说明文档

（一）文件说明

本论文使用的数据代码压缩包命名为“2024-00250\_**数据**”，解压之后里面有“2024-00250\_**程序代码**”、“2024-00250\_**日志文件**”和“2024-00250\_**说明文档**”3个部分。

1、“2024-00250\_**程序代码**”里面包括2个文件，分别是“2024-00250\_表格程序代码”和“2024-00250\_图片程序代码”文件：

“2024-00250\_表格程序代码”是表格运行文件，主要生成正文中表1的监测站数量变化趋势图、表2-表6的运行结果，以及在线附录中表III1变量定义与描述性统计、表IV1、表IV2、表VII1、表VIII1、表IX1、表X1和表XI1的运行结果。

“2024-00250\_图片程序代码”是图运行文件，主要生成正文中图1a-b、图2a-b，以及在线附录中图V1-图V4的平行趋势检验、图VI1-VI4的安慰剂检验。

2、“2024-00250\_**日志文件**”中的“table1”是正文中表1的监测站数量变化趋势表；“table2”-“table6”分别是正文中表2-表6的运行结果；“tableIII1”是附录中表III1的变量定义与描述性统计表；“tableIV1”和“tableIV2”是附录中表IV1和表IV2的运行结果；“tableVII1a”-“tableVII1d”是附录中表VII1的运行结果；“tableVIII1”、“tableIX1”、“tableX1” 和“tableXI1”分别是附录中表VIII1、表IX1、表X1和表XI1的运行结果；“图1a”和“图1b”是正文中图1a-b的结果；“图2a”和“图2b”是正文中图2a-b的运行结果；“图V1”-“图V4”是附录中图V1-图V4的运行结果；“图VI1”-“图VI4”是附录中图VI1-图VI4的运行结果。

需要说明的是：1.与正文中结果相比，图1a-b、图2a-b、图V1-V4以及图VI1-VI4在颜色和实虚线等方面做了细节调整。2.附录I图A1和图A2使用ArcGIS Pro直接生成，因此上述文件中不包含两图的生成代码。

3、“2024-00250\_**说明文档**”即是本说明文档。

（二）数据来源

本文全部图表的生成主要使用了4个数据集，分别对应“2024-00250\_表格程序代码”和“2024-00250\_图片程序代码”文件中的data1、data2、data3和data4。其中，data1是本文用于主要实证分析的核心数据集，使用该数据生成的图表包括：正文中表2-表6和图2a-b；附录中表III1、表VII1、表VIII1、表IX1、表X1、表XI1、图V1-图V4和图VI1-图VI4。data2用于生成正文中表1。data3用于生成正文中图1a-b。data4用于生成附录中表IV1和表IV2。以下是4个数据集的数据来源和数据处理介绍：

1、data1区县整合数据集：

data1的数据主要由以下六个部分所组成：

（1）工商企业注册信息数据

本文使用的工商企业注册信息数据来源于国家企业信用信息公示系统。该系统公布了1949年以来中国历史上曾经注册过的企业注册信息，包括企业名称、企业类型、行业门类、注册资本及办公所在地等信息。我们首先根据《2017年国民经济行业分类（GB/T4754-2017）》90个二位行业码，对2000年1月-2020年12月的工业企业注册信息进行了爬取（主要包含采矿业、制造业以及电力、热力、燃气及水生产和供应业），然后分区县、行业、年份对2003年1月至2019年12月的工业企业注册信息进行了整理，最终获得了新增工业企业注册数的区县-行业-年份面板数据。在数据清洗过程中，我们删除了：（1）由于注册地址缺失、注册地址无法准确定位到区县级的企业；（2）行业类型或相关重要信息缺失的企业；（3）注册时间即注销/吊销时间，或在其之后的企业，最终获得883万条新增工业企业注册信息。

（2）中国县域统计数据

本文使用的县域经济、社会数据来源于《中国县域统计年鉴》的统计数据。县域统计年鉴数据覆盖全国超过两千个县域单位，内容涉及了县（市）的社会经济主要指标、分区域的社会经济基本情况以及按照不同经济指标分类的县（市）数据。

需要说明的是，县域统计年鉴在2012年以前并没有对“第二产业从业人员数”这一指标进行单独统计，本文通过以下流程进行补充：（1）通过2003年-2013年各省、市、县地方统计年鉴获取各区县一二三产业的“从业人员数”数据；（2）对收集的数据进行统计口径确认（主要通过年鉴中已公布的数据进行纵向对比）；（3）确认统计口径一致的数据进行直接补充（填补后有效样本占总样本54.84%）；（4）在统计口径不一致的情况下，进一步使用收集数据测算各区县第二产业从业人员数的年度变化率，然后根据地方统计年鉴中已公开的数据进行向上推演（填补后有效样本占总样本83.32%）；（5）对剩余缺失值，本文参考Chao et al.（2018）的做法，根据数据的时间演变趋势，使用移动平均法进行插补[[1]](#footnote-1)。

（3）环境监测站数据

本文的环境监测站数量数据来源于中国环境监测总站的全国城市空气质量实时发布平台。该平台公布了各城市的实时空气质量数据以及城市内各监测站的所在位置和运行情况。我们基于该平台的公开信息获取了2013年至2019年全国监测站的具体信息并进行了以下处理：（1）对该平台实时发布的全国城市空气质量监测站数据[[2]](#footnote-2)进行地理信息编码；（2）使用Google地图的坐标拾取功能获取各站点所在的地理位置信息；（3）剔除停运站点后，利用ArcGIS Pro将多组附带精确地理位置信息的坐标数据投射到中国县级矢量地图上；（4）根据县域边界，统计各区县的监测站数量并形成本文所需的系列变量。

（4）多时期土地利用遥感监测数据

中国科学院资源环境科学数据中心的中国多时期土地利用/土地覆盖遥感监测数据库（CNLUCC）提供了1990年、1995年、2000年、2005年、2010年、2015年、2018年以及2020年中国各类型土地利用状况的具体信息，监测类型包括耕地、林地、草地、水域以及居民点用地等。我们首先通过ArcGIS Pro对上述各年度的土地利用栅格数据进行矢量化处理（根据各年度的县域行政区划边界），获得各区县不同类型土地利用的具体面积；然后通过计算得出各区县城镇用地以及工交建设用地[[3]](#footnote-3)的面积占比；最后，本文参考Chao et al.（2018）提出的渐进时空预测法（PST），利用2000年-2020年中国县域统计年鉴数据对缺失值进行插补，进而获得各区县由“城镇用地占各区县总面积的比例”表征的“城镇化水平”以及由“工交建设用地占各区县总面积的比例”表征的“交通便利性”的区县-年份面板数据。

（5）环保行政处罚数据

环保行政处罚数据来源于北大法宝网。本文首先通过北大法宝网爬取了2006-2019年全国的环保行政处罚案件及处罚对象的所在区县信息，然后通过分区县、年份对环保处罚案件进行整理，最后得到“环境规制强度”的区县-年份面板数据。

（6）微观工业用地供应数据

微观工业用地供应数据来源于中国土地市场网。本文通过中国土地市场网爬取了2003—2019年全国土地出让数据，爬取对象包括供地面积、供地价格、供地方式、土地用途等。对于获得的数据我们进行了以下处理：（1）根据土地用途筛选出工业用地；（2）根据《2017年国民经济行业分类（GB/T4754-2017）》中的二位码行业分类筛选出工业企业用地[[4]](#footnote-4)；（3）分区县、年份对工业用地出让面积、不同方式出让宗数进行加总以及对出让价格求平均，从而得到工业用地出让面积、不同方式出让宗数以及平均出让价格的区县-年份面板数据。

2、data2全国城市空气质量监测站信息数据集：

该数据包含了2013年至2019年各监测站的所在位置和运行情况，为data1中“环境监测站数量数据”的原始数据。可以直观地考察各年份各区域的的监测站数量和运行情况，本文附录的图A1和图A2也是基于该数据所生成。

3、data3经济指标的时间序列数据：

该数据主要包含了3个指标：城区制造业就业人员数、城区人均工业增加值和县域人均工业增加值。其中，“城区制造业就业人员数”为《中国统计年鉴》中的城镇制造业就业人员数据；“城区人均工业增加值”和“县域人均工业增加值”，则是通过对《中国县域统计年鉴》中公布的区县工业增加值进行计算后获得，计算公式为：城区人均工业增加值=总的城区工业增加值/总的城区户籍人口；县域人均工业增加值=总的县域工业增加值/总的县域户籍人口。

4、data4中国流动人口动态监测调查数据（CMDS）与区县数据的整合数据集：

中国流动人口动态监测调查（CMDS）是国家卫生健康委流动人口司组织开展的流动人口专项监测调查，调查范围涉及31个省（市、区）和新疆生产建设兵团，涵盖全国1800余个县，上万个居（村）民小组。年均样本量约18万。内容涉及流动人口及家庭成员人口基本信息、流动范围和趋向、就业和社会保障、收支和居住、健康与公共卫生服务、婚育和计划生育服务管理、子女流动和教育、社会融合与心理健康等。

中国流动人口动态监测调查（CMDS）为我们提供了大量的流动劳动力微观数据，其所包含的流入地和迁出地信息，可以让我们更有效地识别环境监管强度对流动劳动力选址行为的影响。对此，本文通过将2015-2017年中国流动人口动态监测调查数据（CMDS）与上述data1的区县整合数据集进行合并，以进一步考察和提供监管强度对劳动力流动行为影响的微观证据（详见附录V）。需要说明的是，与data1使用的数据相比，data4中，我们进一步加入了区县层面的城镇和农村居民人均可支配收入均值来反应劳动力在该区县就业可能获得的预期收入，城镇和农村居民人均可支配收入数据来源于各省统计年鉴。

（三）Dofile里面使用的指标说明

（1）data1部分：

工业企业密集度：新增工业企业注册数（个/平方公里）；

劳动力密集度：第二产业从业人员数（人/平方公里）；

规上工业企业数：规模以上工业企业数（个/平方公里）；

第三产业劳动力密集度：第三产业从业人员数（人/平方公里）；

制造业密集度：新增制造业企业注册数（个/平方公里）；

采矿业密集度：新增采矿业企业注册数（个/平方公里）；

水电煤业密集度：新增水电煤业企业注册数（个/平方公里）；

高污染企业密集度：新增高污染企业注册数（个/平方公里）；

低污染企业密集度：新增低污染企业注册数（个/平方公里）；

监测网设立：地市层面设立监测网=1，否则=0；

dummy\_0\_5：各区县到最近监测站的距离小于等于5km=1，否则=0；

dummy\_0\_10：各区县到最近监测站的距离小于等于10km=1，否则=0；

dummy\_n\_m：距离虚拟变量，各区县到最近监测站的距离大于nkm小于等于mkm则=1，否则=0；

event\_city：当年年份-监测网设立年份；

邻近地市的区县：监测网设立地市的邻近地市的区县=1，否则=0

监管强度：各区县的监测站数量（个）；

监管强度区位落差：城区和县域监测站的数量差值（个）；

监测网设立\_去对照点：去掉属于对照点的监测站后，地市层面设立监测网=1，否则=0；

监管强度\_去对照点：去掉属于对照点的监测站后，各区县的监测站数量（个）；

监管强度区位落差\_去对照点：去掉属于对照点的监测站后，城区和县域监测站的数量差值（个）；

人口规模：户籍人口数（万人）；

城镇化水平：城镇用地占区县总面积的比；

交通便利性：工交建设用地占区县总面积的比；

第二产业规模：第二产业产值占区县总产值的比；

第三产业规模：第三产业产值占区县总产值的比；

财政状况：（财政支出-财政收入）/财政收入；

福利保障：各种社会福利收养性单位（个）；

省直管县：属于省直管县=1，否则=0；

环境规制强度：环保处罚案件数（件）；

出让面积：当年工业用地出让面积占县域面积的比；

出让价格：工业用地的出让均价（万元/公顷）；

出让方式：总出让宗数中招拍挂出让宗数的占比；

区域经济：人均GDP（元）；

空气质量：PM2.5浓度（微克/立方米）。

（2）data2部分：

城区：城区=1；县域=0。

（3）data3部分：

城区制造业就业人员数（万人）：城镇单位制造业就业人员（万人）；

城区人均工业增加值（万元）：城区工业增加值/户籍人口（万元）；

县域人均工业增加值（万元）：县域工业增加值/户籍人口（万元）。

（4）data4部分：

Q101N1：流动原因是什么？“务工/工作”=1；

Q206：现在工作的行业？“制造业”=3；

Q205：现在的主要职业？；

Q207：现在的就业单位性质？；

choice：为劳动力选择流入区县的哑变量，其所在就业区县=1，其他区县=0。备选集为劳动力流入城市的所有区县；

监管强度：各区县的监测站数量（个）；

预期收入：各区县的人均可支配收入（元）；

人口规模：户籍人口数（万人）；

城镇化水平：城镇用地占区县总面积的比；

交通便利性：工交建设用地占区县总面积的比；

第二产业规模：第二产业产值占区县总产值的比；

第三产业规模：第三产业产值占区县总产值的比；

财政状况：（财政支出-财政收入）/财政收入；

福利保障：各种社会福利收养性单位（个）。

1. 对于其余变量缺失值，本文也使用相同的方法进行补齐。 [↑](#footnote-ref-1)
2. 详见：<https://www.cnemc.cn/sssj/>。 [↑](#footnote-ref-2)
3. 工交建设用地主要包含大型工业区、交通道路、机场、码头等工业交通用地。 [↑](#footnote-ref-3)
4. 爬取对象参考周玉龙等（2024）的做法，根据中国工业二位码行业进行爬取，包括化学原料及化学制品制造业、医药制造业、金属制品业、有色金属冶炼及压延加工业、电力、热力的生产和供应业、化学纤维制造业、家具制造业、交通运输设备制造业以及石油加工、炼焦及核燃料加工业等。 [↑](#footnote-ref-4)