详细说明文档

本文所使用的数据主要为：交通事故裁判文书数据、地级市贸易流量数据、其他经济地理数据等。对于本文所有的原始或衍生数据的相关信息，如来源、描述、收集方式等，均在本文档、正文或附录中有具体描述。数据指标的使用可参见do文档、logfile文档。

一、交通事故裁判文书数据

（一）数据来源

2013年1月至2021年10月中国裁判文书网中的交通事故民事一审案件

（二）数据描述

本文提取其中交通事故判决书的相关信息 ，主要为基层人民法院审理的交通事故民事一审案件 。对每份判决文书，提取以下信息：法院地址、审判和裁决日期、原告地址、被告地址、案件基本事实、法官提供的司法推理、原告和被告所需承担的案件受理费等信息。具有这些完整信息的案件约为126万件。

（三）收集方式

通过网络爬虫获得原始数据，并使用Python对判决文书进行文本分析，提取相应指标。各指标的使用见do文件和logfile文件，得到的主要指标包括：

1. 外地原告：虚拟变量，该变量为0时表示原告、被告均来自本省（即和法院所在省相同），为1表示原告来自外省，被告来自本省，为-1时表示其他情况
2. 外地原告\_city：虚拟变量，该变量为0时表示原告、被告均来自本市（即和法院所在市相同），为1表示原告来自外市，被告来自本市，为-1时表示其他情况
3. 外地被告：虚拟变量，该变量为0时表示原告、被告均来自本省（即和法院所在省相同），为1表示被告来自外省，原告来自本省，为-1时表示其他情况
4. 外地被告\_city：虚拟变量，该变量为0时表示原告、被告均来自本市（即和法院所在市相同），为1表示被告来自外市，原告来自本市，为-1时表示其他情况
5. 外原外被：虚拟变量，该变量为0时表示原告、被告均来自本省（即和法院所在省相同），为1表示原告来自外省，原告来自外省，为-1时表示其他情况
6. 外原外被\_city：虚拟变量，该变量为0时表示原告、被告均来自本市（即和法院所在市相同），为1表示原告来自外市，被告来自外市，为-1时表示其他情况
7. win\_rate\_yuangao：原告胜率，为0到1的数值变量
8. win\_rate\_beigao ：被告胜率，为0到1的数值变量
9. win\_rate\_二审\_yuangao：原告胜率，为0到1的数值变量，以二审胜率替换一审胜率。
10. 电动车原告 ：原告驾驶非机动车，01虚拟变量，为1表示原告驾驶了非机动车
11. 行人原告 ：原告是行人，01虚拟变量，为1表示原告是行人
12. death\_and\_injury ：原告有伤亡，01虚拟变量，为1表示原告有伤亡
13. 上诉 ：是否上诉，01虚拟变量，为1表示上诉
14. jianyichengxu ：是否采用简易程序，01虚拟变量，为1表示采用
15. free\_code ：是否采用自由裁量，01虚拟变量，为1表示采用
16. zhengren ：是否传唤证人，01虚拟变量，为1表示传唤
17. words\_body ：文书正文字数
18. city\_court:法院所在城市
19. qu\_court:法院所在县区
20. year:案件时间（年份）

二、地级市贸易流量

（一）数据来源

本文数据来源于2017年中国所有企业的增值税发票数据。原始数据为国家税务总局系统收录的各地级市入库企业增值税发票，发票上注明交易的销售金额、增值税额、销售方和购买方的公司名称及其公司所在地，实际上记录了地区之间的贸易流量。

（二）数据描述

本文将原始数据加总到“卖出地-买入地”层面，涵盖的地级市为342个，加上4个直辖市，共计346个地区119,716个“卖出地-买入地”组合，去掉本地对本地的样本及部分极端值，共得到106,755个有方向的“卖出地-买入地”的贸易量组合。得到的指标主要有：

1. 贸易总额\_双边 ：城市对的贸易量
2. 贸易金额\_工业品\_双边 ：城市对工业品的贸易量
3. barrier\_贸易总额：城市对的贸易成本，计算公式见原文

三、其他经济地理变量

（一）数据来源

地理变量来源于百度地图开放平台，经济数据来源于历年《中国城市统计年鉴》。

（二）数据描述

本文还涉及的地理变量主要包括距离、海拔差、地势崎岖度之差、所处河流流域、偏远程度；经济变量主要为各地级市的人均GDP、人口数、公路密度等。主要指标包括：

1. 政府支出比例:各城市一般预算支出与地区生产总值的比值
2. 公路密度：各城市公路里程和土地面积的比值
3. 人均\_FDI：各城市人均FDI
4. 人均地区生产总值元：各城市的人均GDP
5. 年底总人口户籍人口万人：各城市年底总人口户籍人口
6. distance：城市对之间的距离
7. altitude\_diff：城市对之间的海拔差距
8. mean\_remote：城市对的偏远度均值
9. tri\_diff：城市对的崎岖度差距