ADBI Working Paper Series

# SHARING CREDIT DATA WHILE RESPECTING PRIVACY—A DIGITAL PLATFORM FOR FAIRER FINANCING OF MSMEs

Jin-Chuan Duan

No. 1280
August 2021

Asian Development Bank Institute

Jin-Chuan Duan is executive director of the Asian Institute of Digital Finance and Jardine Cycle & Carriage professor at the NUS Business School of the National University of Singapore.



Suggested citation:

Duan, J.-C. 2021. Sharing Credit Data While Respecting Privacy—A Digital Platform for Fairer Financing of MSMEs. ADBI Working Paper 1280. Tokyo: Asian Development Bank Institute. Available: https://www.adb.org/publications/sharing-credit-data-while-respecting-privacy

Please contact the authors for information about this paper.

Email: bizdjc@nus.edu.sg

The author thanks the NUS Credit Research Initiative team for providing the SME data and able research assistance.

Asian Development Bank Institute
Kasumigaseki Building, 8th Floor
3-2-5 Kasumigaseki, Chiyoda-ku
Tokyo 100-6008, Japan

Tel: +81-3-3593-5500
Fax: +81-3-3593-5571
URL: www.adbi.org
E-mail: info@adbi.org

**Abstract**

Lending institutions' reluctance to lend to MSMEs or to offer them competitive interest rates stems from the relatively costly information acquisition for small loans. The central idea is to bridge the information gap between the demand and the supply side by creating a credit analytics sharing infrastructure through federated learning, which completely respects data privacy. Pooling credit information across multiple lending institutions, particularly rare default events, enables the construction of a more informative credit model for MSMEs, which can then serve as a common good among lenders. The technology also allows for lender-specific models, which in essence share the model's parameters on the common prediction variables while differing in their respective alternative data fields. The lenders in the MSME space can work like a coopetition and continue to compete with their varying risk appetites, loan rates, and banking services. We use real MSME credit data to demonstrate the feasibility of the sharing technology and to study the impact of the COVID-19 pandemic via a portfolio that we assembled from four hypothetical banks operating in six ASEAN countries.

**Keywords**: COVID-19, coopetition, alternative data, federated learning, default

**JEL Classification**: C1, C8, G21

# Contents

# 1. INTRODUCTION

Micro, small, and medium-sized enterprises (MSMEs) play a key role in contributing to economic growth and creating employment, yet they face disproportionally large challenges in financing. DiCaprio, Beck, and Daquis (2014), for example, reported that "Gaps in trade finance affect SMEs more negatively than other company respondents. This is a particular problem in Asia where more than 90% of all firms are SMEs." The Asia-Pacific Trade Facilitation Report 2019 of the Asian Development Bank stated that "SMEs are most affected as they tend to have higher rejection rates than larger firms. Banks have higher transaction and information costs when dealing with smaller companies." Evidence for the MSME financial gap abounds in the literature and is a repeated theme in numerous studies.

To put it simply, MSMEs worldwide have always faced hardships in financing, but the difficulties are particularly pronounced for those in emerging economies because the national authorities lack the necessary financial resources to help them. This paper will extend beyond the typical argument to contend that the current way of assisting the financing of MSMEs is ineffective and fails to address the fundamental impediment of informational asymmetry between MSMEs and lending institutions. We will focus the discussion on the Association of Southeast Asian Nations (ASEAN), but the idea and technology that we will describe are universally applicable to other regions, including advanced economies.

According to some estimates, ASEAN is poised to become the world's fourth-largest economy by 2030,[1] and its member nations view digital transformation as a way of growing their economies. This is an opportune time to help the MSME sector to function more effectively with digital technologies. While many national authorities in ASEAN deeply appreciate the importance of helping MSMEs to access financing, among other policy measures, effective and practical solutions to enable effective flows of capital to the sector are still lacking.

Table 1 summarizes, not exhaustively, the financial assistance programs for MSMEs in six of the 10 ASEAN member countries. The MSME financing program in Singapore serves as a good example. Being an advanced economy with solid national finance, the Singapore Government is in an enviable fiscal condition to channel substantial resources to the MSME sector. Enterprise Singapore, among many assistance schemes, offers to share risk with lending institutions on working capital loans to eligible SMEs normally at 50% of a loan default loss, rising to 90% during the COVID-19 pandemic period. Also available is a loan insurance scheme that typically co-pays 50% of the commercial insurance premium, which again has increased to 80% to respond to the pandemic. Such proactive and generous assistance programs have no doubt helped many MSMEs to secure financing that they would otherwise be unable to access.

Leveraging the expertise of lending institutions or insurers to facilitate the financing of MSMEs may instinctively appeal to all as an intelligent and effective way of managing assistance. We would contend, however, that it fails to address the fundamental informational asymmetry between lending institutions and MSMEs. In fact, it may disincentivize lending institutions to invest in better credit assessments of their potential borrowers and thus inadvertently create a perverse consequence by widening the information gap.

1 Singapore PM Lee Hsien Loong's speech. https://www.pmo.gov.sg/newsroom/pm-lee-hsien-loong-27th-worldeconomic-forum-asean-hanoi-vietnam.

**Table 1: A Non-exhaustive Summary of the MSME Financial Assistance Programs in Six of the 10 ASEAN Countries**

| Country | Key Government Policies/Schemes to Help Finance MSMEs |
|---|---|
| Indonesia | **Pre-COVID**<br>• Credit for Business (Kredit Usaha Rakyat)—providing credit, working capital, and investment financing schemes dedicated to micro enterprises, SMEs, and co-operatives.<br>• Program for Eastern Indonesian Small and Enterprise Assistance—collaborating with the International Finance Centre to provide financial assistance to SMEs in the poorest areas of Indonesia.<br>**Additional policy/stimulus post-COVID**<br>• Stimulus for SMEs: IDR123.46 tn—interest subsidies for microcredit (KUR), SME financing, guarantees, and placement of funds in banks. |
| Malaysia | **Pre-COVID**<br>• Shariah-compliant SME financing scheme: subsidy rate of 2%, soft loans, grants, and training under SME Corp and insurance coverage credit facility for SME exporters.<br>• Credit Guarantee Corporation (CGC), in collaboration with SME Corp and Credit Bureau Malaysia—offers loan guarantee and financing facilities and advisory, credit information, and credit-rating services. Through the Bureau, the CGC helps SMEs to improve their creditworthiness.<br>**Additional policy/stimulus post-COVID**<br>• A special grant of RM3,000 for each qualifying micro SME, which must register with the Malaysian Inland Revenue Board.<br>• Enhanced financing schemes for SMEs as follows:<br>• Abolition of the 2% interest rate for the RM500 million Micro Credit Scheme under Bank Simpanan Nasional.<br>• Extension of the easy financing scheme to the TEKUN Nasional Scheme with a fund of RM200 million at an interest rate of 0%. The maximum loan amount is RM10,000 for each micro company. |
| Thailand | **Pre-COVID**<br>• SME Transformation Loan for Thailand 4.0—offering SMEs access to credit of up to THB15 million (USD0.45 million). SMEs applying for a loan of less than 5 million baht (USD 0.15 million) can make fixed interest payments for the first 3 years without collateral, with the Thai Credit Guarantee Corporation acting as guarantor.<br>**Additional policy/stimulus post-COVID**<br>• SME loan restructuring:<br>• Pre-emptive measure against non-performing loans (NPL) through interest reduction and an extensive payment period. This is to avoid being classified as Troubled Debt Restructuring (TDR), with the Credit Bureau, and to be classified as an ordinary loan.<br>• Loan restructuring to promote NPL to ordinary loans when restructuring loans, with three consecutive installments paid off (from 12 instalments).<br>• Measures to support FIs and SFIs in the classification of liquidity loans as ordinary loans (ordinary terms and conditions and a lower interest rate).<br>• Measures to support FIs to maintain unused credit lines.<br>• Financial institutions to monitor closely and report monthly milestones according to the measures, including outstanding loans for SMEs, 21 days after the end of each month.<br>• Soft loans not exceeding THB3 million per business, with a 3% interest rate for the first 2 years, for affected SME entrepreneurs until 30 December 2020. |
| Singapore | **Pre-COVID**<br>• Six categories of loan facilities for Singaporean SMEs: SME Working Capital Loan, SME Fixed Asset Loan, Project Loan, Venture Debt Loan, Trade Loan, and M&A Loan.<br>• Loans are subject to a cap with a default risk share of at least 50%; for example, the cap for the SME Working Capital Loan is 300,000 per borrower with a 50% risk share.<br>• Loan insurance schemes co-pay 50% of the premium.<br>**Additional policy/stimulus post-COVID**<br>• Increased cap and risk share; for example, the SME Working Capital Loan increases the cap to S$1 million, and the government's risk share increases to 90%. SMEs may request deferment of principal repayment for 1 year.<br>• Additionally, FIs may apply for low-cost funding through a new MAS Singapore Dollar facility (extended for another 6 months until September 2021), provided that they pass the savings onto the borrowers.<br>• Loan insurance schemes' premium co-payment has risen to 80%.<br>• Deferment of principal payments on secured term loans. |

*continued on next page*

**Table 1** *continued*

| Country | Key Government Policies/Schemes to Help Finance MSMEs |
|---|---|
| Philippines | **Pre-COVID**<br>• Credit Surety Fund—helps cooperatives to manage and administer credit surety funds to enhance access to finance for micro and SME entrepreneurs, cooperatives, and non-government organizations.<br>• Pondo Para sa Pagbabago at Pagasenso (P3) (President Duterte's flagship microfinancing initiative)—sets aside USD20 m with lower lending rates to eradicate the underground money lender schemes (56 schemes) and shift to micro businesses and other legal microfinancing facilities.<br>**Additional policy/stimulus post-COVID**<br>• Non-application of interest, fees, and charges to future payments and/or amortization of individuals, households, micro, small, and medium enterprises (MSMEs), and corporate borrowers. |
| Viet Nam | **Pre-COVID**<br>• Global Company Partnership Grant and Market Readiness Assistance Grant—offer SMEs up to 70% funding support for overseas expansion projects in capability building, market access, and manpower development. The grants support overseas setup, business partner identification, and marketing.<br>**Additional policy/stimulus post-COVID**<br>• 30% corporate income tax (CIT) reduction—entitling businesses with total revenue in 2020 not exceeding VND200 billion (around USD8.5 million) to a 30% reduction of CIT payable in 2020. |

Governments in both advanced and emerging economies worldwide have rolled out special assistance programs for MSMEs in response to the COVID-19 pandemic. These specific financial assistance schemes that the six ASEAN countries and elsewhere have offered, which Table 1 describes, will phase out once the outbreak is under control. The improved economies will in a natural course help to restore many MSMEs to their pre-pandemic operations.

However, the structural financial difficulties facing MSMEs in the pre-pandemic period will not simply vanish with the coronavirus unless either government or public/private-sector efforts put effective structural measures in place. As for the impact of the special COVID policy measures, it is imperative for national authorities to conduct post-mortem analyses of these measures' efficacy and learn from the experience to aid future policy formation.

Before proceeding, we contend that not all MSMEs deserve or should receive financial assistance. When a business idea is unsound and the operation is fundamentally nonprofitable, making subsidized financing available not only misallocates capital but also deepens the unnecessary losses that the owner-entrepreneur incurs. Focusing on the likelihood of success/failure with an evidence-based approach must therefore be a key part of the solution. There should be an incentive for lending institutions to obtain better information on the credit quality of borrowers as opposed to becoming further disengaged due to the government's loss-sharing assistance scheme. In short, our view is that it would be better for the policy objective to focus on building a shared infrastructure to enhance the quality of credit assessments, through which lenders' competition can naturally achieve fairer financing of MSMEs.

With the abundant capital and liquidity in today's financial markets, a lack of information rather than scarce capital lies at the heart of the MSME financing challenge. Building up an MSME information-sharing infrastructure and treating it as a *common good* among lending institutions, in our view, constitute a more productive way to remove the key impediment to channeling much-needed capital to the MSME sector, particularly to those small operations that are in a better position to create jobs and contribute to economic prosperity.

Digital technology enables us to contemplate a new-style soft infrastructure that facilitates the sharing of data across lending institutions and helps to harness alternative data relevant to credit risk assessment. We will elaborate on the idea and its implementation principles later. Such a soft infrastructure has the same spirit as the physical infrastructure, much like a fiber-optic or high-speed train network, which enhances the overall productivity of an economy. When a lending institution can ascertain at a rather low cost whether the credit quality of an MSME borrower has met its credit standard, it will make business sense to lend without needing a third-party's encouragement. If the credit assessment of a potential borrower is costly, typical MSME loan sizes would not be large enough to justify incurring significant costs to undertake information acquisition.

The central idea underpinning our proposed solution is to create a digital MSME credit analytics platform, a *common good*, for lending institutions to share. Members differing in their risk appetite can compete in loan pricing and services to form, in essence, a *coopetition* business model, which will stand a much better chance of leading to fairer financing of MSMEs.

Defaults are rare events, and therefore data sharing can obviously improve the quality of, say, a probability of default (PD) model. The calibration of a credit model to pooled data from multiple lending institutions needs to respect the data privacy of individual sites. "Federated learning" underlies the technical approach to calibrating a credit model only using the highly aggregated functional values of the member institutions so that there is no need to transmit privacy-sensitive data to another party.

It is necessary to train the credit models under federated learning on the data residing in a distributed network of multiple lending institutions. The technical design needs to utilize both edge and soft computing to gain operational robustness over network latency and local data site failures in a distributed network. Each data site acts as edge storage and performs edge computing to generate and transmit back highly aggregated functional values to serve as the basis for calibrating a model's parameters. Inversion from these highly aggregated functional values back to the values of the input variables is impossible. This design thus ensures the preservation of total data privacy. It is then possible to share the calibrated model as a *common good* among the member institutions.

The development of the *i*CASS (**i**ntelligent **C**redit **A**nalytics **S**haring **S**ystem)[2] software has already taken place to realize the calibration of large-scale parametric credit models over multiple privacy-protected distributed data sites. The optimization method underpinning this software is the density-tempered sequential Monte Carlo (SMC) technique, which is capable of locating the global optimum for large identifiable parametric models, distinguishing itself from, say, the stochastic gradient descent method that researchers have commonly used for obtaining heuristic solutions for neural network models. The test results show that this new federated learning system is indeed robust to network latency and tolerant of localized data site failures during a calibration session.

We will demonstrate in a shared-data setting how to calibrate a common default prediction model involving four credit portfolios, each corresponding to a hypothetical MSME bank operating in three of the six ASEAN countries. The data, inclusive of the COVID-19 period, come from real exchange-traded SMEs, but the credit portfolios are hypothetical. Each portfolio does not have enough default cases in its own sample to pin down parameter estimates reliably, particularly regarding the COVID impact. This

[2] *i*CASS is a joint effort of the Asian Institute of Digital Finance (AIDF) at the National University of Singapore (NUS) and CriAT, a Singapore-registered FinTech firm and an NUS spin-off. This author leads the development team, consisting of members from both AIDF and CriAT. In the interest of full disclosure, this author is a co-founder and the non-executive Chairman of CriAT and concurrently serves as the Executive Director of AIDF.

study design intends to show how data sharing can aid policy analysis by emulating the real-world MSME lending market. The special COVID relief measures, such as the concessional loans and deferment of loan payments that Table 1 describes, can cause typical MSMEs to experience worsening leverage but improved liquidity. Lowering the short-term default likelihood and raising the longer-term credit risk in effect twist the shape of the term structure of default probabilities for MSMEs. Pooling together the data of the four lending institutions helps to shed light on the issue.

Beyond the privacy-protected data-sharing technology, we lay out some general principles and vital components that can guide the formation of a consortium of lending institutions and the development of the user-support infrastructure. The days of fairer financing of MSMEs in ASEAN or elsewhere through coopetition may become a reality in a not-too-distant future.

# 2. DATA PRIVACY AND FEDERATED CREDIT MODEL CALIBRATION

Data privacy protection is a typical issue that data usage agreements and/or laws/regulations dictate. How and under what conditions companies can use the credit data pertaining to an obligor, a natural or legal person, are in principle clear. Lending institutions have explicit consent to use a customer’s specific information for internal operation purposes. Pooling data falling under different lenders, that is, guardians of the data, without anonymization is obviously impermissible.

When an individual or corporation seeks credit facilities from a bank, it voluntarily provides sensitive information to the lending institution, which in turn must ensure respect of data privacy. Aggregating the credit data of multiple lending institutions into a single database is likely to encounter insurmountable legal complications. We thus need to think of an alternative route through the utilization of digital technology.

A decentralized database may be technologically superior in some contexts but not for calibrating credit models over many decentralized databases because doing so involves many network-related issues, such as network latency and occasional individual data site failures. However, facing distributed credit databases in the possession of multiple lending institutions is an operational reality that data privacy concerns dictate.

The main uses of credit models include the prediction of default and recovery rates. The task may also involve developing tools for portfolio credit analysis or pricing credit derivative contracts. Here, we focus on default prediction, that is, estimating the probability of default (PD), which is the most likely area in which lending institutions would need to share data. This is because default events are rare and data sharing has the obvious advantage of materially affecting the quality of a model.

## 2.1 Federated Model Calibration

The issue and technical challenges that we are facing in essence fall under an increasingly popular term, “federated learning,” which Konečný, McMahan, and Ramage (2015) introduced. Federated learning aims to train a model iteratively over multiple distributed data sites without explicitly exchanging data samples. Not exchanging data holds the key to preserving data privacy. Its typical usage is in machine learning models, such as neural networks.

The construction of credit models can take place through various approaches. For example, neural networks and several other machine learning techniques have gained popularity in recent years. It is possible to deploy these machine learning models to classify borrowers into risk categories. Notwithstanding their popularity, neural network credit models are fundamentally deficient both scientifically and in practical usage in terms of credit modeling. Because a neural network model inherently has numerous local optima and saddle points, its optimization in practice always settles for a heuristic solution.

Research has often shown that such heuristically obtained neural network models are powerful in making predictions in various applications. However, those models are fundamentally uninterpretable and thus ill-suited to managerial usage beyond making simple predictions. It is also well known that these machine learning tools are seriously inadequate for situations that require extrapolation, such as stress testing. In short, predictions in a region that the training data have not previously covered will be entirely unreliable. Risk classification is also insufficient for practical credit risk management because users often need granularity to the level of PD. Real-time usage in banks, for example, also requires the aggregation of individual borrowers into credit portfolios.

For the above reasons, we contend that the preferred credit models should take advantage of conventional parametric approaches building on the accumulated financial and economic knowledge and insights. Naturally, this parametric approach needs to incorporate modern big data techniques to combine the strength of the established theory/intuition on credit risk and the information embedded in a large quantity of data.

The credit models that this paper covers are along the lines of the forward-intensity corporate default prediction approach of Duan, Sun, and Wang (2012), which the Credit Research Initiative (CRI) at the National University of Singapore implemented on exchange-listed firms globally.[3] The purpose here is to extend the usage of this line of models or others that research has proven to perform robustly in applications to the MSME space through a new federated learning design.

It is possible to view a PD model as a mathematical function linking the chance of seeing the realization of an outcome, which we denote as $Y^{(i)}_{t+\tau,t+\tau+q}$, over a future time period $(t+\tau, t+\tau+q]$, to a borrower's many attributes, $\boldsymbol{X}_{i,t}$, available at the prediction time $t$. Specifically, it is possible to express borrower *i*'s forward PD at time *t* for such a future period as

$$\mathrm{Prob}_t\left(Y^{(i)}_{t+\tau,t+\tau+q}=1\right)=f(\boldsymbol{X}_{i,t};\tau,q,\boldsymbol{\theta})$$

where $Y^{(i)}_{t+\tau,t+\tau+q}=1$ represents a borrower default in the specified time period and $f\left(\boldsymbol{X}_{i,t};\tau,q,\boldsymbol{\theta}\right)$ is a positive nonlinear function. The forward starting time, $\tau$, must enter into the consideration because a future credit event can occur at different points of time, when, for example, a lending contract is for 2 years with 1 month representing a basic time interval. The functional form $f(\cdot;\boldsymbol{\theta})$ determines the type of model, whereas the multidimensional parameter value $\boldsymbol{\theta}$ fixes the model.

3 See NUS-CRI Staff (2021). The CRI deploys the model of Duan, Sun, and Wang (2012) to generate daily updated PDs on over 80,000 exchange-listed corporations in 133 economies globally and distributes them free of charge through its website (https://www.nuscri.org).

Apart from default, an MSME borrower may terminate its banking relationship for various reasons (acquisition by another firm, banking with a different institution, or dissolving to stop losses), which we denote as $Y^{(i)}_{t+\tau,t+\tau+q} = 2$, and we need to model this as a different function. An MSME that does not experience either a default or another form of exit over a period is a complementary event, $Y^{(i)}_{t+\tau,t+\tau+q} = 0$, which does not require another modeling function simply because the three events must add up to a 100% probability. Hence, we only need a second function,

$$\text{Prob}_t\left(Y^{(i)}_{t+\tau,t+\tau+q} = 2\right) = g(\boldsymbol{X}_{i,t}; \tau, q, \boldsymbol{\vartheta})$$

to describe the dynamic system for a firm that may survive multiple periods or default (exit due to other reasons) in one of the periods. These two forward probability functions—$f(\boldsymbol{X}_{i,t}; \tau, q, \boldsymbol{\theta})$ and $g(\boldsymbol{X}_{i,t}; \tau, q, \boldsymbol{\vartheta})$—form the basis for constructing a term structure of PDs to serve various needs in credit risk management.

Figure 1 depicts conceptually the configuration of this federated calibration system. The Asian Institute of Digital Finance (AIDF), say, operates the Calibration Central, which interacts with multiple lending institutions, which the schema describes as consortium members. On receiving parameter values $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ from the Calibration Central, member *m* computes and submits an aggregated quantity reflective of its contributed credit data pool with $N_m$ borrowers over multiple historical time points $\boldsymbol{T} = \{t_1, t_2, \cdots, t_j\}$. This quantity in the current context is the log-likelihood of the data sample that member *m* contributes with its pool of borrowers who have survived up to some $t \in \boldsymbol{T}$.

The specific prediction reflects a forward starting time $\tau$ and a targeted prediction duration *q*; that is,

$$\begin{aligned} L_m(\boldsymbol{\theta}, \boldsymbol{\vartheta}; \tau, q) = \sum_{t \in \boldsymbol{T}} \sum_{i=1}^{N_m} \Big( & 1_{\{Y^{(i)}_{t+\tau,t+\tau+q}=1\}} \ln\left[f(\boldsymbol{X}_{i,t}; \tau, q, \boldsymbol{\theta})\right] + 1_{\{Y^{(i)}_{t+\tau,t+\tau+q}=2\}} \ln\left[g(\boldsymbol{X}_{i,t}; \tau, q, \boldsymbol{\vartheta})\right] \\ & + 1_{\{Y^{(i)}_{t+\tau,t+\tau+q}=0\}} \ln\left[1 - f\left(\boldsymbol{X}_{i,t}; \tau, q, \boldsymbol{\theta}\right) - g(\boldsymbol{X}_{i,t}; \tau, q, \boldsymbol{\vartheta})\right] \Big) \end{aligned}$$

In the above expression, $1_{\{\cdot\}}$ denotes an indicator function that returns 1 if the condition is true and 0 otherwise. It is evident from the above equation that inverting the process to find an individual $\boldsymbol{X}_{i,t}$ or $Y^{(i)}_{t+\tau,t+\tau+q}$ from $L_m(\boldsymbol{\theta}, \boldsymbol{\vartheta}; \tau, q)$ will be impossible, thus preserving data privacy.

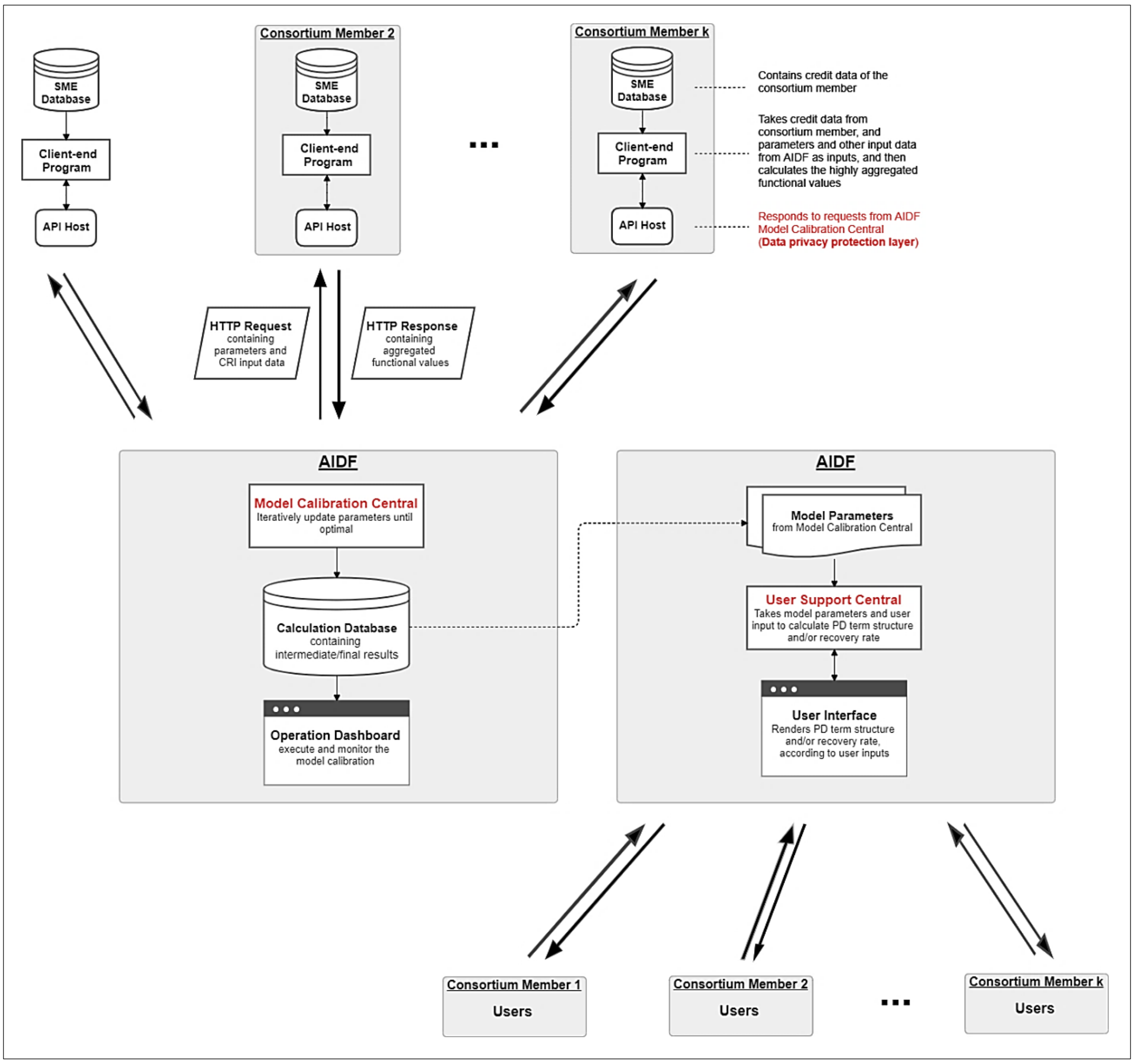
Adding the log-likelihood values from all *K* member institutions together allows the Calibration Central to compute the overall target function value at the parameter values $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$, which is

$$L(\boldsymbol{\theta}, \boldsymbol{\vartheta}; \tau, q) = \sum_{m=1}^{K} L_m(\boldsymbol{\theta}, \boldsymbol{\vartheta}; \tau, q)$$

This quantity then serves as the basis on which to update the parameter values.

One must decide on a robust way of finding $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\vartheta}}$ that maximizes $L(\boldsymbol{\theta}, \boldsymbol{\vartheta}; \tau, q)$. A key factor in the consideration is the fact that it is necessary to compute $L(\boldsymbol{\theta}, \boldsymbol{\vartheta}; \tau, q)$ repeatedly at different parameter values over a distributed data system that is likely to encounter network latency and some individual data site failures.

**Figure 1: The Schema of the Proposed Federated Credit Model Calibration over Multiple Privacy-Protected Distributed Data Sites along with a User-Support Function**




AIDF = Asian Institute of Digital Finance.

## 2.2 Sequential Monte Carlo Optimization

We rely on the density-tempered SMC technique to perform robust federated optimization over the distributed data sites. Density-tempered SMC is a category of sampling techniques that Del Moral, Doucet, and Jasra (2006), Duan and Fulop (2015), and Duan, Fulop, and Hsieh (2020), among others, advanced. In a nutshell, optimization becomes a sampling problem in which the objective function converts into a density function just short of the norming constant.

It should be clear that $\exp\,[L(\boldsymbol{\theta},\boldsymbol{\vartheta};\tau,q)]$ is always a positive function and its maximizer is exactly the same as the maximizer of $L(\boldsymbol{\theta},\boldsymbol{\vartheta};\tau,q)$. Moreover, $\exp\,[L(\boldsymbol{\theta},\boldsymbol{\vartheta};\tau,q)]$ becomes a density function over $(\boldsymbol{\theta},\boldsymbol{\vartheta})$ if we divide it by the norming constant so that it can be integrated to 1. Although this norming constant is unknown and can be highly complex, importance sampling is a way to bypass the need to know it. The density-tempered SMC can be viewed as a sequential way of reliably conducting importance sampling over multiple steps.

Operationally, SMC runs on a sample of $(\boldsymbol{\theta}, \boldsymbol{\vartheta})$, say, 1000 particles. This SMC sample empirically represents the target density function, that is, $\exp[L(\boldsymbol{\theta}, \boldsymbol{\vartheta}; \tau, q)]$. Sequentially updating the SMC sample aims to improve the quality of representation. At the end of the self-adaptive SMC run, the particle in the final sample that yields the highest value of $\exp[L(\boldsymbol{\theta}, \boldsymbol{\vartheta}; \tau, q)]$ is the Monte Carlo solution to the original optimization problem.

Under the shared-data structure, one may not be able to compute $\exp[L(\boldsymbol{\theta}, \boldsymbol{\vartheta}; \tau, q)]$ successfully due to a failure of some data sites to submit their computed results in time for aggregation. Updating the SMC sample of parameters will not be possible without introducing approximation of the missing components. Since the previous round generated the SMC samples corresponding to different data sites, they serve as the natural base on which to perform approximations if necessary.

To make the matter concrete, we use an approximated value, $\hat{L}_m(\boldsymbol{\theta}, \boldsymbol{\vartheta}; \tau, q)$, to replace $L_m(\boldsymbol{\theta}, \boldsymbol{\vartheta}; \tau, q)$ if member *m* fails to deliver its computed result in time for the next round of parameter updating. Many approximation tools are available when a sample of 1,000 particles is in place, for example the use of the Nadaraya–Watson kernel regression[4] to link $L_m(\boldsymbol{\theta}, \boldsymbol{\vartheta}; \tau, q)$ to $(\boldsymbol{\theta}, \boldsymbol{\vartheta})$. Because the parameter may be high dimensional, the initial approximation quality is likely to be poor. As the SMC run progresses, the quality naturally improves. As we mentioned in the introduction, software known as *i*CASS implemented this new federated model calibration.

# 3. ALTERNATIVE DATA

There is a common belief these days that artificial intelligence knows us better than we know ourselves. Digital footprints open new ways for lending institutions to assess the credit quality of MSME borrowers. Utility usage, conventional media coverage, social media chatter, mobile GPS locations, and public records are some examples of alternative data. Harvesting such information solely for the purpose of discriminately pricing borrowers to enhance a lender's return could put MSME borrowers in an even more disadvantaged position.

When many alternative data become available, naively incorporating them into a shared system will become increasingly difficult for three reasons. First, some lending institutions may have the facilities/resources to gather alternative data informative of credit risk, but others may not. Second, individual lenders may place high value on such data and view them as a way of gaining a competitive edge over others. Finally, the creation of these alternative data is likely to have lacked suitable homogeneity in the variable definition.

It is therefore necessary to modify the credit model for the shared system to accommodate individualities. Thus, we can break up the model parameters into those conventional variables that are common to all lending institutions and those alternative data that are specific to an institution.

Returning to the notation that we introduced earlier, there is a need to partition a borrower's attributes into conventional data, $\boldsymbol{X}_{i,t}^{c}$, and alternative data, $\boldsymbol{X}_{i,t}^{a}$. Hence, we can rewrite the PD model specifically to accommodate alternative data individually

[4] Nadaraya (1964) and Watson (1964) produced this well-known kernel regression technique independently in the same year.

for member $m$; that is, $\mathrm{Prob}_t\left(Y^{(i)}_{t+\tau,t+\tau+q}=1\right)=f(\boldsymbol{X}^c_{i,t},\boldsymbol{X}^a_{i,t};\tau,q,\boldsymbol{\theta}^c,\boldsymbol{\theta}^a_m)$ and $\mathrm{Prob}_t\left(Y^{(i)}_{t+\tau,t+\tau+q}=2\right)=g(\boldsymbol{X}^c_{i,t},\boldsymbol{X}^a_{i,t};\tau,q,\boldsymbol{\vartheta}^c,\boldsymbol{\vartheta}^a_m)$. It is evident that it is possible to modify the federated optimization that we discussed in the preceding section slightly to accommodate institution-specific alternative data; that is,

$$\begin{aligned}
&L_m(\boldsymbol{\theta}^c,\boldsymbol{\theta}^a_m,\boldsymbol{\vartheta}^c,\boldsymbol{\vartheta}^a_m;\tau,q)\\
&\qquad=\sum_{t\in\boldsymbol{T}}\sum_{i=1}^{N_m}\Big(1_{\left\{Y^{(i)}_{t+\tau,t+\tau+q}=1\right\}}\ln\left[f\left(\boldsymbol{X}^c_{i,t},\boldsymbol{X}^a_{i,t};\tau,q,\boldsymbol{\theta}^c,\boldsymbol{\theta}^a_m\right)\right]\\
&\qquad+1_{\left\{Y^{(i)}_{t+\tau,t+\tau+q}=2\right\}}\ln\left[g(\boldsymbol{X}^c_{i,t},\boldsymbol{X}^a_{i,t};\tau,q,\boldsymbol{\vartheta}^c,\boldsymbol{\vartheta}^a_m)\ \right]\\
&\qquad+1_{\left\{Y^{(i)}_{t+\tau,t+\tau+q}=0\right\}}\ln\Big[1-f(\boldsymbol{X}^c_{i,t},\boldsymbol{X}^a_{i,t};\tau,q,\boldsymbol{\theta}^c,\boldsymbol{\theta}^a_m)\\
&\qquad-g(\boldsymbol{X}^c_{i,t},\boldsymbol{X}^a_{i,t};\tau,q,\boldsymbol{\vartheta}^c,\boldsymbol{\vartheta}^a_m)\Big]\Big)
\end{aligned}$$

Again, adding the log-likelihood values from all $K$ member institutions gives rise to the overall target function value at parameter values $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$, which is

$$L(\boldsymbol{\theta}^c,\boldsymbol{\vartheta}^c,\boldsymbol{\theta}^a_1,\boldsymbol{\vartheta}^a_1,\boldsymbol{\theta}^a_2,\boldsymbol{\vartheta}^a_2,\cdots,\boldsymbol{\theta}^a_K,\boldsymbol{\vartheta}^a_K;\tau,q)=\sum_{m=1}^{K}L_m(\boldsymbol{\theta}^c,\boldsymbol{\theta}^a_m,\boldsymbol{\vartheta}^c,\boldsymbol{\vartheta}^a_m;\tau,q)$$

It is notable that the above expression runs through the index for all member institutions, but not all institutions need to have alternative data because it is easy to switch off alternative data for a member by setting the corresponding parameter values to zero.

In summary, when a member institution has sufficient credit events to support the introduction of its member-specific alternative data, the calibrated credit model can benefit from the sharing of the conventional credit data while retaining its competitive advantage of utilizing the alternative data.

# 4. THE IMPACT OF THE COVID-19 PANDEMIC ON MSME DEFAULTS

To emulate the real-world MSME lending situation, we conceive four hypothetical financial institutions, which we refer to as Banks A to D. Each operates in three of the six ASEAN countries, as Table 1 shows. We assign the three countries randomly. All the MSMEs in the NUS-CRI database appearing in the sample period from January 1996 to May 2021 inclusive in the six ASEAN countries, 2,856 in total, enter the common pool for sampling.[5] We assign each MSME in the pool randomly without replacement to one of the four banks operating in that country until we have exhausted all 2,856 MSMEs. Table 2 provides some summary statistics on the emulated data sample.

[5] The exchange-listed MSMEs in the NUS-CRI database naturally tilt toward relatively larger firms. Micro enterprises are clearly absent from this database. The SME definition varies across jurisdictions. The adopted upper threshold is based on the annual revenue that each authority has defined: IDR50 billion (Indonesia), MYR50 million (Malaysia), THB500 million (Thailand), SGD100 million (Singapore), PHP100 million (Philippines), and VND300 billion (Viet Nam).

**Table 2: Summary Statistics on the Four Hypothetical Banks that Lend to Real Exchange-Listed SMEs in Six ASEAN Countries**

| | Bank A | Bank B | Bank C | Bank D | Total |
|---|---|---|---|---|---|
| Countries | Indonesia, Thailand, Viet Nam | Malaysia, Singapore, Viet Nam | Philippines, Singapore, Thailand | Indonesia, Malaysia, Philippines | 6 ASEAN countries |
| # SME borrowers | 538 | 1,062 | 690 | 566 | 2,856 |
| Time period | Jan 1996–May 2021 | Jan 1996–May 2021 | Jan 1996–May 2021 | Jan 1996–May 2021 | Jan 1996–May 2021 |
| # defaults | 14 | 60 | 37 | 51 | 162 |
| # other exits | 117 | 230 | 140 | 113 | 600 |
| # defaults—COVID period | 1 | 4 | 2 | 0 | 7 |
| # other exits—COVID period | 11 | 14 | 7 | 4 | 36 |
| # firm-month observations | 27,244 | 72,099 | 49,876 | 34,249 | 183,468 |

Note. The definition of the COVID period is January 2020 to the end of the sample.

It is clear from Table 2 that any bank alone will fall short of the number of defaults necessary to estimate a default prediction model that has many parameters. Needless to say, pooling all four banks together will still be insufficient to identify the impact of the COVID pandemic on each of the prediction variables. Some simplification in the model specification is necessary, and pooling data is the only practical way to conduct such an analysis.

For simplicity, we deploy the logistic regression to model the MSME forward term structure of 1-month PDs from the current time onward: that is, the 1-month PD immediately ahead all the way to the 1-month PD 11 months ahead. We stop at 12 forward months because the COVID period is not long enough to enable a meaningful analysis for longer terms. Using the notation that we described earlier, we treat $f(\boldsymbol{X}_{i,t};\tau,q,\boldsymbol{\theta})$ as a logistic function at a different month-end, *t*, where the prediction duration $q$ is always set to 1 month and the forward starting time $\tau$ varies from 0 to 11 months. Similarly, $g(\boldsymbol{X}_{i,t};\tau,q,\boldsymbol{\vartheta})$ is a logistic function for modeling other exits.

With a limited number of default events in the COVID period, we single out its potential impact on $f(\boldsymbol{X}_{i,t};\tau,q,\boldsymbol{\theta})$ through the intercept and two prediction variables in $\boldsymbol{X}_{i,t}$ because the design of the COVID relief measures aimed to raise liquidity [6] and indebtedness concurrently (i.e., lowering distance-to-default (DTD)[7]). Furthermore, the relief measures involved default suspension. Hence, one can expect some changes to the parameter values in the COVID period, that is, how defaults react to, say, liquidity. We measure these two variables in terms of the level, that is, their 12-month moving averages. Such incorporation of the COVID dummy variable adds four parameters (intercept, DTD, and two liquidity measures, respectively for financial and non-financial SMEs) to each of the 12 PD forward functions.

To avoid introducing too many parameters into the system, we follow the NUS-CRI practice of imposing the Nelson–Siegel function on the forward starting time to smooth the parameters over 12 forward periods on all the variables, including the COVID

6 We measure liquidity as the ratio of cash and cash equivalent over total assets for financial SMEs and the ratio of current assets over current liabilities for non-financial SMEs. This follows the implementation that NUS-CRI Staff (2021) described.

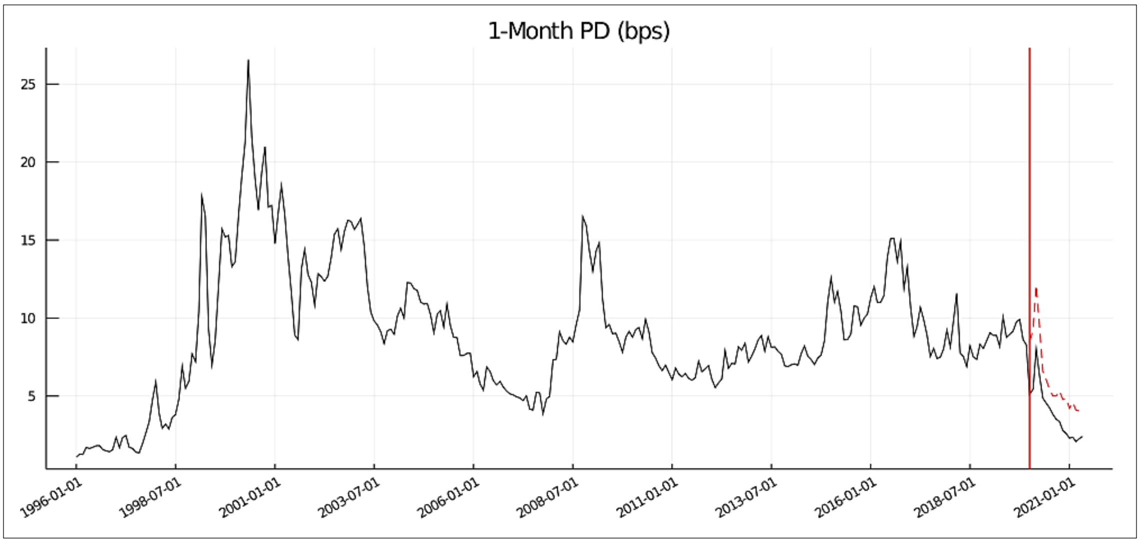
7 We can interpret DTD as the asset volatility adjusted leverage, that is, the ratio of book value of debt over the market value of assets, which we further adjust using asset volatility. The theoretical model of Merton (1974) derived DTD, and we implemented it empirically in accordance with NUS-CRI Staff (2021).

dummy variable.[8] As a result, the simplified specification only adds 12, instead of 48, parameters to the system.

Ideally, we should treat the COVID dummy variable as country specific, but insufficient data prevent its adoption. Notwithstanding potentially differential intensities across these six ASEAN countries, there is little doubt that their special policy measures all work in the same direction. We do not subject all the other variables in $f(\boldsymbol{X}_{i,t}; \tau, q, \boldsymbol{\theta})$ to the COVID dummy variable.[9] Neither do we introduce the COVID dummy variable into the probability of other exit function, $g(\boldsymbol{X}_{i,t}; \tau, q, \boldsymbol{\vartheta})$.

Figure 2a depicts the MSME portfolio's averages of 1-month PDs at different month-ends over the sample period for the six ASEAN countries, that is, the four bank portfolios pooled together, whereas Figure 2b shows the same overall portfolio's averages of 1-year PDs. We deduce each of the 1-year PDs for an MSME at a month-end with the survival-default formula, using that MSME's 12 estimated forward PDs and POEs at the time. To these figures, we also add the counterfactual PDs as if the COVID pandemic did not affect the parameter values.

**Figure 2a: The 1-Month PDs (Portfolio Average) before and during the COVID Period for the Four-Bank Aggregate Portfolio in the Six ASEAN Countries**

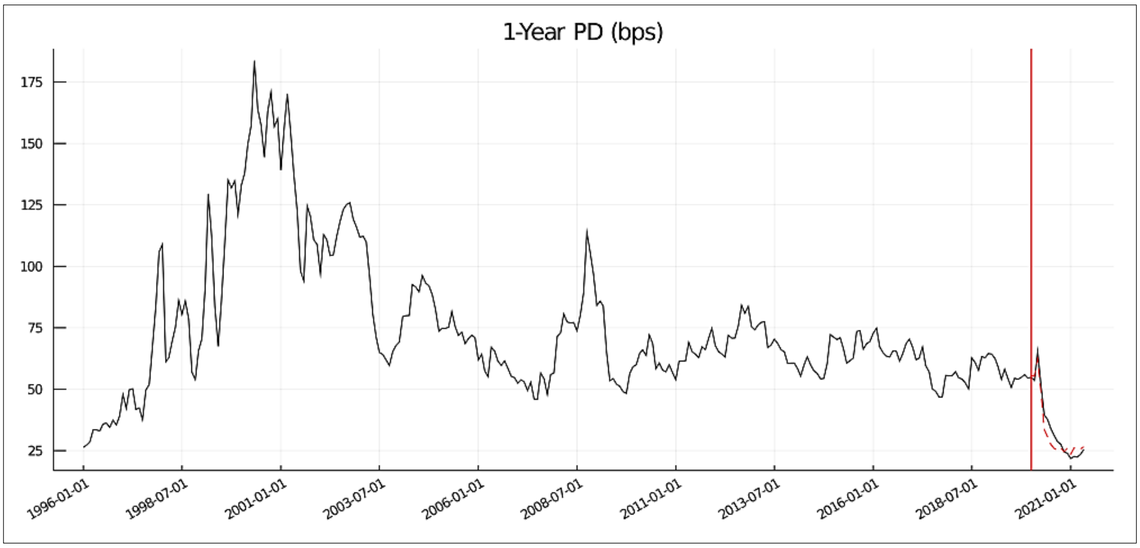



Notes: The graph measures the PDs on the vertical axis in basis points. The dashed line depicts the counterfactual PDs in the COVID period by switching off the COVID dummy variable.

[8] Please refer to NUS-CRI Staff (2021) for a discussion on the use of the Nelson–Siegel smoothing function. The NUS-CRI implementation classifies variables into two categories—vanishing vs non-vanishing types. It gives the former type three parameters to characterize the whole forward curve because its impact eventually decays to zero. For the latter, it uses four parameters because the impact does not converge to zero. The COVID dummy variable clearly falls into the vanishing type because the current COVID status should not affect a distant forward period.

[9] Other firm-specific prediction variables are the net income over total assets, relative firm size, idiosyncratic equity volatility, and market-to-book ratio, whereas the common risk drivers are the country-specific interest rate, stock market return, and aggregate DTD.

**Figure 2b: The 1-year PDs (Portfolio Average) before and during the COVID Period for the Four-Bank Aggregate Portfolio in the Six ASEAN Countries**




Notes: The vertical axis measures the PDs in basis points. The dashed line depicts the counterfactual PDs in the COVID period by switching off the COVID dummy variable.

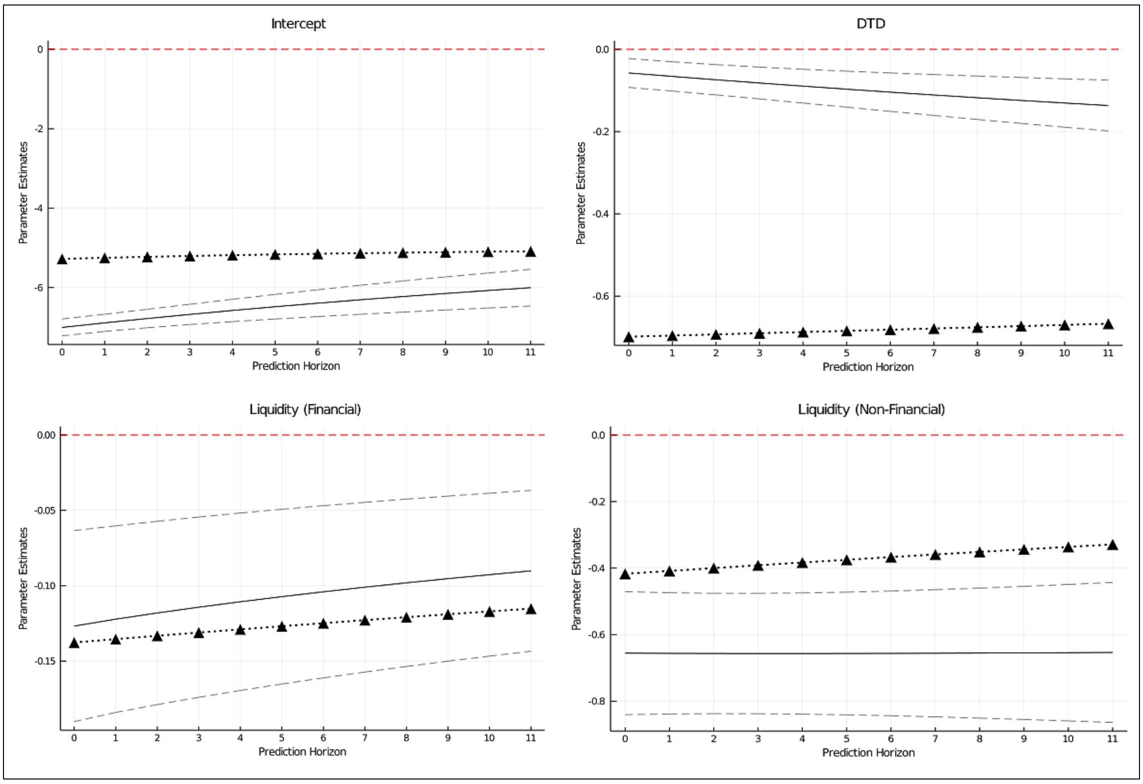
At the first glance, these figures reveal a seemingly counterintuitive conclusion that the COVID pandemic has lowered the MSME credit risk, from either a short-run or a longer-term perspective, that is, 1-month vs 1-year PD. Factoring in the special COVID relief measures that the respective national authorities adopted, the results suggest the impact of the policy measures and become quite understandable. Instead of a comparison with the PDs prior to the onset of the COVID pandemic, the counterfactual PDs offer a different angle. They are evidently higher for the short term, signaling the realization of the intended policy outcomes, but the structural impact is likely to fade.

To understand the impact of the special COVID relief measures, we need to recognize two channels. We can understand the direct channel by considering, for example, concessional loans, which directly change the financial variables of MSMEs even if the structural relationship between the default and the financial variables remains unaltered. The indirect channel pertains to the change in the structural relationship via, in a model sense, the coefficients defining the relationship. Mandated default suspension and/or extended grace periods can have an impact on defaults even if financial variables remain affected by the relief measures.

Pooling the data of the four banks leads to the finding that the COVID pandemic has influenced the model parameter values. Due to too many model parameters in the system of 24 forward functions—12 for PDs and 12 for POEs—we present the estimation results in Figure 3 with four subplots on four selected parameters.

Figure 3 succinctly reveals the impact of the COVID pandemic by plotting the four sets of 12 coefficients affected by the COVID dummy variable. We plot their corresponding 95% confidence intervals to assess their statistical significance quickly. Once again, these four sets correspond to the intercept term, DTD, and liquidity (separately for financial and non-financial SMEs) in the 12 forward PD functions. The horizon axis shows the forward starting time running from 0 to 11, corresponding to 12 forward periods, each having a 1-month duration.

**Figure 3: The Impact of the COVID Dummy Variable on the Intercept and the Coefficients of DTD and Two Liquidity Measures (Financial and Non-financial Firms) in the 12 Forward PD Functions along with their Corresponding 95% Confidence Intervals**




Notes. The horizon axis is the forward starting time running from 0 to 11 corresponding to 12 forward periods of 1-month duration. A triangle marks the pre-COVID coefficients.

We add the 12 pre-COVID parameter values (without plotting their corresponding 95% confidence intervals) to each of the above four subplots in Figure 3. Evidently, the COVID pandemic has a severe impact on the parameter values, except for liquidity for financial firms. Their directions are sensible considering the knowledge of these COVID policy measures. The lowering of SMEs' default risks, other things being equal, is apparent in the lower intercepts. They also increase the sensitivity to liquidity (non-financial), that is, becoming more negative, and lower the sensitivity to DTD, that is, becoming less negative.

In summary, the special COVID measures in the six ASEAN countries have operated in a way that is consistent with the policy intention. Note that the SMEs in the sample exhaust all real exchange-listed SMEs in the six ASEAN countries even though we emulate the four banks with the purpose of illustrating the power of data sharing via modern digital technology. Our conjecture is that the above conclusion would have been stronger if we could have tapped into the closely guarded real banks' SME portfolios comprising many smaller non-exchange-listed SMEs.

# 5. ESTABLISHING AN MSME DATA-SHARING CONSORTIUM

Data-sharing technology, such as *i*CASS, enables MSME lenders to form, for example, a consortium. The design of the consortium can benefit its members in two ways—an improved credit model in prediction quality and a shared support service infrastructure to lower the operating costs. The natural consortium members include conventional/ digital banks, finance companies, P2P lending platforms, and any FinTech companies possessing credit-related data.

Through a consortium, it is possible to treat the improved credit model as a common good. Members still compete with one another via their differences in risk appetite, services, and operational efficiency. In short, the consortium can become a realization in the spirit of *coopetition*.

What incentivizes lending institutions to join a consortium? How can they prevent free riders? Addressing these issues rests with a contract/institution design that extends beyond the technology of data sharing. Here we offer a few thoughts.

As the earlier discussion suggested, the data-sharing platform allows the construction of credit models tailored to individual members if they possess unique alternative data. A lending institution alone may not have sufficient data instances to identify its own credit model when the data features that the prediction uses have expanded to cover alternative data. This member-specific potential may prove to be attractive to some lending institutions. The shared data help to pin down the parameters in the data fields common to all members, which in turn frees a member's own data to work on nailing down those parameters associated with the alternative data. Therefore, the consortium design should encourage members to leverage the shared data and support the infrastructure in deriving member-specific credit models. This benefit may prove to be a strong enough motivator for some lending institutions to join the consortium.

We envision a successful consortium as observing a few guiding principles and key components to address the incentive and other practical issues. Naturally, we expect the variants and refinements to reflect different circumstances and needs. The five general points are as follows:

1. Set up a Governing Board to determine policies and a Secretariat to support the operation of a consortium.
2. The Governing Board determines the formulas for membership fees and query charges to support the operation.
3. The Secretariat maintains the Model Calibration Central and the User Support Central. The former executes the federated learning (initial and subsequent recalibration) of the credit models. The latter facilitates the members' easy utilization of the calibrated credit model through shared implementation of the calibrated model in response to PD queries based on a member's submission of obligor attributes.
4. The members contribute credit data on the predefined variables that the Governing Board has agreed and commit, say, to updating the data quarterly. The contributed data remain at members' own data sites under total privacy protection. The members consent to a third-party audit to ensure the integrity of the contributed data.

5. A consortium may address free ridership by adopting tiered membership to reflect different levels of data contribution. The privilege that the consortium grants to the highest-tiered members can, for example, be exclusive access to the construction of a member-specific credit model that combines the shared data with the member's own alternative data. A more favorable query fee schedule may also serve as a privilege.

# 6. CONCLUDING REMARKS

The COVID-19 pandemic has exacted a toll on many MSMEs worldwide. However painful this might have been, the difficulties arising from the pandemic are only transitory in nature. Those MSMEs that survive the pandemic will continue to face financing challenges with structural roots in the informational asymmetry between themselves and the lending institutions.

Assisting MSMEs with subsidized financing rates and/or risk–share losses, as typical government programs reflect, will not fundamentally alter the pooling equilibrium resulting from the lack of incentives for lending institutions to invest in costly information acquisition on small loans. This paper advocates building a new-style infrastructure for sharing credit information using digital technology for which the small setup and running costs can in a fundamental way help lending institutions to level their credit information acquisition costs on MSMEs vis-à-vis larger corporations.

With this credit information infrastructure serving as a *common good*, lending institutions can still compete by offering different loan rates and banking services and/or by specializing in certain market niches. In our view, this *coopetition* model provides a realistic and productive way to achieve fairer financing of MSMEs.

## REFERENCES

Del Moral, P., A. Doucet, and A. Jasra. 2006. "Sequential Monte Carlo Samplers." *Journal of the Royal Statistical Society: Series B* 68 (3): 411–36.

DiCaprio, A., S. Beck, and J. C. Daquis. 2014. *ADB Trade Finance Gap, Growth, and Jobs Survey*. Asian Development Bank Briefs No. 25, Asian Development Bank, the Philippines.

Duan, J. C., and A. Fulop. 2015. "Density-Tempered Marginalized Sequential Monte Carlo Samplers." *Journal of Business and Economic Statistics* 33 (2): 192–202.

Duan, J. C., A. Fulop, and Y. W. Hsieh. 2020. "Data-Cloning SMC2: A Global Optimizer for Maximum Likelihood Estimation of Latent Variable Models." *Computational Statistics and Data Analysis* 143 (March): 106841.

Duan, J. C., J. Sun, and T. Wang. 2012. "Multiperiod Corporate Default Prediction—A Forward Intensity Approach." *Journal of Econometrics* 170 (1): 191–209.

Konečný, J., B. McMahan, and D. Ramage. 2015. "Federated Optimization: Distributed Optimization beyond the Datacenter." arXiv preprint, arXiv:1511.03575, arxiv.org.

Merton, R. C. 1974. "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates." *Journal of Finance* 29: 449–70.

Nadaraya, E. A. 1964. "On Estimating Regression." *Theory of Probability and Its Applications* 9 (1): 141–2.

NUS-CRI Staff. 2021. *NUS-RMI Credit Research Initiative Technical Report Version: 2021 Update 1*. The Credit Research Initiative at the National University of Singapore. http://nuscri.org.

Watson, G. S. 1964. "Smooth Regression Analysis." *Sankhyā: The Indian Journal of Statistics, Series A* 26 (4): 359–72.