EP-JOURNALS

Available on: https://jefrr.ep-journals.org/ E-: 3051-3650 P-ISSN: 3051-3642

Journal of Economic Finance Research and Review

Volume: 02 Issue: 03 March 2026

Uncovering the Latent Structure of Liquidity Risk Using Machine Learning Clustering Algorithms: Evidence from Vietnamese Commercial Banks

Tran Thi Thuy Huyen
Thuongmai University, Hanoi, Vietnam

Hoang Thao Nhi
Thuongmai University, Hanoi, Vietnam

Pham Ngoc Khue
Thuongmai University, Hanoi, Vietnam

Nguyen Thi Lan Huong
Thuongmai University, Hanoi, Vietnam

Nguyen Thu Ha
Thuongmai University, Hanoi, Vietnam

Vu Thi Thu Huong
Thuongmai University, Hanoi, Vietnam ,Orcid https://orcid.org/0000-0002-1958-8567

Published Date: March 07, 2026 **Article DOI: 10.65150/EP-jefrr/V2E3/2026-02**

ABSTRACT: This study explores the latent structure of liquidity risk in the Vietnamese commercial banking system through the application of unsupervised machine learning clustering algorithms. Using panel data collected from the financial statements of 31 Vietnamese commercial banks over the period 2009–2024, the study constructs a set of indicators capturing banks' liquidity characteristics, profitability, and funding structure. Representative clustering techniques, including K-means clustering and Hierarchical clustering, are employed to identify groups of banks exhibiting similar levels of liquidity risk. The empirical findings reveal a relatively clear segmentation within the banking system in terms of liquidity stability, associated with differences in operational scale, business performance, and reliance on short-term funding sources. The applied clustering algorithms effectively highlight the underlying structures embedded in the data, thereby reflecting the heterogeneity of liquidity risk across the system. This study contributes to the growing literature by incorporating a machine learning–based approach into liquidity risk analysis and provides empirical evidence to support supervisory practices and prudential risk management in the banking sector.

KEYWORDS: Commercial banks, Clustering, Hierarchical clustering, K-means, Liquidity risk.
JEL Classification Codes: C58, G21, M21

INTRODUCTION

In the operations of commercial banks, liquidity risk is widely recognized as one of the core risk categories, characterized by a high degree of contagion and a close linkage to the stability of the overall financial system. By nature, banking activities are grounded in the process of maturity transformation, whereby short-term liabilities are used to finance medium- and long-term assets. While this mechanism enables banks to enhance profitability, it simultaneously exposes them to liquidity pressures when cash flows evolve unfavorably.

Compared with credit risk or market risk, which are often reflected more explicitly in profitability indicators and asset quality measures, liquidity risk is considerably more difficult to detect under normal operating conditions. However, once market confidence deteriorates, liquidity pressures can escalate rapidly and propagate across institutions within a short period of time. Evidence from the 2007–2008 global financial crisis, as well as more recent disturbances in Vietnam, demonstrates that a bank may maintain positive business performance yet still face the risk of collapse if it fails to meet its payment obligations in a timely manner, thereby generating adverse spillover effects on the broader macroeconomy.

In the Vietnamese context, most existing studies primarily rely on individual financial ratios or linear regression models to identify the determinants of liquidity risk. Although these approaches provide important empirical insights, the increasing complexity of financial data has exposed several technical limitations of traditional methods. A fundamental weakness lies in the assumption of stable and homogeneous relationships among variables across the entire sample. In reality, the Vietnamese commercial banking system exhibits substantial heterogeneity in terms of size, capital structure, and business strategy. Applying a uniform analytical framework to the entire system may obscure risk differentiation and overlook critical characteristics specific to particular groups of banks. Moreover, liquidity risk typically arises from multidimensional interactions among financial factors, where the effect of a given variable may vary depending on the bank's condition, dynamics that linear models are often unable to capture adequately.

Against this backdrop, the application of machine learning techniques, particularly unsupervised learning, has emerged as a promising avenue for uncovering hidden patterns and latent structures in data without relying on predefined labeling assumptions. Unlike conventional forecasting or causal inference models, clustering algorithms focus on identifying similarity among entities based on multiple characteristics simultaneously. This approach allows for a more comprehensive and intuitive assessment of liquidity risk across the banking system, while reducing dependence on restrictive linearity assumptions.

Motivated by these academic and practical considerations, this study aims to uncover the latent structure of liquidity risk among Vietnamese commercial banks through representative clustering algorithms. Using financial statement data from 31 banks over the period 2009–2024, the study classifies banks into distinct risk groups and identifies the financial characteristics that shape the differences among them. The findings are expected to contribute methodologically by extending the application of machine learning in risk analysis, and empirically by providing a robust foundation for managers and regulators to design risk-based supervisory policies rather than applying a uniform regulatory framework to the entire system.

The remainder of the paper is organized into five sections. Following this introduction, Section 2 reviews the theoretical foundations of liquidity risk and synthesizes prior research on the application of machine learning in financial risk management. Section 3 describes the research methodology, including data collection procedures and the implementation of clustering algorithms. Section 4 presents the empirical results and provides an in-depth discussion of the latent risk structure, including a comparative evaluation of the performance of K-means clustering and Hierarchical clustering. Section 5 concludes with key findings, policy and managerial implications, and directions for future research.

LITERATURE REVIEW

Theoretical Foundations of Liquidity Risk in Banking Operations

Within the financial system, liquidity risk is defined as the condition in which a financial institution is unable to secure sufficient funding to meet its obligations as they fall due without adversely affecting its daily operations or overall financial position (Basel Committee on Banking Supervision, 2008).

According to Duttweiler (2009, 2016), liquidity should be examined from two perspectives: natural liquidity, arising from contractual cash flow maturities, and artificial liquidity, generated through the ability to convert assets into cash prior to maturity. Liquidity risk materializes when an institution is unable to meet its payment obligations at a given point in time or is forced to obtain funding at excessive cost. This issue is intrinsically linked to the core function of commercial banks as financial intermediaries engaged in maturity transformation, whereby short-term deposits are used to finance longer-term, less liquid assets. Although this mechanism enhances profitability, it simultaneously exposes banks to liquidity pressures when cash flows deteriorate or market confidence erodes.

The central objective of liquidity risk management is to ensure that banks consistently maintain sufficient liquid assets to meet immediate financial obligations, ranging from customer withdrawals to interbank settlements. Effective management helps prevent maturity mismatches and enables business continuity even during periods of stress. Fundamentally, liquidity management involves optimizing the trade-off between liquidity and profitability. Excessive liquidity buffers generate opportunity costs, while insufficient reserves may force banks into costly emergency funding. A sound liquidity management framework identifies optimal reserve thresholds, minimizes funding costs, and maximizes returns on assets. Importantly, liquidity risk exhibits strong contagion characteristics; even minor payment delays can trigger bank runs. Hence, effective management not only safeguards reputation but also sustains depositor confidence, allowing banks to access funding at lower cost and maintain competitive advantage.

For Vietnamese commercial banks, liquidity risk management serves as a "shock absorber" against major disruptions such as the COVID-19 pandemic or periods of monetary tightening. Through stress-testing scenarios, management can anticipate adverse outcomes and proactively restructure funding sources. The sources of liquidity risk may arise from both sides of the balance sheet or from off-balance-sheet activities. Nguyen Van Tien (2010) identifies three primary drivers: maturity mismatches between assets and liabilities, sensitivity of financial assets to interest rate changes, and the pressure to meet liquidity demands flawlessly to preserve public confidence. When interest rates rise, depositors may withdraw funds in search of higher returns, while loan cash inflows may simultaneously weaken, creating dual liquidity pressures. In the context of financial integration, compliance with international standards such as Basel II and Basel III enhances operational safety, improves credit ratings, and facilitates access to international capital markets.

Measurement and Determinants of Liquidity Risk

To quantify liquidity risk for supervisory and managerial purposes, prior studies combine traditional liquidity ratios with composite measures of financial stability. The Loan-to-Deposit Ratio (LDR) is the most basic indicator, reflecting the extent to which banks utilize customer deposits to finance lending activities, while also signaling profitability and institutional credibility. It is calculated as:

$$LDR = \frac{\text{Total Customer Loans}}{\text{Total Customer Deposits}}$$

A high LDR indicates aggressive lending relative to deposit funding. While this may improve capital utilization, it reduces the bank's ability to accommodate sudden withdrawals and increases liquidity risk. Conversely, a low LDR suggests stronger liquidity but may reflect suboptimal profitability. In Vietnam, Circular No. 22/2019/TT-NHNN caps the maximum LDR at 85% for commercial banks from January 1, 2020, with the 80–85% range considered optimal for balancing profitability and systemic safety.

Beyond static ratios, the Z-score is widely employed as a composite indicator of financial stability and insolvency risk. Theoretically, the Z-score measures the number of standard deviations by which returns must decline before equity is fully depleted, resulting in insolvency. It is defined as:

Corresponding Author: Vu Thi Thu Huong

$$Z\text{-score} = \frac{ROA + CAP}{\sigma(ROA)}$$

Where, ROA denotes return on assets (net income divided by total assets), CAP represents the equity-to-asset ratio, and σ(ROA) is the standard deviation of ROA capturing earnings volatility.

A higher Z-score indicates a larger capital buffer and stronger financial stability. Because the Z-score is typically right-skewed and inversely related to risk, the transformed variable $ln_z = -\ln(Z\text{-score})$is frequently employed to reduce skewness, mitigate outlier effects, and ensure a positive association with risk (Lepetit & Strobel, 2013).

The Z-score captures liquidity risk through the interaction between profitability, capital buffers, and earnings volatility. A higher value reflects stronger financial fundamentals, reducing vulnerability to bank runs and facilitating access to interbank funding. Empirical studies by Hannan and Hanweck (1988), Cihak and Hesse (2010), and Lepetit and Strobel (2013) demonstrate that the Z-score provides superior predictive power for financial distress compared with traditional single-indicator measures, as it integrates both capital structure and operational volatility.

The variability of these liquidity indicators is influenced by a complex interplay between bank-specific characteristics and macroeconomic conditions. Internally, liquidity positions are directly shaped by liquidity ratios (e.g., liquid assets, LDR) and capital adequacy. Credit risk factors and liability structure variables, including loan loss reserves (LLR), leverage ratios, credit growth, and net interest margin (NIM), play critical roles in stabilizing cash flows (Beatty et al., 2018). Operational efficiency indicators such as return on equity (ROE), cost-to-income ratio (CIR), operating expense ratio (OER), earnings quality (e.g., receivables-to-assets ratio), and bank size further determine financial autonomy and resilience (Dietrich & Wanzenried, 2011).

Simultaneously, macroeconomic conditions exert significant influence on systemic liquidity risk. These include economic growth (GDP), inflation (CPI) (Myra, V. De L., 2020), and exogenous shocks such as financial crises (Dietrich & Wanzenried, 2011) and, more recently, the COVID-19 pandemic. Particular attention should be paid to depositor confidence indicators and the interaction between inflationary pressures and liquidity structure to assess bank resilience during periods of market stress. The multidimensional interaction among these internal and external determinants forms the empirical foundation for applying machine learning algorithms to uncover the latent structure of liquidity risk within the banking system.

Application of Machine Learning and Clustering Algorithms in Financial Risk Analysis

In the context of the digital era, the rapid expansion of Big Data has triggered a profound shift in financial risk management approaches, gradually moving from traditional statistical models to advanced Machine Learning techniques. Internationally, the application of machine learning to risk analysis has demonstrated clear advantages over classical econometric methods, particularly due to its ability to efficiently process high-dimensional data and capture complex nonlinear relationships (Valentin et al., 2025).

While supervised learning algorithms such as Random Forest and XGBoost enable the development of highly accurate early-warning models, unsupervised learning algorithms, especially clustering techniques, play a crucial role in uncovering latent risk structures without relying on predefined data labels. Specifically, methods such as K-Means clustering and Hierarchical clustering allow financial institutions to be identified and classified based on similarities in their risk profiles. This provides managers with a more multidimensional and objective perspective, rather than depending solely on isolated financial ratios (Valentin et al., 2025).

However, in the Vietnamese research context, there remains a significant gap in the application of these techniques to liquidity risk analysis. Most domestic studies primarily employ traditional econometric models, such as panel data regressions (FEM, REM), to identify determinants of liquidity risk (Nguyen Duc Trung & Tran Trong Huy, 2024; Truong Quang Thong et al, 2014). Some recent research has adopted more advanced estimation techniques such as GMM or GLS to address endogeneity issues, yet these approaches fundamentally rely on linear assumptions (Bui Dan Thanh et al., 2022; Nguyen Duc Trung & Tran Trong Huy, 2024). Although such methods are advantageous in explaining causal relationships, they exhibit notable limitations in forecasting performance and in detecting nonlinear risk patterns, particularly during periods of heightened market volatility such as the COVID-19 pandemic.

Moreover, the application of clustering algorithms to classify liquidity risk profiles remains a relatively novel and underexplored direction in Vietnam. While some studies have begun applying machine learning models such as Random Forest or XGBoost, most focus on stock index prediction or corporate financial distress in general, rather than examining the specific liquidity structure of the banking system (Truong Thi Thuy Duong & Le Hai Trung, 2023). In addition, many existing studies categorize banks based on qualitative criteria or simple asset size classifications, resulting in a lack of rigorous quantitative grouping grounded in multidimensional data structures. Therefore, implementing a study that integrates modern machine learning models with clustering algorithms to analyze liquidity risk in Vietnamese commercial banks is both necessary and timely. Such an approach helps bridge the theoretical gap and provides more effective practical risk management tools in an increasingly complex financial environment.

METHODOLOGY

Data

The primary data source is derived from publicly disclosed financial statements of commercial banks in compliance with transparency regulations. Financial indicators were directly extracted from consolidated financial statements. In addition, several research variables were constructed from relevant statistical sources to better capture operational characteristics and liquidity risk exposure.

The dataset includes variables reflecting liquidity conditions, funding structure, and operational efficiency, such as the Loan-to-Deposit Ratio (LDR), liquidity asset ratio, equity-to-total-assets ratio, Return on Equity (ROE), total asset size, and other financial stability indicators. These variables were selected to highlight differences in liquidity characteristics among banks within the research sample.

Prior to clustering analysis, the dataset was carefully reviewed and processed to remove missing observations and outliers that could distort the results. All variables were standardized to a common scale to ensure that no single variable disproportionately influenced the clustering formation process. This preprocessing procedure enhances the robustness and reliability of the clustering outcomes.

Through comprehensive data collection and processing, the study ensures adequate sample size, strong representativeness, and a relatively comprehensive reflection of the operational characteristics of the Vietnamese commercial banking system during the 2009–2024 period. This provides a solid empirical foundation for uncovering the latent structure of liquidity risk using clustering algorithms in subsequent sections.

Clustering Features

In the context of Vietnam's financial system facing increasingly complex macroeconomic fluctuations and digital transformation pressures, understanding data characteristics plays a pivotal role in constructing accurate risk prediction models. Before implementing more sophisticated machine learning algorithms, this study first focuses on analyzing the empirical data features, thereby establishing a foundation for the application of machine learning techniques in later stages.

Table 1. Input Features

No.	Variable	Calculation Formula	Description
Liquidity			
1	AR	Liquidity assets / Total assets	Reflects the proportion of highly liquid assets within total assets.
2	LDR	Loan-to-Deposit Ratio	A mandatory prudential supervision indicator used to control the extent to which mobilized funds are used for lending.
Capital Adequacy			
3	EAR	Equity / Total assets	A core measure of financial soundness.
Credit Risk and Debt Structure			
4	LLR	Loan loss provisions / Total liabilities	Reflects the bank's prudence against deterioration in asset quality.
5	LOAN_TA	Total liabilities / Total assets	Indicates the bank's dependence on external funding sources (customer deposits, interbank borrowing, debt securities issuance) to finance assets.
6	Credit_G	Annual credit growth rate	Reflects the pace of expansion in lending activities over time.
7	NIM	Net Interest Margin = Net interest income / Earning assets	Evaluates the bank's core operational efficiency.
8	ln_z	ln_z = −ln(Z-score)	Proxies the level of liquidity risk/bankruptcy risk of financial institutions.
Operational Efficiency			
9	ROE	Return on Equity = Net profit after tax / Total equity	Measures managerial performance and the profit generated per unit of shareholders' equity.
10	CIR	Cost-to-Income Ratio = Operating expenses / Operating income	Measures cost management efficiency.
11	OER	Operating Expense Ratio = Operating expenses / Revenue	Reflects the operating cost burden relative to income scale.
12	RTR	Accounts receivable / Total assets	Reflects earnings quality.
Bank Characteristics			
13	SIZE	SIZE = ln(Total assets)	Represents market share and access to capital markets.
Macroeconomic Variables			
14	GDP	GDP growth rate	GDP growth is expected to be negatively associated with risk.
15	CPI	Consumer Price Index	CPI is expected to be positively associated with risk.
16	Covid	Dummy variable: equals 1 for 2020–2021, 0 otherwise	COVID-19 is expected to increase liquidity risk.
New Variables			
17	Stress	Inflation stress index = LDR × CPI	Measures the impact of liquidity structure (LDR) under inflationary pressure.
18	Trust	Depositor confidence index = Growth in customer deposits / Total customer deposits	"Confidence" is a critical factor in preventing bank runs.
19	Liquidity	Liquidity growth rate = Net cash flow from operations / Beginning-of-year total cash	Measures the effectiveness of cash generation from core activities.
20	Cash_Flow	Net cash flow from operations / Total assets	Reflects actual cash-generating capacity per unit of invested assets.
21	AC	Asset conversion ratio = Customer loans / Investment securities	Measures asset structure between less-liquid assets and more liquid assets.

22	Interbank	Interbank funding dependence ratio = Deposits and borrowings from other credit institutions / Total liabilities	Measures reliance on interbank funding sources.

Source: Author's compilation (2025) from literature.

Procedure for Applying Clustering Algorithms to Banking Data

The research process is conducted in four main steps as follows:

Step 1: Data Preparation and Feature Construction

The study employs panel data extracted from the financial statements of 31 Vietnamese commercial banks over the period 2009–2024. The input variables are developed based on the theoretical framework of risk management, capturing multiple dimensions including liquidity conditions, capital adequacy, credit risk, operational efficiency, and institutional size.

All data processing procedures are implemented using STATA and Python, including aggregation of accounting indicators, treatment of missing values, and data standardization (Z-score normalization). Standardization is a prerequisite step to eliminate differences in measurement units, ensuring that distance-based algorithms are not biased by variables with large magnitudes.

The clustering features are described in Table 1.

Step 2: Selection and Training of Clustering Models

Based on the preprocessed and standardized dataset, the study simultaneously applies two clustering algorithms to explore different aspects of liquidity risk structure:

- **K-Means**: The K-Means algorithm is employed as a foundational clustering method, in which banks are grouped according to feature distances in a multidimensional space. The optimal number of clusters is determined using the Elbow method, based on the change in within-cluster sum of squares.
- **Hierarchical Clustering**: Hierarchical clustering is implemented to observe the nested relationships among groups of banks, thereby supporting the evaluation of consistency and the hierarchical structure of clustering outcomes.

Step 3: Model Performance Evaluation

To ensure objectivity and comprehensiveness in selecting the appropriate clustering model, the study develops a multidimensional evaluation framework. Given the nature of liquidity risk data—where no predefined class structure exists—the performance evaluation relies on internal validation metrics.

These metrics measure clustering quality based on geometric distances in the feature space, without requiring reference labels:

- **Silhouette Coefficient (S)**, proposed by Peter J. Rousseeuw (1987), measures the similarity of a data point to other points within the same cluster compared to those in neighboring clusters. The value ranges from −1 to 1. Values closer to 1 indicate well-separated and compact clusters.
- **Davies–Bouldin Index (DB)** measures the ratio of within-cluster dispersion to between-cluster separation. A lower DB index indicates better clustering performance, reflecting a balance between cluster compactness and inter-cluster separation.

Step 4: Model Selection and Cluster Characterization

Based on performance evaluation metrics and visualization results in reduced-dimensional space using Principal Component Analysis (PCA), the study conducts cross-comparisons between algorithms. This combined approach allows for a more accurate reconstruction of the intrinsic characteristics of the banking system, ensuring a balance between technical precision and practical interpretability.

Cluster naming is determined by comparing the most prominent characteristics of each cluster, analyzing directional patterns and visualization outcomes, and incorporating supplementary SHAP analysis to enhance interpretability.

RESULTS AND DISCUSSION

Determination of the Optimal Number of Clusters

To ensure objectivity and accuracy in classifying risk structures, the study applies two quantitative methods simultaneously: the Elbow method and the Silhouette coefficient.

The Elbow method is used to observe the reduction in the within-cluster sum of squared errors as the number of clusters increases, while the Silhouette coefficient evaluates the degree of separation and cohesion among cluster members.

Empirical results from the Elbow plot indicate a clear inflection point at $k = 2$, suggesting that increasing the number of clusters beyond this threshold does not significantly reduce model error. Concurrently, the Silhouette coefficient reaches its maximum average value at $k = 2$, indicating that at this level of partitioning, bank groups achieve the highest degree of risk differentiation with minimal data overlap.

Accordingly, the study selects two clusters as the optimal number for subsequent in-depth analysis.

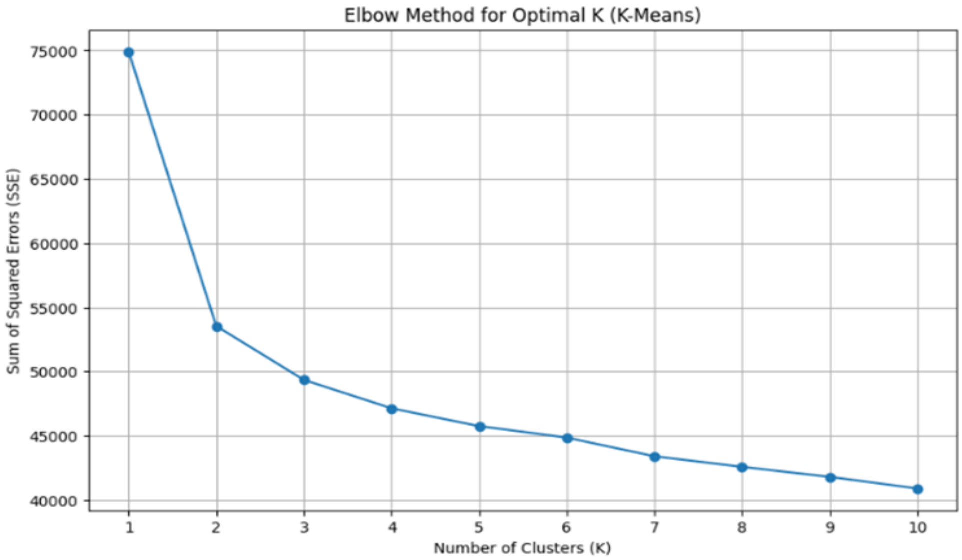

Figure 1: Determination of the Optimal Number of Clusters (k), Elbow Method
Source: Author's calculation and visualization using Python

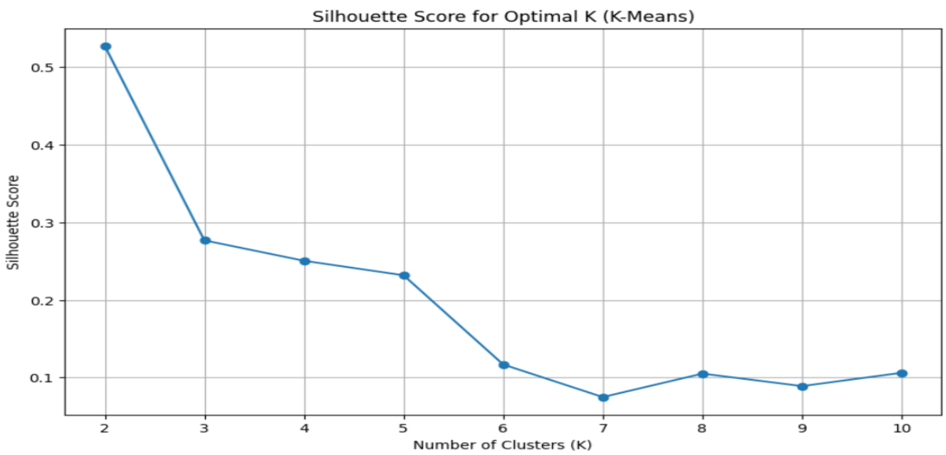

Figure 2: Variation of the Average Silhouette Score by Number of Clusters (k)
Source: Author's calculation and visualization using Python

Clustering Results Using the K-Means Algorithm

Cluster Profiling

Empirical results from the K-Means algorithm divide the research sample into two groups of banks with clearly contrasting risk profiles and operational strategies. Specifically:

- **Cluster 0**: This group represents banks with a "Safe/Low-Risk" profile, operating under sustainable standards. The most prominent characteristic is a very high market credibility index (Trust), with an average value of 8.10, reflecting strong confidence from customers and counterparties. The financial structure of this group remains well balanced, with a Loan-to-Deposit Ratio (LDR) of 82.88, ensuring harmony between credit profitability and liquidity safety. Notably, reliance on the interbank market (Interbank) is maintained at a very low level (0.098), indicating high funding autonomy and limited exposure to external shocks. Supported by a sound financial foundation, Cluster 0 achieves solid performance, with an average Return on Equity (ROE) of 0.17.
- **Cluster 1**: Identified as the "High-Risk" group, this cluster is characterized by signs of aggressive growth and an unstable funding structure. Although it records a rapid credit growth rate (Credit_G) of 3.98, its credibility level is extremely low, with a Trust index of only 0.22. Limited reputation constrains the ability of these banks to attract stable retail deposits, forcing them to rely heavily on short-term interbank funding, as reflected by a sharp increase in the Interbank ratio to 1.31. While the LDR is reported at a low level (1.28), excessive dependence on volatile external funding sources results in persistently high potential liquidity risk.

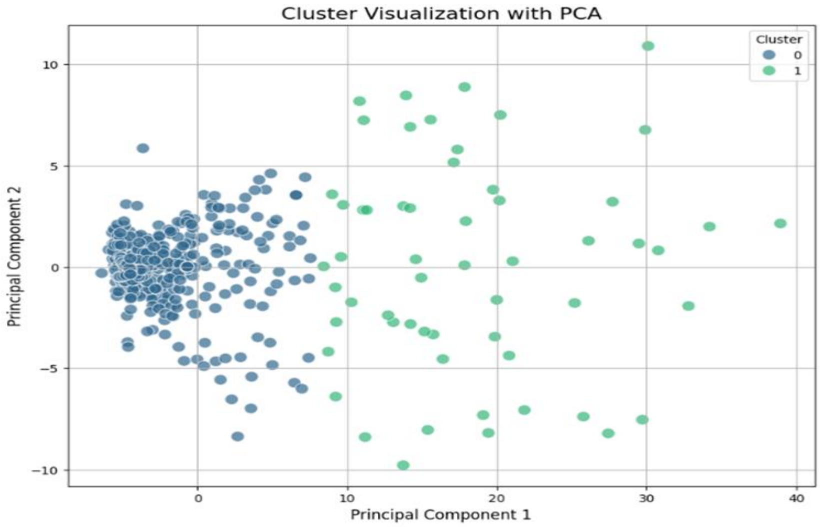

Figure 3: K-Means Clustering Results
Source: Author's calculation and visualization using Python

In addition, visualization results through Boxplot charts provide empirical evidence of profound structural divergence between the two bank groups.
For the Interbank variable (interbank funding dependence), Cluster 0 (Safe Group) shows strong data concentration near zero, demonstrating funding autonomy. In contrast, Cluster 1 (Risk Group) exhibits wide dispersion with a high median, reflecting structural instability and significant reliance on short-term borrowed funds.
This contrast is further reinforced by the Trust and Credit_G variables. Cluster 0 maintains consistently high credibility levels. Conversely, Cluster 1 not only experiences severe reputational deterioration but also displays multiple upper outliers in the Credit Growth distribution. The presence of these outliers implies a high-risk asset expansion strategy, suggesting a willingness to trade capital safety for rapid growth.

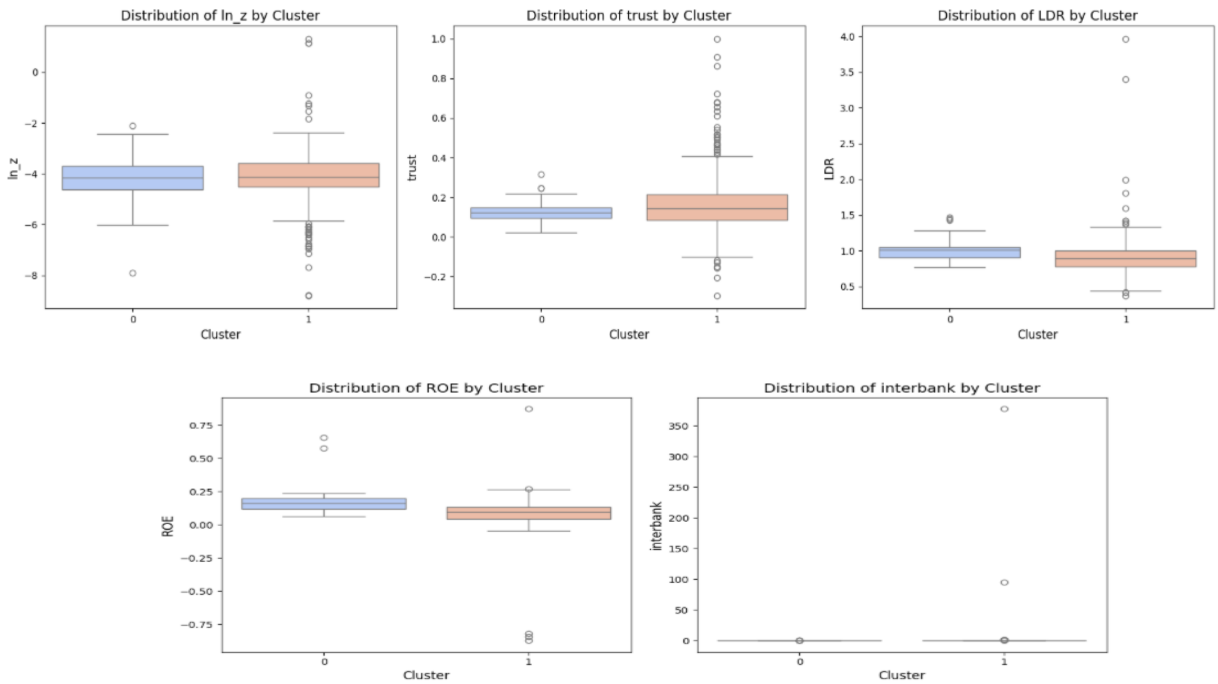

Figure 4: Visualization Results Using Boxplot Charts
Source: Author's calculation and visualization using Python

Clustering Results Using the Hierarchical Clustering Algorithm
The Hierarchical Clustering algorithm constructs a hierarchy of clusters, with results typically represented in the form of a Dendrogram. Using an agglomerative strategy, the hierarchical clustering method divides the banks into two distinct groups based on the processed dataset. Specifically:

- *Cluster 1 (Safe/Low-Risk Group):* This cluster consists of large-scale banks with consistently strong financial indicators. The group operates efficiently, with an average Return on Equity (ROE) of 17.3% and a Loan-to-Deposit Ratio (LDR) of 100.6%. In addition, these banks maintain a high level of financial stability, characterized by low liquidity risk (ln_z = −4.235), stable credit growth (15.8%), and notably limited dependence on the interbank market (coefficient of 0.102).

- *Cluster 0 (High-Risk Group):* This cluster exhibits pronounced risk characteristics, with relatively higher liquidity risk (ln_z = −4.114). The most striking feature is its extremely rapid credit growth (substantially higher than the other group, with a coefficient of 4.063), accompanied by heavy reliance on interbank borrowing (coefficient of 1.338).

Based on the boxplots derived from the Hierarchical Clustering results, the data are clearly divided into two groups of banks with markedly different financial characteristics:

- Cluster 0 stands out with larger scale and stronger operational activity. The values of cash and gold holdings, customer deposits, and customer loans are significantly higher than those of Cluster 1, indicating that this group consists of banks with substantial funding sources and loan portfolios. The LDR of this cluster is approximately equal to or above 1, reflecting a high level of fund utilization. At the same time, the median ROE is higher, suggesting stronger profitability. Variables such as trust and ln_z are also more stable, implying relatively solid safety and reputation levels.

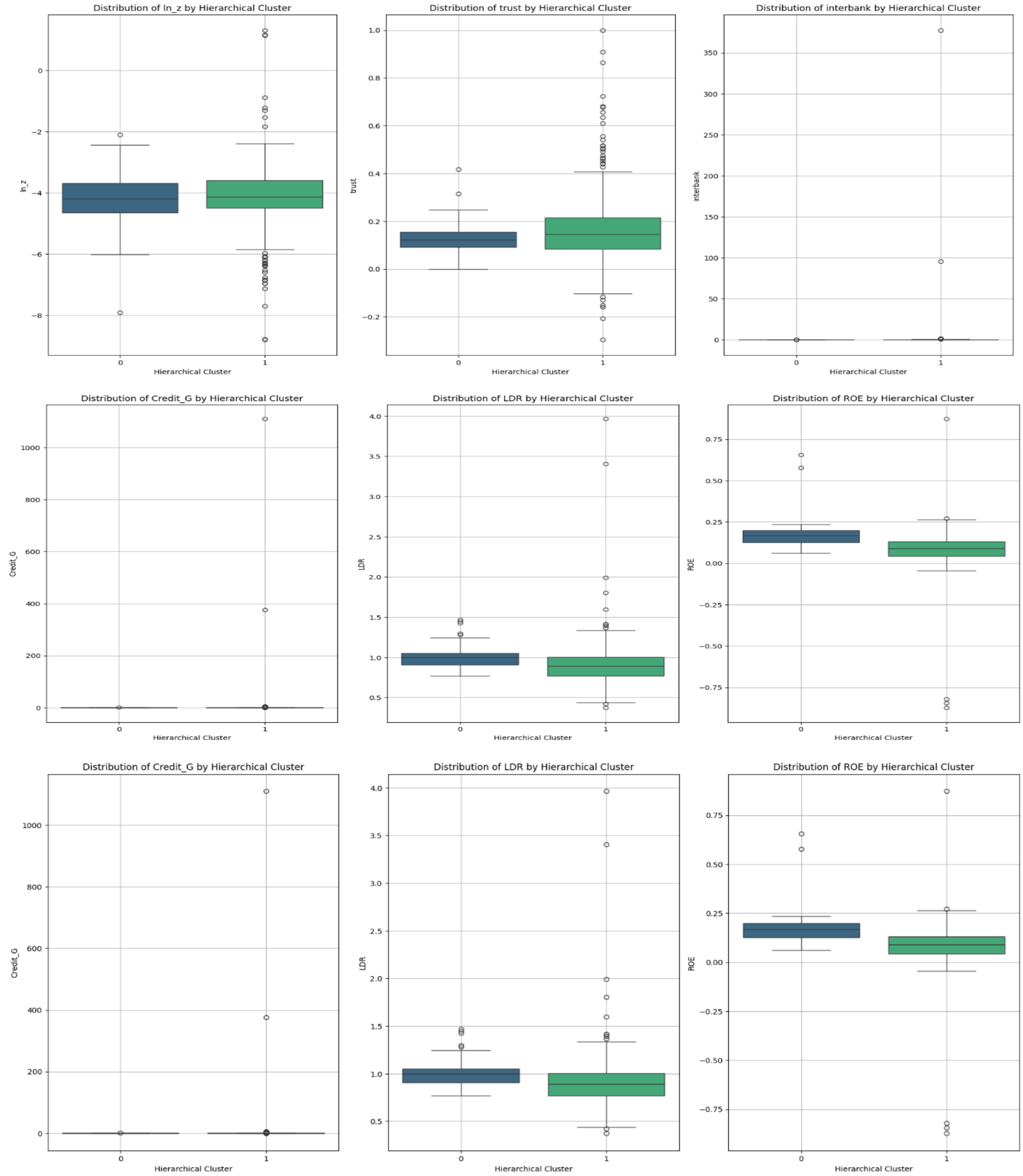

Figure 5: Statistical Test Results for Hierarchical Clusters

Source: Author's calculation and visualization using Python

- Cluster 1, in contrast, exhibits characteristics of smaller-scale banks. Customer deposits, interbank deposits, and customer lending are all considerably lower. The distribution of ROE is lower and more dispersed, indicating unstable profitability. The LDR is lower than that of Cluster 0, implying either a more conservative funding utilization strategy or weaker credit demand. Certain variables, such as interbank and Credit_G, display numerous outliers, reflecting heterogeneity in operational behavior among banks within the same cluster.

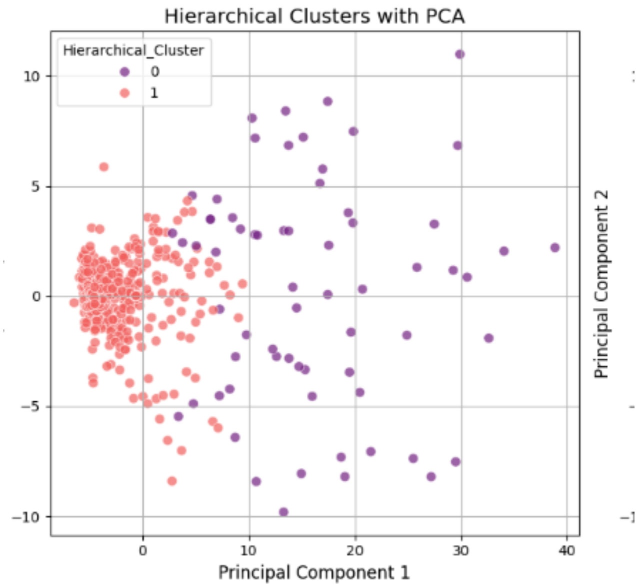

Figure 6: Visualization of Hierarchical Clustering Results in PCA Space
Source: Author's calculation and visualization using Python

The results show that the data are divided into two relatively distinct clusters, primarily differentiated along Principal Component 1. The red cluster is densely concentrated around the origin, reflecting a group of banks with relatively homogeneous financial characteristics and limited variation in liquidity-related indicators. In contrast, the purple cluster is more widely dispersed and extends toward higher values of the first principal component, indicating greater heterogeneity in asset structure, funding sources, and liquidity ratios among these banks.

Comparison of Algorithm Performance

To obtain an objective assessment of clustering quality, the study compares algorithms using a multi-metric evaluation framework, including internal evaluation measures and validation metrics. The detailed results are summarized below:

Table 2: Performance Comparison Between K-Means and Hierarchical Clustering

Algorithm	Silhouette Score	Davies–Bouldin
K-Means	0.2815	1.8838
Hierarchical	0.4238	1.8662

Source: Author's calculation using Python

Evaluation of Internal Cluster Structure

From the perspective of geometric structure in feature space, Hierarchical Clustering demonstrates superior performance among distance-based clustering algorithms. This is reflected in a higher Silhouette Score (0.4238) and a lower Davies–Bouldin Index (1.8662). These results indicate that Hierarchical Clustering achieves a relatively good balance between within-cluster cohesion and between-cluster separation, thereby capturing the latent multidimensional structure of the data more clearly. Methodologically, the hierarchical structure allows greater flexibility in describing complex and nonlinear relationships among banking financial indicators.

In contrast, K-Means records lower performance, with a Silhouette Score of 0.2815 and a Davies–Bouldin Index of 1.8838. Although these metrics suggest that K-Means can still generate reasonably structured clusters, visualization in PCA space reveals that cluster boundaries are not sharply defined. This outcome reflects an inherent structural limitation of K-Means: the algorithm forces every observation into a predefined cluster, even when internal similarity among observations is relatively weak.

Model Selection

The results indicate that Hierarchical Clustering provides relatively better separation performance, as evidenced by superior internal evaluation metrics compared to K-Means. Furthermore, its clustering outcomes generate bank groups with clearer financial characteristics and stronger interpretability.

The hierarchical clustering results suggest that the Vietnamese commercial banking system is primarily divided into two groups with distinct differences in financial stability and liquidity risk levels. The first group comprises large-scale banks with relatively strong profitability, low dependence on the interbank market, and lower liquidity risk. This group plays a stabilizing role within the system, maintaining relatively sustainable funding structures.

Conversely, the second group consists mainly of small- and medium-sized banks characterized by volatile credit growth, greater reliance on short-term funding sources, and higher liquidity risk. Banks in this group tend to be more sensitive to macroeconomic fluctuations and more vulnerable under adverse market conditions.

Overall, empirical findings reveal that the latent structure of liquidity risk in the Vietnamese commercial banking system is shaped by differences in operational scale, profitability, reliance on short-term funding, and cash flow stability. These factors interact to form diverse and clearly differentiated risk patterns across banks.

CONCLUSION AND RECOMMENDATIONS

Conclusion

This study successfully achieved its objective of exploring the liquidity risk structure of the Vietnamese commercial banking system during the 2009–2024 period through the application of unsupervised machine learning clustering algorithms. The empirical results confirm that the use of clustering techniques, particularly the combination of K-Means and Hierarchical Clustering, provides superior effectiveness in identifying multidimensional risk profiles that traditional statistical methods often overlook.

The most significant methodological contribution of this study lies in demonstrating the superiority of the Hierarchical Clustering model in capturing complex financial data structures. This approach more clearly delineates the differentiation among bank groups based on size, profitability, and dependence on the interbank market.

From both theoretical and practical perspectives, the study provides evidence that liquidity risk in Vietnam is not randomly distributed but instead forms distinct cluster structures. The existence of bank groups with contrasting characteristics, one relatively safe group with high funding autonomy and another high-risk group characterized by aggressive growth and reliance on short-term capital flows, highlights the necessity of shifting from a uniform management approach to a risk-segmentation-based governance framework.

Policy and Managerial Implications

Based on the clustering results of risk characteristics, the study proposes specific managerial implications for different market participants.

First, for the State Bank and regulatory authorities, the findings suggest the need to establish a tiered, data-driven supervisory mechanism. Instead of applying a uniform regulatory framework, authorities should develop differentiated prudential standards tailored to each identified cluster. For the high-risk group (characterized by rapid credit growth and strong interbank dependence), stricter controls on the loan-to-deposit ratio (LDR) and higher liquidity buffer requirements should be imposed. Conversely, for the safer group, certain regulatory indicators may be moderately relaxed to encourage operational efficiency, while using this group as a benchmark reference for the entire system.

Second, for commercial banks, managers should clearly recognize their institution's position within the overall risk landscape. Banks in the high-risk cluster should prioritize restructuring their funding sources by reducing reliance on short-term interbank borrowing and increasing medium- and long-term deposits from customers to enhance stability. Credit growth strategies should align with risk management capacity rather than purely pursuing expansion in scale. Meanwhile, banks in the safer cluster should maintain financial discipline and leverage their reputational advantages to expand modern banking services and diversify non-interest income streams in order to reduce concentration risk.

Third, regarding the interbank market, the clear differentiation in funding dependence across clusters signals potential systemic contagion risks. Therefore, greater transparency in capital transactions is necessary, along with the establishment of early warning mechanisms to detect unusual liquidity shifts among bank groups. This would ensure that the interbank market functions as an effective liquidity adjustment channel rather than a shelter for structural risks.

Limitations and Future Research Directions

Although the study provides initial insights into latent liquidity risk structures using clustering models in machine learning, several limitations remain. The dataset primarily focuses on quantitative financial indicators and does not incorporate qualitative governance factors or unstructured data (such as market news), which may significantly influence liquidity risk.

Future research could expand the analysis by integrating textual data from annual reports and media sources. Additionally, combining clustering techniques with supervised machine learning models such as XGBoost or Random Forest to predict the probability of transition between risk clusters represents a promising direction, contributing to more proactive and forward-looking risk management capabilities.

REFERENCES

1) Basel Committee on Banking Supervision. (2008). *Principles for Sound Liquidity Risk Management and Supervision*. Bank for International Settlements.
2) Beatty, Anne L. and Liao, Scott and Zhang, Haiwen (2018). The Effect of Banks' Financial Reporting on Syndicated Loan Structures. *Journal of Accounting & Economics* (JAE), http://dx.doi.org/10.2139/ssrn.3256916
3) Bui Dan Thanh, Nghiem Hoang Thao Vy, & Nguyen Ngoc Huyen. (2022). Factors affecting liquidity risk of commercial banks in Vietnam. *Asian Journal of Economics and Banking*. https://doi.org/10.63065/ajeb.vn.2022.198.81408
4) Cihak, M., Hesse, H (2010). Islamic Banks and Financial Stability: An Empirical Analysis. *J Financ Serv Res* 38, 95–113. https://doi.org/10.1007/s10693-010-0089-0
5) Dietrich, A., & Wanzenried, G. (2011). Determinants of bank profitability before and during the crisis: Evidence from Switzerland. *Journal of International Financial Markets, Institutions and Money*. https://doi.org/10.1016/j.intfin.2010.11.002
6) Duttweiler, R. (2009). *Managing Liquidity in Banks: A Top Down Approach.* John Wiley & Sons Ltd. DOI:10.1002/9781119206415
7) Duttweiler, R. (2016). *Liquidity Risk Management: A Practitioner's Perspective*. John Wiley & Sons. https://lascoco.lagosstate.gov.ng/package/liquidity-risk-management_-a-practitioners-perspective-pdfdrive-pdf/

8) Hannan, T. H., & Hanweck, G. A. (1988). Bank insolvency risk and the market for large certificates of deposit. *Journal of Money, Credit and Banking. Vol 20(2), pp.203-211.* https://doi.org/10.2307/1992111
9) Lepetit, L., & Strobel, F. (2013). Bank insolvency risk and time-varying Z-score measures. *Journal of International Financial Markets, Institutions and Money.* Vol 25, pp.73-87. https://doi.org/10.1016/j.intfin.2013.01.004
10) Milojevi´ C, N., & Redzepagic, S. (2021). Prospects of artificial intelligence and machine learning application in banking risk management. *Journal of Central Banking Theory and Practice*, 10(3), 41–57. https://doi.org/10.2478/jcbtp-2021-0023 .
11) Myra, V. De L. (2020). The impact of credit risk and macroeconomic factors on profitability: the case of the ASEAN banks. *Banks and Bank Systems*, 15(1), 21-29. http://dx.doi.org/10.21511/bbs.15(1).2020.03
12) Nguyen Duc Trung & Tran Trong Huy. (2024). Determinants of Vietnam commercial banks liquidity risk. *Banking Review.* https://tapchinganhang.gov.vn/cac-nhan-to-tac-dong-den-rui-ro-thanh-khoan-cua-cac-ngan-hang-thuong-mai-viet-nam-12199.html
13) Nguyen Van Tien. (2010). *Risk Management in Banking Business.* Statistics Publishing House.
14) Peter J. Rousseeuw (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20, p*53-65. https://doi.org/10.1016/0377-0427(87)90125-7
15) Truong Quang Thong & Pham Minh Tien. (2014). Factors affecting liquidity risk: The case of Vietnamese joint-stock commercial banks. Issue 21, *Financial and Monetary Market Review*
16) Truong Thi Thuy Duong & Le Hai Trung. (2023). Application of machine learning in identifying corporate financial distress. *Journal of Economics and Development*. DOI: https://doi.org/10.33301/JED.VI.1066
17) Valentin Lennart Heß, Bruno Damásio (2025). Machine learning in banking risk management: Mapping a decade of evolution, *International Journal of Information Management Data Insights,* Volume 5, Issue 1, https://doi.org/10.1016/j.jjimei.2025.100324

Corresponding Author: Vu Thi Thu Huong