (see Table 2) shows that actual (observed) defaults are below the upper bound of those expected by the model for each CQS.[31]

[30] There are in principle two ways of reading the results of a backtesting test (see also Annex 1). One way is to adopt the risk aversion profile of a '"supervisor' and make sure there is no underestimation of the PDs in the various rating classes, estimated by the whole ICAS system by adopting a risk aversion profile for default risk typical of a supervisor). Another possibility is to adopt the approach of a 'production unit' with business objectives and check whether the whole ICAS system is too conservative or too loose. In a way, the BI-ICAS' internal validation combines both approaches.

[31] In this study, reference is made only to the upper bound of each CQS, but in principle each CQS consists of both an upper and a lower bound of expected default rates against which the actual defaults should be compared: for instance, for CQS3, the range of default rates is 0.1 to 0.4 per cent and, as a consequence, in Table 2 the lower bound is 73 defaults and the upper bound is 293.

Table 2. Backtesting for the ICAS Statistical Module

(PDs as of 31 October 2018; defaults in the following 12 months. Source: BI-ICAS database)

CQS	**CQS firms numbers**	**expected defaults (upper bound)**	**observed defaults**
1&2	7412	7	4
3	73128	293	164
4	66339	663	457
5	33272	499	369
6	51562	1547	1216
7	27457	1373	1345

Source: Banca d'Italia's own calculations based on BI-ICAS data

Similarly, the results of the predictive power analysis for the full ICAS (considering the Statistical and the Expert modules together) are very good: Table 3 shows that also in this case the number of actual defaults is below the upper bound of expected ones.

Table 3. Backtesting for the Expert System of the ICAS

(PDs at 01.01.2019; defaults in the following 12 months)

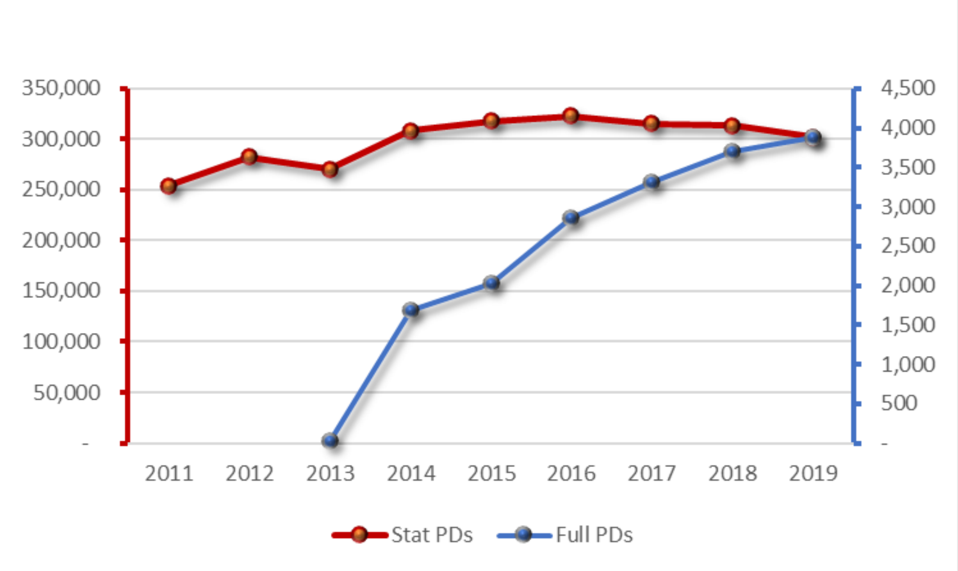
CQS	**Number of firms per CQS**	**Expected defaults (upper bound)**	**Observed defaults**
1&2	92	0	0
3	675	3	0
4	654	7	1
5	385	6	1
6	278	8	4
7	188	9	6

Source: Banca d'Italia's own calculations based on BI-ICAS data

5. Some data on the BI-ICAS system: coverage and rating distribution

The BI-ICAS production has been increasing over time, reaching over 300,000 automatic ratings (i.e. purely statistical PDs) at the end of 2019. As regards the full ratings (i.e. ratings including an expert assessment), their number increased from 1,700 ratings in 2014 to almost 3,900 in 2019 (Figure 6).

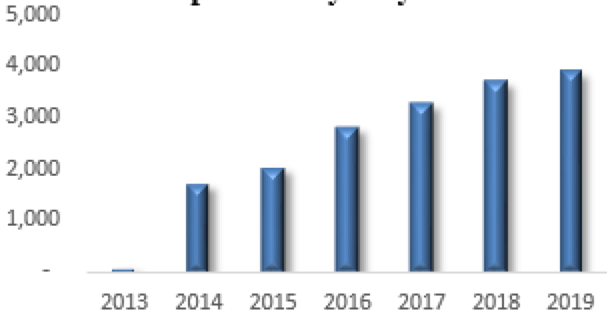
Figure 6. BI-ICAS full ratings and automatic PDs production over time



Source: BI-ICAS database

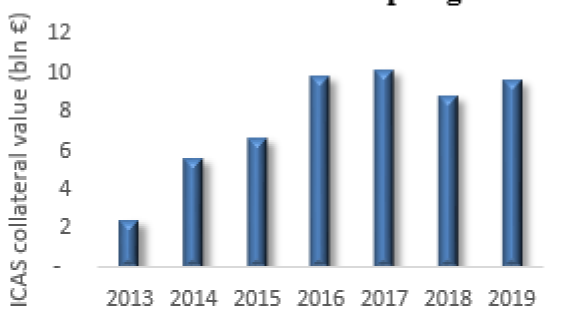
The production of the ICAS ratings is aimed at allowing banks (as monetary policy counterparties) to use an ICAS assessment in order to access monetary policy refinancing. In addition, statistical model ratings provide a virtually complete coverage of the Italian non-financial firms registered as limited liability corporations, and such ratings are also being used for financial stability analysis and research. While the number of BI-ICAS ratings has increased almost linearly over time, in terms of the nominal value of collateral pledged by banks relying on BI-ICAS, the peak was reached in 2017 (Figures 7.a and 7.b), also reflecting the collateral choices and the financing needs of counterparties.

Figure 7a. Number of full ratings produced yearly



Source: BI-ICAS database

Figure 7b. Security value of BI-ICAS assessed collateral pledged[1]



[1]After haircut end-of-period value

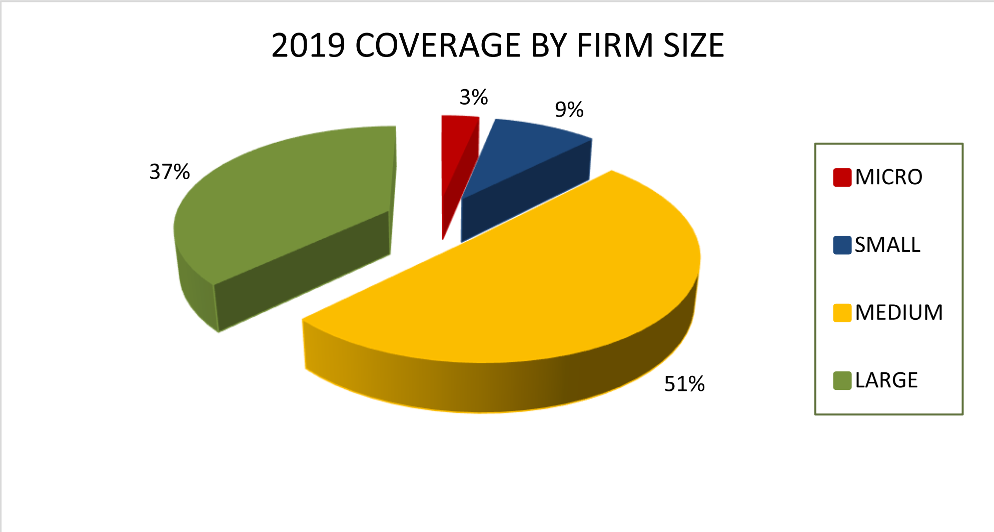
Source: ABACO database

5.1 The universe of companies assessed by BI-ICAS through a full rating

The sample of non-financial companies for which a full rating is available provides a picture of the non-marketable collateral, namely Italian firms' credit claims, pledged by monetary policy bank counterparties that do not have an internal based system (IRB).

Around half of the universe of debtor firms assessed by BI-ICAS with a full rating are medium-sized companies, while a smaller share (37 per cent) is represented by large companies.[32] The remaining portion is composed of small (9 per cent) and micro size (3 per cent) firms (Figure 8).

Figure 8. Rating distribution across firm size in 2019



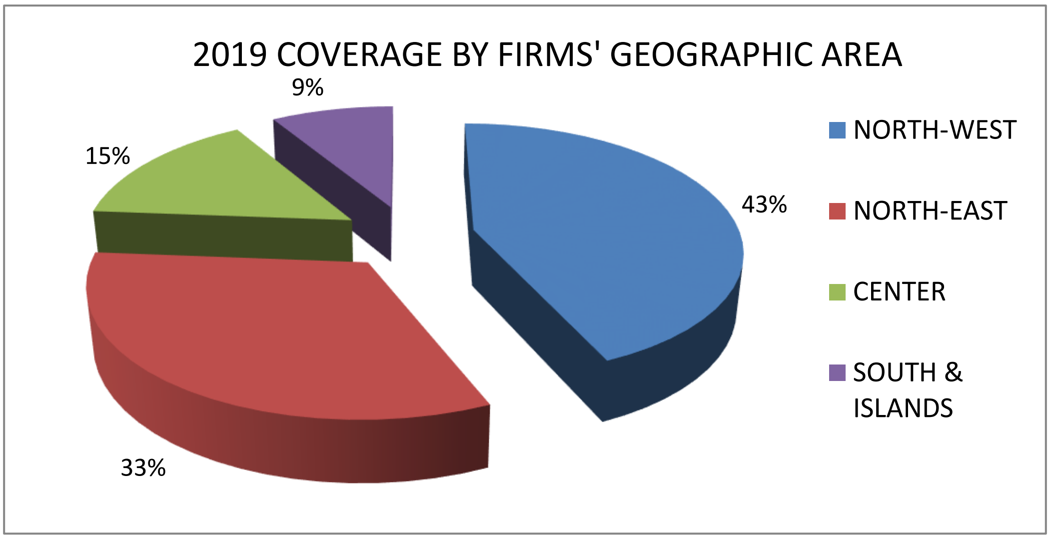
Source: BI-ICAS database

The distribution of firm size over time has remained virtually stable, with the exception of an increase in the full rating production for medium-sized firms.

As regards geographical distribution, less than half (43 per cent) of the overall full rating production refers to companies located in the North-West of Italy, while a slightly lower share of ratings (33 per cent) belongs to companies located in the North-East. The remaining firms are located in Central Italy (15 per cent) and in the South & Islands (9 per cent) (Figure 9).

[32] For firm size, the European Commission definition is used, referring to staff headcount and either turnover or balance sheet total. Based on such a definition, a firm is medium-sized if the staff headcount is fewer than 250 people, turnover is lower or equal to €50 million or the balance sheet total is lower or equal to €43 million; similarly, a firm is small if staff headcount is fewer than 50 people, turnover is lower or equal to €10 million or the balance sheet total is lower or equal to €10 million.

Figure 9. Rating distribution by firms' geographic area in 2019



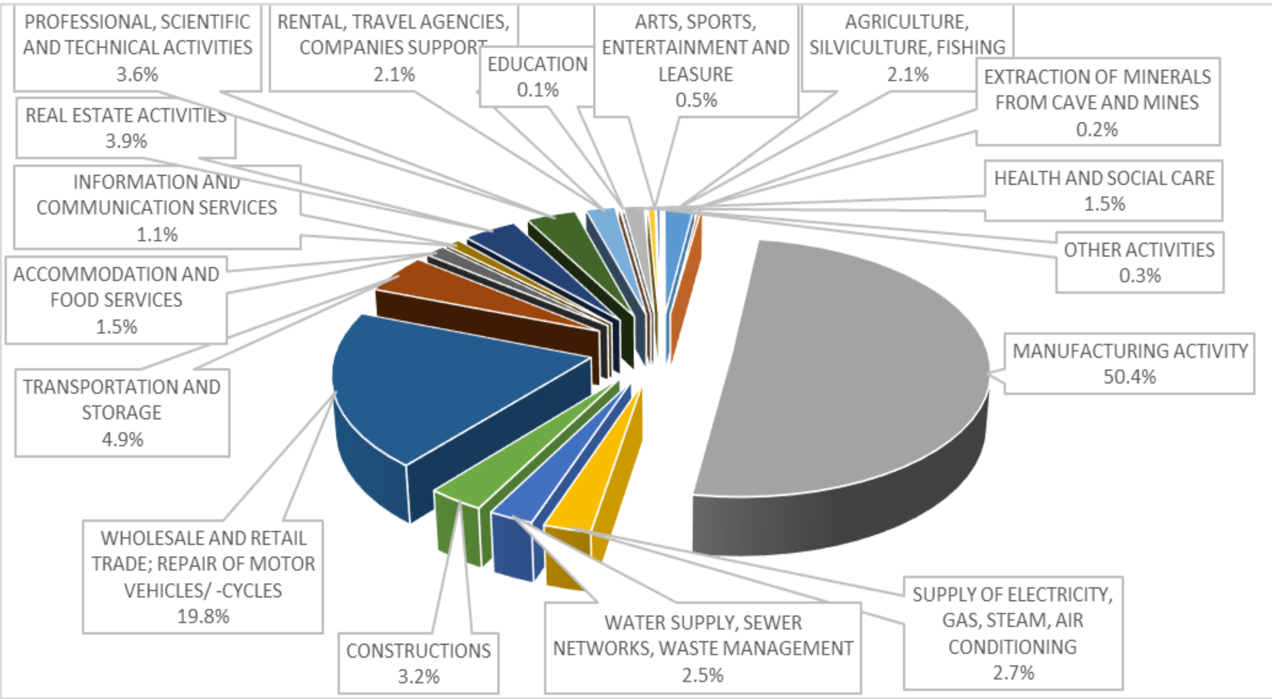
Source: BI-ICAS database

As in the case of firm size, the rating distribution by region has been almost stable over time, with a slight increase of firm ratings in the North-West compared with those in the North-East.

From an industry perspective, around half of the ratings refer to manufacturing companies (50.4 per cent), while the second biggest group is wholesale and retail companies (19.8 per cent) (Figure 10).

The most significant changes in industry coverage over the BI-ICAS life have been the decrease in the manufacturing firms' share and the increase in the wholesale share.

Figure 10. Rating distribution by sector of economic activity in 2019



Source: BI- ICAS database

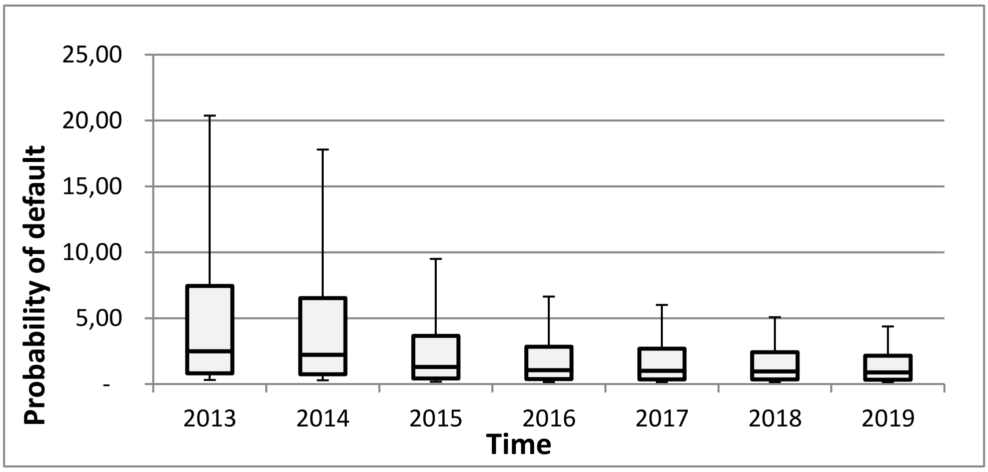
5.2 The evolution of credit risk in the ICAS universe of firms

BI-ICAS' wide coverage of the universe of Italian non-financial firms, through an automatic rating (or purely statistical PD), makes it possible to track the evolution over time of firms' financial situations, also at an aggregate level.

Based on ICAS data, an improvement in firms' financial situations can be observed starting from 2013 onwards. The negative effect of the still fragile cyclical conditions on firms' ability to repay the debts they owe to the financial system was partly offset by the rebalancing of the capital structure carried out in recent years and by still low interest rate levels.

Based on the ICAS of Banca d'Italia, the median one-year-ahead default probability fell from 2.5 per cent in 2013 to less than 1.0 percent in 2019 (see Figure 11). The marked decrease in the values corresponding to the 75th and 90th percentiles (respectively, the upper ends and the top of the vertical lines of each box in Figure. 11) of the sample distribution indicates that the improvement also extended to weaker firms.

Figure 11. Distribution of the probability of default *(per cent)*

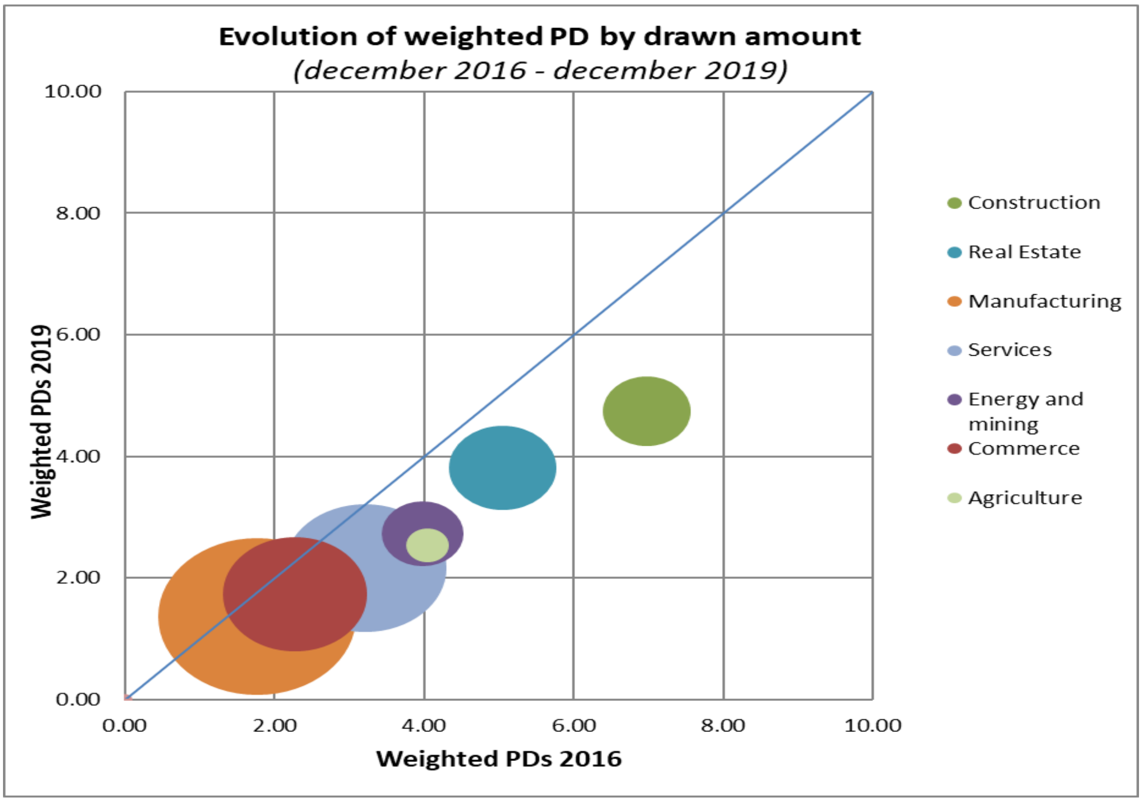


The graph shows from top the 90th, 75th, 50th, 25th and 10th percentiles of the distribution for each year.

Source: BI-ICAS database

In 2019, as a consequence of five years of economic growth and historically very low interest rates, solvency ratios improved. According to BI-ICAS data, between 2016 and 2019, non-financial firms' probability of default decreased in all the main economic sectors (Figure 12).

Figure 12. Probability of default by sector of economic activity *(weighted averages; per cent)*



The size of the circles corresponds to the amount of loans granted to firms in each sector.

Source: BI-ICAS database

5.3 The contribution of BI-ICAS to monetary policy refinancing in the COVID-19 crisis

In ordinary economic conditions, the BI-ICAS ratings allow banks that do not have an IRB system to increase their access to Eurosystem refinancing. This is particularly true for smaller banks, specialized in lending to small and medium-sized enterprises. The number of bank counterparties using BI-ICAS ratings for pledging their bank loans has been increasing over time: from fewer than 20 banks at inception in 2013, to more than 30 in 2019. The net value of bank loans pledged as collateral with a BI-ICAS rating increased as well over the same period, rising from €2.5 to around €9 billion (see above, Figure 7b). In terms of the sheer size of the 30 banks that make use of BI-ICAS, at the end of 2019, the total value of refinancing granted to these banks (including the liquidity received against marketable assets, relying on other credit risk assessment sources such as rating agencies) amounted to €64 billion.

The role of BI-ICAS has become even more relevant since the financial crisis triggered by the COVID-19 pandemic of 2020, by helping to preserve the transmission mechanism of monetary policy. Following the easing of national Additional Credit Claims (ACC)[33] collateral

[33] Additional credit claims are those bank loans that do not fulfil the ordinary framework's eligibility requirements but satisfy the wider criteria set by each national central bank, which bears the related financial risks.

frameworks approved by the ECB Governing Council in April 2020 and the ensuing measures adopted by Banca d'Italia in order to enlarge collateral availability, the net value of bank loans pledged as collateral with a BI-ICAS rating reached €16 billion in June 2020, compared with €9 billion at the end of 2019. Roughly two thirds of this increase are the effect of the reduction of haircuts decided by the ECB, whereas the remaining part is the result of the Banca d'Italia's targeted measures aimed at enlarging collateral availability (see Banca d'Italia (2020)). Despite this increase, the potential amount of eligible loans that may be assessed by ICAS is still much larger than the amount that is currently pledged.

These targeted measures include an easing of the eligibility criteria for pledging pools of bank loans to non-financial firms and households: among other things, maximum PD thresholds have been raised and smaller non-financial firms have been added to the eligible universe. In addition, while the credit risk of these pools of bank loans has always been assessed with BI-ICAS, the new measures include an enlargement and a more flexible use of this system.

6. Conclusions

Over the last decade, Banca d'Italia has made major efforts to broaden the range of eligible collateral and improve banks' access to monetary policy refinancing, with a view to supporting bank lending to firms and households.

An important tool for this purpose is Banca d'Italia's In-House Credit Assessment System, which has allowed Italian banks, in particular smaller ones, which do not have an internal model for assessing credit risk (IRB), to increase the available collateral by pledging loans granted to small and medium-sized enterprises.

This tool has proven particularly valuable since the Italian economy was hit by the severe economic shock caused by the COVID-19 pandemic. In this context, BI-ICAS has allowed Italian banks to take full advantage of one of the measures adopted in spring 2020 by the Eurosystem and, by increasing the availability and value of collateral, has been able to support the provision of bank credit to firms, in particular small and medium-sized ones.

References

1. Banca d'Italia (2020), Le misure di espansione delle attività a garanzia delle operazioni di politica monetaria dell'eurosistema, *Note Covid-19*, 10 giugno https://www.bancaditalia.it/media/notizie/2020/Nota-Covid-collaterale-10062020.pdf
1. Banco d'España (2020), The Banco de España in-house credit assessment system, *Financial stability review*, issue 38, https://repositorio.bde.es/bitstream/123456789/13555/1/Banco_Espana_in_house_credit.pdf
2. Banque de France (2010), Le système de cotation de la Banque de France: un atout pour la Banque centrale et un outil pour les banques, *Bulletin de la Banque de France* n.180, p.69, https://www.banque-france.fr/fileadmin/user_upload/banque_de_france/archipel/publications/bdf_bm/bdf_bm_2010/bdf_bm_180.pdf
3. Banque de France (2015), Le système de cotation de la Banque de France https://hautsdefrance.cci.fr/wp-content/uploads/sites/6/2015/12/A4-LES-DESSOUS-DE-LA-COTATION-BANQUE-DE-FRANCE-CCICO-19112015.pdf
4. Crosbie, P.J. and J. R. Bohn (2002), Modeling default risk, *Technical Report, KMV*, LLC
5. Deutsche Bundesbank (2015), The Common Credit Assessment System for assessing the eligibility of enterprises, *Monthly Report*, January https://www.bundesbank.de/resource/blob/625900/18104dad43ccef12d20419b8f79be11b/mL/2015-01-enterprises-data.pdf
6. Duffie, D. and K. Singleton (1999), Modeling Term Structures of Defaultable Bonds, *Review of Financial Studies*, Special 1999
7. EBA (2017), Guidelines on PD estimation, LGD estimation and the treatment of defaulted Exposures, 20 November
8. European Central Bank (2015), The financial risk management of the Eurosystem's monetary policy operations, July https://www.ecb.europa.eu/pub/pdf/other/financial_risk_management_of_eurosystem_monetary_policy_operations_201507.en.pdf
9. Giannozzi, O., Altman E., Roggi A. and G. Sabato (2013), Building SME ratings: is it necessary for lenders to monitor financial statements of the borrowers?, *Bancaria*, n.10, 54-71.
10. Hillegeist, S., Keating, E., Cram, D. and K. Lundstedt (2004), Assessing the probability of bankruptcy, *Review of Accounting Studies*, vol. 9, pp. 5-34.
11. Jarrow R., Lando D. and S. Turnbull (1997), A Markov Model of the Term Structure of Credit Spreads, *Review of Financial Studies*, n.10
12. Kealhofer, S. (2003), Quantifying credit risk I: default prediction, *Financial Analysts Journal*, January–February, 30–44.

13. King, G. and L. Zeng (2001), Logistic Regression in Rare Events Data, *Political Analysis*, Volume 9, Number 2, 137-163.
14. Merton R. (1974), On the Pricing of Corporate Debt: The Risk Structure of Interest Rates, *Journal of Finance*, n. 29
15. Moscatelli, M., Narizzano, S., Parlapiano, F. and S. Viggiano (2019), Corporate Default Forecasting with Machine Learning, *Temi di Discussione (Working Papers)*, Banca d'Italia, n. 1256.
16. Ohlson, J.A. (1980), Financial ratios and the probabilistic prediction of bankruptcy, *Journal of Accounting Research*, 18 (1), pages 109-31.

ANNEX 1: Statistical module validation

Discriminating power analysis

The main purpose of a rating system is to distinguish between "healthy" and "sick" units (firms), depending on whether or not the occurrence of the default event is considered probable for each of them.

In most cases, the model is expected to draw a line between the two types of units; the most common procedure to do this is to define a cut-off probability and to consider “healthy” those units that have an estimated default probability lower than the cut-off one and consider “sick” the firms with a higher one.

The model therefore must have discriminating power, that is to say, precision in assigning to the healthy companies a default probability inferior to the cut-off and to the sick companies a higher value.

The discriminating power consists of two characteristics:

- specificity, i.e. the ability to correctly classify the units for which the event does not take place;
- sensitivity, which is the ability to correctly classify units for which, instead, the event occurs.

As the cut-off increases, the model will be more effective in correctly classifying healthy companies, and less will be its ability to identify sick ones. Then, there is a trade-off between specificity and sensitivity.

A common way of representing the discriminating capacity of a model is the Receiver Operating Characteristic (ROC) curve, which displays the above trade-off. One way to verify the discriminating power of the model is to calculate the Under Roc Curve Area (AUROC). By construction, the AUROC assumes values ranging from 0 ("the model is completely wrong") to 1 ("the model discriminates perfectly"), whereas 0.5 indicates a purely random model. Practically, a rating system with an AUROC ≥ 0.7 is considered adequate.

It is important to underline that the evidence about the units gone into default and those ones that survived in the time period under consideration represents only one of the possible realizations from the probability distributions of the defaulters and the non-defaulters. In other words, the "default" phenomenon has fundamentally a stochastic nature, meaning that, in theory, if the same experiment is repeated, different realizations of these distributions would be obtained, with every realization being characterized by a different value of the AUROC (i.e. the AUROC is distributed according to a particular functional form with specific parameters).

The stochastic nature of the “default” phenomenon can be investigated with the “U of Mann-Whitney” statistics, the latter being linked to the AUROC through an equality relation; the statistical properties of the Mann-Whitney U are then applicable to the study of the stochastic behaviour of the AUROC.

Against this background, the Mann-Whitney statistics can help in quantifying how much the value achieved by the AUROC is far from its expected value, by calculating appropriate confidence intervals; furthermore, this instrument allows to carry out useful hypothesis tests, for example in order to verify whether the discriminating power of the rating system under validation is significantly different from the value of perfect randomness.

Predictive power analysis

Another important characteristic that is investigated in a rating system is its predictive capacity (namely the capacity to possess a good quality of calibration).

In a broad sense, the calibration quality refers to the ability to identify the real probability of default for an individual debtor or for a class of homogeneous borrowers.

However, the real PD is unknown, so it is not possible to exactly estimate it. Therefore, to test the quality of the PD calibration the observed (ex-post) default frequencies are compared with the estimated (ex-ante) probabilities of default.

In order to carry out the aforementioned comparison it is necessary to use a test whose null hypothesis basically states that the ex-ante estimates of the different PDs are correct. Typically, a rating system includes various classes of risk and the validator needs to evaluate PD forecasts for all risk classes.

One possible approach for testing multiple risk classes at the same time is that of multiple comparisons. As an alternative, a single statistical test is used to compare risk classes simultaneously (joint testing).

When multiple testing is applied, at a first stage of the analysis each risk class is assessed individually. Under the usual assumption of independence among the defaults, the number of defaults by risk class follows a binomial distribution. In a second stage of the analysis the intersection of the results of the various individual tests is considered, to check if the rating system adopts appropriate PDs estimates for all the classes.

Up to now, multiple testing was implemented in order to: 1) first of all, verify the absence of an underestimation of the PDs on the various rating classes, estimated by the whole ICAS *system*; 2) then, understand if the whole ICAS *system* is too conservative or too loose (i.e., if the PDs estimated via the ICAS system were in line with the effective defaults); 3) subsequently, ascertain if the PDs estimated via the ICAS statistical *model* were equal to the empirical frequencies of default. In other words: multiple tests were carried out, respectively: 1) on the whole ICAS, by adopting a risk aversion profile for default risk typical of a "supervisor"; 2) on the whole ICAS, by adopting the risk profile of a “production unit” with business objectives; 3) only on the “statistical model” part of the ICAS (i.e., not considering the adjustment of the model-estimated PDs to the upper PDs in every rating class), to verify if there are some problems underlying the purely quantitative part of the Bank of Italy’s rating system.

Accepting as null hypothesis the assertion "*the expected defaults are higher than those achieved, for each rating class*" (i.e. adopting the "supervisor" view), a one-tailed binomial test is conducted for each rating class. In this case, the distribution of each test has as its average the PD value of the upper end of the probabilities range that identifies each rating class. Each test is evaluated at two different levels of confidence (95% and 99%).

When the number of actual defaults is significantly lower than that of the expected defaults for all rating classes, the ICAS *system* is considered "prudential", i.e. it never leads to underestimating the risk of default.

On the other side, accepting as null hypothesis the statement "the expected defaults are equal to those achieved, *for each rating class*" (i.e. adopting a "production unit" view) – the statement is equivalent to say “the ICAS *system* is precise” – a two-tailed binomial test is conducted for each rating class. Also in this case, the distribution of each test has the average value equal to the upper end of the probabilities range that identifies each rating class. Each test is evaluated at two different levels of confidence (95% and 99%).

When the number of actual defaults is significantly upper or lower than that of the expected defaults for some rating classes, the hypothesis “the ICAS system is precise" is rejected, i.e. the estimate of the risk of default is not always aligned with reality.

Lastly, accepting as null hypothesis the statement “*the expected defaults coming from the statistical model at the basis of the ICAS rating system are equal to those achieved, for each rating class*” (i.e., investigating whether the statistical model at the basis of the ICAS system has some problems or not) – the statement is equivalent to say “the ICAS *model* is precise” – a two-tailed binomial test is conducted for each rating class. In this case, the distribution of each test has the average value equal to the average of the probabilities of default assigned by the ICAS to the borrowers that are part of each rating class. Each test is evaluated at two different levels of confidence (95% and 99%).

When the number of actual defaults is significantly higher or lower than that of the expected defaults for some rating classes, the hypothesis “*the ICAS model is precise*" is rejected, i.e. the estimate of the risk of default by the statistical model is not always aligned with reality (and the adjustment of the model-estimated PDs to the upper PDs in every rating class can help to improve the precision in the PD forecasting via the BI-ICAS system).However, whenever a multiple testing analysis is conducted, the possible impact on the final results of the “alpha-inflation problem” must be considered. As mentioned above, in this analysis a rating system is considered good if and only if each single rating class is well calibrated; even if only one class shows some problems, then the analyst rejects the hypothesis that the whole system is adequate. It is known, however, that as the number of hypotheses to be tested simultaneously increases, the error of the first species relative to the global hypothesis does not remain constant (i.e. it rises, leading to an "inflation of the alpha").

In order to keep this error under control, different methods can be followed: the most widespread in the literature fall under the strategy of adjusting the p-values obtained on the single class tests, increasing them. Many of these methods were applied also with reference to the BI-ICAS validation, either in the case of adopting the “supervisor” risk profile or in that other of the “production unit” illustrated above.

By resorting alternatively to joint testing, it can be verified whether the deviations between the realized defaults and those expected for the individual classes are in line with the assumed dependency structure between the rating classes; this, however, is obtained at the cost of not being able to ascertain if, in the presence of a good calibration in some classes, there is nonetheless a wrong calibration in others.

Within the family of joint tests, the most commonly used procedures are those proposed by Spiegelhalter and Hosmer and Lemeshow.

The Spiegelhalter test, based on the Brier Score, proposes to overcome the multiple testing gaps by combining the information on the calibration quality available at the individual level. Under the null hypothesis of perfect calibration, for which the expected value of the default distribution is equal to the PD estimate for each unit, the test statistic is distributed according to a standardized normal. The test is based on the debated assumption of independence among the defaults.

It should however be noted that the Spiegelhalter test results have to be read carefully. In particular, the Spiegelhalter test statistic is based on a weighted average PD that is estimated on the entire portfolio of examined debtors. Given that, only an average overestimation/underestimation of the probabilities of default causes the refusal of the null hypothesis of perfect calibration at the individual level. The test therefore fails to correctly identify as unacceptable those situations in which an overestimation of PDs for some units is compensated by the underestimation of PDs for other units of the sample.

To overcome this issue, the Hosmer-Lemeshow test is used, which compares expected defaults and realized defaults within each pre-defined rating class. It should be highlighted, nonetheless, that also the Hosmer-Lemeshow test, like the Spiegelhalter one, uses a normal approximation of the binomial distributions that leads to a chi-square distribution of the statistic; this approximation could be questionable because there might be an issue of bad approximation for rating grades with a small number of borrowers.

ANNEX 2: Validation analysis of the Expert System: influence of the six profiles on the final judgment

The aim of this analysis is to evaluate if all profiles are taken into account by the analysts or some of them are systematically disregarded and, in addition, which profile has the strongest influence in the final assessment. In the current framework, in fact, all the profiles contribute with the same weight to the "full PD", with the exception of the "balance sheet ratio" and the "financial flexibility" that have a double weight. In order to measure which profile is more relevant for the final judgement of the analysts, a multinomial logistic regression has been applied, where the dependent variable is the direction of the change in the rating set by the analyst (upgrading, downgrading, confirmation), while the independent variables are the scores assigned by the analyst on the six profiles.

Role of the second analyst

Once the analyst has made his decision, the latter is verified by another analyst in order to guarantee the respect of the so called "four-eyes principle". The first analyst makes a proposal about the rating that can be either a confirmation of the rating stemming from the statistical model or a proposal to change it. The second analyst checks that the rating assessment has been correctly performed and completed; afterwards, he examines the proposal.

In order to obtain a quantitative and synthetic measure of the agreement between the two analysts, the Cohen's Kappa coefficient can be used, a statistical indicator that measures "inter-rater" agreement for qualitative objects. If the analysts always agree, then $\kappa = 1$. If there is no agreement among the analysts, except what would be expected by chance, $\kappa \leq 0$.

The "weighted Cohen's Kappa" is an evolution of the "simple Cohen's Kappa", since it counts differently the level of disagreement emerging from the study. A disagreement that assigns a confirmation instead of an upgrade is weighted less than a disagreement that subverts the decision and turns a proposal of upgrade in a downgrade. In a decision matrix between analyst 1 and analyst 2, off-diagonal cells contain weights indicating the seriousness of that disagreement, double-weighting the disagreement when there is an inversion in the direction of the first assessment (from an upgrade to a downgrade and vice versa).

Effects of the Expert System on the risk classes attributed by the statistical system

In order to understand how much the sample assessed by the Full PD differs from the one assessed using the PD Stat, it is useful to build a transition matrix. In a transition matrix, where rows represent the risk classes attributed by the statistical system and columns are the risk class attributed by the "expert system", the main diagonal contains all the assessments in which the risk class assigned by the statistical system is confirmed, while downgrades are in the top right triangle and upgrades on the low left. The matrix is useful in order to synthetize in a glance how much the

Full PD diverges from the PD Stat. The bigger the dispersion in comparison to the major diagonal, the bigger is the role of the analysts in the assignment of the Full PD.

Relationship between the Expert System and the statistical risk class 12 months later

This analysis aims to assess the ability of the Expert System to foresee the statistical risk class that the firm will have after one year. The idea is that the Expert System improves the accuracy of the evaluation of a firm exploiting information sources that, despite being available at the moment of the statistical assessment, are hardly usable in a quantitative model. This information, still not incorporated in the variables processed by the quantitative model, is expected to have an impact on the statistical risk class after one year. If the estimated Full PD, in which all the available information is taken into account, is a good proxy of the direction of the statistical risk class after twelve months, then there should exist a relationship between the class attributed by the analyst today and the statistical risk class 12 months later.

After having selected the upgrades, the confirmations and the downgrades of the analysts, one needs to compute if there is a positive relation between the analyst's judgement and the change in the statistical PD after 12 months. One then calculates in how many cases the direction of the expert analysis is the same of the statistical PD after 12 months, and, on the contrary, in how many cases a "total discordance event" is recorded, in which the analyst assigns an upgrade, followed 12 months later by a statistical downgrade and vice versa. The relationship is synthetized by the Cohen's Kappa; also in this case the computation of the "weighted Kappa", that penalizes the cases of total discordance, is useful.

Analysis of defaults

This analysis is restricted to cases in which a default occurred, for which it is possible to investigate how the analysts performed in anticipating the worsening of the credit condition of the assessed firms. Limiting the analysis to cases of defaults, the number of occurrences in which the expert system correctly decided for a downgrade, for a confirmation and for an erroneous upgrade is counted. A statistical test of causality is performed.

A deeper investigation, focused on of the six different profiles of analysis, shows which profile contributed to correctly signal a future default and if there are misleading profiles that assigned a positive score in a case that ended up with a default.

The analysts' behaviour in assigning the final ratings and judgment on the profiles

It may be of interest to understand if there are "biases" among experts in examining firms, for example if there are groups of analysts who favor one or more profiles in the production of the final rating relatively to the others and if there are profiles that are not significant for some analysts in the attribution of the final rating, but that are relevant for others.

In order to answer these questions, the reference dataset is divided by selecting records relating to each analyst; subsequently, a multinomial logistic regression is used to investigate how much each analyst weighs the profiles. This approach provides a regression "beta" matrix available for each profile and for each analyst.

The betas obtained in this way are aggregated using a cluster analysis, based on the "Ward method", to identify groups of analysts who adopt the same type of investigation. A generally desirable feature is that analysts consider as significant all the profiles, with no single profile emerging as the most important one. As a second step, the beta matrix obtained becomes an input that allows creating clusters, as more homogeneous as possible within them and heterogeneous when compared with the other ones; this step is aimed at determining similarities among groups of analysts. In order to ensure a high significance to the results obtained, the analysis is limited exclusively to those analysts that performed at least 100 evaluations.

It can be interesting to analyze the cases in which there is a high convergence among the behaviour of some analysts, checking; for example, if two analysts that show a quite similar behaviour are part of the Central Administration or of the branches network, or if they live in the same geographical area or have something else in common that can explain the similarity in the behaviour. Analyses can also be made on experts clusters in scoring the profiles, more in detail, to understand if there are groups of analysts who evaluate in a conservative way some specific profiles when other analysts systematically provide high scores for the same inputs or if there are analysts who are severe in reference to all profiles, when compared to the other members of the expert team. In this regard it has been investigated, through a one-stage analysis of variance accompanied by a multiple comparison, what is the behaviour of each analyst with respect to each profile.

As a general rule it is desirable to reach the conclusion that there are differences among the analysts in scoring the profiles and even that there are some analysts who, on average, can be considered stricter than the others in expressing their own judgments on the profiles.