数量经济技术经济研究
Journal of Quantitative & Technological Economics
ISSN 1000-3894,CN 11-1087/F

《数量经济技术经济研究》网络首发论文

题目：创新知识溢出的测度与检验——基于机器学习生成专利文本相似度的证据
作者：龙小宁，张帆，易巍
DOI：10.13653/j.cnki.jqte.20260204.008
网络首发日期：2026-02-04
引用格式：龙小宁，张帆，易巍．创新知识溢出的测度与检验——基于机器学习生成专利文本相似度的证据[J/OL]．数量经济技术经济研究.
https://doi.org/10.13653/j.cnki.jqte.20260204.008

DOI：10.13653/j.cnki.jqte.20260204.008 网络首发时间：2026-02-04 18:44:15 网络首发地址：https://link.cnki.net/urlid/11.1087.F.20260204.1449.016

创新知识溢出的测度与检验

——基于机器学习生成专利文本相似度的证据

龙小宁 张 帆 易 巍*

摘要：知识溢出是科技创新发挥辐射带动作用，推动产业结构转型，建设现代化产业体系的关键。然而，知识溢出的测度难题长期存在，成为学界研究焦点。本文首先详细梳理了中国的专利制度安排，指出了现有研究中使用专利引文测度知识溢出所面临的挑战及其原因。为破解知识溢出的测度难题，本文基于机器学习模型，利用专利全文文本生成专利文本向量，并构建专利文本相似度指标。在此基础上，本文使用1985~2023年中国发明专利样本，结合知识产生的内在逻辑，提出了利用专利文本相似度衡量知识溢出的思路，并构造了表征城市间知识溢出的新指标。最后，借助高铁开通促进知识溢出的研究共识，本文运用双重差分方法对新指标的有效性进行了严格的实证检验，结果表明，专利文本相似度作为知识溢出衡量指标是合理有效的。本研究为深入理解中国知识溢出现状提供了新的测度工具和实证经验，有助于创新驱动发展战略的有效实施。

关键词：知识溢出 专利文本相似度 专利引用 高速铁路

一、引 言

创新是经济持续增长的根本动力，而知识溢出不仅是创新发挥作用的主要途径，同时也是未来创新的重要源泉。随着中国经济从高速增长阶段转向高质量发

* 龙小宁，教授，厦门大学知识产权研究院、厦门大学一带一路研究院，电子邮箱：cxlong@xmu.edu.cn；张帆（通讯作者），讲师，浙江万里学院物流与电子商务学院，电子邮箱：zhangfan_evan@163.com；易巍，副教授，集美大学财经学院，电子邮箱：susanyiwei@163.com。本文获得国家自然科学基金青年项目（72203074）、国家自然科学基金面上项目（72573144）和福建省自然科学基金青年项目（2022J05166）的资助。本文未使用AI。感谢匿名审稿专家的宝贵意见，文责自负。

展阶段,技术创新日益成为推动经济高质量发展的决定性因素,知识溢出的重要性日益凸显。特别地,近年来无论中国,还是发达经济体都经历着全要素生产率下降的困境,而其中一个共识是,知识溢出速度放缓是造成全要素生产率下降的重要原因(董昀等,2020)。党的二十届四中全会提出,“加快高水平科技自立自强,引领发展新质生产力”,“提升国家创新体系整体效能”。知识溢出的高效畅通有助于破除跨地区、行业和组织间的技术壁垒,实现创新要素的优化配置,通过降低全社会的研发边际成本,显著提升全要素生产率,从而为新质生产力的涌现提供源源不断的动力。因此,在科技创新引领现代化产业体系建设的背景下,知识溢出领域相关问题的研究更加值得重视。

鉴于知识溢出在推动经济高质量发展中的关键作用,学术界对其理论机制和实证效应进行了深入探讨。知识溢出作为推动经济增长的重要机制,其理论基础可追溯至 Arrow(1962)关于“干中学”的开创性研究和 Romer(1986)的内生增长理论。按溢出双方是否属于同一产业,知识溢出可分为 MAR 溢出和 Jacobs 溢出,前者强调产业内专业化知识积累对经济增长的作用(Marshall,1890;Arrow,1962;Romer,1986),后者则认为跨产业知识交叉融合对地区经济增长的影响更为显著(Jacobs,1969)。围绕两类溢出对经济增长的相对重要性,学界进行了大量实证检验,但结论并未达成一致(Glaeser 等,1992;Henderson,2003)。

上述实证结果的分歧反映了知识溢出研究面临的两大挑战:一是知识溢出的发生机制;二是如何识别和测度知识溢出(Audretsch 和 Feldman,2004)。鉴于知识溢出发生时没有明显的痕迹,其测度成为研究知识溢出活动与效应的最大障碍,进而导致知识溢出产生过程难以被清晰揭示(赵勇和白永秀,2009)。Jaffe 等(1993)提出专利引文可刻画知识流动轨迹,并基于该数据测度了美国创新主体间的知识溢出,使测度困境得以缓解。该研究范式在后续研究中得到广泛应用,推动知识溢出成为创新领域的热门话题(Peri,2005)。近年来,随着中国专利引文数据的可获得性增强,基于该数据刻画中国知识溢出活动的研究正在兴起(刘修岩和王峤,2022;易巍和龙小宁,2023;卢福财等,2024;陈骁等,2025)。然而,专利引文作为知识溢出的衡量指标一直存在争议,被认为是含有“噪音”的度量(Jaffe 等,2000;Jaffe 和 De Rassenfosse,2017)。针对美国专利引文数据的研究已指出其存在审查员引用、申请人策略性引用等局限性(Sampat,2010;Lampe,2012;Kuhn 等,2020;Schuster 和 Valentine,2022)。鉴于各国专利制度差异较大,专利引文在中国的适用性如何,有待进一步研究。

本文对中国专利法律和行政法规进行了系统梳理,发现中国专利引文数据在测度知识溢出方面面临更严峻的挑战。除了申请人披露现有技术的义务可能和申请人渴望获取专利权的利益存在冲突(杨德桥,2019),从而导致策略性引用问题之外,更关键的是,中国专利引文数据主要源于审查员检索报告。中国专利法虽规定

申请人负有提供审查所需信息资料的义务，但并未规定不履行该义务的法律后果，导致专利引文以审查员引用为主。然而，只有申请人引用才能反映发明人在创新过程中的知识学习与启发，而审查员引用无法体现真正的知识流动。因此，使用专利引文数据衡量中国知识溢出的有效性明显不足，亟须构建更科学的测度指标。

近年来，随着自然语言处理技术的发展成熟，基于专利文本相似度测度知识溢出的研究范式逐渐兴起。研究者通过计算专利文本间的相似性来追踪知识流动（Younge 和 Kuhn，2016；Feng，2020），其理论基础在于新知识的产生依赖已有知识（Kogut 和 Zander，1992、1996；Jaffe 等，1993）。如果一个专利的公开时间早于另一专利的申请时间，且二者文本相似度较高，则可推断发生了知识溢出。该方法直接基于专利文本内容，无须依赖专利引文数据，从而规避了引文数据的固有缺陷。然而，该方法目前主要应用于美国等发达国家，其在不同专利制度环境下的有效性尚未得到充分检验。因此，在专利引文数据存在明显不足的中国情境下，专利文本相似度方法能否有效测度知识溢出，成为亟待探索的关键问题。

本文围绕中国的知识溢出测度进行深入分析。基于国家知识产权局记录的1985~2023年全样本发明专利，借助谷歌专利团队依据专利全文文本，利用机器学习模型训练所得的专利文本向量，并使用余弦相似度方法计算专利文本相似度。在此基础上，将专利文本相似度数据汇总至城市层面，构建城市间知识溢出新指标，并借助高铁开通促进知识溢出的研究共识，对新指标的有效性进行了严格检验。实证结果表明，高铁直通促进了城市间的知识溢出，该结论在多重检验下保持稳健；机制分析进一步表明，专利知识能够通过学术会议和异地投资的方式向外传播，且高铁开通显著促进了这一知识外溢过程。上述结果充分证明本文基于专利文本相似度构建的知识溢出指标是合理有效的。

相较以往文献，本文的边际贡献如下：

第一，本文为知识溢出测度难题提供了数字时代的创新性解决方案。一方面，通过系统梳理既有研究以及中国专利法律、行政法规与部门规章，本文深入剖析了使用专利引文数据测度知识溢出所面临的制度性约束及其成因；另一方面，本文创新性地引入基于专利全文文本的机器学习语义向量，结合专利文本相似度方法和知识生产的底层逻辑，构建了衡量知识溢出的新指标，从根本上突破了专利引文数据缺陷引致的测度瓶颈。

第二，作为针对中国技术创新的基础性研究，本文对新指标的有效性进行了严格检验。借助高铁开通促进知识溢出的研究共识，本文在严谨的因果推断框架下系统验证了基于专利文本相似度构造的知识溢出指标的有效性和稳健性。上述工作为研究者提供了更科学的知识溢出测度工具，对于完善区域创新政策、优化创新资源配置、促进知识要素高效流动具有重要参考价值。

第三，本文在方法论应用方面实现了双重学术突破。近年来，谷歌专利文本向

量已成为全球创新经济学研究的前沿工具,但现有研究存在两大局限性(Higham等,2021;De Rassenfosse和Palangkaraya,2023;De Rassenfosse等,2024):一是研究对象局限于发达国家,已有文献大多基于美国等发达经济体数据,而不同国家的专利制度安排差异较大,谷歌专利文本向量的跨国适用性和有效性有待检验;二是应用范围局限于控制组构造,现有研究主要通过相似度计算识别与目标专利技术相近的对照样本,尚未被系统性地应用于知识溢出测度。针对上述研究缺口,本文实现了两个维度的开创性突破:一是在研究对象上,本文首次将谷歌专利文本向量应用于发展中国家的因果推断研究,基于中国的现实情境,系统检验了该工具的跨国有效性,并为中国创新经济学研究引入了新的分析工具和数据资源;二是在应用范围上,本文首次将谷歌专利文本向量系统性地应用于知识溢出测度,拓展了该前沿工具在创新经济学中的应用边界,为后续研究提供了可复制和可推广的研究范式。

二、知识溢出测度的挑战、原因与改进思路

专利引文作为知识流动衡量指标的争议长期存在,在中国情境下更面临独特的制度性挑战。本节首先梳理知识溢出的传统测度方法,深入剖析中国专利引文数据存在的制度性约束及其成因,进而提出基于专利文本相似度的改进方案,并阐述其有效性检验思路。

(一)知识溢出测度与专利引文数据

知识溢出作为技术扩散的重要方式,是指知识从创造者向其他主体扩散的过程(Agarwal等,2010),知识接收者无须支付或仅支付远低于知识价值的报酬(Caniëls,2000)。知识溢出具有显著的外部性,在微观层面,非竞争性特征使得技术、管理经验等知识发生流动与扩散,使其他主体获益(Griliches,1998);在宏观层面,知识溢出被视为创新推动经济增长的核心机制(Jaffe,2022),对区域或国家的经济增长发挥重要作用。

囿于知识溢出的重要性,研究者一直尝试量化知识溢出的程度和影响。然而,知识扩散的无形性令其难以捕捉,知识溢出的测度成为一大挑战。针对这一难题,学者发展了多种测量方法,这些方法经历了从间接指标到直接指标的演进历程。在间接测量方法中,学者主要将进出口贸易(Keller,2002;孙洋等,2021)与外国直接投资(Haskel等,2007;李盛楠等,2021)作为知识流动的代理指标。尽管这些方法在一定程度上能够表征知识溢出发生的可能性,但无法直接反映知识的具体内容和传播路径。随着对知识溢出机制理解的不断深化,近年来学者越来越多地采用能够直接体现知识存在的指标,如学术论文合作和专利合作(梁琦等,2019;Dong等,2020;Hanley等,2022)。然而,合作网络方法虽然能够通过合作频次来刻画知识溢出的强度,却难以准确捕捉知识溢出的方向性。在直接测量方法的发展过程中,Jaffe等(1993)的开创性研究为知识溢出测度提供了重要突破。该研究指出,专

利引文作为知识流动的书面记录，能够有效捕捉知识流动的轨迹，并首次运用专利引文数据对美国创新主体间的知识溢出进行了定量分析，显著推进了知识溢出的测度方法。这一研究范式将知识溢出问题推向研究前沿，并在后续研究中得到广泛使用（Peri，2005；刘修岩和王峤，2022）。

（二）专利引文数据的问题及其成因

专利引文作为知识流动衡量指标长期存在争议，核心问题在于其被视为含有“噪音”的度量（Jaffe等，2000；Jaffe和De Rassenfosse，2017）。

第一，专利引文的产生涉及申请人、专利代理师和专利审查员等主体（Meyer，2000）。根据知识溢出的定义，只有申请人添加的引文才能反映发明人之间的知识流动。然而，研究者在使用专利引文数据时，往往将其视同学术论文参考文献，认为所有引文均反映发明人在创新过程中接触到的专利知识。但一项针对专利发明人的调查发现，仅不到四成的受访者在发明前或过程中了解了所引用的专利，更多发明人是在发明完成后才了解，甚至并不了解所引用的专利（Jaffe等，2000）。此外，专利代理师或审查员添加的引文更无法反映发明人间的实际知识流动，进一步削弱了基于专利引文的知识溢出指标的有效性。随着多个发达经济体的专利引文开始区分发明人引用和审查员引用（Criscuolo和Verspagen，2008；Alcácer等，2009），更直接的证据出现，例如，一项针对美国专利引文的研究显示，在2005年到2014年申请的专利中，约四分之一的引文是由专利审查员而非发明人添加（Kuhn等，2020）。

第二，专利引文存在策略性引用问题。一方面，专利申请人可能故意不引用相关专利。Sampat（2010）的研究发现，当专利位于技术复杂度低的领域时，申请人会更多地检索和引用先前技术；在技术复杂度高的领域，申请人对现有技术的检索意愿较低。进一步的研究表明，申请人平均隐瞒了21%至33%的相关引文（Lampe，2012）。另一方面，专利申请人可能引用大量不相关专利，以增加审查难度，从而“隐藏”关键相关专利，提高授权概率（Schuster和Valentine，2022）。Kuhn等（2020）的研究发现，约5%的专利平均每件引用超过100个专利，接近全部专利引文量的一半，而这些引文与施引专利的相关性低。策略性引用问题也削弱了基于专利引文衡量知识溢出的有效性。

以上研究主要基于发达国家，尤其是美国的专利现状，而不同国家专利制度的巨大差异意味着上述现象可能在中国并不适用。通过对中国专利法律和行政法规的梳理，本文发现在中国情境下，使用专利引文衡量知识溢出的问题更为突出，具体表现为以下几点。

1.中国专利申请人的现有技术披露义务虚置化

中国专利制度在引用信息披露方面呈现独特的制度特征。根据《中华人民共和国专利法》（以下简称《专利法》）第三十六条第一款的规定，发明专利的申请人应当提交在申请日前与其发明有关的参考资料，但是该条未规定不履行现有技术披

露义务的后果,缺乏法律强制约束力(殷戈等,2024)。而且《专利法》的其他条文、《中华人民共和国专利法实施细则》(以下简称《专利法实施细则》)以及《专利审查指南》均未有相关补充规定。这导致举证责任制度在专利诉讼和其他程序中难以发挥有效作用。举证责任制度的关键在于负举证责任的一方主体若无法提供证据证明其主张,则需要承担相应的不利后果,然而中国现行专利制度未规定违法技术披露义务的法律后果,缺少对于申请人证明责任的规定(黄国群和朱然然, 2023)。相较之下,美国的专利实施细则37 CFR 1.56完整地规定了承担现有技术披露的主体、具体内容和违反该义务的法律后果。

此外,根据《专利法实施细则》关于专利引证的立法表述变化,申请人披露现有技术义务存在弱化的倾向。具体而言,在1985年和1992年的《专利法实施细则》中,现有技术披露义务被表述为"就申请人所知……并且引证",而2001年、2002年、2010年和2023年的文本表述则被修改为"有可能的,并引证"。虽然立法语言相似,但是前者规定的是对现有技术的必然引证,而后者则是可能引证,反映了申请人披露现有技术义务的弱化(曾清华,2023)。因此,无论是法律还是行政法规,均对专利申请人披露现有技术的约束性不足,使该义务呈现虚置化状态(杨德桥,2019),专利引文难以有效表征申请人之间的知识流动。

2.中国专利引文数据主要记录的是审查员引用

根据中国国家知识产权局的释义,专利引文由引用参考文献和审查对比文件组成。其中,引用参考文献是由专利申请人提供,而审查对比文件则是由专利审查员提供。世界知识产权组织颁布的《ST.9关于及有关专利和补充保护证书的著录项目数据的建议(2013)》和《ST.14在专利文献中列入引证参考文献的建议(2016)》规范着各成员国的专利引文,其中建议专利文献扉页著录项(56)记录"专利引证参考文献清单"。由于申请人现有技术披露义务的制度缺陷,中国专利的引用信息主要来源于专利审查员的检索工作,其实质是核查专利创新性要求的"对比文件"(米晋宏和张南,2022)。因此,具有可得性和可比性的中国专利引文数据主要为审查员引用,反映的是审查员基于技术判断的客观评价,而非发明人的主观引用行为。相比之下,美国专利和商标局从2001年开始在著录项(56)同时列示引用参考文献和审查对比文件,使得申请人引用可被清晰识别。因此,以审查员引用为主的中国专利引文难以有效反映申请人之间真实的知识传播。

3.中国专利引文数据只包含授权发明专利的引证信息

国家知识产权局没有向公众公开审查员检索报告的法律义务,仅将已授权发明专利的审查员引证信息记录在授权版本的专利文献扉页著录项(56)栏目中。因此,如果一项发明专利未被授权,即无法观察到其引证专利。根据智慧芽专利分析系统提供的1985年到2020年中国发明专利数据,该期间内共有13149246件发明申请专利,其中5741935件被授权,授权率约44%。这意味着超过一半的发明专利未

能记载引文信息，如此规模的信息缺失导致专利引文难以全面反映申请人之间的知识流动。

（三）知识溢出测度的改进思路

如前文所述，以专利引文为代表的直接测量方法已成为知识溢出研究的主流。然而，专利引文虽在理论上是有效的知识流动追踪器，但在中国情境下因制度性障碍导致测度有效性不足。相比之下，基于专利文本相似度的方法能够有效规避上述障碍。专利文本直接由发明人撰写，记载着发明人的技术思路、创新逻辑和知识结构，无论专利最终是否获得授权，其技术内容都得以完整保存。因此，通过专利文本相似度能够更直接、更全面地捕捉发明人之间的知识关联性。近年来，随着自然语言处理技术的发展成熟，不断有文献尝试利用文本内容，并结合文本分析方法对溢出效应进行评估（Myers 和 Lanahan，2022）。与本文联系更为直接的进展是，研究者开始通过专利文本间相似性来追踪知识流动（Younge 和 Kuhn，2016；Feng，2020）。尽管专利文本相似度方法衡量知识溢出在国际学术界已有应用，但在目前的公开文献中，尚未系统地用于中国的知识溢出测度。国内学者虽已将专利文本相似度分析引入创新经济学研究，但主要集中在技术关联识别、创新质量评估等方面（孙震等，2025；张同斌等，2024），对其作为知识溢出测度指标的理论基础和实证有效性的系统探讨仍显不足。

第一，专利文本相似度测度知识溢出具有坚实的理论基础。根据知识溢出理论，知识溢出本质上是一种学习过程（Griliches，1992），当产生于某个主体的知识被另一个主体捕捉并吸收时，即认为两者间发生了知识溢出（Peri，2005）。作为创新发明成果，专利的产生意味着新知识的创造（Romer，1990；Jaffe 等，1993），专利文本则是创新主体知识结构和技术思路的重要载体（Ramani 等，2008）。从知识创造的内在规律来看，新知识的产生依赖现有信息、知识和思想（Kogut 和 Zander，1992、1996），完全独立于已有知识体系进行知识创造的可能性较低。认知邻近性理论进一步揭示了知识溢出的深层机制，知识并非自动从一个主体“溢出”到另一个主体，而需要接收方具备足够的认知基础来理解和吸收外部知识（Cohen 和 Levinthal，1990）。这种认知基础体现为共享的知识基础、相似的专业语言和概念框架。基于上述理论逻辑可以推断，当知识溢出发生时，接收方会将吸收的外部知识融入自身知识体系，这一过程必然在其专利文本中留下痕迹，表现为与知识源在技术概念、解决方案逻辑等方面的相似性。因此，专利文本相似度能够有效反映知识溢出的发生。具体而言，如果两个专利满足以下条件：一是一个专利的公开时间早于另一专利的申请时间，二是两个专利的文本相似度较高，那么可以推断从已公开专利到新申请专利发生了知识溢出的可能性较大。基于这一逻辑，结合专利文本和专利的时间特征来测度知识溢出。相较于使用专利引文的传统方法，新指标直接依据专利文本本身的表述，与专利引文无关，因而可以避免专利引文数据缺陷所带来的问题。

第二,作为一项探索性和基础性工作,如何检验新指标的有效性成为关键问题。现有研究表明,知识传播依赖面对面交流(易巍等,2021),因此,消除流动性障碍,降低人员流动成本,推动面对面交流的举措对于扩大知识溢出至关重要。基于这一认识,交通基础设施建设被认为是促进知识溢出的重要措施,且这一促进作用在多种交通方式中得到了实证支持(Agrawal 等,2017;Dong 等,2020)。具体到中国现实情境,大量文献表明高速铁路可以通过便利人员流动,增进面对面交流,显著促进了知识流动(Dong 等,2020;Hanley 等,2022;Long 和 Yi,2024)。鉴于高铁开通促进知识溢出已成为研究共识,我们可以借助这一共识来检验基于专利文本相似度构造的新指标的有效性。检验逻辑在于,如果新指标能够重现高速铁路对知识流动的促进作用,则可为新指标的有效性提供实证支持。

三、基于专利文本相似度的知识溢出测度指标

本节即以中国发明专利为样本,实施基于专利文本相似度来衡量知识溢出的实证分析方案。首先,简要介绍专利文本相似度的计算思路,包括专利文本向量化与相似度计算方法两大关键要素;其次,围绕专利文本获取及向量化面临的挑战,阐述引入谷歌专利文本向量的原因,并以中国发明专利为样本,检验基于谷歌专利文本向量计算的专利文本相似度的有效性;最后,提出使用专利文本相似度衡量知识溢出的建议。

(一)专利文本相似度计算思路

专利文本相似度指的是两件专利所对应技术方案的相似程度。计算专利文本相似度需要两个步骤:一是要将专利文本转换为计算机可以理解和处理的数字表示形式,即专利文本向量化,所得结果称为专利文本向量;二是基于专利文本向量计算专利文本相似度,从而比较不同专利所对应技术方案的相似程度。常用的相似度计算方法有余弦相似度、欧氏距离、曼哈顿距离和切比雪夫距离等,其中余弦相似度解释力较强,是目前最常用的方法(程新生和王向前,2023;欧阳志刚和胡雯华,2024)。

(二)专利文本向量的选择依据

专利文本向量的生成依赖于全面完整的专利文本以及基于海量专利文本训练的向量化模型,然而二者均难以实现。首先,获取全面完整的专利文本信息难度大。现有研究往往使用专利摘要,甚至仅使用专利标题作为专利文本的表征(陈强远等,2022;黄先海等,2023;林建浩等,2025)。尽管专利摘要是全文的概括性总结,包含技术领域、技术问题及其解决方案等信息(程新生和王向前,2023),但是在与专利代理律师访谈的过程中,获知专利摘要所传达的技术信息十分有限①。实际上,专利的文本信息主要体现在标题、摘要、权利要求书和说明书四个部分,其中说明书的信息最

① 《专利审查指南》对专利摘要部分的撰写严格,明确规定:"文字部分(包括标点符号)不得超过300个字。摘要超过300个字的,可以通知申请人删节或者由审查员删节。"《专利审查指南》是专利法及其实施细则的具体化,是国家知识产权局和专利复审委员会依法行政的依据和标准。

为全面[①]。然而，由于专利文件数量数以百万计，逐一人工提取不具有可行性。其次，生成高质量专利文本向量的另一大挑战在于向量化模型的训练。将专利文本转换为高质量向量需要基于数以百万计专利语料训练的机器学习模型，该模型的训练需要海量算力，对研究者的资金和技术均提出了挑战（Higham等，2021；Kelly等，2021）。

针对上述挑战，谷歌专利团队基于海量专利全文文本，利用机器学习模型生成了专利文本向量，并将其公开供科学研究使用[②]。谷歌专利文本向量在现有研究中得到了广泛应用。例如，Higham等（2021）在分析高质量专利特征时，发现即使将专利按照技术类别划分，类别内部的专利仍存在技术异质性。为了减少技术异质性带来的偏差，借助谷歌专利文本向量，通过计算专利文本相似度，将目标专利与语义相似的专利进行匹配，从而更科学准确地比较专利差异。同样，De Rassenfosse和Palangkaraya（2023）在研究专利开源对创新的影响时，利用该向量计算专利文本相似度，构造与开源专利技术相近但未开源的对照组样本，进而评估专利开源对创新活动的因果影响。

（三）专利文本相似度的有效性检验

为初步说明基于谷歌专利文本向量的相似度测度的有效性，即验证文本相似度更高的专利之间确实存在更多的技术信息共享，本文从原始文本、技术类别和专利引文三方面提供经验支持。研究样本为1985~2023年中国发明专利。

1.基于专利原始文本的检验

通过直接对比专利原始文本进行检验。从相似度接近于1的专利对中随机抽取样本深入分析。例如，发明专利CN107519921B和发明专利CN107519922B的相似度为0.998。附表1列出了这两件专利的标题和摘要文本，经仔细对比，可以发现两件专利原始文本间的差异微小，这验证了根据谷歌专利文本向量所得专利文本相似度的准确性[③]。

2.基于专利技术领域的检验

结合专利的技术类别信息展开分析。本文计算每件专利与两组对照样本的专利文本相似度均值，一是同技术领域内（CPC四位数）随机抽取的100件其他专利，二是不同技术领域随机抽取的100件其他专利。结果如附图1所示，目标专利与所在技术领域内其他专利的相似度均值明显高于与不限技术领域随机抽取的其他专利的相似度均值。这一结果表明，专利文本相似度能够有效表征技术信息的相似程度，从而进一步验证了相似度测度的有效性。

① 依据《中华人民共和国专利法》第二十六条规定："说明书应当对发明或者实用新型做出清楚、完整的说明，以所属技术领域的技术人员能够实现为准。"

② 关于谷歌专利文本向量的构建方法和技术细节，详见附录1。本文附录详见《数量经济技术经济研究》杂志网站，下同。

③ 为分析技术术语密度差异对专利文本相似度的影响，本文进行了稳健性检验，结果表明相似度计算在不同技术术语密度水平下保持稳健，详见附录2。

3.基于专利引文类型的检验

从专利引用关系的视角展开分析。虽然专利引文在衡量知识溢出方面存在不足,但在一定程度上也可表征与专利密切相关的技术信息,因此,可作为验证相似度有效性的参照。本文计算每件专利与其引用专利的专利文本相似度,并提出两个研究预期:一是目标专利与引用专利的相似度均值应显著超过与同技术领域其他专利的均值,这是因为引用专利通常代表与目标专利技术关联更密切的专利;二是目标专利与自引用专利的相似度均值应明显高于与他引用专利的相似度均值,这是因为自引用代表发明人引用自己先前的专利,技术连贯性更强;他引用则是指向其他发明人的专利引用,技术关联相对较弱。附图1和附表2的结果支持了上述两个预期,从而进一步验证了专利文本相似度的有效性。此外,专利文本相似度可以实现对自引用专利和非自引用专利的区分,由于自引用专利可以被视作是申请人自己添加的专利引用,因此这一发现在一定程度上缓解了专利引用无法区分申请人引用和审查员引用的问题。

(四)基于专利文本相似度的知识溢出测度建议和优势分析

知识溢出是指通过信息交换获得知识的过程(Caniëls,2000)。作为创新的重要载体,专利翔实记录了新产品生产、新工艺实施的关键知识和技术细节,因而成为知识传播的重要媒介。从知识创造的规律来看,知识的产生依赖于现有信息、知识和思想(Kogut和Zander,1992、1996),完全独立于已有知识体系进行知识创造的可能性较低。因此,先前的技术知识对于开发新知识是有用的,新专利的发明者往往会从已公开的专利中汲取信息(Jaffe等,1993)。基于上述逻辑,假设有两件专利,其中一件专利的公开时间早于另一专利的申请时间,同时两件专利的文本相似度较高,那么可以推断从已公开专利到新申请专利存在知识溢出的可能性较大。专利文本相似度可以在不同层级以均值形式汇总,从而衡量行业间、地区间的知识溢出(Younge和Kuhn,2016;Feng,2020)。

相较于基于专利引用的传统测度方法,基于文本相似度的知识溢出测度具有两方面的显著优势。一是测量范围更完整。传统引用指标仅能衡量已授权发明专利的影响力,因为未授权专利的引文信息无法观察。而文本相似度方法能够同时测量已授权和未授权专利的知识溢出效应。通过随机抽样检验发现,已授权专利与后续专利的平均文本相似度显著高于未授权专利,表明新指标能够有效区分不同质量专利的影响力。二是测量对象更准确。由于中国专利引用数据主要源于审查员检索,反映的是法律审查需求而非发明人的知识学习过程。相比之下,文本相似度直接基于技术内容的实际关联,能够更准确地反映发明人之间真实的知识传播①。

① 关于新方法在测量范围完整性和测量对象准确性两方面的优势,本文进行了详细论证,限于篇幅,具体内容参见附录3。

四、知识溢出指标构建与有效性检验：基于高铁开通的准自然试验

前文的分析已初步显示，利用谷歌专利文本向量计算所得的专利文本相似度可以有效衡量专利间的技术关联。然而，要将其作为知识溢出的测度指标，还应进一步验证其能否捕捉到真实的知识传播过程。理想的验证策略是找到一个已被证实能够促进知识溢出的外生冲击，观察基于专利文本相似度构造的新指标能否重现这一已知的因果关系及其作用机制。如果新指标能够复现已有文献的实证规律，则可为该指标作为知识溢出测度的有效性提供有力的实证支持。这正是本节实证分析的核心目的，即借助已知的因果关系来验证新测度指标的有效性，而非使用新指标来重新论证某一因果关系的存在。

在中国情境下，高速铁路开通为这一验证提供了理想的准自然实验场景。从理论机制来看，知识溢出受到地理距离的制约（易巍等，2021），特别是高度复杂的技术领域，其中涉及较多的隐性知识，更加依赖面对面交流（Agrawal 和 Goldfarb，2008）。近十五年来，中国高铁的飞速发展恰好作用于这一关键环节，通过压缩城市间的时空距离、降低人员旅行成本，显著提高了面对面交流的可能性（叶德珠等，2020；易巍等，2021）。相对于传统的交通运输方式，高铁开通不仅进一步提高了区域之间的地理通达性，其所具有的载客量大、速度快、准点率高、安全性好的优势，更加能够满足那些对于时间具有较强敏感性的高素质人才流动的需求（杜兴强和彭妙薇，2017）。此外，从实证识别的角度来看，高铁主要用于运送旅客的这一特征天然地将人员流动和产品流动分离开，为观察知识溢出提供了良好的识别情景。

基于上述理论机制和识别优势，大量研究已形成共识，中国高速铁路通过便利人员流动、增进面对面交流，显著促进了地区间的知识流动（Dong 等，2020；Hanley 等，2022；Long 和 Yi，2024）。这一稳健的研究共识为验证新指标的有效性提供了可靠的基准。如果基于专利文本相似度构造的知识溢出指标是有效的，那么它应当能够识别并重现高速铁路对知识溢出的促进作用。

基于这一验证逻辑，本节利用中国高铁系统的快速扩张作为城市间旅行时间的外生冲击，通过分析城市间知识溢出的规模变化和作用机制，来检验基于专利文本相似度所构造的知识溢出指标的有效性。具体而言，本节以在2002~2015年提出申请且申请人地址位于中国大陆的发明专利为样本。在前文所述新知识如何产生的逻辑基础上，依据谷歌专利文本向量和文本相似度方法刻画中国城市间知识溢出的规模；进而考察高铁系统扩张与知识溢出规模之间的实证关系是否与已有研究发现相吻合，并借此检验基于专利文本相似度构造的知识溢出指标的有效性。

（一）基于专利文本相似度的城市间知识溢出指标构建

如附图2所示，假设有两个城市 a 和 b，其中城市 a 有3个专利 p_1、p_2 和 p_3，城市 b

有2个专利q_1和q_2。此外,城市a所有专利的公开时间均早于城市b所有专利的申请时间。

结合知识产生的逻辑,城市a到城市b的知识溢出测度思路如式(1)所示:

$$knowledge\ spillover_{ab}=\frac{\sum_{i=1}^{3}\sum_{j=1}^{2}similarity\left(p_i,q_j\right)}{6} \tag{1}$$

其中,分子为城市a中已公开专利与城市b中新申请专利间的专利文本相似度之和,分母为两组专利两两组合形成的组合数量。对于分子部分,以$similarity(p_1,q_1)$为例,其表示城市a中已公开专利p_1与城市b中新申请专利q_1的专利文本相似度。

在构造城市间知识溢出指标的过程中,最小计算单元是专利对层级。由于专利数以百万计,计算需求巨大。以2002~2015年城市间知识溢出的测度为例,涉及的发明专利超过390万件,共需计算超过6万亿个专利对的相似度①。为此,本研究首先依据专利申请年和公开年信息,构造专利对数据集;其次使用Python软件中机器学习库Scikit-learn计算专利对间的余弦相似度②;最后将专利对层级的相似度数据按照专利申请人地址汇总至城市对层级,进而获得城市间的知识溢出程度。

(二)知识溢出的时空演变特征

为直观展示所构建指标的有效性与特征事实,本文采用核密度估计刻画跨地区知识溢出水平的分布特征,该方法在相似度指标的实证研究中已得到广泛应用(Bahar等,2014)。具体而言,首先以城市对为基本分析单位,计算每对城市间的专利文本相似度;其次根据知识来源地和流入地的区域属性对城市对进行分组,如"东部地区→其他地区"包含所有来源地位于东部、流入地位于中部或西部的城市对;最后对各组城市对的专利文本相似度值进行核密度估计,刻画不同区域知识溢出水平的分布特征。

图1显示,中国跨地区知识溢出格局呈现显著的收敛效应与均质化趋势。观察2002年的分布特征,各区域间存在明显的梯度差异。东部地区的核密度曲线峰值最高且略右偏,表明其作为知识溢出源的强度最大,且区域内部各省市的溢出能力具有较高的同质性。相比之下,中部和西部地区的曲线形态较为扁平,峰值明显左移。这不仅反映了当时中西部地区向外辐射知识的整体能力弱于东部,更揭示了这些区域内部存在显著的异质性,仅有少数中心城市具备较强的溢出能力,而广大外围地区的知识输出能力相对薄弱,区域内部发展呈现明显的不平衡特征。

然而,到2015年,分布形态发生了根本性的结构转变,呈现出高度的趋同性与

① 各年度专利对的数量分布信息参见附表4。

② 关于专利文本相似度阈值的确定,本研究基于现有文献经验和实证校准验证,详见附录4。需要说明的是,考虑到高铁开通对跨领域知识交流的促进作用,本研究关注整个相似度分布的系统性变化,而非仅聚焦高相似度专利对。

集聚性。一方面，纵轴的密度峰值从2002年的约12上升至25以上，三条曲线均演变为显著陡峭的“尖峰”形态，方差大幅缩减；另一方面，东、中、西三个地区曲线的中心位置几乎完全重叠，西部和中部地区的“长尾”特征基本消失。这些统计特征的变化共同表明，各地区知识溢出能力的离散程度显著降低，地理区位对知识溢出的制约作用显著减弱。西部地区不仅在平均水平上实现了对东部的追赶，其内部异质性也明显收敛，中国跨地区知识溢出能力实现了从梯度差异向全域均衡的转变。

总体而言，这一时序演变证实了中国跨地区知识溢出网络实现了从“中心—外围”等级结构向“全域协同”扁平化网络的转型。随着全国统一大市场的构建以及高铁、互联网等基础设施的全面覆盖，知识流动的空间摩擦显著降低，区域间知识溢出能力趋向均衡。

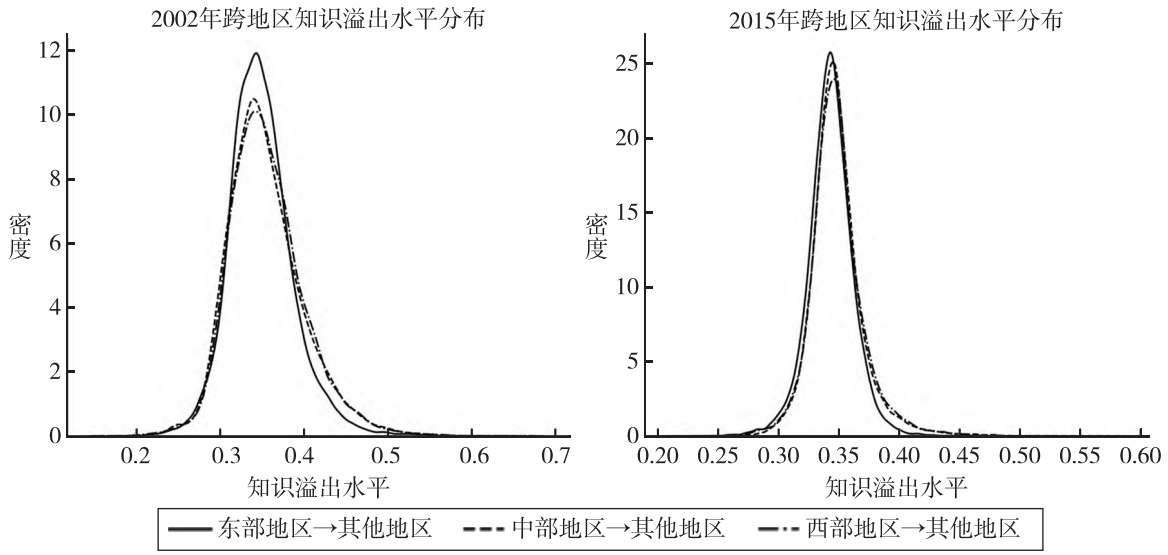

图1 2002年和2015年不同区域知识溢出水平分布

（三）模型设定

为了估计高铁直通对知识溢出的影响，本文构造“城市对—年份”层级的面板数据，并采用如下双重差分模型进行实证分析：

$$similarity_{ijt} = \beta HSR_{ijt} + \theta_{it} + \mu_{jt} + \varphi_{ij} + \lambda X_{ijt} + \epsilon_{ijt} \quad (2)$$

其中，被解释变量$similarity_{ijt}$表示的是第t年城市i到城市j的知识溢出，通过城市i内创新主体在t年之前3年内已公开的发明专利与城市j内创新主体在第t年新申请的发明专利的文本相似度均值衡量。核心解释变量高铁直通HSR_{ijt}为虚拟变量，若两地在第t年直通高铁，那么高铁开通后取1，否则为0。θ_{it}表示知识来源地区（已公开发明专利所在地）的城市—年份固定效应，μ_{jt}表示知识接收地区（新申请

发明专利所在地)的城市—年份固定效应,实现了对知识接收地和来源地随时间变化影响因素的控制,如地区经济发展水平、创新活力等。式(2)还加入了城市对层面的固定效应φ_{ij},对城市对层面不随时间变化的特征进行控制,如地理距离。此外,本文还控制了城市对随时间变化且可能对知识溢出产生影响的其他因素X_{ijt},比如城市i和城市j的互联网发展水平以及产业结构相似程度。ϵ_{ijt}为随机误差项,回归标准误均在城市对层面聚类。本文主要关注高铁直通HSR_{ijt}的回归系数β,若系数显著为正,表明高铁直通促进了城市间的知识溢出,说明本文使用专利文本相似度构造的知识溢出指标是合理有效的①。

(四)数据来源和变量构造

为了评估基于专利文本相似度计算得出的知识溢出指标的有效性,本文利用高铁开通促进知识溢出的研究共识,综合多方数据构造2002~2015年涵盖189个城市共12922个城市对层级的面板数据。

被解释变量。本文选取城市间知识溢出为研究对象。该数据源自国家知识产权局的专利著录项信息及谷歌专利团队提供的专利文本向量数据。具体构造步骤为:首先,计算新申请专利与三年窗口期内已公开专利的文本相似度;其次,将专利对的相似度信息与其基础数据相匹配;最后,将2002~2015年的专利对数据在"城市对—年份"层面汇总。

主要解释变量。为了深入分析高铁直通的影响,本研究借鉴了易巍等(2021)的做法,采用网络爬虫技术从《全国铁路旅客列车时刻表》、高铁网和114票务网中获取了6218个以C/D/G为首字母的车次信息,并通过维基百科获取了相关线路和站点的开通时间,截至2015年底,共有189个城市开通高铁,其中12922对城市通过高铁直接相连。

其他解释变量。为了控制城市对随时间变化且可能对知识溢出产生影响的其他因素,如城市间互联网发展水平和产业结构相似度,本文采用构造并控制相应变量的方式予以处理②。此外,本文还控制了城市间地理距离、城市专利数量、人口规模以及航空客运量。其中,城市间地理距离是通过计算两城市市政府之间的球面距离得出。关于城市专利数量,则根据国家知识产权局提供的专利信息,将发明专利数量按城市—年份层面汇总,并转换为自然对数形式。最后,城市人口数和航空客运量的数据均来源于《中国城市统计年鉴》,同样转化为自然对数形式。主要变量的描述性统计结果如附表3所示。

(五)基准回归

为考察高铁直通对城市间知识溢出的影响,并借此检验所构建的知识溢出指

① 限于篇幅,关于内生性问题的详细讨论和解决方案,参见附录5。

② 限于篇幅,关于城市间互联网发展水平和产业结构相似度的具体测度方法,参见附录6。

标的有效性，本文依据式（2）进行了回归分析，其结果详见表1[①]。本文将高速铁路的直接连通视为外生性冲击，进一步分析高铁直通后城市间知识溢出的规模变化。本文在列（1）至列（5）分别引入了不同的固定效应。具体而言，列（1）至列（2）逐步纳入年份固定效应和城市层面的固定效应，并对地理距离以及随时变化的城市特征变量进行控制；列（3）至列（5）则纳入城市对固定效应和城市—年份固定效应，以最大限度地控制潜在的不可观测因素。在五列回归结果中，高铁开通（*HSR*）的系数均显著为正，其中列（5）在控制内生性问题方面最为严格，提供了最为保守的估计值。为了便于对估计系数进行解释，本文借鉴了Bahar等（2014）的方法，对因变量即城市间知识溢出进行了标准化处理。在此基础上，解释变量的估计系数反映了当解释变量增加一个单位时，城市间知识溢出的程度围绕其均值发生的标准差变化量。据此，列（5）的回归结果意味着高铁直通使城市间的知识溢出增加了0.021个标准差。上述结果表明，高铁连接能够有效促进专利知识的跨区域传播，同时城市间距离的系数显著为负，表明距离的增加会阻碍知识流动，与知识溢出的本地化特征相吻合（Jaffe等，1993）。这些发现均说明本文基于专利文本相似度所构造的知识溢出指标是有效的[②]。

表1　　高铁直通与城市间知识溢出：基准回归

变量	（1）	（2）	（3）	（4）	（5）
HSR	0.022***	0.076***	0.078***	0.042***	0.021***
	（0.007）	（0.007）	（0.008）	（0.005）	（0.005）
控制变量	是	是	是	是	是
年份固定效应	是	是	是	否	否
新申请专利所在地固定效应	否	是	否	否	否
已公开专利所在地固定效应	否	是	否	否	否
新申请专利所在地—年份固定效应	否	否	否	是	是
已公开专利所在地—年份固定效应	否	否	否	是	是
城市对固定效应	否	否	是	否	是
调整R^2值	0.272	0.474	0.526	0.761	0.837
样本量	435663	435663	435663	441728	441540

注：*、**、***分别表示在10%、5%、1%的水平上显著，括号内为聚类至城市对层面的稳健标准误。

① 限于篇幅，表1仅报告核心解释变量的系数，控制变量系数和完整回归结果参见附表5。

② 为了验证基准回归结果的稳健性，本文开展了一系列检验，包括平行趋势检验、安慰剂检验、异质性处理效应检验以及工具变量法估计。限于篇幅，具体的检验结果与图表汇报于附录7。

(六)基于作用机制的检验

前文基于专利文本相似度构造的知识溢出指标成功识别出高铁直通对知识溢出的促进作用,与已有研究共识一致,初步验证了新指标的有效性。然而,有效的知识溢出测度指标,不仅应当识别出总体效应,还应捕捉到机制路径。若新指标能够识别出符合理论预期的知识溢出作用机制,将为其有效性提供更深层次的实证支持。因此,本节进一步检验高铁直通影响知识溢出的作用机制。

从高铁如何改变知识传播成本的视角出发,本文聚焦学术会议和异地投资两种关键的知识溢出渠道。一方面,高铁已成为支持学术界日益频繁交流与合作的重要基础设施(Dong 等,2020;Wang 和 Cai,2020)。特别是学术会议,其作为一种知识交流的重要平台,为学者提供了深入讨论、合作以及分享创新成果的机会,既促进了跨学科知识的融合,也为来自不同研究领域的学者提供了共同研讨和解决问题的平台,从而加速了知识传播(Chai 和 Freeman,2019)。因此,对于大学和研究机构而言,高铁通过便利学术交流,成为促进知识传播的重要推动力。另一方面,相较于学术机构,追求创新的企业有着不同的知识传播机制。企业间的竞争和合作都离不开新知识及新技术的应用和掌握。高铁等交通基础设施的发展为企业扩大其影响范围和到达较远投资地提供了便捷的途径,进而促进了知识流动。

1.学术会议

为了探讨高铁开通是否通过学术会议的渠道促进城市间知识溢出,本文收集了各地区相关会议的资料,并分析了高铁开通对学术会议举办强度的影响。回归模型如下:

$$meetings_{it} = \beta_0 HSR_{it} + X_{it} + \phi_i + \rho_t + \epsilon_{it} \quad (3)$$

其中,$meetings_{it}$ 表示地区 i 在年份 t 所有的大学或科研机构举办的与学术会议有关活动的强度,使用学术会议数量、参会人数和提交的论文数量衡量。HSR_{it} 表示地区 i 在年份 t 是否开通高铁,若开通则为 1,否则为 0。X_{it} 表示一系列控制变量,包括国民生产总值、人口规模、专利存量以及航空客运量。ϕ_i 和 ρ_t 分别表示地区固定效应和年份固定效应,ϵ_{it} 表示随机误差项。回归结果如附表 6 的子表 A 所示,高铁开通显著提升了地区学术会议有关活动的强度。

进一步地,本文检验学术会议作为高铁开通促进城市间知识溢出的渠道作用。设定基准回归模型如式(4)所示,其中左手侧变量为城市间知识溢出,dis_{ij} 为城市 i 和城市 j 间的地理距离,α_{jt} 是知识来源地区(已公开发明专利所在城市)—年份固定效应,ϕ_i 和 ρ_t 分别表示知识接收地区(新申请发明专利所在城市)固定效应和年份固定效应,其余变量设定同式(2)。

$$similarity_{ijt} = \beta_1 HSR_{ijt} + \lambda X_{ijt} + \delta dis_{ij} + \alpha_{jt} + \phi_i + \rho_t + \epsilon_{ijt} \quad (4)$$

然后，从广延边际的视角，在式（5）中纳入知识来源地区的学术会议变量$meetings_{it}$，形成回归模型式（5）。根据附表6子表B的（1）至列（4）回归结果显示，高铁和学术会议均对城市间知识溢出产生了显著的正向影响，证实了学术会议在促进知识溢出中的积极作用。值得注意的是，相较于子表B的列（1）结果，当控制了学术会议强度后，高铁直通的估计系数显著减小，说明通过学术会议增强信息流通是高铁开通促进城市间知识溢出的重要机制。

$$similarity_{ijt} = \beta_1 HSR_{ijt} + \beta_2 meetings_{it} + \lambda X_{ijt} + \delta dis_{ij} + \alpha_{jt} + \phi_i + \rho_t + \epsilon_{ijt} \quad (5)$$

此外，从集约边际的视角，本文将高铁直通与学术会议变量的交互项纳入式（6），以探讨通过提升参会者的互动效率促进知识传播的潜在机制。具体而言，高铁直通显著降低了参会者的旅行成本，使其在会议中能更有效率地进行讨论和沟通。如表2的列（1）至列（3）所示，所有的交互项均正向显著，表明作为直通高铁的两城市，若其中知识来源城市举办的学术会议越多，那么从其向知识接收城市传播的知识也越多。上述分析表明，高铁开通在促进城市间知识溢出方面的作用渠道既包括广延边际上增加地区学术会议数量，也包括集约边际上提升学术会议的信息传播效率。上述分析证实了高铁直通通过学术会议渠道促进城市间知识溢出，进一步验证了基于专利文本相似度构造的知识溢出指标的有效性。

$$similarity_{ijt} = \beta_1 HSR_{ijt} + \beta_2 HSR_{ijt} \times meetings_{it} + \lambda X_{ijt} + \delta dis_{ij} + \alpha_{jt} + \phi_i + \rho_t + \epsilon_{ijt} \quad (6)$$

2.异地投资

为探究企业跨地区投资机制的实际效应，本文依据国泰安数据库提供的上市企业子公司信息库，整理了上市企业在异地设立子公司的时序信息，用于衡量上市企业总部所在地到异地子公司所在地（投资目的地）的年度投资活动强度。本文将该数据与城市高铁开通数据相结合，并进行如下估计：

$$subsidiary_{ijct} = \beta_0 HSR_{ijt} + \varphi_{ct} + \sigma_{jt} + \rho_{cj} + \epsilon_{ijct} \quad (7)$$

其中，$subsidiary_{ijct}$表示位于城市i的上市企业c是否在年份t于城市j设立子公司，若成立则该变量取值为1，否则为0。HSR_{ijt}设定与式（2）一致，表示城市i和城市j间高铁直通的情况。φ_{ct}是公司—年份固定效应，用于控制随时间变化的企业特征；σ_{jt}表示子公司所在地—年份固定效应，用于控制目的地城市随时间变化的未观测特征；ρ_{cj}表示上市企业—目的地固定效应，用于捕捉上市企业—目的地的未观测特征；ϵ_{ijct}表示随机误差项。附表7中列（1）显示了式（7）的估计结果，高铁直通系数显著为正，表明当城市i和城市j直通高铁后，位于城市i的上市公司c在城市j设立子公司的可能性显著增加。

为进一步探讨企业跨区域投资的作用机制，本文构建了基准回归模型，如式（8）所示。在此模型中，对控制变量的选取以及固定效应的设定，均与学术会议渠

道检验保持一致。然后,本文从广延边际和集约边际两个维度,分别纳入异地投资强度变量及其与高铁直通的交互项,探讨异地投资作为高铁开通促进知识溢出作用渠道的存在性,相应估计方程分别为式(9)和(10)。

$$similarity_{ijt} = \beta_1 HSR_{ijt} + \lambda X_{ijt} + \delta dis_{ij} + \alpha_{jt} + \phi_i + \rho_t + \epsilon_{ijt} \quad (8)$$

$$similarity_{ijt} = \beta_1 HSR_{ijt} + \beta_2 subsidiary_{it} + \lambda X_{ijt} + \delta dis_{ij} + \alpha_{jt} + \phi_i + \rho_t + \epsilon_{ijt} \quad (9)$$

$$similarity_{ijt} = \beta_1 HSR_{ijt} + \beta_2 HSR_{ijt} \times subsidiary_{it} + \lambda X_{ijt} + \delta dis_{ij} + \alpha_{jt} + \phi_i + \rho_t + \epsilon_{ijt} \quad (10)$$

附表7列(3)纳入异地投资变量后,相较于列(2),高铁直通系数显著减小,说明从广延边际上看,企业跨区域投资数量增加是高铁开通促进城市间知识溢出的重要渠道。表2列(4)内高铁直通和异地投资的交互项显著为正,说明从集约边际上看,若两城市直通高铁,当城市间跨区域投资增加时,知识溢出的规模也会变大,即异地投资提高了跨地区知识传播的效率。上述结果证实了高铁直通通过异地投资渠道促进城市间知识溢出,进一步验证了基于专利文本相似度构造的知识溢出指标的有效性①。

表2　　高铁影响知识溢出的机制检验:学术会议和异地投资②

变量	(1)	(2)	(3)	(4)
	会议数	参会人数	参会论文数	异地投资
HSR	0.018***	0.017***	0.007	0.018***
	(0.005)	(0.005)	(0.005)	(0.005)
HSR × *Number of meetings*	0.130***			
	(0.005)			
HSR × *Number of meeting attendances*		0.017***		
		(0.001)		
HSR × *Number of meeting papers*			0.031***	
			(0.001)	
HSR × *Subsidiary*				0.051***
				(0.005)
Number of meetings	0.040***			
	(0.003)			

① 此外,本文从专利合作这一知识溢出的直接行为表现角度深化对作用机制的论证,并结合城市间地理距离、创新能力差异、城市规模以及技术领域差异等维度进行拓展性分析,进一步检验基于专利文本相似度构建的知识溢出指标的有效性,限于篇幅,具体讨论参见附录8。感谢匿名审稿人的启发性意见。

② 限于篇幅,关于高铁影响知识溢出机制检验的完整结果见附表6和附表7。

（续）

变量	（1） 会议数	（2） 参会人数	（3） 参会论文数	（4） 异地投资
Number of meeting attendances		0.001 （0.001）		
Number of meeting papers			0.010*** （0.001）	
Subsidiary				0.008 （0.007）
控制变量	是	是	是	是
新申请专利城市—年份固定效应	是	是	是	是
已公开专利城市固定效应	是	是	是	是
年份固定效应	是	是	是	是
调整R^2值	0.691	0.690	0.690	0.685
样本量	372472	372472	372472	438777

注：同表1。

五、结论与启示

在新一轮科技革命和产业变革加速演进、大国科技竞争日趋激烈的时代背景下，准确测度知识溢出效应对于理解创新传播规律、优化创新政策设计具有重要意义，但现有基于专利引文的测度方法存在明显局限。本研究结合中国专利法律制度，系统分析了专利引文测度方法的问题及其制度根源。为突破这一瓶颈，本文引入基于机器学习方法构造的专利文本向量，结合知识产生的内在逻辑，提出利用专利文本相似度衡量知识溢出的新思路，并使用2002~2015年中国发明专利样本构造了测度城市间知识溢出的新指标。进而，本研究借助高铁开通这一已知能够促进知识溢出的外生冲击，检验新指标能否准确捕捉这一因果关系及其作用机制。实证结果表明，新指标不仅显著识别了高铁开通对知识溢出的促进作用，而且准确捕捉了其通过促进学术会议参与和企业异地投资的作用机制，稳健性检验和拓展性分析结果均符合理论预期，充分验证了新指标的有效性。

基于以上结论，本文提出以人工智能技术重塑科技创新评价、深化专利数据开发利用、优化区域创新空间布局的政策建议。

第一，推广人工智能技术在科技创新评价中的应用，构建多维精准的科技情

报监测体系。传统专利评价指标难以全面深刻洞察技术内容本质,而人工智能技术为深度评估提供了可能。应推动科技评价从“计量统计”向“智能语义分析”转型。一要拓展人工智能技术在技术生命周期与质量评估中的应用。利用文本向量技术测度知识溢出、评估技术突破性,通过语义比对识别颠覆性创新,为遴选高价值专利和重大科技项目提供支撑。二要利用智能算法开展关键核心技术攻关与对外依赖度评估。针对国家重大战略需求,利用机器学习算法对全球专利文本进行挖掘,构建产业链安全图谱,动态监测重点领域全球竞争力及对外技术依赖度,精准识别“卡脖子”风险点和技术空白点。三要建立适应智能化评价的标准规范与试点机制。在国家自主创新示范区先行先试,将基于文本挖掘的智能化指标纳入创新绩效考核,逐步形成涵盖知识溢出、技术安全、产业竞争力的多维评价体系。

第二,深化专利数据资源的智能化开发,提升数据要素对科技创新的供给质量。中国虽然已建立专利数据公共服务平台,但面对人工智能时代对数据“颗粒度”和“语义化”的高要求,现有数据供给在非结构化文本处理等方面仍有提升空间。一要推进专利全文本数据的标准化清洗与深度结构化。针对专利说明书等非结构化数据利用难度大的问题,重点加强对专利全文文本、法律状态、引文信息的深度清洗与标准化加工,建设高质量专利数据集,为人工智能模型训练和深度研究提供数据底座。二要构建“政产学研用”协同的数据价值共创生态。改变单纯由政府提供查询服务的模式,设立专利大数据应用创新专项,支持高校和科技企业利用人工智能技术挖掘专利数据中的隐性知识,开发高附加值的数据产品。三要完善数据分级分类开放与安全治理机制。在保障国家安全和商业秘密的前提下,有序扩大高价值专利数据的科研开放范围,通过技术手段和反馈机制持续提升数据质量,将专利数据资源转化为驱动创新的战略资产。

第三,强化交通基础设施与创新布局的协同联动,构建高效互联的区域创新共同体。畅通知识跨区域流动,推动创新要素在空间上的优化配置与产业链、创新链的跨区域深度融合,是缩小区域发展差距、发挥中心城市辐射带动作用、实现区域协调发展的关键路径。鉴于此,交通基础设施规划应超越单纯的“物理通达性”,向服务创新要素流动的“创新连通性”转变。一要将“创新流”纳入综合交通规划的核心评估指标。在规划建设交通基础设施时,应突破传统的人流、物流评价视角,重点考量项目对区域创新要素流动的边际贡献。通过加强城市群内部以及核心城市之间的快速交通联系,降低面对面交流的成本,提升知识溢出的空间效率。二要优化以核心创新城市为枢纽的快速交通网络。重点向长三角、粤港澳等主要城市群的中短距离高铁和城际快线倾斜,强化上海、深圳等创新中心城市对周边城市的辐射带动作用,通过物理时空的压缩促进“中心—外围”的创新协同与产业承接。三

要注重“硬联通”与“软机制”的配套建设。依托高铁站点等交通枢纽配套建设科技成果交易中心和跨区域产学研合作基地，通过完善异地科研合作资助、人才柔性流动等配套政策，消除行政区划导致的隐性壁垒，形成基础设施“硬环境”与制度创新“软环境”的共振效应。

参 考 文 献

[1]陈强远，张醒，汪德华.中国技术创新激励政策设计：高质量发展视角[J].经济研究，2022，(10)：52~68.

[2]陈骁，冯敬宇，高超，等.产研适配性与公共研究的创新溢出[J].数量经济技术经济研究，2025，(4)：135~156.

[3]程新生，王向前.技术并购与再创新——来自中国上市公司的证据[J].中国工业经济，2023，(4)：156~173.

[4]董昀，张明，郭强.美国技术扩散速度放缓：表现、成因及经济后果[J].经济学家，2020，(7)：119~128.

[5]杜兴强，彭妙薇.高铁开通会促进企业高级人才的流动吗?[J].经济管理，2017，(12)：89~107.

[6]黄国群，朱然然.专利审查证明责任合理分配及制度完善探讨[J].科技与法律(中英文)，2023，(6)：52~60.

[7]黄先海，王瀚迪，孙涌铭，等.数字技术与企业出口质量升级——来自专利文本机器学习的证据[J].数量经济技术经济研究，2023，(12)：69~89.

[8]李盛楠，许敏，林周周.研发人力资本效应下国际知识溢出对高技术产业创新绩效的影响研究[J].管理学报，2021，(9)：1354~1362.

[9]梁琦，李建成，夏添，等.知识交流合作的空间溢出与邻近效应——来自长三角城市群的经验证据[J].吉林大学社会科学学报，2019，(2)：41~51+219~220.

[10]林建浩，罗挺威，王茂森.开发区升级能带来创新质量提升吗？——基于异质性创新的视角[J].数量经济技术经济研究，2025，(5)：26~47.

[11]刘修岩，王峤.知识溢出的边界效应——来自专利引用数据的证据[J].经济研究，2022，(11)：84~101.

[12]卢福财，王雨晨，徐远彬.头部企业在数字化转型中的作用[J].数量经济技术经济研究，2024，(5)：92~112.

[13]米晋宏，张南.中国的科技政策与科技赶超——来自全球专利质量测算的视角[J].财贸研究，2022，(6)：36~46.

[14]欧阳志刚，胡雯华.央行沟通公告有助于提升政策利率的传导效率吗?[J].数量经济技术经济研究，2024，(10)：69~88.

[15]孙洋，李春艳，成蕾.国际知识溢出对我国工业行业创新产出的影响[J].税务与经济，2021，(3)：80~88.

[16]孙震,郭佳钰,李习保.中国专利奖的光环与溢出效应[J].科学学研究,2025,(9):2005~2016.

[17]杨德桥.专利申请人之信息披露义务的价值、模式与规则重构[J].科技管理研究,2019,(18):154~163.

[18]叶德珠,潘爽,武文杰,等.距离、可达性与创新——高铁开通影响城市创新的最优作用半径研究[J].财贸经济,2020,(2):146~161.

[19]易巍,龙小宁,林志帆.地理距离影响高校专利知识溢出吗——来自中国高铁开通的经验证据[J].中国工业经济,2021,(9):99~117.

[20]易巍,龙小宁.行政边界与专利知识传播[J].数量经济技术经济研究,2023,(10):159~180.

[21]殷戈,张晓波,李力行.中国专利质量——测度和发展趋势[J].经济科学,2024,(6):5~30.

[22]曾清华.专利引证客体研究——基于《专利法实施细则》第17条第1款规定[D].厦门大学硕士学位论文,2023.

[23]张同斌,刘文龙,王蕾.高质量创新的溢出效应:企业供应链的视角[J].经济研究,2024,(11):38~54.

[24]赵勇,白永秀.知识溢出:一个文献综述[J].经济研究,2009,(1):144~156.

[25] Agarwal R., Audretsch D., Sarkar M., 2010, *Knowledge Spillovers and Strategic Entrepreneurship* [J], Strategic Entrepreneurship Journal, 4(4), 271~283.

[26] Agrawal A., Galasso A., Oettl A., 2017, *Roads and Innovation* [J], The Review of Economics and Statistics, 99(3), 417~434.

[27] Agrawal A., Goldfarb A., 2008, *Restructuring Research: Communication Costs and the Democratization of University Innovation* [J], American Economic Review, 98(4), 1578~1590.

[28] Alcácer J., Gittelman M., Sampat B., 2009, *Applicant and Examiner Citations in U.S. Patents: An Overview and Analysis* [J], Research Policy, 38(2), 415~427.

[29] Arrow K. J., 1962, *The Economic Implications of Learning by Doing* [J], The Review of Economic Studies, 29(3), 155~173.

[30] Audretsch D. B., Feldman M. P., 2004, *Handbook of Regional and Urban Economics Vol.4* [M], Amsterdam: North-holland.

[31] Bahar D., Hausmann R., Hidalgo C. A., 2014, *Neighbors and the Evolution of the Comparative Advantage of Nations: Evidence of International Knowledge Diffusion?* [J], Journal of International Economics, 92(1), 111~123.

[32] Caniëls M. C. J., 2000, *Knowledge Spillovers and Economic Growth: Regional Growth Differentials across Europe* [M], Cheltenham: Edward Elgar Publishing.

[33] Chai S., Freeman R. B., 2019, *Temporary Colocation and Collaborative Discovery: Who Confers at Conferences* [J], Strategic Management Journal, 40(13), 2138~2164.

[34] Cohen W. M., Levinthal D. A., 1990, *Absorptive Capacity: A New Perspective on Learning*

and Innovation [J], Administrative Science Quarterly, 35(1), 128~152.

[35] Criscuolo P., Verspagen B., 2008, *Does It Matter Where Patent Citations Come from? Inventor vs. Examiner Citations in European Patents* [J], Research Policy, 37(10), 1892~1908.

[36] De Rassenfosse G., Palangkaraya A., 2023, *Do Patent Pledges Accelerate Innovation?* [J], Research Policy, 52(5), 104745.

[37] De Rassenfosse G., Pellegrino G., Raiteri E., 2024, *Do Patents Enable Disclosure? Evidence from the Invention Secrecy Act* [J], International Journal of Industrial Organization, 92, 103044.

[38] Dong X., Zheng S., Kahn M. E., 2020, *The Role of Transportation Speed in Facilitating High Skilled Teamwork across Cities* [J], Journal of Urban Economics, 115, 103212.

[39] Feng S., 2020, *The Proximity of Ideas: An Analysis of Patent Text Using Machine Learning* [J], PLOS ONE, 15(7), e0234880.

[40] Glaeser E. L., Kallal H. D., Scheinkman J. A., Shleifer A., 1992, *Growth in Cities* [J], Journal of Political Economy, 100(6), 1126~1152.

[41] Griliches Z., 1992, *The Search for R&D Spillovers* [J], The Scandinavian Journal of Economics, 94, S29~S47.

[42] Griliches Z., 1998, *R&D and Productivity: The Econometric Evidence* [M], Chicago: University of Chicago Press.

[43] Hanley D., Li J., Wu M., 2022, *High-speed Railways and Collaborative Innovation* [J], Regional Science and Urban Economics, 93, 103717.

[44] Haskel J. E., Pereira S. C., Slaughter M. J., 2007, *Does Inward Foreign Direct Investment Boost the Productivity of Domestic Firms?* [J], The Review of Economics and Statistics, 89(3), 482~496.

[45] Henderson J. V., 2003, *Marshall's Scale Economies* [J], Journal of Urban Economics, 53(1), 1~28.

[46] Higham K., De Rassenfosse G., Jaffe A. B., 2021, *Patent Quality: Towards a Systematic Framework for Analysis and Measurement* [J], Research Policy, 50(4), 104215.

[47] Jacobs J., 1969, *The Economy of Cities* [M], New York: Vintage Books: A Division of Random House.

[48] Jaffe A. B., 2022, *Elgar Encyclopedia on the Economics of Knowledge and Innovation* [M], Cheltenham: Edward Elgar Publishing.

[49] Jaffe A. B., De Rassenfosse G., 2017, *Patent Citation Data in Social Science Research: Overview and Best Practices* [J], Journal of the Association for Information Science and Technology, 68(6), 1360~1374.

[50] Jaffe A. B., Trajtenberg M., Fogarty M. S., 2000, *Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors* [J], American Economic Review, 90(2), 215~218.

[51] Jaffe A. B., Trajtenberg M., Henderson R., 1993, *Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations* [J], The Quarterly Journal of Economics, 108(3), 557~

598.

[52] Keller W., 2002, *Trade and the transmission of technology* [J], Journal of Economic Growth, 7(1), 5~24.

[53] Kelly B., Papanikolaou D., Seru A., Taddy M., 2021, *Measuring Technological Innovation over the Long Run* [J], American Economic Review: Insights, 3(3), 303~320.

[54] Kogut B., Zander U., 1992, *Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology* [J], Organization Science, 3(3), 383~397.

[55] Kogut B., Zander U., 1996, *What Firms Do? Coordination, Identity, and Learning* [J], Organization Science, 7(5), 502~518.

[56] Kuhn J., Younge K., Marco A., 2020, *Patent Citations Reexamined* [J], The RAND Journal of Economics, 51(1), 109~132.

[57] Lampe R., 2012, *Strategic Citation* [J], The Review of Economics and Statistics, 94(1), 320~333.

[58] Long C. X., Yi W., 2024, *Information Effects of High-speed Rail: Evidence from Patent Citations in China* [J], China Economic Review, 84, 102115.

[59] Marshall A., 1890, *Principles of Economics* [M], London: Macmillan.

[60] Meyer M., 2000, *What Is Special about Patent Citations? Differences between Scientific and Patent Citations* [J], Scientometrics, 49(1), 93~123.

[61] Myers K. R., Lanahan L., 2022, *Estimating Spillovers from Publicly Funded R&D: Evidence from the US Department of Energy* [J], American Economic Review, 112(7), 2393~2423.

[62] Peri G., 2005, *Determinants of Knowledge Flows and Their Effect on Innovation* [J], The Review of Economics and Statistics, 87(2), 308~322.

[63] Ramani S. V., El-aroui M. A., Carrère M., 2008, *On Estimating a Knowledge Production Function at the Firm and Sector Level Using Patent Statistics* [J], Research Policy, 37(9), 1568~1578.

[64] Romer P. M., 1986, *Increasing Returns and Long-run Growth* [J], Journal of Political Economy, 94(5), 1002~1037.

[65] Romer P. M., 1990, *Endogenous Technological Change* [J], Journal of Political Economy, 98(5), S71~S102.

[66] Sampat B. N., 2010, *When Do Applicants Search for Prior Art?* [J], The Journal of Law and Economics, 53(2), 399~416.

[67] Schuster W. M., Valentine K. G., 2022, *An Empirical Analysis of Patent Citation Relevance and Applicant Strategy* [J], American Business Law Journal, 59(2), 231~279.

[68] Wang J., Cai S., 2020, *The Construction of High-speed Railway and Urban Innovation Capacity: Based on the Perspective of Knowledge Spillover* [J], China Economic Review, 63, 101539.

[69] Younge K. A., Kuhn J. M., 2016, *Patent-to-patent Similarity: A Vector Space Model* [R], SSRN Working Paper, No. 2709238.

The Measurement and Test of Knowledge Spillovers: Evidence from Generating Patent Text Similarity Based on Machine Learning

LONG Xiaoning[1,2] ZHANG Fan[3] YI Wei[4]

(1.Intellectual Property Research Institute, Xiamen University;
2.The Belt and Road Research Institute, Xiamen University;
3.School of Logistics and E-commerce, Zhejiang Wanli University;
4.Finance and Economics College, Jimei University)

Summary: Innovation is the fundamental engine of sustained economic growth, and knowledge spillovers serve not only as the primary channel through which innovation exerts its impact but also as a vital source of future technological advancement. As China moves from high-speed growth to high-quality development, technological innovation has become a decisive driver. However, a "productivity paradox" has emerged: both China and advanced economies have experienced declining total factor productivity (TFP). Increasingly, this trend is attributed to a slowdown in knowledge diffusion. The "Proposal for the 15th Five-Year Plan," recently adopted by the Central Committee, explicitly identifies "steady improvement of TFP" as a strategic objective. Therefore, accurately measuring knowledge spillovers within the broader framework of industrial modernization is both an academic imperative and of pressing practical relevance.

However, measuring knowledge spillovers in China encounters distinctive institutional constraints. Although patent citations constitute the global standard for tracing knowledge flows, this study reveals systemic limitations in their applicability in the Chinese context. A systematic examination of China's patent legislation and administrative rules reveals that while applicants are required to disclose prior art, non-compliance carries no legal penalties. This absence of enforcement generates a clear conflict of interest, incentivizing applicants to deliberately withhold citations. Thus, citation records are dominated by "examiner citations," which reflect administrative searches conducted by patent authorities rather than an inventor's genuine learning or knowledge absorption process. This institutional distortion undermines the reliability of citation-based measures, rendering them inadequate for capturing authentic knowledge spillovers in China.

This study tackles the measurement challenge by introducing a novel artificial intelligence (AI)-based approach. Using the full text of Chinese invention patents from 1985 to 2023, we adopt high-dimensional semantic vectors trained by the Google Patents

team. By computing cosine similarity across these vectors, we develop a new indicator of inter-city knowledge spillovers. Leveraging a panel of Chinese cities from 2002 to 2015, we exploit the staggered opening of high-speed rail (HSR) as an exogenous shock to validate this indicator. The difference-in-differences (DID) estimates demonstrate that the text-similarity measure reliably captures the causal effect of HSR on knowledge spillovers, revealing a statistically significant and robust positive impact.

Mechanism analysis further corroborates the indicator by identifying specific transmission channels. Evidence reveals that spillovers primarily arise from intensified face-to-face academic interactions and expanded cross-regional corporate investments. The indicator's ability to detect these micro-level mechanisms-facilitated by lower travel costs due to HSR-proves it reflects authentic economic behavior rather than statistical numbers. Additionally, extensive heterogeneity analyses assess the indicator's validity. The results reveal that spillover effects detected by the metric vary systematically across cities with different characteristics, which is fully consistent with theoretical expectations. These robust findings affirm that the text-based similarity measure is a sound tool for quantifying knowledge flows in the Chinese context. This study makes three contributions to the literature. First, it solves the measurement challenge by identifying institutional roots of citation data failure and introducing semantic vector analysis to circumvent systemic biases. Second, it rigorously validates the new metric using causal inference, offering a verified tool for research. Third, it achieves a methodological breakthrough as a novel study to apply Google Patent vectors to causal inference in a developing country, extending their use from control group construction to direct measurement.

Based on these findings, we propose the following policy recommendations: promote the "intelligent transformation" of technology evaluation by incorporating AI-based semantic analysis to identify disruptive innovations; deepen the development of patent data as a production factor by cleaning and structuring unstructured texts; and optimize regional planning by prioritizing "innovation connectivity" in transport infrastructure.

Keywords: Knowledge Spillovers; Patent Text Similarity; Patent Citations; High-speed Rail

JEL Classification: D83; O33; R41

(责任编辑:李兆辰;数据编辑:朝　阳)