基于自动化手段的破坏社会主义市场经济秩序罪案件的分析

———以四川省为例

姓名： 李森洋

学号： 201811051123

专业： 地理信息科学

年级： 2018

日期： 2021.02.10

摘　要：分析案件的判决文书社会的犯罪治理有参考作用。利用自动化手段可以更加高效、精确地实现这一过程。本文基于自动化手段，对四川省 2016-2020 年的破坏社会主义市场经济秩序罪的案件进行了时间、空间、罪行、犯罪人群等方面的分析。结果显示，2016-2020 年四川省案件数量先上升后下降，每一年的案件数量从年初递增到年末，东部规模大于西部，大中城市大于小城市，且有向小城市转移的趋势。罪行以破坏市场经济秩序罪最多，犯罪的处罚相对较轻。犯罪人群年龄分布呈 30 岁和 45 岁为峰值的马鞍形，男性罪犯多于女性，各民族罪犯与自身人口所占比例相近，且青年犯罪更集中到少数民族上。最后本文总结了案件规律的成因，针对其特点提出了治理方式和研究方法的相关展望。

关键词：破坏社会主义市场经济秩序罪；自动化；时空分布；经济犯罪；

目　录：

1 引言

破坏社会主义市场经济秩序罪是我国在 1997 年于刑法中提出的专有概念，是指违反国家经济管理法规，破坏国家经济管理活动，严重扰乱社会主义市场经济秩序的行为[1]。按照具体的犯罪方式，可以分为生产销售伪劣商品罪、走私罪、妨碍对企业管理罪、破坏金融管理秩序罪、金融诈骗罪、危害税收征管罪、侵犯知识产权罪、扰乱市场秩序罪等八大种类，各大类下又包含具体的犯罪类型。改革开放以来我国的经济得到快速发展，也带来了大量经济犯罪问题。对已经发生的案件进行分析，能更好地帮助我们了解各种案件的发生规律、分布规律，对于精准合理地打击犯罪、预测犯罪、预防犯罪都有极强的指导意义，促进社会和谐稳定运转。

根据既有犯罪案件进行分析的研究目前已经发展得非常成熟，近几十年来，结合 GIS 技术的发展，研究者也能实现在大量数据的基础上的时空综合分析，犯罪研究的重点也逐渐转移到更深层次的成因分析，如李欣竹[2]在盗窃案件的时空分析基础之上进一步探讨了其成因，刘玲[3]等分析了四川省拐卖儿童犯罪的时空演进规律，并探讨了其影响因素。而在关于以破坏社会主义市场经济秩序罪为代表的经济犯罪的研究方面目前涉及的工作不多，尤其是在省域尺度下结合大量数据的相关探讨。鉴于此，本文将对四川省 2016-2020 年公布的破坏社会主义市场经济秩序罪的案件进行时空分析和罪行人群分析，并进一步探讨其背后的成因，以冀实现一定的补充作用。

2 研究区与数据

2.1 研究区概况

四川省位于东经 97°21′~108°31′，北纬 26°03′~34°19′之间，位于中国西南部，总面积 48.5 万平方公里，总人口 8375 万（2019 年），下辖 21 个地级行政区，183 个县区级行政区。2019 年 GDP 总量为 46615.82 亿元，位居全国第六位。四川省经济规模庞大，人口众多，区域内差异明显，东部为盆地，西部为高原，少数民族众多。对于经济犯罪案件来说有较好的代表性。同时，省内地级/区县级单位较多，以之作为基本单位进行分析可以得到较为细致的结果。

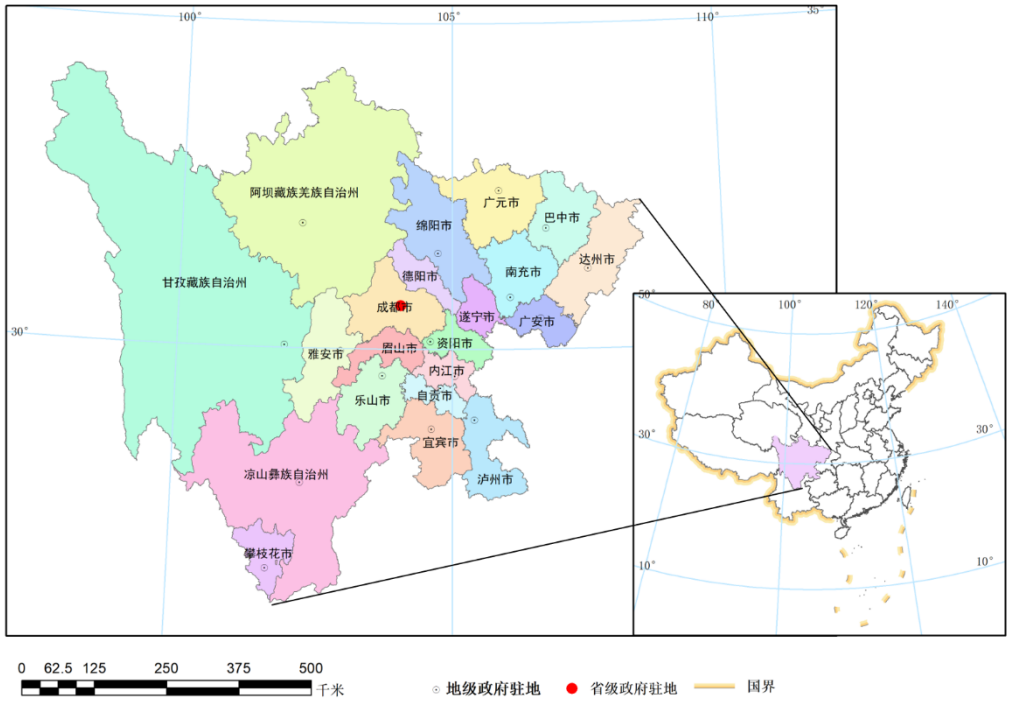


图 1　研究区概况

2.2 数据源

本文使用案件数据来自于中国裁判文书网（https://wenshu.court.gov.cn，在此承诺，所有获取的数据仅供研究分析使用）。该网站有海量的判决文书可供下载。该网站上线于 2013 年 7 月 1 日。考虑到上线初期时间段数据量不完整问题，我选取的研究时间段为 2016-2020 年。案件类型为刑事案件，文书类型为判决书，为了实现区县级的分析，法院层级均为基层法院。区域限定在四川省。最后下载全部相关文书共计 7530 份。

本文相关空间数据来自于高德地图提供的行政区域查询 API 获取的区县及以上行政区域的矢量边界，并按照民政部发布的“2019 年 1 月中华人民共和国县以上行政区划代码”进行查找核实。

3 研究方法

本文首先利用自动化手段，下载了相关文书及其附属信息，再针对文本特点进行关键信息的自动提取。最后综合空间信息进行各维度上的分析，并进行成因等方面的讨论。

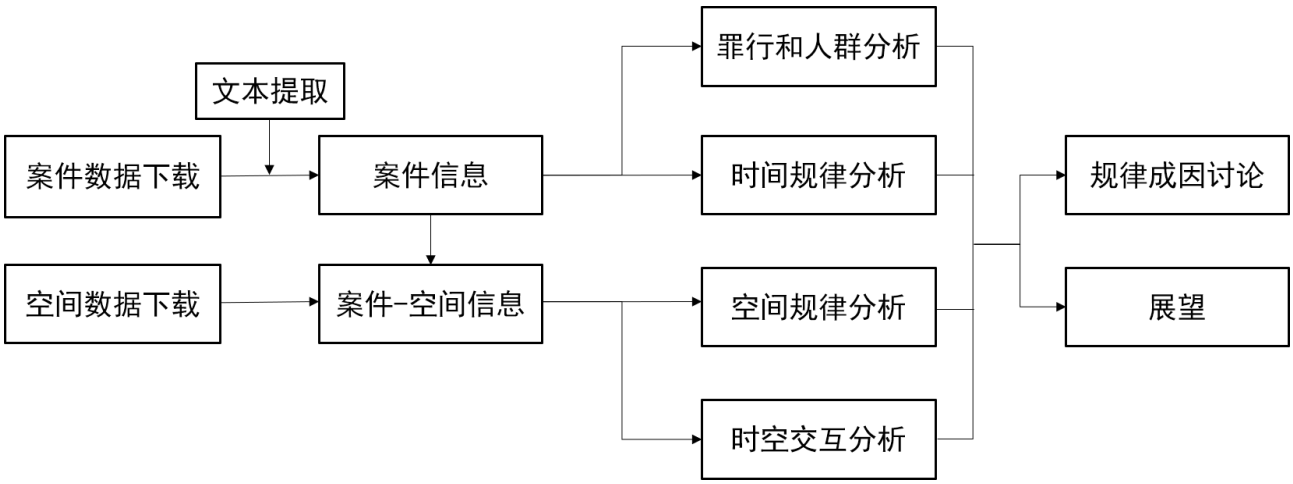


图 2　研究流程图

3.1 案件数据下载

本文利用 web scraper 工具提取每个文书的发布时间、案号、文书概要、标题、下载链接，基于提取的下载链接在 python 中利用 pyautogui 模块模拟鼠标键盘操作将链接输入到浏览器中实现自动下载，以获取每份文书的 doc 文件。具体操作步骤详见附录中的【8.1 案件数据自动下载具体步骤】。

3.2 文本提取

首先利用 python 的 win32com 模块将下载的所有 doc 格式的文书转换为 docx 格式，然后利用 docx 模块读入文书，文书中每一段都成为了一个字符串。根据文书的文字布局结构构建对应正则，引入 re 模块，循环匹配每一个字符串和正则。引入 pandas 模块，将匹配结果存入 dataframe 数组，获取罪犯姓名、性别、出生年份、判决年份、民族、文化水平、罪名、刑期、罚金共九方面的信息（如果一个文书里不存在某些信息，写入空值）。最后导出到 xlsx 格式的表格中，获得罪犯信息 12888 份。具体操作步骤详见附录中的【8.2 文本自动提取具体步骤】。

3.3 分布规律分析

3.1 中主要获取了案号和发布时间的信息，根据案号可以将每一份文书的地址归类到某个地级市和某个区县，结合空间数据可以在 arcgis 和 python 中进行空间分析；发布时间信息可以用以进行时间分析；以上两类分析结合可以进行更一步的时空分析，主要是不同空间的时间规律分异和不同时间的空间分异规律。

3.2 中主要获取了犯罪人及其罪行的信息，将其统计归类可以对破坏社会主义市场经济秩序罪的犯罪人群、犯罪特点、承受的刑事责任等进行分析。以上分析都为进一步的成因分析和相关展望打下基础。

4 研究结果

4.1 时间规律分析

图 3 为根据文书发布时间绘制的按年份推进的案件数量折线图，文书发布时间虽然会受到录入网站等操作的影响，但一定程度上可以表示这些案件的所处大致时间段。图中显示，2016 年的案件数量为 1000 左右，到 2017 年突然增长到 1000 左右，接着 2017、2018、2019 年都保持该水平，2018 年达到峰值，2020 年有所下降，回落到 1000 左右。总的趋势是先上升，稳定一段时间再下降。

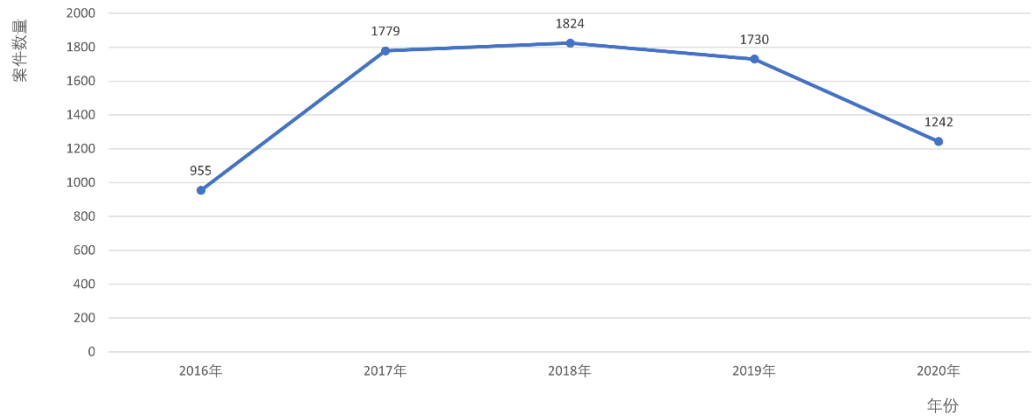


图 3　2016-2020 年四川省按年度破坏社会主义市场经济秩序罪案件数量折线图

我进一步将时间单位细分到季度（图 4），发现在之前的规律的基础上，每一年从第一季到第四季案件数量基本上都是呈上升趋势（2020 年年内不太一样，是呈先上升后下降趋势的）。而每一年第四季到第二年第一季总会呈一个下降趋势，且这个下降是越来越剧烈的。

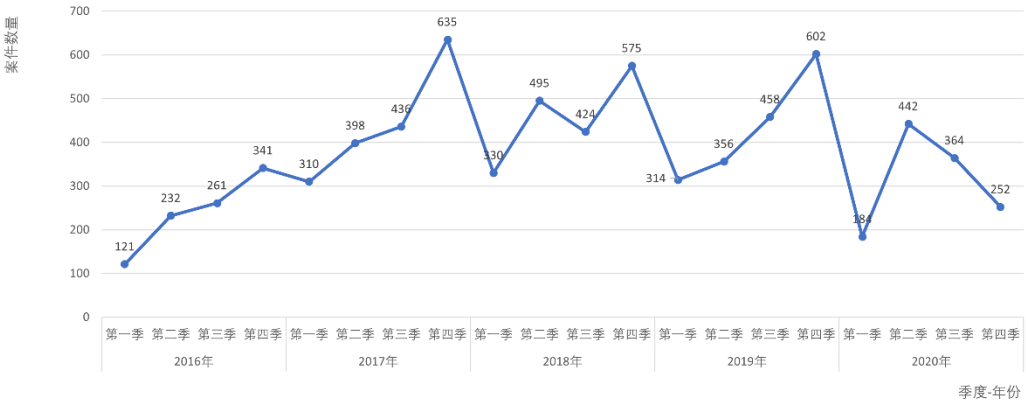


图 4　2016-2020 年四川省按季度破坏社会主义市场经济秩序罪案件数量折线图

更加细致地将时间单位下探到月份（图 5），可以发现每一年内案件数量的上升集中在 2 至 3 月份和 10 至 12 月份，4-9 月份案件数量保持稳定，峰值都在 12 月。2020 年的规律仍然跟前几年不太一致，呈先上升后下降的趋势。之前总结的每一年第四季到第二年第一季的一个下降趋势集中在 12 月至第二年 1 月。

。

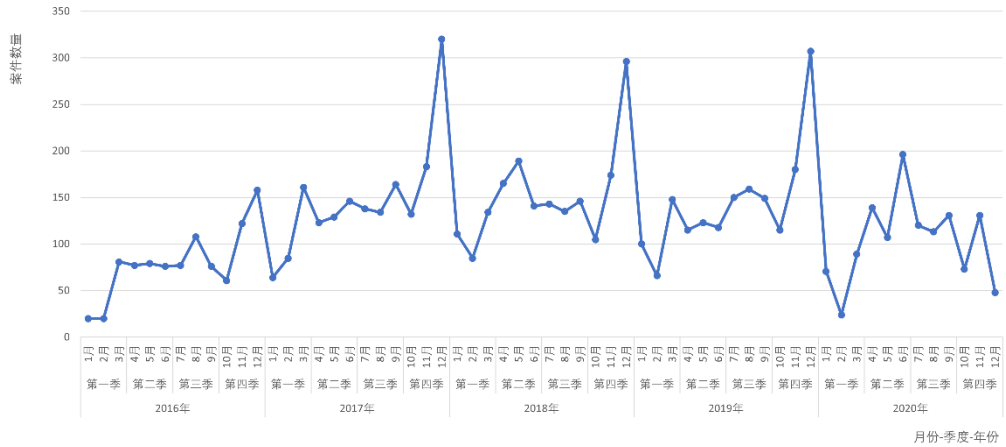


图 5　2016-2020 年四川省按月份破坏社会主义市场经济秩序罪案件数量折线图

总的来说，在时间上，四川省内 2016-2020 年破坏社会主义市场经济秩序罪的案件数量先上升后下降，且除了 2020 年以外每一年的案件数量从年初到年末递增，增长主要发生 2 至 3 月和 10 至 12 月，使得峰值都在 12 月。前一年年末到第二年年初案件数量会剧烈下降。2020 年的规律则是先上升后下降，峰值在年中，谷值在年初和年末。

4.2 空间规律分析

图 6 是依据各案件的审理法院所归属的地级单位汇总 2016 至 2020 年所有数据而成的结果，一定程度上可以认为在该地审理的案件发生在这里。我将案件按数量分为六级，并在图中也标注出了相应市州的案件数量。图中显示，案件发生集中在四川省东部盆地地区，西部高原地区分布极少，相差十分悬殊。四川西部的甘孜藏族自治州和阿坝藏族羌族自治州案件数量分别为 34 和 38，远少于其它地级单位；四川中部的省会成都市案件数量为 2382，远多于其它地级单位；四川北部的绵阳市和南充市以及四川南部的宜宾市案件数量位于 400-500 之间，多余剩余地级单位。

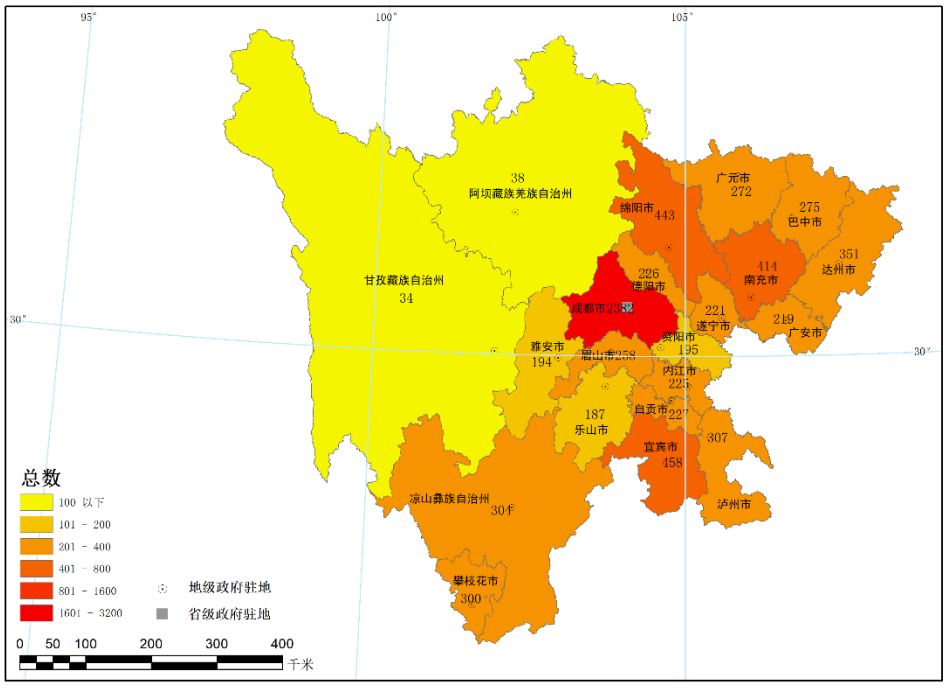


图 6　2016-2020 年四川省各市州破坏社会主义市场经济秩序罪案件数量分布

为了探究更深层次的规律，我将尺度进一步下探到县区级，根据据各案件的审理法院所归属的县区级单位汇总 2016 至 2020 年所有数据，如图 7。从图中可以进一步认识到案件发生集中在四川东部盆地地区和南部攀西地区的规律。另外一处非常明显的规律是，在地级政府驻地所在的地级市中心辖区案件数量明显多于周边县城，且省级政府驻地所在的省会中心辖区的案件数量又明显多于地级市。

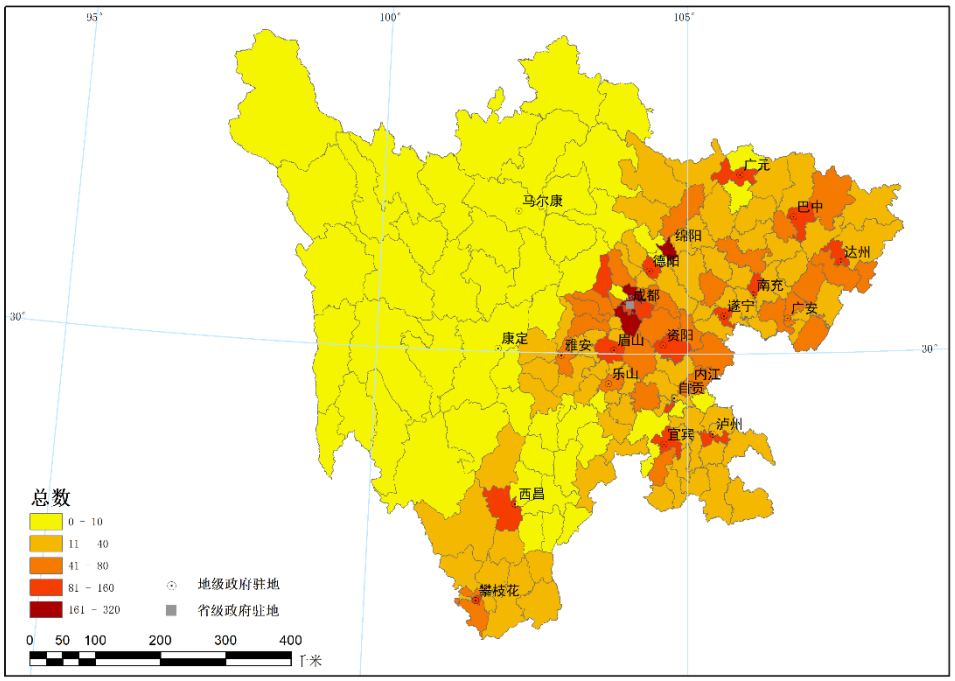


图 7　2016-2020 年四川省各县区破坏社会主义市场经济秩序罪案件数量分布

在空间方面，总而言之，四川省内 2016-2020 年破坏社会主义市场经济秩序罪的案件数量东部明显多于西部，且又聚集在地级市等较大城市辖区范围内，而这个聚集效应在成都市最为显著，案件数量较其它地市有绝对优势。

4.3 时空交互分析

4.3.1 不同空间下的时间规律分异

在【4.1 时间规律分析】的分析中，总的规律是案件数量先上升，保持平稳后再下降。这个规律可能随空间的变化而变化。因此我分地级单位统计了每一年的案件数量变化情况，如图 8。可以看见，2016 年到 2017 年案件数量的上升是全省整体性的，没有一个地级单位发生了数量下降，且上升集中在中部。2017 至 2019 年总体是保持案件数量稳定，体现在部分地市增加而部分地市减少。2019 年至 2020 年的案件数量下降则比较特别，只有成都、南充、宜宾、泸州四市数量下降，但这四市下降程度十分剧烈，使得全省总量都随之下降。

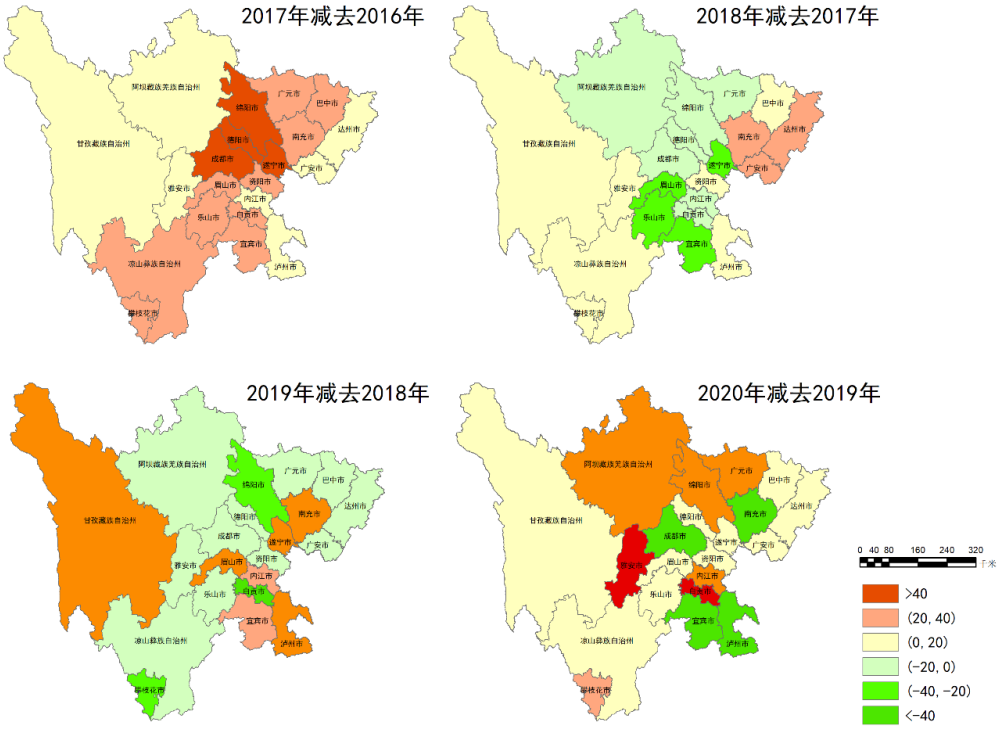


图 8　2016-2020 年四川省各市州破坏社会主义市场经济秩序罪案件变化数量分布

而进一步将空间分级到县区（图 9），就可以得知四川西部大部分县区在这五年案件数量波动其实不大，导致变化的主要是东部盆地区内的县市。无论是 2016 年到 2017 年案件数量的大量增加还是 2019 年到 2020 年案件数量的减少，主要驱动的仍然是地级/省级驻在的省内大中城市市辖区，普通县城的作用有限。此外，2019 年至 2020 年的案件数量下降规律在县区层级上于在地市层级上有些区别，起关键作用的东部单位大致都是有一个下降趋势的。

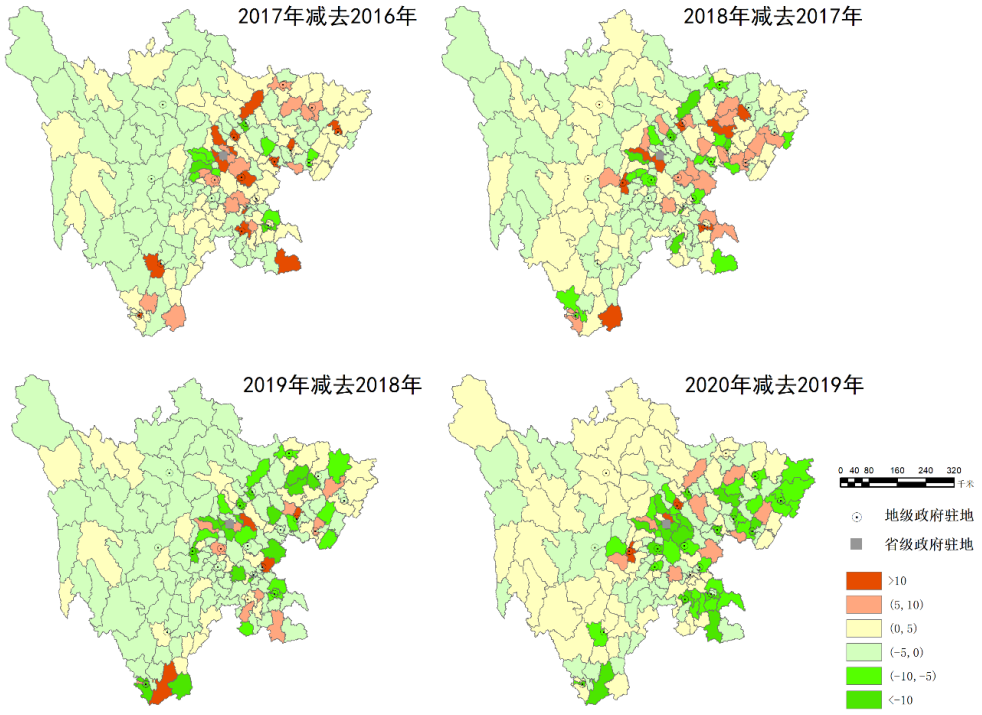


图 9　2016-2020 年四川省各县区破坏社会主义市场经济秩序罪案件变化数量分布

因此，对于整个四川省案件变化的时间规律，东部地市尤其是大中城市起决定性的作用。不同地市的时间规律在前期与全体较为一致，案件数量均为增长；中期产生分异，但都使总体保持稳定；后期分异更为严重，案件数量下降的单位占优势，使得总体的数量发生下降。

4.3.2 不同时间下的空间规律分异

图 10 是不同年份四川省各县区破坏社会主义市场经济秩序罪的案件数量分布，可以观察到，自 2016 到 2020 年，案件数量东部远多于西部的格局没有改变，且案件都保持集中在地级以上城市市辖区内，但结合图 11 可知，随着时间变化，地级以上城市案件数量占全省总数是短暂地上升后就保持下降的，说明案件发生也正在向小城市转移。

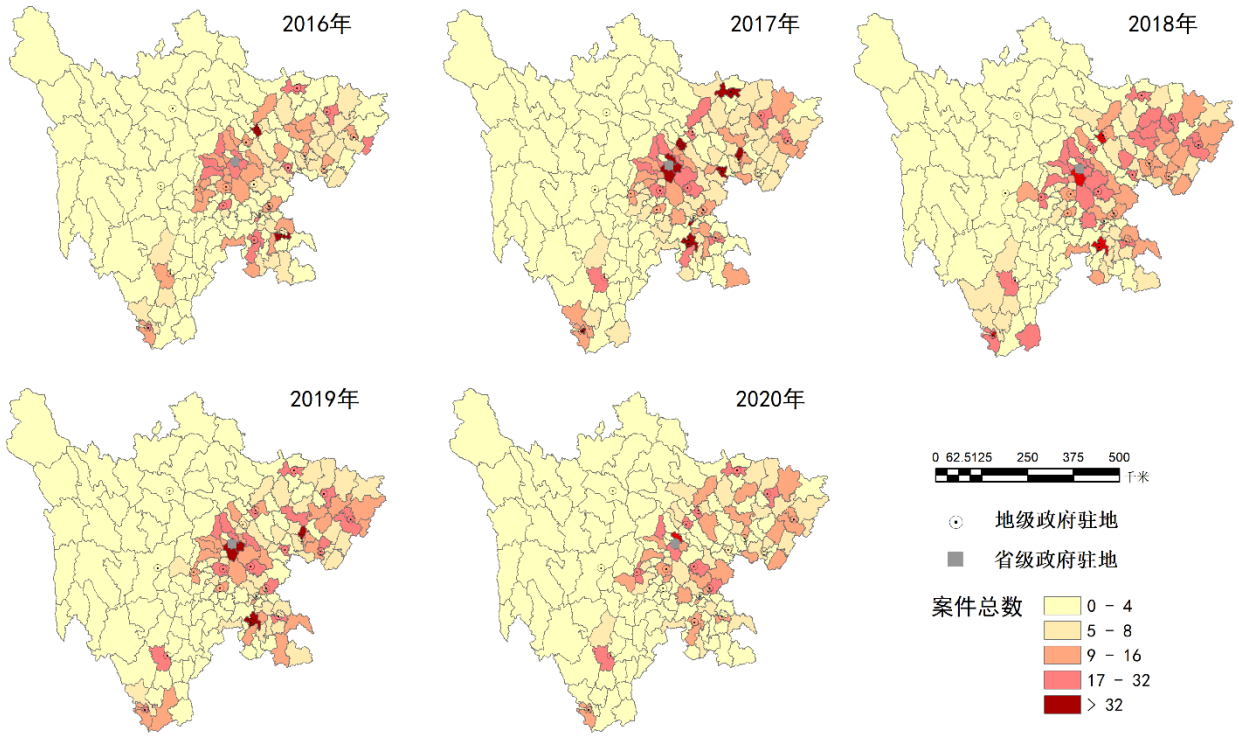


图 10　2016-2020 年四川省各县区破坏社会主义市场经济秩序罪案件变化数量分布

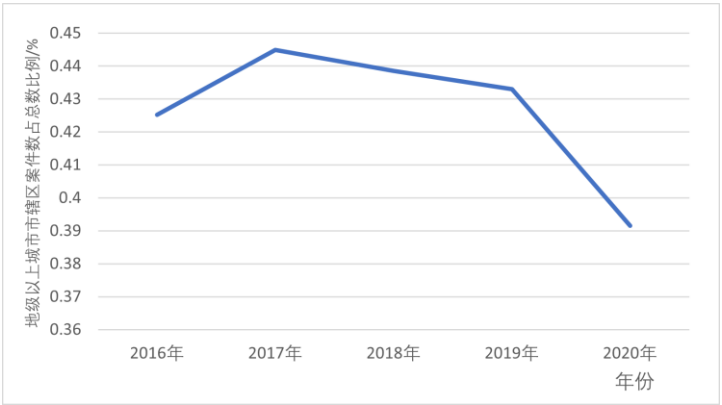


图 11　2016-2020 年四川省地级以上城市市辖区破坏社会主义市场经济秩序罪案件数量占全省总量之比

4.4 罪行和人群分析

4.4.1 罪犯人群及其属性分析

根据从文本中提取的有效年龄信息（判决时间减去出生时间），绘制分布图如图 11。图中显示，罪犯年龄分布广泛，最小 17 岁，最大 83 岁，且集中在 20 至 60 岁之间。典型的两个峰值出现在 30 岁和 45 岁左右。在 30 岁之前，随年龄增长罪犯数量增多；之后到 40 岁左右保持下降，并再上升至 44 岁的第二个峰值；最后一路下降，到 60 岁左右开始维持较低水平。

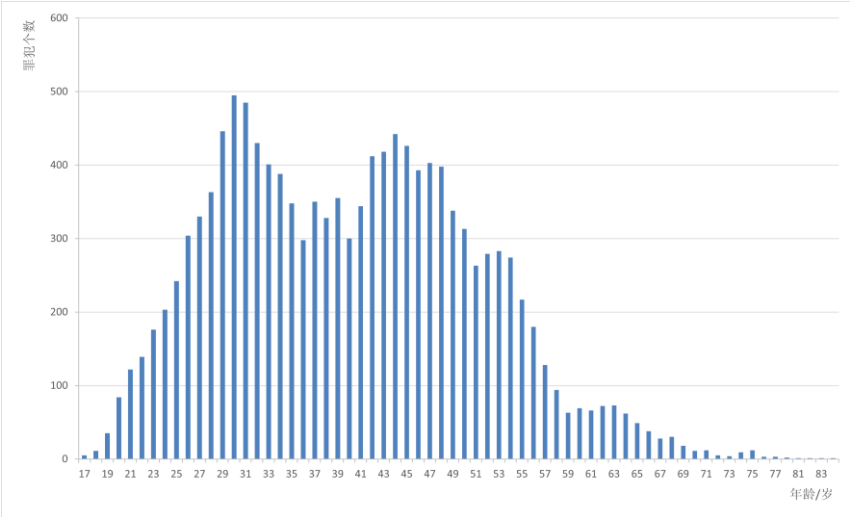


图 11　2016-2020 年四川省破坏社会主义市场经济秩序罪罪犯年龄分布

接下来由提取的有效性别信息（图 12）可知，犯罪人数男性远多于女性，比例接近 4：1。

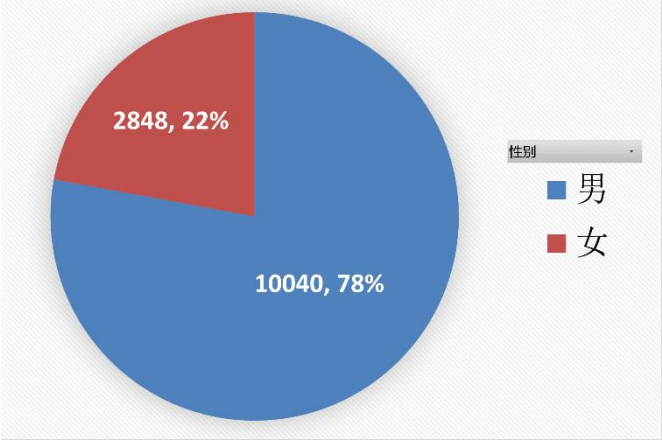


图 12　2016-2020 年四川省破坏社会主义市场经济秩序罪罪犯性别分布

将年龄与性别信息交叉进行分析，如图 13 表示随年龄变化不同性别罪犯分布绝对情况，无论是男性罪犯还是女性罪犯，都保持着以上总结的基本规律，男性罪犯的变化规模大于女性罪犯。图 14 是随年龄变化不同性别罪犯的比例，随年龄变大，女性罪犯的占比呈一个较为明显的上升趋势，且在 70 岁左右达到峰值，女性罪犯占比接近 50%。

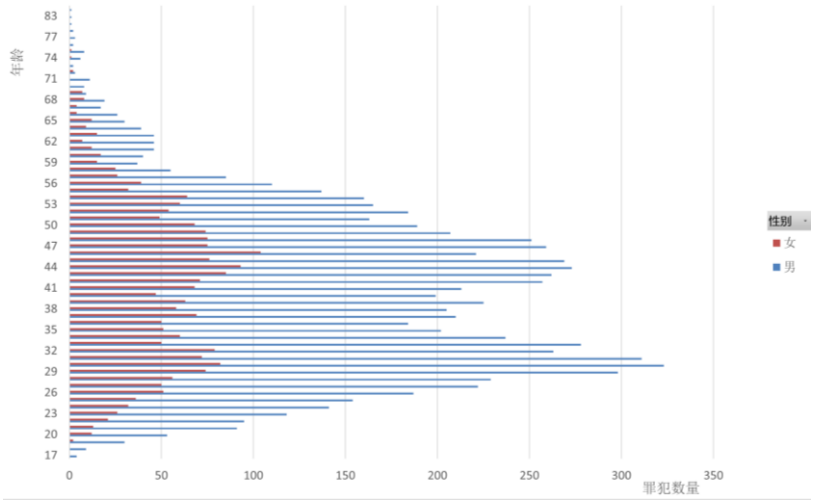


图 13　2016-2020 年四川省破坏社会主义市场经济秩序罪不同年龄罪犯性别分布

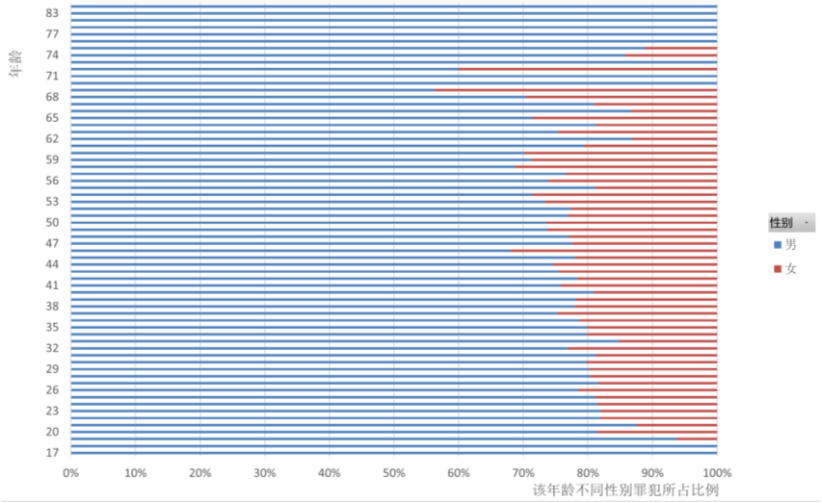


图 14　2016-2020 年四川省破坏社会主义市场经济秩序罪不同年龄罪犯性别占比

在民族构成方面，图 15 是根据所有可获取民族数据绘制的总体上的罪犯民族占比图。汉族罪犯的比例占 94%，为绝对多数。少数民族罪犯占 6%，低于中国少数民族人口比例 8.5%（2015 年），与四川少数民族人口比例 6%（2014 年）十分吻合。细分各少数民族罪犯的占比（图 16），发现彝族罪犯最多，壮族、羌族、藏族、苗族次之，排位与四川省内各少数民族数量的顺序较为一致。

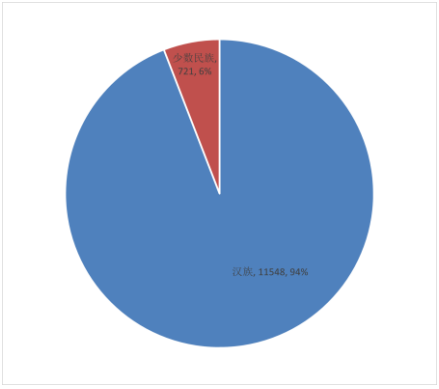


图 15　2016-2020 年四川省破坏社会主义市场经济秩序罪罪犯的民族总体分布

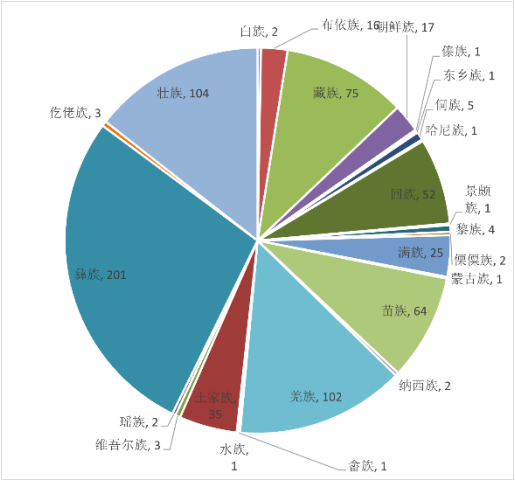


图 16　2016-2020 年四川省破坏社会主义市场经济秩序罪各少数民族罪犯分布

将年龄与民族构成进行交叉分析（图 17），可知在 20-29 岁阶段，少数民族罪犯占比明显高于其他年龄段，青年犯罪更加地集中到少数民族。而相反地，在 65 岁以上的阶段，罪犯完全由汉族构成。

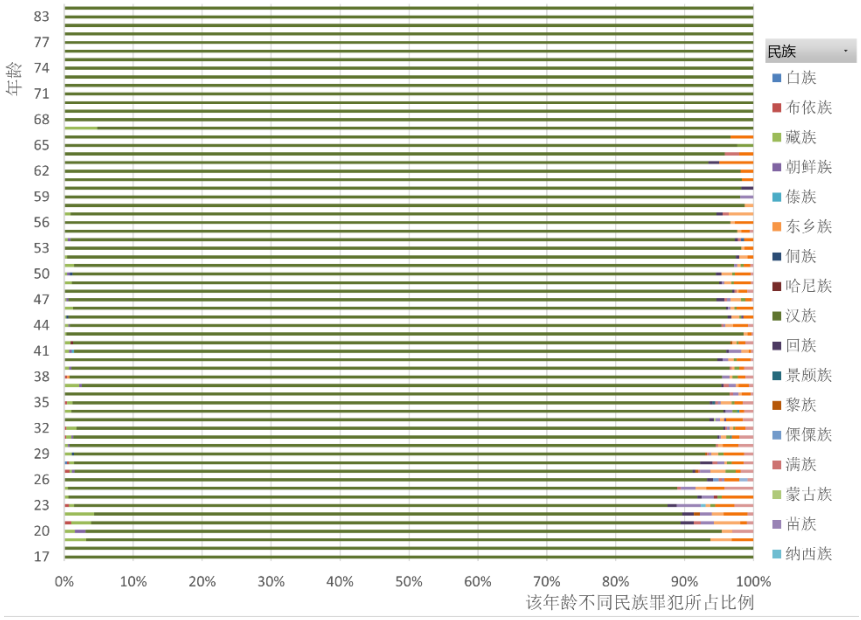


图 17　2016-2020 年四川省破坏社会主义市场经济秩序罪不同年龄罪犯民族分布

罪犯的文化程度也有很明显的分异，如图 18 所示。由学历水平从低到高排列，罪犯数量先上升后下降。文盲和研究生水平的罪犯数量都非常少，而初中水平的罪犯占很大一部分，其余文化水平中小学和高中水平的罪犯数量也成较大的规模。

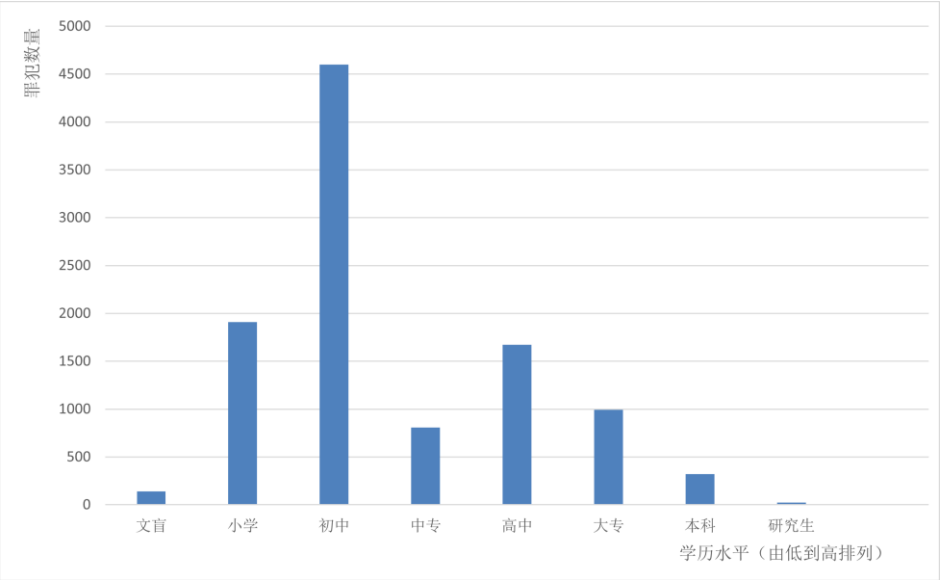


图 18　2016-2020 年四川省破坏社会主义市场经济秩序罪罪犯文化程度分布

罪犯的文化程度分布随年龄的变化也很明显（图 19）。抛开受数据量影响的 70 岁以上阶段和尚未有获取高学历条件的 20 岁以下阶段，年龄越大，较低的文化程度（小学、文盲）占比一般越高。对应地，高文化程度（大专、本科）占比先上升，保持稳定后再下降。占比最大的初中文化程度则一直保持稳定。

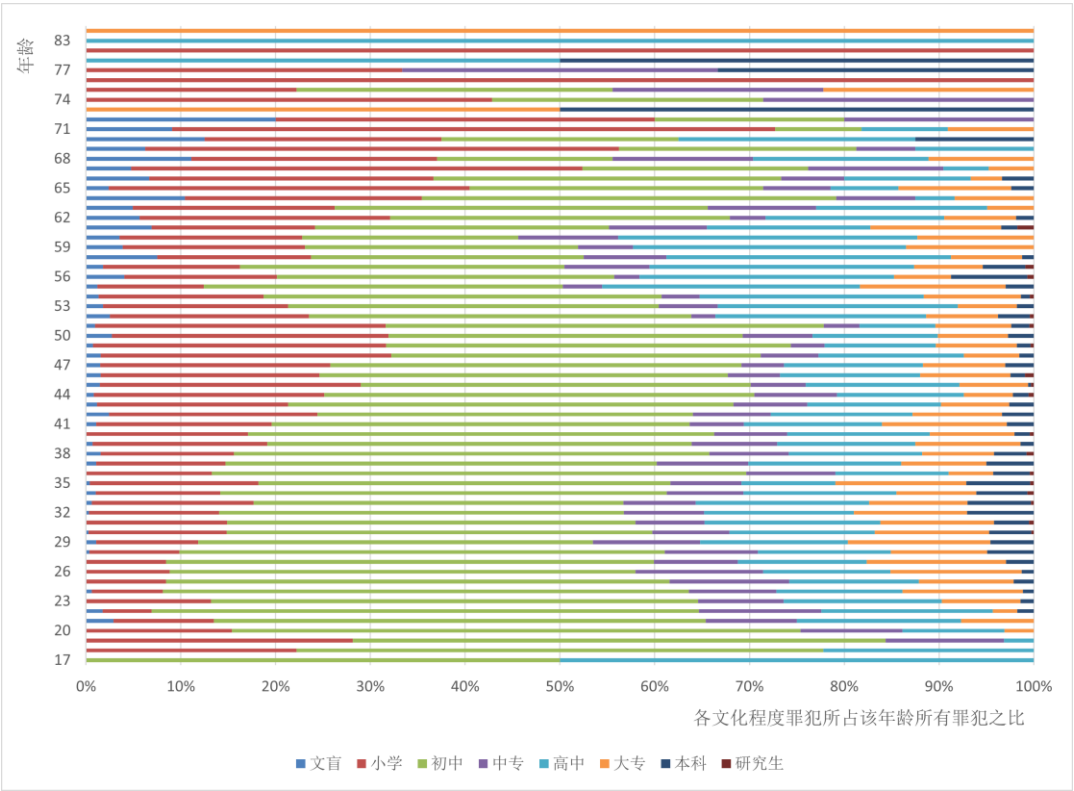


图 19　2016-2020 年四川省破坏社会主义市场经济秩序罪不同年龄罪犯文化程度分布

4.4.2 罪犯罪行及刑事责任分析

所有罪犯所犯罪行及其判刑刑期和罚金的规律都很突出。罪行方面，根据所有罪犯的罪行生成的部分高频罪行的词频云图（图 20）显示，非法经营罪、组织领导传销活动罪、非法吸收公众存款罪是最常见的一类罪行，其次是合同诈骗罪、信用卡诈骗罪、诈骗罪、强迫交易罪。相当成规模的还有虚开增值税专用发票罪、销售伪劣产品罪、假冒注册商标罪等等。

图 20　2016-2020 年四川省破坏社会主义市场经济秩序罪主要罪行词频云图

将所有的罪名按照破坏社会主义市场经济秩序罪的次级分类汇总，结果中如图 21 所示，除了走私罪类，其它七类均有涉及。其中扰乱市场秩序罪占比最大，超过 1/3，这也是出现频率最高的非法经营罪的母类；其次是金融诈骗罪和破坏金融管理秩序罪；危害税收征管罪、生产销售伪劣产品罪、侵犯知识产权罪占比在 4%至 8%，有一定分量；而妨碍对企业管理罪的占比最少，规模不足。

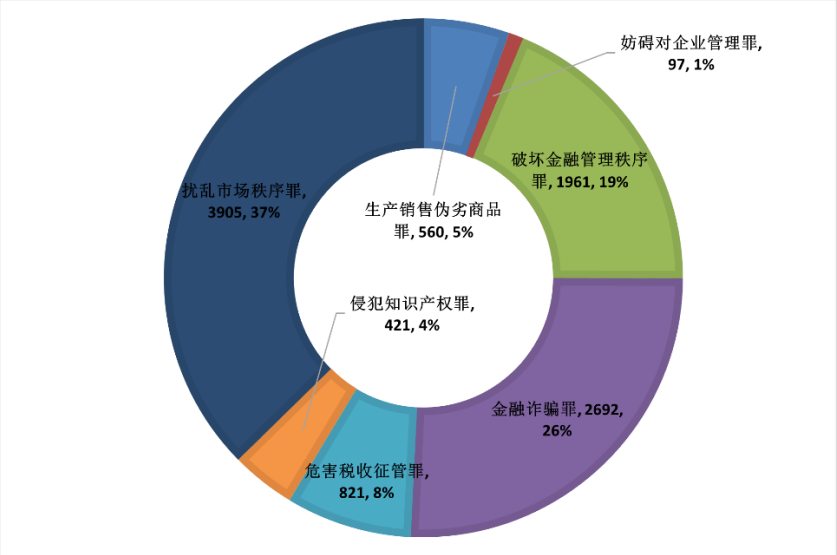


图 21　2016-2020 年四川省破坏社会主义市场经济秩序罪次级分类占比

对于罪犯的刑期来说，有效信息中判管制的案件共 2 件，拘役共 665 件，有期徒刑（缓刑）共 4759 件，有期徒刑共 6499 件。有期徒刑占比一般出头，说明量刑方面破坏社会主义市场经济秩序罪一般都是轻于其它刑事犯罪的。图 22 和图 23 是有期徒刑的刑期分布，刑期最低半年，最高 15 年，范围不长。两年以内的刑期占多数，两年到六年又占了一小半，六年乃至十年以上的刑期可以说都是相对较少的。

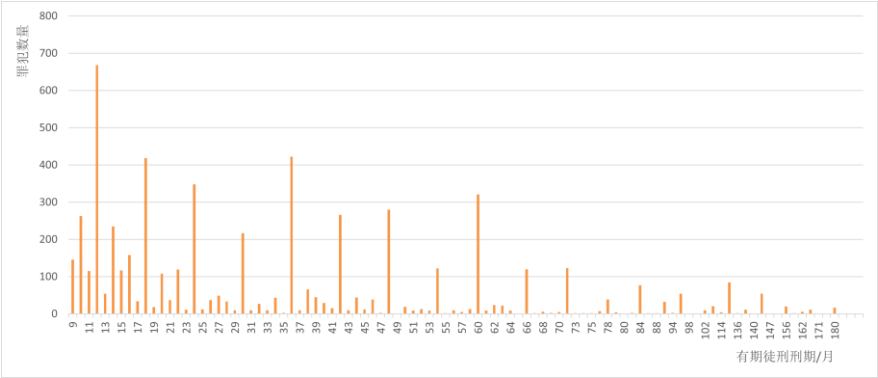


图 22　2016-2020 年四川省破坏社会主义市场经济秩序罪判有期徒刑刑案件刑期分布

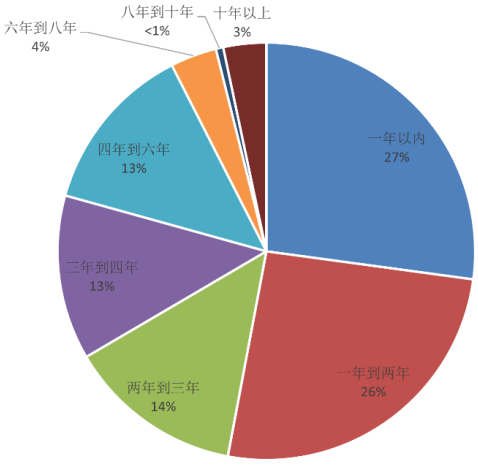


图 23　2016-2020 年四川省破坏社会主义市场经济秩序罪判有期徒刑刑案件刑期分布（2）

破坏社会主义市场经济秩序罪的案件宣布的刑事处罚中一般都包含罚金的部分，文本信息显示最低罚金为 360 元，最高为 2.8 亿，可见罚金的上限和下限都极高。根据文本信息绘制图 24，发现大半案件（51%）的罚金均位于 1-5 万这一区间，1 万以下和 5-20 万其次，分别为 19%和 22%。罚金的范围虽然延展性极强，但绝大多数情况还是处于一个较为稳定的区间。过高过低的罚金往往是个案。

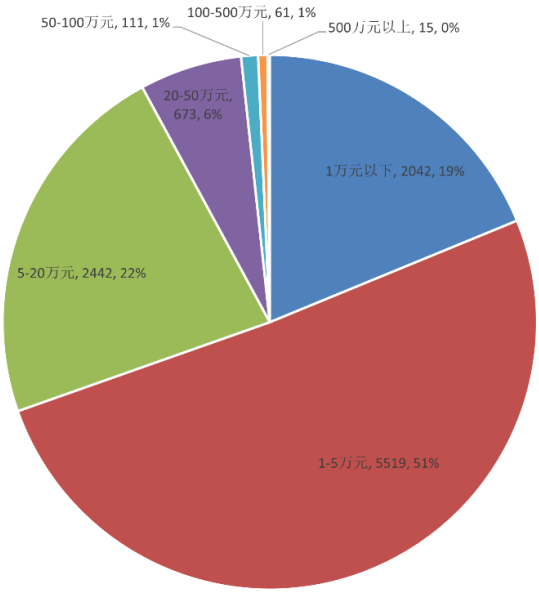


图 24　2016-2020 年四川省破坏社会主义市场经济秩序罪处罚罚金情况分布

5 讨论

5.1 成因讨论

以上种种案件相关规律的出现，一般都离不开一定的社会经济要素的推动。讨论其背后的成因，是对这些规律的进一步挖掘，也是获取经验的关键途径。

而这些案件规律的成因又脱胎于破坏社会主义经济秩序罪的成因。研究指出[4]，经济犯罪与市场经济规模、人口素质、社会治理水平、地方腐败程度等要素息息相关。图 25 是四川省各地市在 2016-2020 年的总案件数量与对应 GDP（来自四川省统计年鉴）总量关系图，呈明显的线性分布，计算得到相关系数高达 0.989，相关性十分显著，很大程度说明破坏社会主义市场经济秩序罪受到市场经济规模的强烈影响。人口素质一方面是在于群众的法制意识，文化程度较低时可能会有欠缺[14]；另一方面，部分经济犯罪需要一定的经济知识，这可能也需要稍高的文化程度。社会治理水平[5]决定犯罪监控和犯罪控制的水平，进而影响犯罪的规模。地方腐败程度可以理解为对社会治理水平的瓦解，为犯罪提供可乘之机。

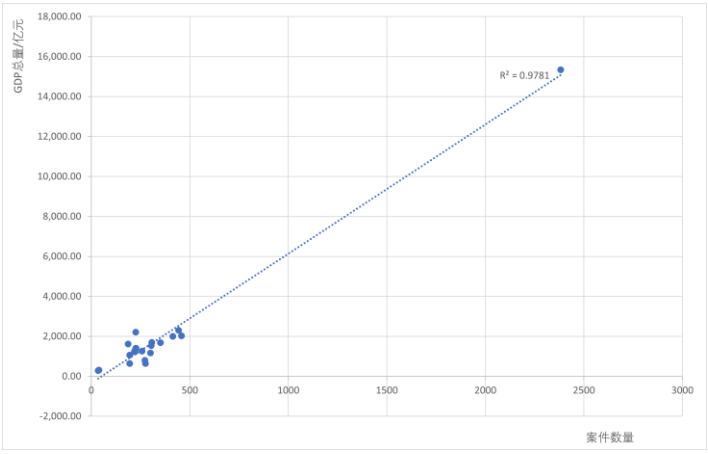


图 25　2016-2020 年四川省各地市破坏社会主义市场经济秩序罪数量与 2018 年 GDP 总量关系

基于此，探讨影响破坏社会主义市场经济秩序罪的成因的分布与变化可以更好说明相关案件的时间空间分布规律和罪行人群属性规律。

5.1.1 时空规律成因讨论

在时间上，四川省内 2016-2020 年破坏社会主义市场经济秩序罪的案件数量先上升后下降，这个数值上升应该与中国裁判文书网的文书录入情况有关,据财新网[6]，2016 年 8 月 30 日，修订后的《最高人民法院关于人民法院在互联网公布裁判文书的规定》正式发布,并于 2016 年 10 月 1 日实行。该规定进一步扩大了上网文书的范围，这之后中国裁判文书网的文书数量得到了进一步完善。与此同时，四川省的经济增长可能也会带动案件数量的上升。2020 年案件数量的下降可能跟新冠疫情对经济活动的破坏有关。另外，2018-2020 年的扫黑除恶斗争[7]可能也对案件数量下降有积极影响。

另外，大部分时间案件数量从年初到年末递增，12 月达到峰值，这一现象很可能是我国司法机构片面追求年终结案率造成的[8]。结案率是法院内部评价审判工作的主要指标，法院系统会赶在年底尽可能审结大量案件，并停止建立新案件，因此第二年新的一季度审结的案件也就相当有限。

综上，在时间规律方面，起更多影响的还是在案件审理程序和文书发布程序这一部分上。

而在空间上，案件数量东部明显多于西部的规律跟四川省的经济格局密不可分。图 25 的线性对应关系也可以说明这一点。据四川省统计年鉴，2018 年四川西部所属甘孜州、阿坝州、凉山州、攀枝花市的 GDP 总量之和仅为 3304 亿，占全省总量的 7.88%。人口分布也可以进行补充说明：上述四个西部地级单位 2018 年的常住人口为 828.4 万，仅占 9.93%。经济规模不足使得经济活动受限，经济犯罪相对减少，加上人口的不足，案件数量自然就下降。

而案件集中在地级行政以上单位市辖区的规律也可能是因为四川省的经济要素过分地集中在地级市的市辖区。四川城镇的空间是典型的单中心、离散的点状空间结构，这种空间结构制约了县域经济的发展[9]，经济要素向更高一级流动，县城经济普遍较弱，因此提供的所谓经济犯罪的机会也就对应较少。最高等级的成都市经济规模远超其它城市，其案件数量因此也占据了绝对优势。而时空结合分析得出的相关犯罪向小城市转移的结论，可能是因为相比小城市，大中城市的治理水平得到了更好的完善，贪腐数量有效减少，市民素质提高，因此这些案件的发生更有可能向虽然经济不强，但其他方面更弱的小城市转移。

所以空间方面的规律的原因可能集中在经济格局的这一块，涉及到时空交叉方面更多地在于多因素的相互影响。

5.1.2 罪行规律成因讨论

在罪行方面，非法吸收公共存款罪数量最多，主要在于大量中小企业缘于贷款难而不得不走非法吸收公众存款的道路。与此同时,非法吸收公众存款罪本身也存在不确定性，法律界限模糊，判罪容易[10]。同样数量很多的非法经营罪关键之处可能也在其适用范围非常广泛[11]，涉及到非法出版有害期刊，未经授权经营食盐或烟草类制品，非法经营保险业务，非法经营电信类业务，互联网业务违法经营，彩票违法经营，非上市公司股票违法经营，违反国家规定、非法使用终端 POS 机进行现金交易或者套现活动，擅自发行基金份额募集基金非法生产、销售赌博机或者其专用软件等等[12]。学者认为刑法典第 225 条第（4）项为非法经营罪设置了一种高度抽象的空白罪状,这为非法经营罪的扩张留下了巨大的自由空间，使之成为了一个新的"口袋罪"[13]。诈骗罪规模庞大应该也是同理。而四川省社会治理水平低于发达水平、人口素质不够高的现实也促进了组织领导传销活动罪、虚开增值税专用发票罪的增长，此类犯罪存在手段隐蔽、过程复杂、牵涉关系多、证据难以收集、犯罪隐蔽等[14]特征，适合生长在四川省这类人口众多、发展治理水平有待提高、本土各方势力盘根错节的地方。

在破坏社会主义经济秩序罪的八个次级分类中，走私罪没有出现可能因为四川省基本不具备边贸条件，也就基本不会发生走私时间。扰乱市场秩序罪因为包含了非法经营罪、组织领导传销罪这两个数量意义上的大罪，具有适用范围广泛的特征，占比最大。金融诈骗罪和破坏金融管理秩序罪占比居于其次一部分可能是罪行适用范围的原因，一部分也是与四川省的特征高度匹配的原因。

因此，在罪行这一部分，我认为两个因素的影响最大，首先是罪行适用范围广泛与否，其次是区域特征。

5.1.3 罪犯人群属性规律成因讨论

由于我认为破坏社会主义市场经济秩序罪跟人口的素质有关，那么罪犯文化程度的分布规律也可以在这个层面进行解释。研究指出[15]，文化程度较低，法制意识相对更加淡薄，犯罪的几率越高；而部分经济犯罪具有文化程度的门槛，不能完全没有文化。这样综合下来，在某种维度上可以理解为什么初中水平的罪犯数量最多，而小学、中专、高中次之。不同年龄段的人群由于本身的文化水平不一致，老一辈大多数文化程度较低，新一代相对较高，因此对应不同年龄段的罪犯产生相应文化程度上的差异。

而更进一步，罪犯文化程度在不同年龄段的差异又可以说明罪犯的年龄分布。由于全民教育的普及，在一定年龄段之前，年龄较小的人群中文化程度随年纪增大而增大，逐渐有了犯罪的文化基础，在高技术含量经济犯罪中出现得越来越频繁，这可能就是 30 岁左右的第一个峰值的成因；然后在一定年龄段之后，文化程度随年纪增大而降低，高技术含量经济犯罪减少，而因为法律意识淡薄引起的低技术含量经济犯罪增多，因此案件数量随年龄下降后又达到第二个峰值。而随年龄再次上升，由于生理上更加衰弱[16]、物质上相对富足等方面的原因，犯罪概率降低。

性别方面，女性犯罪远少于男性，是多种因素的共同作用。心理学研究[17]表明，女性参与社会生活的深度和广度不及男性，对公共生活的关注程度逊于男性。而且，某些体质和心理的因素也限制了女性的犯罪活动，女性在扮演自身社会角色的过程中对自身的欲望和冲动有更好的控制。但女性更容易受到情感因素的驱动，随着年龄增长女性罪犯的占比越来越大的规律就可能是因为部分女性中年时期因离异、家庭负担等原因[18]所受到了较多情感因素的影响，中年女性因感情原因造成的个人性格和社会关系的中断，很可能就会诱导其实施犯罪行为[19]。

民族构成方面，少数民族罪犯与四川数民族罪犯占比十分相近，证明人口基数起到的重要影响。而对于青年犯罪更加地集中到少数民族这一现象，有学者[20]认为在汉族青年对相关法规更加适应的背景下，少数民族青年仍然更容易依照其成长过程中遵循的本民族的习惯法来行事，对制定法的变化感受迟缓，犯法而不自知。同时，少数民族地区经济相对落后，会有大量青年外出务工，成为流动人口，在生活上缺少保障，不少人走上犯罪道路[21]。这可以说是地区经济规模、人口素质、社会治理水平的共同作用引发的后果。

5.2 方法评价

本文基于数据爬取、文本自动提取的相关手段实现了 2016-2020 年四川省所有基层法院判决的破坏社会主义市场经济秩序罪的案件的分析，具有较好的效率和概括性、合理性。数据爬取解决了数据的来源问题，迅速、快捷、省事地获取了研究所需的全部文书与概要，提供的充足的信息支撑。文本自动提取进一步抓取了我想要关注的信息，有效利用了数据爬取获得的大量信息，相对于传统的精读文本的分析模式可以提供更多维的分析模式，以正则提取为基础的文本自动提取也使信息变得标准化，提升了概括性。而合理性方面，本文评估的结果基本都对应了比较客观的成因，基本符合相关事实，在其它科学研究上具有一定参考性。

5.3 研究不足

本文的自动化手段以及分析方法仍然存在一些问题。首先，在数据源方面，为了能够以县区为单位进行分析，我提取的都是基层法院的判决书，而忽略了相对高层的法院，四川省的案件全貌并不完整。且中国裁判文书网也存在判决书不能完全录入到系统的问题。第二，在自动化手段方面，我使用的办法仍然是不完整、效果仍有很大进步空间的自动化，仍然会受到网站只显示 600 份文书的限制而要将一个自动化过程分割为十几个。文本提取时构建的正则也存在信息缺失或者不能完全提取的问题，获得数据后还要进行一定的补充修复工作。第三，分析过程中，囿于相关条件限制，我以单要素分布规律的分析为主，最多结合两个要素的分析，难以获得这些案件的一个综合特征，在揭示完整的规律时有较大苦难。因此，在未来的探索在需要综合更多的数据和分析手段。

6 结论与展望

本文基于自动化提取手段，分析了 2016-2020 年四川省基层法院的破坏社会主义市场经济秩序罪的所有判决书。主要结论是：时间上，2016-2020 年四川省破坏社会主义市场经济罪数量先上升后下降，每一年案件数量从年初到年末递增，这基本取决于案件审理程序和文书发布程序。空间上案件数量东部明显西部，并聚集在地级市等较大城市辖区范围内，而这个聚集效应在成都市最为显著，主要可能是受到四川省经济格局的影响。而对于整个四川省案件变化的时间规律，东部地市尤其是大中城市起决定性的作用，但近年来呈案件向小城市转移的趋势。在罪行方面，扰乱市场经济秩序罪规模最大，其次是金融诈骗罪和破坏金融管理秩序罪，我认为这是罪行适用范围和四川省地域特征导致的综合结果。破坏社会主义市场经济市场经济秩序罪刑期不长，大部分在无期徒刑五年以内，且基本都具有罚金的处罚，一般不超过人民币 20 万元。而在犯罪人群方面，犯罪者年龄以在 30 和 45 岁左右最多，文化程度集中在初中水平，中专、小学、高中次之，年龄和文化程度两个因素内在也有一定能关联。男性犯罪远多于女性，但随年龄增长，女性犯罪占比上升。少数民族犯罪比例与其人口比例基本一致，青年犯罪逐渐向少数民族集中。

基于研究结论，未来在破坏社会主义市场经济秩序罪的控制方面应该要注重经济活动规模较大的地区内文化程度较低的人群的教育与关注，且要注重空间内的全面发展，就近治理问题，减少犯罪的流动情况。罪行的划分上要更加精确，制定更为合理的惩治准则。最本质的还是要解决有游离于社会规则之外迹象人群的适宜性问题。而对于未来的研究中，互联网文书系统可以更加完善，不同的研究机构可以推进数据间的共享，在深度学习越来越进入人们视野的背景下可以以此推动自动化水平的进步，特别是现在还难以做到的自动化分析的层次。研究和实践之间也需要更加实时的互相反馈，更加科学地促进社会向好向善发展。

7 参考文献

[1] 孙丽萍. 破坏社会主义市场经济秩序罪的过去、现在和未来[D].中国政法大学,2010.

[2] 李欣竹. 盗窃类案件时空特征及成因分析与预测[D].中国人民公安大学,2020.

[3] 刘玲,李钢,薛淑艳,马雪瑶,周俊俊,徐婷婷,王皎贝.四川省拐卖儿童犯罪的时空演变过程及影响因素分析[J].地理科学进展,2020,39(05):853-865.

[4] 郭文才,孙光焰.关于经济犯罪原因的思考[J].武汉交通科技大学学报(哲学社会科学版),1995(01):41-45.

[5] 赵海燕,漆泽民.社会治理体系和治理能力现代化背景下网络犯罪防控研究[J].法制与社会,2020(28):119-120.

[6] 财新网.中国裁判文书网访问量破百亿 公开文书逾 3200 万篇[EB/OL]. http://china.caixin.com/2017-08-24/101134967.html, 2017-08-24/2020-02-02.

[7] 彭新林.扫黑除恶的时代意义和法治内涵[J].紫光阁,2018(03):61-62.

[8] 姜大伟.困境与进路:年终法院立案难现象之反思[J].西部法学评论,2013(06):87-91.

[9] 贺泽凯,戴宾.四川县域空间结构及其增长极的培育[J].西南民族学院学报(哲学社会科学版),2003(05):103-105.

[10] 姜首领. 非法吸收公众存款罪相关问题研究[D].中国政法大学,2009.

[11] 刘靖宇,任志贤.非法经营罪的适用探讨[J].山西省政法管理干部学院学报,2020,33(04):81-85.

[12] 张明楷. 刑法学(下册)[M]. 北京:法律出版社，016.

[13] 徐松林.我国刑法应取消“非法经营罪”[J].法学家,2003(06):111-120.

[14] 杨曼.论危害税收征管犯罪案件侦查难点及对策[J].企业科技与发展,2018(03):173-174.

[15] 刘小平.从农民工思想状况视角看教育公平的重要性——以武汉市农民工为例[J].法制与社会,2008(22):277-278.

[16] 郭晓红.当代老年犯罪研究[M].北京:中国政法大学出版社, 2011.

[17] 张远煌.论性别对犯罪的影响[J].刑侦研究,1998(04):16-19.

[18] 王方.女性职务犯罪的特点及预防—以北京市朝阳区人民检察院近两年办理的案件为例[J].中国犯罪学学会年第十八届学术研讨会论文集(下册),2009:1061—1064.

[19] 熊谋林,江立华,陈树娇.生命周期研究:性别、年龄与犯罪[J].青少年犯罪问题,2013(01):75-82.

[20] 全波. 我国少数民族青少年犯罪及其预防对策[D].中央民族大学,2009.

[21] 薛智辉,陈宁.少数民族青少年犯罪与社会控制[J].法制与社会,2008(04):58-59.

8 附录

8.1 案件数据自动下载具体步骤

8.1.1 基本信息爬取

在开始我的工作之前，需要对网站进行一定的分析。首先发现在网页源代码中关键文本被加密，另外，更换页码/进行条件筛选时网页的 url 完全不发生变化，最为关键的一点，网站每次只加载 600 条结果，其余符合条件的结果遭到隐藏。

（如图）

因此如果使用 python 爬取数据则会面临反爬虫措施带来的大量困难，我便选择了 web scraper 工具进行操作。只要浏览器中显示了信息，该工具就可以获取，解决了加密问题。其工作逻辑是建立一个 sitemap 路线图，按路线图的逻辑顺序获取 html 中的每个子模块的内容。针对更换页码时网页 url 不变的问题，使用该工具的动态加载翻页功能（模拟点击页码模块）可以解决。最后，最大的挑战还是网站只加载 600 条结果，解决方案是分地级市搜索并爬取内容。四川省内除成都以外的地级市文书都不足 600 份，可以轻松获得。至于成都市，我进一步细分为分 5 年获取。

下面是获得某一组基本信息的操作步骤：

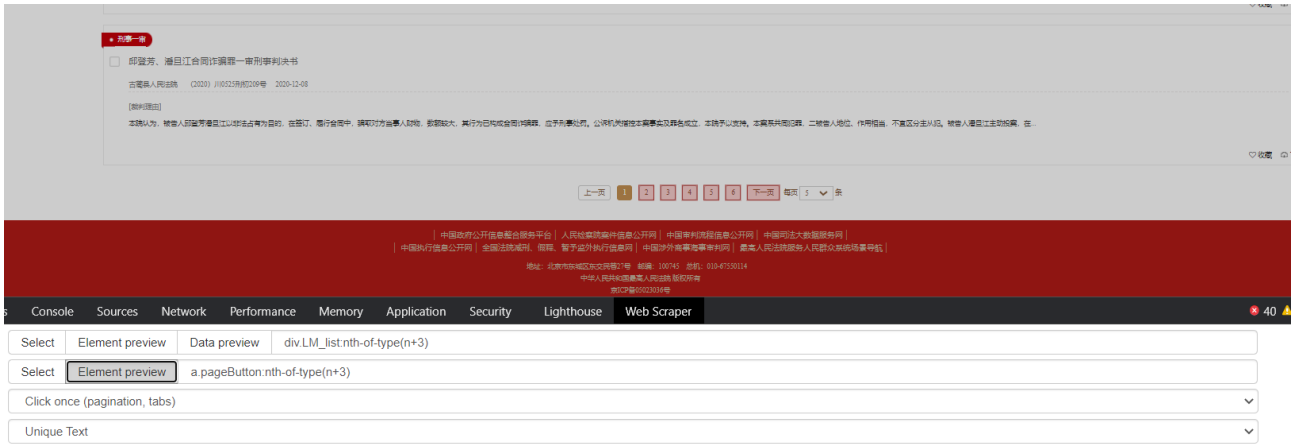
首先按相关条件筛选，根据案号筛选出某一个地级市的所有文书。（川 05 为泸州市）

可以看见，一共有 414 份文书。

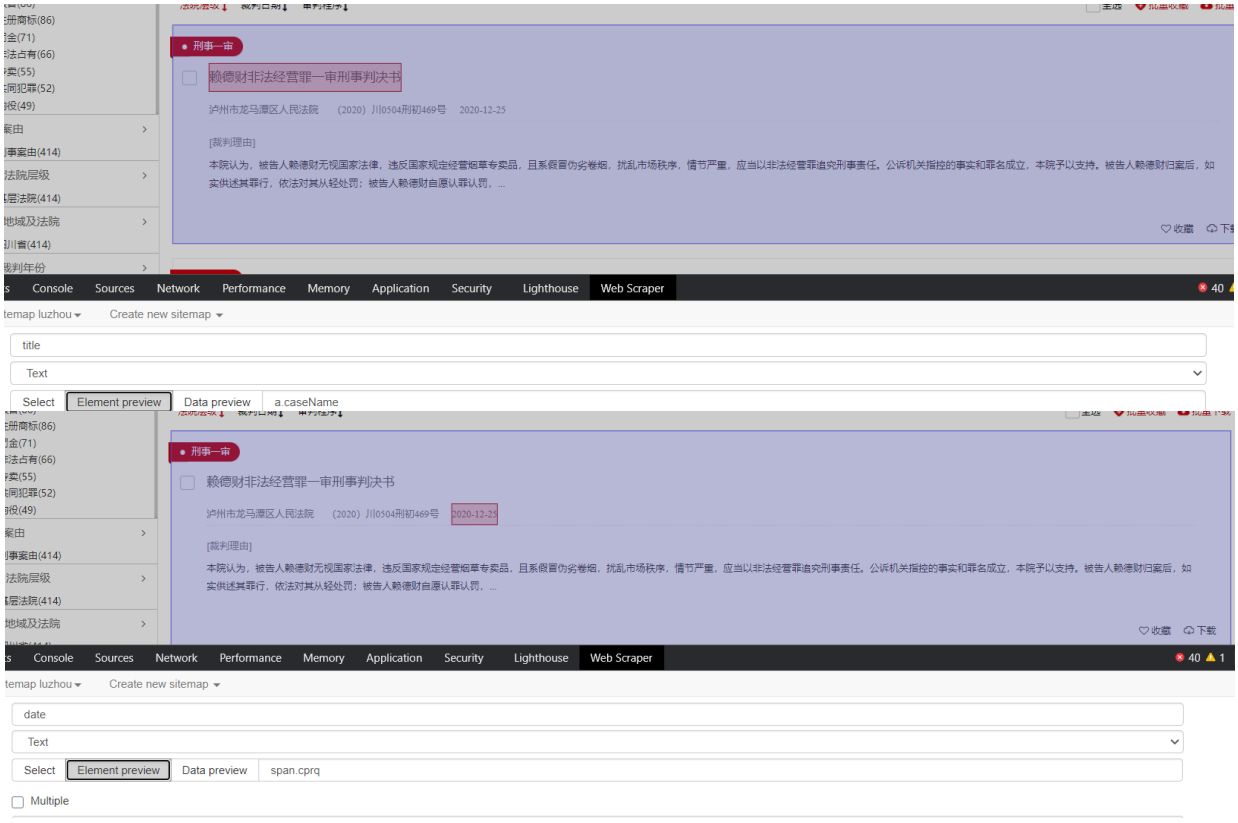
在工具中新建一个 sitemap，输入起始页 url：

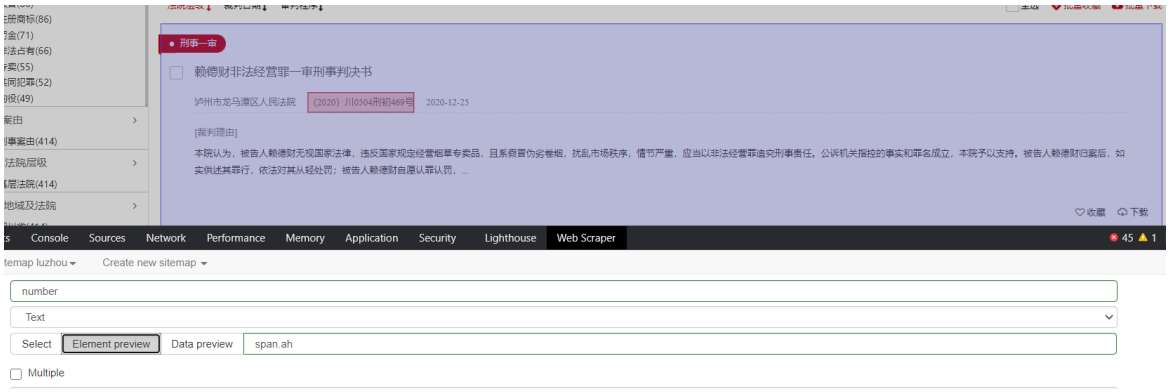
接着建立一个子结点，选定为 element click 模式（实现动态加载翻页功能），选定所有的案件信息子模块：

在 click selector 中选择所有的页码子模块，实现自动翻页

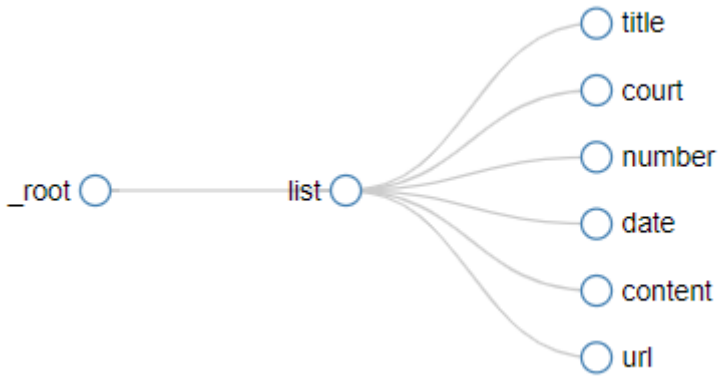


在该结点下建立多个子结点获取标题 • 文书链接 • 案号 • 文书发布时间 • 法院名 • 裁判理由等信息（下图为步骤节选）：





获得的 sitemap 路径图如下：



点击 scrape 开始爬取，获得结果导出为 csv，结果节选如下：

web-scraper-order	web-scraper-start-url	title	court	number	date	content	url	url-href
1610524560-1879	https://wenshu.court.gc	郭都都非法吸收	四川省泸	(2016)	2016/6/21	本院认为	郭都都非	https://wenshu.court.gov.cn/website/wenshu/181107ANFZ0BXSK4/index.html?docId=ac61272a2abe49548c0ca84400c83a7d
1610524560-1829	https://wenshu.court.gc	江达文合同诈骗	四川省泸	(2017)	2017/5/23	本院认为	江达文合	https://wenshu.court.gov.cn/website/wenshu/181107ANFZ0BXSK4/index.html?docId=95bab65205e643a5ad46a84700eced1f
1610524560-1735	https://wenshu.court.gc	李强、牟雪梅非	泸州市龙	(2018)	2018/9/1	根据《最	李强、牟	https://wenshu.court.gov.cn/website/wenshu/181107ANFZ0BXSK4/index.html?docId=b79917cab32b40e98a24a96900d86348
1610524560-1784	https://wenshu.court.gc	金鑫小贷公司及	泸州市江	(2016)	2018/1/17	综上，本	金鑫小贷	https://wenshu.court.gov.cn/website/wenshu/181107ANFZ0BXSK4/index.html?docId=599adec250fc4d9882a2a9c000cef4bc
1610524560-1606	https://wenshu.court.gc	四川兴远翔海科	泸州市江	(2019)	2020/6/22	综上，本	四川兴远	https://wenshu.court.gov.cn/website/wenshu/181107ANFZ0BXSK4/index.html?docId=1c9cd00ae6cf42b3afa8ac3e017d3517

更改案号重新进行筛选，获取别的地区文书的起始 url，输入到已经建立的 sitemap 中，就可以获取其余的文书的基本信息了。

8.1.2 文书自动下载

在 8.1.1 中我已经获取了每个文书的一些基本信息和其 url。但这个 url 并不是文书的下载地址，我需要通过这个 url 内的 docId 进行构造，在其前面加上 http://wenshu.court.gov.cn/down/one?。比如说，有一份文书的 url 链接为：https://wenshu.court.gov.cn/website/wenshu/181107ANFZ0BXSK4/index.html?docId=ac61272a2abe49548c0ca84400c83a7d 我需要提取其 docId 部分：docId=ac61272a2abe49548c0ca84400c83a7d，在前面加上 http://wenshu.court.gov.cn/down/one?，最终获取其下载链接：http://wenshu.court.gov.cn/down/one?docId=b2a02cf3cadd42d796a167010e02214a。

然而通过各种测试，这个链接只有在输入浏览器后才能实现下载。因此为了实现下载的自动化，我在 python 中引入 pyautogui 模块来实现鼠标键盘的模拟操作。

如下图所示，先构建一个自动操作函数，使用 moveTo 定位到屏幕某一点，然后使用 click 函数模拟点击这里，typewrite 函数输入下载链接，最后 press 函数模拟点击 enter 键。

```
def input_id(url): #自动下载
    pyautogui.moveTo(180, 60, duration=0.2)  #定位到某一点
    pyautogui.click(button='left') #点击
    time.sleep(0.5)
    pyautogui.typewrite(url,0.01) #输入字符，0.01表示输入每个字符间隔的时
    time.sleep(0.5)
    pyautogui.press("enter")  #点击确定
```

在主函数方面，利用 pandas 读取 csv 到 dataframe 中，循环获取表中 url 一栏，使用 replace 函数替换相应部分形成下载链接，代入自动操作函数即可实现自动下载，获取 doc 格式的文书。

```
a=pd.read_csv('概要/泸州.csv')    #读取文档
row=a.shape[0]                   #循环获取url
for i in range(row):
    w=a.iloc[i,8]
    w=w.replace("website/wenshu/181107ANFZ0BXSK4/index.html", "down/one", 1)  #替换
    input_id(w)
```

8.2 文本自动提取具体步骤

8.2.1 格式转换

对于我获得的 doc 文档，python 不能很好地解析，而如果将 doc 转化为 docx，python 就可以利用 docx 模块实现标准、高效的读取。

首先在 python 中引入 win32com 模块，使用 client 内的 Dispatch 调出 word 接口，并打开 doc 文件。通过原文件名构建 docx 后缀的新文件名，并使用 SaveAs 函数完成转换。最后关闭文件，推出接口，删除原文件。

```
def doc2docx(path):      #转化函数
    w = client.Dispatch('Word.Application')  #word接口
    doc = w.Documents.Open(path)  #打开
    newpath = os.path.splitext(path)[0] + '.docx'
    doc.SaveAs(newpath, 12, False, "", True, "", False, False, False, False)  #转化
    doc.Close()  #关闭
    w.Quit()
    os.remove(path)
    return newpath
```

为了实现对所有文件的批处理，需要引入 os 模块，使用 walk 函数可以获取一个文件夹下的所有对象的名字 • 绝对路径前缀，通过组合文件名和绝对路径前缀获得文件的绝对路径，代入转换函数，就可以实现批量自动转换了。

```
g = os.walk(r"C:\Users\risen\Desktop\FIN\GISSA大\文书") #浏览
for path, dir_list, file_list in g:  #循环
    for file_name in file_list:
        try:
            a = os.path.join(path, file_name) #获得文件名
            print(a)
            doc2docx(a)  #转化
        except:
            print("oops")
            continue
```

8.2.2 正则构建

根据文本结构构建正则可以实现对文本关键信息的自动·批量提取。由于 docx 模块读取后文书每一段构成了一个字符串，因此我也以段为单位来发掘特征，构建正则，方便接下来的循环匹配。阅读了部分文书发现：

每一个文书总有一段表示案号，构建正则"(20..).*川.*刑.*号"即可提取判决年份

（2020）川 3221 刑初 51 号

每个文书也会有一段专门介绍罪犯基本信息，基本都会披露姓名-性别-出生年-民族-文化信息，其中姓名-性别-出生年顺序非常固定，分割也非常清楚，可以构建正则"被告人(.*)，([男女]).*(19..)年"获取，其中出生年可能会受到其它数字干扰，考虑到刑事犯罪罪犯一定大于 14 岁，且文本中的干扰数字基本都大于 2010，我在大部分情况使用“19”匹配，匹配不成功则使用“200”匹配（正则："被告人(.*)，([男女]).*(200.)年"），将误差减弱到最低。

被告人阿旺群彭，男，1983 年 12 月 7 日出生，藏族，小学文化，经商，户籍所在地四川省阿坝县，住阿坝县。因涉嫌犯资助危害国家安全罪，于 2020 年 7 月 25 日被阿坝县公安局刑事拘留，经阿坝县人民检察院批准，于 2020 年 9 月 1 日被逮捕，现羁押于阿坝县看守所。

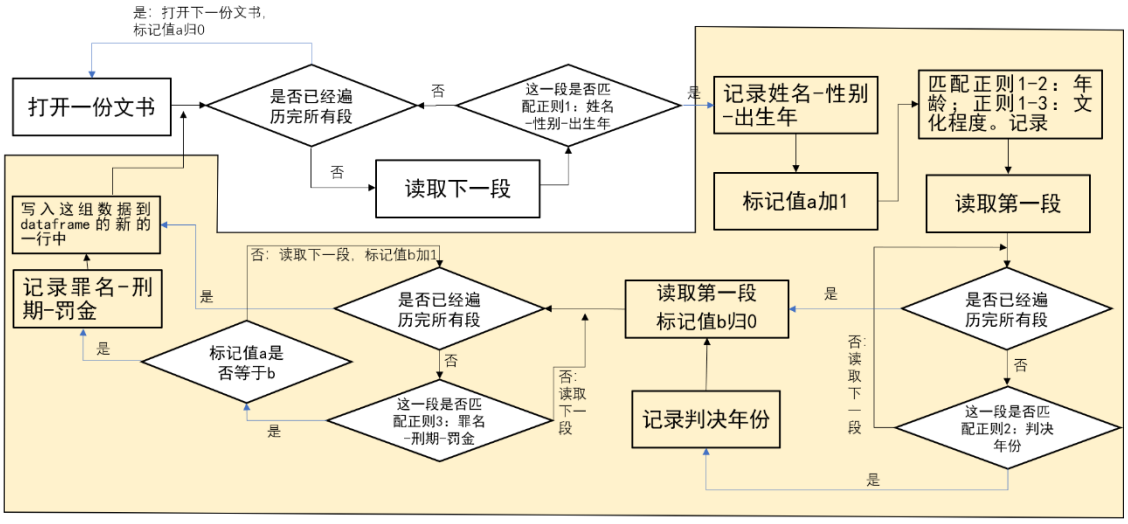
因为民族和文化信息顺序不固定，因此需要分别写入正则匹配，在某一段确定姓名-性别-出生年结果存在后，分别匹配正则"被告人.*(.{1,4}族)"和"被告人.*(..)文化"来获取，文化上存在文本是“文盲”的可能性，因此使用正则"(文盲)"来辅助判定。

另外，在文书后半部分，也会有专门的一段来写判决结果，披露罪名-刑期-罚金信息，可以依据 被告人/犯/判处/并处罚金（人民币） 进行分割，构建正则"被告人.*犯(.*罪)，判处(.{2,20})[；|，]并处罚金人民币(.{1,10})元"则可以获取，如果文本中不含“人民币”，则可使用"被告人.*犯(.*罪)，判处(.{2,20})[；|，]并处罚金(.{1,10})元"补充判定。

一、被告人阿旺群彭犯非法经营罪，判处有期徒刑四年，
并处罚金人民币 32000 元。

通过以上正则可以提取大部分文件的相关信息，部分不符合条件的文件或者是信息不足的文件则囿于条件限制，只能战略性放弃。

8.2.3 自动提取



首先在 python 中创建一个 dataframe 来存入相关信息，并且建立一个标记值。

```
#dataframe数组
df=pd.DataFrame(columns=('名字','性别','出生年','判决年','民族','文化','判罪','刑期','罚金'))
```

我使用 docx 模块读入文书，每一段都是一个字符串，由于一份文书存在一个或多个罪犯，所以我也需要以段为单位进行姓名-性别-出生年的匹配，匹配成功的话说明发现了一个新的罪犯，现在要做的就是获取这个罪犯的所有相关信息，然后在 dataframe 内新增一行来存储，即上方流程框图中的黄色部分。具体来说，匹配成功之后第一步要记录相关信息，并且标记值要加 1，这样这个罪犯在这份文书里就有了序号：

```
for i in range(len(file.paragraphs)):  # 基本正则（正则1）
    # print("第"+str(i)+"段："+file.paragraphs[i].text)
    matchObj1 = re.search(regex1, file.paragraphs[i].text)
    matchObj12 = re.search(regex12, file.paragraphs[i].text)
    if matchObj1:
        con = con + 1
        a = matchObj1.group(1);  # print(matchObj1.group(1))
        b = matchObj1.group(2);  # print(matchObj1.group(2))
        c = matchObj1.group(3);  # print(matchObj1.group(3))
```

然后在这一段紧接着匹配正则 1-2（民族）、1-3(文化程度)，匹配成功的话记录民族和文化程度信息，不成功记录为空值。值得注意的是文化信息有文盲可能，不符合正则 1-3 标准，因此需要代入辅助正则进行补充判断。

```
matchObj2 = re.search(regex2, file.paragraphs[i].text)
if matchObj2:  # 正则1-2民族
    d = matchObj2.group(1);  # print(matchObj2.group(1))
else:
    d = '';# print('no people')
```

```
matchObj3 = re.search(regex3, file.paragraphs[i].text)
if matchObj3:  # 正则1-3文化
    e = matchObj3.group(1);  # print(matchObj3.group(1))
else:
    e = '';# print('no study')
    matchObj32 = re.search(regex32, file.paragraphs[i].text)
    if matchObj32:  # 文盲可能
        e = matchObj32.group(1);  # print(matchObj3.group(1))
```

接下来，由于我要计算罪犯年龄，除了出生年之外，我还需要判决年。需要循环每一段匹配正则 2（判决年份），匹配成功就记录判决年信息。

```
for k in range(len(file.paragraphs)):
    matchObj4 = re.search(regex4, file.paragraphs[k].text)
    if matchObj4:  # 正则2：判决年
        z = matchObj4.group(1)
        break
    else:
        z = ''
```

最后，关于罪名-刑期-罚金，仍然是新开一个 for 循环，使每一段都匹配正则 3(罪名-刑期-罚金)。由于一个文书里可能有多个罪犯，匹配成功的段落可能也有多个，但我发现，这些段落里罪犯的顺序和最开始介绍罪犯的顺序是一致的，也就是说只要我在这里也引入一个标记值，每匹配成功一次标记值加 1，只要这个标记值和最初的第一个标记值（即这个罪犯在文书中的序号）一致，就说明这个段落的罪名-刑期-罚金信息是属于我在研究的这个罪犯。同时也记录下相关信息。

```
con2 = 1  # 标记
for j in range(len(file.paragraphs)):  # 正则3：罪名-刑期-罚金
    matchObj5 = re.search(regex5, file.paragraphs[j].text)
    if matchObj5:
        if con == con2:
            f = matchObj5.group(1);  # print(matchObj5.group(1))
            g = matchObj5.group(2);  # print(matchObj5.group(2))
            h = matchObj5.group(3);  # print(matchObj5.group(3))
            break
        con2 = con2 + 1
```

某个罪犯信息聚齐后，使用 append 函数添加到 dataframe 中

```
df = df.append({'名字': a, '性别': b, '出生年': c,'判决年': z, '民族': d, '文化': e, '判罪': f, '刑期': g, '罚金': h},
               ignore_index=True)
```

然后继续在相同文书中遍历匹配正则 1（姓名-性别-出生年），找到可能存在的下一个罪犯。正则 1 相当于一把钥匙，存在一个新罪犯才能成功匹配，成功匹配之后才能开启我接下来所有其余的信息提取操作。

当一个文书遍历结束之后，遍历下一个文书 ，直到文书遍历全部完毕。这里我仍是引入 os 模块来获取文书的所有路径。

```
lis = os.walk(r"C:\Users\risen\Desktop\FIN\GISSA大\文书")          #获取文档对象
for path, dir_list, file_list in lis:
    for file_name in file_list:
      try:
        a = '';b = '';c = '';d = '';e = '';f = '';g = '';h='';z='' #初始化参数
        pa = os.path.join(path, file_name);print(pa)
        file = docx.Document(pa)  #读入文档
```

自动提取方式十分高效便捷，7530 份文书左右的数据在 15 分钟内就提取完毕，最终获得了 12888 份罪犯的信息。由于信息缺失/部分文书段落正则不匹配，可能存在部分项目空缺的情况，但总体是数据十分完整的。

8.3 感想

这次作业从前期的数据获取和文本提取，到后期的数据分析，满打满算也花了一个月了，虽然蛮辛苦的，但成就感是非常强的，感谢助教师姐和老师给了我一个人完成这次大作业的机会。最开始看到作业的内容，发现每一部分都是不一样的，就有一种想要一个人把所有事都挑战一下的冲动，虽然这样可能锻炼不了团队协作能力，但是在大学生涯中团队合作的机会很多，而一个人考虑到事情的方方面面去完成的项目却很少。这次作业前半部分的要求更偏理科一些，后半更偏文科，走一遍这样的流程想必对我思维的全面性的拓展有所助益。这次作业的难点的确就在数据获取和文本提取上，尤其是数据网站设置了大量反爬虫机制，而我的时间和精力又有限，因此我使用的方法也不是那么底层，在部分关键节点还是会手动控制。而文本提取方面在经过摸索之后学会了正则提取这一方法，算是给未来的学习开了个头。其实在开始写作业的时候我还有把整个作业流程都自动化起来的宏心壮志，但在完成的过程中就觉得问题如山，最后呈现出来的就是这一次的效果。虽然，我还是对我弄出来的这个成果不太自信，但我还是为有这么一段体验感到满足的。