Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Equity Market Structure and Trading Diversification: Insights from Panel Data, Clustering, and Machine Learning

Angelo Leogrande

leogrande.cultore@lum.it

LUM University Giuseppe Degennaro

Fabio Anobile

LUM University Giuseppe Degennaro

Alberto Costantiello

LUM University Giuseppe Degennaro

Carlo Drago

Unicusano University

Massimo Arnone

University of Catania

Research Article

Keywords: Stock market diversification; Market structure; Remittances; Financial development; Machine learning

Posted Date: March 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8947864/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Equity Market Structure and Trading Diversification: Insights from Panel Data, Clustering, and Machine Learning

Angelo Leogrande*, LUM University Giuseppe Degennaro, Casamassima, leogrande.cultore@lum.it

Fabio Anobile, LUM University Giuseppe Degennaro, Casamassima, anobile.phdstudent@lum.it

Alberto Costantiello, LUM University Giuseppe Degennaro, Casamassima, costantiello@lum.it

Carlo Drago, Unicusano University, Rome, carlo.drago@unicusano.it

Massimo Arnone, University of Catania, Catania, massimo.arnone@unict.it

*Corresponding Author

Abstract

This article aims to contribute to a relatively understudied area of financial development, namely, the internal dispersion of trading activity. The focus is not on overall financial development measures such as total market capitalization and liquidity but rather on trading diversification, defined as the proportion of trading volume contributed by firms outside the VTX, representing the top ten most frequently traded firms. The article uses data from the World Bank's Global Financial Development Database. The sample is constructed as a balanced panel of 23 countries over the period 2002-2021, starting with a sample of 38 countries. The article uses four key explanatory variables, namely, relative size of deposit-taking banks (DBS), remittance inflows (REM), market capitalization excluding the top ten firms (MCX), and outstanding international public debt (IPU). The article uses a combination of panel econometrics, hierarchical clustering, and machine learning methods. The econometric results show that a diversified financial system structure and remittance inflows are strongly, positively related to overall and less concentrated trading activity, while bank dominance and reliance on international public debt are related to more concentrated trading activity. The clustering results show significant cross-country heterogeneity and a core-periphery structure. The machine learning results show that, using all models, equity market structure is again found to be the most important explanatory variable, with external financial flows being important as well. The article concludes that equity market structure is key to understanding internal dispersion, with important policy implications.

Keywords: Stock market diversification; Market structure; Remittances; Financial development; Machine learning

JEL Codes: G12; G15; C23; C38; O16

1. Introduction

In the course of the last few decades, financial markets have witnessed significant structural changes in response to globalization, technological progress, and the increasing integration between the real and financial sectors of the economy. These changes have revived research and policy interest in not just the size and liquidity of markets, but also in their structure in terms of the distribution of trading activity in markets. Market depth, breadth, and concentration are now recognized as critical components of financial development with significant implications for financial markets and their development. However, despite the extensive research on stock market development, most of the extant literature has focused on aggregate measures such as market capitalization, total value traded, and turnover ratios, with relatively little attention given to the structure of trading activity in terms of the distribution of trading between firms. In practice, stock markets tend to be characterized by high trading concentration in a small group of large firms, while the majority of firms tend to be thinly traded, with some even being small and medium-sized enterprises. This has critical implications for informational efficiency, diversification benefits, and access to finance for small firms. From a policy perspective, it is therefore critical to understand the determinants of a wider and more diversified trading structure, especially in the context of financial development's increasing association with notions of inclusive growth, financial resilience, and sustainable development. Meanwhile, recent research has also shed new light on the multidimensional, heterogeneous, and complex nature of financial development. For example, research on the finance-growth nexus has reinforced the message that the impact of financial deepening on growth and stability varies depending on the institutional context, market structures, and the coexistence of bank-based and market-based finance. Other recent research has pointed out that financialization could be a source of macroeconomic volatility, while the complex relationship between entrepreneurship, credit, and market capitalization is critical in shaping long-run growth performance. In parallel, a new wave of research has been devoted to digital finance, financial inclusion, the role of technology in shaping financial systems, as well as the opportunities and challenges that these phenomena create in terms of governance, regulation, and market power. In this context, a growing number of recent research studies have made significant use of advanced quantitative methods, as well as machine learning algorithms, in the analysis of financial markets, prediction of asset prices, and estimation of risks. In this regard, these advanced methods have been successful in improving the accuracy of predictions, as well as capturing the complex, nonlinear, and multidimensional relationships in financial data. However, the vast majority of these research studies have focused on price prediction, volatility forecasting, or portfolio optimization, while the market structures, in terms of the distribution of trading activities by firms, remain understudied. In this context, there is a lack of empirical research on the joint impact of market structures, financial system configurations, and external financial flows on the scope, depth, and diversity of stock market trading. This paper attempts to address the information gap by focusing explicitly on stock market trading diversification, defined as the percentage of trading activity attributed to firms other than the top ten trading firms (VTX). This figure captures an important and commonly neglected aspect of stock market development: the degree to which trading is dominated by a few large firms rather than more evenly distributed among all firms. A high VTX figure reflects a less concentrated and more diversified stock trading structure in which smaller firms play a greater role in stock trading activities. A low VTX figure, conversely, reflects a highly concentrated stock trading structure dominated by a few large firms. The key research question, therefore, is simple but innovative: "What are the key structural and financial factors driving the diversification of the trading activity in the stock markets across countries and through time?" More concretely, the paper seeks to answer the following questions: "What are the effects of the relative size of the deposit money banks in the financial system (DBS), the inflows

of remittance to the GDP (REM), the market capitalization net of the top ten companies (MCX), and the outstanding international public debt to the GDP (IPU) on the trading activity in the stock markets?" The variables are used to capture four different aspects: the structure of the financial system, the inflows of external private financial flows, the internal composition of the equity markets, and the level of reliance on external public debt. What is innovative in this approach is, first, the fact that the existing literature has extensively studied the development of the stock markets but has not focused enough on the composition of the trading activity, i.e., the entities trading in the market. Focusing on the VTX, the paper therefore addresses the "who" by looking at the "how." Second, the paper brings together different variables, such as the structure of the financial system, the inflows of remittances, the composition of the equity markets, and the level of reliance on external public debt, and analyzes the joint effect of all these factors on the trading activity in the stock markets, providing an innovative approach by considering whether the level of trading activity is driven by the internal structure of the markets, the structure of the financial system, or the inflows of external financial flows. Third, the paper employs a multi-method empirical approach, using a combination of traditional panel data analysis and unsupervised machine learning techniques, specifically clustering, and supervised machine learning techniques. The main empirical approach involves the implementation of the Fixed Effects and Random Effects panel data models, using a balanced panel dataset with 23 countries and 307 observations, allowing for the control of unobserved heterogeneity and the identification of robust relationships across the panel data set. Unlike the more traditional approach, the present study recognizes the potential for financial system heterogeneity and extends the analysis using hierarchical clustering techniques, allowing for the identification of unique country groupings with different combinations of market structures, banking system importance, remittance dependence, and exposure to public debt at the international level. Furthermore, the present study extends the analysis of the empirical relationships using a range of machine learning techniques, with the objective of identifying the predictive potential of the variables included in the analysis, using a range of supervised machine learning models. The implementation of this approach not only adds robustness to the analysis but also enables the identification of the key drivers of the trading diversification across countries and over time, extending the analysis beyond the more traditional panel data approach, and bridging the gap between the two approaches, providing a further contribution to the literature on the analysis of financial market structures, using a combination of machine learning and traditional panel data analysis techniques. The contribution of this paper can be seen in three ways. First, it presents and analyzes a measure of trading diversification in stock markets, which represents an important, though underinvestigated, aspect of financial development. Second, it offers new evidence on the roles of financial systems' structure, external private flows, market concentration, and international public debt in influencing internal trading distribution. Third, it shows the promise of combining panel econometrics, clustering, and machine learning in analyzing stock markets, highlighting their complementarities in uncovering average relationships and structural heterogeneity. From a policy perspective, the focus on trading diversification is particularly relevant. In fact, a market with trading dominated by a few large firms could be more prone to shocks, less supportive of small and medium-sized enterprises, and less effective in intermediating savings towards a wide range of productive investments. On the contrary, a trading market that is more diversified could be more liquid, with better price discovery mechanisms for a wider range of firms, thereby fostering more comprehensive financial development. In this regard, the research results on the structural determinants and financial determinants of broader trading activities are particularly relevant from a policy perspective. The remainder of the paper is structured as follows: in the next section, the relevant literature is reviewed. The data and methodology are presented in the third section. The fourth section presents the econometric model. The results from the hierarchical clustering analysis are presented in the fifth section. The sixth section presents the results from the machine learning performance comparison as well as the results from the analysis on the importance of the variables. The seventh section integrates the results from the

econometric, clustering, and machine learning analyses in order to identify the most important determinants of stock market trading diversification. The eighth section presents the policy implications, while the final section concludes.

2. Literature Review

In the last few years, there has been significant progress in the field of research that lies at the crossroads of financial markets, technological innovations, sustainability, and the macroeconomic system. A significant observation in the recent literature is the increased tendency towards the application of sophisticated quantitative techniques, such as artificial intelligence and machine learning, for the analysis and interpretation of economic and environmental systems. A relevant piece of research is the one presented by Miao et al. (2025), where the authors apply a spatiotemporal hybrid deep learning framework for the estimation and analysis of carbon stocks in Jiangsu Province, China. This piece of research is relevant to the overall framework presented by the Handbook of Quantitative Sustainable Finance (Tankov and Zhang, 2025). Within the overall framework, one of the significant areas of research is the potential for finance to provide stability to economic growth. Mahmoudi and Torra (2025) have made significant contributions in this area, while Ullah (2025) argues that the volatility in economic growth might be attributed to the phenomenon of financialization, and the development of the financial system might not necessarily lead to economic development. Abid (2025) has analyzed the relationship between entrepreneurship, investment, credit, and market capitalization, and its impact on economic development in the United States, arguing that economic development is the result of the complex interplay between different variables, and not a single dimension of finance and development. Another important aspect of current financial research is the phenomenon of digital transformation. The main factors that determine investment in financial services in the course of digital financial ecosystem transformation are identified by Manoylenko & Kuznetsova (2025), highlighting their significance. Tsybuliak et al. (2025) prove that digital financial inclusion could be a catalyst for sustainable development, together with environmental protection, supporting the argument that digital finance could be a catalyst for sustainable development. However, the normative aspects in the development of digital finance need to be considered. Nieborak (2025) critically addresses the problem of digital coercion in financial markets and the right to digital opt-out, reminding financial scholars that the development of technology could lead to financial constraints, as well as increased efficiency. Other areas in current financial research include market efficiency, information, and institutional communication. Budiarso & Pontoh (2025) research the market efficiency of firms that pay dividends in a hawkish monetary policy environment in Indonesia. Jurkšas & Kaminskas (2025) research the impact of the ECB Governing Council members' communications on intraday financial markets. Rostek & Yoon (2025) review the recent advances in imperfect competition models in financial markets, providing a conceptual framework that could be used to understand frictions, strategic interactions, and deviations from perfect efficiency in financial markets. The growing complexity of financial systems has accompanied the development and application of more advanced forecasting and risk analysis techniques. Liu and Yang (2025) suggest

a hybrid forecasting model for A-share stock index based on frequency domain decomposition and temporal fusion transformers. Abdullayev et al. (2025) have developed a forecasting model based on quasi-recurrent neural networks and crown porcupine optimizer, aiming at stock market risk forecasting. Zuo (2025) has applied a combination of Monte Carlo simulations and LSTM networks in forecasting Tesla stock prices. These works demonstrate the ongoing application of machine learning in financial forecasting. Tang et al. (2025) suggest topological data analysis as a new forecasting model, which can reveal hidden structures in financial systems that are not easily accessible through other forecasting approaches. Jiang and Yang (2025) have contributed a GSADF-based model in detecting firm bubbles in the photovoltaic industry. Another significant category of research focuses on interconnectedness, contagion, and the transmission of shocks. Soltani et al. (2025) explore the dynamic connectedness between economic sanctions sentiment, uncertainty factors, and financial assets by employing a Quantile VAR model. Gong et al. (2025) examine the contagion between international stock markets and geopolitical risks by employing a two-layer network. Bisiriyu et al. (2025) examine the role of policy uncertainty in stock markets in India. From the above research, it is evident that the current stock market is highly interconnected. The research on the stock market's structure, participation, and determinants is the next significant category. Ma (2025) explores the impact of income distribution on participation in risky financial markets in China. Mu et al. (2025) explore the determinants of limited participation in financial markets by employing explainable machine learning. Adamolekun et al. (2025) establish a link between national energy generation capacity and stock market development by employing econometric models. This research establishes a significant link between stock market development and the real sector. Ben Salem et al. (2025) extend the research on quantile connectedness between stock market development and macroeconomic factors in the context of African countries. The research on the stock market's structure, participation, and determinants suggests that the current stock market is highly interconnected. Environmental sustainability is a significant component in the current financial research. Ridwan (2025) explores the role of financial market efficiency in green development in the United States. Wang et al. (2025) explore a green development model in the prefabricated building industry by employing intelligent technology. Wu et al. (2025) extend the research on the green development model by employing climate risk, focusing on the impact of extreme climate events on international stock markets in the tourism sector. Afshan et al. (2025) extend the research on the stock market by focusing on the environmental sustainability component. Institutional, social, and governance issues are also given significant prominence. In this regard, Saucedo Loera et al. (2025) reveal a strong association between corporate social responsibility and corporate reputation in the Mexican stock market, highlighting the growing importance of ESG issues. Another study by Li and Guo (2025) focuses on the potential of government funds in facilitating a transition from relief to governance empowerment through improvements in labor investment efficiency. Horton (2025) provides a broader reflection, based on a medical societal perspective, on the need to monitor those individuals who manage financial systems, particularly in light of current debates on regulations, oversight, and accountability in an era dominated by technological advances.

Macro-theme	Key References	Main Methodologies	Main Findings / Results
1. AI, Machine Learning and Advanced Quantitative Methods	Miao et al. (2025); Liu & Yang (2025); Abdullayev et al. (2025); Zuo (2025);	Deep learning (spatiotemporal models, LSTM, QRNN); Transformers; Monte Carlo	AI-based and data-driven methods significantly improve prediction accuracy (stocks, risk, indices) and enable detection of complex patterns (bubbles, nonlinear

in Finance & Environment	Tang et al. (2025); Jiang & Yang (2025)	simulation; Topological Data Analysis; GSADF bubble tests	dynamics, spatial–temporal environmental processes). These approaches outperform traditional models and allow better risk and sustainability assessment.
2. Financial Development, Market Structure and Growth/Stability Nexus	Mahmoudi & Torra (2025); Ullah (2025); Abid (2025); Rostek & Yoon (2025); Adamolekun et al. (2025); Ben Salem et al. (2025); Ma (2025); Mu et al. (2025)	Econometric models (ARDL, panel regressions, quantile methods); Theoretical models of imperfect competition; Explainable ML; Micro and macro empirical analysis	Financial development affects growth in a nonlinear and heterogeneous way. Financialization may increase volatility; market structure and participation matter for stability; household participation depends on income and constraints; macro-financial linkages vary across countries and quantiles.
3. Interconnectedness, Uncertainty, Policy and Information in Financial Markets	Soltani et al. (2025); Gong et al. (2025); Bisiriyu et al. (2025); Jurkšas & Kaminskas (2025); Budiarso & Pontoh (2025)	Quantile VAR; Network models; Time-frequency analysis; Event studies; Market efficiency tests	Financial markets are highly interconnected and sensitive to geopolitical risk, sanctions, and policy uncertainty. Institutional communication (e.g., ECB) has immediate market effects. Monetary policy regimes shape market efficiency and price discovery.
4. Sustainability, Green Finance, ESG, Digital Finance and Policy	Tankov & Zhang (2025); Tsybuliak et al. (2025); Ridwan (2025); Wang et al. (2025); Wu et al. (2025); Afshan et al. (2025); Saucedo Loera et al. (2025); Li & Guo (2025); Manoylenko & Kuznetsova (2025); Nieborak (2025); Horton (2025)	Quantitative sustainable finance models; Panel econometrics; Case studies; Policy analysis; ESG and governance frameworks; Legal-institutional analysis	Financial markets play a key role in supporting green development, environmental resilience, and energy security. Digital finance promotes inclusion but raises governance and rights issues. ESG and CSR affect reputation and market outcomes. Public policy and guided funds shape investment efficiency and sustainable transitions. Climate risks increasingly impact financial markets.

3. Methodology and Data

In order to fulfill the research objective, the study will use the Global Financial Development Database, a standard source of information that is frequently employed in empirical studies because of its extensive coverage of financial structure, depth, accessibility, and efficiency indicators. The initial sample is comprised of 38 OECD countries, covering the period from 2002 to 2021. However, owing to the constraints of the database, the study will eventually use a sample of 23 countries with 307 observations. The study will use the following formula to determine the dependent variable, VTX, which is equal to the value traded by all firms, excluding the top ten firms, divided by the total value traded on the stock exchange. The study will use four explanatory variables that will cover four different dimensions of financial structure and external financial integration. First, DBS will be determined as the ratio of assets of deposit-taking banks to the combined assets of deposit-taking banks and central banks, which will show the importance of commercial banks within the overall structure of the financial sector. A higher level of DBS will show a higher level of bank intermediation compared with central banks. Second, the study will use REM, determined as remittances divided by GDP, which will show the importance of remittances within the overall structure of the country's income. This will also show the overall

contribution of remittances to the national income, along with their spillover effects on overall financial market depth and national development. Third, MCX will show the market capitalization of all firms, excluding the top ten firms, divided by the total market capitalization, which will show the level of concentration within the overall structure of the equity market. Finally, IPU will show outstanding international public debt securities divided by GDP, which will show countries with higher levels of government debt and, hence, higher exposure to exchange rate risks within a global and local environment owing to disparities in financial market development.

	Variable name	Acronym	Description (≈35 words)
Y	Value traded excluding top 10 traded companies to total value traded (%)	VTX	Measures stock market trading activity excluding the ten largest firms, capturing market breadth and liquidity dispersion. Higher values indicate more diversified trading, lower concentration, and a greater role of smaller and mid-cap companies in market turnover.
	Deposit money bank assets to deposit money bank and central bank assets (%)	DBS	Indicates the relative size of deposit money banks within the overall banking system, compared to central bank assets. Higher values suggest greater financial intermediation by commercial banks and a more market-oriented, less central-bank-dominated financial structure.
	Remittance inflows to GDP (%)	REM	Measures the importance of workers' remittances relative to economic output. It reflects external private financial flows, household income support, and potential effects on consumption, investment, financial development, and macroeconomic stability in recipient countries.
	Market capitalization excluding top 10 companies to total market capitalization (%)	MCX	Captures stock market concentration by excluding the ten largest firms from total capitalization. Higher values indicate a more diversified equity market structure, reduced dominance of large firms, and broader participation of smaller listed companies.
X	Outstanding international public debt securities to GDP (%)	IPU	Measures the stock of public sector debt issued on international markets relative to GDP. It reflects government reliance on external bond financing, exposure to international capital markets, and potential vulnerability to global financial conditions and exchange rate risks.

4. The Econometric model

Specifically we have estimated the following equation:

$$VTX_{it} = \alpha + \beta_1(DBS)_{it} + \beta_2(REM)_{it} + \beta_3(MCX)_{it} + \beta_4(IPU)_{it}$$

Where i=38 and t=[2001;2022]**.**

Moreover, the regression analysis provides an unambiguous and economic interpretation for the determinants of VTX, where VTX represents the proportion of trading activity in the stock market by firms not in the top ten industry segment, thus reflecting the level of trading activity and diversification in the stock market. Furthermore, the evidence from the FE and RE estimations presents a strikingly

similar pattern, suggesting that the findings are not conditional upon unobserved unit-specific heterogeneity.As expected from the findings presented in the paper, the DBS variable has a negative and statistically significant coefficient, significant at the 10 percent level, in both the FE and RE estimations. This result supports the hypothesis that the relative prominence of deposit-taking banks vis-à-vis the central bank is associated with lower levels of VTX, implying a more concentrated trading structure dominated by a few large industry firms. From an economic point of view, the bank intermediation financial system might be related to the reduced reliance on the stock market, thus leading to a more concentrated trading structure dominated by a few large industry firms. Moreover, the REM variable is statistically significant, with a robust and strongly significant positive coefficient in both the FE and RE estimations, thus confirming the significant impact of the variable on VTX. This result supports the hypothesis that an increase in the level of remittance inflows relative to the level of GDP is associated with a more diversified trading structure in the stock market, possibly due to the positive impact of remittances on the level of income and savings of households, thus leading to a wider trading structure beyond the dominant industry firms. Likewise, the MCX variable reaches a positive and highly significant value in both equations. The MCX variable measures market capitalization excluding the top ten firms as a proportion of total capitalization. The positive and significant value of the MCX variable is consistent with the VTX results and points to the fact that higher levels of local market diversification, proxied by the higher weight of smaller firms in local market capitalization, are associated with higher levels of diversification in the trading activity structure beyond the top ten firms. On the other hand, the IPU variable reaches a negative and highly significant value in both equations. The negative value of the IPU variable points to the fact that, in the sample of economies, those with a higher dependence on international public debt markets are associated with lower VTX values, i.e., with less diversified trading activity structures. Finally, the constant term is positive and significant in both equations and captures the base level of VTX when all regressors are zero. On the whole, the parallelism between the FE and RE results reinforces the main findings, which point to the fact that the main drivers of trading activity structure diversification are the local market structure (MCX), external private flows (REM), and the level of external public indebtedness (IPU), with the bank-dominated financial system (DBS) associated with less diversified trading activity structures.

	Fixed Effects				Random Effects			
Variable	Coeff.	Std. Err.	t	p-value	Coeff.	Std. Err.	z	p-value
const	39.9813	13.0676	3.060	0.0024 ***	34.6465	13.2080	2.623	0.0087 ***
DBS	-0.238296	0.126997	-1.876	0.0616 *	-0.231146	0.124323	-1.859	0.0630 *
REM	11.3645	2.09810	5.417	<0.0001 ***	9.57534	1.94036	4.935	<0.0001 ***
MCX	0.553872	0.0797729	6.943	<0.0001 ***	0.650882	0.0709468	9.174	<0.0001 ***
IPU	-0.949573	0.252883	-3.755	0.0002 ***	-1.01572	0.228937	-4.437	<0.0001 ***

On one hand, summary statistics of the Fixed Effects (FE) and Random Effects (RE) models reveal characteristics of the structure of the data set and shed light on the relative performance of each model in managing variability in the dependent variable, VTX, which captures the percentage of stock market trading volume attributed to firms outside of the top ten. Both models are applied to the same data set of panel data, which consists of 307 observations and 23 cross-sectional units, thus making them comparable in their scope and sample characteristics. The mean and standard deviation of VTX remain constant across models, with an average of approximately 43.25 and an appreciable standard deviation of 23.69, which captures considerable variability in the structure of stock markets across countries and

over time. Moreover, diagnostic statistics reveal that the FE model has performed substantially better than the RE model in representing the data set, as reflected by its lower values of summary statistics, particularly the sum of squared residuals (SSR), which is 26,145 for the FE model and 60,837 for the RE model, thus revealing that more of the variability of VTX is captured by the FE model. This is further confirmed by lower standard error of the regression values of 9.66 for the FE model and 14.17 for the RE model, which reveal that FE-based predictions of VTX tend to be closer to observed values than RE-based predictions. Furthermore, the log-likelihood values also support the idea that the FE model performs better than the RE model, with a higher log-likelihood value of -1,117.857 recorded in the FE model compared to -1,247.493 in the RE model. This shows that it is more likely that the FE model actually generated the observed data. More model selection criteria also support this result, with lower values of the Akaike Information Criterion (AIC), Schwarz/Bayesian Information Criterion (BIC), and Hannan-Quinn Information Criterion (HQIC) recorded in the FE model than in the RE model, indicating that, despite having more parameters, the FE model provides a more efficient and accurate description of the data. With regards to R-squared values, it can be noted that, in the LSDV model, the R-squared value is approximately 0.848, indicating that the FE model explains approximately 85 percent of the total variation in VTX, controlling for regressor effects as well as unit-specific effects. The within R-squared value is approximately 0.237, indicating that, controlling for cross-sectional effects, approximately one-fourth of the total temporal variation in the data can be explained by the regressor effects, which is considerable in panel data, where unit effects often explain a significant amount of total variation. With regards to the RE model, it can be noted that, in the variance decomposition of the error component, the between variance is approximately 246.9, which is considerably higher than the within variance, which is approximately 85.2. This shows that, in the case of this panel, there is considerable cross-sectional variation in comparison with temporal variation in the data. The mean theta value in this model is approximately 0.813, indicating considerable weighting towards the random effects transformation. Both models have identical values in terms of the Rho statistic, which is 0.574, and in terms of the Durbin Watson statistic, which is 0.825 in both cases. This value is considerably low, indicating considerable positive serial correlation in the residuals, which may be worthy of further investigation. On the whole, the statistics point to the fact that the Fixed Effects model provides a much better and more reliable framework for analyzing the determinants of stock market trading diversification.

Statistic	Fixed Effects (FE)	Random Effects (RE)
Observations	307	307
Cross-sectional units	23	23
Dependent variable	VTX	VTX
Mean dep. var	43.25002	43.25002
S.D. dep. var	23.68932	23.68932
Sum squared resid	26145.35	60837.35
S.E. of regression	9.663138	14.16980
Log-likelihood	-1117.857	-1247.493
Akaike	2289.714	2504.985
Schwarz	2390.339	2523.620
Hannan-Quinn	2329.953	2512.437

rho	0.574173	0.574173
Durbin-Watson	0.825432	0.825432
R^2 (LSDV)	0.847746	—
Within R^2	0.236521	—
Between variance	—	246.893
Within variance	—	85.164
Mean theta	—	0.813493

On one hand, summary statistics of the Fixed Effects (FE) and Random Effects (RE) models reveal characteristics of the structure of the data set and shed light on the relative performance of each model in managing variability in the dependent variable, VTX, which captures the percentage of stock market trading volume attributed to firms outside of the top ten. Both models are applied to the same data set of panel data, which consists of 307 observations and 23 cross-sectional units, thus making them comparable in their scope and sample characteristics. The mean and standard deviation of VTX remain constant across models, with an average of approximately 43.25 and an appreciable standard deviation of 23.69, which captures considerable variability in the structure of stock markets across countries and over time. Moreover, diagnostic statistics reveal that the FE model has performed substantially better than the RE model in representing the data set, as reflected by its lower values of summary statistics, particularly the sum of squared residuals (SSR), which is 26,145 for the FE model and 60,837 for the RE model, thus revealing that more of the variability of VTX is captured by the FE model. This is further confirmed by lower standard error of the regression values of 9.66 for the FE model and 14.17 for the RE model, which reveal that FE-based predictions of VTX tend to be closer to observed values than RE-based predictions. Furthermore, the log-likelihood values also support the idea that the FE model performs better than the RE model, with a higher log-likelihood value of -1,117.857 recorded in the FE model compared to -1,247.493 in the RE model. This shows that it is more likely that the FE model actually generated the observed data. More model selection criteria also support this result, with lower values of the Akaike Information Criterion (AIC), Schwarz/Bayesian Information Criterion (BIC), and Hannan-Quinn Information Criterion (HQIC) recorded in the FE model than in the RE model, indicating that, despite having more parameters, the FE model provides a more efficient and accurate description of the data. With regards to R-squared values, it can be noted that, in the LSDV model, the R-squared value is approximately 0.848, indicating that the FE model explains approximately 85 percent of the total variation in VTX, controlling for regressor effects as well as unit-specific effects. The within R-squared value is approximately 0.237, indicating that, controlling for cross-sectional effects, approximately one-fourth of the total temporal variation in the data can be explained by the regressor effects, which is considerable in panel data, where unit effects often explain a significant amount of total variation. With regards to the RE model, it can be noted that, in the variance decomposition of the error component, the between variance is approximately 246.9, which is considerably higher than the within variance, which is approximately 85.2. This shows that, in the case of this panel, there is considerable cross-sectional variation in comparison with temporal variation in the data. The mean theta value in this model is approximately 0.813, indicating considerable weighting towards the random effects transformation. Both models have identical values in terms of the Rho statistic, which is 0.574, and in terms of the Durbin Watson statistic, which is 0.825 in both cases. This value is considerably low, indicating considerable positive serial correlation in the residuals, which may be worthy of further investigation. On the whole,

the statistics point to the fact that the Fixed Effects model provides a much better and more reliable framework for analyzing the determinants of stock market trading diversification.

Test	Fixed Effects (FE)	Random Effects (RE)
Joint test on regressors	$F(4,280) = 21.6855$, $p = 1.33e-15$	$\chi^2(4) = 114.292$, $p = 8.84e-24$
Group intercepts test	$F(22,280) = 10.523$, $p = 1.06e-25$	—
Heteroskedasticity (Wald)	$\chi^2(21) = 2228.89$, $p = 0$	—
Breusch–Pagan	—	$\chi^2(1) = 225.171$, $p = 6.74e-51$
Hausman	—	$\chi^2(4) = 10.8839$, $p = 0.0279$
Normality of residuals	$\chi^2(2) = 159.239$, $p = 2.64e-35$	$\chi^2(2) = 3.4995$, $p = 0.1738$
Wooldridge autocorrelation	$F(1,20) = 36.6197$, $p = 6.47e-06$	$F(1,20) = 36.6197$, $p = 6.47e-06$
Pesaran CD	$z = 6.15745$, $p = 7.39e-10$	$z = 6.00375$, $p = 1.93e-09$

5. Hierarchical Clustering Results and Financial Structure Segmentation

The results obtained for the normalized indicators clearly demonstrate the trade-off between compactness, separation, and structure for the considered algorithms. Density-Based clustering has high values for maximum diameter, minimum separation, and the Dunn index, which means good cluster compactness and separation, while it has poor results for Pearson's gamma, entropy, and the Calinski-Harabasz index, which indicate poor global structure and balance in cluster partitions. Hierarchical clustering has the best results for Pearson's gamma and reasonably good results for entropy and the Dunn index, which indicate good global ordering and acceptable internal validity, though it does not dominate in separation and compactness results. k-Means has the best results for the Calinski-Harabasz index and high values for Pearson's gamma, which indicate good global variance separation, while it has poor results for the Dunn index and minimum separation, which indicate poor cluster separation in the data space. Model-based and Random Forest clustering methods demonstrate balanced results, which are neither good nor poor for any of the considered indicators, though they are close to the best results for entropy and moderate results for separation and compactness. Fuzzy C-Means has poor results for all separation and compactness indicators, despite good results for entropy, which makes it difficult to evaluate its quality. Considering all indicators together, Hierarchical clustering demonstrates the best balance between high values for Pearson's gamma and good results for other indicators, avoiding poor results for other methods, which makes it the best compromise for clustering validity.

Indicator	Density Based	Fuzzy C-Means	Hierarchical	Model Based	k-Means	Random Forest
Maximum diameter	1.000	0.135	0.099	0.500	0.000	0.360

Minimum separation	1.000	0.000	0.192	0.051	0.036	0.049
Pearson's γ	0.000	0.383	1.000	0.512	0.710	0.235
Dunn index	1.000	0.000	0.392	0.059	0.088	0.065
Entropy	0.000	1.000	0.616	0.984	0.911	0.984
Calinski–Harabasz	0.000	0.505	0.479	0.578	1.000	0.421

From the hierarchical clustering results, it is evident that there is a significant unevenness in the structure of the clusters, with one major cluster being much larger than the rest. Specifically, cluster 1, with 186 observations, represents the majority of the dataset, while the rest of the clusters are significantly smaller, with some of them having only one observation each, i.e., clusters 5 and 10. This suggests that, in the multi-dimensional space defined by VTX, DBS, REM, MCX, and IPU, most of the units share similar financial and market characteristics, while a small number of units have unique combinations of these variables, requiring smaller clusters. This interpretation is supported by the results in terms of the explained within-cluster heterogeneity. Specifically, cluster 1 captures approximately 69.5% of the total within-cluster heterogeneity, indicating that most of the total heterogeneity of the dataset is contained within this major group. Clusters 2, 3, and 4 have smaller proportions, ranging approximately from 7% to 13%, while the rest have negligible proportions. This implies that most of the structure of the dataset is driven by a small number of broad partitions, with cluster 1 being the largest, reflecting what can be defined as "typical" in terms of combinations of stock market breadth (VTX), banking sector size vis-à-vis central bank size (DBS), remittances (REM), market concentration (MCX), and international public debt (IPU). Further insights can be gained with reference to the within-cluster sum of squares. Cluster 1, which consists of the highest number of observations, has the highest value of the within-cluster sum of squares, which is naturally related to the cluster's size, but it is also an indication of the cluster's internal dispersion. This implies that, although the observations included in this cluster are closer to those in the same cluster than to those in the other clusters, they are characterized by a high degree of financial structures and market development diversity. In contrast, the smaller clusters are characterized by significantly lower values of the within-sum-of-squares, which, in some cases, are close to zero, possibly because of the tight internal homogeneity of the observations included in these clusters or, in the case of the single-observation clusters, because of the absence of internal variation by definition. This implies that the smaller clusters are characterized by highly specific financial structures, possibly corresponding to the financial structures of the outliers regarding market concentration, remittance dependence, and exposure to international public debt. The silhouette is another measure of cluster quality and separation, and it provides an alternative view of the cluster configuration and separation. Cluster 1 is characterized by a relatively low silhouette value, equal to 0.271, which implies the presence of a large proportion of observations close to the cluster boundaries with the neighboring clusters. This is consistent with the idea of a cluster including a large and relatively heterogeneous set of observations with smooth differences between the cluster and the other clusters. In contrast, the silhouette values for the other clusters, i.e., for clusters 2, 3, and 4, are slightly higher, possibly reflecting the relatively higher quality and separation of these clusters, although not at a very high level. In contrast, the silhouette values for the smaller clusters, i.e., for clusters 6, 7, 8, and, above all, for cluster 9, are very high, equal to 0.905, possibly reflecting the relatively high separation of these clusters from the rest of the observations and the relatively high

internal cohesion of the observations included in the same cluster. On the whole, hierarchical clustering indicates that the data has a core-periphery structure. The majority of the observations fall into a large, somewhat diffuse core cluster that is characterized by broadly similar combinations of VTX, DBS, REM, MCX, and IPU, while a few units fall into sharply distinct groups with very specific profiles. From the point of view of economic and financial analysis, this means that while a large number of countries or units display broadly similar patterns of financial development or market structures, there are nevertheless significant exceptions that display markedly different characteristics in terms of market concentration, dependence on remittance flows, the structure of the banking system, or the reliance on international public debt, etc., that are sufficiently different to justify their isolation into distinct clusters.

Cluster	1	2	3	4	5	6	7	8	9	10
Size	186	28	42	24	1	6	9	5	5	1
Explained proportion within-cluster heterogeneity	0.695	0.071	0.134	0.072	0.000	0.008	0.009	0.010	6.498×10^{-4}	0.000
Within sum of squares	331.713	33.816	64.125	34.549	0.000	3.664	4.416	4.755	0.310	0.000
Silhouette score	0.271	0.407	0.310	0.339	0.000	0.646	0.607	0.530	0.905	0.000

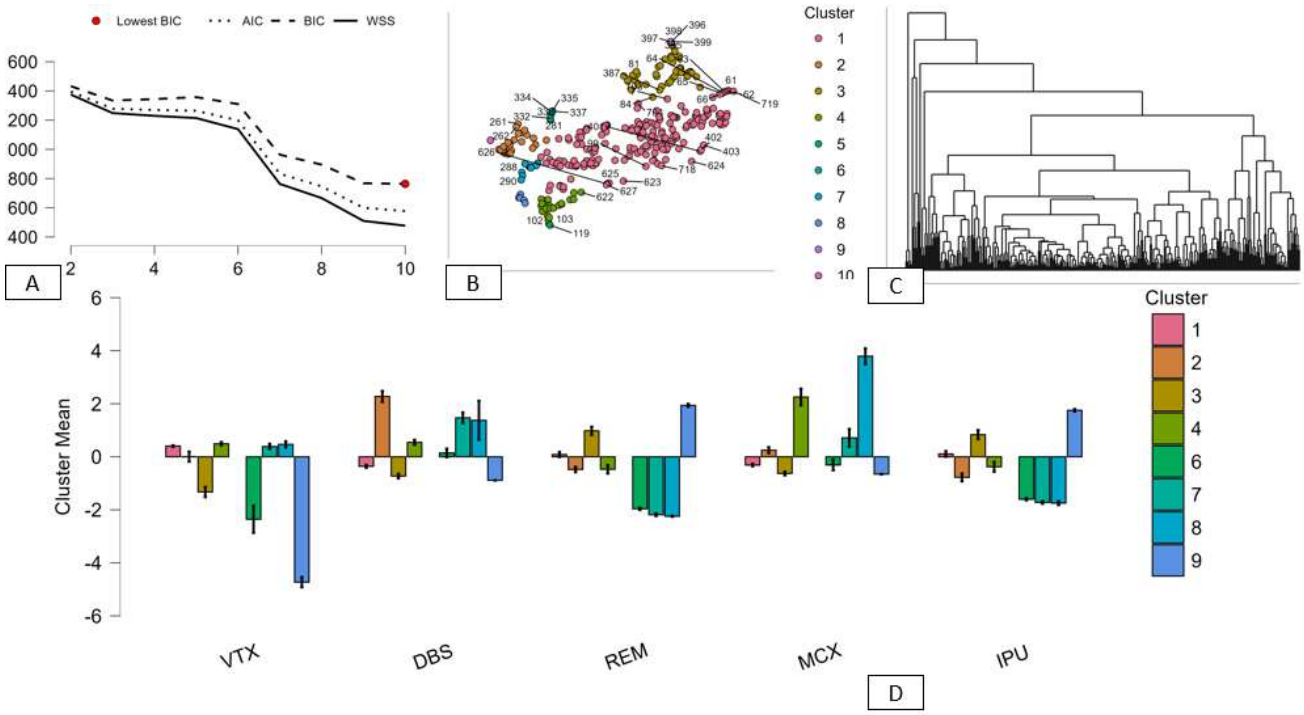
The cluster centroids derived from the hierarchical clustering process provide a rich representation of how differentiated groups of observations can be defined based on their characteristics along dimensions of stock market structure, banking system composition, external financial flows, market concentration, and international public debt reliance. The standardized nature of the variables means that positive and negative values indicate deviations from the sample mean, facilitating cross-cluster comparison of relative positions. For Cluster 1, which also represents the largest proportion of the sample, there is a moderately positive value for VTX and a slightly positive value for REM, along with slightly negative values for DBS and MCX, with IPU close to zero. This represents countries with somewhat broader and more diversified trading activity within the stock market beyond the dominant firms, along with a somewhat smaller role played by deposit-taking banks compared with the central bank. There is a slightly lower level of market concentration and a level of reliance on international public debt close to the mean. This can be interpreted as a mainstream financial structure, with balanced characteristics and no extreme positions along the dimensions considered. Cluster 2, on the other hand, shows a very different profile, with a very positive value for DBS and negative values for REM and IPU, with VTX close to zero and MCX slightly positive. This represents financial systems with dominant deposit-taking banks compared with the role of the central bank within the financial sector, with lower levels of remittance inflows, and a lower reliance on international public debt markets. The slightly positive value for MCX also suggests somewhat higher levels of diversification within the equity market compared with the mean. Cluster 3 is characterized by negative values of VTX and DBS, and positive values of REM and IPU, as well as negative values of MCX. The characteristics of this cluster indicate that the stock exchange is narrow and concentrated, and that banking is relatively less developed, while remittances and international public debt markets are significant. Therefore, Cluster 3 seems to indicate that remittances and international public debt markets play an important role in the overall economy of the country. Cluster 4 is differentiated by its high value of MCX and positive values of VTX and DBS, as well as negative values of REM and IPU. The characteristics of Cluster 4 indicate that the stock exchange is highly diversified and extends beyond the largest firms, and banking is relatively more developed, while remittances and international public debt markets play relatively less important roles. Therefore, Cluster 4 seems to indicate that remittances and international public debt markets do not play an important role in the overall

economy of the country. On the other hand, Clusters 5, 8, and 10 represent more extreme and specialized systems. In Cluster 5, MCX and IPU levels are extremely high, with DBS being positive and REM being strongly negative. This suggests a financial structure with highly diversified equity markets, strong banking systems, high reliance on international public debt, and minimal remittance significance. In Cluster 8, MCX levels are extremely high, with REM being strongly negative, DBS being positive, and VTX being moderately positive. This again points towards highly market-oriented systems with minimal dependence on remittance. In Cluster 10, with exceptionally high levels of DBS and REM, but with IPU being negative, it appears that in this configuration, the banking sector is dominant, remittance is highly significant, while reliance on international public debt is relatively modest. The most extreme cases in terms of negative VTX, as well as high levels in REM and IPU, especially in the case of cluster 9, are covered by clusters 6 and 9. The economies in these clusters are characterized by extremely narrow stock market structures, weak or less dominant banking systems, and a strong dependence on external financial flows, both from the public and the private sector. In conclusion, the results based on the cluster means indicate that the countries are divided into different categories in terms of financial development and external financial integration, ranging from market-based diversified systems to more externally dependent and structurally concentrated systems.

	VTX	DBS	REM	MCX	IPU
Cluster 1	0.405	-0.360	0.085	-0.315	0.107
Cluster 2	0.003	2.275	-0.479	0.249	-0.777
Cluster 3	-1.328	-0.729	0.977	-0.631	0.835
Cluster 4	0.490	0.545	-0.473	2.251	-0.379
Cluster 5	0.128	0.957	-1.201	3.353	2.360
Cluster 6	-2.357	0.145	-1.964	-0.310	-1.604
Cluster 7	0.387	1.469	-2.187	0.713	-1.723
Cluster 8	0.469	1.375	-2.244	3.791	-1.749
Cluster 9	-4.729	-0.889	1.938	-0.655	1.751
Cluster 10	0.495	3.412	2.168	0.412	-1.358

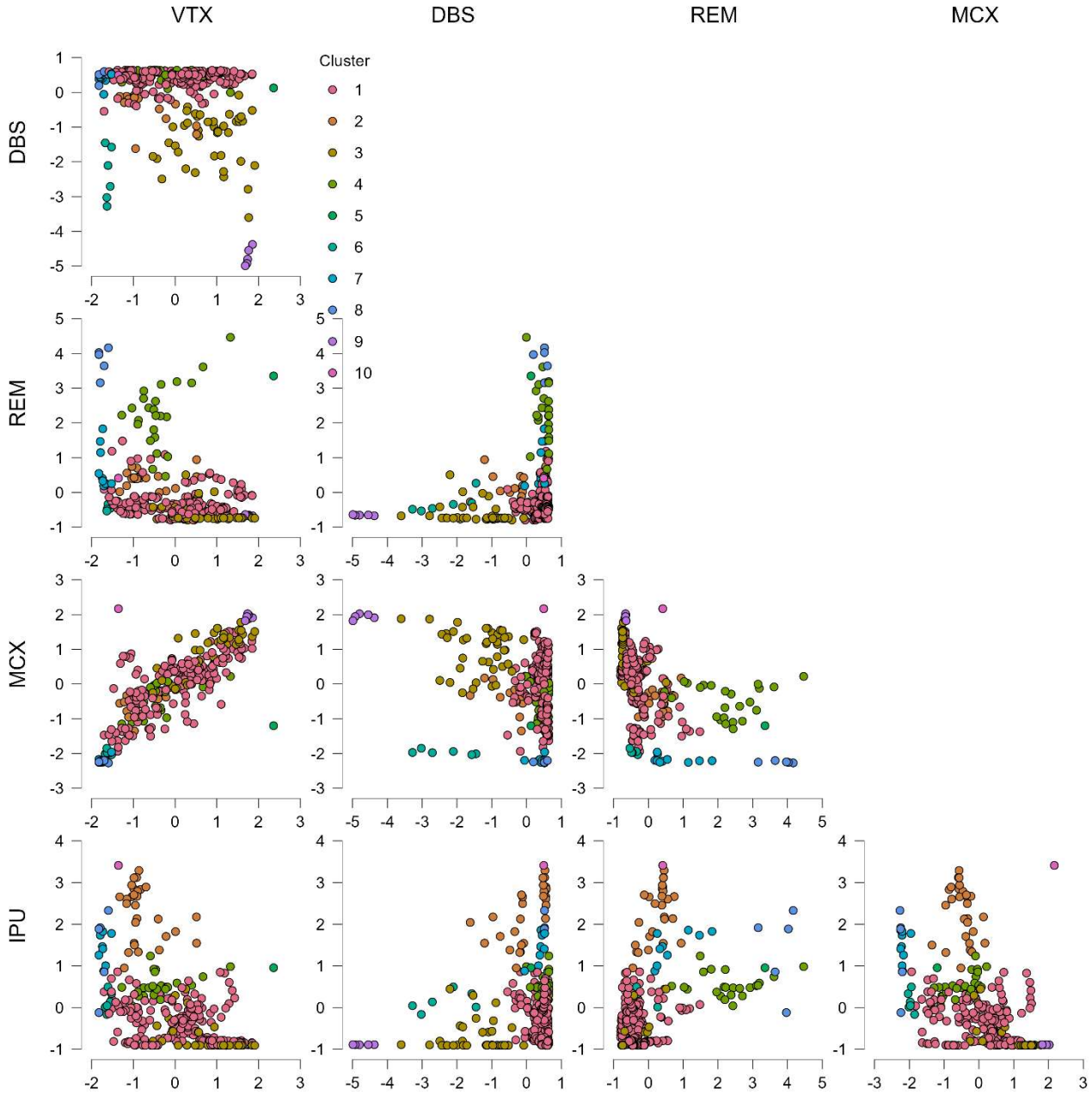
The figure presents a comprehensive visual interpretation of the results obtained from the hierarchical clustering for the model with the inclusion of VTX, DBS, REM, MCX, and IPU, bringing together the information obtained for cluster selection, cluster structure, and economic interpretation. Panel A presents the evolution of information criteria and the sum of squares for each cluster, depending on the number of clusters considered for the partition. The downward trend represents the improvement in the goodness-of-fit measure for the model, while the highlighted minimum represents the optimal number

of clusters, near ten, where the model balances goodness-of-fit and parsimony considerations for the partition. This result supports the selection of a rich cluster structure, capable of capturing the heterogeneity present in the data set. Panel B presents the clustered observations in a reduced dimensional space, where different colors are used to differentiate the clusters. The clear differentiation between some of the clusters supports the evidence that the hierarchical algorithm does not partition the sample mechanically but recognizes patterns in the data set. Some clusters are more compact and well differentiated, while others show more overlapping patterns, consistent with the evidence of a large, heterogeneous core group, along with smaller, more differentiated groups of observations. Panel C presents the dendrogram, where the hierarchical structure of the clustering algorithm is presented, showing the merging of the observations and the different groups obtained during the process, where broad branches are associated with more general patterns, while smaller branches are associated with more idiosyncratic patterns. The level at which the branches merge represents the distance between the groups, while the existence of some long vertical jumps suggests that the clusters are genuinely different in terms of the underlying financial and market characteristics. Panel D displays the standardized cluster means, which are the economic interpretation of the clusters. The differences in the clusters are evident in terms of the stock market breadth (VTX), relative importance of deposit money banks (DBS), remittance inflows (REM), market concentration (MCX), as well as the relative importance of international public debt (IPU). Clusters are characterized by diversified structures, with high values in terms of market concentration (MCX) and relative importance of deposit money banks (DBS), as well as low values in terms of remittance inflows (REM), while other clusters are characterized by the opposite: narrow structures, with low market concentration (MCX) and relative importance of deposit money banks (DBS), as well as high values in terms of remittance inflows (REM) and relative importance of international public debt (IPU). The extreme positive or negative values in the clusters indicate the existence of very specific financial structures, as suggested by the small, well-separated groups in the other panels. The figure suggests that the methodology of hierarchical clustering detects a rich segmentation in the data, highlighting the coexistence of a large group of similar observations, as well as a few smaller, more differentiated clusters, driven by different financial development, market structures, and external financial integration, as captured by the underlying five variables.



The resulting scatter plot matrix provides a detailed view of the relationships between these five standardized financial variables, VTX, DBS, REM, MCX, and IPU, with the data points colored according to their clusters determined using hierarchical clustering. This form of visualization is useful not only for evaluating the internal consistency of these clusters but also for understanding the ways in which distinct financial dimensions interact with each other. Several distinct financial relationships can be determined. For instance, there is a strong positive relationship between MCX and VTX, indicating that an expansion of stock exchange trading activities beyond the largest firms is related to a less concentrated and more diversified financial system. This is true across all clusters, with some clusters residing at more extreme levels. In terms of DBS, there is a more nuanced relationship. Clusters with high levels of DBS, indicating a stronger financial system dominated by deposit money banks relative to the central bank, are typically found at moderate to high levels of MCX and low levels of REM. This indicates a stronger domestic capital market and less reliance upon remittance flows. By contrast, clusters with low DBS levels, indicating a less important role played by domestic banking, are more frequently associated with high levels of REM and, at times, high levels of IPU. From the REM panels, we can see that there is an obvious distinction between clusters that have high remittance dependence and those that have relatively lower remittances. Clusters that have high REM values tend to be located in areas that have lower values of VTX and MCX, indicating that they have relatively narrower and more concentrated market activity and tend to have higher IPU values, indicating more dependence on international public debt. Again, this suggests that remittance-dependent countries tend to have less developed domestic financial markets and more developed links with external sources of finance. The IPU relationships also support this view of segmentation into remittance-dependent and less remittance-dependent countries. Clusters with high IPU values tend to group together with high REM values and lower values of VTX and MCX, while clusters with low IPU values tend more often to group together with stronger values of

domestic market activity and DBS. The colored point clouds of each of the variables also suggest that the hierarchical clustering has successfully captured significant multivariate structure and not simply arbitrary groupings of countries. Each of the clusters is located in relatively distinct areas of the variable space, with considerable overlap in the central or more "average" areas of each distribution. In summary, the scatterplot matrix of the data set visually confirms that the hierarchical clustering has successfully captured significant and coherent financial and market structure, distinguishing between market-oriented systems that have well-developed banking systems and relatively well-developed equity markets and systems that tend to be more externally driven and have relatively higher remittances, relatively higher values of international public debt, and relatively lower values of domestic market activity.



6. Machine Learning Performance and Variable Importance Analysis

The normalized performance metrics provide a clear and consistent basis upon which all the algorithms can be compared along various dimensions of predictive accuracy and goodness of fit. All error-based performance metrics are standardized such that higher values indicate better performance, and the same is true of the R^2 metric. It is, therefore, possible to evaluate all algorithms along a common framework with regard to their predictive accuracy and goodness of fit. It is possible to immediately rule out the neural networks based on their performance metrics, as they indicate a score of zero across all metrics, implying comparatively inferior performance compared to all other algorithms. The boosting regression model is found to possess moderate performance, with all error-based metrics indicating values closer to the midpoint, as well as a low value of R^2, indicating that it does not perform better than basic models. The decision trees indicate comparatively better performance along dimensions of MSE, RMSE, and MAPE, but with a low value of R^2. However, the relevant comparison here is with KNN, Random Forest, and SVM, as they are the ones that are ranked the highest. KNN achieves the maximum or near-maximum score for a number of error metrics, such as MAE, MAPE, etc., and has exceptional scores for MSE, RMSE, and R^2, implying local prediction accuracy with minimal average error. Similarly, the Random Forest achieves the maximum score for scaled MSE and R^2, and near-maximum scores for RMSE and MSE, implying high overall prediction accuracy. Although the SVM achieves high scores for MSE and RMSE, along with high R^2, its performance for MAPE is extremely poor, with a zero score, implying significant problems in terms of percentage error. Thus, if all the metrics are considered, the overall best-performing algorithm is the Random Forest, as it avoids zero scores for any error metrics, is near the top for all error metrics, and achieves the maximum score for R^2, implying the maximum proportion of variance explained by the model. Although the KNN might have a slight edge over the Random Forest for local error metrics such as MAE and MAPE, the overall performance of the Random Forest is significantly better than the other two, with minimal problems in terms of error, unlike the SVM and KNN, and therefore can be considered the best-performing model for those problems where overall prediction accuracy along with R^2 is important.

Metric (normalized)	Boosting	Decision Tree	KNN	Random Forest	Linear Reg.	Neural Net	Lasso	SVM
MSE	0.496	0.838	0.970	0.944	0.848	0.000	0.851	1.000
MSE (scaled)	0.447	0.705	0.958	1.000	0.733	0.000	0.838	0.951
RMSE	0.503	0.834	0.974	0.947	0.850	0.000	0.853	1.000
MAE / MAD	0.486	0.726	1.000	0.794	0.798	0.000	0.760	0.918
MAPE	0.000	0.995	1.000	0.000	0.000	0.928	0.969	0.000
R^2	0.355	0.631	0.941	1.000	0.657	0.000	0.783	0.931

The feature importance metrics offer a vivid and informative view of the contribution of the four explanatory variables, namely MCX, REM, IPU, and DBS, towards the predictive capability of the model. The model is presumably a tree-based ensemble model, e.g., a Random Forest. The metrics offer a nuanced view of the contribution of each variable towards the dependent variable. The Mean Decrease in Accuracy (MDA) is a measure of the decrease in predictive accuracy when the values of a particular variable are randomly permuted. The higher the value, the more important is the variable. MCX has a much higher value of 495.125, more than three times higher than REM and IPU, and nearly nine times higher than DBS. This indicates that the variable representing market capitalization excluding the top ten companies has a significant contribution towards the predictive accuracy of the model. When information is fed into MCX at random, the capability of the model to predict or explain the dependent variable decreases correspondingly. The Mean Decrease in Accuracy of DBS is 55.195, indicating that DBS is the least important of the four explanatory variables and, therefore, has played a relatively less decisive role. The Total Increase in Node Purity (TINP) is usually computed as the decrease in residual sum of squares over all regression tree splits for a particular variable. MCX has again recorded the highest value of 24,459.853, indicating its significance. REM and IPU recorded values of 11,351.301 and 10,534.754, respectively, indicating their contribution towards the model. The importance of remittance flows (REM) and international public debt (IPU) is reflected here. The contribution of these two variables towards splitting the trees towards homogeneity is high, but not as high as MCX. DBS has recorded a much lower TINP of 6,851.254, indicating a relatively minor role played by the domestic business sector. The Mean Dropout Loss, which is computed from the root mean squared error when the variable is dropped from the model, is another measure of variable importance. The higher the dropout loss, the more significant is the effect of dropping that variable from the model. Again, MCX has recorded the highest importance value of 22.182, which reconfirms its importance in the model. REM and IPU have recorded an identical importance value of 12.734 and 12.111, respectively, indicating their relatively comparable importance value. DBS has recorded the lowest importance value of 10.703. These three importance values of variables give an overall picture of variable importance, and they clearly indicate that MCX is the most important variable in influencing the model's prediction, while REM and IPU have recorded intermediate importance values, which is in conformity with their relatively lower contributions as external private and public financial flows, respectively. DBS has recorded the lowest importance value among all variables, though its importance is not negligible. The three importance values of variables converge, which gives more confidence in their interpretation and clearly indicates that the model's prediction is not unduly affected by any one of the importance values of variables, which may not always be true in certain situations.

Variable	Mean Decrease in Accuracy	Total Increase in Node Purity	Mean Dropout Loss
MCX	495.125	24,459.853	22.182
REM	142.043	11,351.301	12.111
IPU	144.458	10,534.754	12.734
DBS	55.195	6,851.254	10.703

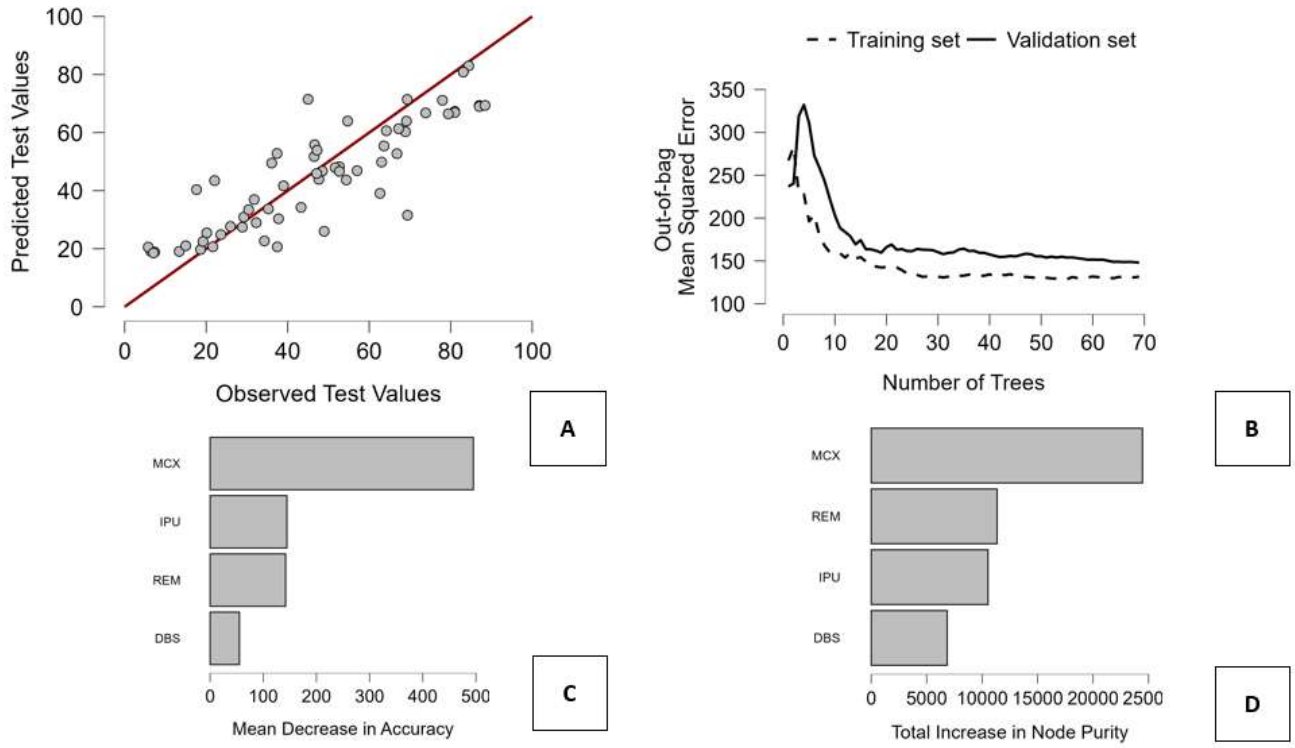
Note. Mean dropout loss (defined as root mean squared error, RMSE) is based on 50 permutations.

The table reports a decomposition of the predicted values of the dependent variable into a baseline component and the individual contributions of the explanatory variables DBS, REM, MCX, and IPU for five representative cases. The baseline value is constant across all cases at 42.517, which can be interpreted as the reference prediction in the absence of deviations in the explanatory variables, while the final predicted value is obtained by adding the variable-specific contributions to this base.Case 1 illustrates a situation in which all variables contribute positively to the prediction. Starting from the baseline of 42.517, the combined positive effects of DBS (0.681), REM (1.081), MCX (7.497), and IPU (8.518) raise the predicted value to 60.295. The largest contributions come from MCX and IPU, confirming the strong influence of market structure and international public debt in shaping the outcome. This case represents a profile where both equity market diversification and external financing conditions significantly boost the predicted level of the dependent variable. Cases 2 and 3 show the opposite pattern, with predicted values well below the baseline. In Case 2, negative contributions from all variables, especially MCX (−11.650) and REM (−4.830), drive the prediction down to 19.734. A similar configuration appears in Case 3, where large negative effects from REM (−5.938), MCX (−9.263), and IPU (−5.183) reduce the predicted value to 20.671. These two cases highlight how unfavorable positions in market diversification and external flows can substantially depress the outcome relative to the baseline level. Case 4 presents the highest predicted value, 66.774, driven by very strong positive contributions from REM (5.970) and especially MCX (17.398), alongside a positive effect from DBS (1.092). The small negative contribution from IPU (−0.202) is negligible in comparison. This case underscores the dominant role of equity market diversification in lifting the prediction far above the baseline, with remittances also playing an important supporting role. Finally, Case 5 shows a more mixed configuration. Although DBS contributes negatively (−0.735) and IPU also exerts a downward effect (−3.285), positive contributions from REM (1.168) and MCX (7.075) are sufficient to keep the predicted value slightly above the baseline at 46.740. Overall, the decomposition confirms that MCX and REM are the most influential drivers of variation around the baseline, while DBS and IPU play more secondary but still meaningful roles in shaping the final predictions.

Case	Predicted	Base	DBS	REM	MCX	IPU
1	60.295	42.517	0.681	1.081	7.497	8.518
2	19.734	42.517	-0.966	-4.830	-11.650	-5.338
3	20.671	42.517	-1.462	-5.938	-9.263	-5.183
4	66.774	42.517	1.092	5.970	17.398	-0.202
5	46.740	42.517	-0.735	1.168	7.075	-3.285

The figure presents the main results obtained from the random forest model and also offers additional information regarding the model's performance in terms of prediction and the relative importance of the explanatory variables used in the model. Panel A in the figure demonstrates the relationship between observed and predicted test values. The dispersion of the points around the 45-degree line in the graph suggests that the model's performance in reproducing observed values is reasonably high. Panel B in the figure presents the evolution of the out-of-bag mean squared error for increasing numbers of trees in the random forest model for both the training and the validation sets. It is clear from the graph in Panel B that initially, the error is high and fluctuates significantly for a small number of trees in the model, but it decreases sharply and stabilizes gradually for increasing numbers of trees in the model, which is consistent with the random forest model's performance and its ability to benefit from the results obtained

from increasing the ensemble of decision trees used in the model. Moreover, the close proximity of the curves for the validation and training sets in the graph in Panel B suggests good generalization performance and little overfitting in the model. However, beyond a certain point, the rate of improvement slows, indicating that the model has reached a level of stability and reliability. The variable importance plots, as represented in Panels C and D, provide a similar, albeit distinct, reading. Both plots show MCX as having the greatest influence, with a significant lead over the other predictors. This again points towards the fact that the structure and diversification of the equity market are significant contributors towards explaining the dependent variable. REM and IPU follow with a fair degree of distance, indicating that external private flows as well as international public debt contribute towards explaining the dependent variable. DBS has the least influence, indicating that, although relevant, the relative size of deposit-taking banks has a lesser impact. The four plots provide a cohesive picture. The random forest model has shown satisfactory levels of predictive capability and stability, improving with an increase in ensemble size while still maintaining a reasonable gap between training and validation error. It has also provided an economically relevant reading, with MCX having the greatest influence, indicating the relevance of the structure and diversification of the equity market. The secondary level of influence of external financial flows is also relevant. The random forest approach can thus be said to be a reliable method of analyzing the dependent variable.



7. Integrated Evidence on the Determinants of Stock Market Trading Diversification

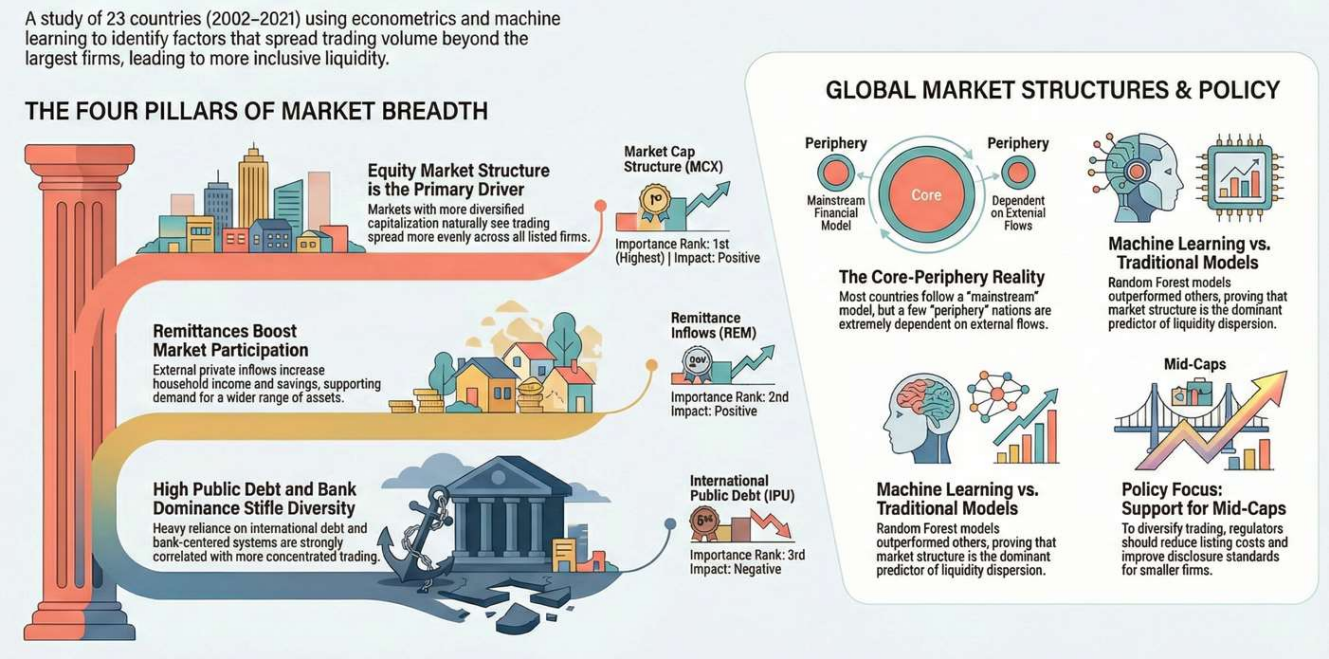
The results of the empirical analysis, which combines panel econometrics, hierarchical clustering, and machine learning, provide a comprehensive evaluation of the determinants of stock market trading diversification, defined in terms of the share of trading activity accounted for by firms outside of the top ten, or VTX. These three approaches collectively offer consistent, complementary, and mutually supporting evidence on the structural determinants of stock market trading diversification and the heterogeneity of financial systems across countries. The panel econometric results clearly reveal a set of significant insights. In particular, with both the Fixed Effects and Random Effects model specifications, MCX, or market capitalization excluding the top ten firms, and REM, or remittance inflows to GDP, show strong, highly significant positive effects on VTX. These results suggest that a more diversified stock market structure, along with stronger external private financial inflows, are positively correlated with stock market trading diversification. By contrast, DBS, or the relative size of deposit money banks, and IPU, or outstanding international public debt, show negative, significant effects on VTX, indicating that more bank-based financial systems, along with stronger external public financial inflows, are positively correlated with stock market trading concentration dominated by a small number of firms. Overall, the panel econometric results suggest that market structure is the most important determinant of stock market trading diversification, with external private and public financial flows playing a secondary, supporting role. The clustering analysis provides a significant structural complement to these results, highlighting significant heterogeneity in financial systems and markets. Hierarchical clustering is identified as having the best performance according to several validity criteria, with a clear uneven distribution of cases across clusters. Most countries are found in a single large cluster with average levels of VTX, DBS, REM, MCX, and IPU, while several small clusters contain countries with very distinctive financial systems. These results suggest significant segmentation in the data, with one group of systems having high levels of MCX and diversified markets, another with high levels of DBS and low dependence on external sources, and a third with high levels of REM and IPU, along with low levels of VTX and MCX. This structure suggests that, while many countries have similar financial structures, a number have extreme or unique configurations that may be economically significant. These results complement the econometric analysis, which found average relationships across countries, with significant heterogeneity in financial development systems. The results of the machine learning analysis further support these results from a predictive standpoint. Among the models considered, the Random Forest model stands out as the most balanced and reliable model, achieving the best performance in terms of fit and error metrics. Although the other models, such as KNN and SVM, perform well in some metrics, they perform poorly in others, while the Random Forest model performs well in all metrics considered. Moreover, the importance of the variables also points to the same conclusion, where MCX is clearly the most important predictor of VTX, followed by REM and IPU, and finally, the importance of DBS is relatively minor. These results are consistent with the results obtained from the econometric model and clearly support, from a purely data-driven standpoint, the importance of equity market structure in explaining trading diversification.The results of the decomposition of the predictions also support the importance of MCX and REM in explaining the deviations from the baseline, while the importance of DBS and IPU is relatively minor. Collectively, the three methods produce a coherent and robust result. The econometric analysis determines the direction and significance of major relationships, the clustering analysis detects structural heterogeneity and variety of financial models across countries, and machine learning validates the importance of market structure while improving prediction accuracy. The agreement of three approaches gives stronger support to the main finding: stock market trading diversification is mainly driven by the internal structure of equity markets, supplemented by external private financial flows, and limited by bank dominance and use of international public debt.

Method	Main Findings	Economic Interpretation

Panel Econometrics (FE/RE)	MCX (+) and REM (+) significantly increase VTX; DBS (−) and IPU (−) reduce VTX. Fixed Effects preferred.	A more diversified equity market and higher remittances broaden trading; bank dominance and external public debt are associated with more concentrated markets.
Hierarchical Clustering	Core–periphery structure: one large cluster with average profiles and several small clusters with extreme configurations (high MCX, high REM/IPU, or high DBS).	Countries follow different models of financial development: market-oriented, bank-centered, or externally dependent systems with distinct trading structures.
Machine Learning (Random Forest)	Best overall predictive performance. Variable importance: MCX ≫ REM ≈ IPU > DBS.	Equity market structure is the dominant driver of trading diversification; external flows matter, banking structure plays a secondary role.
Prediction Decomposition	MCX and REM explain the largest deviations from baseline predictions; DBS and IPU have smaller but non-negligible effects.	Confirms the central role of market structure and the supporting role of external financial flows.

The results provide a clear and textured picture of the determinants underlying the diversification of trading activity in the stock market, and they suggest several interesting economic implications. The econometric analysis suggests that the internal structure of the equity markets plays an important role in explaining the distribution of trading activity across firms, and the strong and statistically significant positive impact of the MCX term supports the close link between market depth and breadth implied by the positive and significant MCX term. Similarly, the positive impact of the remittance term suggests the potential for financial flows to facilitate market participation, possibly through their impact on household incomes, savings, and demand for financial assets. The negative and significant coefficients on the DBS and IPU terms, however, highlight the structural constraints associated with bank-based financial architectures and the importance of international public debt, and suggest the potential for financial architectures based on bank-based financial systems or those with significant external debt to fail to generate adequate incentives for market participation, a result supported by the clustering analysis, which suggests significant cross-country heterogeneity in the average relationships. The presence of a substantial core cluster and other smaller and more specialized clusters points to the fact that countries are following their own path of financial development rather than converging towards a common financial structure. Financial systems with a high level of MCX tend to show diversified financial trading, whereas bank-oriented and externally dependent financial systems show a narrower financial market. This dual nature points to the significance of institutional and financial system differences, and policy recommendations on financial market diversification need to be made at the national level. The results obtained through the machine learning methodology also corroborate the findings from a different perspective. The superior performance of the RF model, along with the dominance of MCX in the variable importance plots, points towards the fact that the structure of the financial equity market is the most important predictor of financial trading diversification. The significant contribution of REM and IPU, with DBS playing a secondary role, also corroborates the earlier findings from a different perspective, thereby reiterating the earlier conclusions and lending more support to the overall

interpretation. The fact that the findings obtained through three different methodologies show a high degree of consistency points towards the fact that the findings are not likely to be a result of a particular model specification, rather they strengthen the earlier findings and reiterate the role of financial structure, external private flows, and bank dominance and external public debt in determining the stock market trading activities.



8. Policy Implications for Promoting Broader and More Inclusive Equity Markets

The policy implications of the present study's findings are significant, and the key ones are discussed below: The internal structure of equity markets is seen to be at least as important as the size of the markets, and this is a very significant policy implication, especially considering the fact that the persistent and significant influence of the market capitalization excluding the largest firms suggests that any policy initiatives focused on raising the size of the markets, measured by either the total market capitalization or the trading volumes, might not be enough if the trading activity continues to be dominated by a small number of large companies. Hence, there is a need for the authorities to attach greater policy importance to initiatives that might promote the listing and visibility of small- and medium-sized companies, and this might include initiatives such as lowering the listing costs, proportionately improving the disclosure norms, and promoting the market-making activity for such companies, so that the overall conditions are created for the trading activity to be more evenly distributed. The positive relationship between remittances and market participation suggests that external private financial flows have the capacity to sustain broader market participation, provided these flows are effectively integrated into the domestic financial system. This result reinforces the relevance of policies that reduce transaction costs in the

remittance system, promote the financial inclusion of remittance-receiving households, and increase the propensity of these households to utilize formal financial instruments in saving and investment. Strengthening the link between remittance flows and the domestic capital market, optionally through the creation of appropriate savings or investment vehicles, could help these flows become a catalyst for the development of deeper, more diversified financial systems. The converse, however, is that the negative relationship between trading system diversification and bank dominance as well as international public debt suggests that there could be structural trade-offs in the design of the financial system. Banks and sovereign public debt are critical components in the financing system, but a dominant position in the financial system by banks or a significant reliance on international public debt could be linked with lower levels of development in the domestic equity market as well as the trading system. This suggests that a policy environment that encourages a balanced financial system, in which banks, bond, and equity markets co-evolve, could be beneficial in terms of market participation. This could be achieved, in practice, through the establishment of a regulatory environment that avoids biasing the system in favor of the banking sector, as well as a debt management strategy that minimizes the risks of over-reliance on external public debt. The significant heterogeneity across countries in terms of their levels of financial development, which was revealed in the above clustering analysis, also carries significant policy messages. Countries do not appear to be moving towards a unique model of financial development; instead, they have their own unique configurations. This suggests that there is no one-size-fits-all solution in terms of policies that could be used to enhance trading diversification. Market-based, bank-based, and external dependency systems have their own unique constraints and opportunities, and policies should be fashioned in light of these. For some countries, it may be necessary to enhance domestic equity markets, whereas for others, it may be necessary to enhance the integration of external financial systems with domestic financial systems. Lastly, the convergence of the econometric, clustering, and machine learning findings suggests that policymakers might gain from the integration of traditional analytical tools with data-driven methods in the design and evaluation of financial development strategies, particularly in the identification of the structural factors that are most informatively related to market outcomes. In conclusion, the findings suggest that the development of a more diversified and inclusive trading structure calls for a comprehensive policy approach, addressing the structure of the markets, the balance of the financial system, and the efficient use of external financial flows, rather than the level of overall market expansion.

9. Conclusions

The current research focuses on the relatively neglected aspect of financial development, which is related to the distribution of trading activities in stock markets. By examining the proportion of trading activities controlled by firms other than the ten largest ones, the current research moves beyond the conventional indicators of financial market size and liquidity and offers a more structural approach to financial markets. The main objective of the current research is to identify the main financial and structural determinants of trading diversification and examine their robustness over alternative specifications and data-driven methods. The main method used in the current research is based on the data set obtained

from the World Bank's Global Financial Development Database, which starts with 38 OECD countries for the 2002-2021 period and results in 23 balanced panel data due to data availability constraints. Moreover, the current research selects the variables in such a way that they cover the main financial systems in terms of their related dimensions, which include the role of deposit-taking banks, the importance of external private financial flows in terms of remittance income, the structure of the stock market, and the importance of international public debt. The results obtained from the econometric methods used in the current research clearly indicate that the main message conveyed by the results is related to the positive and negative associations between financial diversification in the structure of stock markets and the level of remittance income and between financial concentration in the structure of stock markets and the level of deposit-taking banks and international public debt. Aside from the average effects, the analysis of financial heterogeneity suggests that countries do not follow a single path in their development. The results from the clusters indicate a core-periphery pattern, in that most countries tend to have broadly similar characteristics, while a few countries have more differentiated financial structures. The underlying cause of these differences is the existence of different financial development models, ranging from market-based systems to bank-based or external-oriented ones. This, in turn, suggests that the reason why different instruments may have different effects in different countries is due to the existence of financial heterogeneity. The analysis based on the machine learning technique confirms the results based on the explanatory analysis. The superior performance of the RF model, as well as the prominent position of equity market structure in the list of the most important variables, again confirms that the most important factor in explaining the level of trading diversification is the internal configuration of the stock market. The fact that, aside from the equity market structure, the most important variables are remittances and international public debt, while the banking structure plays a minor role, again confirms the results based on the econometric analysis. The fact that the results based on the explanatory and the predictive analysis are consistent suggests that the results are driven by the underlying financial structures, rather than the methodology used. In combination, all these findings point towards a coherent conclusion. The major driver of stock exchange trading diversification is internal structure, supported by external private financial flows, with bank dominance and reliance upon international public debt as limiting factors. By combining traditional econometric methods with clustering and machine learning, this article has shown how complementary approaches can contribute to our understanding of average relations as well as underlying structural heterogeneity. Extensions of this framework could include expanding country coverage, incorporating dynamics, or using firm-level data to shed light upon the microeconomic channels through which financial structure affects participation and liquidity distribution. All these approaches could contribute towards a better understanding of financial system evolution towards more inclusive capital markets.

10. References

Abdullayev, I., Akhmetshin, E., Hajiyev, E., Khorolskaya, T., & Lydia, E. L. (2025). A financial time series forecasting model using quasi-recurrent neural networks and the crown porcupine optimizer for stock market risk prediction. Engineering Technology and Applied Science Research, 15(6), 29035–29040.

Abid, I. (2025). The interplay of entrepreneurship, investment, credit, and market capitalization in shaping sustainable economic growth: An ARDL approach for the United States. Scientific Annals of Economics and Business, 72(4), 633–649.

Adamolekun, G., Sakariyahu, R., & Ahmed, A. (2025). National energy generation capacity and stock market development. Review of Development Finance, 15(2), 1–15.

Afshan, S., Yaqoob, T., Ben Zaied, Y., & Ul-Durar, S. (2025). Pathway to environmental resilience: Analyzing financial dimensions to curb energy security risks. Journal of Environmental Management, 395, 127745.

Ben Salem, M., Alsagr, N., Belkhaoui, S., & Farhani, S. (2025). Quantile connectedness between stock market development and macroeconomic factors for emerging African economies. International Journal of Financial Studies, 13(4), 224.

Bisiriyu, S. O., Ismail, N. B. M., & Ramachandran, R. (2025). Highs, lows, and uncertainty: A deep dive into India's stock market and policy uncertainty. Discover Sustainability, 6(1), 1024.

Budiarso, N. S., & Pontoh, W. (2025). Market efficiency of dividend-paying firms under hawkish monetary policy: The case of Indonesia. Investment Management and Financial Innovations, 22(4), 335–344.

Gong, X.-L., Ning, H.-Y., & Xiong, X. (2025). Research on the cross-contagion between international stock markets and geopolitical risks: The two-layer network perspective. Financial Innovation, 11(1), 23.

Horton, R. (2025). Offline: Watching the watchers (part 1). The Lancet, 406(10520), 2613.

Jiang, K., & Yang, Q. (2025). Firm-level bubble detection in the PV industry via GSADF: A data-driven perspective. In Proceedings of the 2025 International Conference on Economic Management and Big Data Application (ICEMBDA 2025) (pp. 314–318).

Jurkšas, L., & Kaminskas, R. (2025). Communication of ECB Governing Council members: Impact on intraday financial markets from media messages. Journal of Behavioral and Experimental Finance, 48, 101117.

Li, Y., & Guo, J. (2025). From financing relief to governance empowerment: How do government-guided funds activate labor investment efficiency? International Review of Economics and Finance, 104, 104749.

Liu, X., & Yang, J. (2025). Research on A-share index prediction based on frequency domain decomposition and temporal fusion transformer model. In Proceedings of the 2025 International Symposium on Artificial Intelligence and Computational Social Sciences (AICSS 2025) (pp. 507–511).

Ma, X. (2025). Income changes and participation in risky financial markets: Evidence from China. Asian Economic Journal, 39(4), 529–559.

Mahmoudi, A., & Torra, M. (2025). Shaping stability: Can the finance-growth nexus achieve it? Managing Global Transitions, 23(4), 389–421.

Manoylenko, O., & Kuznetsova, S. (2025). Identifying factors impact on investment in financial services under digital financial ecosystem transformation. Technology Audit and Production Reserves, 6(4), 22–30.

Miao, L., Wang, J., Wu, K., Lu, G., & Kwan, M.-P. (2025). Spatiotemporal hybrid deep learning for estimating and analyzing carbon stocks: A case study in Jiangsu province, China. International Journal of Digital Earth, 18(1), 2534008.

Mu, Y., Fu, B., & Hu, Q. (2025). The determinants of limited household participation in risky financial markets: Evidence from China using explainable machine learning. Journal of Risk and Financial Management, 18(12), 686.

Nieborak, T. (2025). Digital coercion? The financial market and the right to digital opt-out between fiction and reality. Białostockie Studia Prawnicze, 30(4), 119–136.

Ridwan, M. (2025). Artificial intelligence and green development: The role of financial market efficiency in the United States. Development and Sustainability in Economics and Finance, 8, 100099.

Rostek, M., & Yoon, J. H. (2025). Imperfect competition in financial markets: Recent developments. Journal of Economic Literature, 63(4), 1191–1243.

Saucedo Loera, L. A., Oropeza Tagle, M. Á., & Montesinos Julve, V. (2025). Corporate social responsibility and corporate reputation in the Mexican stock market. Revista de Métodos Cuantitativos para la Economía y la Empresa, 40.

Soltani, H., Ben Ameur, A., & Abbes, M. B. (2025). Evaluating dynamic connectedness among economic sanctions sentiment, uncertainty factors, and financial assets: A quantile VAR approach. Scientific Annals of Economics and Business, 72(4), 595–613.

Tang, Z., Cao, K., Huang, Q., Li, X., & Zhu, B. (2025). Topological data analysis: Technical principles, financial applications, and future developments. In Proceedings of the 2025 International Conference on Economic Management and Big Data Application (ICEMBDA 2025) (pp. 238–243).

Tankov, P., & Zhang, R. (Eds.). (2025). Handbook of quantitative sustainable finance (pp. 1–498). Springer.

Tsybuliak, A., Panchenko, V., Shepel, O., Minkovska, A., & Buiak, L. (2025). Evaluating digital financial inclusion's role in sustainable development and environmental conservation. African Journal of Applied Research, 11(7), 66–84.

Ullah, W. (2025). Economic growth volatility: Is financialization a culprit? Russian Journal of Economics, 11(4), 381–402.

Wang, Y., Li, Y., & Zhuang, J. (2025). Research on the green development path of prefabricated building industry based on intelligent technology. Engineering, Construction and Architectural Management, 32(12), 8390–8422.

Wu, R., Zeng, H., Abedin, M. Z., & Ahmed, A. D. (2025). The impact of extreme climate on tourism sector international stock markets: A quantile and time-frequency perspective. Tourism Economics, 31(8), 1598–1628.

Zuo, J. (2025). Tesla stock price dynamic prediction model based on Monte Carlo simulation and LSTM neural network. In Proceedings of the 2025 International Conference on Economic Management and Big Data Application (ICEMBDA 2025) (pp. 360–366).