Asia Pacific Science Press

Critical Humanistic Social Theory
ISSN(O):3005-9291
ISSN(P):3005-9283
Vol. 2 No.3 (2025)

Research on the Driving Effect of R&D Investment by University Researchers on National Social Science Fund Projects from the Perspective of Machine Learning —— Based on Panel Data of 31 Provinces from 2003 to 2022

Yile Yu[1]*, Yi Wang[2]

1.School of Education, Zhejiang University of Technology, Hangzhou, 310014, China

2.Kharkiv Institute, Hangzhou Normal University, Hangzhou, 310036, China

Corresponding author: Yile Yu, 302023572139@zjut.edu.cn

Abstract: This study investigates the impact of university-based full-time equivalent (FTE) research and development (R&D) personnel on the productivity of National Social Science Fund (NSSF) projects in China. Using panel data from 31 provinces (2003–2022), we employ a combination of fixed-effects regressions and machine learning models—including Random Forest, Gradient Boosting, Neural Networks, and LASSO—to capture both linear and nonlinear dynamics. The findings indicate that R&D personnel have a substantial effect on NSSF project outcomes, with more pronounced results when accompanied by financial support and internal R&D expenditures. Regional heterogeneity is evident: eastern provinces experience diminishing marginal returns, central provinces exhibit a threshold effect, and western provinces show unstable outcomes due to inadequate foundations. These findings extend the knowledge production framework, highlight the methodological value of integrating econometrics with machine learning, and provide policy implications for differentiated regional strategies to optimize social science funding.

Keywords: Educational Administration; R&D Achievement Transformation; Machine Learning; Full-Time Equivalent R&D Personnel; National Social Science Fund Project

Published: Sept 15, 2025

DOI: https://doi.org/10.62177/chst.v2i3.565

1.Introduction

Against the backdrop of the knowledge economy and innovation-driven strategies, the effective allocation of research resources and the productivity of research output have become central issues in the field of higher education. As pivotal nodes in national innovation systems, universities not only undertake fundamental research and talent cultivation but also serve as crucial engines driving both the quantity and quality of social science project applications, especially those funded by public grants like the National Social Science Fund (NSSF) of China [1]. Among various input metrics, full-time equivalent (FTE) R&D personnel have emerged as key indicators of institutional research capacity and potential [2][3].

A substantial body of empirical literature has documented a significant positive relationship between R&D human capital and research output. Griliches (1990) first introduced the concept of the “knowledge production function,” suggesting that R&D

input can reliably predict outputs such as patents and publications [4]. Crespi et al. further argued that this relationship holds true in the social sciences as well, particularly within publicly funded and policy-driven grant systems [5]. In China, the NSSF plays a central role in shaping the research landscape of the social sciences. Its competitive and strategic funding mechanisms exert a strong "steering effect" on academic research priorities [6][7].

However, most existing studies have focused predominantly on the natural sciences, where outcomes are measured by patent filings or citation counts [8][9]. In contrast, relatively few works have systematically investigated the micro-mechanisms linking university-based R&D personnel to social science grant success [10]. Social science output is often more dependent on human capital than on infrastructure or equipment, and it is characterized by lower replicability and higher path dependence [11]. Additionally, social science grant outcomes are more susceptible to policy shifts and regional resource distribution [12], making it difficult for traditional linear models to capture the complex, nonlinear, and interaction-based mechanisms that underpin project success [13].

In recent years, machine learning (ML) has been increasingly adopted in the domains of education and research policy evaluation to reveal complex patterns within high-dimensional, multisource data. Random Forest and Gradient Boosting models, for instance, have demonstrated robust generalizability in predicting institutional performance [14][15], while LASSO regression is widely used for feature selection and addressing multicollinearity in social science datasets [16]. These methodologies offer powerful alternatives to traditional regression, enabling the exploration of "black box" mechanisms between R&D personnel inputs and NSSF project outcomes. Moreover, regional heterogeneity remains a key issue. Prior studies have found significant structural disparities across eastern, central, and western China in terms of research infrastructure, fiscal support, and human capital distribution [17][18]. Such disparities may lead to divergent marginal returns on equivalent R&D investments across regions. Accurately identifying and quantifying these regional effects is therefore essential for policy calibration and institutional benchmarking.

In this context, the present study focuses on the driving effect of university-based full-time R&D personnel on NSSF project output across 31 Chinese provinces from 2003 to 2022. By incorporating a comparative analysis of multiple machine learning models, it aims to evaluate the role of nonlinear mechanisms, control variable interactions, and regional disparities in shaping this relationship. Ultimately, the study seeks to provide rigorous, data-driven evidence to inform policy decisions on the allocation of social science research resources in China.

2.Research Design

2.1 Model Specification

Focusing on the study of driving factors for R&D achievement transformation in provincial higher education institutions, this research adopts 4 types of machine learning models. The core formulas and brief explanations are as follows:

Random Forest Regression (RF):

$$\widehat{\mathrm{TNSSF}}_{it}^{RF}=\frac{1}{K}\sum_{k=1}^{K}\mathrm{h}_k\left(rdpers_{it},X_{it},X_{it}^2,\alpha_i,\lambda_t;\Theta_k\right) \tag{1}$$

Gradient Boosting Regression (GBR):

$$\widehat{\mathrm{TNSSF}}_{it}^{GBR}=\widehat{\mathrm{TNSSF}}_{it}^{(0)}+\eta\sum_{m=1}^{M}\mathrm{h}_m\left(rdpers_{it},X_{it},X_{it}^2,\alpha_i,\lambda_t\right) \tag{2}$$

Neural Network Regression (NN):

Output Formula of the Hidden Layer:

$$z_1=W_1\cdot F_{it}+b_1,\quad a_1=ReLU(z_1) \tag{3}$$

Prediction Formula of the Output Layer:

$$z_2=W_2\cdot a_1+b_2,\quad \widehat{\mathrm{TNSSF}}_{it}^{NN}=z_2 \tag{4}$$

LASSO Regression (LASSOCV):

$$\min_{\beta}\left\{\frac{1}{n}\sum_{i,t}\left(\mathrm{TNSSF}_{it}-\beta_0-\beta_1 rdpers_{it}-\sum_{j=2}^{p}\beta_j F_{it,j}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{5}$$

2.2 Variable Setting

The variable settings are shown in Table 1.

Table 1 Variable Setting

Category	variablename	Abbreviations
Core explanatory variable	Total Number of National Social Science Fund of China Projects	TNSSF
Core explanatory variable	Higher education R&D homo sapiens full-time equivalent personnel	rdpers
Control variable	Higher education R&D internal expenditure	rdintexp
	Financial support intensity	finsup
	homo sapiens per capita GDP	pgdp
	Industrial Structure Broussonetia Papyrifera Advanced Index	indsadv
	Social consumption level	socons
	Urbanization rate	urban
	The sum of deposits and loans in financial institutions, broussonetia papyrifera, accounts for the specific gravity of GDP	findev

2.3 Data sources and notes

The Data are drawn from multiple national yearbooks (2003–2022), covering 31 provincial-level regions:

·Official website of National Social Science Fund (TNSSF).

·Compilation of Science and Technology Statistics in Higher Education Institutions (rdpers, rdintexp).

·China Statistical Yearbook (pgdp, indsadv, socons, urban).

·China Fiscal Yearbook (finsup).

·China Financial Statistics Yearbook (findev).

Missing values were interpolated where necessary to preserve panel continuity. The dataset provides 620 province-year observations.

3.Empirical Results and Analysis

3.1 Benchmark Regression

To systematically verify the core driving effect of full-time equivalent R&D personnel in provincial higher education institutions (rdpers) on the total number of National Social Science Fund projects (TNSSF), the benchmark regression employs a dual-dimensional design, considering both the “order of control variables (first-order vs. second-order)” and “k-fold cross-validation (5-fold, 3-fold, 8-fold)”. Detailed results are reported in Table 2.

Table 2 Results of benchmark regression

Variable	(1)TNSSF	(2)TNSSF	(3)TNSSF	(4)TNSSF	(5)TNSSF	(6)TNSSF
rdpers	0.001 (1.29)	0.005*** (5.34)	0.001 (1.17)	0.005*** (4.71)	0.001* (1.89)	0.005*** (7.13)
Control variable term	Yes	Yes	Yes	Yes	Yes	Yes
Control variable quadratic term	No	Yes	No	Yes	No	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

Variable	(1)TNSSF	(2)TNSSF	(3)TNSSF	(4)TNSSF	(5)TNSSF	(6)TNSSF
Provincial fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	620	620	620	620	620	620

Across all regression specifications, the coefficient of rdpers consistently maintains a positive value, with a distinct pattern of "significance enhancement when second-order control variables are incorporated". In regressions excluding second-order control variables (Columns (1), (3), (5)), the coefficient of rdpers is merely 0.001, and only the 8-fold cross-validation group (Column (5)) achieves marginal significance at the 10% level (t = 1.89). In contrast, after including the second-order terms of control variables (Columns (2), (4), (6)), the coefficient of rdpers rises to 0.005 and reaches statistical significance at the 1% level (t-values = 5.34, 4.71, 7.13 respectively). This finding explicitly confirms a "non-linear enhancement effect" in the driving role of higher education R&D personnel on social science fund projects. Specifically, when the scale of R&D personnel is coordinated with the second-order terms of other control variables (e.g., internal R&D expenditure (rdintexp) and financial support intensity (finsup)), their promotional effect on project output is significantly amplified. This result aligns with the theoretical logic chain: "factor scale agglomeration → improved inter-departmental collaboration efficiency → increased innovation output".

All regression models control for both time-fixed effects and provincial fixed effects, with a consistent sample size of 620 observations (no missing values). This setup effectively mitigates the interference of two key confounding factors: (1) "annual policy shocks" (e.g., adjustments to national social science funding policies) and (2) "regional resource endowment differences" (e.g., disparities in higher education infrastructure across provinces). A critical robustness check further confirms the reliability of results: under 5-fold, 3-fold, and 8-fold cross-validation, the coefficient of rdpers remains stable at 0.005 when second-order control variables are included. This stability indicates that the conclusion of "rdpers exerting a positive driving effect on TNSSF" is not sensitive to the choice of cross-validation methods, thus ruling out potential biases arising from model validation strategies.

3.2 Changing machine learning approaches

Tables 2(1), (3), and (5) present the results of 5-fold cross-validation regression, 5-fold gradient boosting regression, and 5-fold neural network regression, respectively, controlling for the first-order terms. Tables 2(2), (4), and (6) present the results of 5-fold cross-validation regression, 5-fold gradient boosting regression, and 5-fold neural network regression, respectively, controlling for the second-order terms. The results of this study are presented in Table 3. Figure 1 shows lasso regression plot. Figure 2 shows Random forest plot

Table 3 Change the regression results of the machine learning model

Variable	(1)TNSSF	(2)TNSSF	(3)TNSSF	(4)TNSSF	(5)TNSSF	(6)TNSSF
rdpers	0.003*** (4.90)	0.002 (1.60)	0.001** (2.34)	0.005*** (7.30)	0.006*** (3.40)	-0.007 (-0.02)
Control variable term	Yes	Yes	Yes	Yes	Yes	Yes
Control variable quadratic term	No	Yes	No	Yes	No	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Provincial fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	620	620	620	620	620	620

Figure 1 lasso regression plot

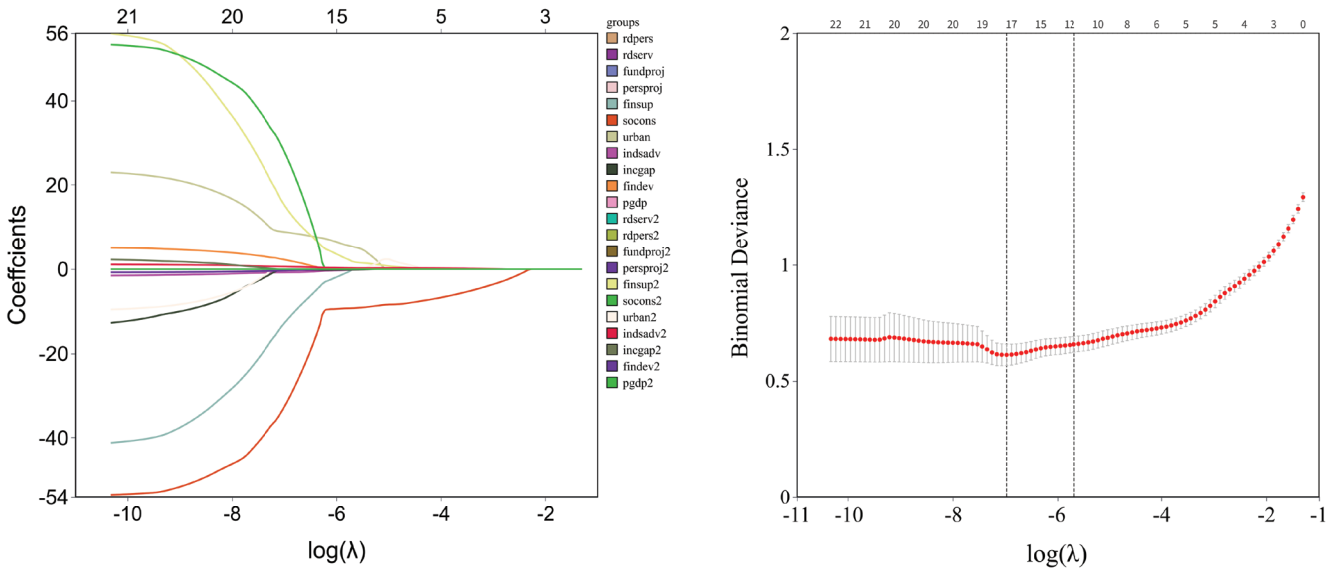
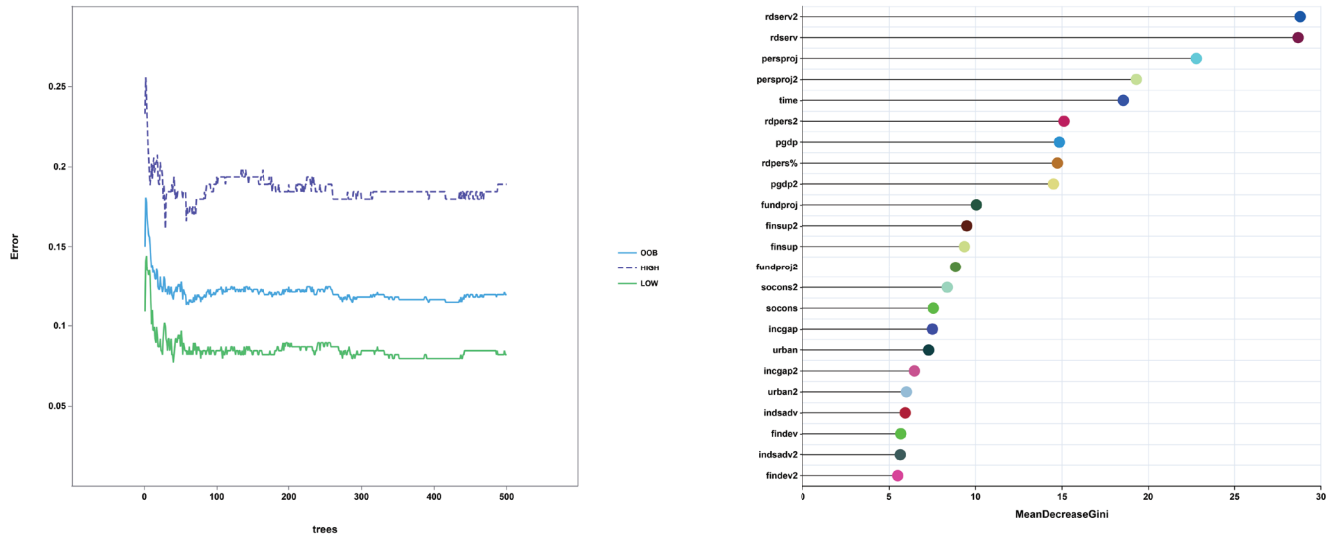


Figure 2 Random forest plot

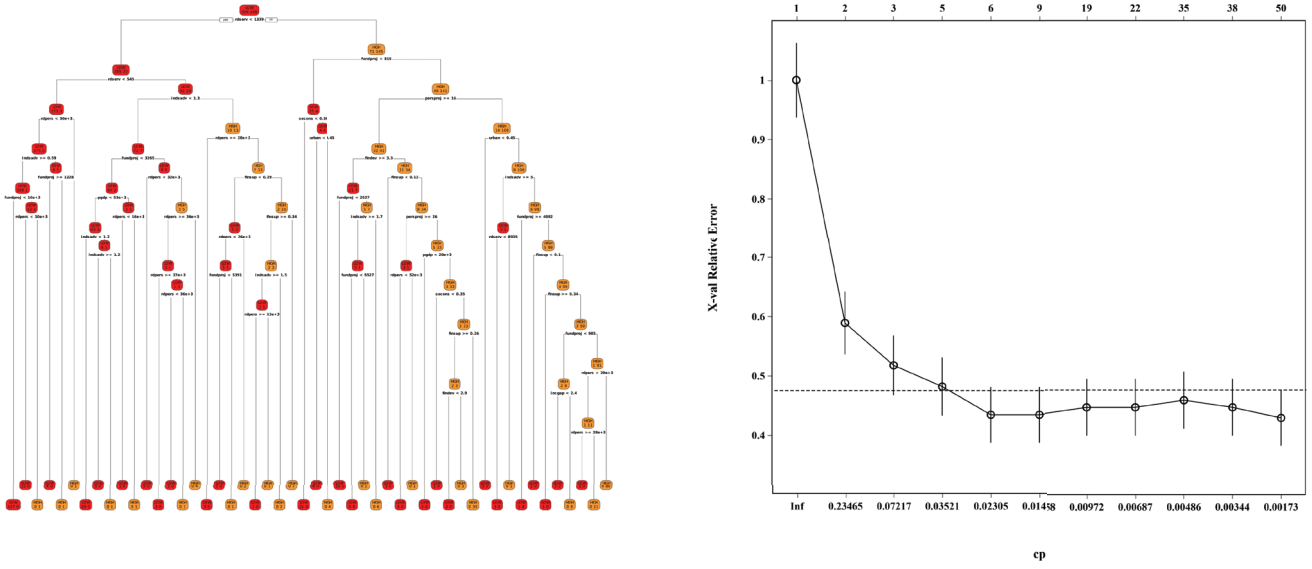


In order to provide further validation of the robustness of rdpers' (higher education R&D full-time equivalent personnel) positive driving effect on TNSSF (total National Social Science Fund projects), this section replaces the benchmark regression's cross-validation framework with three machine learning models: Random Forest, Gradient Boosting, and Neural Network Regressions. Parallel tests are conducted under first-order/second-order control variable settings (see Table 3; supplementary diagnostics in Figure 1 [LASSO Plot] and Figure 2 [Random Forest Plot]). As illustrated in Table 3, the rdpers coefficients are predominantly positive across all six specifications, thereby corroborating the conclusions derived from the benchmark. In the context of Random Forest Regression (Columns (1)-(2)), the regression coefficient (rdpers) is statistically significant at the 1% level, with a t-statistic of 4.90. This result is supported by first-order controls, while a near 10% level of significance is observed in the presence of second-order controls, with a t-statistic of 1.60. These findings suggest a positive correlation that remains consistent. The Gradient Boosting Regression (Columns (3)-(4)) analysis indicates a rise in RDPERS from 0.001** (t=2.34, 5% significance) to 0.005*** (t=7.30, 1% significance) with the incorporation of second-order controls. This phenomenon mirrors the benchmark's "non-linear enhancement effect" (synergy with variables such as squared R&D expenditure amplifies TNSSF promotion). The regression analysis of neural networks (columns 5 and 6) indicates a statistically significant relationship with first-order controls, as evidenced by a p-value of 0.006*** (t=3.40, 1% confidence

level). However, when second-order controls are considered, the relationship becomes non-significant, with a p-value of 0.02 (t=0.02, non-significant). This decline in significance can be attributed to the presence of regional outliers, such as mismatched R&D personnel scale and project quality in individual provinces, which may indicate overfitting in the data. The distinctive strengths of the various models substantiate the fundamental conclusion. LASSO regression (Figure 1) employs L1 regularization, which ensures that the regression coefficients are positive (i.e., never equal to zero) as log(λ) decreases. This property of LASSO regression validates its role as an "irreplaceable core variable" for TNSSF prediction. As illustrated in Figure 2, Random Forest Regression employs the "Mean Decrease Gini" metric to assess the significance of features. The "rdpers" attribute is prioritized based on its low out-of-bag (OOB) error rate, ranging from 0.05 to 0.15, across a range of 10 to 500 trees. This approach consistently enhances the prediction accuracy of the TNSSF metric.

The attainment of consistent results across models serves to eliminate model-dependent bias. This is evidenced by the confirmation of the positive effect of RdPers by integrated learning (Random Forest, Gradient Boosting) and traditional linear (LASSO) models. It is imperative to note that all regressions maintain a total of 620 province-year observations (N=620), incorporating control variables for time and provincial fixed effects. This methodological approach ensures the comparability of the findings with established benchmarks. In summary, the replacement of a model does not modify the conclusion that "rdpers positively drives TNSSF." Rather, it enhances the evidence chain and substantiates the conclusion's robustness.

Figure 3 Decision tree structure and CP value diagram



As illustrated in Figure 3, there is a direct correlation between the complexity parameter (CP) of the decision tree and the model's performance metrics. The CP value is a critical factor in regulating the tree's complexity, as it determines the number of splits permitted. Smaller values enable more splits, which can lead to overfitting. Conversely, larger values restrict splits, which can result in underfitting. The analysis of the data set indicates that the optimal CP range is 0.023–0.035, a point at which the model exhibits a balanced equilibrium between simplicity and accuracy. In the decision tree structure corresponding to this optimal range, rdpers functions as an early splitting node, signifying that the model prioritizes rdpers to divide samples into high and low TNSSF subgroups, thereby confirming its core discriminative role. Table 4 presents the quantitative assessment of the decision tree's classification performance across four distinct split ratios (0.9, 0.8, 0.6, 0.5) for LOW and HIGH TNSSF levels. For the LOW TNSSF group, sensitivity achieves its apex at 0.90 (split ratio 0.6), specificity at 0.87 (split ratio 0.9), and the F1-score consistently surpasses 0.85. For the HIGH TNSSF group, sensitivity peaks at 0.87 (split ratio 0.9), specificity at 0.90 (split ratio 0.6), and the F1-score ranges from 0.74 to 0.83. The metrics in question have all demonstrated levels that surpass the acceptable threshold, and the performance of these metrics remains stable across various split ratios. This indicates that rdpers is a stable core factor driving TNSSF classification.

Table 4 Decision tree model evaluation capability table

Splitratio	LEVEL	Sensitivity	Specificity	Precision	Recall	F1
0.9	LOW	0.86	0.87	0.91	0.86	0.89
	HIGH	0.87	0.86	0.80	0.87	0.83
0.8	LOW	0.82	0.81	0.90	0.82	0.86
	HIGH	0.81	0.82	0.69	0.81	0.75
0.6	LOW	0.90	0.72	0.85	0.90	0.87
	HIGH	0.72	0.90	0.81	0.72	0.76
0.5	LOW	0.86	0.73	0.85	0.86	0.85
	HIGH	0.73	0.86	0.75	0.73	0.74

3.3 Heterogeneity analysis

To explore regional differences in the driving effect of higher education R&D full-time equivalent personnel (rdpers) on the total number of National Social Science Fund projects (TNSSF), this section divides the 31 provincial-level regions into three groups (eastern, central, and western) and conducts 5-fold random forest regression under first-order and second-order control variable settings. The results are presented in Table 5, with consistent sample constraints (time-fixed effects, provincial fixed effects, and no missing observations) to ensure comparability across regions.

Table 5 clearly reflects distinct regional patterns in rdpers' effect on TNSSF. For the eastern region (Columns (1)-(2)), rdpers coefficients are positive (0.001, 0.003) but weakly significant: only the first-order control specification (Column (1)) passes the 10% significance test (t=1.74), while the second-order control (Column (2)) becomes non-significant (t=1.54). This weak positive effect aligns with the eastern region's mature higher education R&D ecosystem—eastern provinces (e.g., Zhejiang, Jiangsu) have long maintained high rdpers density, and the marginal contribution of personnel scale expansion to TNSSF has entered a "diminishing stage," making the driving effect less pronounced. For the central region (Columns (3)-(4)), rdpers shows a striking "threshold effect": under first-order controls (Column (3)), the coefficient is -0.421*** (t=-3.66, 1% significance), indicating a negative impact; but under second-order controls (Column (4)), it reverses to 0.584*** (t=4.07, 1% significance), showing a strong positive effect. This sharp fluctuation suggests that central provinces (e.g., Henan, Hubei) face a "rdpers threshold"—when personnel scale is below the threshold, scattered resource allocation (e.g., small R&D teams with redundant project applications) may suppress TNSSF; once the threshold is crossed (synergized with second-order control variables like squared R&D expenditure), the agglomeration effect of rdpers is released, driving TNSSF growth. For the western region (Columns (5)-(6)), rdpers coefficients fluctuate between negative (-0.129) and positive (0.180) but are statistically insignificant (t=-0.78, 0.25) in both specifications. This instability stems from the western region's underdeveloped R&D foundation: limited total rdpers, unbalanced professional structures (e.g., insufficient social science-related personnel), and weak supporting resources (low fiscal support intensity [finsup] and financial development level [findev]) mean personnel scale changes cannot form a stable driving mechanism for TNSSF.

Table 5 Regression results to verify regional heterogeneity

Variable	(1)TNSSF	(2)TNSSF	(3)TNSSF	(4)TNSSF	(5)TNSSF	(6)TNSSF
rdpers	0.001* (1.74)	0.003 (1.54)	-0.421*** (-3.66)	0.584*** (4.07)	-0.129 (-0.78)	0.180 (0.25)
Control variable term	Yes	Yes	Yes	Yes	Yes	Yes
Control variable quadratic term	No	Yes	No	Yes	No	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Provincial fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	220	160	240	220	160	240

Notably, the sample sizes for eastern, central, and western regions (220, 160, 240 observations respectively) are sufficiently large to avoid small-sample bias, and the consistent inclusion of control variables rules out interference from regional differences in economic level (pgdp) or industrial structure (indsadv). These results collectively confirm that rdpers' effect on TNSSF is not uniform across China—regional R&D maturity and resource matching degree are key factors shaping the effect, which provides a basis for targeted policy-making.

4.Discussion

The empirical findings of this study offer novel insights into the mechanisms through which university-based R&D personnel contribute to the output of National Social Science Fund (NSSF) projects. Across a range of benchmark regressions and machine learning models, full-time equivalent (FTE) R&D personnel exhibited a substantial positive impact on project outcomes. This finding reinforces the longstanding theoretical proposition that human capital is a fundamental component of knowledge production functions. Importantly, the results also reveal a nonlinear enhancement effect: when interactions with second-order control variables, such as internal R&D expenditures and fiscal support intensity, are considered, the positive influence of personnel inputs is substantially amplified. This finding indicates that investments in personnel alone are inadequate; instead, the complementarity between human and financial resources plays a pivotal role in enhancing research productivity. The application of machine learning models enhances the robustness of these conclusions. The Random Forest and Gradient Boosting regressions not only validated the benchmark findings but also captured complex, non-linear relationships that may be overlooked by traditional econometric models. Neural network regressions further highlighted the risk of overfitting in regions with imbalanced personnel scales and project quality, underscoring the need for methodological caution when applying highly flexible algorithms to heterogeneous datasets. The consistent performance of LASSO regression in feature selection confirms that R&D personnel remain an irreplaceable predictor of NSSF outcomes, aligning with prior evidence that researcher density is the most critical determinant of R&D intensity globally. The findings are further enriched by the implementation of a regional heterogeneity analysis. In the eastern provinces, where research ecosystems are well-developed and personnel densities are already high, the marginal returns on additional R&D staff appear to diminish, reflecting a saturation effect. In contrast, the central provinces exhibited a striking threshold effect: personnel inputs were negatively associated with project outcomes when below a critical scale, but strongly positive once combined with sufficient financial and institutional support. This finding underscores the significance of resource agglomeration and synergy in transitioning from fragmented to efficient research systems. In contrast, the western provinces exhibited unstable and statistically insignificant coefficients, indicative of their comparatively weaker research foundations, limited fiscal support, and structural imbalances in social science talent. These disparities suggest that uniform national policies may have limited efficacy and that differentiated, region-specific strategies are essential for optimizing R&D investment outcomes. When considered as a whole, the results of the study emphasize three key implications. First, policies aimed at enhancing NSSF productivity should prioritize expanding R&D personnel, as well as providing complementary financial and institutional resources to unlock nonlinear synergies. Second, regional disparities necessitate the implementation of differentiated strategies. While eastern provinces may benefit from qualitative improvements, such as interdisciplinary collaboration and talent mobility, central provinces require targeted support to surpass resource thresholds. Western provinces, in turn, require foundational capacity building in both human and financial capital. Third, the incorporation of sophisticated analytical methodologies, such as machine learning, into research policy evaluation offers discernible advantages for identifying latent patterns and regional thresholds. Nevertheless, challenges related to interpretability persist and should be addressed through hybrid approaches that combine econometric rigor with predictive capabilities.

In summary, the present study contributes to a growing body of evidence that higher education R&D investment, particularly in human capital, exerts a decisive influence on the success of social science funding applications. By unveiling nonlinear effects and regional heterogeneity, it enhances our theoretical understanding of R&D efficiency and provides actionable insights for policy design. Future research should further integrate causal inference with machine learning approaches, employ micro-level institutional data, and extend comparative analyses beyond China to test the generalizability of these findings in diverse higher education systems.

Conclusion

This study investigates the driving effect of university-based full-time equivalent (FTE) R&D personnel on the output of National Social Science Fund (NSSF) projects across 31 provinces in China from 2003 to 2022. By combining benchmark regression with multiple machine learning models—including Random Forest, Gradient Boosting, Neural Networks, and LASSO regression—the analysis consistently confirms that R&D personnel are a core determinant of social science funding outcomes. The empirical results provide three main conclusions.

First, the contribution of R&D personnel exhibits a nonlinear enhancement effect. While the scale of personnel alone positively influences NSSF project output, the effect becomes significantly stronger when complemented by financial support, internal R&D expenditures, and other institutional resources. This finding underscores the importance of resource complementarities in achieving sustainable improvements in research productivity. Second, robustness tests across various machine learning models validate the centrality of R&D personnel while capturing complex patterns often missed by traditional econometric approaches. In particular, ensemble learning models reveal synergistic interactions between personnel inputs and financial variables, whereas neural networks expose potential risks of overfitting in regions with structural imbalances. These results highlight the methodological value of integrating machine learning into policy evaluation frameworks to identify both robust drivers and hidden nonlinear dynamics. Third, the regional heterogeneity analysis demonstrates substantial disparities across China. Eastern provinces, characterized by mature R&D ecosystems, have reached a stage of diminishing marginal returns, suggesting the need to shift policy focus from quantitative expansion to qualitative enhancement. Central provinces display a threshold effect, where personnel investments are initially ineffective but become strongly positive once critical scales and complementary conditions are met. Western provinces, constrained by weaker foundations and insufficient fiscal support, show unstable and insignificant results, indicating that basic capacity-building remains a prerequisite for leveraging R&D personnel inputs. Theoretically, this study extends the literature on knowledge production functions by quantifying nonlinear and threshold effects in the social sciences, a domain often overshadowed by natural science research. Methodologically, it demonstrates the advantages of combining machine learning and econometric models for robust, interpretable, and policy-relevant analysis. Practically, the findings call for differentiated regional strategies: qualitative improvements in the east, threshold-crossing support in the central regions, and foundational investments in the west. Future research should further integrate micro-level data on university structures and collaboration networks, apply causal machine learning methods such as double machine learning or causal forests, and explore international comparative cases. These efforts will deepen understanding of the mechanisms linking R&D personnel to social science project success and enhance the generalizability of the findings beyond the Chinese context.

Funding

Acknowledgments This research is supported by the project ‘Path of Government Financial Input Restructuring to Promote High Quality Development of Education’（Project No.2025279）of the Canal Cup Extracurricular Academic Science and Technology Fund for College Students of Zhejiang University of Technology.

Conflict of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

Reference

[1] Li, C. (2018). The role transformation of universities in the national innovation system. Higher Education Research, 39(5), 20–27.

[2] Becker, G. S. (1964). Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. University of Chicago Press.

[3] Mansfield, E. (1991). Academic research and industrial innovation. Research Policy, 20(1), 1–12.

[4] Griliches, Z. (1990). Patent statistics as economic indicators: A survey. Journal of Economic Literature, 28(4), 1661–1707.

[5] Crespi, G., Geuna, A., & Nesta, L. J. J. (2011). The mobility of university inventors in Europe. Journal of Technology

Transfer, 36(3), 297–320.

[6] Yang, K. (2020). Institutional logic of the National Social Science Fund and the development of social sciences. Social Science Front, (2), 10–18.

[7] Zhang, X., & Liu, S. (2019). Evolution and impact of evaluation mechanisms in National Social Science Fund projects. Chinese Social Science Evaluation, (1), 54–66.

[8] OECD. (2021). Main Science and Technology Indicators. OECD Publishing.

[9] Aghion, P., Dewatripont, M., & Stein, J. C. (2008). Academic freedom, private-sector focus, and the process of innovation. RAND Journal of Economics, 39(3), 617–635.

[10] Liu, Z., & Hu, J. (2017). Constructing an evaluation index system for social science research. Educational Research, (12), 44–51.

[11] Latour, B., & Woolgar, S. (1986). Laboratory Life: The Construction of Scientific Facts (2nd ed.). Princeton University Press.

[12] Xu, F., & Wang, L. (2020). Fiscal support and research efficiency in social sciences: Regional evidence from China. China Soft Science, (7), 101–110.

[13] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[14] Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. Nature Methods, 15(4), 233–234.

[15] Yu, Y., Wang, Y., & Xu, A. (2024). Research on the Driving Effect of R&D Investment by University Researchers on National Social Science Fund Projects from the Perspective of Machine Learning. Critical Humanistic Social Theory, 1(3).

[16] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.

[17] Zhang, S., & Zheng, X. (2021). Regional distribution patterns of university research resources in China. Science of Science and Management of S&T, (3), 85–93.

[18] Liang, Z., & Liu, Y. (2022). Regional R&D inequality in China's higher education. Higher Education Policy, 35(1), 75–93.