Identification of the Causal Effect of Institutional Investor Ownership on Corporate Green Innovation Based on Double Machine Learning

Jin Cui and Xinyue Miao*

Monash University Malaysia, Subang Jaya, Malaysia

*Corresponding author: jincui2021@gmail.com

Abstract. As capital markets place greater weight on environmental governance, the role of institutional shareholders in shaping firms' green patenting has become a central question in corporate finance and environmental economics. This paper develops a causal identification design for high-dimensional pan-el data by combining fixed-effect residualization with Double Machine Learning. Institutional ownership is specified as the treatment, while green patent applications measure the outcome. The covariate set covers firm fi-nancial conditions, ownership and board characteristics, industry infor-mation, local environmental regulation, and macroeconomic indicators. To estimate the nuisance components in the treatment and outcome equations, the framework combines Lasso, random forest, gradient boosting, and neu-ral-network learners, and it applies cross-fitting together with a Neyman-orthogonal score to limit overfitting and shrinkage bias. Firm, year, and in-dustry-by-year effects are included before the machine-learning step so that persistent firm differences and common time shocks are filtered out. The empirical estimates indicate that higher institutional ownership is associated with a positive and statistically meaningful increase in firms' green innova-tion output.

Keywords: Double Machine Learning, Institutional Investor Ownership, Corporate Green Innovation, Causal Effect Identification, Green Patent Output

1 Introduction

Climate pressure, tighter resource limits, and pollution-control requirements have pushed low-carbon growth from a policy slogan to a practical development constraint. Enterprises are not only major participants in production and resource consumption, but also key actors in technological progress and industrial upgrading [1]. Their green innovation activities, including green patents, clean production technologies, energy-saving processes, and environmentally friendly products, play an important role in improving environmental performance and promoting sustainable economic growth [2].

The capital market is one channel through which these incentives may be transmitted. Institutional investors, as professional investors with strong information

A. Bhattacharjya et al. (eds.), *Proceedings of the 2026 4th International Conference on Management Innovation and Economy Development (MIED 2026)*, Advances in Economics, Business and Management Research 405,
https://doi.org/10.2991/978-94-6239-787-3_26

acquisition, analysis, and monitoring capabilities, have become important external governance forces in modern corporations [3]. Compared with individual investors, institutional investors usually have more resources, stronger professional expertise, and greater influence on corporate decision-making. Their shareholding may affect corporate green innovation by improving governance efficiency, reducing managerial short-termism, easing financing constraints, and encouraging firms to pursue long-term sustainable value [4].

The expected direction of this effect is not obvious. A governance-oriented institutional investor may press managers to curb opportunistic behavior and allocate more resources to green R&D, especially when environmental performance contributes to long-term firm value [5]. From this perspective, institutional investor ownership can promote corporate green innovation by strengthening external governance and guiding firms toward long-term strategies [6]. On the other hand, some institutional investors may focus more on short-term financial returns and stock market performance. Since green innovation often requires large investment, involves high uncertainty, and produces delayed returns, these investors may discourage firms from engaging in costly green innovation projects [7].

A positive association may indicate that institutional investors promote green innovation, but it may also reflect that firms with stronger innovation capacity and better environmental performance are more likely to attract institutional investors [8]. Similarly, firms with better governance quality, stronger profitability, or greater growth potential may simultaneously have higher institutional ownership and more green innovation output.

2 Related Work

Recent studies on corporate green innovation have increasingly shifted from conventional correlation-based analysis to causal identification supported by machine learning techniques. Luo and Liu [9] show that institutional ownership can improve green innovation performance, and their argument emphasizes the governance role of investors when formal institutions do not fully price environmental and innovation externalities. Mao et al. [10] investigate how data assets enhance corporate green innovation capabilities using a double machine learning model. Their study, published in Science & Technology Progress and Policy, treats data assets as a key explanatory factor and analyzes their causal impact on firms' green innovation capability. Shi et al. [11] consider data-asset disclosure divergence and green innovation with a double-debiased machine-learning approach. This work indicates that information quality and external monitoring are crucial factors in green innovation decisions.

Cao et al. [12] evaluate the effect of government green subsidies on corporate ESG performance and report that subsidies operate partly through digital technology innovation and conversion efficiency. Building on these studies, the present paper focuses on institutional ownership as the treatment variable and estimates its effect on green patent output in a high-dimensional panel setting.

3 Methodologies

3.1 Model Setting and High-Dimensional Nuisance Function Estimation

The analysis starts from a partially linear panel model that connects institutional ownership to firm-level green innovation. The outcome variable is green patent output, while the treatment variable is institutional investor ownership, denoted as Eq. (1).

$$Y_{it} = \theta_0 D_{it} + g_0(X_{it}) + \alpha_i + \lambda_t + \delta_{s(i)t} + \varepsilon_{it}, \tag{1}$$

where Y_{it} denotes corporate green innovation output of firm i in year t, and D_{it} represents institutional investor ownership. X_{it} is a high-dimensional control vector including firm financial characteristics, governance variables, industry attributes, regional environmental regulation, and macroeconomic factors. α_i, λ_t, and $\delta_{s(i)t}$ represent firm, year, and industry-year fixed effects. Because institutional investors do not randomly choose firms, the treatment variable must also be modeled. Eq. (2) explains how institutional ownership is determined by observable firm-level and external characteristics.

$$D_{it} = m_0(X_{it}) + \alpha_i^D + \lambda_t^D + \delta_{s(i)t}^D + v_{it}, \tag{2}$$

where $m_0(X_{it})$ is the conditional expectation function of institutional investor ownership given high-dimensional covariates. To eliminate unobservable firm heterogeneity and common time shocks, this study applies fixed-effect residualization before machine learning estimation. This transformation in Eq. (3) removes the variation explained by firm, year, and industry-year effects.

$$\tilde{A}_{it} = A_{it} - F_{it}(F'F)^{-1}F'A, \tag{3}$$

where A_{it} can represent Y_{it}, D_{it}, or any element of X_{it}. F_{it} is the fixed-effect design matrix, and $\tilde{A}_{it}$ is residualized variable after controlling for multi-dimensional fixed effects. After fixed-effect adjustment, the model is transformed into a residualized partially linear form. Eq. (4) allows the causal effect to be identified from within-firm and within-time variation rather than from persistent structural differences across firms.

$$\tilde{Y}_{it} = \theta_0 \tilde{D}_{it} + g_0(\tilde{X}_{it}) + \varepsilon_{it}, \tag{4}$$

where $\tilde{Y}_{it}$, $\tilde{D}_{it}$, and $\tilde{X}_{it}$ are the fixed-effect-adjusted outcome, treatment, and control variables. The unknown function $g_0(\tilde{X}_{it})$ captures the nonlinear influence of high-dimensional covariates on green innovation. Eq. (5) is used to isolate the component of institutional ownership that is not predictable from high-dimensional confounding factors.

$$\tilde{D}_{it} = m_0(\tilde{X}_{it}) + v_{it}, \tag{5}$$

where $m_0(\tilde{X}_{it})$ describes how residualized control variables explain residualized institutional investor ownership. The remaining component v_{it} provides the identifying variation for estimating the causal effect. To improve the estimation of nuisance functions, this study uses an ensemble learning strategy than a single algorithm by Eq. (6).

$$\hat{g}(\tilde{X}_{it}) = \sum_{j=1}^{J} \omega_j^g \, \hat{g}_j(\tilde{X}_{it}), \hat{m}(\tilde{X}_{it}) = \sum_{j=1}^{J} \omega_j^m \, \hat{m}_j(\tilde{X}_{it}), \tag{6}$$

where $j \in$ {Lasso,RF,GBM,NN}, and $\hat{g}_j(\cdot)$ and $\hat{m}_j(\cdot)$ are nuisance-function estimates from the j-th learner. The weights ω_j^g and ω_j^m determine the contribution of each machine learning model.

3.2 Cross-Fitting and Neyman Orthogonalization

To avoid overfitting, the full sample is divided into several mutually exclusive folds. For each fold, nuisance functions are trained on the remaining data and predicted only for the held-out observations by Eq. (7).

$$I = I_1 \cup I_2 \cup \cdots \cup I_K, I_k \cap I_l = \emptyset, k \neq l. \tag{7}$$

In Eq. (7), I is the full sample and I_k is the k-th subsample. For observations in fold I_k, the nuisance functions are estimated using the complementary sample I_k^c. This procedure separates the prediction step from the causal estimation step.

$$\hat{g}_{-k}(\tilde{X}_{it}), \hat{m}_{-k}(\tilde{X}_{it}), (i,t) \in I_k, \tag{8}$$

where the subscript $-k$ indicates that the nuisance models are trained outside fold I_k.

The next step constructs residualized outcome and treatment components using the cross-fitted nuisance estimates. These residuals in Eq. (9) remove the influence of high-dimensional controls from both green innovation and institutional ownership.

$$\hat{U}_{it} = \tilde{Y}_{it} - \hat{g}_{-k}(\tilde{X}_{it}), \hat{V}_{it} = \tilde{D}_{it} - \hat{m}_{-k}(\tilde{X}_{it}), \tag{9}$$

where $\hat{U}_{it}$ is the unexplained component, and $\hat{V}_{it}$ is the residual part of institutional ownership. The causal effect is identified by examining whether changes in $\hat{V}_{it}$ systematically explain changes in $\hat{U}_{it}$.

4 Experiments

4.1 Experimental Setup

The empirical sample is drawn from the CSMAR China Stock Market & Accounting Research Database and covers Chinese A-share listed companies during 2010-2022. Financial firms, ST and *ST companies, and observations with missing core variables are removed. Continuous variables are winsorized to limit the influence of extreme values. Green innovation is measured by the number of green patent applications in a firm-year and is transformed with $\ln(1 + GreenPatent)$.

4.2 Experimental Analysis

Figure 1 describes the time-series patterns of institutional ownership and green innovation for Chinese A-share firms from 2010 to 2022. The bars report average green patent applications and distinguish invention patents from utility-model patents; the line tracks average institutional ownership over the same period. The Figure 1 suggests that both institutional participation and green patenting increased over the sample window.

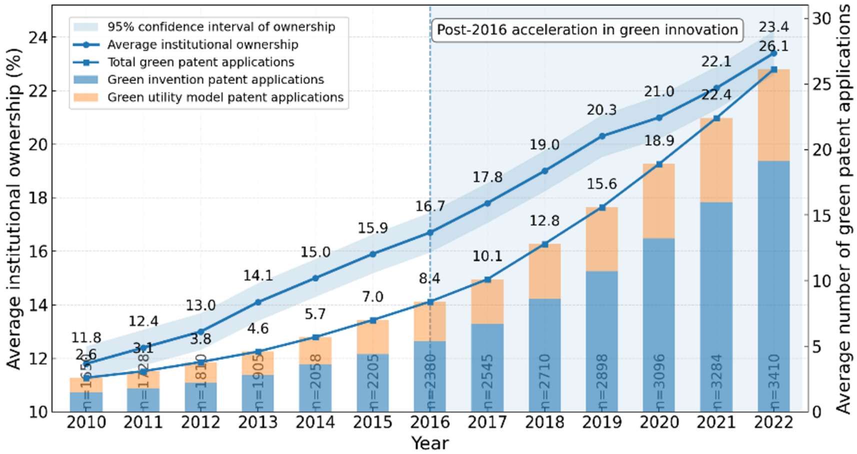

Fig. 1. Annual Trends in Institutional Ownership and Corporate Green Innovation.

Figure 2 summarizes the DML estimates across model specifications. Each coefficient is positive, and the reported confidence intervals remain above zero. This pattern indicates that institutional ownership has a statistically significant positive effect on corporate green innovation under the alternative learners and sample definitions considered here.

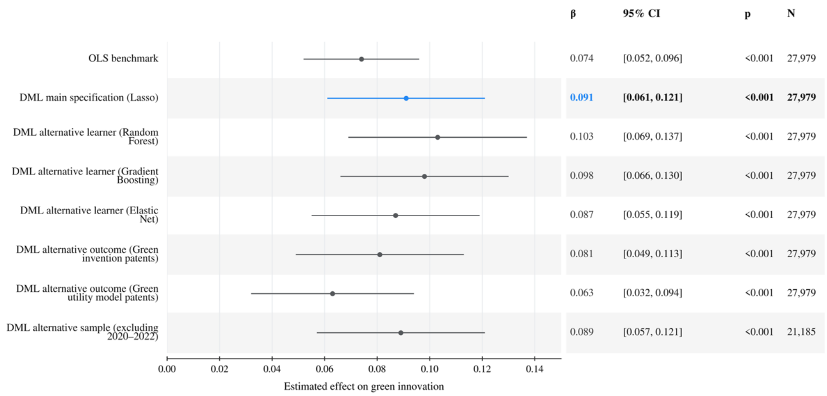

Fig. 2. Estimated Causal Effect of Institutional Ownership under DML Specifications.

Table 1 provides the numerical robustness checks corresponding to the graphical evidence. It reports point estimates, standard errors, and 95 percent confidence bounds for different first-stage learners, patent-output definitions, and sample restrictions. The estimates remain positive across all specifications, and the confidence intervals are narrow enough to support a stable positive conclusion.

Table 1. Robustness Comparison of Double Machine Learning Estimates.

Model	Estimate	Std. Error	CI Lower	CI Upper
Lasso	0.09	0.02	0.06	0.12
Random Forest	0.10	0.02	0.07	0.14
Gradient Boosting	0.10	0.02	0.07	0.13
Elastic Net	0.09	0.02	0.06	0.12
Lasso (Green invention)	0.08	0.02	0.05	0.11
Lasso (Green utility)	0.06	0.02	0.03	0.09
Lasso (2020-2022)	0.09	0.02	0.06	0.12
Random Forest (2020-2022)	0.10	0.02	0.07	0.14
Gradient Boosting (SOEs)	0.08	0.02	0.05	0.12

5 Conclusion

In conclusion, this study constructs an improved framework to identify the causal effect of institutional investor ownership on corporate green innovation. By treating institutional ownership as the core treatment variable and green patent output as the outcome variable, the proposed model integrates high-dimensional firm, governance, industry, regional environmental, and macroeconomic controls. Future research may further extend this framework by incorporating dynamic treatment effects, textual environmental disclosure data.

References

1. Bai, T., Xu, D., Yang, Q., Dudás Piroska, V., Dénes Dávid, L., & Zhu, K. (2023). Paths to low-carbon development in China: the role of government environmental target constraints. *Oeconomia Copernicana*, *14*(4), 1139-1173.
2. Zhang, S., & Wang, M. (2025). Green and Low-Carbon Development. *Environmental Law Codification in China: Paradigm Change in Environmental Legislation*, 269.
3. Huang, C., Wang, C., Rani, T., & Rehman, S. A. U. (2024). Digitalization's role in shaping climate change, renewable energy, and technological innovation for achieving sustainable development in top Asian countries. *Energy & Environment*, 0958305X241258799.
4. Onyshchenko, S., Hlushko, A., Maslii, O., & Chumak, O. (2024). Digital transformation of the national economy in the context of information environment development in Ukraine. *Transformations of national economies under conditions of instability*, 169-197.
5. Mesioye, O., & Bakare, I. A. (2024). Evaluating financial reporting quality: Metrics, challenges, and impact on decision-making. *Int J Res Public Rev*, *5*(10), 1144-1156.

6. Zhao, J., Qu, J., Wei, J., Yin, H., & Xi, X. (2023). The effects of institutional investors on firms' green innovation. *Journal of Product Innovation Management*, *40*(2), 195-230.
7. Yang, Z., Su, D., Xu, S., & Han, X. (2024). Institutional investors and corporate green innovation: Evidence from China. *Pacific Economic Review*, *29*(2), 230-266.
8. Khan, F. U., Zhang, J., Saeed, I., & Ullah, S. (2024). Do institutional contingencies matter for green investment?—An institution based view of Chinese listed companies. *Heliyon*, *10*(1).
9. Luo, S., & Liu, C. (2025). Institutional Investors and Green Innovation Under Double Externalities: A Machine Learning Optimization Perspective. *Mathematics*, *13*(22), 3718.
10. Chunmei, M., Yibo, Y., Junjun, N., & Qing, W. (2025). How Data Assets Enhance Corporate Green Innovation Capabilities: Causal Inference Based on a Double Machine Learning Model. *Science & Technology Progress and Policy*, *42*(20), 1-10.
11. Shi, H., Xia, Y., Li, Y., Hua, Y., & Fu, Y. (2025). Data asset information disclosure divergence and green innovation: causal evidence from double-debiased machine learning. *Finance Research Letters*, *84*, 107757.
12. Cao, Y., Hizam-Hanafiah, M., Fahmi Ghazali, M., Ab Razak, R., & Zheng, Y. (2025). Estimating the Impact of Government Green Subsidies on Corporate ESG Performance: Double Machine Learning for Causal Inference. *Sustainability*, *18*(1), 281.