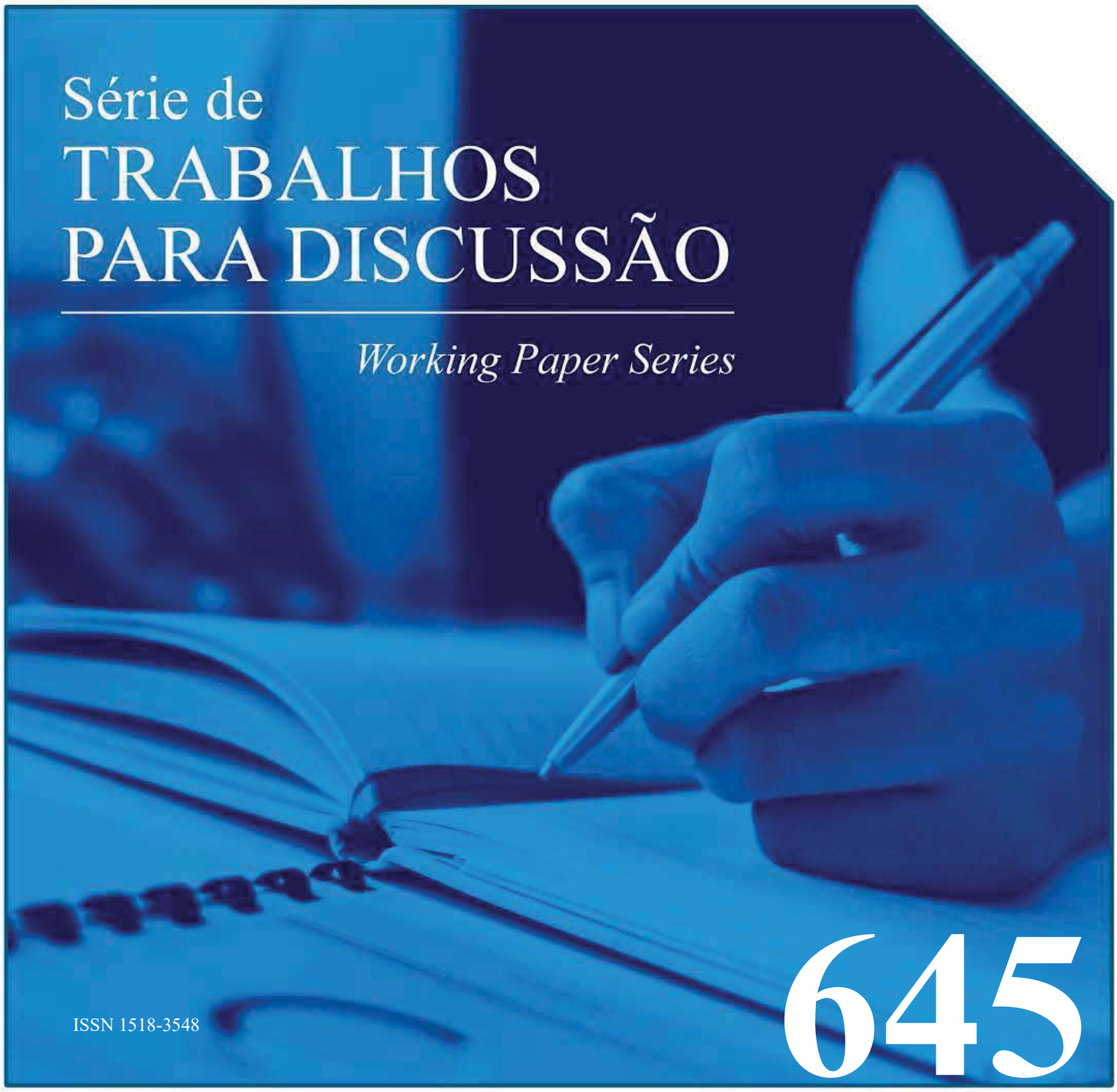


Abril 2026

# Forecasting Out-of-Time Credit Scoring Model Risk

Valter T. Yoshida Jr., Rafael Schiozer, Alan de Genaro, Toni R.E. dos Santos

BANCO CENTRAL DO BRASIL

ISSN 1518-3548
CGC 00.038.166/0001-05

| Working Paper Series | Brasília | no. 645 | Abril | 2026 | p. 3-47 |
|---|---|---|---|---|---|

# *Working Paper Series*

Edited by the Research Department (Depep) – E-mail: workingpaper@bcb.gov.br

Editor: Rodrigo Barbone Gonzalez

Co-editor: Eurilton Alves Araujo Jr

Head of the Research Department: Euler Pereira Gonçalves de Mello

Deputy Governor for Economic Policy: Paulo Picchetti

The Banco Central do Brasil Working Papers are evaluated in double-blind referee process.



**Citizen Service Division**

Banco Central do Brasil

Deati/Diate

SBS – Quadra 3 – Bloco B – Edifício-Sede – 2º subsolo

70074-900 Brasília – DF – Brazil

Toll Free: 0800 9792345

Fax: +55 (61) 3414-2553

Internet: http//www.bcb.gov.br/?CONTACTUS

## Non-Technical Summary

This study presents a procedure to identify, *ex ante*, the best-performing credit scoring model in an out-of-time setting. The procedure aims to assist end-users (practitioners and regulators) in deciding between segmented and full data models, while considering and mitigating potential misuse in these settings. We utilize the Credit Scoring Model Risk (*CSMR*) measure, introduced by Yoshida et al. (2025), along with complementary approaches, to identify the best-performing model.

The *CSMR* is estimated as one minus the correlation ($\rho_{Y,\hat{Y}}$) between the default variable ($Y$, representing the dependent variable) and its prediction ($\hat{Y}$). Yoshida et al. (2025) empirically tested the CSMR in LASSO credit scoring models. They found that adding loans from different banks to increase the number of observations is not optimal on an in-sample basis, challenging the generally accepted assumption that more data lead to better predictions.

However, evaluating model performance using in-sample data may exhibit instability in out-of-time estimation. Therefore, decision-making (choosing a model among a variety of possibilities) based exclusively on in-sample measures may be problematic because banks loan portfolios change over time, models may be uncalibrated (or not well-fitted to the current portfolio), and they can behave differently under new macroeconomic conditions or in response to exogenous and stochastic events. We show that it is crucial to compare out-of-sample correlations and consider indexes such as the Population Stability Index (PSI) to identify possible population changes and overfitted models.

This study proposes a procedure to forecast the best-performing model in out-of-time estimations. Three (complementary) approaches help model users choose between the segmented or full data models for out-of-time applications by predicting which model tends to have a higher correlation (or lower model risk).

The first approach compares estimated out-of-sample correlations based on Copas (1983)'s shrinkage concept. The second approach employs a Monte Carlo simulation that compares the models average predictions. The third approach uses Bayesian estimation of covariances.

Our primary objective was not to determine which type of model (full or segmented data) would be dominant or consistently deliver higher correlations (lower model risk) in out-of-time datasets. Our contribution is to determine how we can reliably anticipate which credit scoring model – segmented or full data – will sustain predictive performance over time, even as loan portfolios evolve and macroeconomic conditions shift.

The proposed approaches can assist practitioners and regulators in selecting and evaluating models during the decision-making process.

## Resumo Não Técnico

Este artigo apresenta um procedimento para identificar, *ex ante*, o modelo de *credit scoring* com melhor desempenho em um contexto *out-of-time*. O procedimento visa auxiliar usuários finais (profissionais e reguladores) a decidir entre modelos segmentados e modelos com dados completos, ao mesmo tempo em que considera e mitiga potenciais usos indevidos nesses cenários. Utilizamos a métrica de Risco de Modelo de *Credit Scoring* (CSMR), introduzida por Yoshida et al. (2025), juntamente com abordagens complementares, para identificar o modelo de melhor desempenho.

O CSMR é estimado como um menos a correlação ($\rho_{Y,\hat{Y}}$) entre a variável de inadimplência ($Y$, representando a variável dependente) e suas previsões ($\hat{Y}$). Yoshida et al. (2025) testaram empiricamente o CSMR em modelos LASSO de *credit scoring*. Eles descobriram que adicionar empréstimos de diferentes bancos para aumentar o número de observações não é ideal em uma base *in-sample* (dentro da amostra), desafiando a suposição geralmente aceita de que mais dados levam a melhores previsões.

No entanto, avaliar o desempenho do modelo usando dados dentro da amostra pode apresentar instabilidade na estimação *out-of-time*. Portanto, a tomada de decisão baseada exclusivamente em medidas dentro da amostra pode ser problemática porque as carteiras de empréstimos dos bancos mudam ao longo do tempo, os modelos podem estar descalibrados e podem se comportar de maneira diferente sob novas condições macroeconômicas ou em resposta a eventos exógenos e estocásticos. Mostramos que é crucial comparar correlações *out-of-sample* (fora da amostra) e considerar índices como o Índice de Estabilidade da População (PSI) para identificar possíveis mudanças populacionais e modelos sobreajustados (com *overfitting*).

Este trabalho propõe um procedimento para prever o modelo de melhor desempenho em estimativas *out-of-time* (fora do tempo). Três abordagens (complementares) ajudam os usuários do modelo a escolherem entre os modelos de dados segmentados ou completos, prevendo qual modelo tende a ter uma correlação mais alta (ou menor risco de modelo).

A primeira abordagem compara correlações estimadas fora da amostra com base no conceito de contração (*shrinkage*) de Copas (1983). A segunda abordagem utiliza uma simulação de Monte Carlo que compara as previsões médias dos modelos. A terceira abordagem emprega a estimação bayesiana das covariâncias.

Nosso objetivo principal não é determinar qual tipo de modelo seria dominante ou consistentemente entregaria uma correlação mais alta (risco de modelo mais baixo) em conjuntos de dados *out-of-time*. Nossa contribuição é o desenvolvimento de abordagens que podem ajudar o usuário final dos modelos de *credit scoring* a decidir entre os dados segmentados ou completos para cada situação específica.

As abordagens propostas auxiliam os usuários na seleção e na avaliação de modelos.

# Forecasting Out-Of-Time Credit Scoring Model Risk*

Valter T. Yoshida Jr. Rafael Schiozer Alan de Genaro

Toni R.E. dos Santos

March 31, 2026

**Abstract**

This paper addresses the challenge of forecasting the best-performing credit scoring model in out-of-time settings, focusing on the decision between segmented (bank-specific) and full data (financial system-wide) models. Building upon the Credit Scoring Model Risk (CSMR) metric, defined as one minus the correlation between observed defaults and predicted scores, we highlight the instability of in-sample performance measures when applied to evolving loan portfolios and changing macroeconomic conditions. We propose three complementary approaches to predict out-of-time model performance: (i) an analytical method based on Copas shrinkage concept utilizing estimated covariances and prediction variances; (ii) a Monte Carlo simulation leveraging average model predictions to simulate default events; and (iii) a Bayesian estimation framework for covariances grounded in conditional expectations of predictions given default. Empirical analysis using a large Brazilian loan dataset reveals that segmented models outperform full data models in in-sample contexts but not consistently out-of-time. Among the approaches, the Monte Carlo simulation achieved the highest accuracy (70.8%) in forecasting the superior out-of-time model, followed by the Bayesian method (66.7%) and the analytical shrinkage approach (54.2%). The study underscores the importance of considering population shifts via the Population Stability Index (PSI) to detect model decalibration and overfitting. The proposed methodologies offer practitioners and regulators practical tools for informed model selection, enhancing predictive reliability over time amid portfolio and economic dynamics.

Keywords: Model Risk; Model Selection; Credit Risk; Credit Scoring; Big Data; Machine Learning.

JEL Classification: C52;C55.

*The views expressed in this Working Paper are those of the authors and do not necessarily reflect those of the Banco Central do Brasil.*

*Yoshida Junior: Banco Central do Brasil, valter.yoshida@bcb.gov.br. Dos Santos: Banco Central do Brasil, toni.santos@bcb.gov.br. Schiozer: FGV, rafael.schiozer@fgv.br. Genaro, FGV, alan.genaro@fgv.br.We thank Banco Central do Brasil, Clodoaldo Annibal, Eduardo Kazuo Kayo, Eduardo Vieira Paiva, Fernando Chertman, Guilherme Yanaka, Gustavo França, Leonardo Alencar, Leonardo Rondon, Sergio Koyama, Theo Martins, Tony Takeda, Vinicius Brunassi, the WPS anonymous referee and VII Workshop da Rede de Pesquisa do Banco Central. Rafael Schiozer acknowledges the financial support from Fapesp and CNPq. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES) Finance Code 001.

# 1 Introduction

Quantifying model risk in credit-scoring models is essential for effective risk and capital management by financial intermediaries. Yoshida et al. (2025) showed that models trained on more specific data subsets capture patterns more efficiently, resulting in lower model risk, $i.e.$, a higher correlation between the dependent variable and predictions. However, segmentation also limits the models ability to learn broader, system-wide patterns. In contrast, a model trained on a more diverse dataset (such as the entire financial system) may capture a wider range of structural relationships, although its predictive accuracy within individual subsets may be lower. This reflects the bias-variance tradeoff: specialization reduces local bias but increases variance out of segment, whereas broader models reduce global bias but can suffer from higher variance (Hastie, Tibshirani, and Friedman, 2009). Consequently, performance metrics based solely on in-sample data may become unstable when evaluated in out-of-time (later-period) samples. Using in-sample model risk measures to determine which model to deploy among several candidates can therefore be problematic, especially as loan portfolio characteristics evolve over time. Moreover, models may be uncalibrated even at the beginning of their application ($i.e.$, not properly fitted to the current portfolio), and their behavior can vary under different macroeconomic conditions, exogenous shocks, and stochastic events.

This study proposes a procedure to identify, *ex ante*, the best-performing model in an out-of-time context. The procedure assists end users in choosing between segmented and full-data models. As in Yoshida et al., 2025, we compare models estimated using data from a single bank with models estimated using data from the entire financial system.[1] However, unlike that paper, we explicitly address out-of-time performance while mitigating potential model misuse. We employ the Credit Scoring Model Risk (*CSMR*) measure, introduced by Yoshida et al. (2025), together with complementary approaches, to identify the best-performing specification between full and segemented data models. The *CSMR* is defined as one minus the correlation ($\rho_{Y,\hat{Y}}$) between the observed default ($Y$) and its prediction ($\hat{Y}$). To evaluate the proposed procedure, we use a large contract-level dataset comprising approximately 200,000 loans per month to micro and small businesses (MSBs) granted by 425 financial institutions in Brazil between 2014 and 2017.

One of the main concerns in predictive models, such as credit scoring models, is overfitting. Overfitting implies that the in-sample (or training) set performs significantly better than the out-of-sample (or holdout) set (Shmueli, 2010). If a model is trained (or fitted) too well on the in-sample dataset, it tends to perform poorly on new, unseen data. Overfitted models fit too closely to noise and randomness in the in-sample data (Ying, 2019), leading to effectiveness loss in databases other than the training set. Unstable performance over time can be a sign of overfitting. This can be caused by changes in the underlying data generation process (for example, due to changes in the bank's credit granting policies) or the relationship between the explanatory variables and the response variable (in our case, the default indicator in a credit

[1]The procedures developed in this paper can be applied more broadly to compare models estimated using on a full loan portfolio with models estimated using any subset of that portfolio. For instance, they may be used to contrast a model estimated with all loans originated by a bank with models estimated using loans from a single geographic region or from a specific loan category (such as receivables-backed loans or working capital loans).

scoring model), due to changes in the macroeconomic setting. Indeed, in-sample measures tend to be very optimistic about the error rate (Hastie, Tibshirani, and Friedman, 2009).

Our proposed procedure for establishing out-of-time credit scoring model risk comparisons consists of three steps: i) calculating the correlation ($\rho_{Y,\hat{Y}}$) on an in-sample basis (or the credit portfolio at time $t_0$); ii) estimating predictions for the out-of-time sample (or the credit portfolio at time $t_1$); and iii) comparing the predicted correlation in out-of-time samples using three exploratory approaches.

The first approach for step (iii) is an analytical comparison of potential out-of-time correlations based on Copas (1983)'s shrinkage concept and on the estimated models' covariances ($\mathrm{Cov}[Y, \hat{Y}]$). The second is a Monte Carlo simulation based on the average of the predictions of the compared models. The third approach is based on the Bayesian estimation of covariances.

When using the first approach for decision-making, we selected models with higher out-of-time correlation (or lower *CSMR*) in 26 out of 48 (eight months and six cross-sectional banks) model comparisons, which corresponds to 54.2%, the same percentage as selecting the model with lower in-sample *CSMR* (or higher correlation). Applying the second approach to decision-making, higher out-of-time correlation models were chosen 34 times, corresponding to 70.8% of our empirical comparisons. Finally, when using the third approach, models with higher correlations were chosen 32 times, representing two-thirds of the comparisons.

This study does not seek to deterministically identify which type of model (full or segmented data) is superior or likely to produce consistently higher correlations. The model with a higher correlation in out-of-time datasets depends on various factors, such as the nature of the problem, data structure, sampling, and vector of independent variables. This study contributes to the development of approaches to aid in deciding, for each specific situation, the model most likely to have a lower model risk based on higher correlations between future observable defaults and predicted scores.

Research on credit scoring (Wang et al., 2011; Zhou, Lai, and Yen, 2014; Barboza, Kimura, and Altman, 2017; Oskarsdóttir et al., 2019; Agarwal et al., 2020; Huang et al., 2020; Alonso and Carbó, 2021) has largely relied on traditional performance metrics[2] and has emphasized technique and variable selection as well as in-sample accuracy, while largely overlooking predictive performance over time and the risks associated with in-sample overfitting.

This study contributes to the literature by proposing complementary approaches for comparing models and supporting decision-makers, particularly in Big Data contexts where practitioners and academics face increasing difficulty in selecting the most appropriate model among numerous alternatives. We address a gap in the literature by examining in-sample overfitting in machine learning models—particularly LASSO—estimated on large datasets, and by proposing methods to identify the model that is most likely to exhibit higher correlation in out-of-time samples.

Although some studies have incorporated out-of-time performance evaluation in credit scoring models (Doumpos and Pasiouras, 2005; Malik and L. C. Thomas, 2012; Provenzano et

[2]Credit risk models are typically evaluated using measures such as the KolmogorovSmirnov statistic (KS), the Area Under the Receiver Operating Characteristic Curve (AUC), or likelihood-based indices such as the Mahalanobis Distance (L. Thomas, Crook, and Edelman, 2017). Yoshida et al. (2025) introduce the Credit Scoring Model Risk (CSMR), based on Barrieu and Scandolo (2015).

al., 2020; Maldonado, López, and Iturriaga, 2022), they have important limitations. Doumpos and Pasiouras (2005) relies on a relatively small dataset, and Malik and L. C. Thomas (2012) does not employ machine learning techniques. While Provenzano et al. (2020) and Maldonado, López, and Iturriaga (2022) use out-of-time samples, their approaches do not allow for direct model comparison or for identifying the specification with the highest predictive correlation or, equivalently, the lowest model risk in out-of-time settings.

We also build directly on Yoshida et al. (2025), who found that segmented data models present lower model risk than full-data models for in-sample applications. These results suggest that decision-makers should choose to develop (and use) segmented data models. Although it is an intuitive argument, these results do not necessarily hold for out-of-time datasets.

In addition to this introduction, this paper has six sections: Section 2 presents the literature review. Section 3 presents an overview of the empirical setup. Alternative approaches to compare models on an out-of-time basis are presented in Section 4. Section 5 presents the data sources and outlines the rationale for the empirical applications. Results are in Section 6, and concluding remarks are in Section 7.

## 2 Literature

The principle of parsimony, or *Occam's Razor*[3] suggests using the fewest number of predictors necessary – nothing more. Overfitting occurs when models or procedures violate parsimony by employing a more flexible model than required or by including irrelevant components or too many predictors. Coefficients fitted to irrelevant predictors can introduce random variations and lead to poor predictions. Although overfitting is a widely recognized concern, it is not an absolute characteristic but a relative one, entailing a comparison, which is often overlooked (Hawkins, 2004). Furthermore, the finance literature does not extensively address overfitting in machine learning models (Barboza, Kimura, and Altman, 2017).

The definition of the training data and its variables affects the model's ability to minimize the loss function (such as *MSE* – mean squared error) and influences its performance. If the in-sample data are highly specific and closely match the real historical data, the model can better capture the patterns and relationships within that specific data and achieve a lower loss (and a higher correlation between observations and predictions). However, if the in-sample data are more heterogeneous and include a wider range of cases, the model may be able to capture a wider range of patterns, but it may also be more difficult to accurately predict specific subsets of the data. This is the biasvariance tradeoff (Hastie, Tibshirani, and Friedman, 2009).

The relationship between training data diversity and model performance can be complex and may depend on factors such as model complexity and the amount of data available for training (Dastile, Celik, and Potsane, 2020 Alonso and Carbó, 2021). Predictive performance further depends on the nature of the problem, underlying data structure, sampling design, set of independent variables employed, and classification framework adopted (Duénez-Guzmán and Vose, 2013; Huang et al., 2020).

[3] A principle stated by the philosopher William of Ockham: "Plurality should not be posited without necessity." https://www.britannica.com/topic/Occams-razor, accessed on March 20, 2025.

Cross-validation using datasets from the same in-sample period (out-of-sample) offers a warning sign for overfitting (Hastie, Tibshirani, and Friedman, 2009). However, out-of-time ($i.e.$, from later periods) predictions remain challenging, as relationships and estimated coefficients may no longer be well-calibrated to new data.

The Least Absolute Shrinkage and Selection Operator (LASSO) is a widely used method for modeling high-dimensional data. By introducing a penalty term, LASSO reduces model complexity, mitigates overfitting, and improves generalization performance (McNeish, 2015). However, LASSO may also limit the models ability to capture group-specific patterns and does not address the instability of the model performance across out-of-time datasets.

Out-of-time datasets can be viewed as population partitions, just as cross-sectional represents partitions of panel data model. Consequently, the regression coefficients may vary across reference dates. As formalized by Simpsons paradox, associations between two variables may reverse in subpopulations when conditioned on a third correlated variable.[4]

When comparing in-sample and out-of-time results, the reference period may act as a confounding factor, potentially reversing associations. Therefore, in credit scoring, where temporal generalization is crucial, out-of-time performance offers a more reliable basis for comparing competing models than in-sample accuracy.

The Population Stability Index (PSI) is used to assess distributional shifts in population scoring over time. This could be an indication of the out-of-time model's decalibration, and it is estimated by the Kullback-Leibler Divergence, which measures the difference between two probability distributions. Assuming that prediction distributions are normally distributed, *Divergence* is defined (L. Thomas, 2009) as

$$\begin{aligned} \text{Divergence} = D &= \int \left(f\left(s|A\right) - f\left(s|B\right)\right) \log\left(\frac{f\left(s|A\right)}{f\left(s|B\right)}\right) dS \\ &= \frac{1}{2}\left(\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2}\right)\left(\mu_A - \mu_B\right)^2 + \frac{\left(\sigma_A^2 - \sigma_B^2\right)^2}{2\sigma_A^2\sigma_B^2} \end{aligned} \tag{1}$$

where $f\left(s|A\right)$ and $f\left(s|B\right)$ denote the distributions of predictions $s$ conditional on $A$ and $B$;
$\sigma_A^2$ and $\sigma_B^2$ are, respectively, the variance of predictions given $A$ and $B$; and
$\mu_A$ and $\mu_B$ are the averages of the predictions given $A$ and $B$.

We consider $\text{Divergence}$ measure as portfolio composition changes within each model, or the Population Stability Index (PSI), which captures the distance between in-sample $(A)$ and out-of-time $(B)$ prediction distributions. If the PSI is indicative of model decalibration, models with high PSI are not suitable for decision-making.

However, although the PSI can identify population distribution shifts[5], it cannot be guaranteed that the correlation between observable defaults and their predictions remains the same

[4]See Simpson (1951) and the formal discussion in Cartwright (1979). In linear regression contexts, this phenomenon arises when partial regression coefficients differ in sign from bivariate estimates (Samuels, 1993). Population partitioning is therefore central to the measurement of associations, as illustrated by the credit-scoring application in Yoshida et al. (2025).

[5]As an empirical rule of thumb, L. Thomas (2009) consider a Population Stability Index (PSI) lower than 0.1 as indicative of a stable population, PSI greater than 0.1 as signaling potential population change, and PSI greater than 0.25 as indicative of significant changes in the population. In this study, we treated PSI values above 0.1 as indicative of model decalibration. An uncalibrated model that produces unstable results should not be considered a viable alternative.

(nor does it reveal whether the comparison of the correlations of in-sample models still holds) across different periods.

For our empirical analysis, we employ the credit-scoring model risk metric proposed by Yoshida et al. (2025), which adapts the relative model risk measure of Barrieu and Scandolo (2015) originally developed for market risk. This metric is based on the correlation ($\rho_{Y,\hat{Y}}$) between the observed outcomes and their model predictions. The resulting Credit Scoring Model Risk (*CSMR*) is defined as:

$$\text{CSMR} = 1 - \left|\rho_{Y,\hat{Y}}\right|. \tag{2}$$

Similar to traditional performance measures (such as KS, the AUC, and the Mahalanobis Distance), model comparisons based on *CSMR* are not necessarily stable over time; that is, a model that carries a comparatively lower *CSMR* (or higher $\rho_{Y,\hat{Y}}$) in development samples (in-sample) may have higher *CSMR* in out-of-time samples. Therefore, we developed three approaches to assist credit risk managers in choosing between different models. We exemplify these approaches by applying them to the choice between *segmented data* (in our case, data from a single bank) and a *full-data* model (in our case, a model using loan data from the entire financial system).

## 3 Empirical Setup and Models' Specifications

In this section, we briefly review the empirical setup and the model framework outlined in Yoshida et al. (2025). The study considers two alternative specifications designed to compare models estimated using full data with those based on segmented subsets. Both specifications rely on Least Absolute Shrinkage and Selection Operator (LASSO) regressions, allowing for variable selection and regularization in high-dimensional settings.

For each March and September between 2014 and 2017, one full data model and six segmented data models were estimated. This resulted in eight full data models and 48 segmented data models, yielding a total of 56 models. The segmented models were estimated separately for each of the five largest banks by the number of loans, with the remaining institutions grouped together.

To develop the full data model, the first specification uses loans from all banks and a group of covariates and fixed effects.

$$Y_{l,i.b,g} = \alpha + \beta X_{l,i,b} + \gamma_b + \delta_g \tag{3}$$

where $Y$ is a binary dependent variable that takes the value one if the loan is overdue for 90 days or more and zero otherwise;
the subscripts $l$ refers to loan, $i$ to borrower, $b$ to bank, and $g$ to location;
$\alpha$ is a constant that is to be estimated;
$\beta$ is a vector of coefficients to be estimated for the covariate vector $X$, which comprises 114 variables drawn from the credit bureau of the Central Bank of Brazil (SCR) and linked data sources, including employment data from a Brazilian administrative database on formal employment

and firm-level data from the Brazilian Revenue Service (RFB). These variables vary at the loan, borrower, and time levels.[6]

$\gamma_b$ is a bank-fixed effect; and

$\delta_g$ is a geographical location (borrower's head office) fixed effect.

From a practitioners perspective, if one must predict scores for the loans of a specific bank, is it better to estimate the model using data on loans from that specific bank or also using loans from other banks? To answer this question, the first specification model's results are compared to the other six models' results, one for each of the largest five banks and one for the remaining banks in the financial system.

The second specification is a variation of the first specification (Equation 3), but it segments the database by bank:

$$Y_{l,i,g} = \alpha + \beta X_{l,i} + \delta_g \tag{4}$$

The conditional correlations between default and its predictions for each bank ($\rho_{Y,\hat{Y}_A}|b$) are computed and then compared to the correlations of the segmented data models ($\rho_{Y,\hat{Y}_B}$).

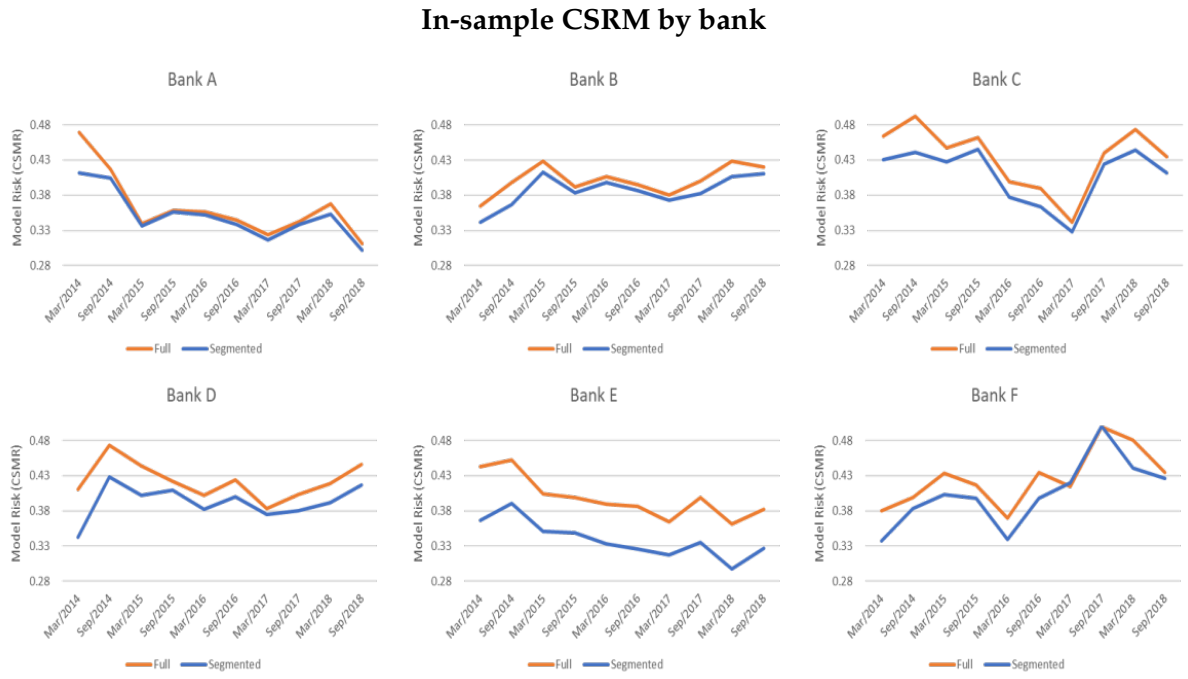


**Figure 1: In-sample Credit Scoring Model Risk *CSMR* by bank**

Note: In each panel, the orange line represents the conditional *CSMR*s of the full data models, while the blue line represents those of the segmented models. In the in-sample datasets, the segmented models (by bank) exhibit lower model risk than the full-data models. Source: Yoshida et al. (2025).

As illustrated in Figure 1, Yoshida et al. (2025) conclude that segmented models exhibit lower *CSMR*s (*i.e.*, higher correlations) in in-sample datasets, indicating that segmentation improves in-sample performance, while the benefits of increasing the number of observations

[6]Each final LASSO specification relies on a vector of 114 explanatory variables, including original and transformed predictors from the SCR database, six employment-related variables, and firm-level characteristics obtained from the Brazilian Revenue Service, such as industry classification, firm size, share capital, and geographic identifiers.

show diminishing returns. This result is consistent with Alonso and Carbó (2021), who show that although classification performance improves with larger sample sizes, the marginal gains for the Logit and LASSO models diminish once the sample size exceeds approximately 10,000 observations.

However, neither Yoshida et al. (2025) nor Alonso and Carbó (2021) address how the models would behave in out-of-time datasets, given transformations in the composition of the credit portfolio or broader economic changes that affect default risk.

Because the *CSMR* measure is a function of the correlation between default observations and their predicted scores, we claim that it is possible to make inferences on the *CSMR* behavior based on out-of-time predictions and its statistical properties. The following section presents three proposed approaches for predicting which model will exhibit a higher out-of-time correlation between observed defaults and scores (or a lower *CSMR*).

# 4 Methodology – Approaches to compare models in out-of-time datasets

In this section, we propose a procedure to identify the best-performing model (full or segmented data model) in out-of-time datasets. The procedure is defined by three sequential steps, as follows:

i) Computing the correlations ($\rho_{Y,\hat{Y}}$) between observed defaults and models' predictions on each in-sample dataset (or the credit portfolio at time $t_0$) for both models[7];

ii) Estimating the predictions from out-of-time datasets (or the credit portfolio at time $t_1$) for both models; and

iii) Comparing the forecasted correlation in out-of-time datasets with three alternative approaches, which we describe in what follows.[8]

## 4.1 First approach – Similar shrinkage

Inspired by Copas (1983) and Copas (1987) shrinkage[9] concept, the first proposed approach involves an analytical comparison of potential out-of-time correlations between default observations and their predicted scores ($\rho_{Y,\hat{Y}}$) based on the estimated covariances ($\mathrm{Cov}[Y,\hat{Y}]$) of the models.

When dealing with out-of-time datasets ($i.e.$, a current loan portfolio or a matrix of vectors $X$, and their default predictions, a vector $\hat{Y}$), it is not possible to compute *ex ante* the correlations between $Y$ and $\hat{Y}$ because $Y$ is not yet observable. However, we argue that it is possible to forecast whether a full data model's correlation would be higher than a segmented model

[7]The correlation between default variable ($Y$) and its predictions ($\hat{Y}$) is conditional to each bank in full data models.

[8]A fourth approach is presented in Appendix A. It is not presented in the main body of the manuscript because it does not offer a reliable approach to forecast the higher correlation model, although it offers interesting inferences on the relevance (Czasonis, Kritzman, and Turkington, 2022) of observations.

[9]In another context, a similar principle of shrinkage was used by Tibshirani (1996) in the development of the Least Absolute Shrinkage and Selection Operator (LASSO).

correlation if we know (or can make reliable estimates of) i) the covariances between future observed values and their predictions and ii) the predictions' standard deviations of both models.

**Lemma 1** *The correlation* ($\rho_{Y,\hat{Y}_A}$) *between the observed default indicator* $Y$ *and the prediction* $\hat{Y}_A$ *from model* $A$ *exceeds the corresponding correlation for model* $B$ ($\rho_{Y,\hat{Y}_B}$) *on the same sample if and only if the ratio of the covariance between* $Y_i$ *and the predictions* $\hat{Y}_i$ *to the standard deviation of the predictions,*

$$\frac{\text{Cov}(Y, \hat{Y}_i)}{\sigma_{\hat{Y}_i}},$$

*is larger for model* $A$ *than for model* $B$*.*

$$\frac{\text{Cov}[Y, \hat{Y}_A]}{\sigma_Y \sigma_{\hat{Y}_A}} = \rho_{Y,\hat{Y}_A} > \rho_{Y,\hat{Y}_B} = \frac{\text{Cov}[Y, \hat{Y}_B]}{\sigma_Y \sigma_{\hat{Y}_B}} \iff \frac{\text{Cov}[Y, \hat{Y}_A]}{\sigma_{\hat{Y}_A}} > \frac{\text{Cov}[Y, \hat{Y}_B]}{\sigma_{\hat{Y}_B}} \tag{5}$$

Covariances are not known *ex ante*. Thus, we need to make an assumption to estimate these values.

Following the definition of overfitting given in Section 2, comparing the correlation between observed values and their corresponding predictions ($\rho_{Y,\hat{Y}}$) in an out-of-sample dataset ($i.e.$, a hold-out sample from the same period as the in-sample data but not used for model training) can help identify overfitted models. Considering that out-of-time datasets have the same fit as out-of-sample datasets, the coefficient ($\beta$) of a regression of default on the model's predictions ($Y = \alpha + \beta\hat{Y}$) will remain the same as the coefficient of the out-of-sample dataset.

**Assumption 1** *If we consider that a model* $m$ *has the same fit in out-of-time* ($t_1$) *and out-of-sample* ($t_0^{OoS}$) *datasets, then* $\beta_m$ *in a regression of observed values on their predictions* ($Y = \alpha_m + \beta_m \hat{Y}_m$) *remains the same as on the out-of-sample dataset: if model* $m$ *has the same fit in out-of-time and out-of-sample datasets then*

$$\beta_m^{t_1} = \beta_m^{t_0^{OoS}} \tag{6}$$

At first sight, Assumption 1 may appear restrictive, as it implicitly requires that the relationship estimated in-sample remains stable when applied to new (out-of-sample) data. However, this type of assumption is quite common in decision-making contexts, especially in predictive modeling, where one typically relies on past data as a proxy for future behavior.[10] The difference between one and the coefficient of the regression of the default variable on the model's predictions in the out-of-sample datasets ($1 - \beta_m^{t_0^{OoS}}$) is called shrinkage by Copas (1983), or the amount by which the out-of-sample fit falls short of the in-sample fit. The coefficient is the estimation of the scale factor by which the predictor should be multiplied to correct the overprediction of the data of interest Copas (1987). The higher the shrinkage, the higher the overfitting.[11]

[10]Models typically rely on past data to predict (or explain) future (or present) events. If there are indicators that a model may not fit as well in out-of-time (or out-of-sample) data as it does in the in-sample dataset, it may be worthwhile to investigate the degree to which the model becomes uncalibrated, or the shrinkage. One possible approach was proposed by Malik and L. C. Thomas (2012) using Markov chains. Another approach, based on an autoregressive model of correlation is outlined in Appendix B.

[11]The $\beta_m^{t_0^{IS}}$, or the $\beta_m$ in a regression of observed values on their predictions ($Y = \alpha_m + \beta_m \hat{Y}_m$) in in-sample datasets, is, by construction, equal to one. Thus, the shrinkage of in-sample datasets ($1 - \beta_m^{t_0^{IS}}$) is equal to zero.

**Corollary 1.1** *Assuming that the models being compared exhibit the same level of fit* (*Assumption 1*), *it follows that the estimated covariances* ($\hat{\text{Cov}}[Y, \hat{Y}_m]$) *are equivalent to* $\beta$ *multiplied by the variance of the predictions* $\sigma^2_{\hat{Y}_m}$ *:*

$$\hat{\text{Cov}}[Y, \hat{Y}_m] = \beta_m \sigma^2_{\hat{Y}_m} \tag{7}$$

**Corollary 1.2** *Considering Lemma 1 and Corollary 1.1, then model A correlation* ($\rho_{Y,\hat{Y}_A}$) *will be higher than an alternative model B correlation* ($\rho_{Y,\hat{Y}_B}$) *for the same sample if* $\beta$ *times the predictions' standard deviation is also higher:*

$$\rho_{Y,\hat{Y}_A} > \rho_{Y,\hat{Y}_B} \iff \beta_A \sigma_{\hat{Y}_A} > \beta_B \sigma_{\hat{Y}_B} \tag{8}$$

Corollary 1.2 provides a proxy for comparing out-of-time sample correlations, which is particularly relevant in credit-scoring applications. We do not need to infer the standard deviation of future default observations ($\sigma_Y$), since this term affects both models symmetrically. As a result, the default rate itself is not determinative for identifying which model achieves the higher correlation.

Portfolio composition changes[12], model decalibration over time or stochastic event fluctuations[13] can alter comparisons when out-of-time samples are the focus, as each of these factors can modify the relationship between predicted and realized defaults.

**Corollary 1.3** *The out-of-time correlation between observed values and their corresponding predictions* ($\rho_{Y,\hat{Y}}$) *can be written as:*

$$\rho_{Y,\hat{Y}_A} = \beta_A \frac{\sigma_{\hat{Y}_A}}{\sigma_Y} \tag{9}$$

Based on Corollary 1.3, the first approach is based on the following two-step procedure (in two steps):

1. For each model, we estimate $\beta_m^{t_0^{OoS}}$, or the covariance of observed values and their correspondent predictions over the variance of predictions in out-of-sample datasets ($\beta_m^{t_0^{OoS}} = (\hat{\text{Cov}}^{t_0^{OoS}}[Y, \hat{Y}_m])/(\sigma^{2OoS}_{\hat{Y}_m})$).

2. Considering Corollary 1.1, we compare, for each pair of models, the product between the estimated coefficient $\beta_m^{t_0^{OoS}}$ obtained in Step 1 and the standard deviation of the model's predictions in out-of-time period $\sigma^{t_1}_{\hat{Y}_m}$. This product serves as a proxy for correlation between $Y$ and $\hat{Y}_m$. Thus, we consider that a model with a higher product is expected to have a higher correlation.

Following Assumption 1, the first approach assumes that the compared models exhibit the same fit in the out-of-time period ($t_1$) and in the out-of-sample dataset ($t_0^{\text{OoS}}$) datasets, or at least the same degree of decalibration, an assumption that may not always hold in practice.

[12]Portfolio composition changes do not necessarily imply model decalibration. Nonetheless, portfolio composition changes indicate that decalibration is more likely because the covariance matrix of the models may change.

[13]We can express changes in portfolio composition as the standard deviation of the model's predictions ($\sigma_{\hat{Y}_A}$); the degree of model decalibration as the coefficient ($\beta_A$) obtained from a regression of observed values on its predictions; and the fluctuations in stochastic events as the standard deviation of defaults ($\sigma_Y = \sqrt{\text{PD}\,(1 - \text{PD})}$).

### 4.2 Second approach – Monte Carlo simulation

The second approach does not depend on the assumption of calibrated models (or on Assumption 1, nor on a better estimation of $\beta_m^{t_1}$ or on the shrinkage). This entails a comparison of Monte Carlo simulations, considering the average predictions of the compared models as the probability default function.

This approach is based on the following four-step procedure:

1. To simulate the stochastic default of each loan in the out-of-time dataset, we take the average of the models' predictions (in each month) as the probability of the default function. The average of the models' predictions, in comparison, has, by construction, the same distance from the predictions themselves. Considering this average as a probability of default (PD), we can simulate the out-of-time correlations of default and predictions ($\rho_{Y,\hat{Y}}$).

2. For each loan, we generate a realization from a uniform $U(0,1)$ distribution. If this realization is smaller than the average predicted probability of default obtained in Step 1, the loan is classified as defaulted in the simulated round ($Y = 1$); otherwise, it is classified as non-defaulted ($Y = 0$).

3. In each simulation round, we compute the correlation between the simulated default outcomes obtained in Step 2 and the predicted scores ($\rho_{Y,\hat{Y}}$).

4. For each bank, we compare the *CSMR* of full and segmented models using a square scatter plot of simulated rounds. The horizontal axis reports the *CSMR* from the full data models, while the vertical axis reports the corresponding *CSMR* for the segmented models. Points above the 45-degree line indicate a lower *CSMR* for the segmented model, whereas points below the diagonal favor the full model. When observations cluster around the 45-degree line, relative performance depends on stochastic default realizations; in such cases, we rely on the median difference between simulated correlations, $\rho_{\text{Simulated Y, Full } \hat{Y}} - \rho_{\text{Simulated Y, Segmented } \hat{Y}}$, to identify the model with lower expected *CSMR*.

The main limitation of the second approach lies in the construction of the simulation process itself. The actual default behavior of the portfolio may differ from the average of the model predictions, as some models may capture out-of-time default dynamics more accurately than others. In other words, the true out-of-time probability-of-default function may be closer to that implied by a particular model. However, if each models predictions are treated as the probability-of-default function itself, it becomes possible to identify stochastic dominance effects, should they arise ($i.e.$, even when simulations are based on one models predicted probabilities, another model may still exhibit a higher simulated correlation).

### 4.3 Third approach – Bayesian estimation of covariances

Following Lemma 1, we can forecast correlations to assess which of the two competing models, denoted $A$ and $B$, is expected to exhibit a higher estimated correlation with the observed

outcome. This comparison is feasible whenever we can predict covariances or, equivalently, estimate the expected values of model predictions for loans that eventually default under each specification. This equivalence arises because the covariance can be expressed as a function of the expected values.

In the case of models with a binary dependent variable, the covariance between the observed outcome and model predictions depends on (i) the expected value of the predictions conditional on the occurrence of the event, $\mathrm{E}[\hat{Y} \mid Y = 1]$; (ii) the unconditional expected value of the predictions, $\mathrm{E}[\hat{Y}]$; and (iii) the expected value of the observed default indicator, $\mathrm{E}[Y]$.

Rearranging Lemma 1, we write Corollary 1.4:

**Corollary 1.4** *Model A correlation will be higher than an alternative model B correlation for the same sample* ($\rho_{Y,\hat{Y}_A} > \rho_{Y,\hat{Y}_B}$) *if the ratio between covariances of models A and B* ($(\mathrm{Cov}[Y,\hat{Y}_A])/(\mathrm{Cov}[Y,\hat{Y}_B])$) *is higher than the ratio between their standard deviations* ($(\sigma_{\hat{Y}_A})/(\sigma_{\hat{Y}_B})$)*:*

$$\rho_{Y,\hat{Y}_A} > \rho_{Y,\hat{Y}_B} \iff \frac{\mathrm{Cov}[Y,\hat{Y}_A]}{\mathrm{Cov}[Y,\hat{Y}_B]} > \frac{\sigma_{\hat{Y}_A}}{\sigma_{\hat{Y}_B}} \tag{10}$$

**Axiom 1** *In binary-response modelssuch as default prediction, where* $Y \in \{0,1\}$*, the covariance between the observed outcome and the model's predict score* ($\mathrm{Cov}\left[Y,\hat{Y}_m\right]$) *can be decomposed into terms that depend on the conditional expectation of predictions for defaulted observations* ($\mathrm{E}\left[\hat{Y}_m|Y=1\right]$)*, the unconditional expectation of the predictions* ($\mathrm{E}[\hat{Y}_m]$)*, and the unconditional expectation of the observed events, or the probability of default* ($\mathrm{E}[Y]$)*:*

$$\begin{aligned}
\mathrm{Cov}\left[Y,\hat{Y}_m\right] &= \mathrm{E}\left[Y\hat{Y}_m\right] - \mathrm{E}\left[Y\right]\mathrm{E}\left[\hat{Y}_m\right] \\
&= 1 \times \mathrm{E}\left[\hat{Y}_m \mid Y=1\right] P\left(Y=1\right) \\
&+ 0 \times \mathrm{E}\left[\hat{Y}_m \mid Y=0\right] P\left(Y=0\right) - \mathrm{E}\left[Y\right]\mathrm{E}\left[\hat{Y}_m\right] \\
&= \mathrm{E}\left[Y\right] \times \left(\mathrm{E}\left[\hat{Y}_m \mid Y=1\right] - \mathrm{E}\left[\hat{Y}_m\right]\right)
\end{aligned} \tag{11}$$

Substituting Axiom 1 in Corollary 1.4, we write Corollary1.5:

**Corollary 1.5** *Model A correlation* ($\rho_{Y,\hat{Y}_A}$) *will be higher than an alternative model B correlation for the same sample* ($\rho_{Y,\hat{Y}_B}$) *if the ratio between the differences of the conditional expected value and the full expected value for each one of the models* ($(\mathrm{E}\left[\hat{Y}_A|Y=1\right] - E\left[\hat{Y}_A\right])/(\mathrm{E}\left[\hat{Y}_B|Y=1\right] - E\left[\hat{Y}_B\right])$) *is higher than the ratio between their standard-deviations* ($(\sigma_{\hat{Y}_A})/(\sigma_{\hat{Y}_B})$)*:*

$$\rho_{Y,\hat{Y}_A} > \rho_{Y,\hat{Y}_B} \iff \frac{\mathrm{E}\left[\hat{Y}_A|Y=1\right] - E\left[\hat{Y}_A\right]}{\mathrm{E}\left[\hat{Y}_B|Y=1\right] - E\left[\hat{Y}_B\right]} > \frac{\sigma_{\hat{Y}_A}}{\sigma_{\hat{Y}_B}} \tag{12}$$

By splitting the predicted values into ranges (or risk ranks), it is reasonable to consider the in-sample default rate of each range, $P_{iS}\left(Y=1|\mathrm{range}=r\right)$, as a parameter to predict the out-of-time conditional probability of a range of given observations being defaulted, $P_{iS}\left(\mathrm{range}=r|Y=1\right)$.

**Axiom 2** *Using Bayes' Theorem, we can derive the out-of-time conditional probability of a range $r$ given observations are defaulted* ($P_{OoT}\left(\text{range} = r|Y = 1\right)$)*:*

$$P_{OoT}\left(\text{range} = r|Y = 1\right) = \frac{P_{iS}\left(Y = 1|\text{range} = r\right) \times P_{OoT}\left(\text{range} = r\right)}{P_{OoT}\left(Y = 1\right)} \tag{13}$$

Using the out-of-time conditional probability of a range given loans will be defaulted ($P_{OoT}\left(\text{range} = r|Y = 1\right)$), we can estimate the expected value of prediction given defaulted loans ($\text{E}\left[\hat{Y}_m|Y = 1\right]$).

**Axiom 3** *Using Bayes' theorem, the expected value of the predicted score estimated by a model* ($\hat{Y}_m$)*, conditional on a loan being in default* ($\text{E}\left[\hat{Y}_m|Y = 1\right]$)*, can be expressed as the weighted average of the conditional expected value within each range* ($\text{E}\left[\hat{Y}_m|\text{range} = r\right]$)*, where the weights are given by the conditional probability of each range in out-of-time sample* ($P_{OoT}\left(\text{range} = r|Y = 1\right)$)*:*

$$\text{E}\left[\hat{Y}_m|Y = 1\right] = \sum_{r=1}^{n} \frac{P_{OoT}\left(\text{range} = r|Y = 1\right) \times \text{E}\left[\hat{Y}_m|\text{range} = r\right]}{n} \tag{14}$$

where $n$ is the number of ranges for each model.

The third approach is based on the following four-step procedure (in four steps):

1. For each in-sample full or segmented model, we assign the observations in the dataset into score ranges[14], by determining thresholds for each of them.

2. For each score range ($r$), we assume the default rates for in-sample datasets, $P_{iS}\left(Y = 1|\text{range} = r\right)$, as the percentual frequency of each range.

3. For each out-of-time full or segmented model, we assign the observations in the dataset into score ranges according to step 1 thresholds.

4. For each score range ($r$) and following Axiom 2, we assume the conditional default probability of a range of given observations will be defaulted in out-of-time datasets,$P_{OoT}\left(\text{range} = r|Y = 1\right)$ , as the percentual frequency of each range.

5. Following Axiom 3, for each out-of-time predictions' vector ($\hat{Y}_m$), we estimate the conditional expected value given loans will be defaulted, $\text{E}\left[\hat{Y}_m|Y = 1\right]$.

6. Following Corollary 1.5, we compare models in pairs using the results in step 5 to predict which model (A or B, full or segmented models) will have, in an out-of-time dataset, a higher correlation between default observations and their respective predictions.

## 5 Data and Results

For our analyses, we used the output generated by the estimations in Yoshida et al. (2025). Unlike that study, our focus is not on the coefficients of variables and the evaluation of the risk

[14]Score ranges were automatically created using R's package "monobin". Manual available at https://cran.r-project.org/web/packages/monobin/monobin.pdf.

model metric. Instead, we depart from the predicted output for each loan using the full and segmented models ($i.e.$, models estimated using loans from all banks versus models estimated using loans from specific banks) and the model risk metric of each model (the *CSMR* measure, described in Equation 2) to estimate 12-month ahead out-of-time default predictions.

### 5.1 Primary Data Sources

We use the same loan-level data from Yoshida et al. (2025), who, in turn, utilize loans to microenterprises and small businesses (MSBs) in Brazil. The primary data source for these estimations is the credit bureau of the Central Bank of Brazil (SCR), which contains detailed monthly loan-level information on loans made to MSBs by banks in Brazil from January 2013 to December 2019. SCR data were matched to other data sources, namely: the database of restructured loans; the database from the Brazilian Revenue Service, containing borrowers industry classification, firms size, share capital, and ZIP code; employment data from RAIS – Annual Social Information Report; GeoSampa geospatial database; and the database from the 2010 Census produced by the IBGE – Brazilian Institute of Geography and Statistics.[15] Similar to that study, default estimations were made for March and September of each year. Because we do not have data to estimate out-of-time models for March and September 2018, our out-of-time estimations span March 2013 to September 2017.

### 5.2 Results

In Yoshida et al. (2025), full and segmented data models were estimated for each bank in every March and September from 2014 to 2017. This approach generated a balanced set of estimations across institutions and time. Table 1 summarizes the in-sample results for the full model, Equation 3, and the segmented models, Equation 4.

Table 1 also summarizes the out-of-sample (evaluated at the same point in time as the in-sample analysis, but using a holdout sample not employed in model estimation) and out-of-time (based on a sample collected 12 months later) results for both full and segmented models. The higher correlation of segmented models observed in the in-sample estimations does not persist in out-of-time applications. As illustrated in Figure 1, segmented models outperform full models in 46 out of 48 in-sample comparisons.[16] However, as shown in Figure 2, the segmented models outperform the original models in 42 of the 48 out-of-sample datasets.

Out-of-sample correlation comparisons (reported in Table 1) were used to detect overfitting. Notably, Bank C's segmented models in September 2014 and March 2015 exhibit a correlation drop of more than 40%. Figure 2 illustrates the out-of-sample *CSMR* ($1 - \rho_{Y,\hat{Y}}$) for each bank, highlighting a pronounced correlation shift in these months and providing clear evidence of overfitting.

[15] See Yoshida et al. (2025) for details on databases, prediction methods, and empirical applications.

[16] According to the correlation coefficient test proposed by Zou (2007), the remaining two comparisons (Bank F, March and September 2017) are not statistically different at the 99% confidence level.

**Table 1: Summary of in-sample results of full and segmented models**

| | Correlation(Y,Ŷ) | Mar-14 | | Sep-14 | | Mar-15 | | Sep-15 | | Mar-16 | | Sep-16 | | Mar-17 | | Sep-17 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* |
| **Bank A** | *In-sample* | 0.5306 | 0.5883 | 0.5828 | 0.5962 | 0.6606 | 0.6634 | 0.6416 | 0.6434 | 0.6441 | 0.6476 | 0.6556 | 0.6617 | 0.6760 | 0.6831 | 0.6572 | 0.6613 |
| | *Out-of-sample* | 0.5385 | 0.5913 | 0.5855 | 0.5967 | 0.6614 | 0.6620 | 0.6339 | 0.6369 | 0.6584 | 0.6588 | 0.6485 | 0.6544 | 0.6882 | 0.6911 | 0.6532 | 0.6563 |
| | *Out-of-time* | **0.6442** | 0.5729 | **0.6102** | 0.5994 | **0.6433** | 0.6327 | 0.6211 | **0.6228** | **0.6773** | 0.6743 | 0.6373 | **0.6390** | **0.5830** | 0.5764 | 0.6687 | **0.6740** |
| | *PSI* | 0.0007 | 0.0019 | 0.0097 | 0.0070 | 0.0084 | 0.0104 | 0.0001 | 0.0001 | 0.0015 | 0.0031 | 0.0249 | 0.0267 | 0.0139 | 0.0330 | 0.0020 | 0.0010 |
| **Bank B** | *In-sample* | 0.6349 | 0.6580 | 0.6023 | 0.6331 | 0.5720 | 0.5871 | 0.6081 | 0.6165 | 0.5938 | 0.6019 | 0.6047 | 0.6136 | 0.6197 | 0.6269 | 0.6002 | 0.6178 |
| | *Out-of-sample* | 0.6330 | 0.6531 | 0.5886 | 0.6155 | 0.5816 | 0.5913 | 0.5948 | 0.5978 | 0.5957 | 0.6037 | 0.6074 | 0.6166 | 0.6118 | 0.6157 | 0.6030 | 0.6146 |
| | *Out-of-time* | 0.5687 | **0.5822** | 0.5710 | **0.5920** | 0.5927 | **0.6006** | **0.6084** | 0.6028 | 0.5996 | **0.6100** | 0.5937 | **0.6046** | 0.5769 | **0.5845** | 0.5756 | **0.5854** |
| | *PSI* | 0.0068 | 0.0071 | 0.0402 | 0.0398 | 0.0507 | 0.0595 | 0.0214 | 0.0674 | 0.0404 | 0.0238 | 0.0171 | 0.0201 | 0.0141 | 0.0151 | 0.0208 | 0.0288 |
| **Bank C** | *In-sample* | 0.5358 | 0.5692 | 0.5080 | 0.5588 | 0.5526 | 0.5725 | 0.5378 | 0.5546 | 0.6011 | 0.6226 | 0.6100 | 0.6361 | 0.6585 | 0.6716 | 0.5604 | 0.5755 |
| | *Out-of-sample* | 0.5316 | 0.5602 | 0.5273 | 0.3934 | 0.5527 | 0.3426 | 0.5269 | 0.5410 | 0.5994 | 0.6176 | 0.6157 | 0.6368 | 0.6336 | 0.6419 | 0.5342 | 0.5449 |
| | *Out-of-time* | 0.5435 | **0.5512** | **0.5145** | 0.3524 | **0.5930** | 0.3118 | **0.6010** | 0.5945 | **0.6162** | 0.6071 | **0.5260** | 0.5154 | **0.5381** | 0.5360 | **0.5406** | 0.5223 |
| | *PSI* | 0.0532 | 0.0677 | 0.0364 | 0.7112 | 0.0072 | 1.2614 | 0.0054 | 0.0016 | 0.0012 | 0.0017 | 0.0823 | 0.1059 | 0.1544 | 0.1829 | 0.0503 | 0.0426 |
| **Bank D** | *In-sample* | 0.5894 | 0.6573 | 0.5269 | 0.5714 | 0.5565 | 0.5978 | 0.5785 | 0.5905 | 0.5978 | 0.6179 | 0.5763 | 0.5999 | 0.6172 | 0.6256 | 0.5972 | 0.6202 |
| | *Out-of-sample* | 0.5586 | 0.6249 | 0.5107 | 0.5498 | 0.5475 | 0.5872 | 0.5521 | 0.5649 | 0.5987 | 0.6106 | 0.5957 | 0.5882 | 0.6136 | 0.6148 | 0.5975 | 0.6094 |
| | *Out-of-time* | 0.5383 | **0.5639** | **0.5379** | 0.5311 | **0.5967** | 0.5608 | **0.5968** | 0.5905 | **0.6100** | 0.4348 | **0.5966** | 0.5220 | 0.5706 | **0.5733** | **0.5552** | 0.5550 |
| | *PSI* | 0.0499 | 0.1032 | 0.0044 | 0.0104 | 0.0277 | 0.0571 | 0.0367 | 0.0670 | 0.0089 | 0.1335 | 0.0099 | 0.0069 | 0.0461 | 0.0341 | 0.0414 | 0.0430 |
| **Bank E** | *In-sample* | 0.5575 | 0.6330 | 0.5474 | 0.6097 | 0.5958 | 0.6489 | 0.6009 | 0.6509 | 0.6108 | 0.6665 | 0.6139 | 0.6738 | 0.6353 | 0.6821 | 0.6008 | 0.6652 |
| | *Out-of-sample* | 0.5455 | 0.6111 | 0.5494 | 0.6082 | 0.5949 | 0.6464 | 0.5855 | 0.6376 | 0.6055 | 0.6547 | 0.6286 | 0.6781 | 0.6333 | 0.6632 | 0.6069 | 0.6621 |
| | *Out-of-time* | 0.5741 | **0.6196** | 0.5600 | **0.6423** | 0.5933 | **0.6582** | 0.6140 | **0.6634** | 0.6088 | **0.6134** | 0.5924 | **0.6486** | 0.5970 | **0.6525** | 0.5985 | **0.6516** |
| | *PSI* | 0.0175 | 0.0238 | 0.0060 | 0.0073 | 0.0088 | 0.0037 | 0.0019 | 0.0041 | 0.0215 | 3.4337 | 0.0431 | 0.0449 | 0.0160 | 0.0311 | 0.0017 | 0.0196 |
| **Bank F** | *In-sample* | 0.6201 | 0.6626 | 0.6007 | 0.6172 | 0.5667 | 0.5972 | 0.5834 | 0.6018 | 0.6299 | 0.6604 | 0.5650 | 0.6018 | 0.5858 | 0.5797 | 0.5010 | 0.4998 |
| | *Out-of-sample* | 0.5881 | 0.5664 | 0.5861 | 0.5995 | 0.5811 | 0.6090 | 0.5383 | 0.5416 | 0.6049 | 0.6269 | 0.5608 | 0.5855 | 0.5697 | 0.5635 | 0.5392 | 0.5363 |
| | *Out-of-time* | **0.5509** | 0.4921 | 0.5290 | **0.5394** | **0.6078** | 0.5907 | 0.5715 | **0.5731** | 0.5513 | **0.5552** | 0.5101 | **0.5307** | **0.4915** | 0.4709 | 0.5411 | **0.5455** |
| | *PSI* | 0.0079 | 0.0016 | 0.0005 | 0.0052 | 0.0300 | 0.0068 | 0.0085 | 0.0063 | 0.0462 | 0.1487 | 0.0093 | 0.0261 | 0.0063 | 0.0046 | 0.0055 | 0.0175 |

Notes: In-sample, Out-of-sample, and Out-of-time (12 months later) correlations between observations (the dependent variable of a credit scoring model) and its predictions (Correlation$(Y, \hat{Y})$). **Black bold** numbers in Out-of-time correlations show the higher number in each cross-section comparison. **Red** (or **green**) numbers sign reversal (or same) order with out-of-time comparison, in In-sample and Out-of-sample datasets. The PSI is a divergence measure, as shown in Equation 1. **Red** numbers in PSI sign PSI greater than 0.25, an indicator of significant changes in the population; and **orange** numbers sign PSI greater than 0.1, an indicator of possible population change (L. Thomas, 2009).

**Out-of-sample CSRM by bank**

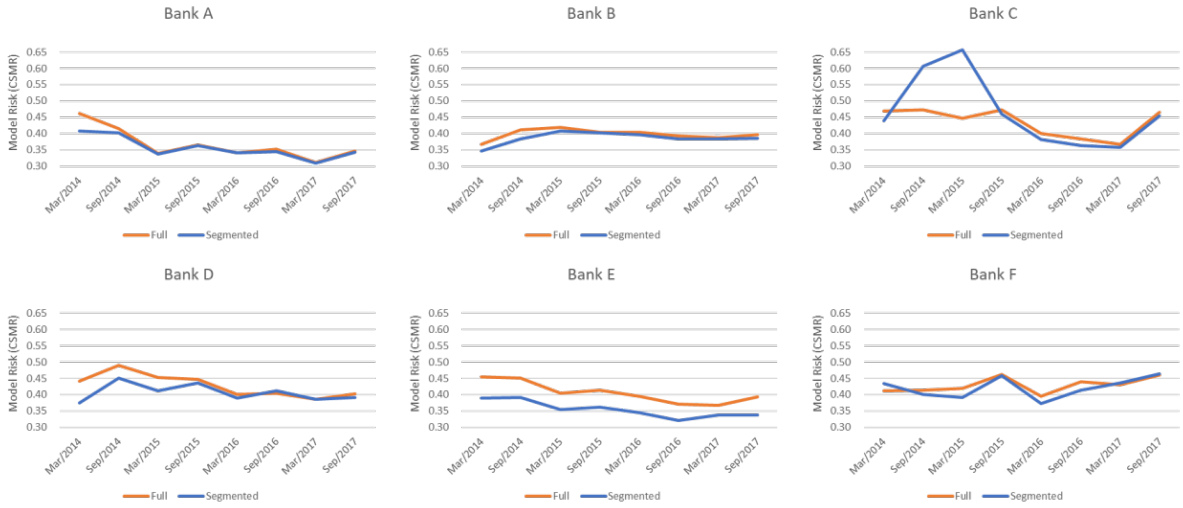


**Figure 2: Out-of-sample Credit Scoring Model Risk (*CSMR*) for each bank**

Note: In each graph, the orange line represents the conditional *CSMR*s of full data models, while the blue line represents those of segmented data models.

**Out-of-time CSRM by bank**

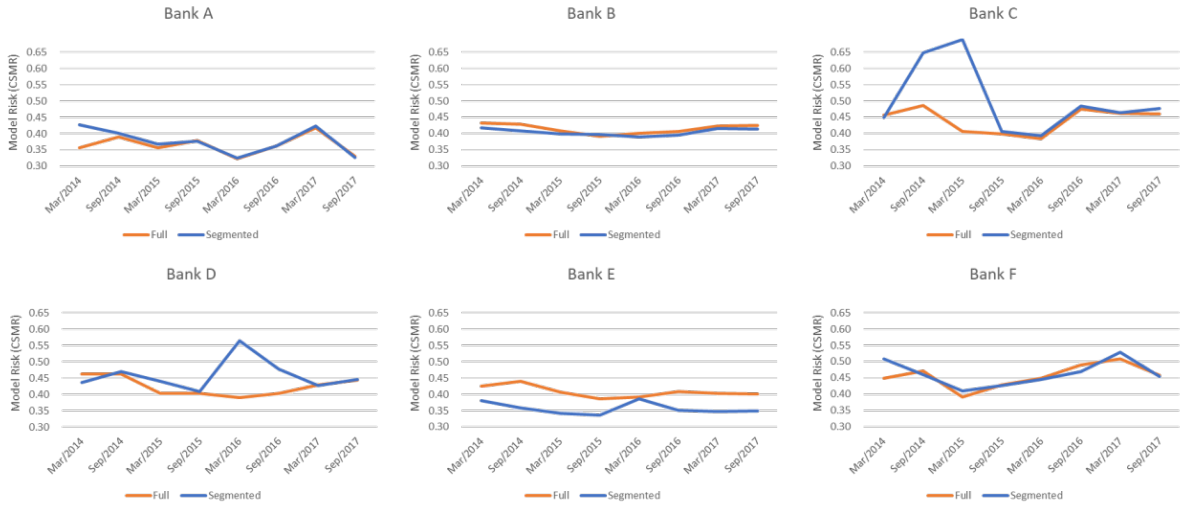


**Figure 3: Out-of-time (12 months later) Credit Scoring Model Risk *CSMR* by bank**

Note: In each graph, the orange line represents the conditional *CSMR*s of full data models, while the blue line represents those of segmented data models. While in in-sample datasets, the *CSMR*s of segmented data models (by bank) exhibit lower model risk than full data models, this pattern is not observed here.

Figure 3 compares the *CSMR* for the full data and segmented data models in out-of-time applications for the six banks. In only two of the six banks (banks B and E), the segmented data models generally have lower *CSMR* than the full data models and consistently exhibit the same behavior as observed in the in-sample measures. For these banks, the inferences are consistent with the in-sample results of Yoshida et al., 2025; *i.e.*, segmented models present a lower *CSMR*. However, for banks C and D, *CSMR* is generally lower for full data models (with a few exceptions), which is at odds with the in-sample results of Yoshida et al., 2025. Finally,

for banks A and F, we do not observe any general dominance between the full and segmented data models.

The takeaway from the examination of Figure 3 is that, unlike for in-sample applications, one cannot observe a consistent stochastic approach dominance between segmented and full data models for out-of-time applications. The practical implication is that we must carefully measure and monitor model risk in out-of-time data.

Using the in-sample results (lower *CSMR*) from Yoshida et al. (2025) would lead one to prefer the segmented models. However, Figure 3 shows that this choice results in a lower out-of-time *CSMR* in only 26 of the 48 cases (54.2%).[17] The remaining 22 cases in which the segmented models exhibited a higher *CSMR* were unevenly distributed across banks: five in Bank A, one in Bank B, seven in Bank C, six in Bank D, none in Bank E, and three in Bank F. This heterogeneity indicates that out-of-time shrinkage (or decalibration) affects some banks more strongly than it does others.

In a visual inspection of Figure 3, it is noticeable that the difference between the full data *CSMR* (orange line) and segmented data *CSMR* (blue line) is economically small, suggesting that the model risk in both the full and segmented models is similar. However, in some model estimations (particularly for banks C and D), the difference between the *CSMR*s is practically large. The results for the segmented models of Bank C for September 2014 and March 2015, and of Bank D for March 2016, suggest considerable overfitting in these models. It is particularly important to predict which model estimations could present overfitting behavior, which leads to a large model risk in out-of-time applications.

Along with the *CSMR* measure, Table 1 also presents the Population Stability Index (PSI) (or the *Divergence* measure) for each estimated model. The PSI can flag potential overfitting, which is problematic for datasets that are out of time. Three models have PSI higher than 0.25, indicating significant changes in the population distribution (the segmented models of Bank C in September 2014 and March 2015, and the segmented model of Bank E in March 2016). Indeed, we flagged potential overfitting in two of these cases in the inspection shown in Figure 3.

The elevated PSI values observed in the segmented models of Bank C (September 2014 and March 2015) are primarily attributed to significant increases in the dispersion of out-of-time predictions compared to the in-sample results. This suggests that the segmented specification intensified the variability of the predicted probabilities when applied to a shifted population. In contrast, the heightened PSI observed in the segmented model of Bank E (March 2016) indicates a change in the mean level of predictions rather than in their dispersion.

If models with PSI above 0.25 are considered ineligible,[18] The lower *CSMR* model was selected in 27 of 48 cases (56.3%), only one more than when PSI was ignored. This criterion, however, identifies more overfitted models, notably the segmented models of Bank C in September 2014 and March 2015.[19]

[17] According to the Zou (2007) correlation test, full and segmented models do not differ statistically at the 99% level in 4 of the 56 comparisons (Bank C in March 2017, Bank D in September 2017, and Bank F in September 2015 and March 2016). Excluding these cases, the lower *CSMR* model would have been selected in 24 of 44 comparisons.

[18] When one model is ineligible, the eligible model is taken as the only option.

[19] If models with PSI above 0.25 were excluded, the higher out-of-time correlation model for Bank E in March 2016 would also be discarded. Although the segmented model in this case exhibits a high PSI, its lower *CSMR* (i.e., higher

Six models present PSI above 0.1, indicating potential changes in the population distribution: Bank C's segmented models (September 2016 and March 2017) and full model (March 2017), Bank D's segmented models (March 2014 and March 2016), and Bank F's segmented model (March 2016).

If we consider models with PSI higher than 0.10 as ineligible for decision-making, then we would have chosen the lower *CSMR* models 28 out of 48 times, or in 58.3% of the correlations' comparisons. This is one more time when we do not consider models with PSI higher than 0.25. However, when we chose the incorrect model (or the model with a lower correlation), the difference between the models' correlations was 0.0198 on average rather than 0.0264. In March 2017, both (full and segmented models) of Bank C presented a PSI value higher than 0.10. We considered the full model in decision-making because it had the lowest PSI.

**Table 2: Overview of the decision-making criteria for select models**

| | correct | | | | | | | | Mean | Max |
|---|---|---|---|---|---|---|---|---|---|---|
| | bank A | bank B | bank C | bank D | bank E | bank F | Total | (%) | (incorrect) | (incorrect) |
| ***Full*** | 5 | 1 | 7 | 6 | 0 | 3 | 22/48 | 45.8% | 0.0222 | 0.0823 |
| ***Segmented*** | 3 | 7 | 1 | 2 | 8 | 5 | 26/48 | 54.2% | 0.0451 | 0.2812 |
| ***Segmented - PSI(0.25)*** | 3 | 7 | 3 | 2 | 7 | 5 | 25/45 | 55.6% | 0.0275 | 0.1752 |
| ***Segmented - PSI(0.10)*** | 3 | 7 | 5 | 2 | 7 | 4 | 23/40 | **57.5%** | **0.0213** | **0.0746** |
| ***In-sample*** | 3 | 7 | 1 | 2 | 8 | 5 | 26/48 | 54.2% | 0.0444 | 0.2812 |
| ***In-sample - PSI(0.25)*** | 3 | 7 | 3 | 2 | 7 | 5 | 25/45 | 55.6% | 0.0267 | 0.1752 |
| ***In-sample - PSI(0.10)*** | 3 | 7 | 5 | 2 | 7 | 4 | 23/40 | **57.5%** | **0.0203** | **0.0746** |
| ***Out-of-sample*** | 3 | 7 | 3 | 3 | 8 | 6 | 30/48 | 62.5% | 0.0222 | 0.1752 |
| ***Out-of-sample - PSI(0.25)*** | 3 | 7 | 3 | 3 | 7 | 6 | 27/45 | 60.0% | 0.0222 | 0.1752 |
| ***Out-of-sample - PSI(0.10)*** | 3 | 7 | 5 | 3 | 7 | 5 | 25/40 | 62.5% | 0.0142 | 0.0713 |

Note: Decision-making criteria. "Full" and "Segmented" refer to the selection of full-data and segmented models. "In-sample" and "Out-of-sample" criteria select the higher correlation models, respectively, as indicated by the in-sample and out-of-sample datasets. The suffix "- PSI(0.25)" indicates the respective criteria excluding models with PSI higher than 0.25, while the suffix "- PSI(0.10)" indicates the respective criteria excluding models with PSI values higher than 0.10. The first six columns present the sum of correct higher-correlation choices by banks, "Total" and "(%)" present correct choices in number and percentage; and "Mean (incorrect)" and "Max(incorrect)" are calculated as the average and maximum of the absolute value of the difference between correlations of the incorrect choices. The red (and green) numbers indicate lower (and higher) measures than those from the criterion without the PSI filter (within each criterion). **Bold green** numbers sign the lowest measure within each criterion.

If out-of-sample *CSMR*s values had been used instead of in-sample *CSMR*s, the lower *CSMR* ($i.e.$, higher correlation) model would have been selected in 30 out of 48 comparisons (62.5%). The 18 incorrect choices were unevenly distributed across banks: five for Bank A, one for Bank B, five for Bank C, five for Bank D, none for Bank E, and two for Bank F.

Table 2 summarizes the decision criteria used to select full or segmented models based on their out-of-time performance. The "Full" and "Segmented" criteria correspond to choosing full data or segmented models, respectively, while "In-sample" and "Out-of-sample" select the model with higher correlation according to in-sample or out-of-sample datasets. Each criterion is also combined with PSI thresholds: criteria labeled "–PSI(0.25)" exclude models with PSI above 0.25, and those labeled "–PSI(0.10)" exclude models with PSI above 0.10.

correlation than the full data model) suggests superior risk ordering relative to the full model. Since the primary objective of credit scoring models is to rank credit risk rather than to estimate individual default probabilities, this model may still be preferable.

In the in-sample data, segmented models outperform full-data models in 46 of 48 comparisons, but this advantage does not persist out of time. Out-of-sample correlation comparisons combined with PSI help filter the models affected by significant population shifts. Among the criteria, "Out-of-sample – PSI(0.10)" performs best overall, yielding 30 correct selections out of 48 (62.5%) and the lowest mean (0.0137) and maximum (0.0713) absolute correlation differences among incorrect choices.

# 6 Results of Approaches to compare models in out-of-time basis

## 6.1 First Approach – Similar shrinkage

Based on Assumption 1, models in production should ideally remain calibrated on out-of-time datasets or, at least, maintain the same level of shrinkage as the out-of-sample datasets. However, ignoring changes in the credit portfolios is unreasonable. Over time, portfolios may show varying default behaviors, densities, and prediction standard deviations ($\sigma_{\hat{Y}}$), leading to different correlations between default observations and predicted scores ($\rho_{Y,\hat{Y}}$), and consequently, different Model Risks, or *CSMR*.

In the first approach (Table 3), we assume that out-of-time datasets have the same calibration or the same shrinkage (same $\beta$ of the out-of-sample estimations, as assumed in Assumption 1) as out-of-sample datasets. We use Corollaries 1.2 and 1.3 (see section 4.1) to make inferences on the potential correlation of models in comparison.

Using Approach 1 (Similar Shrinkage) as the decision rule would lead to selecting the model with lower model risk (*CSMR*), or higher out-of-time correlation, in 26 out of 48 comparisons (54.2%), the same sucess rate obtained by the "In-sample" criterion. However, this performance is weaker than other benchmarks: the "Out-of-sample" criterion correctly selects the higher-correlation model in 62.5%, while the "Segmented–PSI(0.25)" and "Segmented–PSI(0.10)" criteria acheive 56.3% and 58.3% respectively.[20]

Importantly, this aggregated behaviour masks substantial heterogeneity across banks. Errors are unevenly distributed: they occurred three times (out of eight) for Bank A, two times for Bank B, four times for Bank C, seven times for Bank D, once for Bank E, and five times for Bank F. This pattern indicates that the Similar Shrinkage approach performs inconsistently across institutions, likely reflecting differences in portfolio composition, temporal dynamics, and sensitivity to overfitting.

[20]Incorrect decisions or inconsistent inferences about out-of-time correlations using Approach 1 arise from the unrealistic Assumption 1. The regression coefficient $\beta_m$ in $Y = \alpha_m + \beta_m \hat{Y}_m$ and the implied shrinkage $(1-\beta_m)$ are not stable and do not coincide between out-of-time and out-of-sample datasets. If, instead of assuming constant $\beta_m$, we had assumed constant covariance between default and predicted default—$\text{Cov}\left(Y, \hat{Y}\right)$—across out-of-sample and out-of-time datasets, the approach would correctly identify the model with higher out-of-time correlation in 32 out of 48 (two thirds). However, this alternative assumption is even less plausible than Assumption 1.

**Table 3: Results of first approach**

| | | Mar-14 | | Sep-14 | | Mar-15 | | Sep-15 | | Mar-16 | | Sep-16 | | Mar-17 | | Sep-17 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* |
| **Bank A** | *A1: Similar shrinkage* | 0.0811 | 0.0918 | 0.0913 | 0.0946 | 0.1076 | 0.1069 | 0.1052 | 0.1055 | 0.1064 | 0.1056 | 0.0926 | 0.0930 | 0.0954 | 0.0916 | 0.0965 | 0.0963 |
| | *Correlation(Y, Ŷ)* | **0.6442** | 0.5729 | **0.6102** | 0.5994 | **0.6433** | 0.6327 | 0.6211 | **0.6228** | **0.6773** | 0.6743 | 0.6373 | **0.6390** | **0.5830** | 0.5764 | 0.6687 | **0.6740** |
| | *Model Risk* | **0.3558** | 0.4271 | **0.3898** | 0.4006 | **0.3567** | 0.3673 | 0.3789 | **0.3772** | **0.3227** | 0.3257 | 0.3627 | **0.3610** | **0.4170** | 0.4236 | 0.3313 | **0.3260** |
| **Bank B** | *A1: Similar shrinkage* | 0.0793 | 0.0812 | 0.0856 | 0.0899 | 0.0932 | 0.0950 | 0.1030 | 0.1061 | 0.1050 | 0.1031 | 0.1017 | 0.1044 | 0.1078 | 0.1093 | 0.1111 | 0.1154 |
| | *Correlation(Y, Ŷ)* | 0.5687 | **0.5822** | 0.5710 | **0.5920** | 0.5927 | **0.6006** | **0.6084** | 0.6028 | 0.5996 | **0.6100** | 0.5937 | **0.6046** | 0.5769 | **0.5845** | 0.5756 | **0.5854** |
| | *Model Risk* | 0.4313 | **0.4178** | 0.4290 | **0.4080** | 0.4073 | **0.3994** | **0.3916** | 0.3972 | 0.4004 | **0.3900** | 0.4063 | **0.3954** | 0.4231 | **0.4155** | 0.4244 | **0.4146** |
| **Bank C** | *A1: Similar shrinkage* | 0.1716 | 0.1826 | 0.1647 | 0.1297 | 0.1745 | 0.1253 | 0.1650 | 0.1666 | 0.1898 | 0.1997 | 0.1615 | 0.1641 | 0.1266 | 0.1252 | 0.0967 | 0.0987 |
| | *Correlation(Y, Ŷ)* | 0.5435 | **0.5512** | **0.5145** | 0.3524 | **0.5930** | 0.3118 | **0.6010** | 0.5945 | **0.6162** | 0.6071 | **0.5260** | 0.5154 | **0.5381** | 0.5360 | **0.5406** | 0.5223 |
| | *Model Risk* | 0.4565 | **0.4488** | **0.4855** | 0.6476 | **0.4070** | 0.6882 | **0.3990** | 0.4055 | **0.3838** | 0.3929 | **0.4740** | 0.4846 | **0.4619** | 0.4640 | **0.4594** | 0.4777 |
| **Bank D** | *A1: Similar shrinkage* | 0.1501 | 0.1697 | 0.1581 | 0.1717 | 0.1880 | 0.2081 | 0.1843 | 0.1966 | 0.1752 | 0.2372 | 0.1611 | 0.1660 | 0.1800 | 0.1762 | 0.1694 | 0.1713 |
| | *Correlation(Y, Ŷ)* | 0.5383 | **0.5639** | **0.5379** | 0.5311 | **0.5967** | 0.5608 | **0.5968** | 0.5905 | **0.6100** | 0.4348 | **0.5966** | 0.5220 | 0.5706 | **0.5733** | **0.5552** | 0.5550 |
| | *Model Risk* | 0.4617 | **0.4361** | **0.4621** | 0.4689 | **0.4033** | 0.4392 | **0.4032** | 0.4095 | **0.3900** | 0.5652 | **0.4034** | 0.4780 | 0.4294 | **0.4267** | **0.4448** | 0.4450 |
| **Bank E** | *A1: Similar shrinkage* | 0.1570 | 0.1648 | 0.1494 | 0.1680 | 0.1707 | 0.1866 | 0.1587 | 0.1770 | 0.1566 | 0.1445 | 0.1480 | 0.1582 | 0.1394 | 0.1417 | 0.1363 | 0.1390 |
| | *Correlation(Y, Ŷ)* | 0.5741 | **0.6196** | 0.5600 | **0.6423** | 0.5933 | **0.6582** | 0.6140 | **0.6634** | 0.6088 | **0.6134** | 0.5924 | **0.6486** | 0.5970 | **0.6525** | 0.5985 | **0.6516** |
| | *Model Risk* | 0.4259 | **0.3804** | 0.4400 | **0.3577** | 0.4067 | **0.3418** | 0.3860 | **0.3366** | 0.3912 | **0.3866** | 0.4076 | **0.3514** | 0.4030 | **0.3475** | 0.4015 | **0.3484** |
| **Bank F** | *A1: Similar shrinkage* | 0.1109 | 0.1045 | 0.1124 | 0.1101 | 0.1341 | 0.1329 | 0.1174 | 0.1121 | 0.1138 | 0.1038 | 0.1042 | 0.1033 | 0.1046 | 0.1015 | 0.1008 | 0.0984 |
| | *Correlation(Y, Ŷ)* | **0.5509** | 0.4921 | 0.5290 | **0.5394** | **0.6078** | 0.5907 | 0.5715 | **0.5731** | 0.5513 | **0.5552** | 0.5101 | **0.5307** | **0.4915** | 0.4709 | 0.5411 | **0.5455** |
| | *Model Risk* | **0.4491** | 0.5079 | 0.4710 | **0.4606** | **0.3922** | 0.4093 | 0.4285 | **0.4269** | 0.4487 | **0.4448** | 0.4899 | **0.4693** | **0.5085** | 0.5291 | 0.4589 | **0.4545** |

Notes: First approach. "A1: Similar shrinkage" follows Corollary 1.2 and is calculated for each model m by $\beta_m \sigma_{\hat{Y}_m}$, where $\beta_m$ is the coefficient ($\beta$) of regression of default on model ms predictions ($Y = \alpha + \beta \hat{Y}_m$) and $\sigma_{\hat{Y}_m}$ is the standard deviation of predictions in model m. Correlation$(Y, \hat{Y})$ is out-of-time (12 months later) correlation ($\rho_{Y,\hat{Y}}^{t_{1,m,b}}$) between observations (the dependent variable of a credit scoring model) and its predictions. Model Risk is Credit Scoring Model Risk (CSMR) equal to one minus the correlation. Full refers to the full data model and Segmen., to the segmented data models. **Black bold** numbers in Correlation$(Y, \hat{Y})$ show the higher number in each cross-section comparison. **Black bold** numbers in Model Risk (or the CSMR) show the lower number in each cross-section comparison. **Red** (or **green**) numbers in "A1: Similar shrinkage" sign reversal (or same) order with Correlation$(Y, \hat{Y})$.

### 6.2 Second Approach – Monte Carlo simulation

The second approach is independent of Assumption 1 and instead relies on a Monte Carlo simulation, and is based on the average of the full and segmented models' predictions for each bank. We conducted 1,000 rounds of simulations to compare the correlations between the two models.[21]

Figures 4 to 11 plot the simulated correlations of the full and segmented models. Each figure consists of six graphs, with each graph representing the set of correlations between the simulated default events and the predicted scores for a specific month. The horizontal axis shows the correlation for the full data model ($\rho_{\text{Simulated Y, Full } \hat{Y}}$), while the vertical axis shows the correlation for the segmented data model ($\rho_{\text{Simulated Y, Segmented } \hat{Y}}$). When points are concentrated in the upper left (bottom right) of the 45-degree diagonal, the segmented (full) data model is preferable because it presents the highest correlation among them. The color[22] of the plot indicates the simulated frequency of default, named here as "FD". In a visual inspection, there was no relationship between FD and the difference in simulated correlations (or the plot position in the graph).

Using Approach 2 for decision-making, a higher out-of-time correlation model was chosen in 34 of the 48 model comparisons (70.8%).[23] Approach 2, on a general basis, is therefore better than the in-sample or out-of-sample criterion of choice for our set of estimations. It is also better than Approach 1. The wrong choices (14 comparisons, or 29.8%) are due to the assumption of the average of the models' predictions as the probability of default function.

In Table 4, we present the median of the simulated correlations' difference (median of $\rho_{\text{Simulated Y, Full } \hat{Y}} - \rho_{\text{Simulated Y, Segmented } \hat{Y}}$) for each bank for March and September from 2014 to 2017.

The incorrect inferences are not homogeneous across banks and months. They occurred three times (out of eight comparisons) in Banks A, D, and E; once in Bank B; and twice in Banks C and F. When the Monte Carlo simulation approach incorrectly indicates the higher correlation model, the median of the correlation difference is lower than 0.0070, that is, the choice criterion narrowly misses.

### 6.3 Third Approach – Bayesian estimation of covariances

Following Axiom 2, we estimated the expected value of predictions, given that loans would be defaulted, for both full and segmented models ($\text{E}\left[\hat{Y}_{\text{Full}}|Y=1\right]$) and ($\text{E}\left[\hat{Y}_{\text{Segmented}}|Y=1\right]$). Then, following Corollary 1.5, we compare the models to assess which model would have the greater correlation ($\rho_{Y,\hat{Y}}$).

[21] The number of rounds was limited by computational constraints. However, in less extensive exercises, reducing or increasing the number of rounds does not materially affect the overall position of the plots, their general visual shape, or the median difference between the correlations of the full and segmented models, and therefore does not alter decision-making under this approach.

[22] A gradient where blue indicates the lowest and red the highest simulated default rates.

[23] By deviating from the approach (as described in Section 4.2), where we use the average of models' predictions as the probability of default function, and instead considering each models' predictions itself for each simulation round, we observe three comparisons that preserve stochastic dominance (in Bank C's September 2014 and March 2015 and Bank D's March 2016, *i.e.*, full data models present higher correlation even when the segmented model's predictions are used as the probability default function). Graphs not reported.

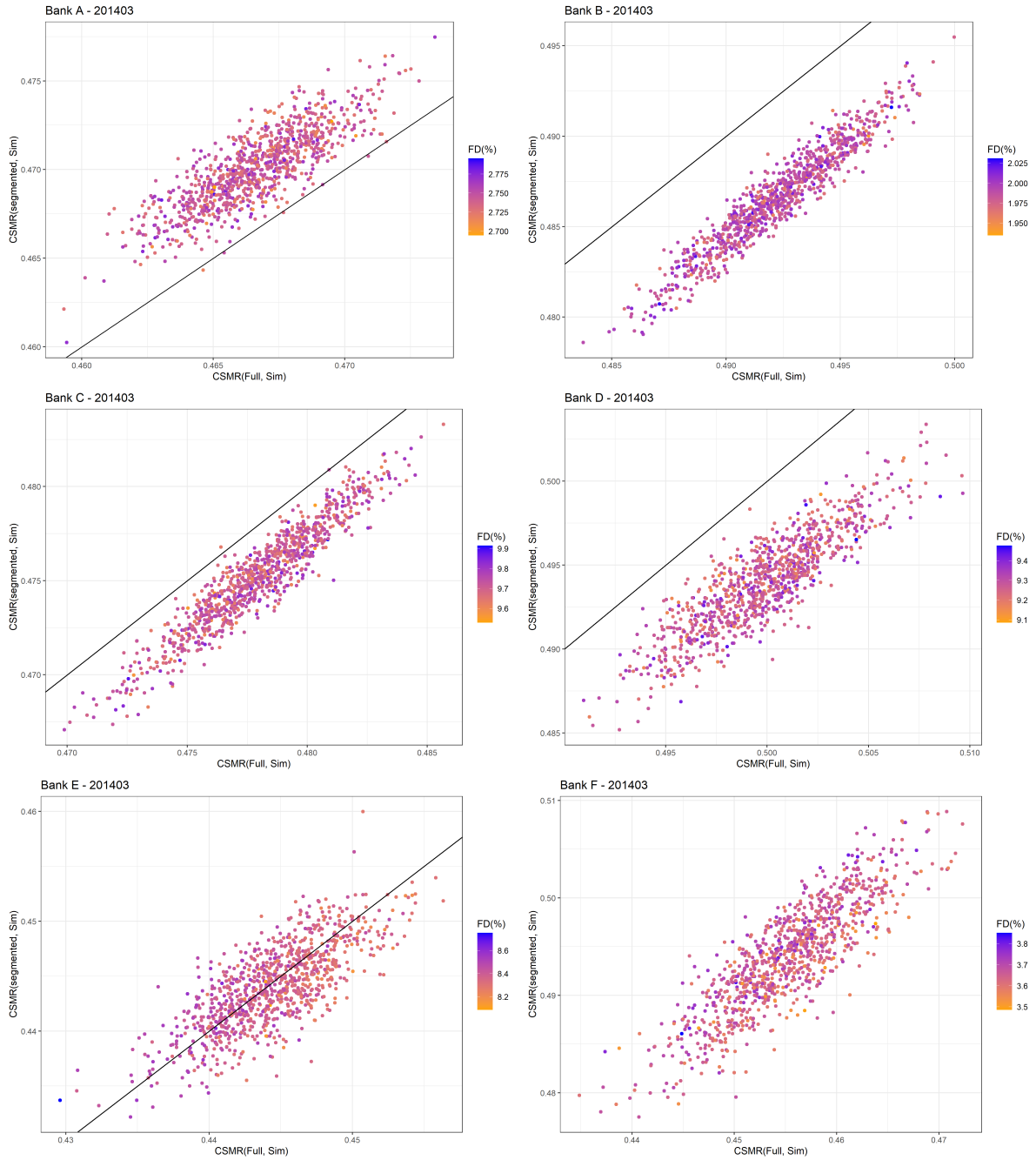


**Figure 4: Simulated correlations for each one of the banks in March 2014.**

Note: The X-axis is the full data model simulated correlation, and the Y-axis is the segmented model simulated frequency of default, named here as "FD."

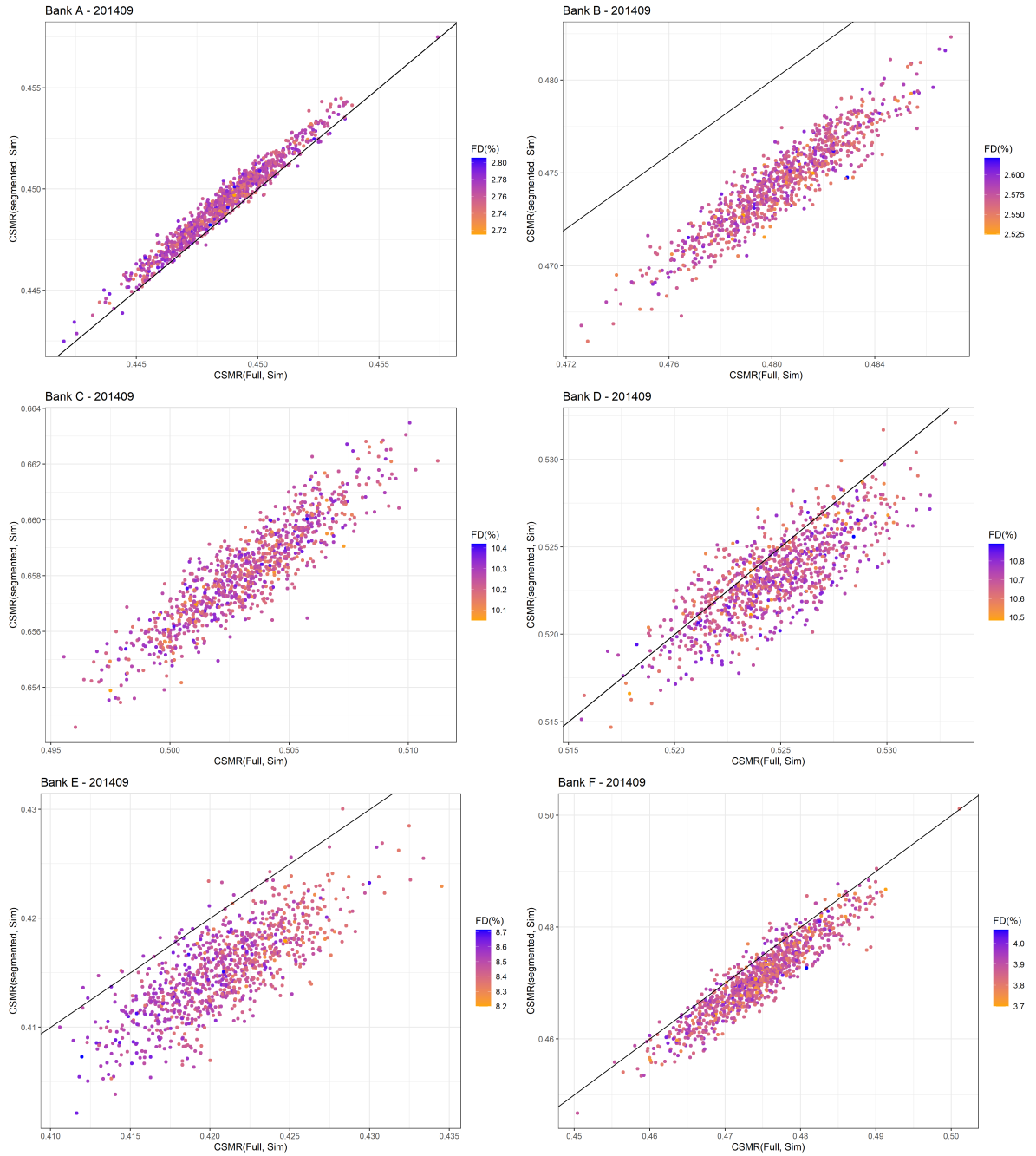


**Figure 5: Simulated correlations for each one of the banks in September 2014.**

Note: The X-axis is the full data model simulated correlation, and the Y-axis is the segmented model simulated frequency of default, named here as "FD".

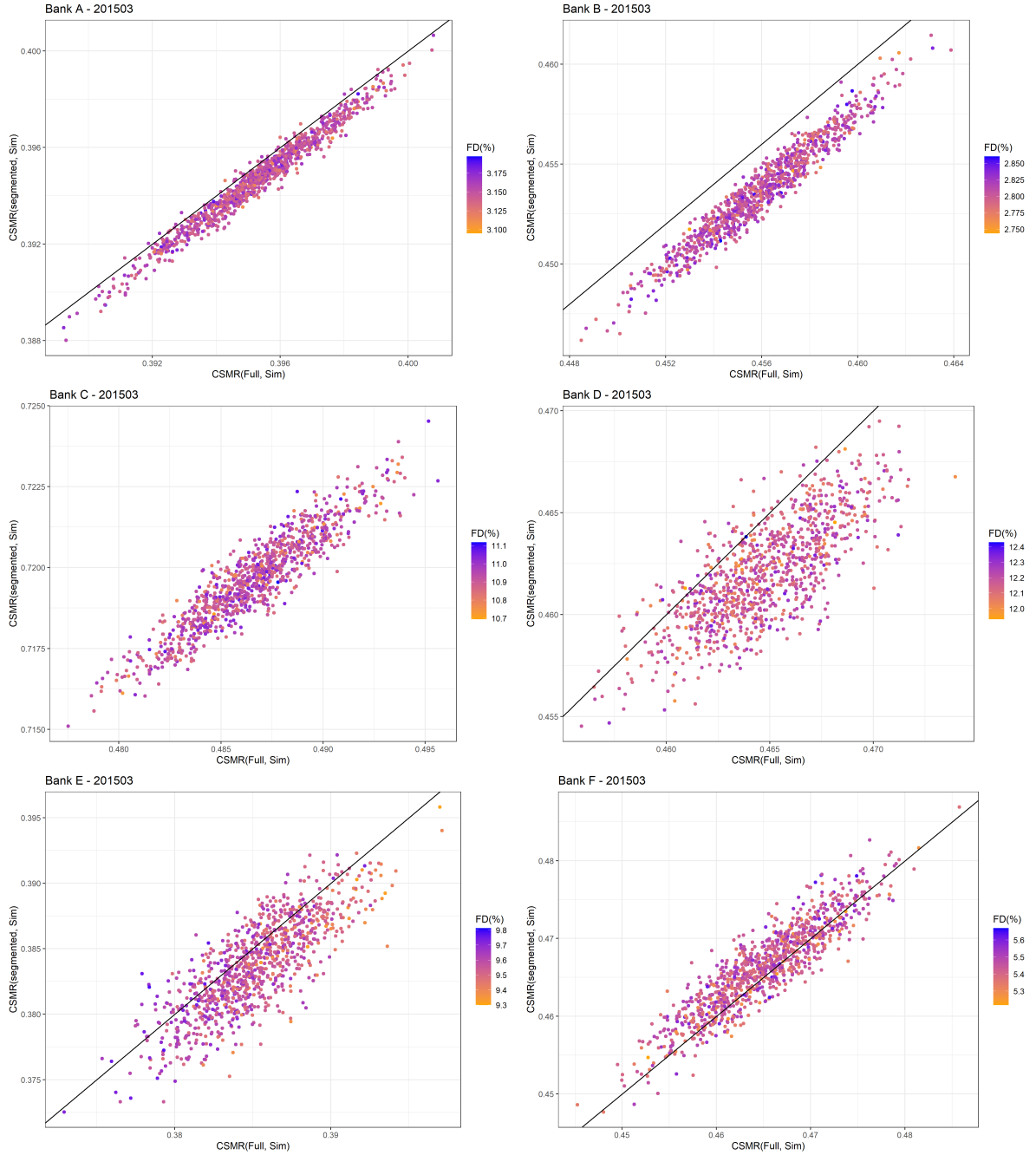


**Figure 6: Simulated correlations for each one of the banks in March 2015.**

Note: X-axis is the full data model simulated correlation, and Y-axis is the Segmented model simulated frequency of default, named here as "FD".

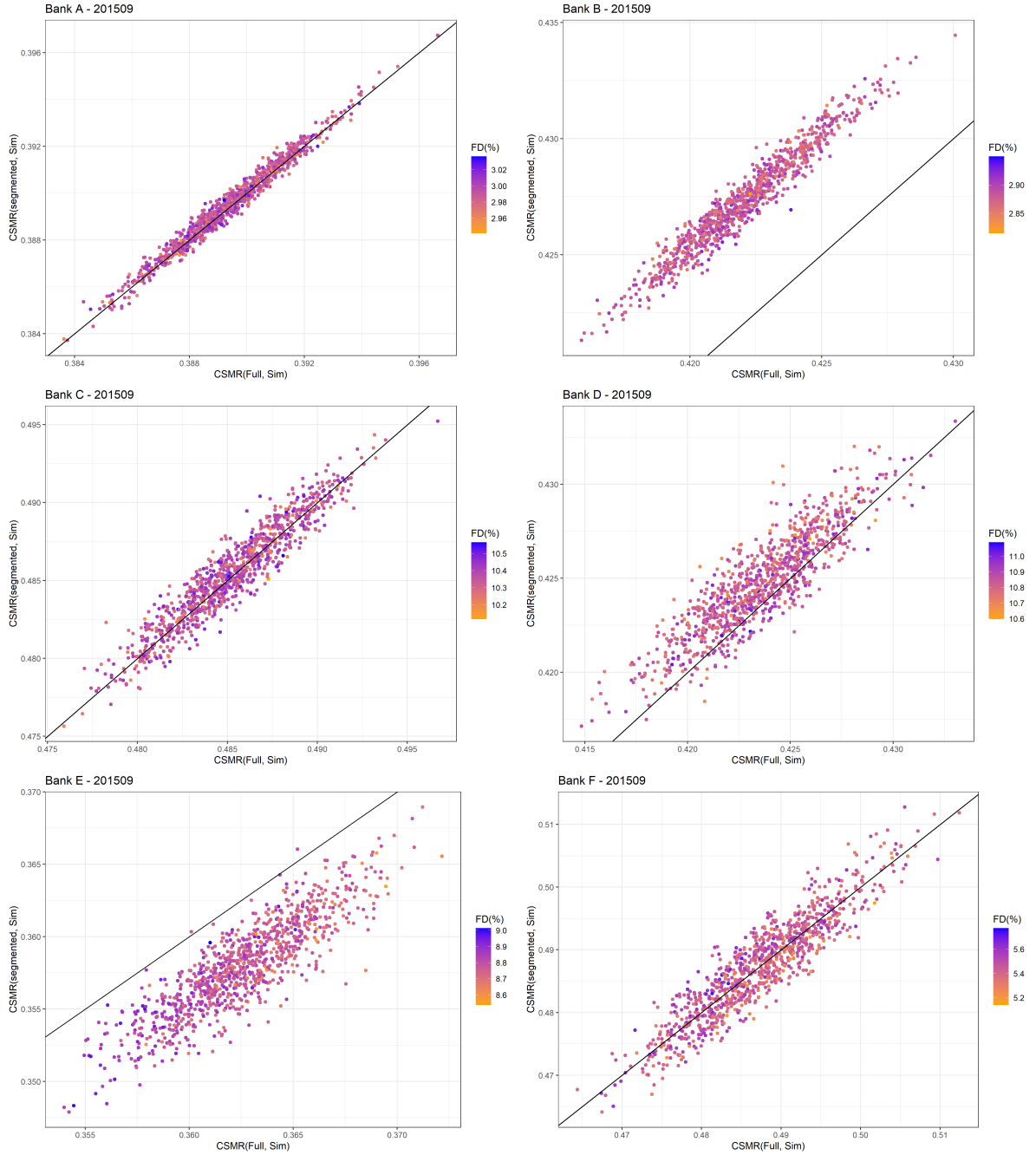


**Figure 7: Simulated correlations for each one of the banks in September 2015.**

Note: X-axis is the full data model simulated correlation, and Y-axis is the Segmented model simulated frequency of default, named here as "FD".

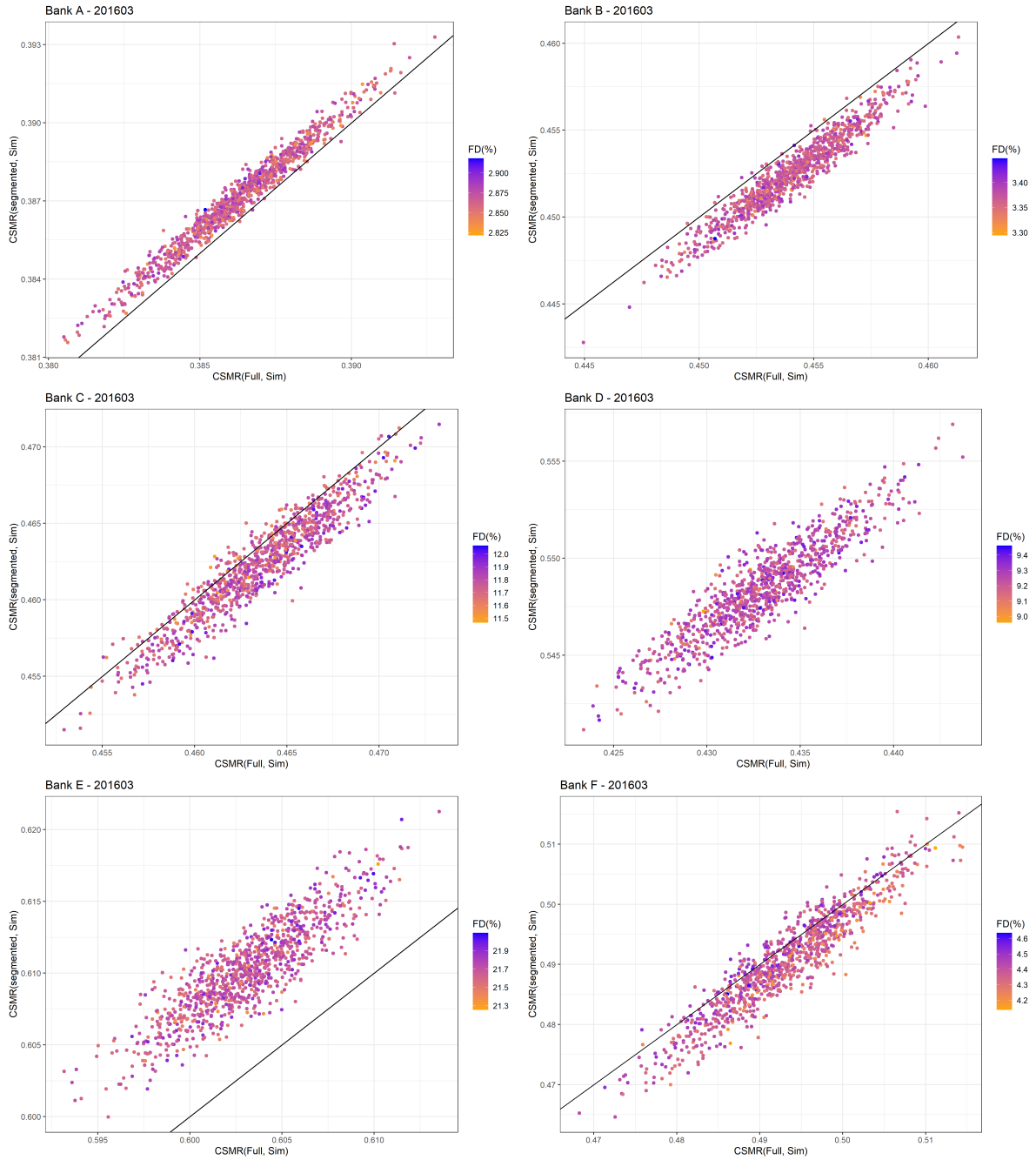


**Figure 8: Simulated correlations for each one of the banks in March 2016.**

Note: The X-axis is the full data model simulated correlation data model simulated correlation, and the Y-axis is the Segmented model simulated frequency of default, named here as "FD".

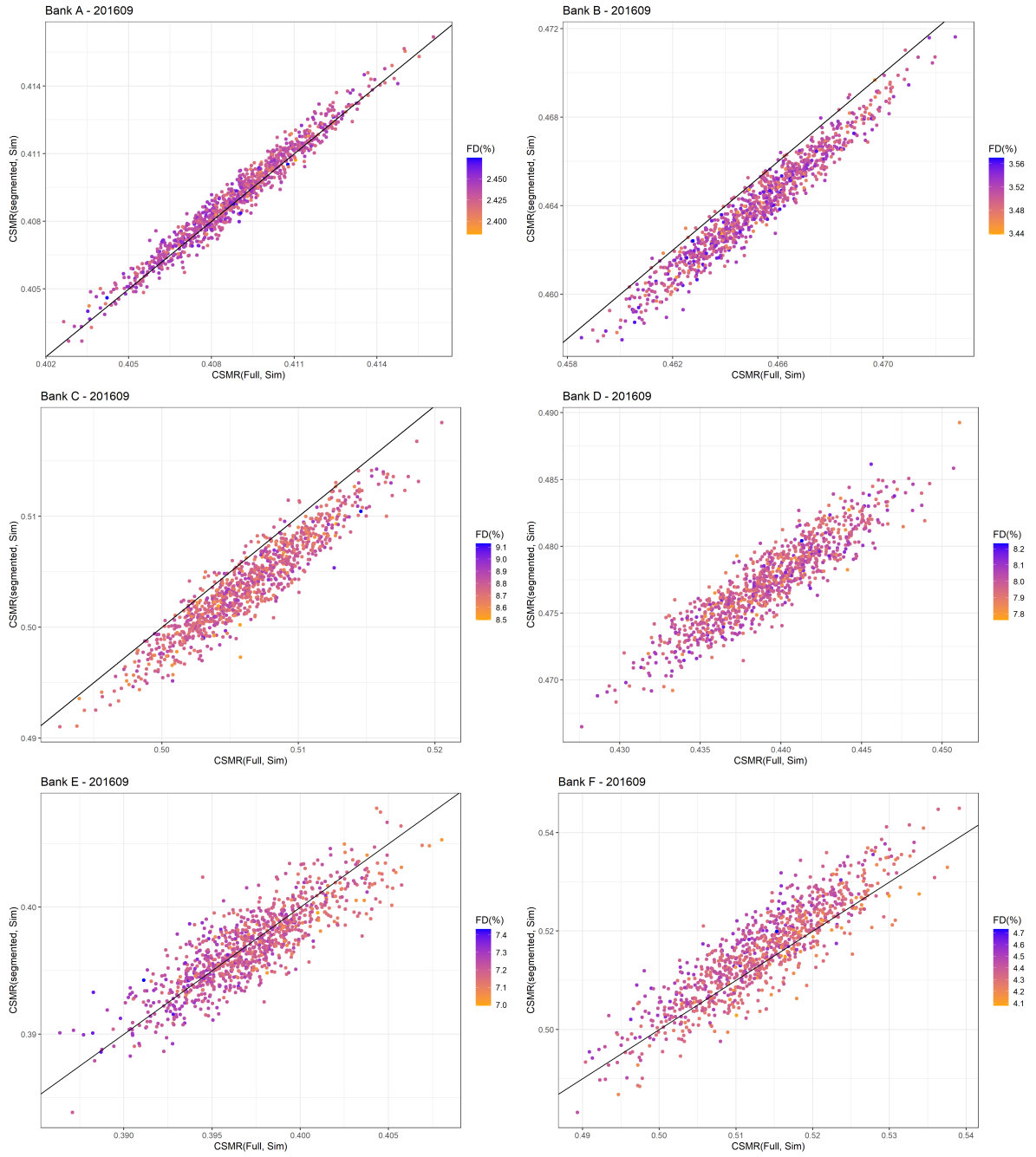


**Figure 9: Simulated correlations for each one of the banks in September 2016.**

Note: The X-axis is the full data model simulated correlation, and the Y-axis is the segmented model simulated frequency of default, named here as "FD".

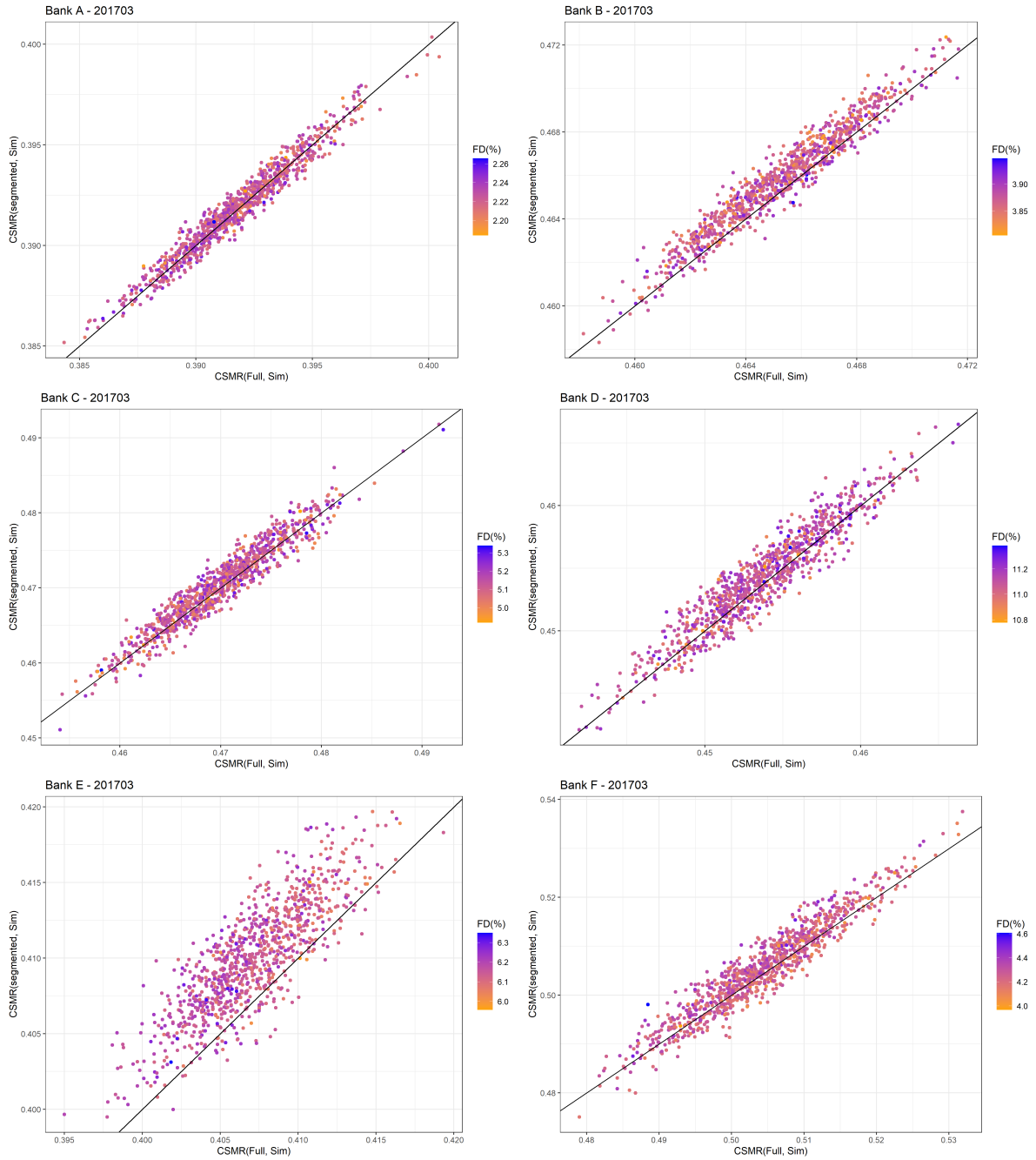


**Figure 10: Simulated correlations for each one of the banks in March 2017.**

Note: The X-axis is the full data model simulated correlation, and the Y-axis is the segmented model simulated frequency of default, named here as "FD".

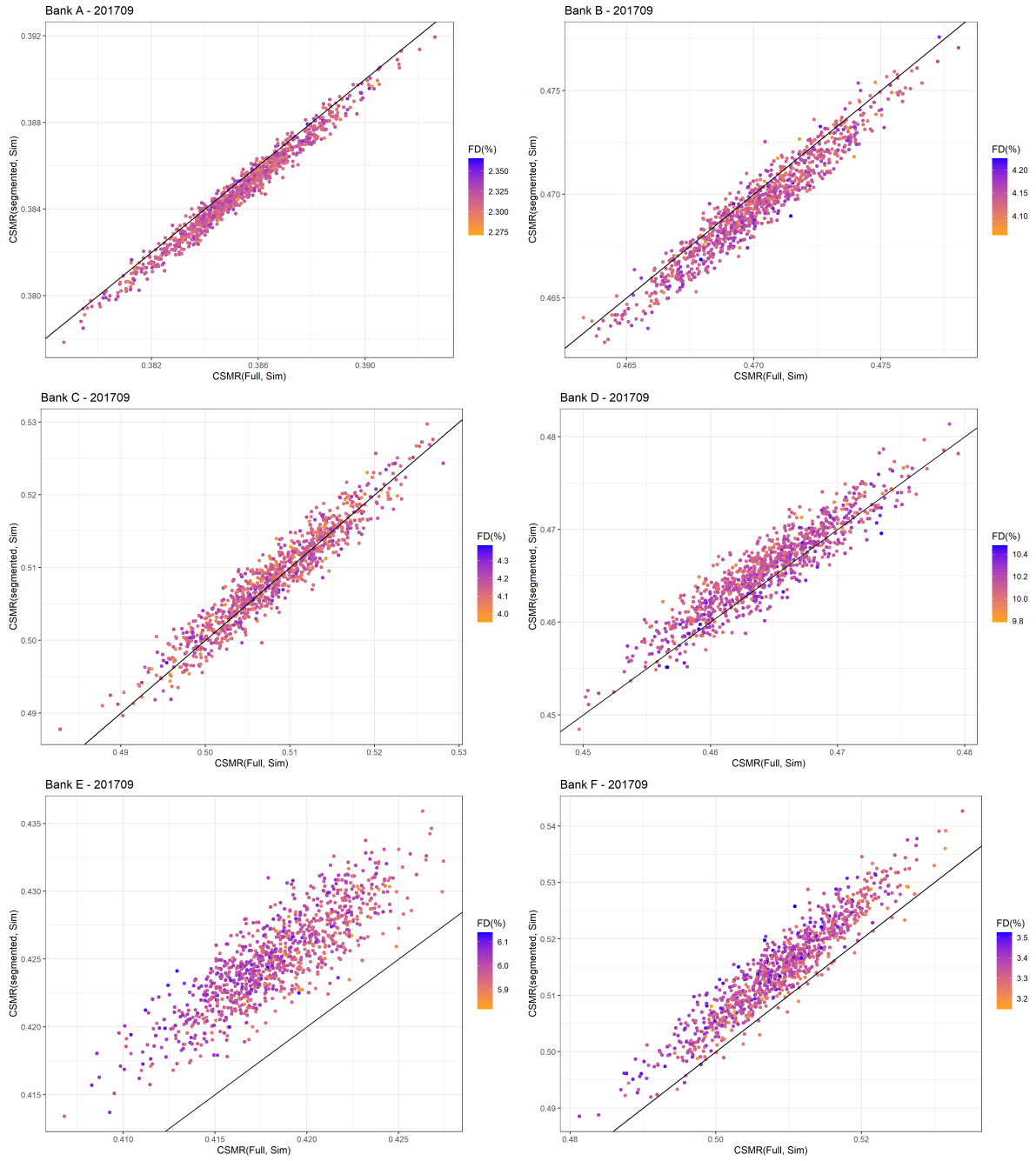


**Figure 11: Simulated correlations for each one of the banks in September 2017.**

Note: The X-axis is the full data model simulated correlation, and the Y-axis is the segmented model simulated frequency of default, named here as "FD".

**Table 4: Simulated correlations' difference**

| Out-of-Time | Mar-14 | Sep-14 | Mar-15 | Sep-15 | Mar-16 | Sep-16 | Mar-17 | Sep-17 |
|---|---|---|---|---|---|---|---|---|
| Bank A | 0.0034 | 0.0007 | -0.0006 | 0.0001 | 0.0008 | 0.0002 | 0.0001 | -0.0005 |
| Bank B | -0.0057 | -0.0059 | -0.0023 | 0.0053 | -0.0015 | -0.0012 | 0.0006 | -0.0007 |
| Bank C | -0.0027 | 0.1550 | 0.2330 | 0.0001 | -0.0012 | -0.0025 | 0.0005 | 0.0004 |
| Bank D | -0.0059 | -0.0014 | -0.0027 | 0.0011 | 0.1155 | 0.0381 | 0.0005 | 0.0014 |
| Bank E | -0.0004 | -0.0056 | -0.0015 | -0.0046 | 0.0070 | -0.0003 | 0.0029 | 0.0067 |
| Bank F | 0.0388 | -0.0035 | 0.0013 | -0.0002 | -0.0025 | 0.0025 | 0.0012 | 0.0059 |

Notes: Median of out-of-time difference of full and segmented data correlations (median of $\rho_{\text{SimulatedY,FullData}\hat{Y}} - \rho_{\text{SimulatedY,SegmentedData}\hat{Y}}$). Correlations are estimated based on simulated default sets and the predictions of both models by the bank. The numbers in red represent incorrect choices when Approach 2 is used for decision-making.

Table 5 presents the differences between the adapted left-hand side of Equation 12, $((\text{E}\left[\hat{Y}_{\text{Full}}|Y=1\right] - E\left[\hat{Y}_{\text{Full}}\right])/(\text{E}\left[\hat{Y}_{\text{Segmented}}|Y=1\right] - E\left[\hat{Y}_{\text{Segmented}}\right]))$ and its adapted right-hand side, the ratio between standard deviations $((\sigma_{\hat{Y}_{\text{Full}}})/(\sigma_{\hat{Y}_{\text{Segmented}}}))$. According to this approach, if the difference is positive ($i.e.$, if the left-hand side is larger than the right-hand side), the full model has a higher correlation than the segmented model.

**Table 5: Differences of the adapted left-hand side of Equation 12 and the ratio between standard deviations.**

| Out-of-Time | Mar-14 | Sep-14 | Mar-15 | Sep-15 | Mar-16 | Sep-16 | Mar-17 | Sep-17 |
|---|---|---|---|---|---|---|---|---|
| Bank A | -0.0309 | -0.0255 | 0.0058 | -0.0028 | -0.0155 | 0.0404 | 0.0368 | 0.0007 |
| Bank B | 0.0431 | -0.0219 | -0.0249 | -0.0961 | -0.0298 | -0.0238 | 0.0072 | -0.0299 |
| Bank C | -0.0177 | 0.1850 | 0.3455 | -0.0668 | -0.0271 | -0.0346 | 0.0283 | 0.0033 |
| Bank D | 0.2182 | 0.0559 | 0.1272 | 0.0438 | 0.1929 | 0.0432 | -0.0271 | -0.0366 |
| Bank E | -0.0773 | -0.1868 | -0.1166 | -0.0743 | 5.5385 | -0.0556 | -0.1157 | -0.1892 |
| Bank F | 0.0423 | 0.0145 | 0.0175 | -0.0922 | -0.0159 | -0.1127 | 0.0919 | 0.0696 |

Notes: Differences of $((\text{E}\left[\hat{Y}_{\text{Full}}|Y=1\right] - E\left[\hat{Y}_{\text{Full}}\right])/(\text{E}\left[\hat{Y}_{\text{Segmented}}|Y=1\right] - E\left[\hat{Y}_{\text{Segmented}}\right]))$ and the ratio between models standard-deviations $((\sigma_{\hat{Y}_{\text{Full}}})/(\sigma_{\hat{Y}_{\text{Segmented}}}))$, from March 2014 to September 2017. If the results are positive, the full model has a higher correlation than the segmented one. The numbers in red represent incorrect choices when considering Approach 3 (Bayesian estimation of covariances) for decision-making.

If we consider Approach 3 (Bayesian estimation of covariances) in decision-making, we will choose higher out-of-time correlation models in 32 out of 48 models' comparisons, or 2/3 (66.7%) of the comparisons. Therefore, in several correct choices, Approach 3 is worse than Approach 2 or the Monte Carlo simulation approach (which leads to higher correlation decisions in 70.8% of the comparisons) for our set of estimations. However, in the incorrect choices ($i.e.$, when the approach criterion leads to choose a lower correlation model in out-of-time datasets), the average of the absolute value of the difference between observed out-of-time correlations ($\rho_{Y,\hat{Y}_{\text{Full}}} - \rho_{Y,\hat{Y}_{\text{Segmented}}}$) is smaller in Approach 3 (0.0119) than in Approach 2 (0.0160) or Approach 1 (0.0225). These comparisons indicate that although Approach 3 leads to a larger number of incorrect choices than Approach 2, the difference between the models correlations is smaller when incorrect choices occur. As such, the magnitude of the mistake and the possible

economic effects of the model's choice are smaller with the use of Approach 3 than with the other approaches.

### 6.4 Comparing approaches

Table 6 summarizes the decision-making criteria based on in-sample and out-of-sample lower *CSMR*s (or higher correlations) for each of the three approaches.

The out-of-time correlations are not homogeneous across banks. In banks B and E (except for Bank B in September 2015), *CSMR* is still lower in segmented models. In banks C and D, it is quite the contrary; in most comparisons, *CSMR* is lower in the conditional full data models. Banks A and F exhibited more unstable behavior. Except for Bank E, there is no consistent stochastic dominance between the models.

**Table 6: Decision-making criteria.**

| | | Mar-14 | Sep-14 | Mar-15 | Sep-15 | Mar-16 | Sep-16 | Mar-17 | Sep-17 | # correct | Mean (incorrect) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Bank A** | *In-sample* | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | 3 | 0.0204 |
| | *Out-of-sample* | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | 3 | 0.0204 |
| | *A1: Similar shrinkage* | Segmented | Segmented | Full | Segmented | Full | Segmented | Full | Full | 5 | 0.0291 |
| | *A2: MC simulation* | Full | Full | Segmented | Full | Full | Full | Full | Segmented | 5 | 0.0046 |
| | *A3: Bayesian cov* | Segmented | Segmented | Full | Segmented | Segmented | Full | Full | Full | 3 | 0.0184 |
| **Bank B** | *In-sample* | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | 7 | 0.0056 |
| | *Out-of-sample* | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | 7 | 0.0056 |
| | *A1: Similar shrinkage* | Segmented | Segmented | Segmented | Segmented | Full | Segmented | Segmented | Segmented | 6 | 0.0080 |
| | *A2: MC simulation* | Segmented | Segmented | Segmented | Full | Segmented | Segmented | Full | Segmented | 7 | 0.0076 |
| | *A3: Bayesian cov* | Full | Segmented | Segmented | Segmented | Segmented | Segmented | Full | Segmented | 5 | 0.0089 |
| **Bank C** | *In-sample* | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | 1 | 0.0699 |
| | *Out-of-sample* | Segmented | Full | Full | Segmented | Segmented | Segmented | Segmented | Segmented | 3 | 0.0093 |
| | *A1: Similar shrinkage* | Segmented | Full | Full | Segmented | Segmented | Segmented | Full | Segmented | 4 | 0.0111 |
| | *A2: MC simulation* | Segmented | Full | Full | Full | Segmented | Segmented | Full | Full | 6 | 0.0098 |
| | *A3: Bayesian cov* | Segmented | Full | Full | Segmented | Segmented | Segmented | Full | Full | 5 | 0.0087 |
| **Bank D** | *In-sample* | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | 2 | 0.0498 |
| | *Out-of-sample* | Segmented | Segmented | Segmented | Segmented | Segmented | Full | Segmented | Segmented | 3 | 0.0449 |
| | *A1: Similar shrinkage* | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Full | Segmented | 1 | 0.0431 |
| | *A2: MC simulation* | Segmented | Segmented | Segmented | Full | Full | Full | Full | Full | 5 | 0.0151 |
| | *A3: Bayesian cov* | Full | Full | Full | Full | Full | Full | Segmented | Segmented | 6 | 0.0129 |
| **Bank E** | *In-sample* | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | 8 | - |
| | *Out-of-sample* | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | 8 | - |
| | *A1: Similar shrinkage* | Segmented | Segmented | Segmented | Segmented | Full | Segmented | Segmented | Segmented | 7 | 0.0046 |
| | *A2: MC simulation* | Segmented | Segmented | Segmented | Segmented | Full | Segmented | Full | Full | 5 | 0.0377 |
| | *A3: Bayesian cov* | Segmented | Segmented | Segmented | Segmented | Full | Segmented | Segmented | Segmented | 7 | 0.0046 |
| **Bank F** | *In-sample* | Segmented | Segmented | Segmented | Segmented | Segmented | Segmented | Full | Full | 5 | 0.0268 |
| | *Out-of-sample* | Full | Segmented | Segmented | Segmented | Segmented | Segmented | Full | Full | 6 | 0.0107 |
| | *A1: Similar shrinkage* | Full | Full | Full | Full | Full | Full | Full | Full | 3 | 0.0082 |
| | *A2: MC simulation* | Full | Segmented | Full | Segmented | Segmented | Full | Full | Full | 6 | 0.0125 |
| | *A3: Bayesian cov* | Full | Full | Full | Segmented | Segmented | Segmented | Full | Full | 6 | 0.0074 |
| **Total # correct** | *In-sample* | 4 | 3 | 2 | 3 | 3 | 4 | 4 | 3 | 26 | 0.0444 |
| | *Out-of-sample* | 5 | 4 | 3 | 3 | 3 | 5 | 4 | 3 | 30 | 0.0222 |
| | *A1: Similar shrinkage* | 5 | 3 | 5 | 2 | 1 | 3 | 5 | 2 | 26 | 0.0225 |
| | *A2: MC simulation* | 6 | 5 | 4 | 5 | 4 | 3 | 3 | 4 | 34 | 0.0160 |
| | *A3: Bayesian cov* | 3 | 4 | 6 | 4 | 3 | 4 | 5 | 3 | 32 | 0.0119 |

Notes: Decision-making considering in-sample and out-of-sample CSMRs and each of the proposed approaches. "A1: Similar shrinkage" is defined in Section 4.1; "A2: MC simulation" (Monte Carlo simulation), in Section 4.2; and "A3: Bayesian cov" (Bayesian estimation of covariances), in Section 4.3. Words in green and red represent correct and incorrect choices (*i.e.*, the approach indicates the correct out-of-time, higher-correlation model). The "# correct" presents the number (of month comparisons) for which each approach indicates the model with a higher correlation in out-of-time datasets. The "Mean (incorrect)" presents the average of the difference between the correlations of the incorrect choices.

The number of correct choices by month is also heterogeneous across approaches. Approach 1 (Similar Shrinkage) indicates five (out of six) models' higher correlation in March

2014, March 2015, and March 2017, but only one in March 2016. Approach 2 (Monte Carlo Simulation) indicates a higher correlation for all six models in March 2014, but only three in September 2016 and March 2017. Approach 3 (Bayesian Estimation of Covariances) indicates higher correlations for all six models in March 2015, but only three in March 2014, March 2016, and September 2017.

In our empirical application, Approach 2 (Monte Carlo Simulation) was found to be more effective for Banks A and C. In Bank A, Approach 2 shows five (out of eight) correct indications of a higher correlation model (and the average difference of correlations of the remaining three incorrect choices is 0.0046). In Bank C, Approach 2 shows six indications of a higher-correlation model. For banks B and E, using segmented models, as indicated by the in-sample and out-of-sample criteria, would be preferable over any of the proposed approaches. This was evidenced by the higher correlation model being selected seven times in Bank B and all eight comparisons in Bank E. The small difference in the correlation of the full data and segmented models in Bank B (0.0056) suggests that the effect of this unique incorrect choice is not economically important. Approach 3 (Bayesian Estimation of Covariances) was more effective for banks D and F, with both banks showing six indications of higher correlation models. Using Approach 3 in Bank D, the average of the difference of the full and the segmented models' correlation for the incorrect choices is 0.0129, and in Bank F, 0.0074. Following those composed criteria, we obtained correct higher correlation models 38 times (out of 48), or 79.2% of the comparisons. The difference in the models' correlation of incorrect choices in the composed criteria was 0.0080.

Table 7 presents a summary of the decision-making criteria.[24] Considering the aggregate results, Approach 2 (Monte Carlo Simulation) presented the highest percentage of correct choices for decision-making. It correctly identified a higher correlation model in 70.8% of the model correlations. However, Approach 3 (Bayesian Estimation of Covariances) presents the lowest average absolute value of the difference between correlations when incorrect identifications are made (0.0119).

**Table 7: Summary of the decision-making criteria.**

| | correct | | | | | | | | Mean | Max |
|---|---|---|---|---|---|---|---|---|---|---|
| | bank A | bank B | bank C | bank D | bank E | bank F | Total | (%) | (incorrect) | (incorrect) |
| *Out-of-sample - PSI(0.10)* | 3 | 7 | 5 | 3 | 7 | 5 | 25/40 | 62.5% | 0.0142 | 0.0713 |
| *A1: Similar shrinkage* | 5 | 6 | 4 | 1 | 7 | 3 | 26/48 | 54.2% | 0.0225 | 0.1752 |
| *A2: MC simulation* | 5 | 7 | 6 | 5 | 5 | 6 | 34/48 | 70.8% | 0.0160 | 0.0554 |
| *A3: Bayesian cov* | 3 | 5 | 5 | 6 | 7 | 6 | 32/48 | 66.7% | 0.0119 | 0.0713 |

Notes: Decision-making criteria. "Out-of-sample—PSI(0.10)" criterion, as presented in Section 5.2. Select the higher-correlation models, as indicated by out-of-sample excluding models with PSI higher than 0.10. Approach "A1: Similar shrinkage" is the Similar Shrinkage criterion, as defined in Section 4.1. The approach "A2: MC simulation" is the Monte Carlo Simulation criterion, as defined in Section 4.2. Approach "A3: Bayesian cov" is the Bayesian Estimation of the Covariance criterion, as defined in Section 4.3. The columns present the sum of correct higher-correlation choices by banks, "Total" and "(%)" present correct choices in number and percentage, "Mean (incorrect)" and "Max(incorrect)" are calculated as the average and maximum of the absolute value of the difference between correlations of the incorrect choices. The gradient of colors from red to green represents the worst to best criteria.

[24] Appendix C presents the complete comparison approaches.

# 7 Concluding Remarks

Yoshida et al. (2025) have shown that, in in-sample datasets, segmented models present a higher correlation (between the observable default variable and its predicted scores) than full data models. Therefore, in in-sample datasets, segmented models present a lower *CSMR* or model risk. However, this behavior does not hold for out-of-time datasets.

In this study, we have shown that in out-of-time contexts, there is no clear dominance in the choice between full and segmented data models. Segmented models of Banks B (seven out of eight models' comparisons) and E (all models' comparisons) still hold higher correlations in out-of-time datasets. However, full data models are preferable for Banks A (five out of eight), C (seven out of eight), and D (six out of eight). In Bank F, there is no preferred model; the full and segmented models have a higher correlation the same number of times (four).

The purpose of our empirical test is not to identify whether full or segmented models systematically dominate or deliver higher out-of-time correlations. Out-of-time performance depends on the nature of the problem, data structure, sampling, and set of independent variables. Instead, this study contributes by developing complementary approaches to support model selection in specific contexts, identifying the model more likely to achieve higher future correlation between observed defaults and predicted scores, and thus lower *CSMR*.

Although Approach 1 (similar shrinkage) is not the best choice for decision-making in aggregate analysis or any specific bank, it would become an effective criterion if we could improve out-of-time shrinkage (or out-of-time $\beta_m$ forecasting). Approach 2 (Monte Carlo simulation) is more effective on a general basis and for banks A and C. Approach 3 (Bayesian estimation of covariance) is more effective for banks D and F, indicating that for these banks, credit portfolio changes play a key role in out-of-time predictions. However, for banks B and E, in-sample or out-of-sample criteria are preferable over any of the proposed approaches, indicating that credit portfolio changes are not important.

The suggested approaches can aid practitioners and regulators in their decision-making processes for selecting and evaluating models. These out-of-time comparison approaches can also be utilized in other prediction and pricing applications, including those designed for public policy and financial systems.

We strongly recommend the development of both (full and segmented models), which is particularly important for some cohorts. The analysis of Approaches 2 (Monte Carlo simulation) and 3 (Bayesian estimation of covariances) in decision-making is also strongly recommended. If one has cost or computational power constraints for developing both models, it is preferable to develop segmented models. However, it is crucial to compare out-of-sample correlations and consider the PSI to identify possible population changes and overfitted models.

Future studies can study different periods for out-of-time analysis (and default definition); improve out-of-time $\beta_m$ forecasting in Assumption 1; develop different probability default functions for Monte Carlo simulations (Approach 2); and propose an alternative Bayesian default distribution for Approach 3.

# 8 Conclusion

The integration of large databases and machine learning techniques heralds a paradigm shift with the power to revolutionize not only the financial sector and banking supervision but also various sectors within the real economy. Previous research on credit scoring models has focused on comparing Machine Learning techniques. However, these studies have relied on traditional measures, emphasizing classification optimization, more accurate prediction technique identification, and in-sample prediction evaluation. The literature has not determined a gold-standard for the best credit machine larning technique.

Although managers and regulators have concerns about the potential risks associated with the discretion of algorithms for variable selection and model building (as well as the lack of causality), insufficient attention has been given to the inappropriate utilization of high-hit-rate credit scoring models or to model risk.

This study focuses on the model risk for credit scoring models. It proposes a measure, the Credit Scoring Model Risk (*CSMR*), to assess the model risk of credit scoring models. It is based on the correlation between the dependent variable ($Y$) of a credit scoring model and its predictions ($\hat{Y}$).

Using *CSMR*, we challenge the conventional belief that more data ($i.e.$, a larger number of observations) will always lead to better-quality inferences. Indeed, for in-sample measures of model risk, bank-specific (or segmented) data models tend to present lower model risk than financial system-wide (or full) data models. We also compared models with different numbers of potential explanatory variables. The results were similar across these models.

Although, in general, a model trained on a more specific subset of the data (segmented by banks) exhibits a greater ability to capture patterns and relationships effectively within that particular subset, a model trained on a more diverse dataset (*e.g.,* , the entire financial system data) might have the advantage of capturing a broader range of patterns and relationships than the segmented dataset. Consequently, despite its lower prediction accuracy in in-sample subsets, the model based on the diverse dataset might deliver better out-of-time predictions owing to its ability to generalize more effectively across various scenarios.

Evaluating model performance solely on in-sample data can lead to instability and unreliable estimation when applied to out-of-sample scenarios. Therefore, basing decision-making (selecting a model from multiple options) solely on in-sample model risk or its performance measures can be problematic for practitioners. Loan portfolios are subject to changes over time, and models might not be appropriately calibrated. Additionally, models may behave differently when confronted with new scenarios, varying macroeconomic conditions or exogenous and stochastic events.

In this paper, we emphasize the importance of utilizing the PSI – the Population Stability Index– to detect potential changes in the population and identify overfitted models. Moreover, we introduce a procedure to identify (or forecast) the best-performing model in out-of-time samples. The main focus of this study is not to establish a dominant model type (full or segmented data) that consistently yields higher correlations in out-of-time datasets. Rather, its contribution lies in creating complementary approaches to assist end-users of credit-scoring models in deciding between segmented or full data for each situation. These approaches fore-

cast the model that is more likely to exhibit a higher correlation between future observable defaults and predicted scores, thereby reducing future *CSMR*.

The proposed approaches can help practitioners and regulators in their decision-making processes for selecting and evaluating models. They can also be used in other prediction and pricing applications, including those aimed at designing public policies.

We conclude by recommending the development of both (full and segmented models), which is particularly important for some cohorts ($i.e.$, some banks). If one has cost or computational power constraints to develop both models, it is preferable to develop segmented models. However, it is crucial to compare out-of-sample correlations and to consider PSI to identify possible populational changes and overfitted models.

A promising avenue for future research is one aiming at optimizing the choice of the best dataset to reduce model risk, both in-sample, as well as out-of-time. In this paper, we use compare between models using the entirety of the loans in the financial system (full-data model) and models estimated using bank-specific loans. One potential problem of the full dataset is that loans from different lenders may present substantial heterogeneity, which may add noise to this model. It is possible that a model estimated using the loans from a subset of selected financial institutions that present some degree of homogeneity among them can outperform both the single-bank and the full-data models. We leave this as a suggestion for future research.

## References


Agarwal, S., S. Alok, P. Ghosh, and S. Gupta (2020). *Financial inclusion and alternate credit scoring for the millennials: role of big data and machine learning in fintech*. Tech. rep. SSRN 3507827. Available at SSRN: https://ssrn.com/abstract=3507827 or http://dx.doi.org/10.2139/ssrn.3507827. Business School, National University of Singapore.

Alonso, A. and J. M. Carbó (2021). *Understanding the performance of machine learning models to predict credit default: a novel approach for supervisory evaluation*. Tech. rep. Documentos de Trabajo N. 2105. Available at https://repositorio.bde.es/handle/123456789/14691. Banco de España.

Ball, C. A. and W. N. Torous (2000). "Stochastic correlation across international stock markets". In: *Journal of Empirical Finance* 7.3-4, pp. 373–388. DOI: 10.1016/S0927-5398(00)00017-7.

Barboza, F., H. Kimura, and E. Altman (2017). "Machine learning models and bankruptcy prediction". In: *Expert Systems with Applications* 83, pp. 405–417. DOI: 10.1016/j.eswa.2017.04.006.

Barrieu, P. and G. Scandolo (2015). "Assessing financial model risk". In: *European Journal of Operational Research* 242.2, pp. 546–556. DOI: 10.1016/j.ejor.2014.10.032.

Cartwright, N. (1979). "Causal laws and effective strategies". In: *Noûs*, pp. 419–437. DOI: 10.2307/2215337.

Copas, J. B. (1983). "Regression, prediction and shrinkage". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 45.3, pp. 311–335. DOI: 10.1111/j.2517-6161.1983.tb01258.x.

Copas, J. B. (1987). "Crossvalidation shrinkage of regression predictors". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 49.2, pp. 175–183. DOI: 10.1111/j.2517-6161.1987.tb01689.x.

Czasonis, M., M. Kritzman, and D. Turkington (2022). "Relevance". In: *Journal of Investment Management* 20.1. Available at https://joim.com/downloads/relevance/, pp. 37–47.

Dastile, X., T. Celik, and M. Potsane (2020). "Statistical and machine learning models in credit scoring: A systematic literature survey". In: *Applied Soft Computing* 91, p. 106263. DOI: 10.1016/j.asoc.2020.106263.

Doumpos, M. and F. Pasiouras (2005). "Developing and Testing Models for Replicating Credit Ratings: A Multicriteria Approach". In: *Computational Economics* 25, pp. 327–341. DOI: 10.1007/s10614-005-6412-4.

Duénez-Guzmán, E. A. and M. D. Vose (2013). "No free lunch and benchmarks". In: *Evolutionary Computation* 21.2, pp. 293–312. DOI: 10.1162/EVCO_a_00077.

Hastie, T., R. Tibshirani, and J. H. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. New York: Springer, pp. 1–758.

Hawkins, D. M. (2004). "The problem of overfitting". In: *Journal of Chemical Information and Computer Sciences* 44.1, pp. 1–12. DOI: 10.1021/ci0342472.

Huang, Y., L. Zhang, Z. Li, H. Qiu, T. Sun, and X. Wang (2020). *Fintech Credit Risk Assessment for SMEs: Evidence from China*. Tech. rep. 2020(193). IMF Working Papers. DOI: 10.5089/9781513557618.001.

Maldonado, Sebastián, Julio López, and Andrés Iturriaga (2022). "Out-of-time cross-validation strategies for classification in the presence of dataset shift". In: *Applied Intelligence* 52.5, pp. 5770–5783.

Malik, M. and L. C. Thomas (2012). "Transition matrix models of consumer credit ratings". In: *International Journal of Forecasting* 28.1, pp. 261–272. DOI: 10.1016/j.ijforecast.2011.01.007.

McNeish, D. M. (2015). "Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences". In: *Multivariate Behavioral Research* 50.5, pp. 471–484. DOI: 10.1080/00273171.2015.1036965.

Oskarsdóttir, M., C. Bravo, C. Sarraute, J. Vanthienen, and B. Baesens (2019). "The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics". In: *Applied Soft Computing* 74, pp. 26–39. DOI: 10.1016/j.asoc.2018.10.004.

Provenzano, A. R., D. Trifirò, A. Datteo, L. Giada, N. Jean, A. Riciputi, and C. Nordio (2020). "Machine learning approach for credit scoring". In: *arXiv preprint arXiv:2008.01687*. DOI: 10.48550/arXiv.2008.01687.

Samuels, M. L. (1993). "Simpson's paradox and related phenomena". In: *Journal of the American Statistical Association* 88.421, pp. 81–88. DOI: 10.1080/01621459.1993.10594297.

Shmueli, G. (2010). "To Explain or to Predict?" In: *Statistical Science* 25.3, pp. 289–310. DOI: 10.1214/10-STS330.

Simpson, E. H. (1951). "The interpretation of interaction in contingency tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 13.2, pp. 238–241. DOI: 10.1111/j.2517-6161.1951.tb00088.x.

Thomas, L. (2009). *Consumer credit models: Pricing, profit and portfolios*. OUP Oxford.

Thomas, L., J. Crook, and D. Edelman (2017). *Credit scoring and its applications*. Society for Industrial and Applied Mathematics.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

Wang, G., J. Hao, J. Ma, and H. Jiang (2011). "A comparative assessment of ensemble learning for credit scoring". In: *Expert Systems with Applications* 38.1, pp. 223–230. DOI: 10.1016/j.eswa.2010.06.048.

Ying, X. (2019). "An overview of overfitting and its solutions". In: *Journal of Physics: Conference Series*. Vol. 1168. IOP Publishing, p. 022022. DOI: 10.1088/1742-6596/1168/2/022022.

Yoshida, Valter T., Rafael Schiozer, Alan de Genaro, and Toni R.E. dos Santos (2025). "A novel credit model risk measure: Do more data lead to lower model risk?" In: *The Quarterly Review of Economics and Finance* 100, p. 101960. ISSN: 1062-9769. DOI: https://doi.org/10.1016/j.qref.2025.101960.

Zhou, L., K. K. Lai, and J. Yen (2014). "Bankruptcy prediction using SVM models with a new approach to combine features selection and parameter optimisation". In: *International Journal of Systems Science* 45.3, pp. 241–253. DOI: 10.1080/00207721.2012.720293.

Zou, G. Y. (2007). "Toward using confidence intervals to compare correlations". In: *Psychological Methods* 12.4, p. 399. DOI: 10.1037/1082-989X.12.4.399.

# Appendix A — Fourth approach — Relevance

In this appendix, we present a fourth possible approach—the relevance sample—to forecast the higher correlation model in out-of-time datasets.

One attempt to incorporate current conditions and perspectives (instead of relying solely on unweighted historical data) is the statistical measure called relevance (Czasonis, Kritzman, and Turkington, 2022). Czasonis, Kritzman, and Turkington (2022) argue that not all historical observations contribute equally to forecasting. They proposed a method to filter observations based on quantified relevance of a set of independent variables. Their framework reconciles classical statistics with intuitive way individuals evaluate experiences and draw inferences about the future, emphasizing contemporaneous conditions and recorded narratives.

Although the relevance approach is conceptually appealing for credit scoring, since portfolios characteristics evolve with macroeconomic conditions and risk appetite, its implementation is limited in large datasets due to substantial computational requirements.

Given these constraints, we adapt the method proposed by Czasonis, Kritzman, and Turkington (2022). Here, we estimate the correlation between the observable default variable and its predictions using samples selected not randomly but based on relevance relative to the out-of-time dataset. We define relevance as the similarity between out-of-time observations ($i.e.$, the vector of model covariates $\mathbf{X}$) and in-sample observations.

The fourth approach follows the procedure below:

1. For each out-of-time set, we generate a random sample of 200,000 observations.
2. For each observation in the sample (from step 1), we select the most similar observation on the in-sample datasets, creating a relevant sample.
3. For each model, we estimate the correlation between the observable default variable and its predictions, considering the relevant samples (selected in step 2).
4. And we compare the full and segmented data correlations (estimated in step 3).

The fourth approach delivers a worse decision-making criterion than Approach 2 (the Monte Carlo simulation) and Approach 3 ( Bayesian estimation of covariances). The results of the fourth approach are presented in Table 8. In an aggregate measure, it is better than Approach 1 (Similar shrinkage); however, it does not capture the highly overfitted models, especially in segmented models of Bank C in September 2014 and March 2015.

The relevant sample cannot capture changes in the covariance matrix (and in the shrinkage) in out-of-time datasets.

**Table 8: Results of the fourth approach**

| | | Mar-14 | | Sep-14 | | Mar-15 | | Sep-15 | | Mar-16 | | Sep-16 | | Mar-17 | | Sep-17 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* |
| **Bank A** | *A4: Relevance* | 0.5930 | **0.6047** | **0.6312** | 0.6187 | **0.6666** | 0.6438 | **0.6483** | 0.6480 | 0.6456 | **0.6696** | 0.6656 | **0.6752** | **0.7055** | 0.6839 | 0.6565 | **0.6631** |
| | *Correlation(Y, Ŷ)* | **0.6442** | 0.5729 | **0.6102** | 0.5994 | **0.6433** | 0.6327 | 0.6211 | **0.6228** | **0.6773** | 0.6743 | 0.6373 | **0.6390** | **0.5830** | 0.5764 | 0.6687 | **0.6740** |
| | *Model Risk* | **0.3558** | 0.4271 | **0.3898** | 0.4006 | **0.3567** | 0.3673 | 0.3789 | **0.3772** | **0.3227** | 0.3257 | 0.3627 | **0.3610** | **0.4170** | 0.4236 | 0.3313 | **0.3260** |
| **Bank B** | *A4: Relevance* | 0.5726 | **0.6221** | 0.5844 | **0.6525** | 0.5942 | **0.6273** | 0.6133 | **0.6559** | 0.5918 | **0.6279** | 0.5757 | **0.5988** | 0.5821 | **0.5991** | 0.5937 | **0.6201** |
| | *Correlation(Y, Ŷ)* | 0.5687 | **0.5822** | 0.5710 | **0.5920** | 0.5927 | **0.6006** | **0.6084** | 0.6028 | 0.5996 | **0.6100** | 0.5937 | **0.6046** | 0.5769 | **0.5845** | 0.5756 | **0.5854** |
| | *Model Risk* | 0.4313 | **0.4178** | 0.4290 | **0.4080** | 0.4073 | **0.3994** | **0.3916** | 0.3972 | 0.4004 | **0.3900** | 0.4063 | **0.3954** | 0.4231 | **0.4155** | 0.4244 | **0.4146** |
| **Bank C** | *A4: Relevance* | 0.6173 | **0.6533** | 0.5740 | **0.5931** | 0.5804 | **0.5865** | 0.5546 | **0.5830** | **0.6171** | 0.5902 | 0.5327 | **0.5709** | 0.6156 | **0.6242** | 0.5260 | **0.6303** |
| | *Correlation(Y, Ŷ)* | 0.5435 | **0.5512** | **0.5145** | 0.3524 | **0.5930** | 0.3118 | **0.6010** | 0.5945 | **0.6162** | 0.6071 | **0.5260** | 0.5154 | **0.5381** | 0.5360 | **0.5406** | 0.5223 |
| | *Model Risk* | 0.4565 | **0.4488** | **0.4855** | 0.6476 | **0.4070** | 0.6882 | **0.3990** | 0.4055 | **0.3838** | 0.3929 | **0.4740** | 0.4846 | **0.4619** | 0.4640 | **0.4594** | 0.4777 |
| **Bank D** | *A4: Relevance* | 0.6100 | **0.6110** | 0.5661 | **0.5799** | 0.6064 | **0.6701** | 0.6130 | **0.6465** | **0.6236** | 0.6229 | 0.5809 | **0.6130** | **0.5875** | 0.5225 | 0.5562 | **0.5793** |
| | *Correlation(Y, Ŷ)* | 0.5383 | **0.5639** | **0.5379** | 0.5311 | **0.5967** | 0.5608 | **0.5968** | 0.5905 | **0.6100** | 0.4348 | **0.5966** | 0.5220 | 0.5706 | **0.5733** | **0.5552** | 0.5550 |
| | *Model Risk* | 0.4617 | **0.4361** | **0.4621** | 0.4689 | **0.4033** | 0.4392 | **0.4032** | 0.4095 | **0.3900** | 0.5652 | **0.4034** | 0.4780 | 0.4294 | **0.4267** | **0.4448** | 0.4450 |
| **Bank E** | *A4: Relevance* | 0.6140 | **0.6507** | 0.5332 | **0.6161** | 0.6184 | **0.7000** | 0.6262 | **0.6810** | **0.6333** | 0.4233 | 0.6084 | **0.6648** | 0.5701 | **0.6663** | 0.5593 | **0.6876** |
| | *Correlation(Y, Ŷ)* | 0.5741 | **0.6196** | 0.5600 | **0.6423** | 0.5933 | **0.6582** | 0.6140 | **0.6634** | 0.6088 | **0.6134** | 0.5924 | **0.6486** | 0.5970 | **0.6525** | 0.5985 | **0.6516** |
| | *Model Risk* | 0.4259 | **0.3804** | 0.4400 | **0.3577** | 0.4067 | **0.3418** | 0.3860 | **0.3366** | 0.3912 | **0.3866** | 0.4076 | **0.3514** | 0.4030 | **0.3475** | 0.4015 | **0.3484** |
| **Bank F** | *A4: Relevance* | 0.5717 | **0.7379** | 0.6051 | **0.6556** | 0.6028 | **0.6276** | 0.5729 | **0.5930** | 0.5008 | **0.6148** | 0.4988 | **0.6732** | 0.5350 | **0.5454** | 0.6031 | **0.6466** |
| | *Correlation(Y, Ŷ)* | **0.5509** | 0.4921 | 0.5290 | **0.5394** | **0.6078** | 0.5907 | 0.5715 | **0.5731** | 0.5513 | **0.5552** | 0.5101 | **0.5307** | **0.4915** | 0.4709 | 0.5411 | **0.5455** |
| | *Model Risk* | **0.4491** | 0.5079 | 0.4710 | **0.4606** | **0.3922** | 0.4093 | 0.4285 | **0.4269** | 0.4487 | **0.4448** | 0.4899 | **0.4693** | **0.5085** | 0.5291 | 0.4589 | **0.4545** |

Notes: Relevant in-sample and out-of-time correlations between default variable and model predictions. Approach 4 is relevant to the in-sample correlations. Correlation$(Y, \hat{Y})$ is an out-of-time correlation. Model Risk is Credit Scoring Model Risk ($CSMR$), equal to one minus the correlation. "Full" refers to full data models, and "Segmen." refers to segmented data models. **Black bold** numbers in Correlation$(Y, \hat{Y})$ show the higher number in each cross-section comparison. **Black bold** numbers in Model Risk (or the $CSMR$)) show the lower number in each cross-section comparison. **Red** (or **green**) numbers in "A4: Relevance" sign reversal (or same) order with Correlation$(Y, \hat{Y})$.

## Appendix B — Correlation autoregressive model

Correlation can be modeled as a stochastic variable (Ball and Torous, 2000). The approach presented in Appendix B uses an autoregressive model based on historical correlations and does not rely on the assumption of calibrated models (Assumption 1) or simulation-based methods (Approach 2). Therefore, we propose a model to predict out-of-time correlations using historical correlation dynamics as follows:

$$\rho_{Y,\hat{Y}}^{t_{1,m,b}} = \alpha_m + \beta \rho_{Y,\hat{Y}}^{t_{0,m,b}} + \gamma_b + \alpha_m \gamma_b \tag{15}$$

where $\rho_{Y,\hat{Y}}^{t_{1,m,b}}$ is the out-of-time ($t_1$) correlation ($\rho_{Y,\hat{Y}}$) in model $m$ of bank $b$;
$\rho_{Y,\hat{Y}}^{t_{0,m,b}}$ is the in-sample ($t_0$) correlation (($\rho_{Y,\hat{Y}}$)) in model $m$ of bank $b$;
$\alpha_m$ is a fixed effect by type of model $m$ (full or segmented data models);
$\gamma_b$ is a fixed effect for bank $b$.

Table 9 presents the results.

**Table 9: Results of the autoregressive model based on historical correlations.**

| | *Estimate* | *Std. Error* | *t value* | *Pr(>\|t\|)* | |
|---|---|---|---|---|---|
| *Full data model* | 0.54870 | 0.08429 | 6.510 | 0.000000 | *** |
| *Segmented data model* | 0.53534 | 0.08585 | 6.236 | 0.000000 | *** |
| *Correl t0* | 0.13776 | 0.13133 | 1.049 | 0.297233 | |
| *Bank B* | - 0.04615 | 0.02201 | - 2.097 | 0.039067 | * |
| *Bank C* | - 0.06819 | 0.02314 | - 2.947 | 0.004169 | ** |
| *Bank D* | - 0.05334 | 0.02274 | - 2.345 | 0.021405 | * |
| *Bank E* | - 0.03845 | 0.02223 | - 1.729 | 0.087465 | . |
| *Bank F* | - 0.08467 | 0.02268 | - 3.733 | 0.000346 | *** |
| *Segmented data model * Bank B* | 0.02075 | 0.03073 | 0.675 | 0.501539 | |
| *Segmented data model * Bank C* | - 0.05030 | 0.03078 | - 1.634 | 0.105967 | |
| *Segmented data model * Bank D* | - 0.02462 | 0.03082 | - 0.799 | 0.426679 | |
| *Segmented data model * Bank E* | 0.05675 | 0.03133 | 1.811 | 0.073732 | . |
| *Segmented data model * Bank F* | 0.00352 | 0.03075 | 0.114 | 0.909142 | |

Notes: Autoregressive Correlation Model $\rho_{Y,\hat{Y}}^{t_{1,m,b}} = \alpha_m + \beta \rho_{Y,\hat{Y}}^{t_{0,m,b}} + \gamma_b + \alpha_m \gamma_b)$.
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

The out-of-time correlations were estimated using the autoregressive correlation model (Equation 15), the results of which are presented in Table 10. The comparison of predictions of out-of-time correlations (full versus segmented data models) behaves in the same manner as real correlation in 36 of 48 comparisons, or 75.0%. Although this approach presents a higher rate of correct inferences, it requires long time-series data to allow for inferences and decision-making.

As in other approaches, incorrect inferences are not homogeneous across banks or months. They occurred three times (out of eight months) in Bank A; once in Banks B and C; twice in

Bank D; none in Bank E; and five times in Bank F.

It is possible to forecast $\beta_m$ in Assumption 1 as an autoregressive model. Using a similar specification, the results (not reported) are worse than those of the correlation autoregressive model.

**Table 10: Comparison of predictions of out-of-time correlations.**

| **Out-of-Time** | Mar-14 | | Sep-14 | | Mar-15 | | Sep-15 | | Mar-16 | | Sep-16 | | Mar-17 | | Sep-17 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* | *Full* | *Segmen.* |
| *Bank A* | 0.6218 | 0.6164 | 0.6290 | 0.6175 | 0.6397 | 0.6267 | 0.6371 | 0.6240 | 0.6374 | 0.6246 | 0.6390 | 0.6265 | 0.6418 | 0.6294 | 0.6392 | 0.6264 |
| *Bank B* | 0.5900 | 0.6006 | 0.5855 | 0.5972 | 0.5814 | 0.5908 | 0.5863 | 0.5949 | 0.5844 | 0.5929 | 0.5859 | 0.5945 | 0.5879 | 0.5963 | 0.5852 | 0.5950 |
| *Bank C* | 0.5543 | 0.4953 | 0.5505 | 0.4938 | 0.5566 | 0.4957 | 0.5546 | 0.4933 | 0.5633 | 0.5026 | 0.5646 | 0.5045 | 0.5712 | 0.5094 | 0.5577 | 0.4961 |
| *Bank D* | 0.5766 | 0.5479 | 0.5680 | 0.5361 | 0.5720 | 0.5397 | 0.5751 | 0.5387 | 0.5777 | 0.5425 | 0.5748 | 0.5400 | 0.5804 | 0.5436 | 0.5776 | 0.5428 |
| *Bank E* | 0.5871 | 0.6408 | 0.5857 | 0.6376 | 0.5923 | 0.6430 | 0.5930 | 0.6433 | 0.5944 | 0.6455 | 0.5948 | 0.6465 | 0.5978 | 0.6476 | 0.5930 | 0.6453 |
| *Bank F* | 0.5495 | 0.5455 | 0.5468 | 0.5392 | 0.5421 | 0.5365 | 0.5444 | 0.5371 | 0.5508 | 0.5452 | 0.5419 | 0.5371 | 0.5447 | 0.5341 | 0.5331 | 0.5230 |

Notes: Appendix B follows an autoregressive model, equation 15. Full refers to the full data model and Segmen., to segmented data models.

# Appendix C - Comparing approaches complete table

Table 11 presents the complete comparison of approaches.

**Table 11: Complete decision-making criteria.**

| | correct | | | | | | | | Mean | Max |
|---|---|---|---|---|---|---|---|---|---|---|
| | bank A | bank B | bank C | bank D | bank E | bank F | Total | (%) | (incorrect) | (incorrect) |
| *Full* | 5 | 1 | 7 | 6 | 0 | 3 | 22/48 | 45.8% | 0.0222 | 0.0823 |
| *Full - PSI(0.25)* | 5 | 1 | 7 | 6 | 0 | 3 | 20/45 | 44.4% | 0.0229 | 0.0823 |
| *Full - PSI(0.10)* | 5 | 1 | 7 | 6 | 0 | 3 | 17/40 | **42.5%** | **0.0236** | 0.0823 |
| *Segmented* | 3 | 7 | 1 | 2 | 8 | 5 | 26/48 | 54.2% | 0.0451 | 0.2812 |
| *Segmented - PSI(0.25)* | 3 | 7 | 3 | 2 | 7 | 5 | 25/45 | 55.6% | 0.0275 | 0.1752 |
| *Segmented - PSI(0.10)* | 3 | 7 | 5 | 2 | 7 | 4 | 23/40 | **57.5%** | **0.0213** | **0.0746** |
| *In-sample* | 3 | 7 | 1 | 2 | 8 | 5 | 26/48 | 54.2% | 0.0444 | 0.2812 |
| *In-sample - PSI(0.25)* | 3 | 7 | 3 | 2 | 7 | 5 | 25/45 | 55.6% | 0.0267 | 0.1752 |
| *In-sample - PSI(0.10)* | 3 | 7 | 5 | 2 | 7 | 4 | 23/40 | **57.5%** | **0.0203** | **0.0746** |
| *Out-of-sample* | 3 | 7 | 3 | 3 | 8 | 6 | 30/48 | 62.5% | 0.0222 | 0.1752 |
| *Out-of-sample - PSI(0.25)* | 3 | 7 | 3 | 3 | 7 | 6 | 27/45 | 60.0% | 0.0222 | 0.1752 |
| *Out-of-sample - PSI(0.10)* | 3 | 7 | 5 | 3 | 7 | 5 | 25/40 | 62.5% | 0.0142 | 0.0713 |
| *A1: Similar shrinkage* | 5 | 6 | 4 | 1 | 7 | 3 | 26/48 | 54.2% | 0.0225 | 0.1752 |
| *A1: Similar shrinkage - PSI(0.25)* | 5 | 6 | 4 | 1 | 7 | 3 | 24/45 | 53.3% | 0.0233 | 0.1752 |
| *A1: Similar shrinkage - PSI(0.10)* | 5 | 6 | 5 | 1 | 7 | 3 | 22/40 | 55.0% | 0.0167 | 0.0746 |
| *A2: MC simulation* | 5 | 7 | 6 | 5 | 5 | 6 | 34/48 | 70.8% | 0.0160 | 0.0554 |
| *A2: MC simulation - PSI(0.25)* | 5 | 7 | 6 | 5 | 5 | 6 | 32/45 | 71.1% | 0.0169 | 0.0554 |
| *A2: MC simulation - PSI(0.10)* | 5 | 7 | 7 | 4 | 5 | 5 | 28/40 | 70.0% | **0.0174** | 0.0554 |
| *A3: Bayesian cov* | 3 | 5 | 5 | 6 | 7 | 6 | 32/48 | 66.7% | 0.0119 | 0.0713 |
| *A3: Bayesian cov - PSI(0.25)* | 3 | 5 | 5 | 6 | 7 | 6 | 30/45 | 66.7% | 0.0124 | 0.0713 |
| *A3: Bayesian cov - PSI(0.10)* | 3 | 5 | 6 | 6 | 7 | 5 | 27/40 | 67.5% | 0.0115 | 0.0713 |
| *A4: Relevance* | 5 | 7 | 2 | 2 | 7 | 5 | 28/48 | 58.3% | 0.0395 | 0.2812 |
| *A4: Relevance - PSI(0.25)* | 5 | 7 | 4 | 2 | 7 | 5 | 28/45 | 62.2% | **0.0201** | 0.0746 |
| *A4: Relevance - PSI(0.10)* | 5 | 7 | 6 | 1 | 7 | 4 | 28/40 | **70.0%** | 0.0220 | 0.0746 |
| *A5: Correlation Autoregresive* | 5 | 7 | 7 | 6 | 8 | 3 | 36/48 | 75.0% | 0.0076 | 0.0256 |

Note: "Full" refers to the selection of full data models. "Segmented" refers to the selection of segmented models. "In-sample" and "Out-of-sample" criteria select, respectively, the higher correlation models as indicated by the in-sample and out-of-sample datasets. Approach "A1: Similar shrinkage" is the Similar Shrinkage criterion, as defined in Section 2.4.1. Approach "A2: MC simulation" is the Monte Carlo Simulation criterion defined in Section 2.4.2. Approach "A3: Bayesian cov" is the Bayesian Estimation of the Covariance criterion, as defined in Section 2.4.3. The suffix "- PSI(0.25)" indicates the respective criteria excluding models with PSI higher than 0.25, while the suffix "- PSI(0.10)" indicates the respective criteria excluding models with PSI higher than 0.10. The columns present the sum of correct higher-correlation choices by banks, "Total" and "(%)" presents correct choices in number and in percentage, "Mean (incorrect)" and "Max(incorrect)" are calculated as the average and maximum of the absolute value of the difference between correlations of the incorrect choices. The red (and green) numbers indicate higher (and lower) measures than those from the criterion without the PSI filter (first line within each criterion). **Bold red** numbers sign the highest measure within each criterion. **Bold green** numbers sign the lowest measure within each criterion.