数据相关信息

1.专利数据

本文使用的专利数据来自于国家知识产权局(SIPO)数据库，其中每项专利包含了专利名称、申请日期、申请地址、申请人具体信息等一系列具体信息。在搜寻数据的基础上，我们进一步对合作申请专利数据进行了细致的处理。具体操作如下：首先，将申请人依托的机构为个人的专利予以剔除，合作申请专利范围被缩小至所有申请人所在地理位置都可被获取的专利，这些申请人主要包括公司、高等院校和科研院所等。其次，对于申请人所在地为3个及以上城市的专利，采用两两交叉的方式进行拆分，以确保合作城市间一一对应的关系。例如，将申请人所在地为北京、上海、广州三个城市的专利数据记录为“北京—上海”、“北京—广州”、“上海—广州”。最后，将处理后的数据按照城市与城市群内所有城市的对应关系进行加总得到“城市—年份”数据库，由此得到2011—2019年间中国城市间合作专利的665495项记录，其中包括发明专利372430项，实用新型专利265404项，外观设计专利27611项。所对应的变量分别为基准回归及后续分析中的合作创新水平（）、内生性处理中的全国层面的专利总量（）、稳健性检验中的以发明专利计算的合作创新水平（）、机制检验中的合作创新水平扩展边际（）、合作创新水平集约边际（）、三年内首次参与合作创新的伙伴数量（、连续三年均有合作创新产出的伙伴数（、）、）、异质型团队（）、同类型团队/异质型团队（）、企业-高校（）、企业-科研院所（、、小组计算的知识宽度（1）、大组计算的知识宽度（2）、大类计算的知识宽度（3）以及进一步分析中的人均群外合作申请专利数量（）、人均总合作申请专利数量（+）、人均城市内部合作申请专利数量（）、人均城市内部独自申请专利数量（）、城市人均申请专利总量（）。

2.城市层面变量的数据

来自各年份的《中国城市统计年鉴》，少数缺失值采用插值法补充。所对应的变量分别为：控制变量中的城市经济发展水平()、教育水平()、基础设施()、信息化程度()、政府支出()、产业结构()和城镇化率()；机制检验中的高新技术产业从业人员（）、生产性服务业从业人员（）、社会性服务业从业人员（）、区位熵-高新技术产业从业人员（）、区位熵-生产性服务业从业人员（）、区位熵-社会性服务业从业人员（）、政府科学支出（）、规模扩散指数-政府科学支出（）。

3.其他变量：

（1）内生性处理中，河流密度原始数据来源于国家基础地理信息中心网站，根据城市群涵盖范围重新计算得到，方言多样化指数原始数据来源于《中国语言地图集》和《汉语方言大词典》，参考徐现祥等(2015)进行测算得到。所对应的变量分别为城市群河流密度（）、方言多样性指数（）。

（2）政策变量中，变量整理自各项城市群发展规划，创新型城市试点政策()和“宽带中国”试点政策()整理自相应政策公开文件，进一步分析中的创新关注度（）根据各项城市群发展规划通过查询创新相关术语词典整理而得。

（3）机制分析中的其余变量：新注册企业数据（、）来源于爱企查网站，按照年份、行业分别检索历年20个行业口径的企业注册数据，进一步对高技术产业新企业数量加总得到“城市—年份”层面的高新技术行业新注册企业数量数据，高新技术行业的行业划分同上。高新技术企业认定(上市公司)数据（、）来源于国泰安数据库，按照历年高新技术企业认定名单整理到城市层面。将上述数据对数化处理后，作为被解释变量代入回归方程。企业(上市公司)研发投入（、）来源于国泰安数据库，按照上市公司所属城市进行加总得到城市层面企业研发投入。专利交易量（、）来源于国家知识产权局，按照每条专利申请人所属城市进行加总得到城市层面专利交易量数据。省级层面的技术合同成交额（）来源于中国科技统计年鉴，按照各城市高技术人才占比计算到城市层面。

（4）异质性分析中的变量：城市群的空间结构定义，一是参照各项规划的划定方式进行二分类划分，有两个及以上中心城市的城市群定义为多中心城市群，多中心城市群中城市赋值为1，其余城市赋值为0；二是使用赫芬达尔指数()对城市群空间结构进行测算。研发禀赋定义，一是参考Agrawal et al.(2017)的设定方法，将前1%参与创新合作的企业定义为明星企业，这些企业所在的城市即为研发禀赋较高的城市，即赋值为1，其余城市为0；二是考虑到顶级高校科研资源的多样性和创新视野的前瞻性，将拥有“985工程”和“211工程”高校的城市定义为研发禀赋较高的城市，即赋值为1，其余城市为0。