European Journal of Operational Research 324 (2025) 590–615

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Monitoring bank risk around the world using unsupervised learning

Mathieu Mercadier [a,*], Amine Tarazi [b,c], Paul Armand [d], Jean-Pierre Lardy [e]

[a] DCU Business School, Dublin City University, Dublin, Ireland
[b] Université de Limoges, LAPE, 5 rue Félix Eboué, 87031 Limoges Cedex, France
[c] Institut Universitaire de France (IUF), 1 rue Descartes, 75231 PARIS Cedex 05, France
[d] Université de Limoges, XLIM UMR CNRS 7252, 123 avenue Albert Thomas, 87060 Limoges, France
[e] JPLC SASU, 54 avenue de la Révolution, 87060 Limoges, France

ARTICLE INFO

JEL classification:
G01
G20
G28
C44

Keywords:
Risk analysis
Decision support tool
Banking risks
Systemic risk
K-means

ABSTRACT

This paper provides a transparent and dynamic decision support tool that ranks clusters of listed banks worldwide by riskiness. It is designed to be flexible in updating and editing the values and quantities of banks, indicators, and clusters. For constructing this tool, a large set of stand-alone and systemic risk indicators are computed and reduced to fewer representative factors. These factors are set as features for an adjusted version of a nested k-means algorithm that handles missing data. This algorithm gathers banks per clusters of riskiness and ranks them. The results of the individual banks' multidimensional clustering are also aggregable per country and region, enabling the identification of areas of fragility. Empirically, we rank five clusters of 256 listed banks and compute 72 indicators, which are reduced to 12 components based on 10 main factors, over the 2004–2024 period. The findings emphasize the importance of giving special consideration to the ambiguous impact of banks' size on systemic risk measures.

1. Introduction

The 2007–08 global financial crisis (GFC) has drawn attention from academics, regulators and politicians regarding banks' supervision and regulation. Over the last decade, Basel III was brought as a response to the GFC (Penikas, 2015; Walter, 2019), and numerous governments had to rescue banks with dedicated programs (e.g., the United States' Troubled Asset Relief Program (TARP) or the European Financial Stability Facility), to avoid disastrous implications on the whole economy. This led researchers to establish new indicators quantifying systemic risk.

Generally, forecasting bankruptcies relied on econometric methods that require complex assumptions. However, the recent advances in computing capacities have propelled machine learning to the forefront of optimization techniques. Supervised learning algorithms can be used to forecast bankruptcies but training them is contingent upon the availability of reliable labels. However, one cannot be certain that banks that received rescue programs (e.g., TARP) would have definitely defaulted without them (Calomiris and Khan, 2015).

Consequently, we propose an adjusted version of an unsupervised learning algorithm, the k-means, to build a dynamic decision support tool for monitoring banks' riskiness. It aims at ranking clusters of listed banks based on diverse risk indicators. The algorithm executes automatically the four following steps in sequence within a single run: i) extraction and preprocessing of raw financial data; ii) risk indicators' computation from these data; iii) reduction of the indicators into representative components based on rational factors; iv) running the adjusted k-means on the factors. This tool offers the possibility to continuously update the values and quantities of indicators, factors, and clusters. It can also present the resulting information on a dedicated website, allowing for the aggregation of the clustering per country and region, which helps with identifying fragile areas and anticipating potential future distress. Furthermore, we analyze the outputs of the tool from an economics perspective.

The strength of this tool lies in its fundamental function of creating and ranking bank clusters. In addition, it provides many risk indicators which are reported on a weekly basis and computed utilizing both balance sheet and financial market data. Stand-alone indicators such as volatility, VaR, expected shortfall, Z-scores and credit spread approximations are computed to assess banks' health. Some financial ratios are

* Corresponding author.
E-mail addresses: mathieu.mercadier@dcu.ie (M. Mercadier), amine.tarazi@unilim.fr (A. Tarazi), paul.armand@unilim.fr (P. Armand), jp.lardy@jplc.fr (J.-P. Lardy).

https://doi.org/10.1016/j.ejor.2025.01.036
Received 12 July 2023; Accepted 25 January 2025
Available online 28 January 2025

also integrated to our tool. Along with correlations and betas, other indicators are chosen to assess banks' size, liquidity, and regulatory requirements. Systemic indicators developed after the GFC are also calculated such as the MES, SRISK, Tail Beta and ΔCoVaR. Once calculated, they are cleaned and summarized into representative components, which are based on rational factors, using principal component analysis.

Empirically, we rank 5 clusters containing 256 listed banks from 43 countries during the 2004–2024 period. To do so, a combination of financial data is transformed into 72 indicators quantifying individual and systemic risks. These indicators are further condensed into 10 economically rational factors, represented by 12 principal components, which are used as input features for our adjusted nested k-means clustering algorithm. We also justify our choice of the clustering algorithm and the selection of hyperparameters.

From an economic standpoint, the analysis of the centroids emphasizes the significance of considering bank size. In line with Varotto and Zhao (2018)'s study that deals with the overshadowing effect of bank size on systemic indicators, our tool distinguishes two circumstances. First, the safety of small banks tends to be exaggerated, according to systematic metrics, namely MES, ΔCoVaR and Tail Beta. Second, SRISK is more inclined to overstate big banks as risky. Our findings also indicate that the Global Financial Crisis (GFC) had a more severe impact on the banking industry compared to the COVID-19 pandemic. Additionally, we investigate the probability of banks transitioning to other clusters using a transition matrix and evaluate the generalizability of the tool.

This study aims to create a transparent, self-contained, and easily editable tool to rank listed banks worldwide. This dynamic tool can be readily adapted for specific research studies, with the inclusion or exclusion of banks or indicators. It has considerable value in assisting academics and practitioners in comparing the riskiness of listed banks.

The rest of the paper is laid out as follows. Section 2 provides a brief interdisciplinary literature review. Section 3 introduces the automatic handling of the raw data, in terms of cleaning and removal of outliers. Section 4 presents the different risk indicators used to assess specific risks and their corresponding factors. Section 5 deals with the dimensionality reduction and the adjusted clustering algorithm. Finally, Section 6 analyzes the clusters' outputs, their behavior, and the tool's generalizability.

2. Related literature

Established literature in the banking sector usually uses clustering algorithms in studying various banks' similarities. Moldovan and Mutu (2015) assess the relationship between banks' default probability and their risk taking incentives for listed and unlisted European banks, using both partitional and hierarchical clustering. Knotek (2014) studies European banking sector similarities with the Ward method. This resembles to the study of Soerensen and Gutiérrez (2006) which detects some basic patterns and trends in the Eurozone banking sector. The grouping of the banking sectors based on the banking ratios shows that the EU countries in similar geographic area and with higher economic partnership tend to group in similar clusters (Ercan and Sayaseng, 2016). Dardac and Giba (2011) create groups of financial crises to highlight the most efficient politics to manage these crises. Furthermore, Dardac and Boitan (2009) gather Romanian credit institutions in function of their risk profile and profitability with an agglomerative hierarchical clustering algorithm. Cernohorska (2017) highlights the similarities between Czech and Slovak banking sectors with hierarchical agglomerative clustering. Zarandi et al. (2014) use a hybrid k-means and Grey relational analysis to measure the relative efficiency of various Iranian banks. Applying different hierarchical and non-hierarchical clustering approaches, Lagasio and Quaranta (2022) identify that a bank business model is based on different endogenous factors and can be linked to exogenous factors, including financial crisis.

The aforementioned studies mostly use hierarchical clustering, that does not require prior knowledge of the number of clusters and produces a visualization of the hierarchy of clusters. But these computationally complex algorithms are generally inefficient for large samples and are sensitive to noise. Partitional algorithms, such as k-means, are faster to partition the best k-clusters of the observations (Hartigan, 1975). Although they require a better understanding of the dataset (Di Lascio et al., 2018), they offer the possibility to set the number of clusters allowing to keep control over the final results. Therefore, we employ the k-means algorithm in this study.

The reviewed studies mostly directly deal with standard financial ratios as input, such as return on assets, equity to assets ratios, etc. Our indicators are summarized into factors to reduce noise and speed up the learning process. Giglio et al. (2016) build systemic indexes from estimators' dimension reduction to predict macroeconomic shocks. As reported by Xu et al. (2015), the k-means is more efficient in low-dimensional space. In addition to filtering noise and reducing computation time, the risk of converging toward local optima is reduced, in line with various recent papers from broad research domains (Danley, 2019; Zhu et al., 2019).

More broadly, our study contributes to the literature on bank failure prediction and risk assessment using machine learning approaches,[1] positioning itself alongside notable works such as Le and Viviani (2018); Kristóf and Virág (2022), and Tuan et al. (2023). While these studies have applied various machine learning techniques to predict bank failures for specific countries, including clustering methods combined with neural networks using simulation data (Negnevitsky, 2017), our approach diverges by emphasizing a transparent and dynamic decision support tool that ranks clusters of listed banks worldwide by riskiness. Unlike studies that use multicriteria analysis to combine multiple indicators into a single risk score (Doumpos et al., 2017; Tsagkarakis et al., 2021), our tool leverages a nested k-means algorithm adjusted to handle missing data, which clusters banks by riskiness and ranks them accordingly. Additionally, our methodology includes the computation and reduction of a large set of stand-alone and systemic risk indicators to fewer representative factors, providing a multidimensional risk assessment that is both flexible and easily editable. This tool not only ranks individual banks but also aggregates results by country and region, identifying areas of fragility. By empirically ranking five clusters containing 256 listed banks and reducing 72 indicators to 10 main factors over the 2004–2024 period, this study offers a novel and comprehensive framework that enhances adaptability and transparency in bank risk assessment, distinguishing itself from prior works in the field.

Individual bank risks are measured by financial ratios, raw balance sheet data, betas, correlations, and other aggregated indicators. For instance, the Z-score (Hannan and Hanweck, 1988; Boyd et al., 1993) and its market version the MZ-score (Lepetit et al., 2008; Saghi-Zedek and Tarazi, 2015) assess bank solvency. Credit spreads are approximated with the CreditGrades model (Finger et al., 2002) and the E2C formula (Mercadier and Lardy, 2019). Systemic risk indicators are the Marginal Expected Shortfall, that was introduced by Brownlees and Engle (2016), its long-term versions the LRMES and the SRISK (Acharya et al., 2012; Brownlees and Engle, 2016), the Delta Conditional Value-at-Risk (Adrian and Brunnermeier, 2016) and the Tail Beta (Straetmans et al., 2008; deJonghe, 2010). Besides being conventionally computed, the Value-at-Risk, the expected shortfall and the systemic indicators integrating them are computed using weighted historical methods (Boudoukh et al., 1998; Hull and White, 1998) to obtain more reactive metrics.

Large banks are related to significantly higher systemic risk (deJonghe, 2010; Banbula and Iwanicz-Drozdowska, 2016; Kleinow et al., 2017). But the role of the size as a major determinant of the systemic risk is still debated (Zhang et al., 2015; Laeven et al., 2016; Sedunov, 2016; Hué et al., 2019), notably regarding its tendency to

[1] For more information on this topic, see Doumpos et al. (2023).

overshadow other factors contributing to banks' systemic importance (López-Espinosa et al., 2012, 2013; Weiß et al., 2014b; Varotto and Zhao, 2018; Elyasiani and Jia, 2019).

3. Data

Requiring a reactive and well-resourced database, the data are extracted from Bloomberg L.P. via the dedicated API. Then, the tool is written in Python 3.6 and calls our C++ adjusted version of the k-means.

3.1. Data overview

The dataset used in this paper provides a comprehensive representation of banking systems across various world regions from January 2004 to May 2024.

A constraint of our dynamic tool is the increased computation time as the dataset grows over time. Therefore, following Hanson et al. (2024), we only include in the dataset the largest global and regional banks with total assets exceeding USD 100 billion. More specifically, we retain all available banks that had total assets exceeding USD 100 billion at the end of any year within the studied period. This dataset covers a total of 256 banks, with an average of 227 (±12) banks per year. Over the period (with data from the Financial Stability Board (2023) only available until 2023), our dataset represents, on average, 78.4% (±5.3%) of total bank assets. More detailed statistics are available in Appendix A (see Table A.1, Fig. A.1, Fig. A.2).

More specifically, we extract the entire universe of listed banks available in the Bloomberg database using the "Equity Screening" function. From the database, we obtain a list of 4363 banks, including active banks as well as those that have been acquired, delisted, or suspended, to mitigate survival bias. From this sample, we retain the banks with total assets exceeding USD 100 billion. To avoid duplication, we keep only one ticker when banks are listed on multiple stock exchanges, such as Nordea (Helsinki, Stockholm, and Copenhagen) and several Chinese banks (Shanghai and Hong Kong). The chosen ticker is the one with the most available information on both balance sheet and financial market data. In the end, we retain 256 listed banks across 43 countries.

3.2. Extraction methodology

Unsupervised learning algorithms are built to deal with any type of digital inputs, and many papers (Soerensen and Gutiérrez, 2006; Dardac and Boitan, 2009; Dardac and Giba, 2011; Knotek, 2014; Zarandi et al., 2014; Moldovan and Mutu, 2015; Ercan and Sayaseng, 2016) use financial report data and ratios as features. However, in this tool, we compute risk indicators requiring both accounting and market data as inputs. An obvious limitation of comparing banks based on both types of data is that it restricts our universe to listed banks. Moreover, such information is released with different frequencies – daily for market data and quarterly, semiannually or yearly for accounting data.

By construction, dealing with automatically extracted listed banks' data potentially convey some issues. For instance, dividend payments, stock splits, or equity offerings, may affect a stock's entire price and volume history when adjusted by the provider. Furthermore, to deal with mergers and acquisitions, our tool verifies and removes all data after the date at which the stocks are delisted or acquired. In addition, the data provider may sporadically change the tickers of banks, resulting in missing information and requiring manual identification of the new corresponding tickers. Thus, when the code is run, automatic audits are conducted on the data to highlight their consistencies. Moreover, the data might contain missing points or outliers. Given all these issues, a complete update of the database is necessary every time the tool is initiated.

To calculate some indicators, the countries in which the banks operate are utilized as common keys for matching them with their respective domestic indexes or for conversion to US dollars. Thus, beyond banks, national indexes and currencies data, the program extracts a few descriptive fields for all tickers, e.g., name, industrial sector, currency, status, name & ISO code of the country of domicile, frequency of data release, etc.

3.3. Dealing with asynchronicity and outliers

The input database is split in two tables (daily and monthly) that are cleaned and merged into a unique weekly table. To handle the daily data, we only require outliers to be removed. However, as our tool is dynamic, we have developed an outlier detection algorithm, which is explained below. This algorithm is designed to be simple and global, capable of handling various frequencies, and operates "on-the-fly" irrespective of the input dataset. Moreover, given Bloomberg L.P.'s specialization in market data, outliers are highly uncommon within this context. In the daily table, 13.35% of the values were missing, and 0.27% of the values were identified as outliers. After the removal of the outliers, the percentage of missing values increased to 13.59%.

However, erroneous data is more frequent in financial accounts. The countries in our sample release similar financial information at different dates and frequencies. Another issue is that accounting data are released at the end of the month that can fall on weekends – only keeping business days might lead to omit some values. Studying the data, we noticed additional inconsistencies between quarterly, semiannually, and yearly data. First, it is important to note that the reference date for a given bank may differ between files for a specific month and year. This issue can be readily resolved by enforcing the reference dates as the calendar's end of the month. However, the situation becomes more complex when the corresponding values are dissimilar for a particular date. We keep the closer value to the mean of the previous and consecutive values, as computing the average of these different values may be problematic given that at least one of them may be too extreme. In addition, these outliers can randomly be on the quarterly, semiannually, or yearly file.

In short, the outliers' removal algorithm parses each time series with two overlapped rolling windows over an automatically defined period – depending upon the number of points. This process sets values as outliers if they are beyond $\mu \pm 5\sigma$, where μ is the mean value and σ the standard deviation. In the final monthly table, 4.1% of the values were missing, 1.32% of the values (excluding the original missing values) were adjusted based on the method described above, and 0.13% of the values were identified as outliers and converted to missing values.

Additionally, we have noticed that even after passing through the aforementioned filters, some of Bloomberg's data remains odd. Notably, some values might be originally wrongly set at zero, instead of as missing value, making no economic sense. This issue is neutralized by replacing the null values by missing data for some specific economic fields, e.g., bid-ask spread, total assets, total equities, tier 1 capital ratio, etc. In the studied sample, the percentage of null values to non-missing values is 0.72%.

At this point, all banks' variables belong either to a monthly or a daily frequency table. To avoid mismatches for the indicators' computations, balance sheet datasets remain monthly and only Fridays' values are kept for daily ones. Then, to keep as much information as possible, prior month-end inputs are forward filled before weekly calculations and those expressed in national currencies are converted in US dollars.

4. Risk indicators and representative factors

4.1. Data pre-processing

It is fundamental to pre-process the data, reduce dimensions and extract hand-crafted, domain-specific features (Bengio et al., 2013). In this respect, we firstly transform the raw data in terms of risk indicators that are then summarized into fewer representative factors. As explained in the previous section, we undertake the first phase of preprocessing on the raw data by removing outliers and matching dates and currencies of market and balance sheet data. After computing risk indicators from raw data, a second phase of features engineering is applied, involving cleaning, orientation, transformation, standardization, and dimensional reduction (cf. Section 5.1). Although the k-means algorithm can directly manage the indicators, it is a useless time-consuming process as many highly correlated indicators are handled,[2] especially as the dataset grows over time. Additionally, interpreting a smaller number of representative factors is more manageable than analyzing each individual indicator. Therefore, 10 chosen factors are derived with principal component analysis from subsets of the 72 banking risk indicators presented below.[3] This dimension reduction method was selected due to its wide usage and transparent characteristics, which align with the objectives of this paper.

Linking some indicators, notably regarding systemic measures, is sensible. In fact, many systemic indicators are built in the same way, only the underlying indexes or tuning parameters differ. However, it requires a greater level of consideration for individual and size-related metrics. Initially, we group the indicators into factors based on economic rationale. These groupings are then validated ex-post through reliability analysis (cf. Section 5.1). In the end, the factors can be classified in three main categories: individual & systemic risks measures and factors encapsulating banks' size.

4.2. Individual risk indicators

4.2.1. Financial ratios and balance sheet information

The factor called "**RatioBS**" measures banks' exposures using balance sheet ratios, such as debt to total assets, deposit to total assets, total loan to total assets, total loan to deposit, non-performing assets to total assets, non-performing loans to total loans & reserve for loan loss to total loans.

The "**CreditBS**" factor assesses banks' credit worthiness based on balance sheet information, such as common equity to total assets, return on assets (ROA), price to book ratio, Texas ratio and Z-score. Including a simple leverage ratio like equity to total assets is still relevant nowadays (Durand and Le Quang, 2022). The Texas ratio was developed by Gerard Cassidy to predict banks' failure during the 1980s recession. It is computed by dividing the bank's non-performing assets by the sum of the reserve for loan loss and the total equity. The Z-score is a simple and popular risk measure of a bank's probability of insolvency. It was initially proposed by (Hannan and Hanweck, 1988; Boyd et al., 1993). This measure is still being modified, improved, and used (Lepetit and Strobel, 2013; Mare et al., 2017; Ardekani Mahdavi et al., 2020; Pamen Nyola et al., 2021). For sake of reactivity, we use the last available ROA and equity over total assets and not their rolling averages. The Z-score only requires accounting data and is a decreasing function of the probability of failure.

$$Z-score_{i,\ t} = \frac{ROA_{i,t} + \frac{total\ \ equities_{i,t}}{total\ \ assets_{i,t}}}{\sigma_{ROA_{i,t}}}$$

Where $\sigma_{\mathrm{ROA}_{\mathrm{i,t}}}$ is the ROA's standard deviation over 4 years.

We also build the "**RegCapital**" factor highlighting the main Basel's capital ratios, like the Tier 1 capital ratio, total risk-based capital ratio, also known as capital adequacy ratio (CAR), and risk-weighted-asset to total assets.

4.2.2. Credit and solvency indicators using market information

Two factors are defined to apprehend banks' riskiness using market data they respectively focus on banks' credit quality and individual market risk.

The most straightforward input to build the "**CreditMkt**" factor would be banks' CDS spreads, however, not all banks have actively traded CDS. Thus, we use two measures coming from the capital structure optional framework. At first, we compute the credit spread approximation developed by RiskMetrics Group's CreditGrades model (Finger et al., 2002), based on the probability of survival.

$$\mathbb{P}(survival)_t = \Phi\left(-\frac{A_t}{2} + \frac{log(d)}{A_t}\right) - d\cdot\Phi\left(-\frac{A_t}{2} - \frac{log(d)}{A_t}\right)$$

Where $d = \frac{S_0+\bar{L}D}{\bar{L}D}\cdot e^{\lambda^2}$, $A_t^2 = \left(\sigma_{S_0}\cdot\frac{S_0}{S_0+\bar{L}D}\right)^2\cdot t + \lambda^2$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function. Thus,

$$\mathrm{CrdGrd} = \frac{-\ln(\mathbb{P}(\mathrm{survival})_t)\cdot(1-R)}{\mathrm{t}}$$

The second credit indicator called E2C (Equity-to-credit) is a credit spread approximation derived by Mercadier and Lardy (2019), that broadly comes from the Gauss inequality applied to a Brownian motion. Its parameters are the recovery rate R, the Market-Adjusted Debt (MAD) ratio – which includes the total debt per share outstanding, the stock price and the average recovery on the debt – and the equity volatility detailed in (Finger et al., 2002).

$$\mathrm{E2C} = (1-R)\cdot\frac{4}{9}\cdot\mathrm{MAD}\cdot\sigma_{S_0}^2$$

Where $\mathrm{MAD} = \frac{\bar{L}D}{S_0+\bar{L}D}$

Except the debt which varies on lower frequencies, these hybrid indicators require market inputs available daily. Lastly, a simple market-based leverage ratio (debt to enterprise value ratio) is added to this factor.

Another market-based factor named "**IndivMktRisk**" measures banks' riskiness through nine indicators: the volatility, three VaR metrics, three Expected Shortfall metrics and the MZ-score. A responsive daily volatility metric σ_{i} of yearly rolling banks' daily stock returns is computed integrating a forgetting factor, computing the weighted sum of squared returns normalized by $1-\lambda$:

$$\sigma_{\mathrm{i,t}} = \sqrt{(1-\lambda)\cdot \mathrm{r}_{\mathrm{i,t}}^2 + \lambda\cdot\sigma_{\mathrm{i},t-1}^2}$$

Where the decay factor λ is set at 0.98 and the initial annualized $\sigma_{\mathrm{i,0}}$ at 0.29 for the banks (at 0.17 for the indexes), corresponding to the median of the standard deviation rolling yearly over the 2004–20 period.

Additionally to the standard Value-at-Risk (VaR) computation, two weighted historical VaR methods are used, the HW VaR (Hull and White, 1998) and the BRW VaR (Boudoukh et al., 1998). According to Jorion (2001), the "VaR measure is defined as the worst expected loss over a given horizon under normal market conditions at a given level of confidence". In the context of a short time horizon, in which losses X are generally assumed as $\mathrm{E}[\mathrm{X}] = 0$, we write the VaR of X at confidence level α:

$$\mathbb{P}(X \geq \mathrm{VaR}_\alpha(\mathrm{X})) = \alpha$$

In our tool, the standard VaR is the 5%-quantile of the daily returns over one year multiplied by -1 to be expressed in terms of losses.

In the operational research literature applied to finance, the VaR remains a widely investigated measure (Chun et al., 2012; Barrieu and Scandolo, 2015; Pesenti et al., 2019; Meng and Taylor, 2020; Mercadier and Strobel, 2021; Leung et al., 2021; Hoga and Demetrescu, 2022; Zou et al., 2023; Liu et al., 2024). However, its use has been criticized for not properly allowing for the potential severity of the risk on a portfolio, notably linked with its lack of subadditivity (Acerbi and Tasche, 2002). Thus, the Expected Shortfall has become increasingly popular as an alternative risk measure. It can be defined, at confidence level α, as the expected value of loss X, conditional on the loss exceeding the

[2] Similar results are obtained by running k-means on all indicators, but the process is 7 times slower. All results are available upon request to the authors.

[3] Summary statistics are available in Table B.1 of Appendix B.

Value-at-Risk:

$$ES_{\alpha}(X) = \frac{1}{\alpha}\int_{0}^{\alpha} VaR_u(X)du$$

An extension from the BRW VaR is derived to evaluate the BRW expected shortfall. Furthermore, we compute two more Expected Shortfall measures, both based on the tail average over one year of the daily returns below the 5%-quantile and below $\sigma_S \cdot \Phi^{-1}(\alpha)$, respectively called “ES” and “ESn”, where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cumulative distribution function. Like for the Value-at-Risk, the results are multiplied by -1 to be expressed as losses. These measures remain widely studied in operational research used in the financial industry.

The last indicator specifically designed for the banking system is derived from the Z-score introduced in Section 4.2.1. The Market Zscore (Lepetit et al., 2008; Saghi-Zedek and Tarazi, 2015), is a market-based indicator updated daily. It is a decreasing function of the probability of failure computed with the yearly rolling stocks returns average $\bar{r}$ and standard deviations σ.

$$MZ - score = \frac{1 + \bar{r}_{i,t}}{\sigma_{i,t}}$$

4.3. Systemic risk indicators

4.3.1. Specificities

In the empirical literature (Idier et al., 2014; Weiß et al., 2014b; Reboredo and Ugolini, 2015), systemic metrics generally rest upon national indexes. Although national indexes are influenced by major global shocks, they may also be impacted by localized events of lesser significance. Therefore, to partially mitigate these idiosyncratic effects, we incorporate global and global banking indexes to the systemic factors. In fact, we define banks' systemic risk indicators with their respective national, global (MSCI World) and international bank sector (MSCI World Bank) indexes. Moreover, we compute multiple systemic indicators to filter contradicting assessments as discussed by Kleinow et al. (2017).

4.3.2. Correlations and betas

“**BetaCorrel**” quantifies basic relations between a bank and its corresponding national indexes and with the MSCI World & World Bank indexes. In line with the responsive volatility computations, a factor decay is also applied to the covariances computing the weighted sum of the cross products the returns normalized by $1 - \lambda$:

$$cov(r_{i,t}, r_{m,t}) = (1-\lambda) \cdot r_{i,t} \cdot r_{m,t} + \lambda \cdot cov(r_{i,t-1}, r_{m,t-1})$$

Two metrics are derived from the covariances, the correlation and beta, as usual expressed as follows: $\beta_{i,t} = cov(r_{i,t}, r_{m,t}) / \sigma^2_{m,t}$. Another indicator called tail beta (Straetmans et al., 2008), is calculated as banks' quantile regressions onto their corresponding national, global and bank sector indexes. All these indicators are computed from daily returns over a yearly rolling window.

4.3.3. Systemic risk measures in percentage

The factor called “**SysRisk**” includes indicators that assess systemic risk from two perspectives: the impact on a specific bank when the market is in distress, and the market's response to a stress event affecting a bank.[4] It is built on eighteen measures: three MES and LRMES measures based on the quantile tail average, three BRW MES and LRMES measures extending the BRW expected shortfall, and three MESBeta metrics based on the beta between the bank and market indexes. This factor also includes three Delta Conditional Value-at-Risk (ΔCoVaR) measures. Each set of measures is calculated against the national, global, and bank sector indexes, respectively. The Marginal Expected Shortfall (MES) is a well-known systemic risk measure. It is defined by Acharya et al. (2012; 2017) as the marginal contribution of firm i to systemic risk, as measured by the Expected Shortfall of the financial system and adjusted as follows for our study. It corresponds to the negative value of the one-day loss expected for a bank $(r_{i,t})$ if market returns $(r_{m,t})$ are less than the five percent quartile of market returns over the last year.

$$MES_{i,t} = -E\left[r_{i,t} | r_{m,t} < \mathcal{Q}_{5\%}(r_{m,t})\right]$$

Benoit et al. (2016) show that “the MES of a given financial institution i is proportional to its systematic risk, as measured by its time-varying beta”; in particular, they show that one can write:

$$MESBeta_{i,t} = -\beta_{it} \cdot ES(r_{m,t})$$

Where β_i is the beta of firm i (cf. Section 4.3.2), and the Expected Shortfall corresponds to the tail average of the index daily returns below the 5%-quantile over one year. A long-term adjustment of the MES by Acharya et al. (2012) is called the Long Run Marginal Expected Shortfall (LRMES). More precisely, it is defined as the firm's expected drop in equity value if the market falls by more than a given threshold within the next 6 months, it is computed as follows:

$$LRMES_{i,t} = 1 - e^{-18 \cdot MES_{i,t}}$$

In addition, we use the Benoit et al. (2013)'s version of the Delta Conditional Value-at-Risk (ΔCoVaR), originally introduced by Adrian and Brunnermeier (2016). According to these authors, the CoVaR corresponds to the VaR of the financial system conditionally on a specific event affecting a given firm. The ΔCoVaR is the difference between its CoVaR when the firm is and is not under financial distress.

$$\Delta CoVaR_{i,t}(\alpha) = -\gamma_{i,t} \cdot \left[VaR_{i,t}(\alpha) - VaR_{i,t}(50\%)\right]$$

Where $\gamma_{i,t} = \rho_{i,m,t} \cdot \frac{\sigma_{m,t}}{\sigma_{i,t}}$ and $\alpha = 1\%$.

4.3.4. Systemic risk measures in dollars

The “**SRISK**” factor similarly estimates distressed markets' impact on banks but expressed in US dollars rather than in percent. Nine SRISK metrics relying either on the LRMES, its BRW version or on the LRMESBeta,[5] composed this factor. The SRISK measure, introduced by Acharya et al. (2012) and Brownlees and Engle (2016), is defined as “the expected capital shortfall of a financial entity conditional on a prolonged market decline”. It represents an extension of the MES embodying the equities and liabilities of the financial institution. Acharya et al. (2012) define it as:

$$\begin{aligned} SRISK_t(X_i) &= E[k(Debt + Equity) - Equity|Crisis] \\ &= kD_{it} - (1-k)[1 - LRMES_t(X_i)]E_{it} \end{aligned}$$

with D_{it} the book value of debt, E_{it} the market value of equity and k the prudential capital ratio (commonly set at 8% and at 5.5% for Europe, cf. V-Lab).

4.4. Size indicators

Our project aims to deliver a global ranking of banks based on both individual and systemic risks. But the impact of bank size regarding both aspects of the risks is ambiguous. On the one hand, large banks are related to significantly higher systemic risk (deJonghe, 2010; Banbula and Iwanicz-Drozdowska, 2016; Kleinow et al., 2017) but, on the other

[4] Separating these two perspectives into distinct factors yields similar results, but they exhibit high multicollinearity. These results are available from the authors upon request.

[5] $LRMESBeta_{i,t} = 1 - e^{\ln(1-d)\beta_{i,t}}$, where d = 0.4, the six-month crisis threshold for the market index decline.

hand, the size may have a positive impact on a stand-alone basis. In addition, the role of the size as a major determinant of the systemic risk is unclear, sometimes accused to cloud other factors contributing to banks' systemic importance (López-Espinosa et al., 2012, 2013; Weiß et al., 2014b; Varotto and Zhao, 2018; Elyasiani and Jia, 2019). Nevertheless, it is difficult to assess the additional stability offered by big banks through market-based indicators such as volatility.

Thus, we added the factor called "**Size**" regrouping balance sheet data in US dollars from the banks' balance sheets, e.g., common equities, total equities, total assets, total deposits, reserves for loan losses, the total debt, risk-weighted assets and total loans, and the market capitalization.[6] As balance sheet and market size indicators are a sign of stability, all the features must be read in an inverted scale and are multiplied by -1 to be increasing functions of the risk, further explanation is available in Section 5.1.

The stability dimension is also considered in "**LiquidityMkt**" which is composed of the rolling average of the stock price bid-ask spreads over five days taken as a percentage of the mid-price. The two other constituents of this factor are the daily volume divided by the number of shares and the daily volume multiplied by the stock price. The liquidity factor must be read in an inverted scale, thus as an illiquidity one.

One could argue that this number of factors is arbitrary and that it could be reduced. However, the correlations among factors remain diversified, shows a lack of redundancy, and as explained above, this distribution makes economic sense considering our objective.

5. Dimensionality reduction and clustering

5.1. Feature engineering and dimensionality reduction

As mentioned in Section 4, a second phase of features engineering is performed on the computed risk indicators. In fact, we found remaining outliers in the indicators' values, originating from the automatically extracted raw data. Therefore, the indicators are cleaned, setting as outliers, values exceeding the upper or lower bound derived from a permissive inter-quartile range method. Among the indicators, 8.32% of the values were missing. The percentage of identified outliers was 0.13%. After removing the outliers, the percentage of missing values increased to 8.45%.

Our final target being to assess banks' risks, the indicators are "oriented" according to this point of view. In other words, when an indicator's value increases, the corresponding bank should be characterized as riskier. Therefore, the indicators that are not increasing functions of the risk are multiplied by minus one. Then, the log-modulus transformation (John and Draper, 1980) is applied to all the risk indicators, to pull in the distribution's tails. Finally, the indicators are standardized, which is more consistent with the k-means Euclidean norm criteria.

Once the data are cleaned, oriented as increasing functions of the risk, transformed via log-modulus and standardized, we group the indicators of a specific factor into a dedicated subsample. In this subset, we remove all rows only having missing values, then we partly fill the remaining missing values. The 10 factors are then derived from the individual, systemic and size subsamples using the principal component analysis, that is used for dimensionality reduction. As mentioned in Section 4.1, the groupings of the indicators are chosen based on economic rationale and validated ex post through reliability analysis. First, we evaluate the degree of interrelatedness among the indicators within a factor, employing Cronbach's alpha as an indicator of internal consistency. All factors have a Cronbach's alpha higher than 0.7, except for CreditBS and LiquidityMkt, which are both around 0.67 (see Table C.1 in Appendix C) and can still be considered reasonable according to Taber (2018) .[7] Second, we retain the first principal component as a single factor if the corresponding explained variance accounts for at least 60 percent of the total variance, which is considered satisfactory and is not uncommon in social sciences (Hair et al., 2019). Otherwise, we retain as many principal components as needed to reach at least 60 percent of the cumulative explained variance. As shown in Table C.1, with the exception of two factors, CreditBS and RatioBS, where the first two principal components are retained, only the first principal component is kept for all other cases. The chosen principal components are rescaled by expressing their explained variances as percentages of the total explained variance of the selected components for a given factor. This rescaling reflects that they are based on economic rationale, and thus, we consider them as representing a complete risk dimension. The explained variance, Cronbach's alpha, and the correlations between the indicators and the factors to which they contribute remained stable over time for all factors.[8] Thus, the 10 factors are represented by 12 components, which are used as input features for the k-means algorithm.

5.2. Adjusted nested k-means clustering

Training a supervised algorithm to forecast bank failures would have required the correct identification of past bankruptcies to label our data. However, there is, for instance, still debate on whether all US firms benefiting from the TARP of the United States government, would have defaulted otherwise (Calomiris and Khan, 2015). Thus, our tool does not involve forecasting, rather, it presents a comprehensive overview of the current situation regarding listed banks' riskiness. To do so, an unsupervised learning algorithm is chosen, the k-means clustering (Steinhaus, 1956; MacQueen, 1967; Lloyd, 1982), which has in our perspective some good properties. This widely-used method is fast, transparent, and produces clusters easy to interpret.

Other unsupervised learning algorithms were tested on our sample, namely agglomerative clustering using Ward's method, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Self-Organizing Maps (SOM). As mentioned in Section 2, the first is a simple and widely used hierarchical clustering method, but it is not very efficient for large datasets, which is an issue for our tool. DBSCAN clusters closely packed points based on density, but it yields poor results in our setup. Finally, SOM maps high-dimensional data onto a low-dimensional grid of neurons through unsupervised learning, preserving topological relationships. The results are similar to those obtained using our method, but SOM is more computationally complex. Overall, except for DBSCAN, the results across the different algorithms are closely aligned, reinforcing the validity of our methodology. Further details on the analysis of these algorithms in our setup can be found in Appendix D.

The k-means algorithm clusters the input dataset through a fixed a priori number of clusters, noted k. The biggest issues are precisely related to the choice of the hyper-parameter k, along with the choice of the centroids at initiation which might lead to different inefficient local optimums. The number of clusters is determined according to the well-known elbow curve applied to inertia and the silhouette criterion

[6] As highlighted by an anonymous referee, some banks can be considered as small by international standards while still playing a major role domestically, particularly in smaller countries. To address this issue, we also introduce a country-relative size factor in our analysis, calculated by dividing each bank's total assets by the average total assets of banks within the same country. This factor is not included in the final analysis because our sample consists only of banks with total assets exceeding USD 100 billion, mitigating concerns about relative size discrepancies. Furthermore, the results remain consistent regardless of its inclusion, and the same conclusion is observed when applying the analysis to the larger sample of banks discussed in Section 6.4. All these results are available from the authors upon request.

[7] The correlations among the factors, as well as between the indicators and the factors, are available from the authors upon request.

[8] Further information is available upon request to the authors.

(Rousseeuw, 1987). The Davies-Bouldin index is also considered (Davies and Bouldin, 1979). Since there are no ground-truth labels for external validation, our clustering is evaluated using these three internal measures. Among several initialization algorithms (Oliveira and Nicoletti, 2019), the k-means++ is chosen to set the centroids at initiation. This algorithm is generally quicker and more robust than randomly selecting k points as initial clusters. Given the potential for obtaining suboptimal outputs, we select the configuration that minimizes the inertia from 60 distinct initiations of the process. For sake of consistency, the first point is chosen randomly for some runs and is fixed for others, e.g., the initial centroid can be the closest or farthest point from the dataset's average, median, min, max, first or third quantile. Then, alternatively the four other points are either the farthest from each other or chosen randomly. Upon determination of the initialization centroids, the process reverts to a conventional k-means clustering.

The conventional k-means algorithm does not handle missing data; however, our dataset contains some missing values. In this case, one is often faced with a choice between removing observations or filling in the missing points. These alternatives entail either a loss of information or making decisions based on potentially strong assumptions about the patterns of missingness. In light of this, we developed our own k-means program, which was inspired by a method relying on the majorization-minimization algorithm called k-pod (Chi et al., 2016). More precisely, our program retains only points with at least 60% non-missing values (i.e., at least 7 out of 12 components). This process removes 12.86% of the observations, leaving us with a total of 198,068. Similar results, available upon request from the authors, are obtained with our retention choice compared to a setup where all missing values are removed. However, working with no missing data would result in losing around 50% of the complete dataset.

Additionally, our program uses a nested version of k-means (Niedzielski et al., 2017) to increase granularity without disrupting the overall clustering. This approach ensures consistency over time, with a nested structure enabling smoother transitions as the dataset grows and facilitating long-term analysis. Further details on the comparison between the standard and nested versions of k-means can be found in Appendix E.

Assigning coherent labels to clusters in k-means involves translating abstract clusters into interpretable categories. K-means algorithm partitions the data into k clusters, with each centroid representing the average values of the data points within that cluster. In our application, we compute, for each centroid, the sum of its 12 standardized coordinates. Since all coordinates are increasing functions of risk, we sort the sums and define the highest as the riskiest. The cluster with the highest centroid sum is labeled as 1, representing the riskiest cluster, while the remaining clusters are labeled accordingly, with the cluster having the lowest sum labeled as k, representing the safest. This straightforward method transforms abstract clusters into practical risk categories, making the analysis of the centroids easier. A flowchart in Appendix F (Fig. F.1) illustrates the different steps of the tool.

6. Applied clustering analysis and empirical illustration

6.1. The choice of the number of clusters

Our dataset is built on 256 international listed banks from 43 countries, among the biggest in terms of total assets based on Barth and Wihlborg (2016, J.R. 2017). The study spans from January 2, 2004 to May 31, 2024, including both crisis & post-crisis periods and the raw sample is composed of balance sheet and market data.

The components are first inputted into a standard k-means algorithm. As explained in Section 5.2, the number of clusters is determined using three metrics: inertia, silhouette score, and the Davies-Bouldin index. Table 1 presents these metrics, and it shows that moving from one to two clusters results in a 32% drop in inertia, with the silhouette score and Davies-Bouldin index both indicating that two clusters are optimal. The initial split creates two sub-clusters, referred to as the riskiest and safest clusters.

Table 1
K-means output metrics for the entire sample split into k clusters.

K	No Obs	Inert	Sil	DBI
1	198,068	1644,286		
2	198,068	1245,740	0.31	1.47
3	198,068	1052,103	0.21	1.62
4	198,068	906,165	0.22	1.52
5	198,068	831,524	0.20	1.58

Notes. The column "k" indicates the number of clusters, with k being the number of clusters and $k \in \mathbb{N} \cap [1, 5]$. The columns labeled "No Obs", "Inert", "Sil", and "DBI" represent the sample size, Inertia, Silhouette, and Davies-Bouldin Index, respectively. The chosen value of k is highlighted in grey.

Running a two clusters k-means algorithm as suggested by these metrics, we end up with an historical timeline (cf. Fig. 1) highlighting four periods globally riskier than the others. The first three points in time match those highlighted by the probability of crisis graph from the recent paper by Engle and Ruan (2019). The first period, the riskier one, corresponds to the 2008 GFC originating in the United States and the next ones are respectively linked to the 2012 European sovereign debt crisis and the 2016 Asian debt crisis. The last turbulent period depicts the COVID-19 pandemic, and to a lesser extent, the 2022 Russian invasion of Ukraine and the 2023 banking crisis. While it seems to efficiently split, for a given date, the riskiest banks (in red, bottom) from the safest ones (in green, top), it is sensible to propose a more detailed analysis.

Higher granularity can easily be achieved by running a k-means algorithm in each cluster, a concept known as nested k-means. As usual, we start by determining the two optimal values of k based on the three metrics. As shown on the left-hand side of Table 2, the decision to split the riskiest cluster into two is supported by a clear analysis, with a silhouette score above 0.40 and the lowest DBI value.

The right-hand side of Table 2 does not provide conclusive figures for selecting the number of clusters. While the highest silhouette score favors two sub-clusters, it is only marginally higher than for three. Conversely, the DBI suggests a preference for three sub-clusters. Given that the safest cluster comprises 77.9% of the sample, we opted to split it into three sub-clusters. Additionally, maintaining 5 clusters overall is consistent with most supervisory ratings, such as CAMELS ratings or BOPEC ratings in the U.S. ,[9] which rate banks from 1 to 5.

In a nutshell, we split the riskier cluster in two sub-clusters and the safest one in three, leading to an overall $k = 5$. More explicitly, this can be represented as $k = 2 \rightarrow 2|3$, as nested k-means first requires two clusters, which are then split into 2 and 3 sub-clusters, respectively. In addition to aligning with the three metrics, the nested algorithm provides steadier results than the standard $k = 5$ version when a component is removed.

By construction, the same overall shape is observed whether using $k = 2$ or $k = 2 \rightarrow 2|3$. However, it is more shaded for $k = 2 \rightarrow 2|3$, where the red and orange clusters split the previous riskier cluster and the other colors divide the safest one. According to Fig. 2, the increase in granularity outputs a lower number of extremely risky banks even during the GFC compared to Fig. 1, although higher than in other periods. Additionally, the European crisis seems to have reached its maximum in 2012, which is consistent with the reality. Comparing the two riskiest clusters, our tool distinguishes a more severe impact on the banking industry of the GFC – a banking crisis – from the COVID-19 pandemic – not directly related to the banking sector. More specifically, the banking sector, although still under pressure, was quickly supported by massive support programs launched worldwide by central banks, e.g., the Fed's monetary policy measures or the ECB's Pandemic Emergency Purchase

[9] Appendix G provides additional information on these rating systems.

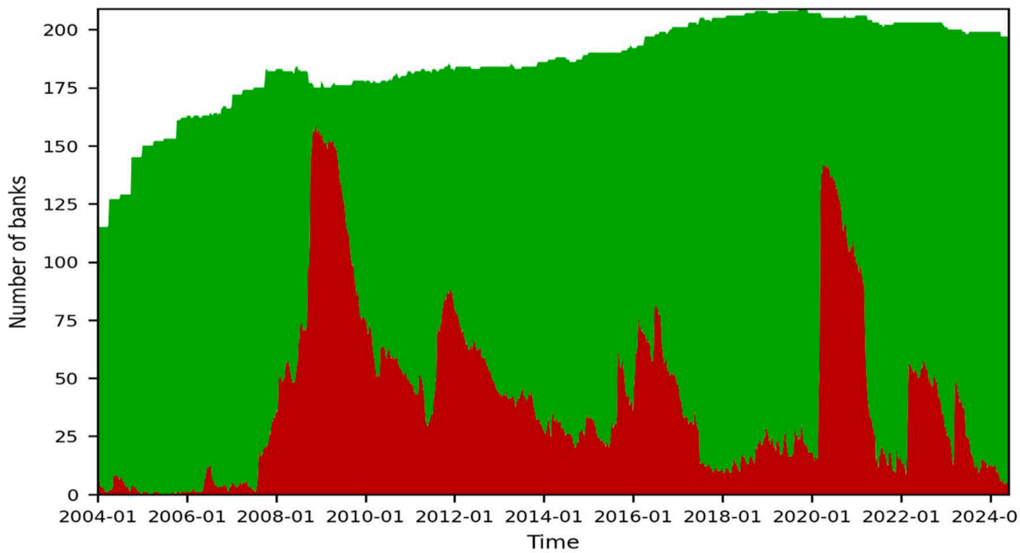
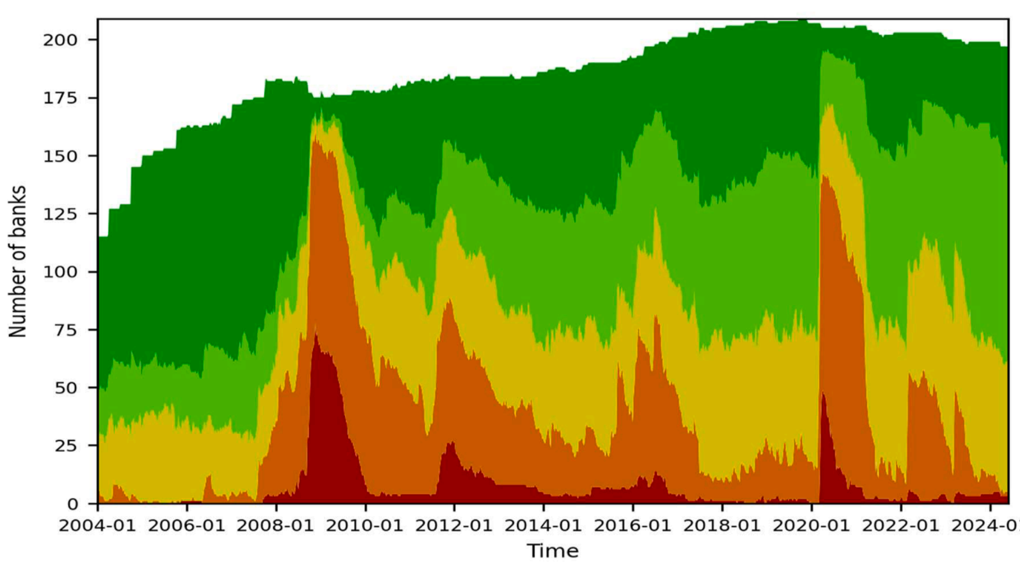


Fig. 1. Clustering timeline ($k = 2$).
Notes. This graph shows how banks are clustered through time. The results notably highlight crisis periods with higher number of banks within the riskiest cluster (red).

Table 2
K-means output metrics for the sub-samples after the initial split in two (2 → •|•).

	Risky cluster				Safe cluster			
K	No Obs	Inert	Sil	DBI	No Obs	Inert	Sil	DBI
2	43,798	296,394	0.42	1.43	154,276	639,920	0.24	1.73
3	43,798	261,088	0.17	1.83	154,276	545,784	0.21	1.62
4	43,798	236,209	0.17	1.78	154,276	498,240	0.20	1.56

Notes. K-means output metrics for the sub-samples belonging to the risky (lhs) and safe (rhs) clusters after the initial split in two (2 → •|•). The column "k" indicates the number of clusters, with k ∈ {2, 3, 4}. The columns labeled "No Obs", "Inert", "Sil", and "DBI" represent the sub-sample size, Inertia, Silhouette, and Davies-Bouldin Index, respectively. The chosen value of k is highlighted in grey.



Notes. This graph shows how banks are clustered through time with a higher granularity than in fig.1. The results still highlight crisis periods but with a consistent lower number of banks within the riskiest cluster (red, bottom).

Fig. 2. Clustering timeline ($k = 2\rightarrow2|3$).
Notes. This graph shows how banks are clustered through time with a higher granularity than in Fig. 1. The results still highlight crisis periods but with a consistent lower number of banks within the riskiest cluster (red, bottom).

Programme (Rizwan et al., 2020; Borri and Giorgio, 2021). Furthermore, the last moderate spike matches the 2023 banking crisis that was quickly handled by regulators. About ten days after the failure of Silicon Valley Bank and Signature Bank, bigger banks (e.g., First Republic Bank and Credit Suisse) were taken over under the supervision of financial authorities (Metrick and Schmelzing, 2023).

In addition to illustrating the temporal distribution of banks across clusters (cf. Figs. 1 and 2), we employ t-distributed Stochastic Neighbor Embedding (t-SNE; cf. Fig. 3) to visualize the structure of data points within each cluster over time. Specifically, the clustering is generated annually using extended data windows, starting from January 2, 2004, to December 31, 2013, and progressively extending to December 31, 2023. These visualizations support the stability of the clusters over the studied period. More details are available in Appendix G.

6.2. Centroids' analysis

From an economic point of view, the analysis of the centroids matters. As introduced in the end of Section 5.2, the k-means algorithm only creates k classes regardless of their meaning. Thus, we set the clusters order in function of the individual risk factors, as they are derived from the simplest and most transparent indicators.

On Figs. 4 and 5, the two worst clusters stick out, and above all emphasizing the riskiest banks' contribution to the first one. In addition, the most solvent institutions stand out in the fifth group, except for the balance sheet ratios factor. However, there is no obvious conclusion regarding the third and fourth clusters. The second principal components of both 'RatioBS' and 'CreditBS' are used in the clustering, but their impact is minimal (see Fig. H.1 in Appendix H). The next paragraph is focused on the "Size" factor, as it is linked to the rest of the analyses.

On Fig. 6 (left), the smallest firms are concentrated in the third cluster (let us remember that the indicators composing the size factors are multiplied by -1) and the biggest ones in the second but mostly in the fourth one. This is an argument in favor of a certain stability brought by big banks. Above all, the riskiest and safest clusters are not impacted by the size, an argument supporting the exclusion of banks' size from the major determinants of their risks, in line with Varotto and Zhao (2018). According to the scale, the regulatory capital factor has a limited impact on the risk and groups banks like the size factor.

While the two worst clusters still emerge for systemic measures, we emphasize a difference between the worst two clusters of the "SysRisk" factor (Fig. 7) leads us to underline that the large banks from the second riskiest cluster seem to contribute the most to the systemic risk. A phenomenon accentuated by the addition of the COVID-19 pandemic data, in line with Borri and Giorgio (2021). In addition, the less systemically risky banks belong to the third cluster and not anymore to the fifth. According to the importance of "BetaCorrel" and from a detailed analysis using the algorithm directly on the 72 indicators, the systemic factors' worst third cluster seems to be particularly related to systematic (beta-based) indicators. By extrapolating from the previous paragraph's size analysis, small banks appear to have, as expected, less relations with the market – through lower betas. A phenomenon also visible, to a lower extent in the third group, for the other systemic factor, the SRISK. But highlighting that bank sizes and betas are not sufficient determinants of their overall riskiness, as they do not belong to the safest cluster.

The SRISK is a systemic measure, expressed in US dollars, having common components with the size factor, namely the debt and the market capitalization. The noteworthy contribution here (see Fig. 8) is the safety brought by banks having a low SRISK. Although this factor is globally well ordered, the biggest banks from the fourth cluster seem to provide relatively high values. Leading us to the same conclusion as in Varotto and Zhao (2018)'s study which points out that systemic indicators primarily driven by firm size lead to an overriding concern for big firms.

Lastly, the illiquid banks have more odds to belong either to the riskiest cluster or the third cluster, grouping small banks according to the size analysis (Fig. 9). As expected, small banks and extremely risky ones tend to be illiquid.

Having acknowledged the limitations of our dataset and tool, we conclude by stating that MES, ΔCoVaR and Tail Beta measures tend to overstate small banks as the safest banks by default. The SRISK has, by construction (expressed in US dollars), rather a tendency to overstate big institutions as riskier. Additionally, using softmax regression, we identify SRISK, Size, BetaCorrel, CreditMkt, LiquidityMKT, SysRisk, and IndivMktRisk as key contributors to the clustering process. Details on their importance at different stages of the process can be found in Appendix I.

6.3. Clusters' behavior

Finally, we draw conclusions from the overall behavior of the clusters. First, Table 3 summarizes the average of consecutive weeks banks spend in a specific cluster.

Except for the safest cluster – in which banks tend to remain for longer time periods – banks generally spend on average around six months in a specific cluster. Additionally, one can be interested in the

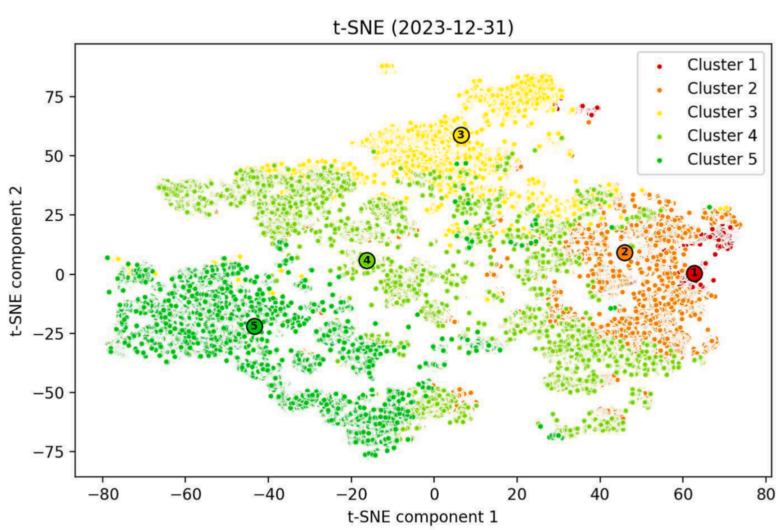


Fig. 3. t-SNE clustering visualization ($k = 2 \rightarrow 2|3$).
Notes. This graph shows the t-SNE output of our clustering from January 2, 2004, to December 31, 2023. The centroids are indicated by their numbered labels.

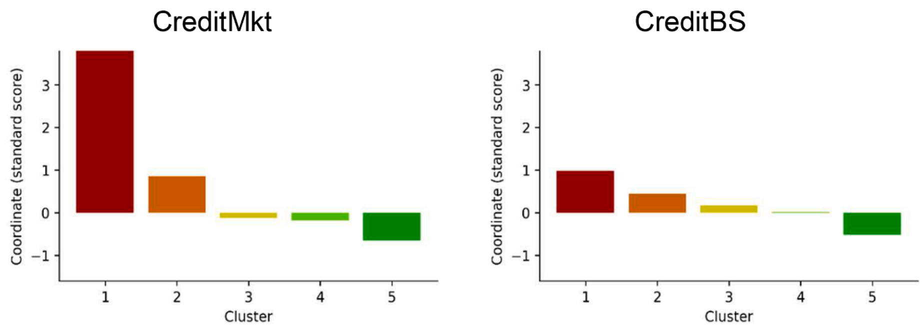


Fig. 4. Credit related individual risks centroids.
Notes. These bar charts display the standardized coordinates of the clusters' centroids ordered from the riskiest (red, lhs) to the safest (green, rhs) per credit related factors. The results notably emphasize the riskiest banks' contribution to the first cluster.

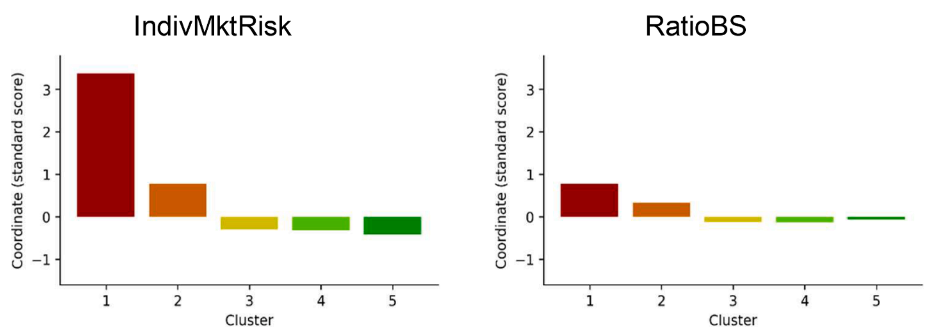


Fig. 5. Other individual risks centroids.
Notes. These bar charts display the standardized coordinates of the clusters' centroids ordered from the riskiest (red, lhs) to the safest (green, rhs) per individual risk factors. The results notably emphasize the riskiest banks' contribution to the first cluster.

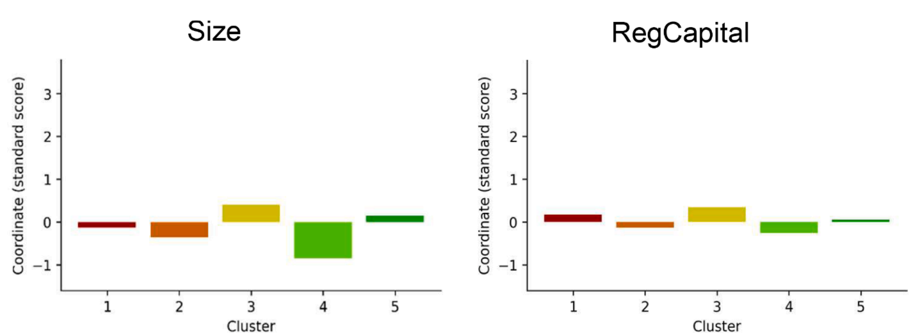


Fig. 6. Size centroids.
Notes. These bar charts display the standardized coordinates of the clusters' centroids ordered from the riskiest (red, lhs) to the safest (green, rhs) per size and regulatory capital factors (inverted scales). Above the fact that the smallest and biggest firms are spread within the middle clusters, the results underline that the size does not impact the most extreme clusters.

probability a bank remains in the same cluster or transfers to another one. Thus, Table 4 is the weekly transition matrix of a Markov chain for all banks over the entire time length.

It is extremely rare that a bank is transferred by more than two clusters (i.e., 4 to 1 or 2 to 5, etc.), and the more likely transitions are either by one cluster or, for a large bank between clusters 2 and 4. Finally, small banks in the third cluster are more likely to become safer than riskier.

In addition, our tool provides an aggregated version of the clustering per country and per region, which can help to identify fragile areas. To do so, weighted averages of banks' cluster numbers are computed in terms of total assets per country and region. The results are displayed on a dedicated website (cf. Appendix L). Beyond the historical graphs (cf. Figs. 1, 2, and K.1) and each bank, country, and region's time frame (cf. Figs. J.1 and J.3), the report includes an exhaustive listing of the banks per cluster, as of the most recent date.

6.4. Generalizability across a broader universe

The original sample of 256 banks has been expanded to include additional banks and countries. Specifically, the dataset was enlarged by lowering the selection criterion to include all banks with total assets exceeding USD 50 billion, as in Weiß et al. (2014a) and Kanno (2015). This adjustment added 138 banks and 8 countries to the sample, resulting in a dataset comprising 394 banks across 51 countries. On average, the dataset includes 347 (21) banks per year. Over the period,[10]

[10] Yearly data is only available until 2023 from the Financial Stability Board (2023).

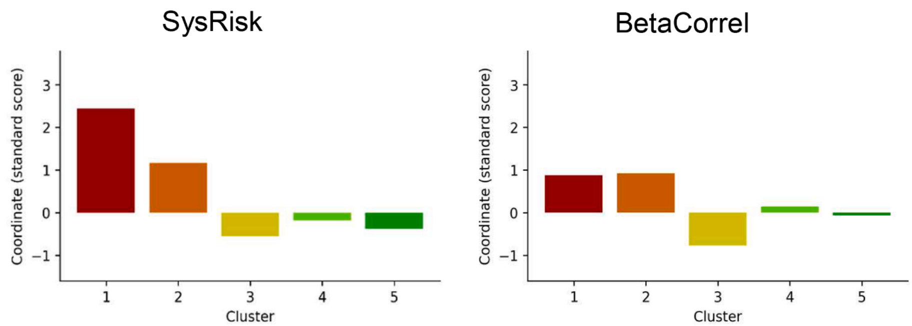


Fig. 7. Systemic risks centroids.
Notes. These bar charts display the standardized coordinates of the clusters' centroids ordered from the riskiest (red, lhs) to the safest (green, rhs) per systemic factors. Beyond the beta's influence clustering the systemically safest banks in the middle cluster, the results underline the riskiest banks' contribution to the first two clusters.

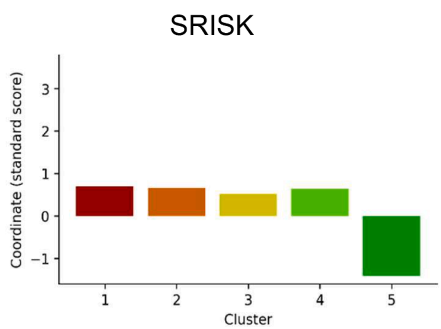


Fig. 8. SRISK centroids.
Notes. This bar chart displays the standardized coordinates of the clusters' centroids ordered from the riskiest (red, lhs) to the safest (green, rhs) per SRISK factor. The result notably highlights the safety brought by low SRISK.

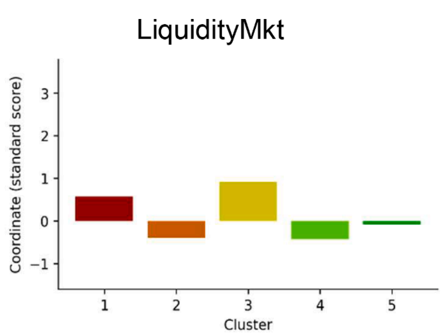


Fig. 9. Market liquidity centroids.
Notes. This bar chart displays the standardized coordinates of the clusters' centroids ordered from the riskiest (red) to the safest (green) per liquidity factor (inverted scale, illiquidity). The result notably emphasizes that small and highly risky banks tend to be illiquid.

Table 3
Average number of consecutive weeks spent by banks in each cluster.

Cluster	1	2	3	4	5
Number of weeks mean (SD)	23.1 (37.6)	20.8 (29.9)	25.5 (67.5)	19.7 (41.1)	48.7 (98.3)

Table 4
Clusters' weekly transition matrix.

Cluster	1	2	3	4	5
1	95.929%	3.870%	0.175%	0.013%	0.013%
2	0.875%	95.202%	0.933%	2.674%	0.316%
3	0.048%	0.700%	96.226%	2.061%	0.965%
4	0.004%	1.788%	1.972%	95.101%	1.135%
5	0.003%	0.328%	0.718%	0.904%	98.047%

Notes. $\mathbb{P}_{i,j}$ is the probability of transitioning from i to j.

Table 5
K-means output metrics for the entire sample split into k clusters.

k	No Obs	Inert	Sil	DBI
1	303,990	2535,797		
2	303,990	1923,555	0.30	1.50
3	303,990	1634,169	0.21	1.68
4	303,990	1419,424	0.21	1.54
5	303,990	1305,373	0.20	1.59

Notes. The column "k" indicates the number of clusters, with k being the number of clusters and $k \in \mathbb{N} \cap [1, 5]$. The columns labeled "No Obs", "Inert", "Sil", and "DBI" represent the sample size, Inertia, Silhouette, and Davies-Bouldin Index, respectively. The chosen value of k is highlighted in grey.

the extended dataset represents 82.6% (5.3%) of total bank assets on average. Tables 5 and 6 correspond to the various stages of the nested clustering process, as described in Section 6.1. Both the silhouette score and the Davies-Bouldin index yield values closely aligned with those reported in Tables 1 and 2, thereby reinforcing the generalizability of the tool as the sample size of banks increases. Additionally, Appendix G addresses temporal generalizability, highlighting the stability of our results across different time periods.

7. Conclusion

In this paper, we provide a dynamic decision support tool designed to cluster listed banks based on representative risk factors. This tool computes and summarizes a large set of financial ratios, stand-alone and systemic risk indicators, combining balance sheet and market information into fewer representative factors. Notably, our analysis reveals that bank size does not consistently impact systemic risk, as the safest and riskiest clusters are not determined by size, highlighting the nuanced relationship between bank size and risk.

We detail the mechanisms of the tool, including dynamic data extraction, cleaning procedures, indicator and factor calculations, and the choice and settings of algorithms. Specifically, we employ a nested k-means algorithm adjusted to handle missing values, ensuring robust clustering results. The tool's ability to aggregate bank clustering data by country and region facilitates the identification of areas of financial fragility contributing to the broader understanding of banking sector stability. Our results demonstrate coherence with established benchmarks, such as the 5-year CDS, and highlight that over 70% of Global

Table 6
K-means output metrics for the sub-samples after the initial split in two (2 → •|•).

	Risky cluster				Safe cluster			
k	No Obs	Inert	Sil	DBI	No Obs	Inert	Sil	DBI
2	71,809	495,112	0.42	1.36	232,181	955,121	0.24	1.8
3	71,809	431,527	0.18	1.82	232,181	836,088	0.19	1.64
4	71,809	393,677	0.17	1.84	232,181	745,517	0.18	1.60

Notes. K-means output metrics for the sub-samples belonging to the risky (lhs) and safe (rhs) clusters after the initial split in two (2 → •|•). The column "k" indicates the number of clusters, with k ∈ {2, 3, 4}. The columns labeled "No Obs", "Inert", "Sil", and "DBI" represent the sub-sample size, Inertia, Silhouette, and Davies-Bouldin Index, respectively. The chosen value of k is highlighted in grey.

Systemically Important Banks (G-SIBs) were in the riskiest cluster during the Global Financial Crisis. Additionally, we observe that banks typically remain in a cluster for an average of six months, with minimal transitions between non-adjacent clusters.

Our results are displayed on a dedicated website, enhancing both accessibility and usability. The tool's flexibility allows for easy customization, enabling the addition or removal of banks, indicators, or factors to suit specific research needs. This adaptability serves both academics and practitioners such as financial analysts and regulators by providing a flexible framework for comparing the riskiness of listed banks. Despite its advantages, the framework has some limitations, notably the need to define certain parameters and thresholds.[11]

The implications of this tool are significant for both academic research and practical applications in financial risk management. Academics can leverage the tool's comprehensive framework for in-depth comparative studies on bank riskiness, while practitioners and specifically bank supervisors can use it to monitor bank health, anticipate potential failures, and make informed decisions regarding regulatory interventions such as for example onsite examinations.

Looking forward, the tool can be expanded to incorporate not only quantitative data but also qualitative factors such as institutional factors, culture differences among countries, climate risk, environmental aspects, and Environmental, Social, and Governance (ESG) criteria. This enhancement would provide a more holistic view of bank risk, acknowledging the growing importance of qualitative factors in financial stability assessments.

Future research could examine alternative clustering algorithms, such as Fuzzy C-Means, to introduce soft clustering properties. Integrating real-time data feeds and automated updates could further improve the tool's usability and relevance in rapidly changing financial environments. Additionally, extending the tool to cover a broader range of financial institutions, including non-listed banks and other financial entities, would provide a more comprehensive risk assessment landscape.

Despite its advancements, the tool is not without limitations. The reliance on historical data may not fully capture future risk dynamics or emerging threats. The tool's effectiveness is also dependent on the quality and completeness of the input data, and missing or inaccurate data could affect the reliability of the rankings. Addressing these limitations through advanced data imputation techniques and the inclusion of forward-looking indicators, such as market sentiment or macroeconomic forecasts, could enhance the tool's predictive capabilities.

All in all, while this study represents a significant step forward in bank risk assessment, there remain opportunities for further refinement and expansion. By addressing the identified limitations and exploring new methodologies, future research can build on this work to develop even more effective tools for evaluating and managing bank risk all around the world.

CRediT authorship contribution statement

Mathieu Mercadier: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Amine Tarazi:** Conceptualization, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing. **Paul Armand:** Conceptualization, Methodology, Supervision. **Jean-Pierre Lardy:** Conceptualization, Data curation, Methodology, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing.

Acknowledgments

The paper has benefited from valuable comments and suggestions received from L. Lepetit, A. Sauviat and participants at CleRMa seminar of the Finance, Information and Corporate Responsibility axis, the Autumn workshop of the Financial Intermediation Network of European Studies, the 11th International Conference of the Financial Engineering and Banking Society and the 7th Vietnam Symposium in Banking and Finance. We also thank L. Xu for her advice. We thank the anonymous reviewers and the editor for their comments and suggestions, which have significantly contributed to improving the quality of the paper. Any remaining errors or omissions are solely the responsibility of the authors.

Appendix A. Dataset summary statistics

Table A.1
Bank sample summary statistics.

Mean	Std.	Min	p25	p50	p75	Max
398	565	26	90	161	410	3099

Notes. These statistics summarize the banks' total assets, with all figures averaged over the full years 2004–2023 period, based on an average of 227 (±12) banks per year. Each bank has had at least one instance of total assets exceeding USD 100 billion. All values are in billions of USD. Data source: Bloomberg.

[11] The limitations of the tool, along with the rationale for the choices made, are provided in Appendix M.

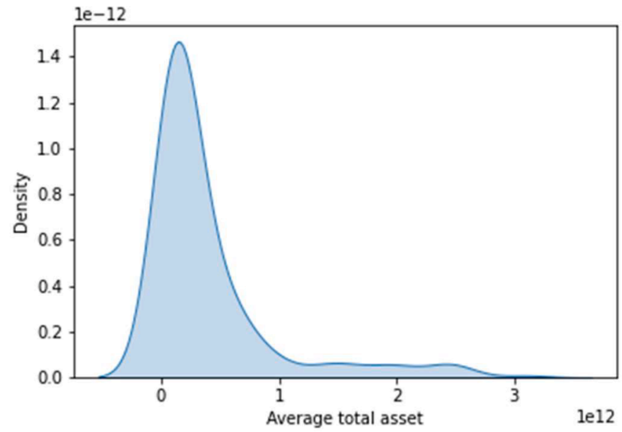


Fig. A.1. Distribution of the banks' average total assets.
Notes. Distribution of the banks' average total assets over the full years 2004–23 period. Each bank has had at least one instance of total assets exceeding USD 100 billion. All values are in USD. Data source: Bloomberg.

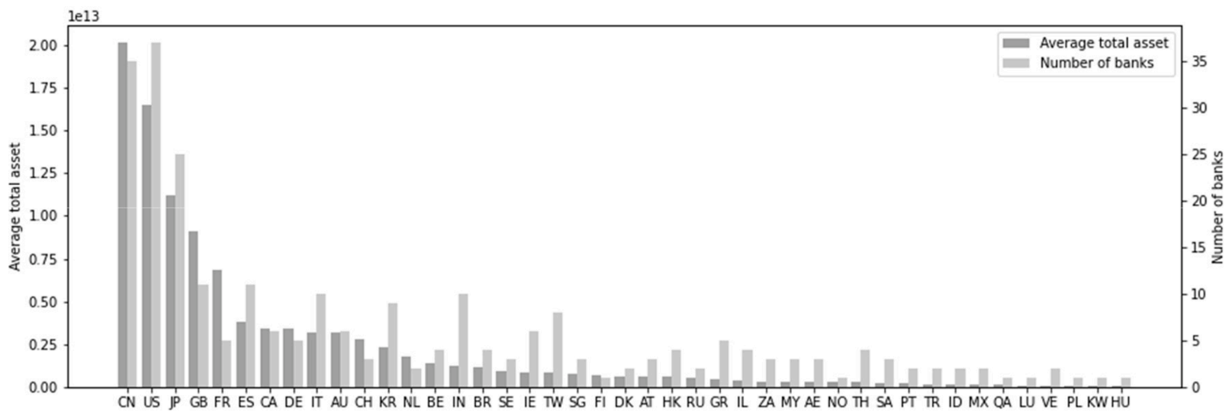


Fig. A.2. Bank distribution per country.
Notes. The dark grey bars (lhs) show the average total assets for each bank over the full years 2004–23 period per country. The number of banks (rhs) corresponds to the total number of available banks per country for the 2004–23 sample. Each bank has had at least one instance of total assets exceeding USD 100 billion. All values are in USD. Data source: Bloomberg.

Appendix B. Indicators summary statistics

Table B.1
Indicators summary statistics.

Indicator	Factor	N	Mean	Std	Median	Min	q25	q75	Max
Volatility	IndivMktRisk	214,752	0.328	0.220	0.276	0.000	0.203	0.382	3.041
Correl	BetaCorrel	213,121	0.610	0.198	0.648	-0.299	0.511	0.753	1.548
CorrelGlb	BetaCorrel	215,214	0.331	0.216	0.310	-0.679	0.171	0.493	1.694
CorrelGlbBk	BetaCorrel	215,214	0.379	0.237	0.356	-0.634	0.197	0.562	1.337
Debt2EV	CreditMkt	209,561	0.670	0.233	0.710	0.000	0.515	0.870	1.000
Debt2Ast	RatioBS	227,415	0.230	0.168	0.190	0.000	0.102	0.318	0.976
ES	IndivMktRisk	212,126	0.044	0.027	0.038	0.005	0.028	0.052	0.399
ESn	IndivMktRisk	211,164	0.048	0.032	0.040	0.004	0.030	0.055	0.438
ESBRW	IndivMktRisk	212,067	0.045	0.030	0.038	0.002	0.028	0.052	0.413
VaR	IndivMktRisk	212,149	0.031	0.018	0.027	0.000	0.020	0.036	0.275
VaRHW	IndivMktRisk	212,044	0.033	0.025	0.026	0.000	0.019	0.038	0.327
VaRBRW	IndivMktRisk	212,114	0.032	0.021	0.027	0.000	0.019	0.037	0.299
ZScore	CreditBS	211,011	67.98	75.08	44.80	-37.49	23.40	85.17	1065.89
MZScore	IndivMktRisk	212,225	8.259	4.458	7.582	-1.027	5.474	10.219	75.282
Beta	BetaCorrel	213,079	1.052	0.472	1.044	-5.858	0.775	1.306	7.903
BetaGlb	BetaCorrel	215,102	0.797	0.651	0.690	-8.086	0.343	1.166	12.811
BetaGlbBk	BetaCorrel	215,113	0.693	0.550	0.597	-7.103	0.300	1.024	11.554
TailBeta	BetaCorrel	211,509	1.066	0.470	1.049	-5.446	0.794	1.309	7.381
E2C	CreditMkt	202,030	0.016	0.027	0.008	0.000	0.004	0.017	0.264
CrdGrd	CreditMkt	204,484	0.026	0.043	0.008	0.000	0.001	0.034	0.675
MES	SysRisk	211,537	0.030	0.020	0.025	0.000	0.017	0.038	0.234
MESGlb	SysRisk	212,233	0.020	0.017	0.016	0.000	0.009	0.025	0.249

(*continued on next page*)

Table B.1 (*continued*)

Indicator	Factor	N	Mean	Std	Median	Min	q25	q75	Max
MESGlbBk	SysRisk	212,233	0.022	0.018	0.017	0.000	0.010	0.028	0.184
MESBRW	SysRisk	211,524	0.031	0.022	0.025	0.000	0.017	0.038	0.345
MESBRWGlb	SysRisk	212,205	0.020	0.018	0.015	0.000	0.009	0.025	0.258
MESBRWGlbBk	SysRisk	212,222	0.022	0.020	0.017	0.000	0.010	0.028	0.305
MESBeta	SysRisk	211,519	0.030	0.022	0.024	-0.195	0.016	0.037	0.343
MESBetaGlb	SysRisk	212,197	0.018	0.019	0.013	-0.154	0.006	0.023	0.233
MESBetaGlbBk	SysRisk	212,207	0.019	0.020	0.014	-0.160	0.007	0.025	0.232
LRMES	SysRisk	211,537	0.388	0.173	0.366	0.000	0.261	0.494	0.985
LRMESGlb	SysRisk	212,233	0.277	0.170	0.244	0.000	0.152	0.362	0.989
LRMESGlbBk	SysRisk	212,233	0.295	0.177	0.259	0.000	0.162	0.394	0.964
LRMESBRW	SysRisk	211,537	0.388	0.177	0.362	0.000	0.259	0.495	1.000
LRMESBRWGlb	SysRisk	212,233	0.275	0.172	0.240	0.000	0.148	0.361	1.000
LRMESBRWGlbBk	SysRisk	212,233	0.295	0.178	0.259	0.000	0.160	0.397	1.000
dCoVaR	SysRisk	211,537	0.019	0.012	0.016	0.000	0.011	0.024	0.117
dCoVaRGlb	SysRisk	212,233	0.008	0.008	0.006	0.000	0.003	0.011	0.066
dCoVaRGlbBk	SysRisk	212,233	0.013	0.012	0.009	0.000	0.004	0.017	0.102
SRISKMES	SRISK	206,334	12,336	29,599	3856	-107,781	-450	12,574	232,408
SRISKMESGlb	SRISK	206,955	9785	28,413	2830	-117,973	-1813	10,678	208,953
SRISKMESGlbBk	SRISK	206,972	10,275	28,799	3031	-116,329	-1641	11,106	214,440
SRISKMESBRW	SRISK	206,331	12,362	29,651	3887	-108,806	-458	12,562	230,955
SRISKMESBRWGlb	SRISK	206,949	9739	28,353	2827	-117,815	-1839	10,661	208,067
SRISKMESBRWGlbBk	SRISK	206,960	10,318	28,761	3059	-112,852	-1615	11,221	216,866
SRISKBeta	SRISK	208,369	13,339	29,079	4227	-98,352	171	13,588	249,008
SRISKBetaGlb	SRISK	208,873	11,209	28,648	3296	-104,218	-843	11,635	220,714
SRISKBetaGlbBk	SRISK	208,830	10,398	28,228	2987	-107,794	-1306	11,063	214,489
ROA	CreditBS	223,122	0.008	0.009	0.008	-0.080	0.004	0.012	0.103
ComEqy2Ast	CreditBS	227,268	0.071	0.037	0.065	-0.137	0.051	0.086	0.607
LLRsrv2Loan	RatioBS	219,325	0.025	0.026	0.017	0.000	0.009	0.031	0.270
NPA2Ast	RatioBS	195,483	0.019	0.031	0.009	0.000	0.004	0.020	0.309
NPL2Loan	RatioBS	191,185	0.030	0.045	0.016	0.000	0.008	0.032	0.453
Loan2Dpst	RatioBS	221,306	0.982	0.506	0.871	0.000	0.726	1.100	6.409
Loan2Ast	RatioBS	222,383	0.542	0.170	0.575	0.000	0.449	0.660	1.308
TexasRatio	CreditBS	193,564	0.192	0.243	0.106	0.000	0.052	0.232	3.267
PBRatio	CreditBS	210,556	1.296	0.997	1.102	-10.227	0.662	1.700	16.001
BidAskSprd	LiquidityMkt	128,543	0.002	0.003	0.001	0.000	0.001	0.002	0.031
VolmPx	LiquidityMkt	213,410	112.00	218.29	37.25	0.00	11.65	120.40	2229.34
VolmNbSh	LiquidityMkt	211,619	0.004	0.005	0.003	0.000	0.002	0.005	0.064
T1CapRatio	RegCapital	200,234	0.124	0.042	0.120	-0.061	0.096	0.144	0.740
Cap2RskCap	RegCapital	202,936	0.150	0.042	0.144	-0.053	0.125	0.169	0.760
RWA2Ast	RegCapital	196,500	0.543	0.194	0.553	0.000	0.406	0.675	1.693
Dpst2Ast	RatioBS	225,706	0.607	0.195	0.644	0.000	0.496	0.755	1.063
RWA	Size	196,500	214,459	338,388	91,086	0	51,081	224,155	3508,708
LLRsrv	Size	218,219	4201	7286	1629	0	610	4252	71,062
Deposit	Size	225,445	231,437	382,378	91,570	0	47,980	238,401	3938,997
TotLoan	Size	222,400	197,250	301,831	82,839	0	43,593	222,873	3649,939
TotDebt	Size	226,894	109,170	181,562	32,401	0	9596	114,328	2064,822
ComEqy	Size	226,586	25,216	40,785	10,969	-106,735	5424	26,247	410,123
TotEqy	Size	226,755	27,756	45,287	11,705	-106,735	5806	28,774	456,953
TotAst	Size	226,932	426,182	659,806	158,224	3	81,088	434,728	6588,877
MktCap	Size	210,849	28,425	42,334	13,617	2	5524	33,209	489,320

Notes. This table includes the factor to which the indicator is affected, the number of observations (N), mean, standard deviation (std), median, minimum (min), first quartile (q25), third quartile (q75), and maximum (max). Note that the LRMES beta related measures are included. Although they are not in the final tools, they still are inputs for the corresponding SRISK measures.

Appendix C. PCA summary

Table C.1
PCA summary.

Factor	No Indic	Cronbach's alpha	ExpVar PC1	ExpVar PC2	Weight PC1	Weight PC2
BetaCorrel	7	0.91	0.67	0.17	1	0
CreditBS	5	0.67	0.44	0.22	0.66	0.34
CreditMkt	3	0.77	0.69	0.26	1	0
IndivMktRisk	8	0.98	0.90	0.04	1	0
LiquidityMkt	3	0.67	0.61	0.26	1	0
RatioBS	7	0.76	0.42	0.32	0.57	0.43
RegCapital	3	0.79	0.72	0.25	1	0
SRISK	9	0.99	0.91	0.03	1	0
Size	9	0.96	0.77	0.08	1	0
SysRisk	18	0.98	0.79	0.07	1	0

Notes. The first and second columns shows respectively the number of indicators and the Cronbach's alpha per factor. Then, the explained variance for the first two principal components is presented. Finally, the last two columns show the explained variance rescaled values. If more than one principal component is needed to reach

60% of the cumulated explained variance, the weights are rescaled as percentage of the sum of the explained variance of the chosen principal components. Otherwise, 100% is distributed to the first principal component.

Appendix D. Comparison of algorithms

Ward's method is a commonly used hierarchical clustering technique that works by minimizing within-cluster variance. However, it is computationally expensive for large datasets due to its quadratic time complexity, which poses a challenge for the scalability of our tool.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering algorithm that groups together points that are closely packed based on a specified density criterion. It identifies regions of high density (clusters) and labels points in sparse regions as noise or outliers.

Self-Organizing Maps (SOM) use a grid of neurons in a low-dimensional space to represent high-dimensional data through unsupervised learning. Neurons compete to match input data; when an input is presented, the most similar neuron and its neighboring neurons are updated to resemble the input. This cooperative learning process allows the network to self-organize over time, effectively mapping high-dimensional data onto the grid while preserving topological relationships—meaning similar data points are mapped to nearby neurons. K-means and SOM work similarly in that both algorithms iteratively adjust representations—k-means adjusts cluster centroids, while SOM adjusts the weight vectors of neurons (representing clusters)—to better capture the underlying structure of the data.

Due to memory limitations, notably when using Ward's method, we compare the various algorithms by computing the average and standard deviation of the Silhouette and Davies-Bouldin scores from 50 randomly generated sub-samples, each containing 20,000 observations from the original dataset (excluding missing values).

To ensure comparability with our other experiments, we fixed the number of clusters to 5. The results in Table D.1 show that both the average Silhouette score and the Davies-Bouldin index favor K-Means. Additionally, we ran hierarchical clustering without predefining the number of clusters, which resulted in significantly better metrics (e.g., a Silhouette score of 0.37). However, this resulted in a large number of clusters—nearly a quarter of the total observations—making it impractical for meaningful data analysis.

Although DBSCAN produces a slightly better average Davies-Bouldin index (indicating more compact and well-separated clusters), its average Silhouette score is significantly lower, suggesting less clearly defined clusters compared to other methods. Additionally, DBSCAN assigns about 92.4% of the data to a single large cluster, whereas the largest cluster in k-means contains only 28.7% of the data. This observation reinforces our decision to employ k-means clustering.

As shown in Table D.1, the results from SOM are very similar to those of k-means, although k-means performs slightly better.

Table D.1
Performance metrics for different clustering algorithms.

Algoritm	Sil	DBI
k-means	0.20 (0.01)	1.55 (0.08)
Ward	0.17 (0.01)	1.59 (0.10)
DBSCAN	0.07 (0.05)	1.35 (0.10)
SOM	0.19 (0.02)	1.59 (0.10)

Notes. The average and standard deviation of the Silhouette and Davies-Bouldin scores are computed from 50 randomly generated sub-samples, each containing 20,000 observations from the original sample (after excluding missing values). The final number of clusters for each experiment was set to 5.

Appendix E. Comparison of Standard and Nested k-means

To avoid disrupting the overall clustering when increasing granularity, we use a nested version of k-means (denoted 2->2|3) instead of the standard k-means (denoted 5->0|0). These two k-means versions were run on samples, allowing observations with at least 60% non-missing values, generated yearly using progressively extended windows of data, starting from January 2, 2004, to December 31, 2013, and continuing until January 2, 2004, to December 31, 2023.

The average metrics from Table E.1 show that our approach provides comparable, though slightly lower, performance metrics, likely because the standard k-means is less constrained in its clustering. Table E.2 shows the average percentage of points in each cluster over the periods of the extended window. The lower standard deviation from the nested algorithm, except for cluster 2, suggests that the data may shift less compared to the standard k-means. This supports the argument for less disruption across time points.

We further investigate the clusters' behavior using the entire dataset. Table E.3 shows the average number of consecutive weeks banks spend in a specific cluster. Using both versions, the conclusion is that, except for the safest cluster—where banks tend to remain for longer periods—banks typically spend an average of about six months in each cluster.

Finally, Tables E.4 and E.5 present the weekly transition matrix of a Markov chain for all banks over the entire time period. The interpretation from Section 6.3 remains valid, i.e., "It is extremely rare that a bank is transferred by more than two clusters (i.e., 4 to 1 or 2 to 5, etc.), and the more likely transitions are either by one cluster or, for a large bank between clusters 2 and 4. Finally, small banks in the third cluster are more likely to become safer than riskier." By design, the standard k-means allows slightly larger movements, such as between clusters 2 and 5, which are more constrained in the nested framework.

Overall, the nested k-means version improves granularity without risking disruption to the clustering and enables smoother transitions across time points.

Table E.1
Average output metrics for standard and nested k-means.

K-means	No Obs	Inert	Sil	DBI
5->0|0	140,851	563,595	0.20 (0.003)	1.56 (0.01)
2->2|3	140,851	569,864	0.19 (0.005)	1.59 (0.01)

Notes. Average output metrics (full years 2013–2023) for standard (5→0|0) and nested (2→2|3) k-means, both with 5 clusters. Silhouette and Davies-Bouldin Index values are shown with standard deviations in parentheses.

Table E.2
Average percentages of data per cluster for standard and nested k-means.

K-means	No Obs	1	2	3	4	5
5->0|0	140,851	4.5% (0.4%)	17.4% (0.6%)	23.1% (0.8%)	21.7% (1.5%)	33.3% (2.1%)
2->2|3	140,851	4.3% (0.3%)	16.5% (1.2%)	22.9% (0.5%)	22.8% (1.2%)	33.4% (2.1%)

Notes. Average percentages of data per cluster (full years 2013–2023) for standard (5→0|0) and nested (2→2|3) k-means, both with 5 clusters. Standard deviations (in parentheses) indicate that, except for cluster 2, the data appears to be less volatile using the nested k-means.

Table E.3
Average number of consecutive weeks banks spend in each cluster using the standard and nested versions of k-means.

Cluster	1	2	3	4	5
5->0|0	23.7 (37.7)	20.9 (32.0)	27.3 (68.3)	22.8 (50.5)	48.1 (95.9)
2->2|3	23.1 (37.6)	20.8 (29.9)	25.5 (67.5)	19.7 (41.1)	48.7 (98.3)

Table E.4
Weekly transition matrix of clusters based on the standard k-means.

5->0|0	1	2	3	4	5
1	95.991%	3.708%	0.263%	0.013%	0.025%
2	0.848%	95.242%	1.253%	2.036%	0.621%
3	0.043%	0.945%	96.497%	1.648%	0.867%
4	0.009%	1.356%	1.837%	95.779%	1.019%
5	0.007%	0.506%	0.695%	0.769%	98.024%

Table E.5
Weekly transition matrix of clusters based on the nested k-means.

2->2|3	1	2	3	4	5
1	95.929%	3.87%	0.175%	0.013%	0.013%
2	0.875%	95.202%	0.933%	2.674%	0.316%
3	0.048%	0.7%	96.226%	2.061%	0.965%
4	0.004%	1.788%	1.972%	95.101%	1.135%
5	0.003%	0.328%	0.718%	0.904%	98.047%

Appendix F. Flowchart of the tool’s steps

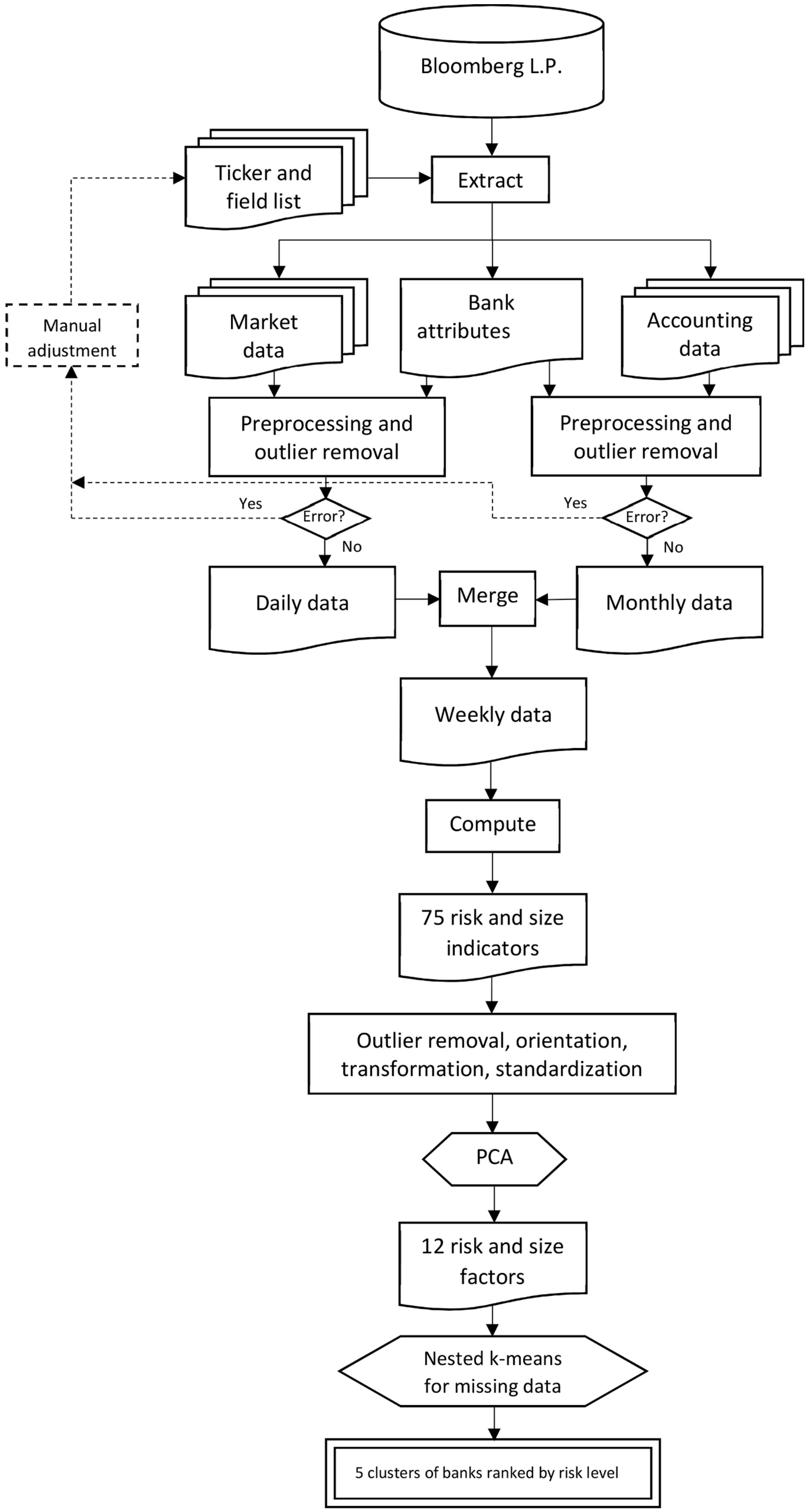


Fig. F.1. Flowchart of the tool for ranking listed banks worldwide.
Notes. The solid arrows and shapes represent what occurs each time the tool is launched. The dashed arrows and shape represent actions that must be handled manually in case of an error. For instance, if the provider modifies a ticker, we need to adjust it in the input ticker and field list. The magnetic disk shape represents Bloomberg L.P., the data provider. The document and multidocument shapes represent single or multiple files containing data, such as the list of tickers and fields to be extracted, the list of financial and accounting values extracted, and the file containing the 75 computed indicators. Rectangles represent actions, like extracting the data or computing the indicators. The elongated hexagons indicate when algorithms are used. The diamond shape represents a test: the tool checks for errors—if none

are found, the process continues; otherwise, we must manually update the input files. Finally, the nested rectangles represent the final output: the clusters of listed banks worldwide, ranked by risk.

Appendix G. Visualization of clusters with t-SNE

The t-distributed Stochastic Neighbor Embedding (t-SNE) is a machine learning algorithm used for visualizing high-dimensional data. It reduces data, here, into two dimensions, preserving the local structure and relationships between nearby points. The algorithm converts high-dimensional distances into conditional probabilities that represent similarities. It then minimizes the divergence between these probabilities in the high-dimensional and low-dimensional spaces.

Here, the t-SNE provides a representation of the clusters that helps us understand the structure of the data points within each cluster. Instead of only showing the number of banks in each cluster over time (cf. Figs. 1 and 2), these visualizations depict the evolution of data instances and centroids within clusters at several time points (cf. Fig. G.1 below). We use an extending window approach, with the first figure displaying clusters based on data from January 2, 2004, to December 31, 2013. Subsequently, each figure extends the window by one year until the last figure displaying clusters based on data from January 2, 2004, to December 31, 2023. These visualizations offer insights into the evolution and stability of each cluster.

Throughout the entire period, the riskiest clusters remain unchanged, as does the safest cluster. Only clusters 3 and 4 appear to swap positions along the x-axis during the COVID-19 period (2020–2021). This slight instability could be mitigated by using 4 clusters instead of 5. However, to remain consistent with most supervisory ratings—such as CAMELS ratings or BOPEC ratings in the U.S., which rate banks from 1 to 5—we maintain 5 clusters.

The acronym CAMELS stands for six key components: Capital adequacy, Asset quality, Management quality, Earnings, Liquidity, and Sensitivity to market risk. The acronym BOPEC stands for the following components: Bank subsidiaries, Other non-bank subsidiaries, Parent company, Earnings, and Capital adequacy. Both are confidential rating systems used by U.S. regulators to evaluate the health and stability of financial institutions, but they are designed for different types of entities and emphasize different aspects of operations. CAMELS focuses on the health of individual banks, evaluating internal operations such as asset quality, management, and liquidity. BOPEC, on the other hand, assesses the overall condition of bank holding companies, including their bank and non-bank subsidiaries, as well as the parent company itself. BOPEC takes a broader view of the performance of all entities under a holding company, including its financial condition and capital structure.

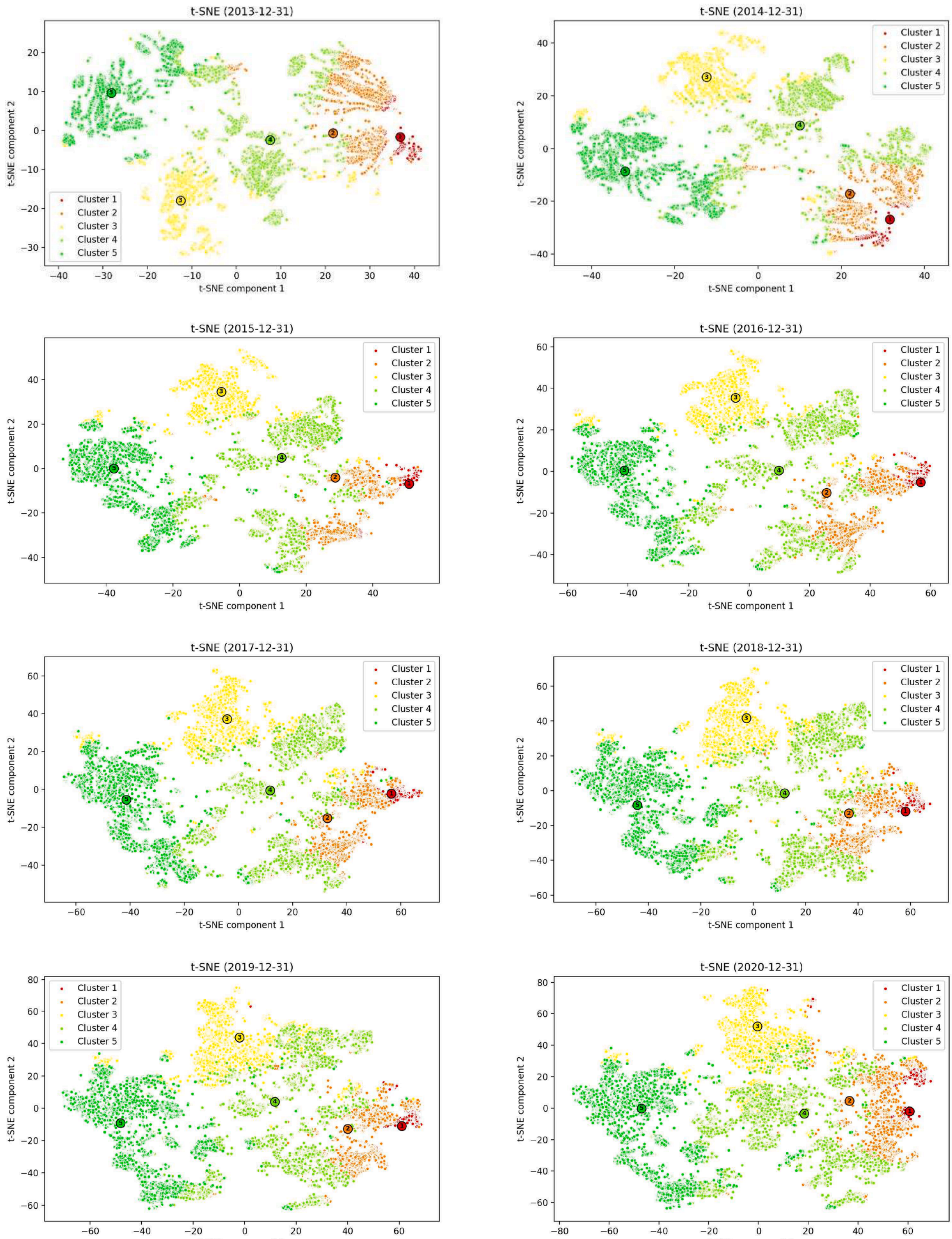


Fig. G.1. t-SNE clustering visualizations ($k = 2 \rightarrow 2|3$).
Notes. These graphs show the t-SNE outputs of our clustering over time, from full years 2013 to 2023. The centroids are indicated by their numbered labels.

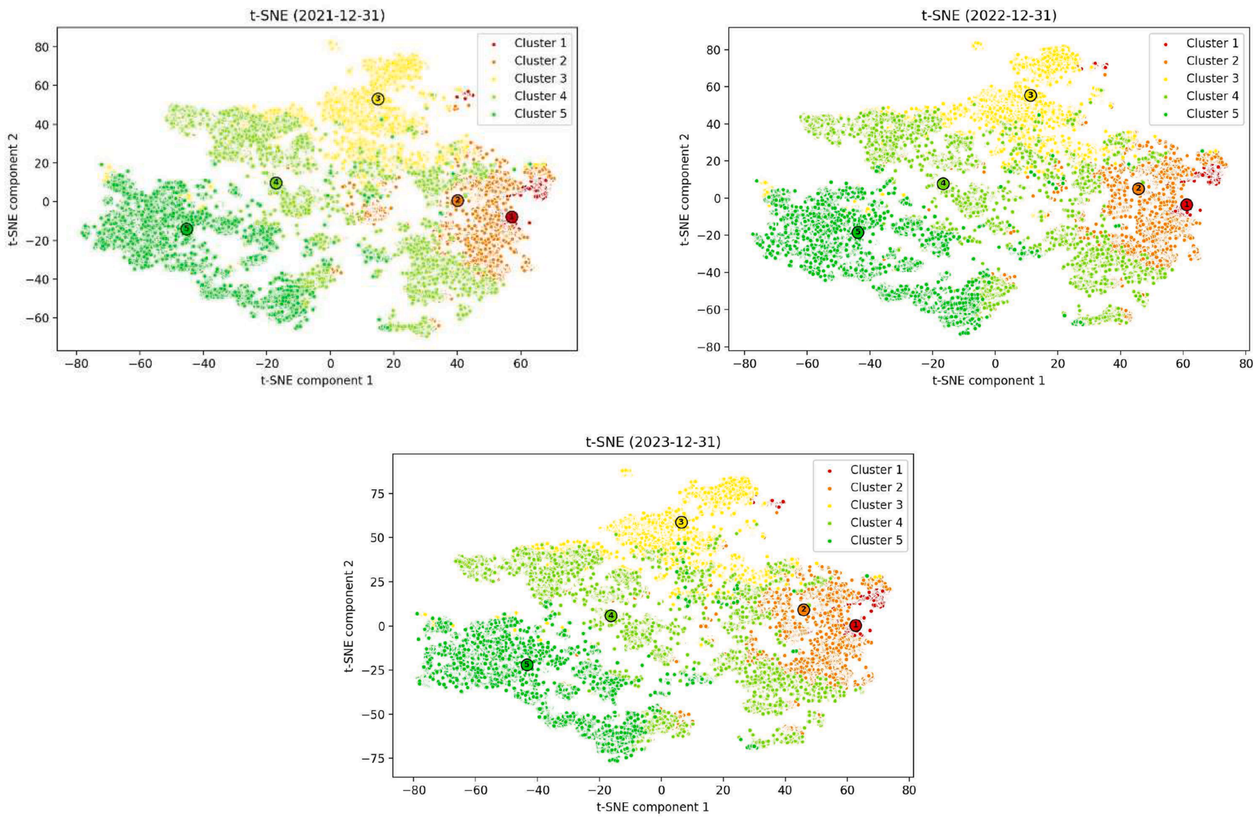


Fig. G.1. (*continued*).

Appendix H. Second principal components

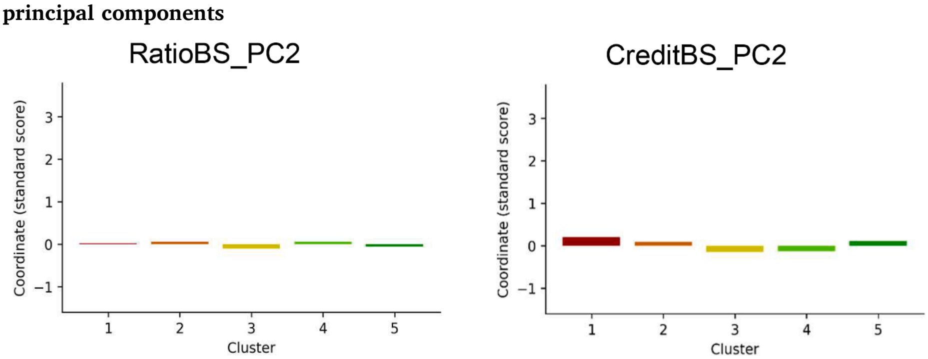


Fig. H.1. Second principal components centroids.
Notes. These bar charts display the standardized coordinates of the clusters' centroids ordered from the riskiest (red, lhs) to the safest (green, rhs) per individual risk factors. The results notably emphasize the riskiest banks' contribution to the first cluster.

Appendix I. Influential factors

Using softmax regression, we identify the factors that influence cluster formation. Despite our efforts to address multicollinearity, the bank risk monitoring tool requires the inclusion of risk factors that may still exhibit some degree of correlation. Therefore, instead of directly interpreting the regression model coefficients, we evaluate variable importance using permutation importance. This model-agnostic method measures variable significance by assessing the impact on model performance when each variable's values are randomly shuffled. The importance score reflects the drop in accuracy, showing how much the feature's information affects the model.

Fig. I.1 illustrates that SRISK is the most significant contributor to the clustering process. When this variable is shuffled, the model's accuracy decreases by 31.6% compared to the initial model. The next six factors, ranging from Size to IndivMktRisk, result in accuracy drops between 13.9% and 7.9%. Notably, the four risk factors, namely BetaCorrel, CreditMkt, SysRisk, and IndivMktRisk, collectively lead to a combined accuracy drop of 41.7%, underscoring their substantial contribution to the overall clustering.

A strength of the nested approach is that it allows for more granular information. Therefore, we further investigate by computing the permutation importance at each step of the process. Initially, the sample is split into two clusters (cf. Fig. I.2.0), with the riskiest cluster further divided into two sub-clusters (cf. Fig. I.2.1) and the safest cluster into three sub-clusters (cf. Fig. I.2.2). Fig. I.2.0 shows that the initial split is primarily influenced by SysRisk and CreditMkt, both contributing to an approximate 8.75% drop in accuracy, followed closely by IndivMktRisk, which results in a 6.4%

decrease. Fig. I.1 highlights SRISK as the most important feature overall. Fig. I.2.2 shows that SRISK is particularly critical for the split within the safest cluster, which contains the majority of observations, leading to a 38.7% drop in accuracy. Additionally, six other factors contribute to accuracy drops, ranging from 16.1% to 1.6%. The main contributors to the split in the riskiest cluster are the individual risk measures, CreditMkt (11.3%) and IndivMktRisk (6.9%). Similar conclusions are reached using permutation importance on random forest classification, and these results can be provided upon request to the authors.

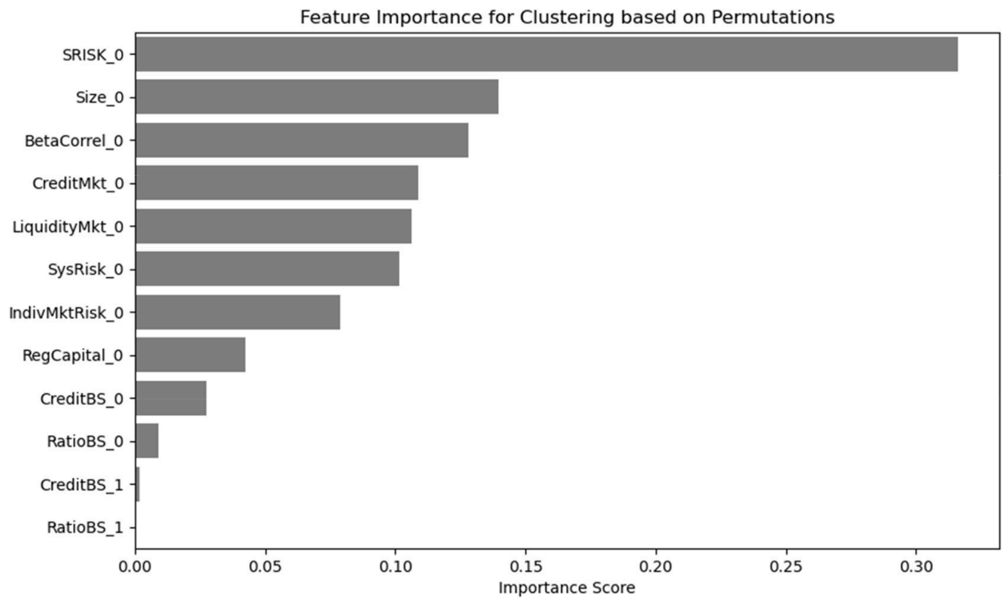


Fig. I.1. Permutation importance (2 → 2|3).
Notes. Permutation importance using softmax regression for assessing the features' impact on the 2 → 2|3 splits, based on the sub-sample. Note: The importance scores represent the drop in accuracy when each feature's values are shuffled. Higher scores indicate a greater impact on the model's performance, highlighting the most influential features in determining the clusters.

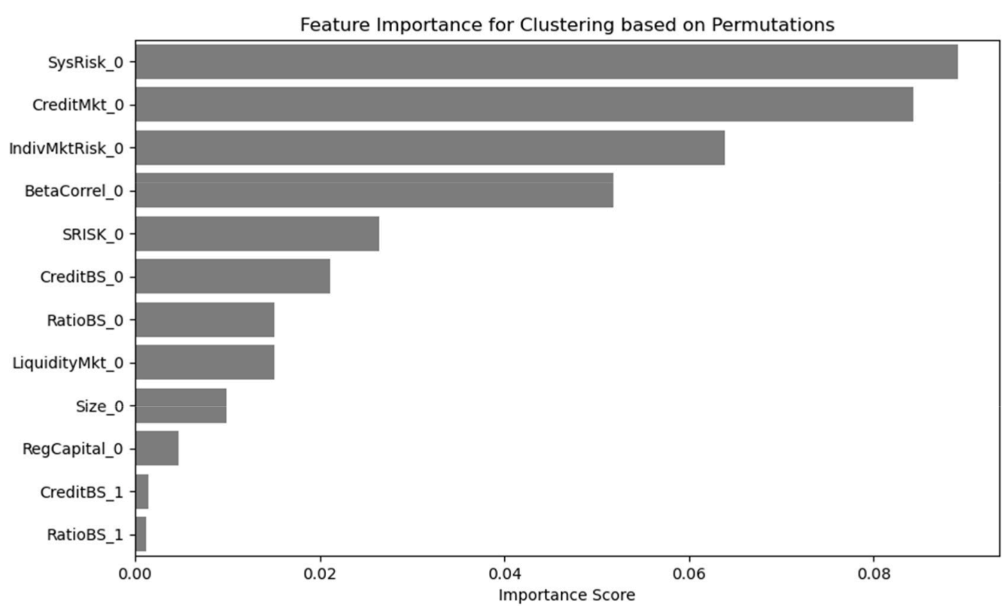


Fig. I.2.0. Permutation importance (2 → •|•).
Notes. Permutation importance using softmax regression for assessing the features' impact on the two-cluster split (2 → •|•), based on the entire sample.

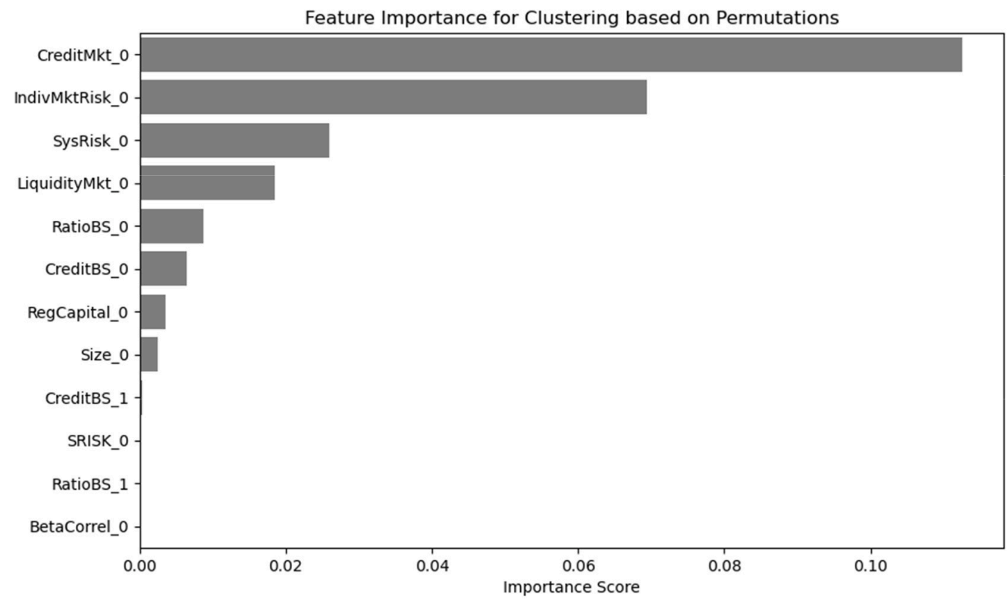


Fig. I.2.1. Permutation importance (2 → 2|•).
Notes. Permutation importance using softmax regression for assessing the features' impact on the two-cluster split, based on the sub-sample belonging to the riskiest cluster after the initial split into two clusters (2 → 2|•).

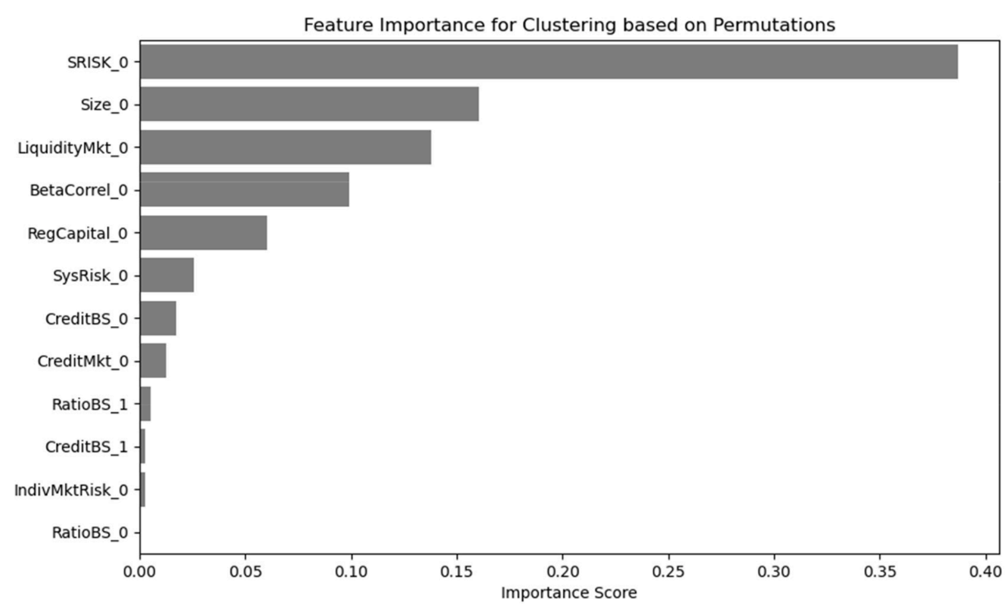


Fig. I.2.2. Permutation importance (2 → •|3).
Notes. Permutation importance using softmax regression for assessing the features' impact on the three-cluster split, based on the sub-sample belonging to the safest cluster after the initial split into two clusters (2 → •|3).

Appendix J. Illustrations with major American banks

As an illustration, we graphically compare two major American banks, the bankrupted Lehman Brothers and JPMorgan Chase, with their respective five-year CDS (Figs. J.2 and J.4). Note that in our tool, traded 5y CDS are not set as inputs.

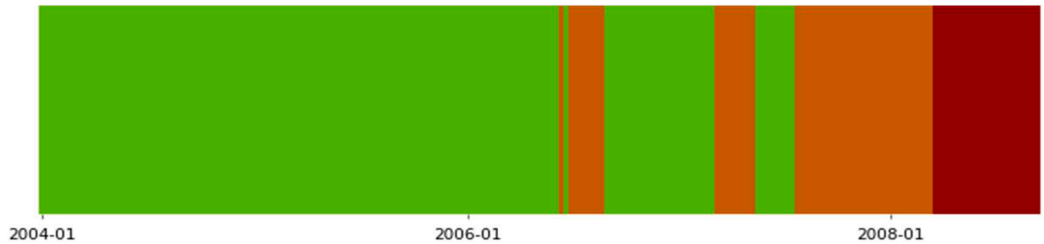


Fig. J.1. Lehman Brothers output.
Notes. This graph presents in which cluster Lehman Brothers belongs through time. It notably highlights its presence in the worst cluster as it gets closer to bankruptcy.

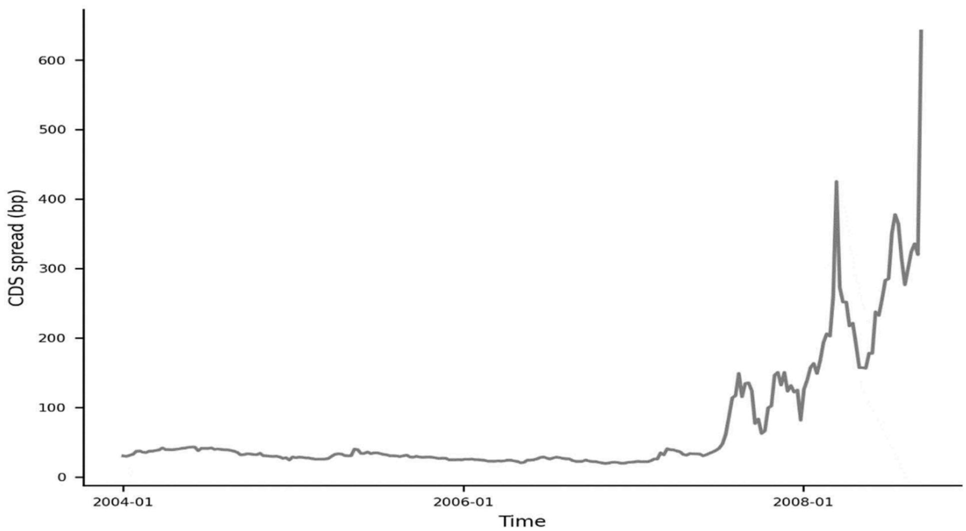


Fig. J.2. Lehman Brothers 5y CDS.
Notes. This graph presents the weekly prices of Lehman Brothers 5y CDS during the period January 2, 2004 to September 12, 2008, matching the dates of Fig. J.1.

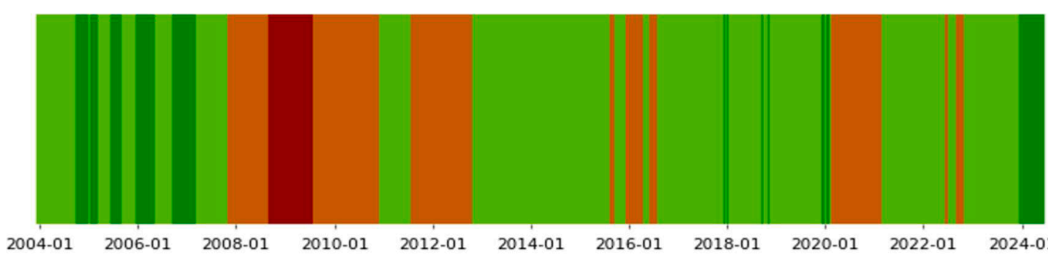


Fig. J.3. JPMorgan Chase output.
Notes. This graph presents in which cluster JPMorgan Chase belongs through time. It notably highlights its presence in the worst cluster during the GFC.

Fig. J.4. JPMorgan Chase 5y CDS.
Notes. This graph presents the weekly prices of JPMorgan Chase 5y CDS during the period January 2, 2004 to May 31, 2024, matching the dates of Fig. J.3.

Lehman Brothers' collapsing period closely linked to the subprime mortgage crisis clearly appears on both graphs. It highlights the path to bankruptcy in 2008.

JPMorgan Chase reached its riskiest moments right after the Lehman Brothers failure, it is marked by the red period from our clustering, matching the 5y CDS highest values. Other less risky periods arise around 2012 and 2020, corresponding as well to the derivative's values.

Appendix K. Focus on G-SIBs

Although more than 70% of the Global Systemically Important Banks (G-SIBs, composed of thirty banks, chosen from the November 11th, 2020 list of G-SIBs (https://www.fsb.org/2020/11/2020-list-of-global-systemically-important-banks-g-sibs/). Groupe BPCE not being listed, Natixis is set as proxy.) were in the worst cluster for nine months during the GFC, they were less than 30% during the European debt crisis. During the COVID-19 pandemic, around 40% of G-SIBs were in the worst cluster for a month before quickly dropping to below 10% (cf. Fig. K.1). The last spike in the riskiest cluster is associated with the difficulties encountered by the G-SIB Credit Suisse that was ultimately taken over by UBS. But this conclusion can be challenged by the fact that the number of banks in the first two clusters remains stable during crisis periods. A phenomenon likely related to the banks' size.

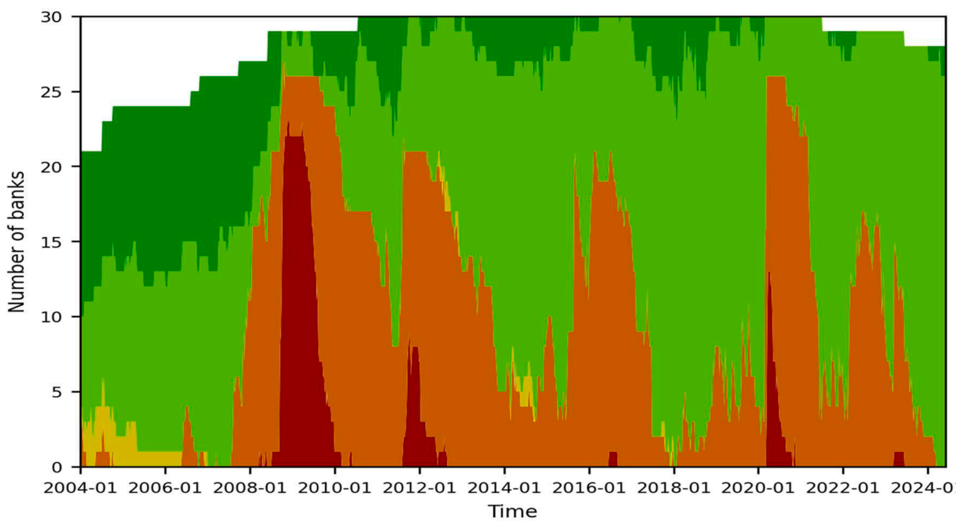


Fig. K.1. G-SIBs' clustering timeline ($k = 2 \rightarrow 2|3$).
Notes. This graph shows how Global Systemically Important Banks are clustered through time. The results notably highlight crisis periods with higher number of banks within the riskiest cluster (red).

Appendix L. Related websites

In this section, we present the websites that inspired our website (http://debtequity.net). The Volatility Laboratory (V-Lab, https://vlab.stern.nyu.edu), developed by the Volatility Institute, provides very complete and detailed information on various aspects of the risks with the following sections: volatility, correlation, systemic risk, liquidity, fixed income, long-run Value-at-Risk, and climate risk. All of their measures employ complex modeling that is closely tied to conditional volatilities. Given their goal of ranking banks according to each indicator, this approach is likely to enhance the accuracy of their measures. Although our indicators show similar values, our calculations are simpler as we mitigate potential noise by aggregating them into representative factors, upon which the clustering is performed. In terms of systemic risks, both V-Lab and our tool assess the impact of market movements on bank outcomes. However, our analysis also considers the reciprocal impact of banks on the market. Considerably, our tool is designed to focus on the banking sector, utilizing a sample of 256 listed banks from 2004 onwards. In comparison, V-Lab's approach involves ranking 947 companies across the entire financial sector since 2000. In summary, V-Lab offers an accurate financial ranking based on a limited number of systemic indicators, whereas our tool involves clustering banks based on various aspects of risk.

Other websites offer similar information, including that of the non-profit credit research initiative conducted at the Risk Management Institute of the National University of Singapore (https://nuscri.org). This initiative publicly provides two metrics, the Corporate Vulnerability Index (CVI) and their Systemically Important Financial Institution ranking (CriSIFI). Contrary to our tool, the CVI measures do not aim to cluster banks per risk level but rather assess an aggregate probability of default per country. The CriSIFI metrics are used to rank banks and insurers in terms of systemic risks that are determined by their size and interconnectedness. It should be noted that while the Risk Management Institute does not display the values of the indicators, the rankings are updated on a monthly basis and include a coverage of 2148 banks and insurance companies since 2000. In addition to the aforementioned sources, the Federal Reserve Bank of St. Louis (https://fred.stlouisfed.org) provides solvency and risk measures related to banks, among numerous economic indicators. For instance, similar to our tool, they calculate banks' non-performing loans to Gross Loans and Z-scores. However, their approach involves aggregated measures at the country level, whereas our analysis also provides each indicator at the individual bank level.

Appendix M. Tool limitations and rationale

This research provides regulators and practitioners with a valuable dynamic tool to rank banks based on risks, though certain limitations should be considered. We present these limitations in process order: data preparation and feature engineering, PCA application, and the use of nested k-means.

For the reasons presented in Section 3.2, the tool does not retain memory of previous runs. Otherwise, this feature could have enhanced data validation processes and ensured consistency with data from prior analyses. Furthermore, the audit process for each run is determined by parameters that we have established. While these parameters have been rigorously stress-tested, they may still be suboptimal. This also applies to the parametrization of the outlier removal algorithm. Additionally, outliers can occur randomly across frequency datasets, preventing the selection of a definitive reference (cf. Section 3.3). Therefore, we apply the standard criterion of averaging the preceding and following values, though some uncertainty remains.

A second outlier removal step is applied to the indicators using the interquartile range with a unique and highly inclusive multiplier, as the tool aims to retain extreme values that are not considered outliers, such as balance sheet data from Chinese banks. Alternatively, a more complex solution could use different multipliers to distinguish between balance sheet indicators and others.

Although applying k-means directly to the indicators produces similar results, the indicators are grouped using PCA, with the groupings based on economic rationale. Nevertheless, alternative combinations of indicators could be suggested. The thresholds we set for both statistics—Cronbach's alpha and explained variance—may be considered somewhat low, though it is supported by the literature.

As explained in Section 5.2, the setting of the two hyperparameters in k-means—the number of clusters and the choice of initial centroids—may be subject to criticism. However, we employed robust methods to minimize the risk of suboptimal outcomes. The k-means algorithm is adjusted to handle missing values, and it could be argued that a threshold other than 60% might have been more appropriate. Additionally, we opted for a nested version of k-means to increase granularity without disrupting overall clustering. This choice is also subject to debate, as standard k-means, by design, yields slightly better performance metrics (see Appendix E). Finally, an alternative method for assigning coherent labels to clusters could be suggested.

To a lesser extent, some risks could slow the process, such as increasing computational demands as the dataset grows and limitations from data providers. These issues could be mitigated by using more powerful computers and, respectively, switching to a different financial data provider.

References

Acerbi, C., & Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking and Finance, 26*, 1487–1503.

Acharya, V., Engle, R., & Richardson, M. (2012). Capital shortfall: a new approach to ranking and regulating systemic risks. *American Economic Review, 102*, 59–64.

Acharya, V. V., Pedersen, L. H., Philippon, T., & Richardson, M. (2017). Measuring systemic risk. *Review of Financial Studies, 30*, 2–47.

Adrian, T., & Brunnermeier, M. K. (2016). CoVaR. *American Economic Review, 106*, 1705–1741.

Ardekani Mahdavi, A., Distinguin, I., & Tarazi, A. (2020). Do banks change their liquidity ratios based on network characteristics? *European Journal of Operational Research, 285*, 789–803.

Banbula, P., & Iwanicz-Drozdowska, M. (2016). The systemic importance of banks - name and shame seems to work. *Finance Research Letters, 18*, 297–301.

Barrieu, P., & Scandolo, G. (2015). Assessing financial model risk. *European Journal of Operational Research, 242*, 546–556.

Barth, J. R., & Wihlborg, C. (2017). Too big to fail: Measures, remedies, and consequences for efficiency and stability. *Financial Markets, Institutions & Instruments, 26*, 175–245.

Barth, J. R., & Wihlborg, C. (2016). Too big to fail and too big to save: dilemmas for banking reform. *National Institute Economic Review, 235*, 27–39.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*, 1798–1828.

Benoit, S., Colletaz, G., Hurlin, C., & Pérignon, C. (2013). *A theoretical and empirical comparison of systemic risk measures*. HEC Paris Res. Pap.

Benoit, S., Colliard, J.-E., Hurlin, C., & Pérignon, C. (2016). Where the risks lie: a survey on systemic risk. *Review of Finance, 21*, 109–152.

Borri, N., & Giorgio, G.di (2021). Systemic risk and the COVID challenge in the european banking sector. *Journal of Banking and Finance, 140*, 1–13.

Boudoukh, J., Richardson, M., & Whitelaw, R. F. (1998). The best of both worlds: A hybrid approach to calculating Value at Risk. *Risk, 11*, 64–67.

Boyd, J. H., Graham, S. L., & Hewitt, R. S. (1993). Bank holding company mergers with nonbank financial firms: effects on the risk of failure. *Journal of Banking and Finance, 17*, 43–63.

Brownlees, C., & Engle, R. F. (2016). SRISK: A Conditional Capital Shortfall Measure of Systemic Risk. *Review of Financial Studies, 30*, 48–79.

Calomiris, C. W., & Khan, U. (2015). An Assessment of TARP Assistance to Financial Institutions. *J. Econ. Perspect., 29*, 53–80.

Cernohorska, L. (2017). The efficiency of selected banking sectors in the European Union. *Financial Assets and Investing, 8*, 5–18.

Chi, J. T., Chi, E. C., & Baraniuk, R. G. (2016). k-POD: a method for k-means clustering of missing data. *American Statistician, 70*, 91–99.

Chun, S. Y., Shapiro, A., & Uryasev, S. (2012). Conditional value-at-risk and average value-at-risk: Estimation and asymptotics. *Operations Research, 60*, 739–756.

Danley, B. (2019). Forest owner objectives typologies: Instruments for each owner type or instruments for most owner types? *Forest Policy and Economics, 105*, 72–82.

Dardac, N., & Boitan, I. A. (2009). A cluster analysis approach for banks' risk profile: The romanian evidence. *European Research Studies Journal, 12*, 109–118.

Dardac, N., & Giba, A. (2011). Systemic financial crisis: A cluster analysis. *European Research Studies Journal, 14*, 53–64.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 1*, 224–227.

deJonghe, O. (2010). Back to the basics in banking? A micro-analysis of banking system stability. *Journal of Financial Intermediation, 19*, 387–417.

Di Lascio, F. M. L., Giammusso, D., & Puccetti, G. (2018). A clustering approach and a rule of thumb for risk aggregation. *Journal of Banking and Finance, 96*, 236–248. https://doi.org/10.1016/j.jbankfin.2018.07.002

Doumpos, M., Hasan, I., & Pasiouras, F. (2017). Bank overall financial strength: Islamic versus conventional banks. *Economic Modelling, 64*, 513–523.

Doumpos, M., Zopounidis, C., Gounopoulos, D., Platanakis, E., & Zhang, W. (2023). Operational research and artificial intelligence methods in banking. *European Journal of Operational Research, 306*, 1–16.

Durand, P., & Le Quang, G. (2022). Banks to basics! Why banking regulation should focus on equity. *European Journal of Operational Research, 301*, 349–372.

Elyasiani, E., & Jia, J.(Jane) (2019). Relative Performance and systemic risk contributions of small and large banks during the financial crisis. *Quarterly Review of Economics and Finance, 74*, 220–241.

Engle, R. F., & Ruan, T. (2019). Measuring the probability of a financial crisis. *The Proceedings of the National Academy of Sciences PNAS, 116*, 18341–18346.

Ercan, H., & Sayaseng, S. (2016). The cluster analysis of the banking sector in Europe. *Economics and management of global value chains* (pp. 111–127).

Financial Stability Board. (2023). *Global monitoring report on non-bank financial intermediation 2023*.

Finger, C. C., Lardy, J. P., Finkelstein, V., Pan, G., Ta, T., & Tierney, J. (2002). *CreditGrades*. RiskMetrics Group.

Giglio, S., Kelly, B., & Pruitt, S. (2016). Systemic risk and the macroeconomy: An empirical evaluation. *Journal of Financial Economics, 119*, 457–471.

Hair, J. F., Babin, B. J., Black, W. C., & Anderson Rolph, E. (2019). *Multivariate data analysis*. Cengage.

Hannan, T., & Hanweck, G. (1988). Bank insolvency risk and the market for large certificates of deposit. *Journal of Money, Credit and Banking, 20*, 203–211.

Hanson, S. G., Ivashina, V., Nicolae, L., Stein, J. C., Sunderam, A., & Tarullo, D. K. (2024). The evolution of banking in the 21st century: Evidence and regulatory implications. In *Brookings Paper on Economic Activity. Presented at the BPEA Conference*.

Hartigan, J. A. (1975). *Clustering algorithms*. Hoboken: John Wiley&Sons.

Hoga, Y., & Demetrescu, M. (2022). Monitoring value-at-risk and expected shortfall forecasts. *Management Science*, 1–18. In press.

Hué, S., Lucotte, Y., & Tokpavi, S. (2019). Measuring network systemic risk contributions: A leave-one-out approach. *Journal of Economic Dynamics and Control, 100*, 86–114.

Hull, J., & White, A. (1998). Incorporating volatility updating into the historical simulation method for VaR. *Journal of Risk, 1*, 5–19.

Idier, J., Lamé, G., & Mésonnier, J.-S. (2014). How useful is the Marginal Expected Shortfall for the measurement of systemic exposure? A practical assessment. *Journal of Banking and Finance, 47*, 134–146.

John, J. A., & Draper, N. R. (1980). An Alternative Family of Transformations. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 29*, 190–197.

Jorion, P. (2001). *Value at risk: the new benchmark for managing financial risk* (3rd ed.).

Kanno, M. (2015). Assessing systemic risk using interbank exposures in the global banking system. *Journal of Financial Stability, 20*, 105–130.

Kleinow, J., Moreira, F., Strobl, S., & Vähämaa, S. (2017). Measuring systemic risk: A comparison of alternative market-based approaches. *Finance Research Letters, 21*, 40–46.

Knotek, P. (2014). Banking sectors in EMU - cluster analysis. *European Scientific Journal, 10*.

Kristóf, T., & Virág, M. (2022). EU-27 bank failure prediction with C5.0 decision trees and deep learning neural networks. *Research in International Business and Finance, 61*, 2–17.

Laeven, L., Ratnovski, L., & Tong, H. (2016). Bank size, capital, and systemic risk: Some international evidence. *Journal of Banking and Finance, 69*, S24–S34.

Lagasio, V., & Quaranta, A. G. (2022). Cluster analysis of bank business models: The connection with performance, efficiency and risk. *Finance Research Letters, 47*, 1–8.

Le, H. H., & Viviani, J.-L. (2018). Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios. *Research in International Business and Finance, 44*, 16–25.

Lepetit, L., Nys, E., Rous, P., & Tarazi, A. (2008). Bank income structure and risk: An empirical analysis of European banks. *Journal of Banking and Finance, 32*, 1452–1467.
Lepetit, L., & Strobel, F. (2013). Bank insolvency risk and time-varying Z-score measures. *Journal of International Financial Markets, Institutions and Money, 25*, 73–87.
Leung, M., Li, Y., Pantelous, A. A., & Vigne, S. A. (2021). Bayesian Value-at-Risk backtesting: The case of annuity pricing. *European Journal of Operational Research, 293*, 786–801.
Liu, X., Yan, X., & Zhang, K. (2024). Kernel quantile estimators for nested simulation with application to portfolio value-at-risk measurement. *European Journal of Operational Research, 312*, 1168–1177.
Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*, 129–137.
López-Espinosa, G., Moreno, A., Rubia, A., & Valderrama, L. (2012). Short-term wholesale funding and systemic risk: A global CoVaR approach. *Journal of Banking and Finance, 36*, 3150–3162.
López-Espinosa, G., Rubia, A., Valderrama, L., & Antón, M. (2013). Good for one, bad for all: Determinants of individual versus systemic risk. *Journal of Financial Stability, 9*, 287–299.
MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *, 1. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297).
Mare, D. S., Moreira, F., & Rossi, R. (2017). Nonstationary Z-Score measures. *European Journal of Operational Research, 260*, 348–358.
Meng, X., & Taylor, J. W. (2020). Estimating value-at-risk and expected shortfall using the intraday low and range data. *European Journal of Operational Research, 280*, 191–202.
Mercadier, M., & Lardy, J.-P. (2019). Credit spread approximation and improvement using random forest regression. *European Journal of Operational Research, 277*, 351–365.
Mercadier, M., & Strobel, F. (2021). A one-sided Vysochanskii-Petunin inequality with financial applications. *European Journal of Operational Research, 295*, 374–377.
Metrick, A., & Schmelzing, P. (2023). The March 2023 bank interventions in long-run context – Silicon Valley Bank and beyond. *Work*.
Moldovan, D., & Mutu, S. (2015). A cluster analysis on the default determinants in the European banking sector. *Business Information Systems Workshops BIS, 228*, 64–74, 2015.
Negnevitsky, M. (2017). Identification of failing banks using Clustering with self-organising neural networks. *Procedia Computer Science, 108*, 1327–1333.
Niedzielski, T., Jurecka, M., Stec, M., Wieczorek, M., & Mizinski, B. (2017). The nested k-means method: A new approach for detecting lost persons in aerial images acquired by unmanned aerial vehicles. *Journal of Field Robotics, 34*, 1395–1406.
Oliveira, A. F.de, & Nicoletti, M.do C. (2019). Favoring the k-Means algorithm with initialization methods. In *International Conference on Intelligent Systems Design and Applications* (pp. 21–31).
Pamen Nyola, A., Sauviat, A., Tarazi, A., & Danisman, G. O. (2021). How organizational and geographic complexity influence performance: Evidence from European banks. *Journal of Financial Stability, 55*, 1–27.
Penikas, H. (2015). History of banking regulation as developed by the basel committee on banking supervision 1974-2014. *Estabilidad Financiera/Banco de España, 28*, 9–47.
Pesenti, S. M., Millossovich, P., & Tsanakas, A. (2019). Reverse sensitivity testing: What does it take to break the model? *European Journal of Operational Research, 274*, 654–670.
Reboredo, J. C., & Ugolini, A. (2015). Systemic risk in European sovereign debt markets: A CoVaR-copula approach. *Journal of International Money and Finance, 51*, 214–244.
Rizwan, M. S., Ahmad, G., & Ashraf, D. (2020). Systemic risk: The impact of COVID-19. *Finance Research Letters, 36*(1–7), Article 101682.
Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.
Saghi-Zedek, N., & Tarazi, A. (2015). Excess control rights, financial crisis and bank profitability and risk. *Journal of Banking and Finance, 55*, 361–379.
Sedunov, J. (2016). What is the systemic risk exposure of financial institutions? *Journal of Financial Stability, 24*, 71–87.
Soerensen, C. K., & Gutiérrez, J. M. P. (2006). Euro Area banking sector integration using hierarchical cluster analysis techniques. In *European Central Bank: Working Paper Series* (pp. 1–42).
Steinhaus, H. (1956). Sur la division des corps matériels en parties. *Bull. Académie Pol. Sci., Classe III, IV*(12), 801–804.
Straetmans, S. T. M., Verschoor, W. F. C., & Wolff, C. C. P. (2008). Extreme US stock market fluctuations in the wake of 9/11. *Journal of Applied Econometrics, 23*, 17–42.
Taber, K. S. (2018). The use of Cronbach's Alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*, 1273–1296.
Tsagkarakis, M.-P., Doumpos, M., & Pasiouras, F. (2021). Capital shortfall: A multicriteria decision support system for the identification of weak banks. *Decision Support Systems, 145*, 1–13.
Tuan, L. Q., Lin, C.-Y., & Teng, H.-W. (2023). Machine learning methods for predicting failures of US commercial bank. *Applied Economics Letters, 31*, 1353–1359.
Varotto, S., & Zhao, L. (2018). Systemic risk and bank size. *Journal of International Money and Finance, 82*, 45–70.
Walter, J. (2019). US Bank capital regulation: history and changes since the financial crisis. *Economic Quarterly, 105*, 1–40.
Weiß, G. N. F., Bostandzic, D., & Neumann, S. (2014a). What factors drive systemic risk during international financial crises? *Journal of Banking and Finance, 41*, 78-96165–181.
Weiß, G. N. F., Neumann, S., & Bostandzic, D. (2014b). Systemic risk and bank consolidation: International evidence. *Journal of Banking and Finance, 40*, 165–181.
Xu, Q., Ding, C., Liu, J., & Luo, B. (2015). PCA-guided search for k-means. *Pattern Recognition Letters, 54*, 50–55.
Zarandi, S. A. T. P., Shahabi, V., Emami, M., Rezaee, A., & Hassanloo, M. (2014). Ranking banks using K-Means and Grey relational method. *Management Science Letters, 4*, 2319–2324.
Zhang, Q., Vallascas, F., Keasey, K., & Cai, C. X. (2015). Are market-based measures of global systemic importance of financial institutions useful to regulators and supervisors? *Journal of Money, Credit and Banking, 47*, 1403–1442.
Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and k-means techniques. *Informatics in Medicine Unlocked, 17*, 1–7.
Zou, Z., Wu, Q., Xia, Z., & Hu, T. (2023). Adjusted Rényi entropic Value-at-Risk. *European Journal of Operational Research, 306*, 255–268.