RISK MANAGEMENT MAGAZINE

Vol. 17, Issue 3
September – December 2022

EXCERPT

https://www.aifirm.it/rivista/progetto-editoriale/

Modello LGSR forward looking

David Cavallini, Francesco Letizia

Modello LGSR forward looking

David Cavallini (Risk Management, BU Governance – CABEL Industry S.p.a.) and Francesco Letizia (Internship at Cabel Industry S.p.a.)

Article submitted to double-blind peer review, received on 22nd September 2022 and accepted on 13th November 2022

Abstract

In this work, we propose a hierarchical model to introduce *Forward-Looking* effects on the Loss Given Default Rate (LGDR) estimate, as required by IFRS9. The Framework consists of two modules: a SURTS satellite model (*Seemingly Unrelated Regressions Model Time Series*), which analyses the dynamics of the systemic LGSR (*bad loans LGDR*) and a set of selected macroeconomic factors, and a *Beta Inflated-(0,1)* model which estimates the LGSR for the single entity. The basic hypotheses for the construction of the hierarchical model will also be illustrated, underlining how this approach is particularly relevant for LSIs (*Less Significant Institutions*). The theoretical aspects are followed by an application on a series released by the Bank of Italy, presenting the LGDR estimation process on an archive of closed bad loans by a set of banks belonging to the CABEL (ICT Service Provide) network. By way of example, we illustrate the forecast results for the three-year period 2022-2024 for the systemic LGDR. Other aspects related to the construction of LGDR models are addressed, such as the segmentation of the portfolios and the selection of individual attributes. In particular, we introduce the NPL *vintage* as an explanatory variable in the LGDR model, outlining the interconnections with the effects of macroeconomic projections.

1. Introduction

The financial crisis of 2008 raised the need to promptly account the quality deterioration of loans portfolio, to align share funds with credit risk. The *International Financial Reporting Standard (IFRS9)* required operators to switch to a lifetime perspective in the accounting of the expected losses for those assets characterized by a significant increase in credit risk (SICR). Furthermore, to reduce the *shortfall* of the provisions, was introduced the necessity to assess the probability distribution of losses from a *forward-looking* perspective.

From the first adoption of the new accounting standard, more or less complex methodologies have been proposed, which underly on different hypotheses. This phenomenon has highlighted an already known aspect in the *risk management* area, linked to model risk and to comparison of results. However, a prevalent school has been consolidated, widespread both in institutional environments and among consulting private company (Deepak Parmani & Analytics.), which envisages the use of macroeconomic factors as a fundamental element to consider the *forward-looking* component in risk parameters.

Among the various attempts, it is worth mentioning (Joubert, et al., 2021), which proposes to include macroeconomic factors as explanatory variables in the LGDR model. This approach, which we call *direct*, seems reasonably applicable to systemic banks as major contributors to national data, but we believe it is not as suitable for *Less Significant Institutions* (LSIs). After having defined LGDR in Section 2, in Section 3.1, we set out in detail the reasons for our proposal, which is embodied in the formulation of a hierarchical model, using basic concepts of conditional independence.

Following the economic-financial laws related to credit recovery a *satellite* model is implemented, which establishes a relationship between the systemic *Recovery Rate* (RR) and macroeconomic variables. In Subsection 3.1, we illustrate the methodological aspects of the *Seemingly Unrelated Regressions Model Time Series* (Harvey, 1989 and Hamilton, 1994).

Empirically, the distribution of LGDR of a banking portfolio is bimodal with points of discontinuity at the extremes $\{0,1\}$, with consistent forms of asymmetry. In this context, *a Beta-Inflated* distribution has been adopted, considered in the literature as a reference density for the estimation of econometric LGDR models (Grzybowska & Karwański, 2020) and (Joubert, et al., 2021). In Subsection 3.3, we illustrate a methodological synthesis of the *Generalized Additive Models for Location, Scale and Shape (GAMLSS)* proposed by (Rigby & Stasinopoulos, 2005), which represents the foundation of the adopted regression model. In a direct approach, in addition to the individual characteristics of NPL, the explanatory variables also consider a set of macroeconomic factors that are empirically significant to explain the RR. The proposed hierarchical model condenses the trend of the economic cycle into a single factor, the systemic LGDR, which represents the only macroeconomic variable considered in the set of regressors. This approach has the advantage of limiting the effects of multicollinearity but, above all, it brings cognitive benefits to the estimation process by facilitating the interpretation of the results.

Once the macroeconomic scenarios that reflect the prospects of the national economy have been defined, the projections of the LGDR are drawn up. In Subsection 3.4 it is shown that, with respect to the specified model, it is not possible to obtain a closed form of the forecasts in terms of expected value. Even though the Monte Carlo method would allow to obtain an adequate approximation in probabilistic terms. Here, we have adopted the *plug-in* technique which considerably reduces the computational efforts.

In Section 3 we present the results obtained from the application to Italian data, related to NPL closed through internal credit recovery (the ones not sold on the market). The system of equations reflects the segmentation that we believe to be minimal in the context of LGDR models, deriving from the intersection of two attributes: type of counterparty (firms/households) and the presence or absence of collateral (secured/unsecured).

Maintaining the level of confidentiality of the data of the banks belonging to the CABEL[1] network, in Section 4 we illustrate the results of the application to a historical archive of losses. From the series of publications of Bank of Italy, Financial Stability and Supervisory Notes (Fischetto, et al., 2021), we show the quite relevant role played by the duration of the recovery process on RRs. In fact, the Italian average recovery rate for NPL closed in the first year is about twice the one on those closed in more than five years. In other words, an inverse relationship exists between recovery rates and duration. This empirical evidence is reflected by regulation in the *Calendar Provisioning* framework introduced by the European Central Bank's (ECB) "Impaired Credit Guidelines" in March

[1] CABEL (ICT Service Provide, Center for Banking and Leasing Assistance) company dedicated to providing services in the banking industry. (www.cabel.it).

2017[2], which establishes progressive minimum levels of prudential provisioning (*Prudential Backstop*) that are based on the NPL vintage[3]. Given the relevant role of this attribute, the duration of recovery processes was introduced as an explanatory variable in the regression model. The study ends presenting in Section 6 some final remarks and future developments.
As detailed above, the following section outlines the basic concepts of LGDR calculation with a brief mention of the regulatory and normative framework. The section on conclusions is devoted to expressing the benefits of the proposal, emphasizing that the model is not an exhaustive solution to the issue addressed. Methodological aspects are subject to improve.

2. Definition of LGDR

Despite the entry into force of the new definition of *default*[4] that banks must adhere to as of January 1, 2021, it is still relevant to split the life cycle of impaired loans (NPLs) into two main stages. The first stage is the administrative states of *Past-Due* and *Unlikely to Pay* (UTP), while the next step is the transition to non-performing. Although in banking practice legal actions and out-of-court settlements are also undertaken for UTP, the recovery of the loan intended in the strict sense is fundamentally to be ascribed to the bad debt status. This distinction is also inevitably reflected in the methodological aspects related to the estimation of the LGDR parameter, which consists of two basic modules: the estimated LGDR model on NPL (LGSR) and the estimation of the probability of transition from *Past-Due* and *Unlikely to Pay* states to non-performing status, which is considered an absorbing state (*Danger Rate*).
This paper is framed in the context of estimating distressed LGSR models following the *Workout approach*, which is based on the measurement of cash flows from the recovery process, appropriately discounted. In formal terms, LGSR is calculated as the complement with respect the RR i.e., LGS $= 1 -$ RR. Given a schedule $\mathrm{t} = (t_1, t_2, t_3, \dots, t_n)$, representing the n relevant time events along the duration of recovery process, we define:

$$\mathrm{RR} = \frac{1}{EAD} \sum_{i=1}^{n} \frac{R(t_i) - C(t_i)}{[1 + r(t_i)]^{t_i}} \quad (2.1)$$

where:

$\mathrm{R}(t_i)$ is the nominal value of the gross amount recovered as at t_i;
$\mathrm{C}\ (t_i)$ is the nominal value of the cost (direct/indirect) related to the recovery procedure incurred at instant t_i;
$\mathrm{r}(t_i)$ rate curve for the discounting observed at the time of bad status entry;
EAD is the initial nominal value of the exposure at the time of bad status entry;
t is the the duration of the recovery procedure.

The choice of interest rate maturity curves, which are useful for discounting cash flows, depends on the scope of application. From a regulatory perspective, using historical rates, the discount factor must consider a risk premium (e.g., by estimating a CAPM - *Capital Asset Pricing Model*). The new accounting standards (IFRS9), on the other hand, require the use of the *Effective Interest Rate* (EIR) calculated at the origination of the loan. For further discussion of aspects related to the selection of discount rates for calculating LGDR we refer to (Gibilaro & Mattarocci, 2006).
The basic principle underpinning the regulatory approach is the principle of *prudence*: the Basel Committee expects LGDR to be estimated during a *downturn* in the credit cycle. The drafting of the new IFRS9 accounting standards follows the *point in time* (PIT) philosophy and, therefore, the LGDR risk parameter must be constructed taking into account expected risk factors (LGDR *forward looking*). In addition, to avoid *double accounting*, costs are derecognized in the calculation of RRs in equation 2.1. For a comprehensive analysis of the main differences emerging between the two regulatory and accounting approaches, please refer to AIFIRM Position Paper No. 8/2016.
The priority source for feeding the data structure is the historical data contained in banks' internal archives, which form the information base for building internal models and for reporting regarding losses (LD matrix[5]). The detail in supervisory reporting often does not provide sufficient information for the segmentation (see subsection 3.2). Therefore, anagraphics additions are required.
Alternatively, following Bank of Italy publications (Ciocchetta, et al., 2017), LGDR can be calculated by accessing data from the Centrale Rischi (CR). The information available in that repository is not sufficient for a detailed calculation as expressed in equation 2.1, but it is possible, through reasonable working assumptions, to achieve an adequate approximation.

3. LGSR model in forward-looking perspective

The inclusion of macroeconomic factors within LGSR models can be tackled by following two alternative approaches.
The first, which we call *direct*, involves relating the entity's LGSR to macroeconomic variables by following *model selection* techniques and/or referring to concepts from a well-defined economic theory. In practical application, problems may be encountered both in the construction phase and in the final interpretation of the selected model; in some cases, parameter estimates may assume a sign that is in contrast with the economic laws of credit recovery. In general, there may be a variety of causes for such difficulties, which can be traced to multicollinearity phenomena in macroeconomic variables and/or the use of metrics that tend to select models affected by *overfitting*. Since credit recovery duration often have long to consider a completely economic cycle, the analyst may encounter the additional difficulty of defining the instant of observation of the macroeconomic factor with respect to the date of entry and closure of bad status[6].

[2] Then supplemented in March 2018 with the Addendum to the Impaired Credit Guidelines and in April 2019 with the publication of Regulation (EU) 2019/630 of the European Parliament and of the Council (the Regulation 2019/630).
[3] Length of permanence in NPL (Non-Performing Loan) status.
[4] European issuance of guidelines on the application of the definition of default pursuant to Article 178 of Regulation (EU) No. 575/2013 (EBA/GL/2016/07) and regulatory technical standards on the materiality threshold for credit obligations in arrears and related Delegated Regulation (EU) 171/2018 of the European Commission of October 19, 2017 (EBA/RTS/2016/06).
[5] Circular No. 284 of June 18, 2013 - First Update of Bank of Italy
[6] Financial Duration seems to be the most congenial instant to take as a reference point for the detection of macroeconomic factors.

Despite the technical complexities associated with model construction, the attempt to identify a relationship between macroeconomic factors and micro phenomena inherent individual items (or portfolio) may be a sustainable solution for systemic banks, as they contribute to the national data on a consistent basis. In our view, *Less Significant Institutions* (LSIs) deserves special attention. By their nature, the portfolios of these institutions are characterized by a narrow, or at least uneven, spatial distribution across the country, as well as by significant idiosyncratic differences in the share of loans allocated to various economic sectors of customers and amount classes. The limited contribution to systemic data of small and medium-sized banks, combined with the limited availability of data from these institutions, makes estimates of LGSR bank elasticities with respect to macroeconomic factors insensitive and often inconsistent in statistical terms. In our view, these considerations assume substantial and sufficient relevance to suggest that a direct approach is not suggested for LSIs.
Here, we have therefore opted for an approach we call *indirect*, which involves the construction of a *hierarchical* model. Below, we summarize the steps followed in constructing the model:

- *Step 1*: A satellite model to estimate the relationships between systemic LGSR and macroeconomic factors following the economic laws of credit recovery. As detailed in subsection 3.1, the model is implemented through the specification of a *Seemingly Unrelated Regressions Model Time Series* (SURTS), constructed following the minimal segmentation that is used in practice by analysts. Specifically, the system of equations reflects the segmentation resulting from the intersection of two dimensions: counterpart (firms/families) and presence or absence of collateral (secured/unsecured);
- *Step 2*: Segmentation of the historical data matrix of closed NPL with internal process of credit recover, based on the bank's *Business Model*. As emphasized in subsection 3.2, the characteristics used to construct the OLAP (*On-Line Analytical Processing*) cube can vary from bank to bank and depends on the information availability (number of items for cell);
- *Step 3*: For each OLAP cell estimate a *Beta Inflated-(0,1) Regression* by considering the systemic LGSR among the explanatory variables. The set of covariates can also consider other individual attributes that are typically continuous, such as, for example, the duration of the recovery process, the collateral coverage rate (LTV - *Loan to Value*), etc.

The LGSR model estimation procedure as shown by *Steps 1, Step 2,* and *Step 3* has an immediate derivation in probabilistic terms, with important conditional independence assumptions[7].
We denote by x_t the vector of macroeconomic factors, s_t is the vector of the systemic LGSR with $t = 1, 2, \dots, T$. The systemic data is organized in a design matrix $X = [x_t]_{t=1}^{T}$ and a matrix of response variables $S = [s_t]_{t=1}^{T}$. For single bank, the internal data are substantiated by two sets of explanatory variables: d_j attributes, which allow the j-th observation to be ranked in an OLAP cell, and z_j, a set of individual characteristics inherent the counterparty or/and the NPL positions being recovered. The symbol l_j is used for indicating the LGSR for j-th NPL closed with internal recovery process, to be calculated according with the methodology outlined in Section 2. Hence, the database held by the bank consists of the triplet: $L = [l_j]_{j=1}^{N}$, $D = [d_j]_{j=1}^{N}$ e $Z = [z_j]_{j=1}^{N}$, where N is the number of observations in the train sample.
Taking advantage of the well-known rules of probability calculus, the joint density of the response variables $S = s$, $L = l$ conditional on the set of exogenous (X, D, Z) can be decomposed as follows:

$$P(s, l|X, D, Z; \theta) = P(l|s, X, D, Z; \theta) \cdot P(s|X, D, Z; \theta) \quad (3.1)$$

where θ is the set of parameters to be estimated.
The underlying probabilistic assumptions that justify the indirect approach for the construction of the LGSR hierarchical model can be summarized basically in two points:

- (*Hypothesis 1*) $P(s|X, D, Z; \theta) = P(s|X; \theta)$: systemic LGSR is not affected by the data of single financial institution. The evidence on macroeconomic factors encapsulates all the information needed to explain the loan loss trends of the entire banking system; the recovery action of a single bank brings no relevant cognitive to the national phenomena. In our view, this assumption is commensurately correct for LSIs but may fail for institutions that are the backbone engine of the financial system;
- (*Hypothesis 2*) $P(l|s, X, D, Z; \theta) = P(l|s, D, Z; \theta)$: conditional on the systemic LGSR, the credit losses of a bank are independent of macroeconomic factors, $L \perp X|S = s$[8]. By considering the systemic LGSR, knowledge of macroeconomic factors adds no information in terms of the economic cycle in modelling credit losses of a single financial institution.

Thus 3.1 can be written as follows:

$$P(s, l|X, D, Z; \theta) = P(l|s, D, Z; \theta) \cdot P(s|X; \theta) \quad (3.2)$$

Furthermore, based on probabilistic assumptions, it is quite natural to assume that the parameter vector θ can be partitioned into two sub-vectors: β and γ such that:

$$P(l \mid s, D, Z; \theta) = P(l \mid s, D, Z; \beta)$$
$$P(s \mid X; \theta) = P(s \mid X; \gamma). \quad (3.3)$$

Therefore, given 3.3, from 3.2 we have that:

$$P(s, l \mid X, D, Z; \theta) = P(l \mid s, D, Z; \beta) \cdot P(s|X; \gamma). \quad (3.4)$$

[7] For a definition of conditional independence and *Graphical Models*, see Whittaker (1990).
[8] The symbol $\perp$ is used for indicating the independence among random variable Whittaker (1990).

Having the observations on the endogenous and exogenous variables, it is immediate to see that the likelihood function, equation 3.4, enjoys the property of separability in the parameters. This ensures that γ and β can be estimated separately according to the likelihood principle. Specifically, from Step 1 (*satellite model*) we obtain the estimation of γ, while Steps 2 and 3 are devoted to the estimation of β.
The indirect approach represents a generalization, because a response variable, systemic LGSR, is introduced into the model, which participates as exogenous variables in the bank LGSR model and takes the role of endogenous in the satellite model. The two methodologies are related; in fact, by marginalizing 3.4 with respect to s, we obtain by reduction the direct approach model:

$$P(l \mid X, D, Z; \theta) = \int_{\mathcal{S}} P(s, l \mid X, D, Z; \theta)\, ds \quad (3.5)$$

where $\mathcal{S}$ is the sample space of S.
The two working approaches have differences not only in terms of construction and estimation but also in the final application i.e., forecasting. Once the macroeconomic scenarios that allow defining the expectations on the explanatory variables X are available, for each OLAP cell and conditional on the individual characteristics of the institution's portfolio it is possible to obtain the expected bank-level LGSR. Since in the direct approach, the relationship between bank LGSR and macroeconomic factors is estimated following a probabilistic approach, Equation 3.5, the forecasting step in terms of expected value become immediate (except for complex functional forms). For the indirect method is not the same thing. In fact, as discussed in detail in subsection 3.4, it is not possible to produce closed-form forecasts, but it is necessary to implement a Monte Carlo solution or to propose an approximation by *plug-ins*. The latter solution, which will be followed to expose the empirical results, has a reduced computational complexity compared to estimation by simulation.

3.1 Satellite Model: SURTS

In econometrics, the SUR (*Seemingly Unrelated Regressions Model*), proposed by Zellner (1962), is a generalization of the linear model and consists of a series of equations each of which has its own dependent variable and a set of explanatory variables.
Denoting by k the number of equations and by m the number of explanatory variables, which define the dimensions of the vectors $r_t = 1_k - s_t$ (systemic RR) and x_t, respectively, the SUR model has the following linear structure:

$$r_t = \Lambda\, x_t + \epsilon_t \quad (3.6)$$

where Λ is the parameters matrix of k rows and m columns and ϵ_t is the vector of k stochastic errors.
Typically, each regressor does not appear in all equations, so the matrix Λ occurs in a sparse form. Some cells of that matrix are set equal to zero, which indicates the exclusion of a regressor from an equation. Therefore, Λ is subject to the following linear constraints:

$$\Lambda = \sum_{i=1}^{q} G_j\, \gamma_j \quad (3.7)$$

where $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$ is the vector of q free parameters and the G_j are sparse matrices[9].
The SUR model can be equivalently reformulated as follows:

$$r_t = X_t\, \lambda + \epsilon_t \quad (3.8)$$

where:

$$X_t = (x_t^T \otimes I_k)$$
$$\lambda = \mathrm{vec}(\Lambda) = \sum_{i=1}^{q} g_j\, \gamma_j \quad (3.9)$$

with $g_j = \mathrm{vec}(G_j)$.[10]

Regarding stochastic errors ϵ_t, we assume that they follow a *Multivariate Autoregressive State-Space* (MARSS) process proposed by Holmes (2021), namely:

$$\epsilon_t = \eta_t + \nu_t$$
$$\eta_t = \Gamma\, \eta_{t-1} + \xi_t \quad (3.10)$$

where ν_t and ξ_t are mutually independent Normal random vectors with zero mean and variance/covariance matrix Σ_ν and Σ_ξ respectively. In 3.10, Γ represents the not necessarily symmetric square matrix of autoregressive parameters. For the process to be stationary, the k eigenvalues of Γ must be in modulus less than 1. By tending $\Sigma_\nu \to 0$, that is, defining ν_t as a degenerate Normal on the zero mean, Equation 3.10 specifies stochastic VAR (*Vector Autoregressive*) disturbances of order 1.
In the context of state-space models, see (Harvey, 1989) and (Hamilton, 1994), the model consisting of equations 3.6 (*Measurement Equation*) and 3.10 (*State Equation*) goes by the name of *Seemingly Unrelated Regressions Model Time Series* (SURTS).
Maximum likelihood (ML) estimation of the parameters β, Γ and the matrices of variances/covariances is then obtained using the *Expectation Maximization* (EM) algorithm proposed by Dempster (1977). For the formal aspects inherent the application of EM to the

[9] The linear constraints defined in 2.7 represent a generalization with respect to our analysis, which focuses on including or excluding a regressor in a given equation. Each matrix G_j is specified sparse with only one value being 1 and everything else being 0. Specifically, if the (i, p) element of a G_j is equal to 1 it implies that the i-th explanatory variable is included in the p-th equation. To make the system identifiable all the matrices G_j must be different from each other.
[10] By $\otimes$ we denote the Kronecker product, while vec is the matrix vectorization operator.

class of MARSS models we refer to Holmes (2013). In this work the numerical part of computation was developed in *R on Cran* with the MARSS package version 3.11.3 (Holmes, et al., 2020 and Holmes, et al., 2021).
The system of equations 3.8 consists of four equations $k = 4$ i.e., the vector r_t has the following elements:

r_t^{cr} recovery rate for loans to firms covered by collateral;
r_t^{cn} recovery rate for loans to firms not covered by collateral;
r_t^{rr} recovery rate for loans to households covered by collateral;
r_t^{rn} recovery rate for households not covered by collateral.

Since 2017, Bank of Italy (Financial Stability and Supervision Notes) has been implementing a series of publications in which the trend of recovery rates on NPL for the Italian banks is in terms of a weighted average. Regarding the calculation methodology, see Ciocchetta (2017), and concerning the annual data update, see Fischetto (2017), Fischetto (2018), Fischetto (2019), Fischetto (2020) and Fischetto (2021). The appendices to these publications show time series not only for recovery rates, but also of other quantities of stock and flow: in particular, the number and volume of closed NPL. From the aforementioned Tables we extracted the RRs not subject to divestment in order to construct the time series of the vector endogenous to the system of equations 3.8. For the design matrix X, we refer to Section 4, which is devoted to data and exposition of the results of model estimation.
From a methodological point of view, another aspect to consider concerns the structure of the matrix Γ. To be able to parsimoniously specify a model that succeeds in capturing state-space correlations, we opted to impose Γ as diagonal. Each latent factor η_{it} with $i = 1, 2, \dots, \kappa$, follows a stationary AR (*Autoregressive*) process of order 1. Moreover, specifying unconstrained and full Σ_ξ allows us to estimate the spatial correlations of the latent factors that contribute to the covariation of the response variables. As for Σ_ν, we impose that it is diagonal i.e., the elements of the random vector ν_t are independent from each other.

3.2 Segmentation (OLAP)

Customer segmentation for loss analysis is closely related to the institution's *Business Model* and from its internal organization. Despite some subjective elements in application, seems widespread practice to consider at least two dimensions: type of counterparty (firms/families) and presence of collateral (secured/unsecured).
In this context, basic segmentation is a minimum and strictly necessary requirement because it allows the systemic LGSR to be assigned to each closed NPL in the train sample. A natural criterion is to proceed to assign the systemic LGSR based on the closed date of the distress loan.
To have an adequate stratification, additional attributes that reasonably influence the credit recovery process can be added. As pointed out by Resti and Sironi (2021), the attributes to be considered can basically be grouped into two categories. The first refers to attributes inherent to the exposure such as, for example, the promptness and liquidity of collateral, the face value to be recovered (EAD), the presence or absence of personal guarantees (sureties issued by third parties and/or consortia), etc. The second category encapsulates the characteristics of the debtor, the sector of economic activity in which the company operates, the presence or absence of *forbearance*, territorial court of jurisdiction, etc.
Having established the dimensions for segmentation, the NPLs are classified: each distress loan, based on observed attributes, is uniquely placed in a single cell of the multidimensional table. Then, the train sample is organized into an OLAP cube: each partition (also called *LGDR grade*) contains several closed NPL from which we observe several measures that are typically continuous variables such as LGSR, EAD, duration of recovery process, etc.
Expert judgment guiding the choice of attributes defining the dimensions of the OLAP table must necessarily be followed by careful data analysis. First, the final decision in adopting a multidimensional structure must be conditioned on the number of items for cell, to ensure the consistency of the estimates of the LGSR models. Another aspect to be checked with appropriate statistical tests concerns the level of separation introduced by segmentation. The main objective is to construct groups that share similar characteristics and exhibit sufficiently small residual variance (i.e., within-grade), but also where a significant explained variance is manifested (i.e., between-cell).
In practice, may also arise the need to discretize a quantitative variable by changing its nature to qualitative ordinal to be able to consider that characteristic among the dimensions of the OLAP cube. A data processing is used to transform a measure into a dimension. Such an operation decreases the degrees of freedom, reducing the impact of volatility and limiting the noise. Most importantly, it improves the interpretability of the results. Some drawbacks appear. The definition of the discrete variable is affected by choice of breakpoints with the possibility of information loss in terms of entropy. Increasing the size of the OLAP cube is always constrained by an appropriate data analysis, but mostly by the number of items for grade that considerably affect the consistency of the LGSR model estimates.
Once the dimensions of the OLAP cube have been outlined, careful maintenance work is empirically required to make the segmentation adherent to one's banking real world. Once the process of constructing the OLAP cube on the train sample has been concluded, it is necessary to proceed a constant testing on existing portfolio, i.e., open NPL has been consistent with the historical data composing the train sample in terms of value and number of items.
In this work we adopt the minimum segmentation required and, at the same time, we maintain the duration of the recovery process as a continuous variable that appear as an explanatory variable of the regressions.

3.3 LGSR Model: Beta Inflated-(0,1)

Given a generic OLAP cell, let n be the number of observations. We assume that observations on the response variable l_j (bank LGSR) with $j = 1, 2, \dots, n$, are generated independently by the probability distributions $f(l_j|\theta_j)$ where θ_j is the vector of p incidental parameters. Each element of θ_j is then additively linked to the vector of explanatory variables $\boldsymbol{w}_j$ by specification of an appropriate monotone function (*link function*), namely:

$$g_i(\theta)_{ji} = \boldsymbol{w}_j^T \beta_i \text{ with } i = 1,2, \dots, q \quad (3.11)$$

The above model goes under the name *Generalized Additive Models for Location, Scale and Shape* (GAMLSS) proposed by Rigby and Stasinopoulos (2005).
As pointed out by Resti and Sironi (2021), the distribution of LGSR for a single financial institution takes a U-shape, i.e., bimodal, with points accumulating on the extremes of the [0, 1] interval. Typically, for exposures secured by residential real estate, recovery rates tend to be high (close to 1) while for all other lines not backed by any collateral, total losses are often experienced (LGSR = 1).
A reduced residual variance is expected for each cell since all items share similar attributes. Typically, in some grades, such as those secured by real estate, the LGSR distribution will exhibit left skewness with accumulation points on the zero in proportion more pronounced than unity. By contrast, for groups that are characterized by unsecured positions, the LGSR distribution will be skewed to the right with an excess of values close to 1. Figure 1 shows two histograms illustrating the described phenomenon by way of example.

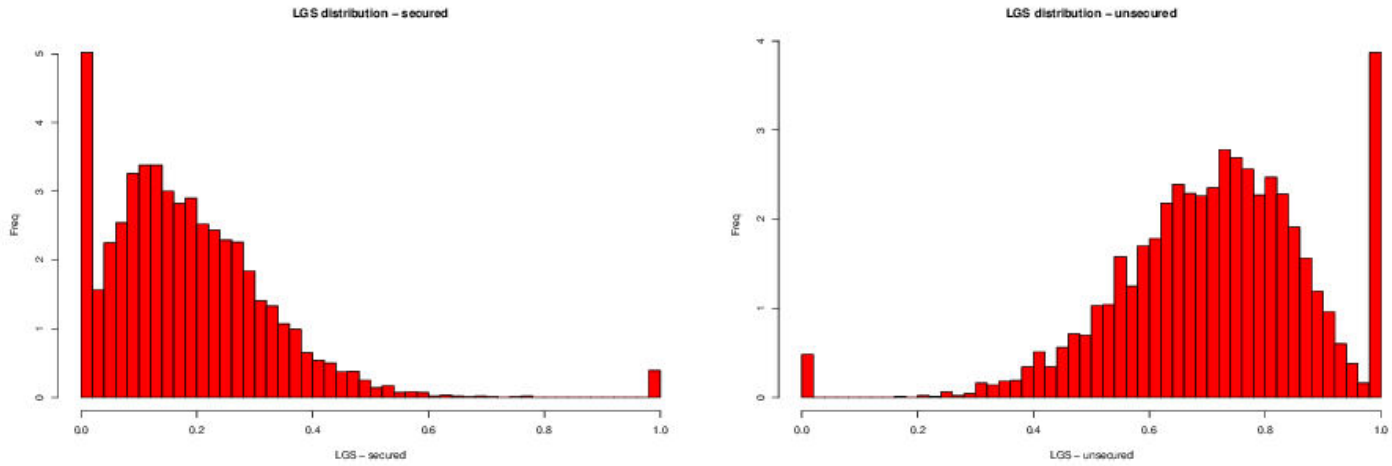

Figure 1: LGSR Distribution for secured and unsecured NPLs (Source: processing on own data)

Based on these empirical considerations, as proposed by Hossain (2016b) (see also Stasinopoulos (2017) and Resti and Sironi (2021)), the *Beta Inflated - (0, 1)* distribution seems to be one among the most suitable ones for constructing a model of LGSR. The use of such a probability distribution succeeds in meeting the two main application needs:

- consider, in inferential terms, the fact that there are two points of discontinuity in 0 and 1;
- the need to capture the forms of skewness and kurtosis typical in the LGSR.

The *Beta Inflated - (0, 1)* distribution is defined as follows:

$$f(l_j|\theta_j) = f(x) = \begin{cases} \eta_{0j} & l_j = 0 \\ (1 - \eta_{0j} - \eta_{1j})\, \mathbb{B}(l_j \mid \alpha_j, \pi_j), & l_j \in (0,1) \\ \eta_{1j} & l_j = 1 \end{cases} \quad (3.12)$$

where $\theta_j = (\alpha_j, \pi_j, \eta_{0j}, \eta_{1j})$, $\mathbb{B}(l_j \mid \alpha_j, \pi_j)$ is the standard Beta density with shape parameters (α_j, π_j) both positive, while $\eta_{0j}, \eta_{1j} \in (0, 1)$ refer to the probabilities on the discontinuity points 0 and 1 with $\eta_{0j} + \eta_{1j} < 1$, respectively. To facilitate the estimation process, it is practical to define monotonic reparameterization:

$$\eta_{0j} = \frac{\delta_{0j}}{1+\delta_{0j}+\delta_{1j}}, \eta_{1j} = \frac{\delta_{1j}}{1+\delta_{0j}+\delta_{1j}}, \alpha_j = \frac{\mu_j(1-\sigma_j^2)}{\sigma_j^2}, \pi_j = \frac{(1-\mu_j)(1-\sigma_j^2)}{\sigma_j^2} \quad (3.13)$$

where δ_{0j}, δ_{1j} are defined on the positive real axis, $\mu_j = \alpha_j(\alpha_j + \pi_j)^{-1}$ is the expected value of standard Beta, and $\sigma_j^2 = (1 + \alpha_j + \pi_j)^{-1}$. We note that the parameter of the mean $\mu_j \in (0, 1)$, with $\sigma_j^2 \in R^+$.[11]
Regarding the specification of link functions, we will proceed as follows:

$$\log(\delta_{0j}) = \boldsymbol{w}_j^T\beta_0, \log(\delta_{1j}) = \boldsymbol{w}_j^T\beta_1, \text{logit}(\mu_j) = \boldsymbol{w}_j^T\beta_\mu, \text{logit}(\sigma_j) = \boldsymbol{w}_j^T\beta_\sigma \quad (3.14)$$

where $\text{logit}(.)$ is the Logit[12] transformation e $\beta = (\beta_0, \beta_1, \beta_\mu, \beta_\sigma)$ is the vector of regression coefficients to be estimated.
The vector of explanatory variables $\boldsymbol{w}_j$ in 3.14 appears in all parameter transformations. This approach seems to be entirely general; in fact, in the strictly application domain, some regressors might be useful in explaining the accumulation of observations on

[11] The variance of the Beta distribution according to the reparameterization in 2.13 is equal to $\sigma_j^2\,\mu_j(1-\mu_j)$ i.e., it depends on the mean μ_j and σ_j^2 representing the variance amplification parameter. In fact, with respect to μ_j the variance function has its maximum value equal to $\sigma_j^2/4$.
[12] For every $x \in (0,1)$, $logit(x) = log(x) - log(1-x)$.

discontinuity points, others might be more suitable for modelling the phenomenon in terms of levels, variability and symmetry within the (0, 1) support.
The direct approach requires macroeconomic factors to be considered in the set of explanatory variables i.e., $\boldsymbol{w}_j$ consists of $\boldsymbol{z}_j$ and $\boldsymbol{X}_j$ or appropriate selection. Moving from a *Long List* to a *Short List* of regressors can be done by referring to expert judgments or by setting up *Model Selection*.
The alternative, consisting of the hierarchical model, considers as a single factor the systemic LGSR that is assigned to each item based on the closing time i.e., $\boldsymbol{w}_j = (\boldsymbol{z}_j, \boldsymbol{s}_j)$. Considering a single synthetic indicator as a thermometer of economic trends makes it possible to streamline the set of explanatory variables by reducing the effects of multicollinearity.
Maximum Likelihood (ML) method is used for estimating the β parameters. As demonstrated in (Hossain, et al., 2016b), the likelihood for the *Beta Inflated - (0, 1)* model enjoys the separability property in the parameters (β_0, β_1) and $(\beta_\mu, \beta_\sigma)$. Specifically, (β_0, β_1) are estimated by means of a *Multi-Logit* model and $(\beta_\mu, \beta_\sigma)$ are obtained by proceeding to estimate a *Beta* regression for only those observations that lie in the interval (0, 1). The statistical computations shown in this paper were developed in R on Cran, GAMLSS package version 5.4-1. For further technical details we refer to (Rigby, et al., 2021).
Once the parameters are estimated $\hat{\beta}$ for a generic OLAP cell $\star$, conditional on the individual characteristics $\boldsymbol{Z} = \tilde{\boldsymbol{z}}$ and the prediction on the systemic LGSR $\boldsymbol{S} = \tilde{\boldsymbol{s}}$, it is possible to construct the probability distribution of the bank LGSR in that grade, which we denote by $L^\star | \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{s}}$. Given the assumptions made about the data generating process, the predictive distribution is also *Beta Inflated - (0, 1)* with expected value:

$$E\left[L^\star \middle| \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{s}}, \hat{\beta} \right] = \hat{\eta}_1 + (1 - \hat{\eta}_0 - \hat{\eta}_1)\, \hat{\mu}. \quad (3.15)$$

The conditional estimate of LGSR for the generic $\star$ cell thus consists of two elements. The first, $\hat{\eta}_1$, represents the estimated probability of the LGSR=1 event: this component tends to increase or decrease the expected loss based on the percentage of positions that empirically closed the recovery process without registering any repayment flow. The second component $\hat{\mu}$ is the expected value of estimated losses from NPL closed for which an LGSR was observed in the support (0, 1). Specifically, $\hat{\mu}$, which is weighted by the corresponding empirical percentage $(1 - \hat{\eta}_0 - \hat{\eta}_1)$, is the expected value of a standard Beta distribution that will be located further to the left or further to the right than the simple mean according to the symmetry and kurtosis of the distribution observed in the grade.
The *Beta Inflated - (0, 1)* regression belongs to the class of nonlinear models; therefore, the sensitivity of the expected loss with respect to a generic explanatory variable does not have an immediate interpretation. In fact, the partial derivative of 3.15 with respect to any element of $\boldsymbol{w}$ is not constant, as is the case of simple linear models, but has a rather complex formulation that depends on both parameter estimates and on all exogenous variables. Therefore, all measures of sensitivity, such as partial derivatives, semi-elasticity, and elasticity, are strictly conditioned by the value that the vector $\boldsymbol{w}$ assumes, which will have to be evaluated numerically.

3.4 FORWARD LOOKING PROJECTIONS

Once *Steps 1-3* have been completed, we can proceed to the last step inherent the application of the model for forecasting. Having one or more macroeconomic scenarios, for each OLAP cell, conditional on individual attributes, the best forecast of the LGSR is expressed in terms of expected value

$$E\left[L^\star_{T+h} \middle| \mathcal{I}_{\mathcal{T}+\hbar}; \hat{\theta} \right] \quad (3.16)$$

where $\mathcal{I}_{\mathcal{T}+\hbar}$ with $h = 1, 2, 3, \ldots$ is the *Information Set* at instant $T + h$ and $\hat{\theta}$ is the parameters estimate from the train sample.
The number of years with respect to which forecasts are made are chosen according to the depth of macroeconomic scenarios[13]. Since the model is specified considering serial correlations, the information set at $T + h$ $(\mathcal{I}_{\mathcal{T}+\hbar})$ is formed by the train sample and the forecasts on exogenous variables until the $T + h$ time.
By using the standard rule of probabilistic calculus, the expected value, 4316 can be rewritten as follows:

$$E\left[L^\star_{T+h} \middle| \mathcal{I}_{\mathcal{T}+\hbar}; \hat{\theta} \right] = E\left[E\left(L^\star_{T+h} \,\middle|\, S^\star_{T+h}, \mathcal{I}_{T+h}; \hat{\theta} \right) \middle|\, \mathcal{I}_{\mathcal{T}+\hbar}; \hat{\theta} \right] \quad (3.17)$$

where the first expected value on the right side of the equation is calculated with respect to the predictive probability distribution of the systemic LGSR (satellite model), and the second expected value is estimated using the density of the bank LGSR conditional on the systemic LGSR equation 3.15.
It is not possible to obtain a closed-form solution of 3.17, since the required integral does not have an analytical solution. Therefore, either numerical integration techniques have been used or estimation by simulation is required. In the context of a Monte Carlo study, one generates random numbers $\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_v$ of replications from the predictive probability distribution of the systemic LGSR (satellite model) and then proceeds to aggregation by using mean, i.e.:

$$E\left[L^\star_{T+h} \middle| \mathcal{I}_{\mathcal{T}+\hbar}; \hat{\theta} \right] = \frac{1}{v} \sum_{i=1}^{v} E\left[L^\star \middle| \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{s}}, \hat{\beta} \right] \quad (3.18)$$

where the expected value within the sum coincides with 3.15. It can be shown that 3.18 is an unbiased and consistent estimator of 3.17.
The proposed implementation of estimation by simulation is certainly an attractive solution because the estimator satisfies good asymptotic properties but may be computationally expensive. The required effort is strictly dependent on the number of replications **v** to which the accuracy of the predictions is related.
An alternative way is to proceed by *plug-in*, which consists of setting **v** = 1 and substituting in 3.15 the prediction coming from the satellite model in term of conditional expected value. This solution has the advantage of being computationally feasible and at the

[13] Typically, government authorities proceed to publish scenarios that at most cover a three-year time frame.

same time sufficiently indicative of the trend of the phenomenon. The use of point forecasts, as opposed to exploring the sample space by simulation, means that the volatility of the phenomenon around the mean is not considered. For purely computational reasons, but aware that this is an approximation, we adopt *plug-in* prediction in this paper.

4 Satellite Model – Estimates and Forecast

Chosen an appropriate econometric methodology (subsection 3.1), the satellite model specification consists of selecting the set of explanatory variables.-An initial univariate exploratory analysis is conducted by using Bayesian approach[14], but the resulting model is built according to causal laws by referring to economic-financial mechanisms related to credit recovery. From a long list of macroeconomics factor the following are selected (short list):

PIL: Annual percentage change in Gross Domestic Product
RISPARM: Marginal propensity of saving
TAXOCC: Unemployment Rate
PREIMM: Annual percentage change in the real estate price index
PRESIMPR: Year-on-year growth rates of loans granted to non-financial firms
PRESFAM: Year-on-year growth rates of loans granted to households
TAXIMPR: Interest rates charged to non-financial firms
TAXABIT: Interest rates charged for home purchases (households)

The time series of the variables PIL, TAXOCC, PREIMM, PRESIMPR, and PRESFAM are retrieved from the publication inherent to the Bank of Italy's Coefficient of Countercyclical Capital Reserve (2022) and the data collection inherent to Italian household savings is downloaded from EUROSTAT (seasonally adjusted and calendared nasq-10-nf-tr series). As for interest rates TAXIMPR and TAXABIT are acquired from the *Statistical Database* (BSD) of Bank of Italy[15].
The time series 31/03/2006 - 31/12/2021 have quarterly frequency, except for loan growth rates (PRESIMPR | PRESFAM) and the unemployment rate (TAXOCC) are published monthly. From this series we have made an aggregate by quarterly moving averaging. Dependent variables evidence, as indicated in subsection 3.1, have been observed as annual averages (2006-2020). We adopted, as a working hypothesis, that the RR of each segment remains constant within each year by applying a *smoothing* through a moving average of order 4. In Table 1 we show the estimation of the regression coefficient[16]. For the NPL covered by collateral a year-on-year increase of about 1 percent in GDP growth, the model estimates, with a lag of two quarters, an increase in recovery rates on firms of 0.2208% on secured and about 0.1339% on unsecured. As mentioned in other work, see Belotti and Crook (2012), the rates practiced by the banks have a negative effect on RRs. Specifically, a 1 percent change in rates charged to firms for term loans results a RR reduction of about 0.8252 percent for secured and 2.4232% for unsecured with two quarters lag. As for households, a 1% increase of rates on housing loans reduces RRs for secured by about 3.0700% and 4.2793% for unsecured, with a two-quarter lag. As pointed out by Bonaccorsi Di Patti and Gascarino (2020), default rates are related to bank loan growth rates according to an inverse relationship, i.e., an expansion of credit tends to reduce the default rates. Lower inflows of assets to impairment states have a positive effect on recovery rates; in fact, following a 1 percent increase in loan growth, an improvement of 0.3294% is estimated for RRs of collateral-backed and 0.3047% for unsecured. Referring to the household, the improvement for unsecured would be 0.0842% with a lag of two quarters. The performance of the housing market plays a significant role in the credit recovery phase, especially for that part of the portfolio covered by collateral. In fact, a 1% increase in the general real estate price index, sign of a market with expanding demand, tends to increase RR by about 0.3379% for firms and 0.2212% for households. Finally, for the household sector, labour market factors (unemployment rate - TAXOCC) and marginal propensity to save affect recovery rates. A 1% increase in the unemployment rate reduces the RRs of secured positions by 1.6963% and by 1.0380% for unsecured positions. An increased marginal propensity to save by households of 1% manifests with two quarters delay in a 0.9094% improvement in recovery rates for NPLs covered by collateral and a 0.7468% for the unsecured ones.
For completeness, in Table 2 we illustrate a brief diagnostic. The first column shows the estimates of the diagonal matrix of Γ i.e., autoregressive parameters of order 1 combined with the latent factors. Being in modulus less than 1, we infer that these processes are stationary (immediate consequence of the EM estimation algorithm). This consideration allows us to estimate the matrix of long-run variances/covariances of the response variables, according to the following formula:

$$\lim_{t \to \infty} vec[V(Y_t)] = (I_{k^2} - \Gamma \otimes \Gamma)^{-1}\mathrm{vec}(\Sigma_v) + \mathrm{vec}\left(\Sigma_\xi\right) \quad (4.1)$$

The second and third columns show the sample and long-run standard deviations estimated by the model. In general, we observe that the estimated variability is higher than observed, especially for the equations related to firms.

As shown in Table 3, the long-run correlations estimated by the model are also found to be higher than the sample correlations. In general, the satellite model tends to over-estimate the log run variances/covariance matrix. This consideration, in combination with the fact that the model fits the data rather egregiously (fifth column of Table 2), does not in general raise any particular concerns. The diagnosis does not seem to be a reason for invalidation, especially for the use for which it was constructed. Indeed, in case the satellite model is used to generate random simulations, once the macroeconomic scenarios are specified, the synthetic samples will be such that they will explore in probability a larger region of the sample space.

[14] The *Bayesian Model Selection* techniques used are similar to those proposed by Bonaccorsi Di Patti and Gascarino (2020) as part of the construction of the dynamic model to explain default trends in the Italian banking system.
[15] Specifically, TAXABIT corresponds to the series "Harmonized interest rates - home purchase loans - flows" while TAXIMPR "Harmonized interest rates - nonc/c loans - nonfinancial companies - flows".
[16] The *s next to the coefficient estimates refer to their significance: (***) 99%, (**) 95% and (*) 90%.

As previously explained in subsection 3.4, defined one or more macroeconomic scenarios, the satellite model makes it possible to produce forecasts of systemic LGSR. By way of example, let us consider the scenario published by the Bank of Italy, dated January 21, 2022, regarding forecasts of the performance of the Italian economy over the three-year period 2022-2024. Table 4 shows the forecasts in terms of values and percentage change from the previous year of the macroeconomic factors used as exogenous in the satellite model.

Equation	Explanatory variables	LAG	Estimates
Firms - secured	INTERCEPT	-	32.2500 (***)
	PIL	2	0.2208 (**)
	PREIMM	-	0.3379 (***)
	log(TAXIMPR)	-	-0.8252 (**)
	PRESIMPR	-	0.3294 (**)
Firms – unsecured	INTERCEPT	-	25.3097 (***)
	PIL	2	0.1339 (**)
	PRESIMPR	-	0.3047 (**)
	log(TAXIMPR)	2	-2.4232 (**)
Householders - secured	INTERCEPT	-	64.4106 (***)
	PREIMM	-	0.2212 (**)
	TAXOCC	1	-1.6963 (**)
	RISPARM	2	0.9094 (**)
	TAXABIT	2	-3.0700 (**)
Householders – unsecured	INTERCEPT	-	48.3380 (***)
	TAXOCC	1	-1.0380 (**)
	TAXABIT	2	-4.2793 (**)
	PRESFAM	4	0.0842 (**)
	RISPARM	2	0.7468 (**)

Table 1: Parameters Estimation of Satellite Model (Source: own computations on macroeconomic data)

Equation	Γ	Std. Obs.	Std. Estim.	R^2
Firms - secured	0.98	2.308	5.306	0.97
Firms – unsecured	0.96	2.735	4.725	0.96
Householders - secured	0.97	9.444	13.299	0.98
Householders – unsecured	0.94	7.751	8.911	0.97

Table 2: Diagnostics of Satellite Model (Source: own computations on macroeconomic data)

Sample	Firms Secured	Firms Unsecured	Hous. Secured	Hous. Unsecured
Firms Secured	1	-	-	-
Firms Unsecured	0.524	1	-	-
Hous. Secured	0.734	0.605	1	-
Hous. Unsecured	0.616	0.676	0.889	1
Estimates	Firms Secured	Firms Unsecured	Hous. Secured	Hous. Unsecured
Firms Secured	1	-	-	-
Firms Unsecured	0.702	1	-	-
Hous. Secured	0.742	0.731	1	-
Hous. Unsecured	0.645	0.735	0.689	1

Table 3: Linear Correlation Indexes Response Variables (Source own computations on macroeconomic data)

The growth of the Italian economy recorded in 2021 (by about 6 percent in terms of GDP) is also confirmed in the three-year period 2022-2024, projecting an increase in GDP and a reduction in the unemployment rate. Despite some inflationary tensions due to rising energy prices, consumption would grow robustly reaching pre-pandemic values in 2024 with a one-year lag in GDP growth. The increase in consumption would be mainly attributable to household spending on durable goods. During the pandemic period a sharp reduction in consummation was observed and consequently a considerable increase in the marginal propensity to save (15% in 2020, 11% approximately 2020 versus 8% in 2019). Growth prospects in terms of GDP and domestic demand would be followed by a reduction in savings which, relative to disposable incomes, is estimated to return to pre-crisis values at the end of 2024.
From industry studies, the outlook for the house market is positive due to change in household[17] preferences and tax breaks from government authorities (e.g., Bonus-110). A stationary general trend is expected in real estate prices with small upward changes, in line with recovery confidence: the scenario depicted in Table 4 has as its underlying assumption an increase in the general index of real estate prices of about 1%, constant in the years 2022-2023.

[17] The analysis conducted by Nomisma points out that for many households there has arisen a need to replace their first home to improve the inadequacies found during lock down periods. Demand for home purchases is mainly heading to the suburbs: a phenomenon that has been taking place for some time of suburbanization towards the search for higher quality spaces and lifestyles.

Variable	2022	2023	2024
PIL	3.8	2.5	1.7
TAXOCC	9.0	8.9	8.7
PREIMM	1	1	1
RISPARM	10	9	8
PRESFAM	3	2.7	2.4
PRESIMPR	1.5	1.3	1.2
TAXIMPR	1.2	1.3	1.4
TAXABIT	1.5	1.6	1.65

Table 4: Macroeconomic Scenario (source: economic bulletin no. 1, 2022, bank of Italy)

The scenario on economic trends[18], assumes that monetary and financial conditions remain favourable despite a slight increase in nominal rates, also to be attributed to inflationary tensions. Rates charged to firms would stand at around 1.4% at the end of the three-year period, versus 1.13% in 2021. The cost of credit for the purchase of the first home, indicative rates for household loans, would stand at around 1.65% in 2024 (at the end of 2021 it is about 1.42%)[19]. Expectations for growth in the Italian economic system are based on the assumption that government policies, particularly the *National Recovery and Resilience Plan* (NRP), will be implemented in a timely and effective manner: private investment would be financed through access to credit, which would cause an increase in terms of growth of 1.5% in 2022 and then settle at 2018 values of around 1.2%. As for the growth in private sector lending, thanks to positive outlooks on the housing market and the fact that about 80% of new home purchases are made through access to credit, it would thirst at about 3% in 2022 versus 3.6% in 2019 and then align with pre-crisis value (2.4% at 2018).

Variable	2019	2020	2021*	2022	2023	2024
Firms - secured	47.89	45.12	43.81	43.60	42.23	41.30
Firms – unsecured	35.43	35.88	33.40	32.91	31.40	30.21
Householders - secured	48.40	63.74	58.93	56.88	56.88	54.14
Householders – unsecured	32.54	48.29	43.82	42.23	42.23	39.75

Table 5: Forecast on Recovery Rates of the Italian System (Source: own computations on macroeconomic data)

In Table 5 we illustrate the satellite model's forecasts of recovery rates for non-disposable NPL in the Italian banking economy. For comparative purposes, we also report the final data for the years 2019 and 2020 while for 2021, which is marked with an asterisk, the values are estimated with the observed covariate.

As also shown in the graphs in Figure 2, after an up and down trend in the first part of the historical series, recovery rates show a downward trend. For the household sector, the series appears to have a cyclical component around the trend, while for firms the recovery rates have a linear downward trend.

The baseline macroeconomic scenario, Table 4, combined with the sign of the regression coefficients allows us to appreciate which elements contribute to increasing/decreasing recovery rates over the 2022-2024 forecast horizon. The growth expectation of the Italian economy is associated with a reduction in the unemployment rate, which is a contributing element to the improved recovery rates for the household segment. A positive contribution to RRs for secured NPL comes from the upward forecast of the general real estate price index. On the other hand, for the household, the contraction in the marginal propensity to consume, the increase in interest rates and the weakening of loan growth in the banking sector compared to the 2020-2021 period contribute to the reduction. The recovery rates of the household presenting a peak in 2020, having a downward trend even if they remain above those observed in the period before the economic crisis. As for the firms, the increase in interest rates and the reduction in terms of growth compared to the year 2021 of gross domestic product and bank loans contribute to a reduction in recovery rates. In the midst of the health emergency (year 2020), a reduction in recovery rates was recorded, and this trend is confirmed downward in terms of expectation in the three-year period 2022-2024.

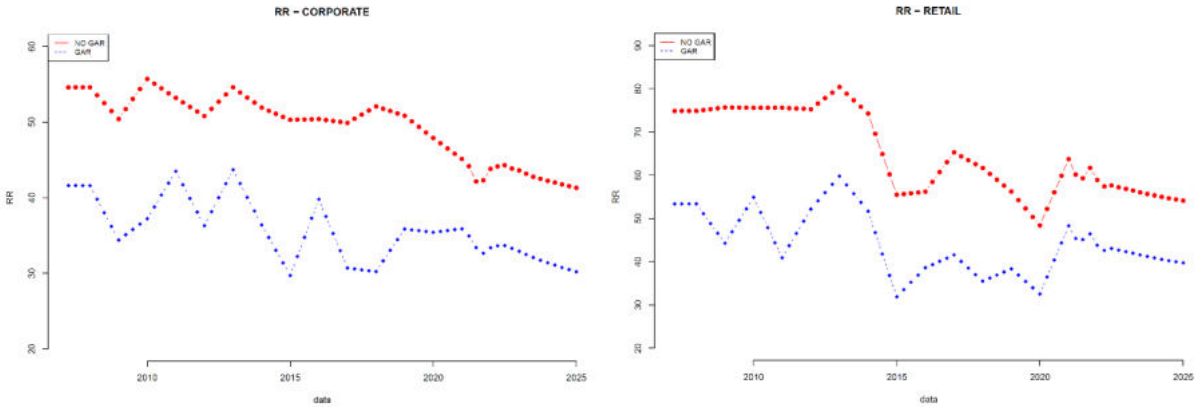

Figure 2: RR Estimates for Positions not Subject to Disposal (source: Bank of Italy (Financial Stability and Supervision Notes and forecasting results by the satellite model)

[18] Bank of Italy, 2022. The Countercyclical Capital Buffer (CCyB) rate. Macroprudential policy decisions of the Bank of Italy.

[19] The interest rate evolutions (TAXIMPR and TAXABIT) presented in Table were obtained by estimating an ECM model on historical data assuming a long-run relationship with the 10-year BTP yield rates and the 3-month EURIBOR.

In conclusion, we would like to point out how the results were obtained as a mere illustrative exercise trying to adhere as closely as possible to the Italian real economy. The set goal of tracing the process of model estimation and use in forecasting seems to have been satisfactorily achieved.

5 LGSR Internal Model

In this section we illustrate the results of the LGSR model, presented in section 3.3, estimated on a data archive of about 5,000 non-performing positions closed in the past 10 years (2011-2021), extracted from a sample of banks from the CABEL network.
The data were processed before estimating the model, so to limit any arbitrage phenomena that banks unconsciously engaged in the allocation of payment flows for unsecured. Specifically, all unsecured NPLs or covered by personal guarantees at the time of the subject's entry into default are aggregated into a single one. The LGS of those items is calculated as a weighted average of the individual LGSR whose weights are represented by the exposures. Respecting the related indirect costs incurred by the bank in the recovery phase, as described in Section 3, were not considered in the calculation of LGSR. The EIRR rates of each transaction are used to discount the nominal cash flows and expenses.
Table 6 shows the percentages of data for which an LGSR value is observed within the open $(0,1)$ range. Complement to 100 represents the fraction of the data on which the LGSR collapses to 1 or 0: LGSR=1 cases are the NPL on which the recovery process had no efficacy whatsoever and in the LGSR=0 cases the bank was able to fully repossess the non-performing exposure. The existence of discontinuity points, which justifies the specification of *inflated* distributions (see section 3.3), shows different intensities in the *grades*; in particular, the phenomenon is more pronounced for unsecured NPLs (about 50 percent for firms and 40 percent for households). The introduction of additional state variables that may be an appropriate proxy for the bank's attention to the recovery process (such as, for example, EAD at the time of entry into non-performing loans), would allow us to circumscribe the intensity of the phenomenon to those segments on which the institution is pessimistic in terms of expectations about debt repayment.
Considering *forward-looking* effects in LGSR model as our main objective, we preferred not to introduce further dimensions of analysis beyond besides the minimal ones: firms/households and secured/unsecured. Each institution can introduce additional stratum variables to build a segmentation that better represents its *Business Model* and, more importantly, its recovery process.

Warranties	Firms	Householders
Secured	57.21	62.14
Unsecured	52.48	38.22

TABLE 6: PERCENTAGE OF LGSR VALUES OBSERVED IN THE INTERVAL (0,1) (SOURCE: OWN DATA).

As explained in Section 3.3, the hierarchical LGSR model involves defining a set of individual attributes (z-vector) to go along with the corresponding systemic LGSR. Aware it does not represent an exhaustive solution, we introduce the duration of recovery as the only continuously characteristic, which nonetheless assumes a rather relevant role in explaining RR trends (Fischetto (2021)). The specified model contemplates that the observed maturity on the closed bad loans enters as an explanatory variable in all the equations in 3.13: in essence, we consider that the duration influences the probabilities of the total ineffectiveness of the recovery action as well as the complete return (parameters δ_0 and δ_1 of equation 3.13, respectively). At the same time, it helps explain the level and volatility of all other items with LGSR within the range $(0,1)$. Empirical results confirm that systemic LGSR is not relevant in explaining the failure or complete recovery of credit. These phenomena should be related purely to the type of recovery and strategies of the institution. For these reasons, the systemic LGSR enters as a regressor in the equations related to μ and σ in Equation 3.13 i.e., explain the trend in the levels and variability of the LGSR distribution for NPLs closed in the interval $(0,1)$.
To maintain the necessary confidentiality of the results, parameter estimates as well as absolute values of prospective projections of loss rates are not published. In our opinion, the reader can observe the consistency of the model by viewing dimensionless numbers, from which the main trends of the phenomenon can be appreciated.
In Table 7 and Table 8 we report the index numbers with 2021 basis of the LGSR *Forward-Looking* estimates for the three-year period 2022-2024, in accordance with some values of duration of the recovery process and in the presence or absence of collateral. The forecasts are obtained following the plug-in methodology: we refer to subsection 3.4 for the reasons for using this method as an alternative to a more correct implementation of a Monte Carlo technique. Index numbers are interpreted according to a comparative key: for instance, a non-performing position with a counterparty belonging to the secured household segment, which is covered by collateral, is estimated to close with an LGSR of 1.0444 times larger in 2023 than in 2021 in case the total duration of the process is around 5 years. Two main reflections are necessary. First, we note that all indices are greater than 1 and have a monotonic increase: that is, the LGSR that considers the systemic perspectives, see Table 5, has an upward expectation. This statement is full agreement with the satellite model results shown in Section 5: an estimated progressive reduction in systemic RRs is reflected in a worsening of the LGSR of each individual bank. A further aspect to note relates to the inverse relationship between percentage increases in LGSR and duration required for closure. This phenomenon is justified by the fact that LGSR values are increasing with respect to the duration of the NPL closed: with higher values of expected losses in 2021, therefore, smaller prospective percentage changes are expected. This rationale is also fully extendable to the fact that, with the same duration, lower LGSR percentage changes are generally observed for secured rather than for unsecured ones.

Years - Duration	Secured			Unsecured		
	1 Year	2 Year	5 Year	1 Year	2 Year	5 Year
2022	1.0085	1.0078	1.0075	1.0020	1.0019	1.0020
2023	1.0642	1.0600	1.0562	1.0081	1.0081	1.0079
2024	1.1031	1.0953	1.0886	1.0129	1.0127	1.0126

TABLE 7: FORWARD-LOOKING LGSR BY DURATION -NON-FINANCIAL FIRMS (SOURCE: OWN DATA)

Years - Duration	Secured			Unsecured		
	1 Year	2 Year	5 Year	1 Year	2 Year	5 Year
2022	1.0252	1.0253	1.0249	1.0070	1.0070	1.0071
2023	1.0450	1.0449	1.0444	1.0139	1.0139	1.0139
2024	1.0650	1.0602	1.0595	1.0192	1.0192	1.0195

TABLE 8: FORWARD-LOOKING LGSR BY DURATION – HOUSEHOLDERS (SOURCE: OWN DATA)

Vintage	Firms		Householders	
	Sec.	Unsec.	Sec.	Unsec.
1 (2022)	1.0256	1.0213	1.0572	1.0130
2 (2023)	1.1494	1.0477	1.1108	1.0261
3 (2024)	1.2346	1.0733	1.1622	1.0379
4 (2025)	1.2781	1.0946	1.1984	1.0442
5 (2026)	1.3219	1.1161	1.2350	1.0507
6 (2027)	1.3654	1.1380	1.2730	1.0571
7 (2028)	1.4086	1.1601	1.3119	1.0635
8 (2029)	1.4515	1.1824	1.3512	1.0700
9 (2030)	1.4932	1.2047	1.3919	1.0766
10 (2031)	1.5342	1.2270	1.4330	1.0830

TABLE 9: LGSR VINTAGE CURVES WITH FORWARD-LOOKING EFFECT (SOURCE: OWN DATA)

Lastly, Table 9 shows the index numbers of the vintage curves with *forward-looking* effect, having 2021 as the base and zero duration. In essence, these curves can be interpreted as the evolution of LGSR with respect to the economy's expectations for non-performing loans originated in 2021. A position belonging to the firm, backed by collateral and with the expectation of closure at 5 years (i.e., 2026), is estimated to have an expected loss 1.3219 times larger than it would be with respect to the 2021 closure event. For the three-year period 2022-2024, LGSR projections consider two elements: the effect of macroeconomic scenarios using the satellite model and the expected duration of NPL closure. From the fourth year onward (i.e., 2025), LGSR estimates were obtained by holding constant the macroeconomic scenario of the third year (2024). The estimated evolution of the vintage curves reflects the considerations mentioned above: namely, increasing monotonicity with respect to the expected time of position closure and larger percentage development for positions covered by collateral. We reiterate that this evidence is a natural consequence of the fact that secured NPL have lower LGSR values than those without collateral.

6 Conclusions

This paper proposes a methodological framework for estimating LGSR from a *forward-looking* perspective. In literature and practice, the use of direct models involving the inclusion of macroeconomic factors as explanatory variables in regressions seems widespread. Here we explored the alternative consisting of implementing a hierarchical model by leveraging conditional probability concepts. The *framework* consists of two modules: the module involving the estimation of a satellite model on the macroeconomic data and the second involving the estimation of the bank's LGSR model.

The choice of an indirect approach involves greater complexity in both estimation and forecasting, for which a closed form is generally not possible. However, this approach has several advantages that we believe to be crucial: on the one hand, the plausible applicability even to LSIs, which tend to have a low contribution to systemic data; on the other hand, the limitation of the effects of multicollinearity on the volatility of estimates.

The proposal is not intended to be definitive but represents a feasible alternative with consistent methodological basis in economic theory supported by established concepts of probability calculus. Two points of potential improvement are highlighted. The first concerns the approximation of forecasts using the *plug-in* technique, which by its nature is unable to capture the volatility of the phenomenon around the mean and detect the presence of any concentration bubbles. One possible resolution to this problem involves the implementation of a Monte Carlo simulation, which requires significant computational effort but ensures consistent estimates. A second element to be introduced is the correlation within and between segments of the classification adopted by the bank through the implementation of mixed-effects models. Regression models, such as the one proposed in subsection 3.3, have an underlying data-generating process with assumptions of independence among observations. In general, it seems implausible to assume stochastic independence among NPLs classified in each segment, i.e., the assumption of no correlation between recovery actions in a same grade and between grade is not a worth hypothesis.

In general, credit risk management models assume stochastic independence between the default event and the random variable "losses" as the basic assumption. However, a set of factors that simultaneously affect both the PD and LGDR parameters may coexist. As pointed out by Resti and Sironi (2021), the factors that could be detected have a macroeconomic nature such as, for example, the value of interest rates but also other indicators that represent the thermometer of the economic cycle. Also worth mentioning is the presence of possible knock-on and sectoral effects. The authors themselves formally illustrate in Merton's model the relationship between PD and RR. In this context, an interesting line of research lies in the attempt to integrate satellite models that are used to propagate macroeconomic effects on PD and LGDR risk parameters. Another element to be tested is to integrate the individual institution's LGSR, Danger Rate and PD models by leveraging the well-established Merton-style model theory. There are several literature contributions that consider default rate as exogenous in LGD model (see Höcht et al. (2022), Wang et al. (2020), Bruce et al. (2010), Khieu, H. D., Mullineaux et al. (2012) and Zhang (2009)), but at the same time doesn't seem to be any explorative study that

estimates the correlations between PD and LGD risk parameters. In such context, the default rates assume the role of endogenous variables in a system of equations and, therefore, can be considered as determinants of bank loan recovery rates.

References

- AIFIRM, 2016. *Il principio contabile IFRS 9 in banca: la prospettiva del Risk Manager*. Position Paper N. 8.
- Bank of Italy, 2022. *Coefficiente della riserva di capitale anticiclica (countercyclical capital buffer, CCyB).* s.l.:Decisioni di politica macroprudenziale della Banca d'Italia.
- Belotti, A. & Crook, J., 2011. *Loss given default models incorporating macroeconomic variables for cards*. International Journal of Forecasting,. In: Vol. 28, no. 1 a cura di s.l.:International Journal of Forecasting, Vol. 28, no. 1, pp. 171-182.
- Bonaccorsi Di Patti, E. & Gascarino, G., 2020. *Modelling the dynamics of nonperforming loans in Italy.* Note di Stabilità Finanziaria e Vigilanza N.19 a cura di s.l.:Banca d'Italia.
- Bruche, M. & Gonzalez-Aguado C., 2010. *Recovery rates, default probabilities, and the credit cycle.* Journal of Banking & Finance, Vol. 24, issue 4, pp. 754-764
- Ciocchetta, F., Conti, F.M., De Luca, R., Guida, I., Rendina, A. & Santini, G., 2017. *I tassi di recupero delle sofferenze.* In: Note di stabilità finanziaria e vigilanza N.7. s.l.:Banca d'Italia.
- Deepak Parmani, Y. P. & Analytics., w. f. M., 2017. *Forward-looking Perspective on Impairments using Expected Credit Loss.* s.l.:Whitepaper for Moody's Analytics.
- Dempster, A., Laird, N. & Rubin, D., 1977. *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society, Series B, 39(1), pp. 1-38.
- Fischetto, L.; Guida, I.; Rendina, A. & Santini, G., 2017. *I tassi di recupero delle sofferenze nel 2016*. Note di Stabilità Finanziaria e Vigilanza, Banca d'Italia*,* Volume 11.
- Fischetto, L.; Guida, I.; Rendina, A.; Santini, G. & Scotto di Carlo, M., 2018. *I tassi di recupero delle sofferenze nel 2017*. Note di Stabilità Finanziaria e Vigilanza, Banca d'Italia*,* Issue 13.
- Fischetto, L.; Guida, I.; Rendina, A.; Santini, G. & Scotto di Carlo, M., 2019. *I tassi di recupero delle sofferenze nel 2018.* Note di Stabilità Finanziaria e Vigilanza, Banca d'Italia, Issue 18.
- Fischetto, L.; Guida, I.; Rendina, A.; Santini, G. & Scotto di Carlo, M., 2020. *I tassi di recupero delle sofferenze nel 2019.* Note di Stabilità Finanziaria e Vigilanza, Banca d'Italia, Issue 23.
- Fischetto, L.; Guida, I.; Rendina, A.; Santini, G. & Scotto di Carlo, M., 2021. *I tassi di recupero delle sofferenze nel 2020*. Note di Stabilità Finanziaria e Vigilanza, Banca d'Italia, Issue 27.
- Gibilaro, L. & Mattarocci, G., 2006. *La selezione del tasso di attualizzazione nella stima della Loss Given Default: un'applicazione al mercato italiano* (working paper).
- Grzybowska, U. & Karwański, M., 2020. *Application of machine learning method under IFRS 9 approach to LGD modeling.* In: s.l.:Acta Physica Polonica A, pp. 116-122.
- Hamilton, J. D., 1994. *Time Series Analysis.* s.l.:Princeton University Press.
- Harvey, A. C., 1989. *Forecasting, Structural Time Series Models and the Kalman Filter.* s.l.:Cambridge University Press.
- Holmes, E. E., 2013. *Derivation of an EM algorithm for constrained and unconstrained multivariate autoregressive state-space (MARSS) models.* s.l.:arXiv:1302.3919 .
- Holmes, E. E., Ward, E. J., Scheuerell, M. D. & Wills, K., 2020. *MARSS: Multivariate Autoregressive State-Space Modeling.*. R package version 3.11.4 a cura di s.l.:s.n.
- Holmes, E. E., Ward, E. J. & Wills, K., 2021. *MARSS: Multivariate Autoregressive State-space Models for Analyzing Time-series Data.* The R Journal, 4(1), pp. 11-19.
- Hossain, A., Rigby, R. A., Stasinopoulos, M. D. & Enea, M., 2016b. *A flexible approach for modelling a proportion response variable: Loss given default.* In: s.l.:In Proceedings of the 31th International Workshop on Statistical Modelling, pp. 127-132.
- Höcht, S., Min, A., Wieczorek, J. & Zagst R. (2022). *Explaining Aggregated Recovery Rates*. Risks, MDPI, vol. 10(1) pages 1-30.
- Joubert, M., Verster, T., Raubenheimer, H. & Schutte, W. D., 2021. *Adapting the Default Weighted Survival Analysis Modelling Approach to Model IFRS 9 LGD.* s.l.:Centre for Business Mathematics and Informatics, North-West University, Potchefstroom 2531, South Africa.
- Khieu, H. D., Mullineaux D. J. & Yi, H. C., (2012). *The determinants of bank loan recovery rates*. Journal of Banking & Finance, Vol. 36, issue 4, pp. 923-933.
- Resti, A. & Sironi, A., 2021. *Rischio e valore nelle banche. Misura, regolamentazione, gestione.* s.l.:Egea.
- Rigby, R. A. & Stasinopoulos, D. M., 2005. *Generalized additive models for location, scale and shape.* In: Applied Statistics, 54. s.l.:Journal of the Royal Statistical Society.
- Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z. & De Bastiani, F., 2021. *Distributions for Modeling Location, Scale, and Shape Using GAMLSS in R.* s.l.:Chapman and Hall/CRC.
- Stasinopoulos, M., Enea, M., Rigby, R. A. & Hossain A., 2017. *Inflated distributions on the interval [0, 1].* s.l.:Working Paper at https://www.gamlss.com/the-books/.
- Wang H., Forbes C.S., Fenech J.P. & Vaz J., 2020. *The determinants of bank loan recovery rates in good times and bad - New Evidence.* Journal of Economic Behavior & Organization, vol. 177, pp. 875-897.
- Whittaker, J., 1990. *Graphical Models in Applied Multivariate Statistics.* Chichester: J. Wiley and Sons.
- Zellner, A., 1962. *An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias.* In: s.l.:Journal of the American Statistical Association, Vol. 57, No. 298, pp. 348-368.
- Zhang, Z., 2009. *Recovery Rates and Macroeconomic Conditions: The Role of Loan Covenants*. AFA 2010 Atlanta Meetings Paper, Available at SSRN: https://ssrn.com/abstract=1346163 or http://dx.doi.org/10.2139/ssrn.1346163