省域赋能知识产权高质量转化：差异识别、机制分析与预测研究

摘要

本文以省域研发投入向高质量知识产权产出的转化过程为研究对象，在区域差异识别基础上讨论其形成机制，并将短期预测作为对前述结论稳定性的补充检验。基于国家统计局、国家知识产权局和科技部等公开资料，构建覆盖 31 个省级地区、2020—2024 年的平衡面板数据。首先，从授权效率、发明专利存量结构、企业主体地位和单位投入产出效率四个维度测算高质量知识产权产出指数，并以熵权法构建研发投入指数和转化落差指标。其次，利用四象限识别和 KMeans 聚类刻画省域创新转化类型，在此基础上运用固定效应模型分析投入规模、投入强度与企业主体变量的作用方向。最后，为检验前述差异结构在短期内是否具有延续性，引入 ElasticNet 等模型对下一期产出进行滚动验证和保留集比较。结果表明，2020—2024 年全国省域平均 R&D 强度由 1.9384% 升至 2.1774%，高质量知识产权产出指数均值由 0.3088 升至 0.4974，企业有效发明专利占比均值由 61.46% 升至 66.25%。2024 年西藏、吉林、黑龙江、广西等省份表现出“低投入高转化”特征，广东、江苏则存在“高投入高产出但超额转化优势不足”的结构性现象。双向固定效应模型的调整后 R^2 达到 0.9523，年份虚拟变量持续显著为正。调优后的 ElasticNet 模型在 2024 年保留集上取得 $R^2 = 0.8861$、$MAE = 0.0319$，综合表现优于随机森林、XGBoost 和加权集成模型。研究表明，将区域差异识别、机制解释与短期预测置于同一分析框架，有助于更完整地把握省域创新转化的结构特征。

目录

表格与插图清单

表格清单

插图清单

1 引言

1.1 研究背景与现实意义

进入高质量发展阶段后，我国创新政策关注点正在由“投入规模持续扩张”逐步转向“创新产出的质量提升与结构优化”。

从全国层面看，研究与试验发展（R&D）经费总量、投入强度以及发明专利存量均保持增长态势[1–4]，这表明创新资源配置能力和知识产权积累能力都在持续增强。然而，投入增长本身并不自动等同于高质量创新成果的同步形成。若仅观察研发经费或专利数量的绝对规模，往往难以进一步辨识这些投入最终沉淀为高价值发明专利和稳定有效专利组合的实际程度。从区域治理角度看，省级地区既是科技资源配置的重要承接单元，也是创新政策落地的关键空间尺度。不同省份在产业基础、科教资源、市场化程度和成果转移条件等方面存在显著差异，因此，相同数量的研发投入进入不同区域后，其转化路径和边际效率并不一致。正因如此，省域比较不能停留在“谁投得更多、谁专利更多”的简单排序上，而应进一步追问“谁的投入更有效率，谁的创新系统更有利于形成高质量知识产权”。因此，知识产权“高质量”并不是一个由单一指标即可完整刻画的概念。仅以申请量或授权量衡量创新绩效，容易把数量扩张、结构改善、企业主体性增强以及单位投入产出效率提升混为一谈。研究省域研发投入转为高质量知识产权产出的过程，既有助于把创新规模、创新质量、创新结构和创新效率纳入同一分析体系，也能为区域分类施策提供更细致的统计依据。围绕这一现实关切，本文将研究对象聚焦于省域层面的投入—产出转化关系，并按“区域差异识别—作用机制解释—短期预测检验”的顺序展开分析。

1.2 已有研究评述与研究切入点

现有研究分别从政府研发支出、数字技术创新、技术转移网络演化和创新质量阶段性特征等角度讨论了创新投入与创新绩效之间的关系[5–8]。国内研究方向逐步从创新数量转向创新质量，既关注资金支持的非线性作用，也强调区域差异、技术转移网络和创新质量阶段演化的重要性[5–9]。国外研究则进一步从专利质量框架和模型方法两个方面提供启示。相关研究指出，专利质量应同时考虑技术内容、法律强度与经济影响，不同政策工具对专利质量提升也可能存在明显

异质性[10–13]。总体来看，现有研究为理解研发投入与创新质量关系提供了重要基础，但仍存在三点不足：一是难以同时捕捉质量、结构与效率三者之间的共同变化；二是更侧重“谁高谁低”的排序比较，而较少识别更具政策含义的类型差异；三是解释模型与预测模型常被分开使用，导致研究要么偏重机制解释而缺乏前瞻性，要么偏重预测效果而缺乏清晰的问题导向。基于上述不足，本文并不把预测作为脱离问题背景的独立任务，而是把它置于区域识别和机制解释之后，用以检验关键结构差异能否在短期外推中得到延续。具体而言，本文既关注哪些地区存在显著的投入—转化错位，也分析这种错位主要由何种统计机制支撑，并进一步比较在当前短面板条件下哪类模型更适合作为下一期高质量知识产权产出的预测工具。

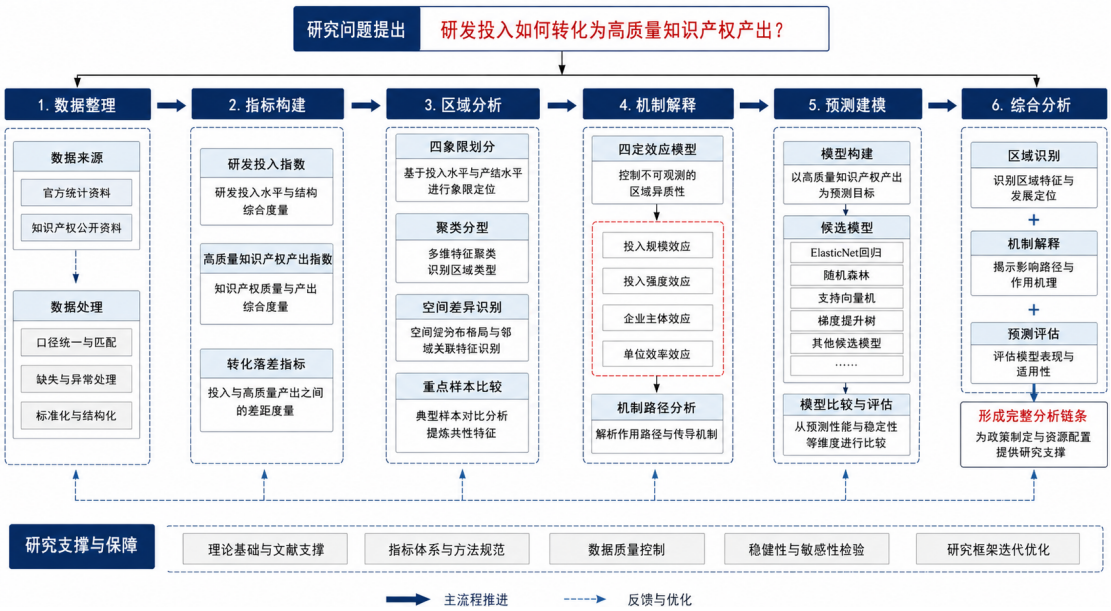

图 1　研发投入转化为高质量知识产权产出的研究思路与技术路线

1.3 研究问题与统计推断

图1展示了本文的总体研究思路。该图强调的是研究对象、分析步骤与方法之间的衔接关系，而不是结果呈现。其核心含义在于，先通过指标构建把“研发投入”和“高质量知识产权产出”转化为可比较的综合度量，再通过区域分析识别不同省份的转化差异，随后借助解释模型分析这些差异由何种因素支撑，最后以前述变量体系为基础开展短期预测比较，以检验相关结构特征是否具有延续性。（1）**研究问题**

在这一思路下，本文将研究问题进一步具体化为三个层面。第一，2020—

2024 年中国 31 个省级地区在研发投入与高质量知识产权产出方面呈现出怎样的整体变化趋势，这种变化是否存在显著区域错位。这个问题主要对应前文的综合指数构建与趋势识别，目的是先把不同地区放到统一口径下进行比较。第二，哪些地区属于“低投入高转化”或“高投入边际转化偏弱”的典型样本，这些差异主要由投入规模、企业主体结构还是单位投入效率所支撑。这个问题强调的不只是地区排序，更在于识别差异背后的结构来源。第三，在省域短面板条件下，如何对下一期高质量知识产权产出进行相对稳健的预测比较。这里的预测并不是与区域识别并列的独立任务，而是作为对前述识别结论的外推检验。围绕这三个问题，本文把区域识别、机制解释和预测比较设计成前后衔接的分析过程，而不是简单叠加若干模型方法。（2）**统计推断**

结合现有研究和中国区域创新现实，本文提出如下统计推断。第一，研发投入规模和投入强度是创新质量形成的基础条件，但不同地区会因数字创新环境、技术转移能力和政策组合差异而表现出不同边际效率[6;7;14]。这意味着在相近投入水平下，地区间仍可能出现明显不同的转化结果。第二，企业层面的专利质量与绩效联系说明，企业在专利申请、维持和转化中的主体性越强，研发投入越可能沉淀为可持续的高质量知识产权产出[11;15;16]。换言之，企业不仅是专利数量扩张的重要承担者，也是高质量专利能否真正进入转化环节的关键主体。第三，在省域短面板、小样本且变量相关性较强的条件下，完全依赖高复杂度树模型未必最优。相反，同时引入 L_1 与 L_2 惩罚的 ElasticNet 更有利于兼顾变量筛选、系数稳定性与跨年份预测能力[17]。这一判断并不是单纯追求算法简化，而是强调模型复杂度应当与样本规模、变量结构和外推任务相匹配。对于当前这类以省域短面板为特征的数据场景，稳定性和可解释性往往比局部拟合优势更重要。本文不预设“投入越高，转化一定越优”的简单线性关系，而是引入综合指数和转化落差指标来识别地区之间的错位结构。若某省份的高质量知识产权产出指数显著高于研发投入指数，则说明其创新系统存在较强的投入转化效率。反之，则说明其创新体系尚有明显的提质增效空间。

2 指标测度与研究方法

2.1 数据来源与样本说明

本文数据全部来自公开、可核验的官方资料。研发投入数据主要来源于国家统计局发布的《2022 年全国科技经费投入统计公报》《2023 年全国科技经费投入统计公报》《2024 年全国科技经费投入统计公报》以及国家统计局、科学技术部、财政部联合发布的《2021 年全国科技经费投入统计公报》[1–3;18]。知识产权数据则主要来源于国家知识产权局发布的《2024 年知识产权统计年报汇编》中分地区发明专利申请、授权、有效量以及企业专利占比等统计表[4]。在统一口径、处理缺失并核对地区名称后，最终得到 2020—2024 年 31 个省级地区共 155 条观测值的平衡面板。

2.2 指标测度

为避免使用单一指标导致的信息损失，本文将“高质量知识产权产出”拆解为授权效率、质量结构、企业主体和单位投入产出四类维度。设 x_{ijt} 表示第 i 个地区在 t 期第 j 个指标的原始值，则正向指标标准化为

$$z_{ijt} = \frac{x_{ijt} - \min(x_j)}{\max(x_j) - \min(x_j)}. \tag{1}$$

在标准化基础上，计算指标占比 p_{ijt}

$$p_{ijt} = \frac{z_{ijt}}{\sum_{i=1}^{n} z_{ijt}}, \tag{2}$$

并进一步得到熵值 e_j 与权重 w_j

$$e_j = -\frac{1}{\ln n} \sum_{i=1}^{n} p_{ijt} \ln p_{ijt}, \qquad w_j = \frac{1 - e_j}{\sum_{j=1}^{m} (1 - e_j)}. \tag{3}$$

据此构建高质量知识产权产出指数 HQ_{it} 和研发投入指数 RD_{it}

$$HQ_{it} = \sum_{j=1}^{m_1} w_j z_{ijt}, \qquad RD_{it} = \sum_{k=1}^{m_2} \omega_k z_{ikt}. \tag{4}$$

最后定义转化落差指标 Gap_{it}

$$Gap_{it} = HQ_{it} - RD_{it}, \tag{5}$$

其中 $Gap_{it} > 0$ 表示高质量产出相对投入存在超额表现，反之则说明投入规模尚未有效转化为对应质量产出。式 (1) 至式 (5) 中，z_{ijt} 表示标准化后的指标值，p_{ijt} 表示标准化指标在同一期样本中的占比，e_j 表示第 j 个指标的信息熵，w_j 与 ω_k 分别表示高质量知识产权产出指数和研发投入指数的熵权权重，HQ_{it} 表示第 i 个地区在 t 期的高质量知识产权产出指数，RD_{it} 表示研发投入指数，Gap_{it} 表示两者之差所刻画的投入—转化落差。表 eftab:weights 给出了指标与熵权结果。高质量知识产权产出指数中，发明专利存量占比、单位研发有效专利量和授权率权重较高，说明“质量结构”和“单位投入效率”是区分地区创新转化能力的关键维度；研发投入指数中，R&D 经费权重为 0.7143，显著高于 R&D 强度，说明在当前样本中，投入规模差异仍是地区创新投入差异的主导来源。

表 1　指标体系与熵权结果

指标组	指标名称	权重
高质量知识产权产出指数	授权率	0.1683
	发明专利存量占比	0.2333
	企业有效专利占比	0.1140
	企业申请占比	0.1062
	单位研发有效专利量	0.2094
	单位研发授权量	0.1688
研发投入指数	R&D 经费	0.7143
	R&D 强度	0.2857

在得到指标权重之后，本文进一步计算了 2024 年 31 个省级地区的研发投入指数、高质量知识产权产出指数和转化落差。表2给出了完整测度结果。可以看到，北京、广东、江苏、浙江、上海等地区在高质量产出指数上处于前列，但其中广东、江苏的转化落差为负；而西藏、吉林、黑龙江、广西等地区虽然研发投入指数较低，却在高质量产出指数和转化落差上表现突出。这说明“投入规模领先”与“转化表现领先”并不完全重合，省域创新转化存在明显的结构性差异。

表 2　2024 年 31 个省级地区创新投入与高质量产出测度结果

地区	研发投入指数	高质量产出指数	转化落差	地区	研发投入指数	高质量产出指数	转化落差
北京	0.7339	0.8654	0.1316	天津	0.2253	0.5023	0.2770
河北	0.2134	0.4371	0.2237	山西	0.0860	0.4729	0.3870
内蒙古	0.0655	0.3332	0.2676	辽宁	0.1888	0.4923	0.3034
吉林	0.0916	0.6100	0.5184	黑龙江	0.0898	0.5679	0.4781
上海	0.5063	0.5679	0.0616	江苏	0.7794	0.6377	-0.1417
浙江	0.5356	0.6167	0.0811	安徽	0.3046	0.6006	0.2959
福建	0.2600	0.4711	0.2110	江西	0.1672	0.3783	0.2110
山东	0.4678	0.5512	0.0834	河南	0.2552	0.3908	0.1356
湖北	0.3159	0.4886	0.1727	湖南	0.2984	0.4981	0.1997
广东	0.8602	0.7089	-0.1513	广西	0.0620	0.5063	0.4443
海南	0.0645	0.3694	0.3048	重庆	0.2086	0.4252	0.2167
四川	0.3000	0.4991	0.1991	贵州	0.0660	0.4588	0.3928
云南	0.0916	0.3624	0.2709	西藏	0.0031	0.6567	0.6536
陕西	0.2323	0.5088	0.2765	甘肃	0.0675	0.3806	0.3131
青海	0.0309	0.3798	0.3488	宁夏	0.0777	0.4018	0.3241
新疆	0.0378	0.2789	0.2410				

2.3 研究方法

在区域识别部分，本文以 2024 年 RD_{it} 和 HQ_{it} 的中位数为阈值进行四象限划分，再结合 Gap_{it} 和企业有效专利占比进行 KMeans 聚类，以捕捉“高投入高产出”“高投入低转化”“低投入高转化”“低投入低产出”等典型模式。之所以在四象限之后继续使用聚类方法，是因为仅凭两个维度还不足以区分位置相近地区内部的结构差异，而 KMeans 能够把投入、产出、转化落差和企业主体结构同时纳入识别。

在机制解释部分，本文使用固定效应模型识别投入规模、投入强度和企业主体变量的作用方向。考虑到省域样本同时存在不可观测的地区异质性和年份共同冲击，本文在混合 OLS、地区固定效应、年份固定效应和双向固定效应之间做分步比较，并采用按地区聚类的稳健标准误。设 HQ_{it} 表示第 i 个地区在 t 期的高质量知识产权产出指数，$RDExp_{i,t-1}$ 表示滞后一期研发经费，$Intensity_{i,t-1}$ 表示滞后一期 R&D 强度，$EntApp_{i,t-1}$ 与 $EntStock_{i,t-1}$ 分别表示企业发明专利申请占比和企业有效发明专利占比，μ_i 表示地区固定效应，λ_t 表示年份固定效应，ε_{it} 为随机扰动项。模型形式为

$$\begin{aligned} HQ_{it} = {} & \alpha + \beta_1 \ln(RDExp_{i,t-1} + 1) + \beta_2 Intensity_{i,t-1} + \beta_3 Intensity_{i,t-1}^2 \\ & + \beta_4 EntApp_{i,t-1} + \beta_5 EntStock_{i,t-1} + \mu_i + \lambda_t + \varepsilon_{it}. \end{aligned} \tag{6}$$

在预测比较部分，本文以前文识别出的区域差异和机制变量为基础，对下一期高质量知识产权产出进行短期外推检验，而不把预测视为脱离前文的独立任务。若某些区域类型在外推中持续出现系统性误差，则说明现有解释变量尚未完全覆盖其结构差异。设 y_i 为被预测地区的高质量知识产权产出指数，X_i 为输入特征向量，β_0 为截距项，β 为系数向量，λ 为正则化强度，ρ 为 L_1 与 L_2 惩罚的权重参数。ElasticNet 目标函数为

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - X_i\beta)^2 + \lambda \left[\rho \|\beta\|_1 + \frac{1-\rho}{2} \|\beta\|_2^2 \right] \right\}, \tag{7}$$

并用滚动年份验证对模型进行比较和调参。由于样本属于短面板，本文不使用随机 K 折，而是令训练集始终早于测试集。具体而言，2022 年和 2023 年作为滚动验证年份，2024 年作为完全未参与调参的保留集。

3 省域创新转化的区域差异识别

3.1 总体趋势与四象限识别

2020—2024 年，全国省域平均 R&D 强度和平均授权率整体向上，说明我国区域创新系统正在由“扩规模”向“提质量”过渡。2024 年，研发投入指数中位数为 0.2086，高质量知识产权产出指数中位数为 0.4923。以这两个中位数为划分标准，31 个省份中有 11 个属于“高投入高产出”，10 个属于“低投入低产出”，5 个属于“低投入高产出”，5 个属于“高投入低转化”。这一结果首先说明，我国省域创新格局仍然具有较明显的梯度特征，高投入高产出地区主要集中在创新基础较强、产业体系较完整的区域，但与此同时，四象限分布并没有简单演化为“投入越高、位置越靠右上”的单一路径，不少地区在投入水平和高质量产出之间仍存在显著错位。

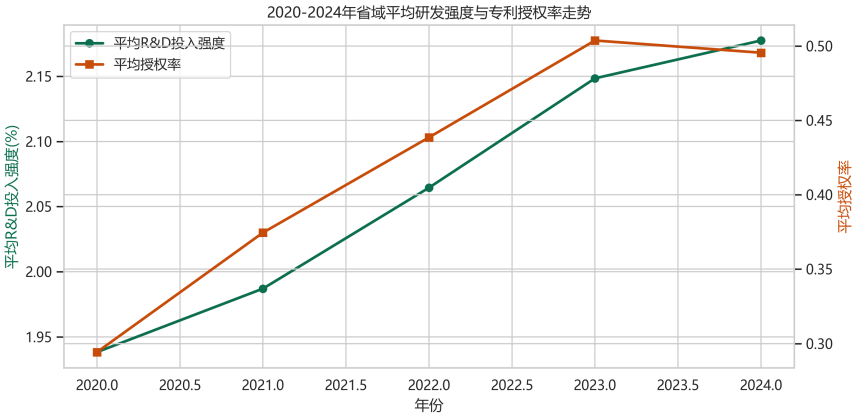

图 2　2020—2024 年省域平均 R&D 强度与授权率变化趋势

从图3可以更细致地看到这种错位。西藏、吉林、黑龙江和广西位于左上象限，意味着这些地区虽然研发投入指数不高，但高质量知识产权产出指数已经超过全国中位水平。就统计含义而言，这类地区的优势更可能来自单位投入效率、专利质量结构或特定产业技术积累，而不是总量扩张。相较之下，广东、江苏虽然仍位于右上象限，说明其投入与产出总体水平均处于全国前列，但两地转化落差为负，表明在综合指数口径下，其高投入并没有同步转化为更强的超额质量优势。也就是说，四象限图并不是简单给地区贴上“先进”或“落后”的标签，而是提示我们，高投入地区需要关注边际转化效率，低投入但高产出地区则值得进一步追问其支撑机制。

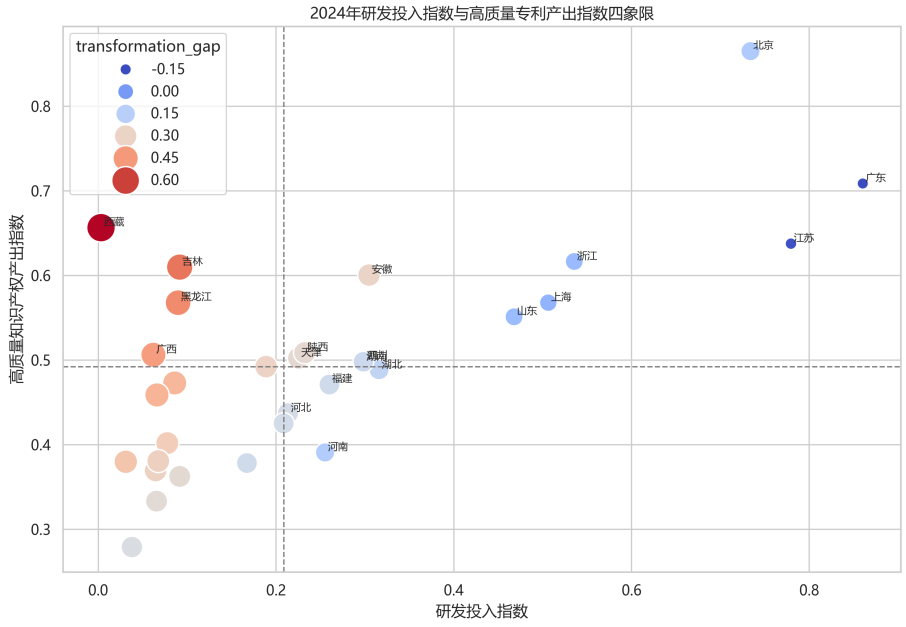

图 3　2024 年研发投入指数与高质量知识产权产出指数四象限

因此，图4将 2024 年转化落差映射到省域地理单元上。与四象限图相比，这张地图的优势不在于再次给出分类结论，而在于把“错位”放回地理空间之中考察。可以看到，西藏、吉林、黑龙江、广西、贵州等地区形成了较明显的正向转化落差集聚，而广东、江苏则表现为负向转化落差较突出的高投入样本。这样的空间分布说明，区域创新转化并不只是单个省份的孤立表现，而可能与产业结构、技术转移网络、科研组织方式以及区域间要素流动共同相关。

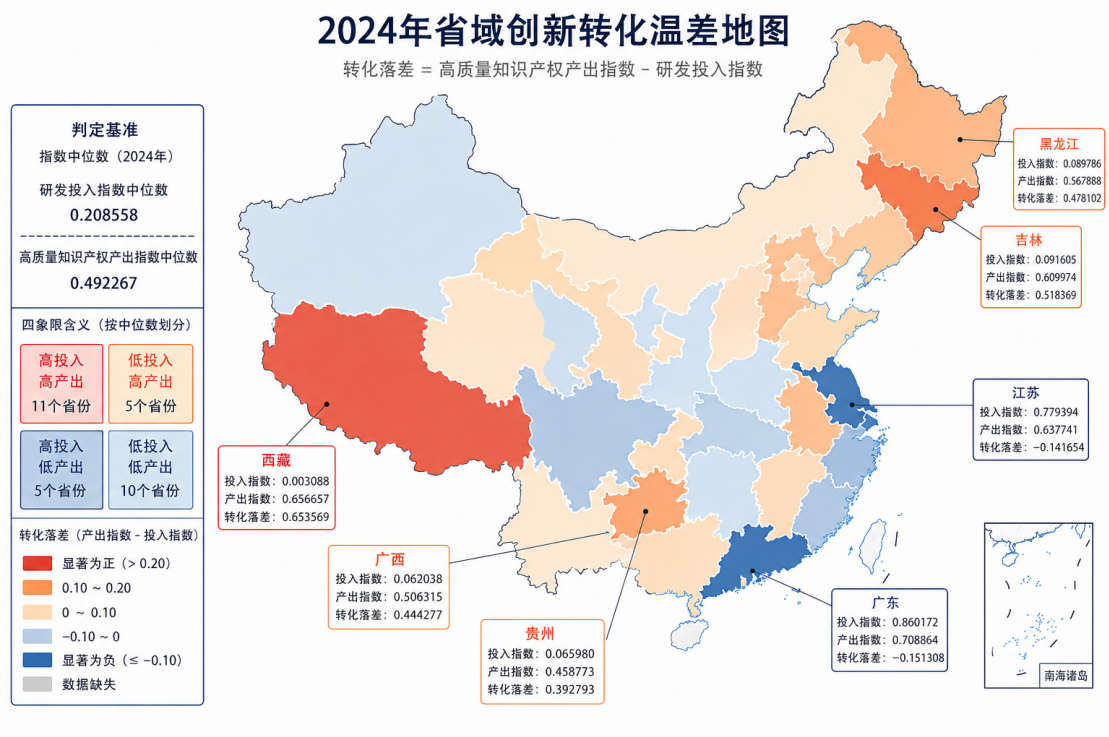

图 4　2024 年省域创新转化温差地图

表 3　2024 年代表省份创新投入—转化表现

地区	研发投入指数	高质量产出指数	转化落差
西藏	0.0031	0.6567	0.6536
吉林	0.0916	0.6100	0.5184
黑龙江	0.0898	0.5679	0.4781
广西	0.0620	0.5063	0.4443
贵州	0.0660	0.4588	0.3928
广东	0.8602	0.7089	-0.1513
江苏	0.7794	0.6377	-0.1417
北京	0.7339	0.8654	0.1316

表3进一步选取了最具代表性的省份进行比较。可以看到，西藏、吉林、黑龙江、广西和贵州共同表现为“低投入但转化落差显著为正”，其共同特点不是

投入规模大，而是单位投入产出效率或质量结构较强。广东和江苏则体现出“高投入高产出但超额转化优势不足”的另一种结构。换言之，同样是高质量产出较高的地区，其支撑机制并不一致，这也是后续需要通过聚类分析和解释模型进一步拆解的问题。

3.2 KMeans 分型与反直觉样本

进一步将 HQ_{it}、RD_{it}、Gap_{it} 和企业有效专利占比联合纳入 KMeans 聚类后，省域创新转化被识别为四类：高能引领型、效率跃升型、均衡追赶型和单点极值型。与四象限图不同，聚类分析并不只是利用两个维度来判断地区位置，而是把投入、产出、转化落差和企业主体结构放在一起综合识别。因此，它更适合回答“看上去位置接近的地区，内部支撑机制是否相同”这一问题。

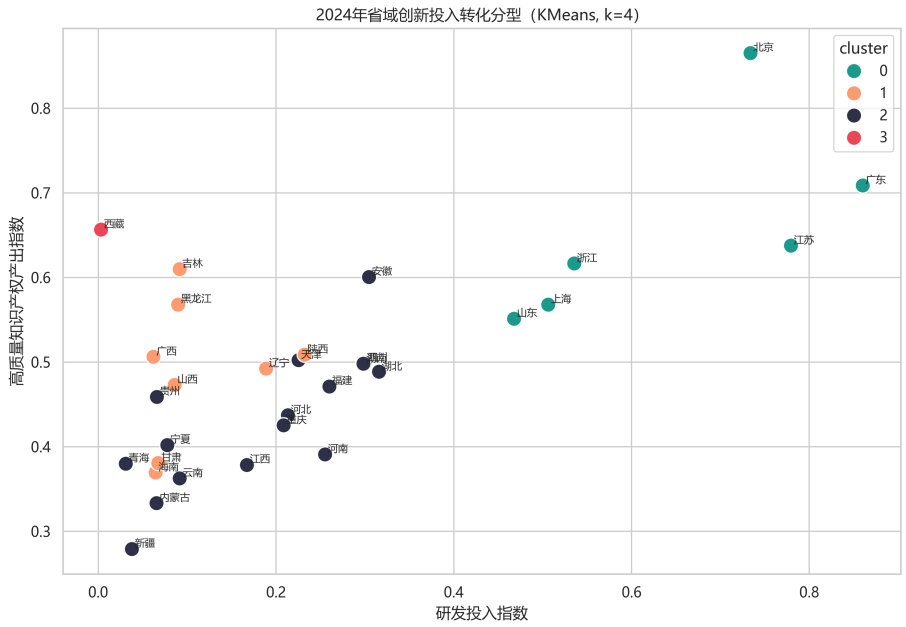

图 5　2024 年省域创新投入转化分型（KMeans）

图5表明，东部沿海省份更多集中于“高投入高产出”区域，但并不意味着其全部具备最优的转化效率；相反，一些中西部与东北地区虽然投入水平不高，却形成了更高的单位投入产出效率。KMeans 结果和四象限分析相互印证，但它的优势在于能把“同属左上象限”或“同属右上象限”的地区进一步拆开。以北京、广东、江苏、浙江为代表的高能引领地区，同时兼具高授权率、高发明存量占比和较强企业主体性；以吉林、黑龙江、广西等为代表的效率跃升地区，则更多依靠单位研发投入对应的高有效专利产出支撑其位置；均衡追赶型省份虽然

在企业参与度上并不弱，但发明专利质量结构和授权效率仍有明显补短空间；西藏则单独构成单点极值型。表4把四类类型对应到具体省份，便于把抽象聚类结果转化为可以直接讨论的地区清单。

表 4　2024 年 KMeans 分型及对应省份

类型	对应省份	主要特征
高能引领型	上海、北京、山东、广东、江苏、浙江	研发投入与高质量产出均处于高位，企业主体性较强，但广东、江苏的转化落差已转为负值
效率跃升型	吉林、山西、广西、海南、甘肃、辽宁、陕西、黑龙江	投入规模偏低，但转化落差整体显著为正，单位投入效率和存量结构表现更突出
均衡追赶型	云南、内蒙古、四川、天津、宁夏、安徽、新疆、江西、河北、河南、湖北、湖南、福建、贵州、重庆、青海	多数指标处于中间区间，具备一定转化能力，但授权效率和专利结构仍有补短空间
单点极值型	西藏	极低投入与极高单位产出组合形成的单独样本，统计位置显著偏离其他类型

3.3 企业主体演化与单位投入效率的探索性画像

如果说四象限图和聚类图主要解决“谁处于什么位置”的问题，那么企业主体热力图和单位投入效率图则进一步回答“这些位置是由什么支撑起来的”。图6显示，2020—2024 年企业有效发明专利占比整体呈上升趋势，样本均值由 61.46% 提高到 66.25%。但这种上升并不平滑，也不是所有地区同步发生：2024 年广东、安徽、宁夏、浙江和西藏的企业有效发明专利占比分别达到 86.86%、80.72%、80.55%、78.91% 和 78.38%，而黑龙江、吉林、甘肃、陕西和辽宁则分别只有 36.11%、43.52%、45.73%、46.19% 和 52.99%。

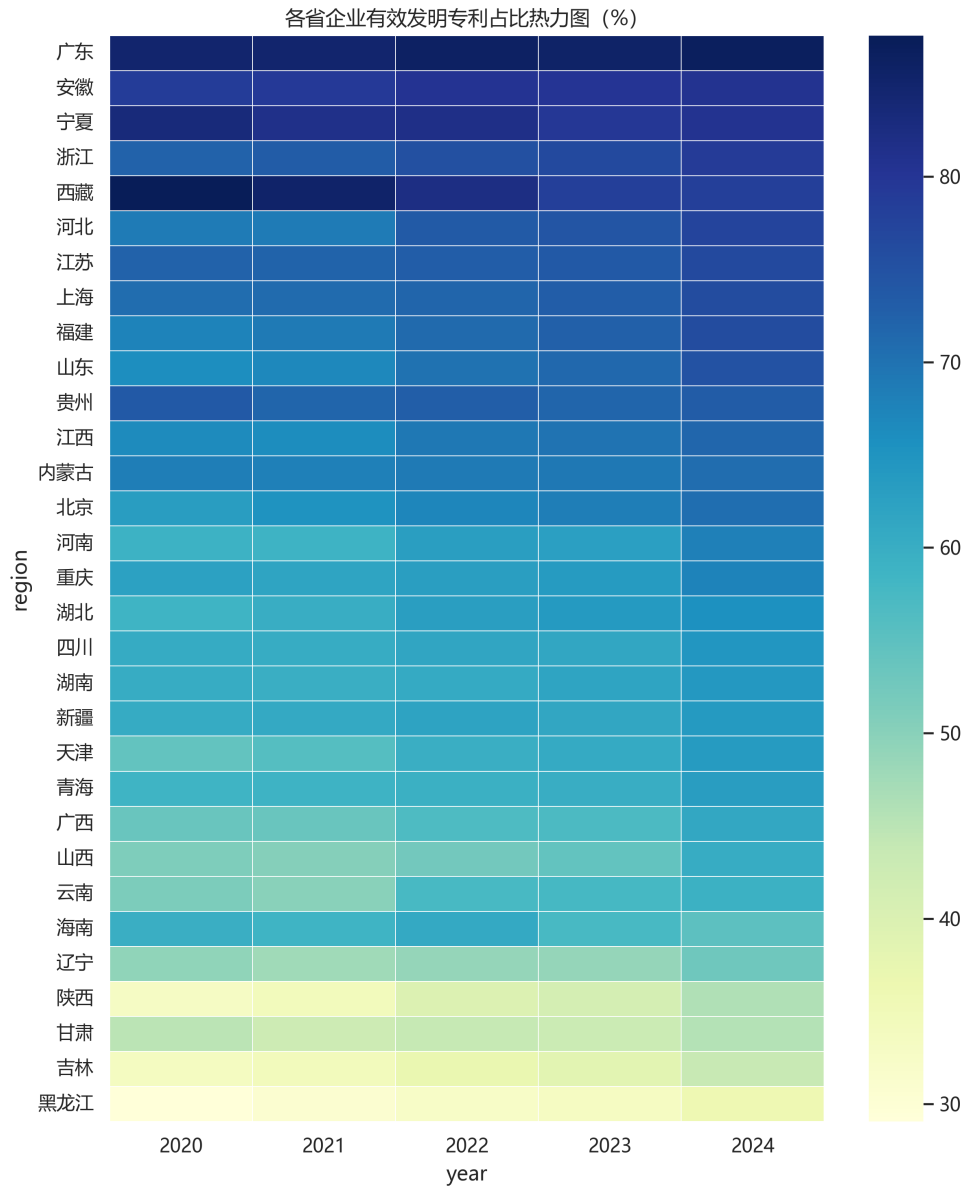

图 6　2020—2024 年企业有效发明专利占比热力图

更值得注意的是，企业主体地位的演化方向同样存在明显差异。2020—2024 年间，企业有效发明专利占比增幅最大的省份并不集中在传统意义上的创新强省，而是陕西、吉林、天津、山西、河南和山东；相反，西藏、海南和宁夏则出现回落。这表明“企业主导”具有明显的结构性和阶段性，有些地区是企业主体持续强化后带动了高质量产出，有些地区则更多依赖科研系统或既有存量形成当前优势。

图7与图8从另一个角度给出了更有解释力的线索：相较于研发投入指数与高质量产出指数之间 0.697 的相关系数，单位研发有效专利量与高质量产出指数之间的相关系数达到 0.848，说明“单位投入产出效率”比“投入规模”更贴近

最终的质量结果。2024 年每亿元研发经费对应有效发明专利量排名前八的地区分别是北京、黑龙江、西藏、吉林、广西、广东、浙江和安徽，这一结果解释了为何一些投入并不突出的地区仍能在高质量产出指数上取得较高排名。

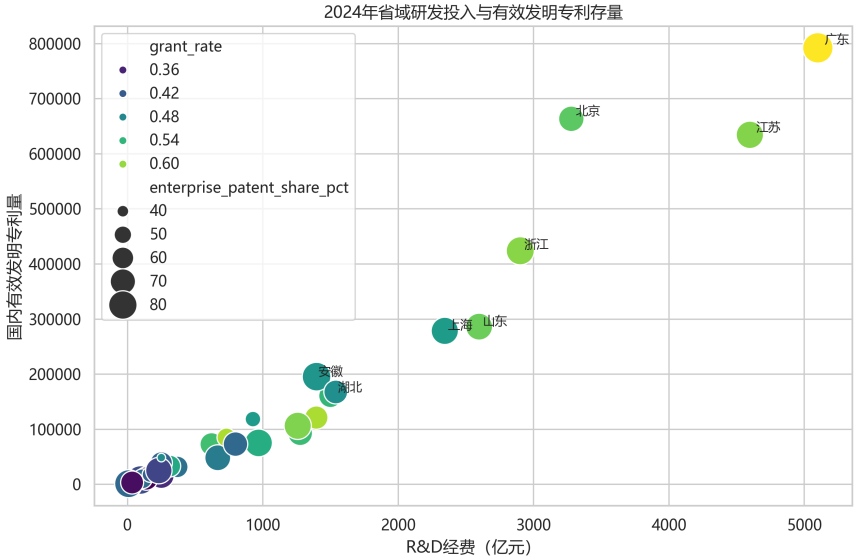

图 7　2024 年研发投入指数与单位研发有效专利量散点图

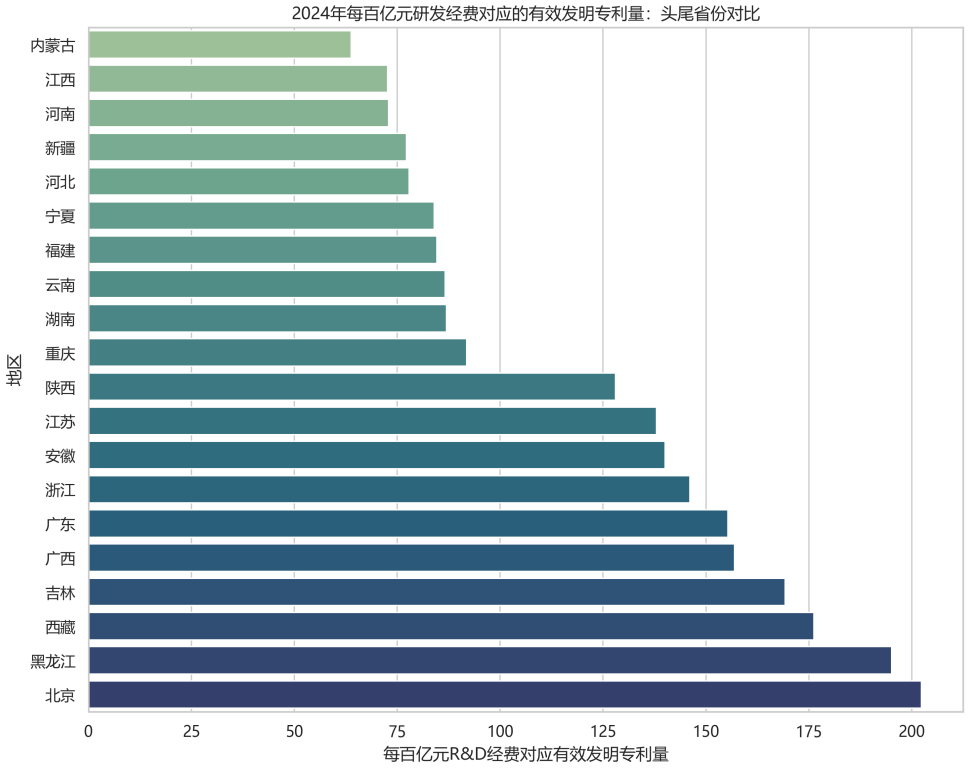

图 8　2024 年每亿元研发经费对应有效发明专利量排序

把企业主体地位和单位投入效率放在一起看，会得到一个比“谁投入多、谁

专利多”更有政策意义的判断：有些地区的问题不是投入不足，而是投入尚未高效沉淀为企业可支配、可转化的高质量专利资产；另一些地区则已经形成较强的效率优势，但需要进一步补足资源供给和产业承载能力。

3.4 聚类画像与区域异质性的补充讨论

图9把各类省份的核心指标均值并列展示出来，使 KMeans 分型的含义变得更直观，表4给出了四类类型的成员省份。高能引领型（6 个省份）同时拥有最高的授权率、较高的发明专利存量占比以及较强的企业主体性，其研发投入指数和高质量产出指数均值分别达到 0.647 和 0.658；效率跃升型（8 个省份）虽然研发投入指数均值只有 0.110，但高质量产出指数均值达到 0.489，转化落差高达 0.378；均衡追赶型（16 个省份）在多数指标上居于中间位置，是当前最需要通过结构优化实现跃升的主体类型；西藏单独形成的单点极值型，则表现出极低投入与极高单位产出的组合特征。

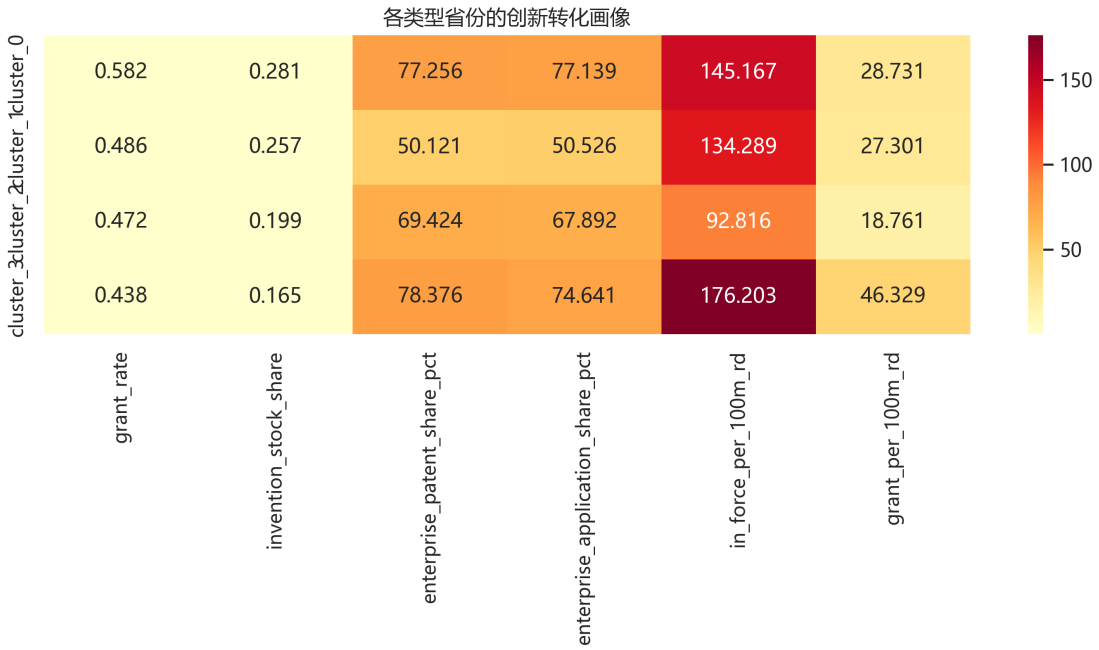

图 9　各类省份的创新转化画像热力图

这一聚类画像的意义在于，它把“区域差异”从平面上的位置差异推进到结构上的机制差异。对于高能引领型地区，政策重点不应只是延续高投入，而应进一步提高高价值专利和产业转化的边际质量；对于效率跃升型地区，重点是放大单位投入效率优势，改善其长期资源约束；对于中间梯队地区，则应同时从授权效率、发明专利存量结构和企业创新组织方式三方面补短。

图10进一步对比了低投入高转化代表省份与高投入但边际转化优势有限地区的结构差异。该图把投入规模、企业主体地位、单位投入产出效率和转化落差四个维度放到同一张结构化对照图中，有助于读者看清不同区域类型背后的支撑机制，而不是只停留在“谁高谁低”的表面比较。

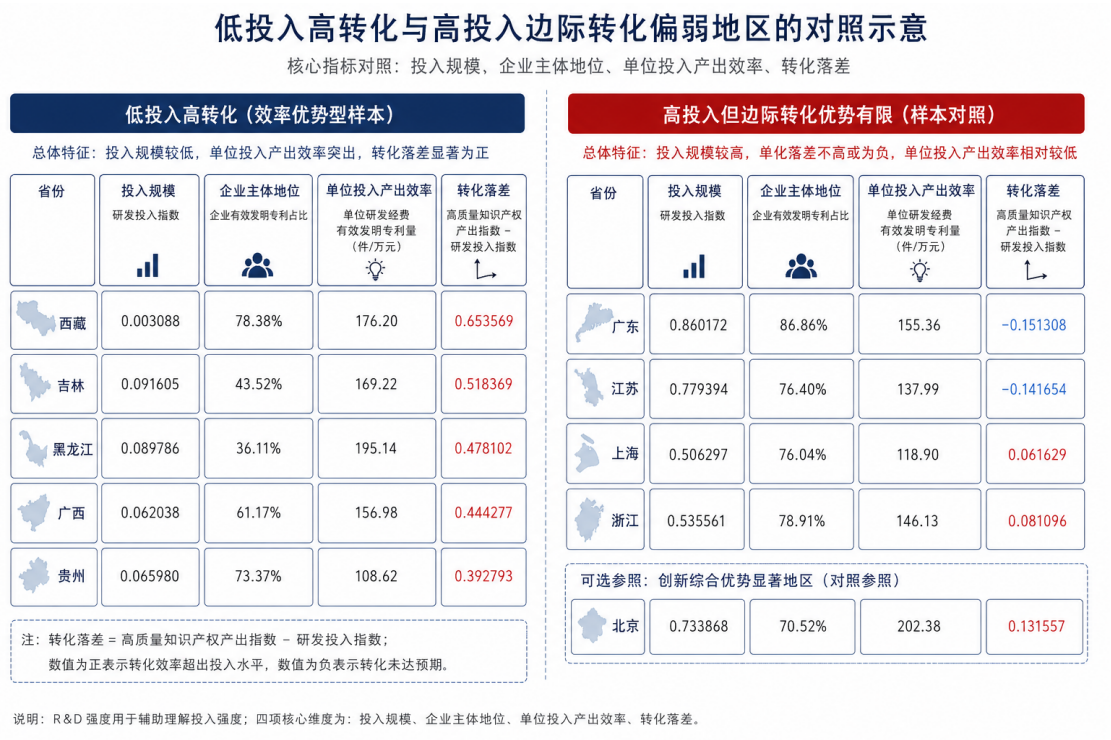

图 10　低投入高转化与高投入边际转化偏弱地区的对照示意

4 作用机制解释与短期预测检验

4.1 建模前的变量关系与可行性检视

在进入解释模型和预测模型之前，有必要先检视变量之间的相关关系。图11显示，2024 年横截面上，研发投入指数与高质量知识产权产出指数的相关系数为 0.697，说明投入扩张总体上仍然与质量提升同向变化；但与此同时，研发投入指数与转化落差的相关系数为 −0.848，这意味着“投入越高，超额转化优势越不明显”反而是一个更强的统计事实。进一步看，单位研发有效专利量与高质量产出指数之间的相关系数达到 0.848，明显强于研发投入指数与单位研发有效专利量之间的相关系数 0.339。

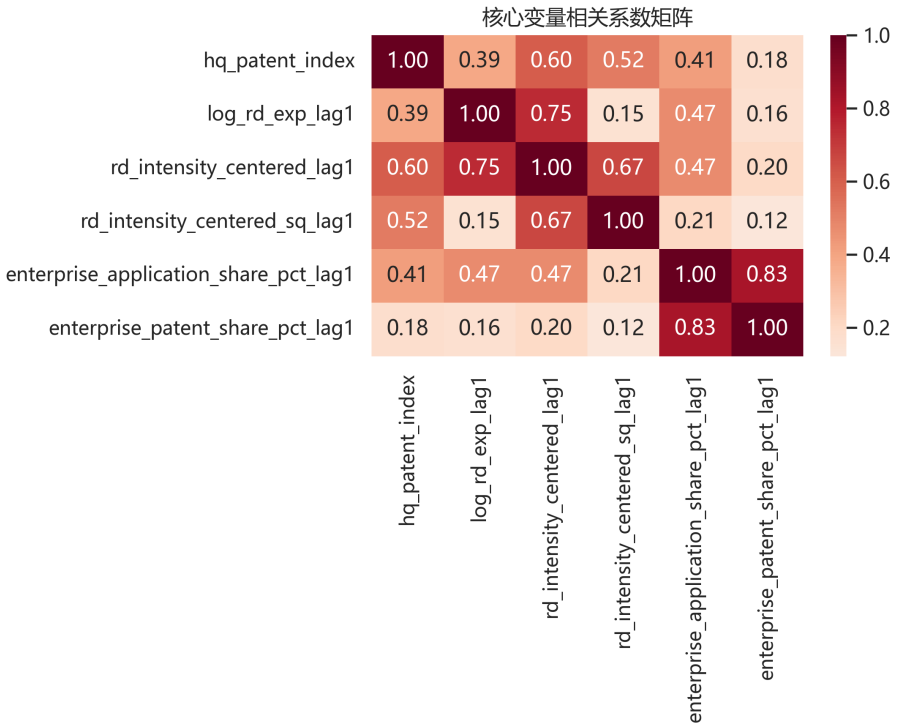

图 11　建模核心变量相关系数热力图

相关矩阵还揭示了若干对建模有直接影响的结构特征。第一，企业有效专利占比与企业申请占比的相关系数高达 0.933，单位研发有效专利量与单位研发授权量的相关系数达到 0.879，这说明企业主体性指标和单位效率指标内部都存在较强共线性。第二，发明专利存量占比与高质量产出指数的相关系数为 0.705，表明“存量结构质量”对综合高质量产出具有稳定贡献。基于这些事实，本文在固定效应模型中只保留最具经济含义的核心变量，在预测模型中则优先采用带正则化约束的 ElasticNet。

4.2 固定效应模型结果

为比较不同控制方式下固定效应模型结果的变化，本文分别估计混合 OLS、地区固定效应、年份固定效应和双向固定效应四种设定。这里的调整后 R^2，是对解释变量数量进行修正后的拟合优度指标，用于避免因控制变量增多而机械抬高模型拟合度。表5显示，仅使用混合 OLS 时，模型调整后 R^2 为 0.4239；仅控制地区固定效应后，调整后 R^2 上升到 0.9179；仅控制年份固定效应时，调整后 R^2 为 0.5455；同时控制地区和年份固定效应后，调整后 R^2 进一步提高到 0.9523。这个对比说明，省域高质量知识产权产出的变化既受到显著的地区异质

性影响，也受到共同年份冲击影响，其中地区固定效应所吸收的不可观测差异尤其重要。

表 5　固定效应模型分步设定比较

变量	混合 OLS	地区固定效应	年份固定效应	双向固定效应
滞后 $\ln(RDExp)$	-0.0077	0.3279^{***}	-0.0010	-0.1581^{*}
	(0.0265)	(0.0709)	(0.0255)	(0.0914)
中心化滞后 R&D 强度	0.0321	-0.1071	0.0433	0.0111
	(0.0335)	(0.0669)	(0.0347)	(0.0565)
中心化滞后 R&D 强度平方	0.0083	0.0501^{**}	0.0074	0.0091
	(0.0060)	(0.0208)	(0.0063)	(0.0095)
滞后企业申请占比	0.0053^{***}	0.0027^{***}	0.0000	-0.0002
	(0.0019)	(0.0010)	(0.0029)	(0.0010)
滞后企业有效专利占比	-0.0033	0.0059^{*}	0.0004	0.0001
	(0.0021)	(0.0034)	(0.0028)	(0.0032)
2022 年虚拟变量			0.0544^{***}	0.0799^{***}
			(0.0062)	(0.0134)
2023 年虚拟变量			0.1170^{***}	0.1630^{***}
			(0.0150)	(0.0320)
2024 年虚拟变量			0.1395^{***}	0.2031^{***}
			(0.0266)	(0.0418)
地区固定效应	否	是	否	是
年份固定效应	否	否	是	是
样本量	124	124	124	124
调整后 R^2	0.4239	0.9179	0.5455	0.9523

注：括号内为按地区聚类的稳健标准误，***、**、* 分别表示在 1%、5% 和 10% 水平上显著。

从这一比较看，如果不控制固定效应，模型会把大量地区间既有差异直接归因于解释变量，从而得到相对粗糙的平均关系；仅控制地区固定效应时，滞后研发经费系数转为显著为正，说明长期地区基础与研发规模存在较强耦合；而在双向固定效应设定下，滞后 $\ln(RDExp)$ 系数变为 -0.1581，且在 10% 水平边际显著，这意味着在扣除共同年份趋势和地区不变特征之后，单纯扩大投入规模并不能保证当期质量产出同步跃升。图12给出了双向固定效应模型核心系数及 95% 置信区间，便于与表格结果相互印证。

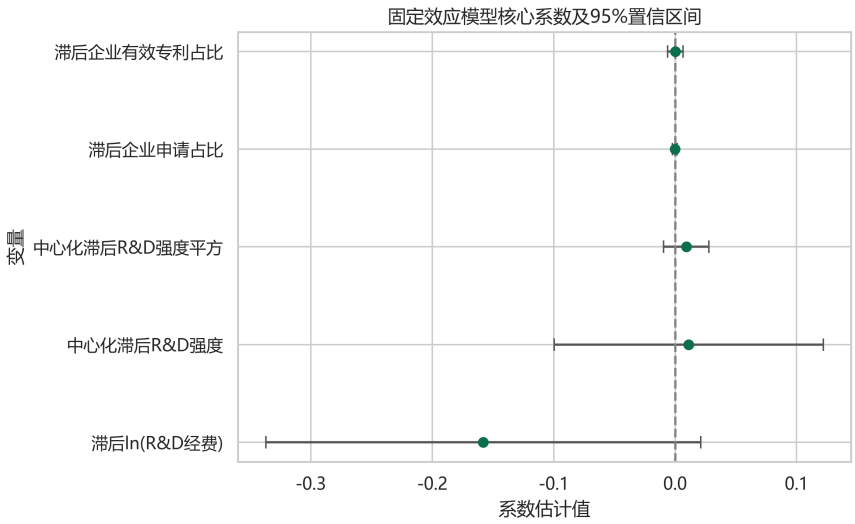

图 12　固定效应模型核心系数及 95% 置信区间

从解释变量方向看，滞后 $\ln(RDExp)$ 系数为 -0.1581，在 10% 水平边际显著，提示单纯扩大投入规模并不能保证当期质量产出同步跃升；中心化后的 R&D 强度和平方项均为正，但统计显著性较弱，说明强度提升更可能通过较长链条和更复杂机制发挥作用。滞后企业有效专利占比系数为正，企业申请占比系数为负，同样反映出“主体规模扩张”和“真正形成高质量存量”之间存在时间错位。

4.3 短期预测比较与调参结果

承接前文的区域识别和机制解释结果，本文进一步比较不同模型对下一期高质量知识产权产出的短期预测能力。评价指标采用滚动年份验证平均 R^2、平均 MAE 及 2024 年保留集表现，并构建 ElasticNet、RandomForest、ExtraTrees、XGBoost 和 CatBoost 五类基线模型，其中对 ElasticNet、ExtraTrees 和 XGBoost 做进一步调参。图13给出了基线模型结果。可以看到，在不调参时，ExtraTrees 和 XGBoost 在 2024 年保留集上的 R^2 略高于 ElasticNet，但 ElasticNet 的滚动验证 MAE 最优，说明它在时间滚动意义下更稳定。树模型在当期拟合上有一定优势，而正则化线性模型在小样本、按年份外推的设定下波动更小。

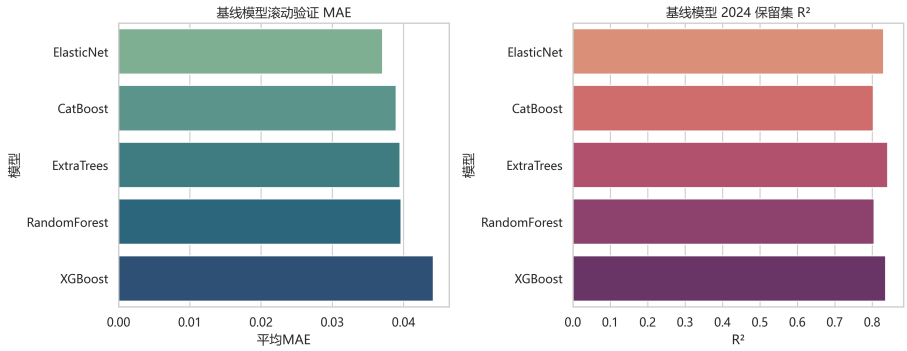

图 13　基线预测模型比较结果

为了避免“只看某一年测试集”的偶然性，本文将调参目标明确绑定在滚动年份验证表现上。ElasticNet 最终选择 $\alpha = 0.005$、$l1\_ratio = 0.6$，意味着模型在保留较强线性解释性的同时，适度引入 L_1 惩罚来压缩冗余变量；ExtraTrees 的最优结构为 228 棵树、最大深度 7、最小叶节点样本数 2；XGBoost 则采用 301 棵树、4 层深度、0.0769 学习率，并结合 0.8176 的子样本比例和 0.7012 的列采样比例，以控制学习过程中的方差。表6汇总了主要候选模型的最终设定。

表 6　主要候选模型的调参与结构设定

模型	参数或结构设定
ElasticNet_tuned	$\alpha = 0.005$，$l1\_ratio = 0.6$；数值变量做中位数填补与标准化，地区变量做独热编码
ExtraTrees_tuned	`n_estimators`=228，`max_depth`=7，`min_samples_leaf`=2，`min_samples_split`=5
XGBoost_tuned	`n_estimators`=301，`max_depth`=4，`learning_rate`=0.0769，`subsample`=0.8176，`colsample_bytree`=0.7012
WeightedEnsemble	对 ElasticNet_tuned、ExtraTrees_tuned、XGBoost_tuned 加权集成；权重分别为 0.3691、0.3223、0.3086

表7和图14显示，调优后的 ElasticNet 在两套标准上均表现最优：滚动验证 $cv\_R^2 = 0.8140$、$cv\_MAE = 0.0330$，2024 年保留集 $R^2 = 0.8861$、$MAE = 0.0319$。与其基线版本相比，ElasticNet 的保留集 R^2 提升了 0.0559，MAE 下降了 0.0082，说明“温和调参 + 严格时间验证”能够在当前样本条件下带来较稳定的增益。这里的预测比较并不是另起一层分析，而是检验基于投入规模、企业主体和单位效率构建的变量体系，能否在短期外推中继续刻画地区差异。相较之下，XGBoost 调参后也有改善，但提升幅度不如 ElasticNet；ExtraTrees 在调参后反而从基线的 $R^2 = 0.8417$ 下降到 0.8130。

表 7　不同预测模型的滚动验证与保留集表现

模型	滚动验证 R^2	滚动验证 MAE	2024 年保留集 R^2	2024 年保留集 MAE
ElasticNet_tuned	0.8140	0.0330	0.8861	0.0319
XGBoost_tuned	0.7516	0.0395	0.8488	0.0376
ExtraTrees_tuned	0.7674	0.0378	0.8130	0.0413
RandomForest	0.7442	0.0396	0.8049	0.0396
WeightedEnsemble	—	—	0.8736	0.0350

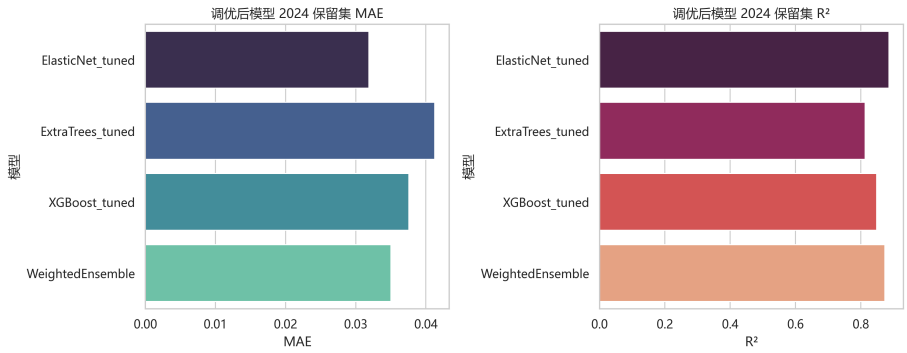

图 14　调参后模型比较结果

这种结果具有明确的方法论意义。在当前样本只有 31 个地区、5 个年份且特征主要由滞后变量构成的情形下，过于复杂的树模型并没有稳定地超越正则化线性模型。ElasticNet 通过同时引入 L_1 和 L_2 惩罚，在保留解释性的同时抑制了过拟合，更适合作为本文的主预测模型。

4.4 最优模型的变量贡献、误差诊断与稳健性讨论

最优模型的标准化系数显示，滞后高质量产出指数（0.0956）是最重要的路径依赖变量，其后依次是滞后单位研发有效专利量（0.0098）、滞后授权率（0.0050）、滞后 R&D 强度（0.0048）和滞后 $\ln(RDExp)$（0.0040）。这说明省域创新质量存在较强的累积效应，提升高质量产出既依赖既有存量基础，也依赖单位投入效率和授权转化效率。

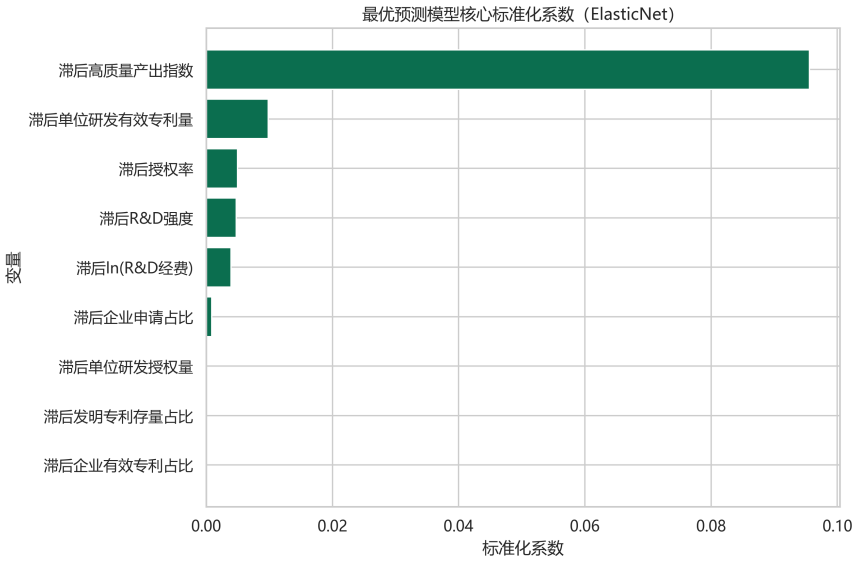

图 15　最优预测模型的核心标准化系数

图16给出了 2024 年保留集上的拟合情况。总体上，ElasticNet 对大多数省份的预测误差控制在较低水平，但对新疆、广西、黑龙江、海南等具有明显结构异质性的地区仍存在一定高估现象。这也提示未来若进一步扩展样本期或补充产业结构、科研平台、技术市场成交额等变量，预测精度仍有提升空间。

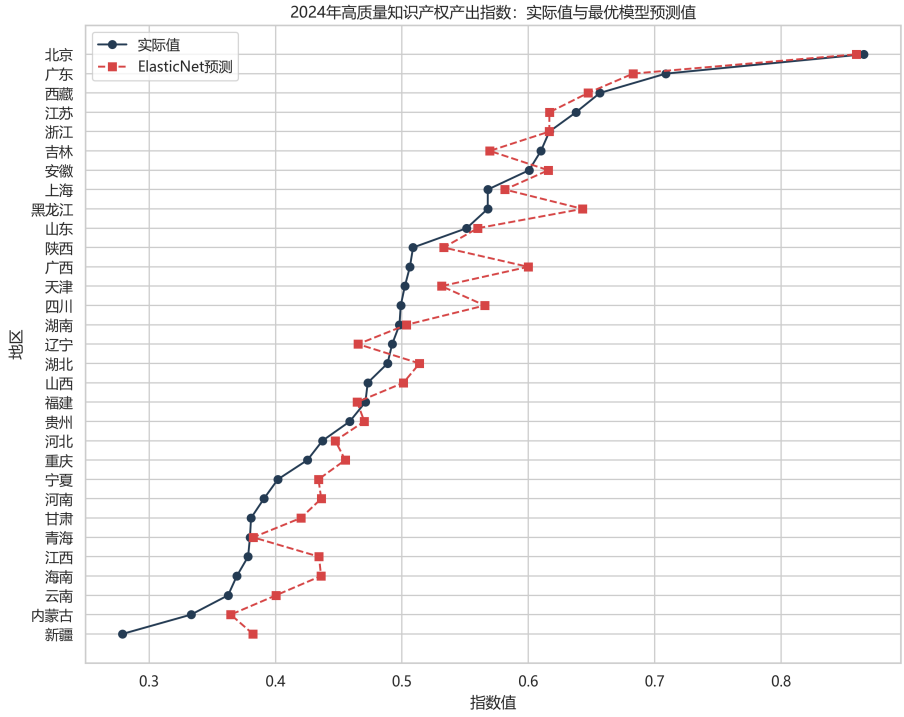

图 16　2024 年最优预测模型拟合效果

从误差结构看，ElasticNet 在 2024 年样本上的平均绝对误差为 0.0319，但

误差分布并不完全对称。表8显示，绝对误差最大的省份依次是新疆、广西、黑龙江、海南、四川、江西、河南和吉林。其中，新疆、广西、黑龙江、海南和四川均表现为模型高估，吉林则属于模型低估最明显的地区。

表 8　2024 年预测误差绝对值较大的省份

地区	实际值	预测值	绝对误差	误差方向
新疆	0.2789	0.3821	0.1032	高估
广西	0.5063	0.5999	0.0936	高估
黑龙江	0.5679	0.6429	0.0750	高估
海南	0.3694	0.4360	0.0666	高估
四川	0.4991	0.5656	0.0665	高估
江西	0.3783	0.4344	0.0561	高估
河南	0.3908	0.4362	0.0454	高估
吉林	0.6100	0.5695	0.0405	低估

图17进一步给出了残差分布。可以看到，残差并未围绕零点形成完全随机、均匀的云团，而是在若干预测值区间表现出轻微的系统性高估，这与前述省份误差表一致。造成这一现象的原因，很可能是当前特征集更擅长刻画“投入—效率—质量”的平均规律，但对少数地区的制度环境、产业结构和资源禀赋冲击反应不足。

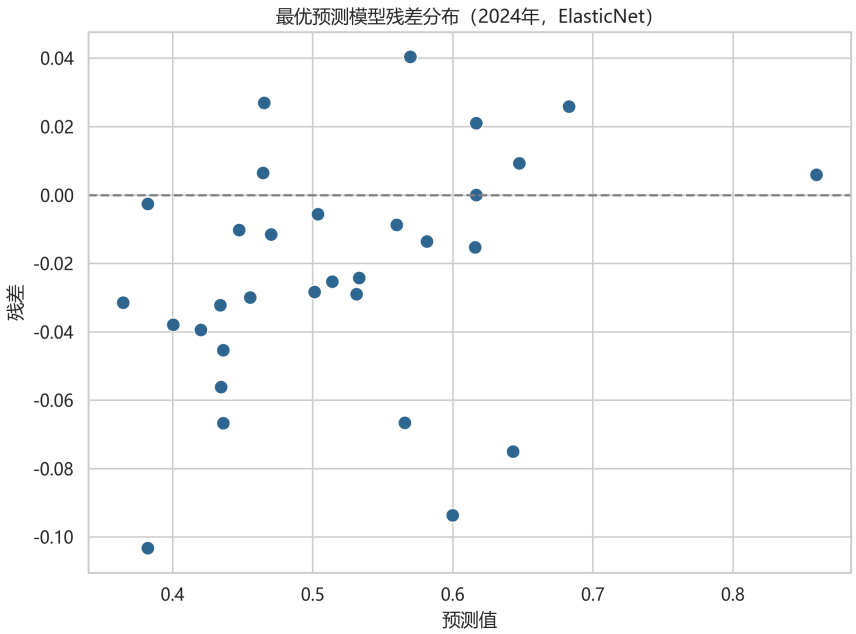

图 17　2024 年最优预测模型残差分布

为了进一步检验结论是否具有跨模型一致性，本文还把随机森林的变量重要性与 ElasticNet 的标准化系数做交叉比较。结果表明，随机森林同样把滞后 $\ln(RDExp)$、滞后 R&D 强度和企业有效专利占比识别为最重要的输入变量；

ElasticNet 则进一步突出滞后高质量产出指数和单位投入效率的作用。这种跨模型的一致性说明，本文得到的核心判断并不是单一方法的偶然产物，而具有一定稳健性。

5 结果讨论与政策含义

5.1 投入规模、主体结构与单位效率的关系重估

综合前文图表与模型结果，省域创新转化至少包含投入规模、主体结构和单位效率三个层面。投入规模决定创新活动能否持续展开，主体结构决定研发资源能否沉淀为发明专利存量，单位效率则直接影响同等研发经费最终形成多少高质量成果。本文的横截面结果表明，真正拉开地区差距的并不只是投入总量，而是这三者能否形成有效耦合。

这也意味着需要重新理解“创新强省”的含义。传统语境更强调经费、强度和专利总量，但在高质量知识产权框架下，更应关注资源、主体和效率能否同步改善。北京更接近“质量与规模同时领先”样本，广东、江苏和浙江体现出高位规模型特征，吉林、黑龙江、广西则更接近效率型样本。把这些地区简单归为“创新强”或“创新弱”，都难以解释其真实结构。

从统计关系上看，研发投入指数与高质量产出指数之间 0.697 的相关系数说明投入仍是必要基础，而投入指数与转化落差之间 -0.848 的强负相关则提示，规模扩张之后更容易出现边际收益下降和配置效率下滑。相对而言，单位研发有效专利量与高质量产出指数的 0.848 高相关系数说明，效率指标更能解释最终质量结果。换言之，规模回答的是“有没有能力做”，效率回答的则是“做得好不好”。

5.2 高投入地区边际转化偏弱的可能原因

广东、江苏等地区虽仍位于投入与产出的高位区间，但转化落差为负，这并不意味着创新能力减弱。其一，创新体系规模越大，常规创新活动越多，单位研发有效专利量、单位研发授权量等效率指标越容易被摊薄；其二，科研、龙头企业、中小企业、专利运营与产业化落地之间的链条更长，投入向产出的传导时滞更明显。固定效应模型中滞后 $\ln(RDExp)$ 系数为负、R&D 强度项显著性不强，

也与这种“高投入未必即时转化为当期高质量产出”的判断相吻合。

此外，高位地区的比较基准本身更高。对于广东、江苏这类已处于全国创新前沿的地区而言，下一阶段的目标不是从 0 到 1 建立创新体系，而是在高位平台上继续提高高价值发明专利的比例、法律稳定性和产业收益能力。因此，“负转化落差”更应被理解为高位地区已进入由规模扩张转向效率精修的新阶段，评价这类地区时也应更多关注结构改进而非短期绝对增量。

5.3 低投入高转化地区的形成机制与持续性边界

西藏、吉林、黑龙江和广西等低投入高转化地区的存在，说明创新质量并不完全取决于投入总量，而可能受到历史积累、专利结构、组织方式和产业耦合模式的共同影响。部分地区可能依托既有工业基础、科研院所积累、特定产业技术沉淀或阶段性成果集中释放，在单位投入效率和专利结构上取得相对优势。尤其对东北地区而言，既有工业门类和科研院所体系仍可能支撑较高的发明专利存量结构；对西藏等极端样本，则需要结合低基数放大效应来理解其单位指标高值。

但这种优势并不必然稳定。低基数容易放大单位指标，阶段性成果也未必能长期延续；若后续资源供给、平台支撑和产业承接不足，高质量专利未必能够持续转化为竞争力。因此，对这类地区既要总结其有效经验，也要警惕把短期高值误判为长期稳态。换言之，低投入高转化值得重视，但不宜被直接等同为长期、普遍和可无条件复制的优势。

5.4 时间演化与指标权重的补充解释

从时间演化看，2020—2024 年的省域创新转化经历了由投入扩张到质量抬升、再到结构分化加深的过程。授权率和企业有效发明专利占比总体上升，但到 2024 年，单纯依靠授权率提升来拉动综合质量指数的空间已明显收窄，地区差异更多体现为存量结构、主体组织和单位效率的差别。也就是说，越到后期，区域比较的关键越不在“有没有增长”，而在“增长主要来自什么”。

熵权结果也说明，综合指标更看重发明专利存量占比和单位研发有效专利量，而不是单纯奖励专利总量或企业占比。正因如此，一些投入规模不占优的地区仍可能凭借结构质量和单位效率取得更高排序，高投入地区若未同步改善效

率和结构，则会出现超额转化优势不足。采用“熵权法 + 转化落差”的意义，也正在于把总量叙事下不易识别的结构差异尽可能显化出来，这也是本文不直接以研发经费或专利数量进行省域排序的重要原因。

6 研究结论与政策建议

6.1 主要结论

基于中国 31 个省级地区 2020—2024 年的官方面板数据，本文构建研发投入指数、高质量知识产权产出指数和转化落差指标，发现我国省域创新体系总体呈现由数量扩张向质量提升过渡的趋势，但地区间仍存在明显错位，研发资源的规模优势并不必然转化为高质量知识产权优势。

区域识别结果显示，低投入高转化与高投入边际转化偏弱并存；机制解释表明，投入规模并非唯一决定因素，企业主体结构和单位效率同样关键；预测比较则表明，在短面板和小样本条件下，ElasticNet 比更复杂的树模型更稳定，更适合作为本研究的主预测工具。北京体现出质量与规模并进特征，广东、江苏更接近高位规模型，吉林、黑龙江、广西则表现出明显的效率型特征。这说明将区域识别、机制解释与预测比较放入同一分析链条，能够更完整地刻画省域创新转化的结构差异。

6.2 政策建议

政策上，应将高投入地区的重点从继续扩张经费转向优化投入结构，强化基础研究、应用研究、专利培育与成果转化之间的衔接，提高高价值专利沉淀效率。

对吉林、黑龙江、广西等转化落差为正的地区，应系统总结其在单位研发效率、企业专利组织方式和区域协同上的有效做法，同时补足平台、人才和产业承接能力，防止优势难以持续。

同时，应进一步强化企业在高质量知识产权形成中的核心地位，重视高价值专利组合培育、知识产权运营和成果转化金融支持，使专利数量、技术先进性和产业收益能够协同提升。

最后，应建立常态化监测框架，持续跟踪 RD_{it}、HQ_{it}、Gap_{it} 及企业专利结构等指标，据此对“高投入低转化”和“低投入高转化”地区实施差异化治理。

参考文献

[1] 国家统计局, 科学技术部, 财政部. 2022 年全国科技经费投入统计公报 [EB/OL]. (2023-09-18). https://www.stats.gov.cn/sj/zxfb/202309/t20230918_1942920.html.

[2] 国家统计局, 科学技术部, 财政部. 2023 年全国科技经费投入统计公报 [EB/OL]. (2024-10-02). https://www.stats.gov.cn/sj/zxfb/202410/t20241002_1956810.html.

[3] 国家统计局, 科学技术部, 财政部. 2024 年全国科技经费投入统计公报 [EB/OL]. (2025-09-29). https://www.stats.gov.cn/sj/zxfb/202509/t20250929_1961429.html.

[4] 国家知识产权局. 2024 年知识产权统计年报汇编 [EB/OL]. https://www.cnipa.gov.cn/tjxx/jianbao/year2024/indexy.html.

[5] 陈晓斌, 冯雅萱. 政府研发支出是否有利于撬动中小企业创新绩效——基于工业行业企业面板数据的非线性门槛效应检验 [J]. 统计研究, 2023, 40(10): 57-68.

[6] 杨名彦, 浦正宁. 我国省际数字技术创新水平测算及区域差异研究 [J]. 统计研究, 2024, 41(2): 15-28.

[7] 王帮娟, 王涛, 刘承良. 中国技术转移枢纽及其网络腹地的时空演化 [J]. 地理学报, 2023, 78(2): 293-314.

[8] 张林, 陈梓慕. 基于 MS-VAR 模型的中国创新质量演化阶段性与区域异质性研究 [J]. 区域经济评论, 2024(3): 52-60.

[9] 彭小宝, 陈文清. 我国高价值发明专利界定标准研究 [J]. 科技与法律 (中英文), 2021(6): 58-64.

[10] Higham K W, de Rassenfosse G, Jaffe A B. Patent quality: Towards a systematic framework for analysis and measurement[J]. Research Policy, 2021, 50(4): 104215.

[11] Wu H, Lin J, Wu H-M. Investigating the real effect of China's patent surge: New evidence from firm-level patent quality data[J]. Journal of Economic Behavior & Organization, 2022, 204: 422-442.

[12] Wang X, Fan L-W, Zhang H. Policies for enhancing patent quality: Evidence from renewable energy technology in China[J]. Energy Policy, 2023, 180: 113660.

[13] Jiang R, Hu S, Su Z, Huang Y, Zhang H. Spatial and Temporal Evolution of Urban Patent Quality: Evidence From China Prefectural Cities[J]. SAGE Open, 2025, 15(1): 21582440251319925.

[14] Li C, Wang Z. Investigating the Impact of Innovation Policies and Innovation Environment on Regional Innovation Capacity in China[J]. Sustainability, 2024, 16(23): 10264.

[15] 刘莹, 王宏新. 研发投入强度对高新技术企业财务绩效的影响——基于时滞性和异质性视角 [J]. 辽宁工程技术大学学报 (社会科学版), 2023(5): 365-372.

[16] Chen H, Qian L, Gu H, Chen Q, Zheng K, Zhang Y, Sha D. Patent quality, R&D investment, and the profitability of technology-based firms[J]. Finance Research Letters, 2025, 76: 106923.

[17] Zou H, Hastie T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(2): 301-320.

[18] 国家统计局, 科学技术部, 财政部. 2021 年全国科技经费投入统计公报 [EB/OL]. (2022-08-31). https://www.stats.gov.cn/xxgk/sjfb/zxfb2020/202208/t20220831_1887783.html.

附录

研究方案说明。本研究属于基于公开统计资料的省域比较与统计建模分析，未设置实地走访、问卷调查和访谈环节，因此不存在调查问卷、访谈提纲和回收样本等附加材料。整体研究流程可概括为“数据采集—口径统一—指标构建—区域识别—机制解释—预测检验”六个步骤，重点在于利用统一口径的省级面板数据识别研发投入向高质量知识产权产出的转化差异。**数据处理与软件说明。**在数据处理阶段，本文先对地区名称、年度口径和指标口径进行一致化整理，再对少量缺失项做核对与补齐；在指数测度中使用极差标准化和熵权法，在聚类分析中对输入变量做标准化处理，在固定效应建模中采用滞后一期设定并使用按地区聚类的稳健标准误，在预测比较中采用按年份滚动验证并保留 2024 年样本作为独立测试集。统计分析和图表绘制主要在 Python 环境下完成，使用的工具包括 pandas、numpy、matplotlib、seaborn、statsmodels、scikit-learn、optuna、xgboost 和 catboost。

本文主要简称与含义对照表

全称	简称
高质量知识产权产出指数	HQ_{it}
研发投入指数	RD_{it}
转化落差指标	Gap_{it}
滞后研发经费	$RDExp_{i,t-1}$
滞后 R&D 强度	$Intensity_{i,t-1}$
滞后企业发明专利申请占比	$EntApp_{i,t-1}$
滞后企业有效发明专利占比	$EntStock_{i,t-1}$

致谢

在本次大赛论文完成过程中，指导教师在选题把握、建模思路、结果解释和文字规范等方面给予了耐心指导；校内外老师、同学在数据核对、图表修改和文稿完善过程中提供了宝贵建议；国家统计局、国家知识产权局等部门公开发布的连续统计资料，为本文开展省域比较和模型检验提供了可靠基础。在此谨致诚挚谢意。