SECTORAL ANALYSIS OF FDI DEPENDENCE RISK STRUCTURE IN VIETNAM: EMPIRICAL EVIDENCE FROM PCA AND MACHINE LEARNING

Do Ngo Duy

Department of Mathematical Economics, Thuongmai University, Hanoi, Vietnam

DOI: https://doi.org/10.56293/IJMSSSR.2026.6207

IJMSSSR 2026
VOLUME 8
ISSUE 3 MAY - JUNE **ISSN: 2582 – 0265**

Abstract: In the process of extensive integration, Foreign Direct Investment (FDI) has become a key driver of Vietnam's economic growth. However, the rapid expansion and concentration of FDI in several core sectors also raise concerns regarding long-term structural dependence risks. This study constructs a sectoral FDI Composite Risk Index (CRI) using Principal Component Analysis (PCA) based on a dataset of 3,932 observations from 1986–2025. The results indicate that the first five principal components explain 87.98% of the total variance, reflecting the existence of a capital accumulation axis that dominates sectoral structures. Testing through Machine Learning models reveals that the relationship between FDI accumulation and dependence risk is non-linear, with the Gradient Boosting model achieving an accuracy of 80.05%, significantly outperforming traditional linear models. The findings imply a need for a sectoral FDI concentration monitoring mechanism to mitigate dependence risks and ensure sustainable development.

Keywords: FDI, dependence risk, sectoral structure, PCA, Machine Learning.

1. INTRODUCTION

After nearly four decades of economic opening, the Foreign Direct Investment (FDI) sector has become a vital catalyst for Vietnam's economic growth and structural transformation. This sector currently plays a leading role in industrial production and exports, facilitating deep integration into global value chains. Theoretically, Alfaro et al. (2004) asserted that FDI only truly promotes growth when the host economy possesses sufficient absorptive capacity to integrate capital flows into its internal structure. Thus, the impact of FDI depends not only on its scale but also on its degree of integration and position within the sectoral structure.

In Vietnam, research by Nguyen Hoang (2025) suggests that FDI can improve domestic productivity through technology spillover channels, though this effect is conditional and uneven across sectors. As FDI increasingly concentrates in strategic sectors like manufacturing, logistics, and infrastructure, the question arises: beyond its contribution to growth, to what extent is the economy's structural dependence on the foreign investment sector increasing?

Most current studies still measure FDI using single indicators like total registered capital or the FDI/GDP ratio. This approach reflects scale but fails to quantify dependence risk—which stems from the simultaneous accumulation of multiple factors such as capital size, project density, duration of presence, and trade integration. When these factors concurrently reach high levels within a sector, the FDI sector may hold a dominant position in the value chain, increasing the sensitivity of the sectoral structure to external fluctuations.

Recent studies on composite risk indices and risk classification in emerging economies, notably Zhang et al. (2022), show that economic risks are often non-linear and exhibit threshold effects. This suggests that FDI dependence may not increase linearly but becomes evident once accumulation reaches a certain combined level.

Based on this reality, this study focuses on constructing a multi-dimensional sectoral FDI dependence risk index while testing the non-linearity of risk formation through Machine Learning models. The results are expected to provide a quantitative basis for designing policies to attract and manage FDI, balancing openness with the enhancement of domestic capacity and ensuring long-term economic structural stability.

2. THEORETICAL FRAMEWORK AND RESEARCH HYPOTHESES

In essence, FDI is not merely a supplementary capital flow; it comes with control rights, technology, and the ability to dominate value chains. Therefore, the impact of FDI does not solely depend on capital volume but on the integration level and the status of the FDI sector within the industry structure. When FDI enterprises hold high-value-added stages, account for a large share of production, and maintain a long-term presence, the sectoral structure may gradually form a state of systemic dependence.

This dependence process is not formed by a single factor but is the result of the simultaneous accumulation of multiple components. Large capital scale enables production expansion; high project density increases presence; long operational duration consolidates market position; while trade integration can amplify spillover effects. It is the interaction between these factors that shifts the power structure within a sector.

Regarding measurement, an approach based on a single indicator like the FDI/GDP ratio does not fully reflect the multi-dimensional accumulative nature of dependence risk. Therefore, the study utilizes a **Composite Risk Index (CRI)** structure to quantify the sectoral dependence state. Before aggregation, variables are normalized to ensure comparability:

$$Z_{ij} = \frac{X_{ij-min(x_j)}}{(x_j)-min(x_j)}$$

Where: Z_{ij} *is the normalized value of variable* j *at unit* i.

The dependence risk index for sector i is defined in a general form:

$$CRI_i = \sum_{j=1}^{p} w_j \cdot Z_{ij}$$

Where: Z_{ij} *is the normalized value of risk variable* j *in the model, and* w_j *is the weight coefficient of risk variable* j.

Economically, the CRI reflects not only the scale of FDI but also the degree of structural dominance when multiple components accumulate simultaneously. A sector with large FDI capital but low project density or short presence may not necessarily create high dependence. Conversely, when component variables have high loadings on a single accumulation axis, the CRI will increase sharply, reflecting the formation of structural dependence.

However, the relationship between FDI accumulation and dependence risk is not necessarily linear. In the early stages, capital increases may only produce a resource-supplementary effect. But once accumulation crosses a critical threshold, the FDI sector may shift from a supportive role to a dominant one. At that point, even a small change in scale or density can cause the risk probability to surge. This is the **threshold effect** in dependence structures.

To test this non-linearity, the study employs the **Gradient Boosting** model—an ensemble learning method based on the additive combination of multiple weak models. The general prediction function is represented as:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

Where: $h_m(x)$ *is the* m *– th weak model (typically shallow decision trees),* γ_m *is the multiplier adjusting the contribution of model* $h_{m(x)}$, *and* M *is the total number of component models.*

This iterative mechanism allows the model to step-by-step correct remaining errors, thereby replicating the non-linear interactions between capital scale, project density, and duration of presence. Consequently, the model can detect **threshold risk zones**—where CRI spikes when multiple factors simultaneously reach high levels.

Based on the theoretical foundation above, the study proposes the following hypotheses:

1. **H1:** The simultaneous accumulation of capital scale, project density, and duration of presence increases the probability of forming a high sectoral dependence risk state.
2. **H2:** The relationship between FDI accumulation and dependence risk is non-linear and exhibits threshold effects.
3. **H3:** The level of trade integration plays a conditional role in the relationship between FDI accumulation and dependence risk.

3. RESEARCH METHODOLOGY

3.1. Research Data

The study utilizes sector-year panel data for the period 1986–2025, comprising a total of 3,932 observations. Data were aggregated from the General Statistics Office (GSO), the Foreign Investment Agency (FIA), and the World Bank, ensuring authenticity, temporal continuity, and cross-sectoral comparability. The variable system was designed to reflect the multi-dimensional presence and accumulation of FDI, including capital scale, project density, participation intensity, and macro-control factors. Since the variables have different units of measurement and amplitudes, all data were normalized using the formula presented in the theoretical section. This processing step not only ensures uniformity in multivariate analysis but also establishes the foundation for constructing the composite index in the subsequent step.

To maintain the stability of the data structure, multicollinearity was tested using the Variance Inflation Factor (VIF) and correlation matrices; variables with high information overlap were adjusted accordingly. Upon completion of pre-processing, the dataset was split into a training set (70%) and a testing set (30%). This approach allows for in-sample model calibration while verifying out-of-sample generalization capability, ensuring the reliability of the empirical results.

3.2. Constructing the FDI Dependence Risk Index

Based on the normalized variable set, the challenge lies in quantifying the level of FDI dependence as a representative index capable of reflecting the overall accumulation structure rather than individual dimensions. To address this, the study employs **Principal Component Analysis (PCA)** to reduce data dimensionality and determine endogenous weights based on explained variance. After constructing the covariance matrix, eigenvalues and eigenvectors were calculated to identify the principal axes of variation. According to the Kaiser criterion, components with eigenvalues greater than 1 were retained to ensure statistical significance. The results show that the first five components explain 87.98% of the total variance, reflecting high representativeness of the reduced structure. Notably, the first component (PC1) accounts for the largest variance proportion and aggregates variables related to capital scale, project density, and FDI duration. This indicates that dependence risk is not formed by a single variable but from a structural convergence of accumulation. Based on the selected components, the CRI values were calculated according to the established synthesis formula and used as the central variable in the subsequent testing phase.

3.3. Model Design

To test the research hypotheses, particularly the relationship between CRI and sectoral risk states, the study implements a classification model system including both linear and non-linear methods. Using multiple models in parallel allows for a comparison of relational structures and an assessment of improvements in predictive performance.

3.3.1. Logistic Regression (Baseline Model)

Firstly, Logistic Regression was used as the baseline model to provide a comparison standard for non-linear models. The general form of the model is represented as:

$$P(X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p)}}$$

Where: Y Is the FDI risk level classification variable (k=0, 1, 2 corresponding to Low, Medium, High levels); X_1, X_2, ..., X_p *are input variables; 0 is the intercept; and 1, 2,…, p are regression coefficients.*

The role of Logistic Regression is to provide a linear benchmark, reflecting the relationship between the logit of risk probability and input variables. If non-linear models achieve significantly higher performance, it serves as empirical evidence confirming the existence of multivariate interactions and threshold effects that linear models cannot capture.

3.3.2. Random Forest

Next, Random Forest was applied to exploit its capability in modeling non-linear relationships through an ensemble of multiple independent decision trees. This method helps reduce predictive variance and enhances stability in the context of multi-dimensional data interactions.

3.3.3. Gradient Boosting

Finally, Gradient Boosting was implemented following the sequential training principle, where each base model is built to correct the residual errors of the previous step. This approach allows the model to flexibly capture complex interactions and separation structures within the data.

3.4. Model Evaluation Criteria

Classification performance was evaluated through **Accuracy** and a **Confusion Matrix** to analyze the ability to identify each risk group in detail. Additionally, the discrepancy between results on the training and testing sets was used to check stability and out-of-sample generalization. This ensures that empirical results reflect not only in-sample fit but also the robustness of the model when applied to new observations.

4. RESEARCH RESULTS

Based on the established framework, this section presents detailed empirical results on the structure and formation mechanism of sectoral FDI dependence risk. The analysis focuses on three key aspects: the stratification of risk across sectors, the role of accumulation factors in shifting risk probability, and evidence of threshold effects in the transition from a supportive state to a structural dependence state.

4.1. Stratified Structure of Sectoral FDI Dependence Risk

Industry Sector Risk Classification Summary Table

Industry Sector	CRI Mean	Risk Level	Total FDI (M$)	Projects
Manufacturing	0.337	HIGH	402214.8	28548
Infrastructure	0.166	HIGH	349141.4	23261
Logistics	0.129	HIGH	300680.7	20585
Agriculture	0.036	HIGH	283030.2	18574
Aquaculture	0.029	HIGH	283358.0	19082
Textile	-0.12	MEDIUM	220648.4	15011
Technology	-0.124	MEDIUM	216243.8	14094
Mining	-0.136	MEDIUM	205583.2	13460
Retail	-0.147	LOW	203115.9	13555
Real_Estate	-0.164	LOW	220467.7	14012

Figure 1: Average CRI Values and Sectoral Risk Classification

Empirical results show that FDI dependence risk is not evenly distributed but forms a distinct stratified structure. The CRI distribution ranges from high positive to deep negative values, reflecting significant differences in the degree of accumulation and the dominant role of the FDI sector within each industry. This confirms that FDI dependence is a structural phenomenon linked to a sector's position in the value chain and the multi-dimensional convergence of capital.

The **Manufacturing** sector stands out with the highest CRI value, forming the upper pole of the distribution. This is an area where capital scale, project density, and duration of presence converge simultaneously. Continuous accumulation over time increases the sector's linkage with the FDI area, thereby raising adjustment costs when investment structures change. In this context, dependence is reflected not only in scale but also in the depth of integration with production networks and cross-border value chains.

Infrastructure and **Logistics** follow in the high-risk group. Although their CRI is lower than manufacturing, their intermediary positions in the production-distribution system mean that FDI accumulation can spill over into linked industries. Dependence risk may thus become systemic when concentrated at the coordination nodes of the value chain.

The medium group, including **Textile, Technology, and Mining**, represents a transition state. These sectors participate in production chains but have not reached a level of convergence large enough to create clear structural dependence. Risk only increases when accumulation factors cross a certain threshold.

At the lower pole, **Real Estate** and **Retail** maintain deep negative CRI values. Their operational structures rely heavily on domestic demand and have high investor dispersion, reducing long-term accumulation potential. Even with large capital scale, fragmentation and short project lifecycles make structural dependence difficult to form.

4.2. Accumulation Mechanism and Risk Non-linearity

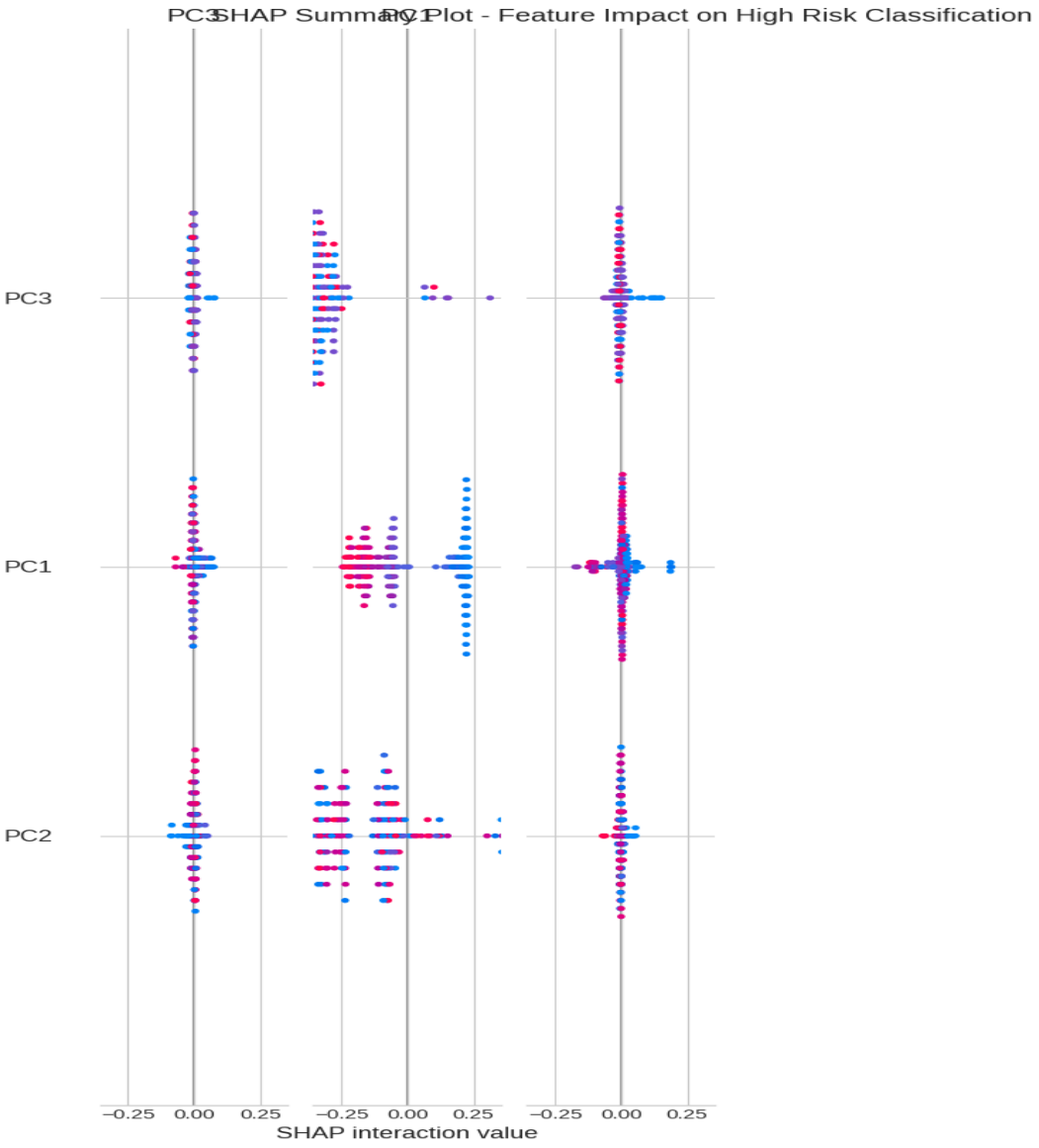

Figure 2: SHAP Summary Plot – Impact of variables on high-risk classification

In PCA, **PC1** emerges as the aggregate accumulation axis, carrying large loadings for capital scale, project density, and duration of presence. This shows that accumulation factors do not move independently but tend to be positively correlated within the same structural configuration.

SHAP analysis clarifies the micro-level operation of the PC1 axis. Capital scale and project density continue to show the largest marginal contributions to high-risk classification. Notably, in high-value regions, SHAP contributions shift sharply toward the positive side with wide dispersion, indicating that as accumulation reaches high intensity, the system becomes more sensitive to additional changes. Conversely, in low-value regions,

contributions cluster around the negative zone with narrow amplitude, implying that reducing accumulation does not diminish risk at the same rate.

This asymmetry indicates that the relationship between accumulation and risk is non-linear. Risk surges only when the accumulation configuration reaches a sufficiently high level on the PC1 axis. This explains why PC1 is both the dominant axis in PCA and the foundation of large positive contributions in SHAP: it reflects a genuine structural mechanism.

The role of **duration of presence** further clarifies this. When scale and density are moderate, time has a limited impact. However, once the core factors have accumulated strongly, time becomes a consolidating factor, deepening integration and reducing adjustment capacity. In other words, time does not create dependence; it makes dependence more durable once the accumulation structure is established.

Meanwhile, **trade integration** (corresponding to PC3) acts as a conditioning factor. In sectors that have already reached high accumulation, deep integration tends to amplify positive contributions due to tighter linkage into global value chains.

4.3. Consistency between Sectoral Stratification and Micro-mechanisms

The compatibility between sectoral CRI distribution and multi-dimensional accumulation mechanisms demonstrates the internal consistency of the results. High-CRI sectors are those where factors with strong positive contributions to risk probability converge simultaneously. FDI dependence is not an inevitable consequence of attracting large capital, but a product of sustained structural accumulation at strategic positions in the value chain. This result also explains why non-linear models outperform linear ones; the dependence structure is formed by multivariate interactions and threshold effects that exceed traditional linear assumptions.

5. CONCLUSION

This study provides clear empirical evidence that FDI dependence risk at the sectoral level is not a direct consequence of large capital scale but the result of a structural accumulation process in sectors occupying central positions in the production system. By integrating the Composite Risk Index (CRI), PCA, and non-linear models, the study identifies risk stratification across sectors and elucidates the mechanism through which dependence is formed and consolidated over time.

Specifically, the study contributes in three main areas. **First**, structurally, risk fluctuations converge around an aggregate accumulation axis (PC1), reflecting the co-movement of scale, density, and duration. **Second**, dynamically, non-linear models confirm the existence of threshold effects: the probability of dependence surges only when factors cross a certain convergence level. **Third**, theoretically, the study refines the traditional interpretation of dependency theory by identifying specific mechanisms at the sectoral level rather than assuming an inevitable relationship between foreign capital and decreased autonomy.

Despite these contributions, the study has limitations. The analysis is primarily at the sectoral level and does not yet delve into ownership structures or corporate-level control networks. Additionally, the exclusion of dynamic institutional factors may limit the full explanation of long-term FDI dynamics. Future research could expand to spatial and firm-level analysis, as well as integrate variables reflecting institutional quality and GVC participation. Overall, the study affirms that FDI dependence is a structural phenomenon of long-term accumulation, requiring an analytical approach beyond single-scale measurement to identify configurations capable of altering the state of the industry and the economy.

REFERENCES

1. **Alfaro, L., Chanda, A., Kalemli-Ozcan, S., & Sayek, S. (2004).** FDI and economic growth: The role of local financial markets. *Journal of International Economics*, 64(1), 89–112.

2. **Foreign Investment Agency (2023).** Summary report on the status of foreign direct investment in Vietnam for the period 2007–2022. Hanoi: Ministry of Planning and Investment.
3. **Nguyen Hoang, M. P. (2025).** How foreign direct investment impacts domestic productivity: The case of Vietnam. *Duke Journal of Economics*, 12(1), 1–35.
4. **General Statistics Office (2019).** Statistical Yearbook of Vietnam 2018. Hanoi: Statistical Publishing House.
5. **World Bank (2022).** World Development Indicators. Washington, DC: World Bank.
6. **Zhang, L., Wang, Y., & Liu, K. (2022).** Integrating composite risk index and machine learning for risk classification in emerging economies. *Technological Forecasting and Social Change*, 178, 121–134.