

同济大学

基于县域数据的中国贫困格局可视化分析

学 号： 1854193

姓 名： 杨幸

专 业： 数据科学与大数据技术

所在院系： 电信学院计算机科学与技术系

任课教师： 王瀚漓

完成日期： 2021年6月20日

摘要：在进入 21 世纪以来，贫困仍然是文明和经济发展需要面对的重大难题。如何准确地在各个时空尺度上评估、鉴别并减轻贫困成为一个迫切的话题。本文基于 2000–2019 年的中国县域年鉴中的普查数据，通过社会剥夺理论提取关键特征，并运用主成分分析构建关于县域地理位置多维贫困指标，在此基础上使用 GIS 的方法可视化分析得到中国 2311 个县市的聚类异质性，并在数据集的时间跨度上展现了类簇的迁徙情况。最后结合“精准扶贫”政策中 832 个国家级贫困县“摘帽”数据，综合展现我国贫困治理工作的进展和成就。

关键词：可视化 社会剥夺 多维贫困指标 精准扶贫

目录

一、绪论.....	3
1. 选题背景和意义.....	3
2. 研究方法.....	3
2.1 可视化方法.....	3
2.2 指标建立理论.....	4
3. 项目结构.....	4
二、数据处理与指标构建.....	5
1. 数据概要.....	5
2. 数据预处理.....	5
2.1 整齐数据.....	5
2.2 数据管理与清洗.....	6
2.3 缺失值处理.....	6
2.4 异常点检测.....	6
3. 多维贫困指标.....	7
3.1 PCA 降维.....	7
3.2 多维贫困指数.....	8
三、几种可视化技术.....	9
1. 地理数据时变可视化.....	9
2. 主题河流与词云.....	9
3. 箱线图.....	9
四、数据分析与可视化应用.....	10
1. 形势与政策.....	10
2. 多维指标时空格局.....	11
3. 单维指标分析.....	13
4. 贫困县摘帽时空演变.....	15
五、总结与展望.....	16
1. 本文总结.....	16
2. 工作展望.....	16
参考文献.....	17
附录.....	18
表格.....	18
表 1 普查指标变化表	18
表 2 指标映射对照表	19
表 3 成分因子贡献表	20
图片.....	21
图 1 多维指标时空格局(F1 – Rank)	21
图 2 多维指标时空格局(F2 – Rank)	22
图 3 多维指标时空格局(IMPI)	23
图 4 多维指标时空格局(人口密度)	24
图 5 贫困县摘帽逐年分布	25

一、绪论

1. 选题背景和意义

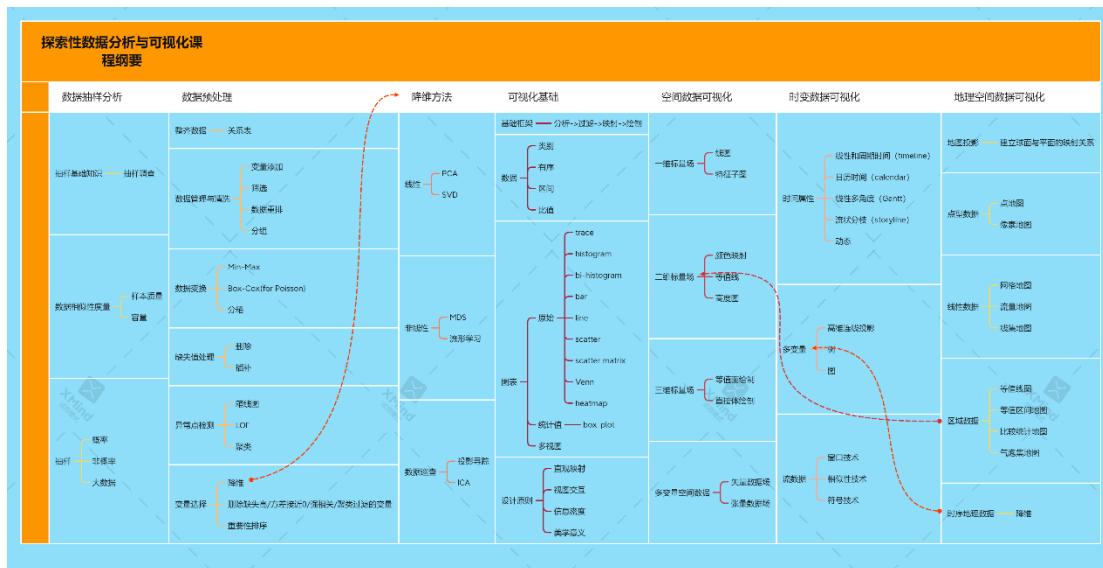
贫困是伴随人类社会发展历史变迁的客观社会现象，同时也是一个全球性的重大社会问题和现实难题。在刚刚过去的 2020 年，我国正式全面建成小康社会，但这并非脱贫工作的终点，而是新的起点，如何保证脱贫不返贫，实现社会公平，如何消除区域发展和条件的差异性，如何以扶贫工作为起驱动乡村振兴战略成为亟待解决的问题。

在这样的形势下，研究中国县域贫困的现状，建立一套评价指标体系，从而对其反映的区域致贫原因进行归纳，对存在的社会问题进行定位，可以为制定全面脱贫、实现社会可持续发展的国家或区域政策提供参考。

2. 研究方法

2.1 可视化方法

本次课程实验方法来自课程讲述内容，将其进行层次树可视化如下：



考虑到研究目的和数据具有时空分布的特点，本次实验主要采用了地理数据可视化中针对区域数据可视化的方法，将指标值二维标量场映射到以县域为粒度的中国地图域上，同时针对数据时变的特点动态地展现了这种变化。针对政策文本采用主题河流的方法进行可视化，展现我国扶贫攻坚战略的演进情况。另外利用箱线图展现多维贫困指标在分布上的差异和变化并采取散点图+线图进行回归分析，进一步阐明这些指标对于地区发展的指示作用。

2.2 指标建立理论

社会剥夺是人们维持日常生活的基本物质需要和参与社会活动的需要未能得到满足的现象。个人水平的社会剥夺主要反映在维持日常生活所需是否得到满足，如食物、衣服、住房等方面，可以直接准确地反映每个人的生活水平。但个人剥夺的研究由于受访者人数和调查区域所限往往受到很大限制。从地理学视角来看，受到剥夺的人群往往在空间分布上具有一定的聚集特性，使得某一地理单元在社会或者经济水平方面呈现普遍处于不利地位的特征，而区域剥夺正是反映研究单元中所有人群总体上都受到某种或多种维度社会剥夺的现象，因此可以说，区域剥夺是个人剥夺的在空间上的体现。个人剥夺和区域剥夺有着密切的关联，但这种关联往往是非线性的，例如，生活质量较好的人不会单纯因为搬到剥夺水平较高的区域而处于个人剥夺之下。大多数情况下，由于数据可获性的限制，个人剥夺的研究相对较少，而区域剥夺相关的研究不仅在数据收集方面更加容易，而且可以方便地通过时空分析进行横向和纵向比较，从而揭示区域剥夺的多维特征以及时空异质性。

3. 项目结构

└── data	处理后的县域数据
└── geojson	县级行政区地理边界数据
└── poverty	
└── accurate	贫困县摘帽数据
└── policy	
└── count	政策文件词频统计
└── text	政策文件文本
└── raw	原始县域数据
└── process	
└── clean	数据清洗相关实现
└── process	数据处理与分析
└── vis	可视化

二、数据处理与指标构建

1. 数据概要

本次实验数据主要包括三个部分：2000–2019 年中国县域统计年鉴，载于[国家统计局网站](#)；832 个国家贫困县历年摘帽名单，载于[国家乡村振兴局信息公开目录](#)，县级行政边界数据，载于[DATAV](#)，扶贫政策数据整理自[国务院网站](#)。

县域统计年鉴中主要包括各县市的基本情况、综合经济、农业、工业及投资、教育、卫生和社会保障这四大类的数据，由于部分数据过于不规整，很难进行数据清洗，这里采用 2011–2017 的普查数据。扶贫政策数据则是国务院网站下主题分类为民政、扶贫、救灾\扶贫的政策条目中的所有正文内容。县级行政边界数据记录全国所有县级行政单位的地理形态。

获取数据的具体方法如下：县域年鉴直接下载完成；贫困县摘帽名单运用 Kettle（一款开源的 ETL 工具）获取对应经纬度；县级行政边界数据，先是使用爬虫在国务院民政部爬取区号名单后再根据区号调用 DATAV 中的 api 下载数据；政策数据通过爬虫获取经分词统计、人工筛查后获得其时序表。

2. 数据预处理

这一部分主要是对县域年鉴数据的处理。

2.1 整齐数据

获取的统计数据是非常不规整的，这种不规整体现在目录结构不统一上，更重要的，是数据表结构不规整，如下图。因此第一步需要将其转换成规整的关系表，经查看选定相对规整的 2011–2017 年总共 215 张数据表普查数据进行处理。

2.2 数据管理与清洗

这部分的工作在于变量列名的确定，根据调研，我国县乡普查数据自 2014 年为界前后普查指标的数量、名称发生了变化，如附录表 1 所示。

经考虑实际意义最后保留了 24 个特征，对应原来列名的映射关系如附录表 2 所示。

2.3 缺失值处理

在普查数据中有一些数值是缺失的，如棉花产量、社会福利院数和床位数等，根据其实际意义进行了填 0 处理。

2.4 异常点检测

在核对数据类型中发现很多变量数据类型不是数值，因此进行一一的查验，部分结果如下，最终发现约 30 处系人为输入时造成的错误（如多余的符号 '*' '/' ']'，错误的 OCR：0 写成 0 等），均已更正。

A	B	C	D	E	F	G
225 第二产业人		96402	31002	76890	27522	98412
226 第三产业人		176134	86132	137893	39890	168772
227 固定电话户		58655	28103	43000	13679	59780
228 二、综合经济						
229 地区生产总值万元	1374477	750357	841372	335383	1194200	
230 第一产业万元	464105	179028	238539	96830	422024	
231 农业增加值万元	281282	95315	176717	43532	256711	
232 牧业增加值万元	146239	52411	48682	39313	147444	
233 第二产业万元	461862	337352	213247	116912	238298	
234 公共财政收入万元	"62080	34553	51752	26963	66589	
235 各项税收万元	60105	24618	39847	24323	51338	
236 公共财政支出万元	422258	265492	298068	189531	344542	
237 居民储蓄余额万元	1477039	663598	925086	411066	1191177	
238 年末金融资产万元	752049	277915	451151	276199	594340	
239 三、农业、工业及投资						
240 农业机械万千瓦时	69	42	46	16	57	
241 机收面积公顷	45921	12250	11652	5100	41850	
242 沿海丘陵面积	402	20	170	221	502	
湖南省						
就绪						

A	B	C	D	E	F	G
37 小学在校人		1513	712	37946	16397	38683
38 医疗卫生床		530	393	4826	3126	2315
39 各种社会个		1	5	34	51	20
40 各种社会床		60	216	5445	8021	3783
41						
42						
43 2016年县(市)社会经济主要指标						
44 江苏省						
45 指 标	单 位	江宁区	六合区	溧水区	高淳区	锡山区
46 一、基本情况						
47 行政区域平方公里		1563	1471	1064	790	399
48 乡个数	个					
49 镇个数	个					
50 街道办事处个	10		11	2	2	5
51 户籍人口万人		103	91	43	44	44
52 第二产业人		346500	235900	163000	163500	300300
53 第三产业人		332700	193900	110600	89800	135500
54 从业人员万人		215010	93421	70627	50279	
江苏省						
就绪						

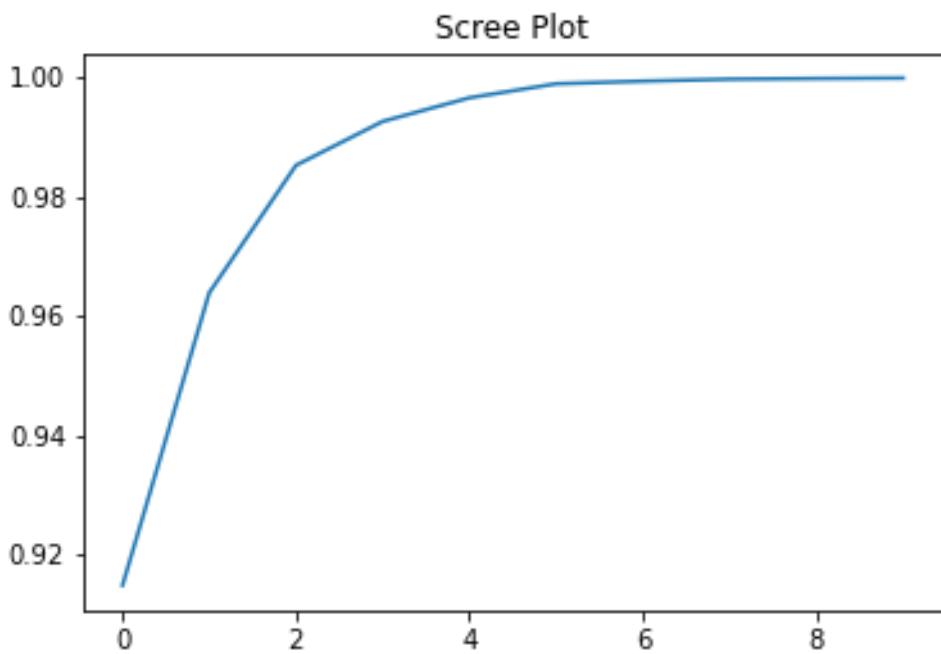
3. 多维贫困指标

3.1 PCA 降维

根据前面提到的社会剥夺理论，结合英国国际发展机构（DFID）建立的脆弱性-可持续生计分析框架，从人力资本、金融资本、工业资本3个维度选取了中国县域2011到2017年24个指标作为原始指标，采用主成分分析对指标进行降维，根据贡献方差累积的碎石图取前3个贡献约98.5%的成分作为主成分，成为测度单个维度社会贫困的指数，对于每一主成分，提取其因子载荷大于0.4的指标作为该主成分的贡献指标，所有主成分的贡献指标的集合成为中国县域多维贫困指标体系，详情见附录表3，可见F1主要表现地区的发展水平，和工业总产值与第二产业（工业）的增加值显著相关，F2主要表现地区民众经济条件，和居民储蓄余额、金融机构余额关系较大，F3则是这两种的综合补充。

$$F_i = \sum_{j=1}^n L_j X_j$$

其中， L_j 表示多维贫困指标j的主成分载荷， X_j 是多维贫困指标j的标准化值，而n为该主成分贡献指标的总数。



3.2 多维贫困指数

根据 F1, F2, F3 构建起的多维贫困指标建立多维贫困指数，即用它们与主成分分析得到的特征向量作内积，在实际处理中对数据大小统一修正使之范围较小，其数学表示为：

$$IMPI = \sum_{j=1}^3 E_j F_j$$

其中 E_j 为第 j 个主成分的特征值。

三、几种可视化技术

1. 地理数据时变可视化

区域数据是一种常见的地理空间数据，涉及地图上不同区域自然或社会经济的基本状况和统计信息。其中，自然数据区域包括自然要素的空间分布及其相互关系，如地质、气象和植被等；社会经济和人文数据反映区域中社会、经济等人文要素的地理分布、区域特征和相互关系，如人口、行政区划和交通等；还有面向其他专业的数据，如航海、旅游和工程设计等。

区域数据的可视化常采用专题地图（theme map）类似的绘制方法。其基本思路是遵循可视化设计原则，给地图上不同区域赋予特定的颜色、形状，或采用特定的填充方式展现其特定的地理空间信息。

在本次实验中主要使用了等值区间地图，采用 `plotly.express` 中的 `mapbox` 实现地理映射绘制和时变动效。

2. 主题河流与词云

主题河流图顾名思义就是形状像河流的图形，实际上是一种特殊的流图，它主要用来表示事件或主题等在一段时间内的变化。主题河流中不同颜色的条带状河流分支编码了不同的事件或主题，河流分支的宽度编码了原数据集中的值。此外，原数据集中的时间属性，映射到单个时间轴上。主要是用来表示事件或主题等在一段时间内的变化，可以用来观察事物的变化，应用于多个生产生活领域之中。

“词云”就是对网络文本中出现频率较高的“关键词”予以视觉上的突出，形成“关键词云层”或“关键词渲染”，从而过滤掉大量的文本信息，使浏览网页者只要一眼扫过文本就可以领略文本的主旨。

在本次实验中采用 `pyecharts.charts` 中的 `ThemeRiver` 实现主题河流绘制，采用 `pyecharts.charts` 中的 `WordCloud` 实现词云绘制。

3. 箱线图

箱线图是一种用作显示一组数据分散情况资料的统计图。因形状如箱子而得名。在各种领域也经常被使用，常见于品质管理。它主要用于反映原始数据分布的特征，还可以进行多组数据分布特征的比较。箱线图的绘制方法是：先找出一组数据的上边缘、下边缘、中位数和两个四分位数；然后，连接两个四分位数画出箱体；再将上边缘和下边缘与箱体相连接，中位数在箱体中间。

在本次实验中采用 `plotly.express` 中的 `box` 实现箱线图绘制。

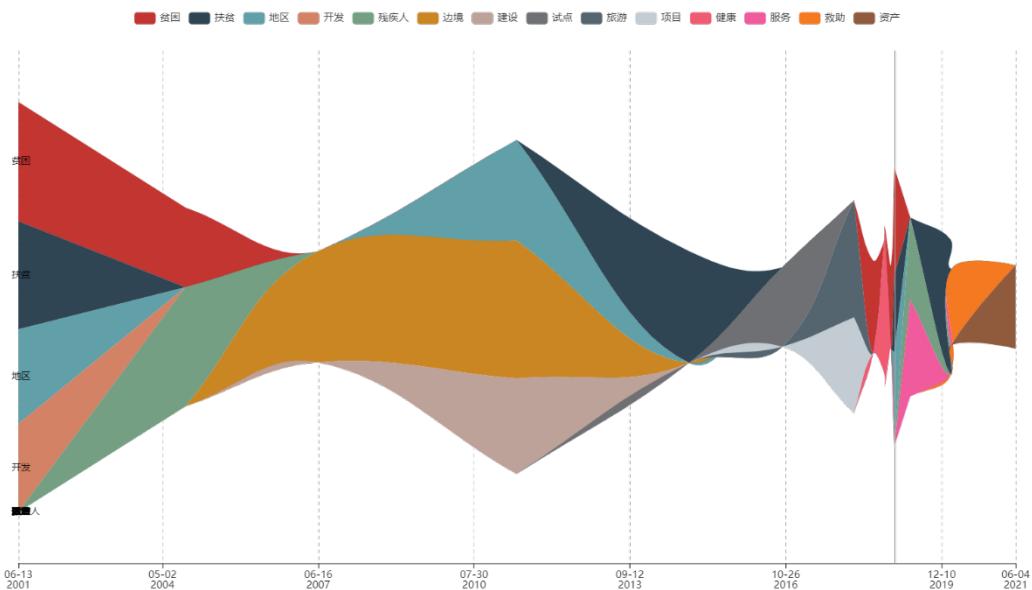
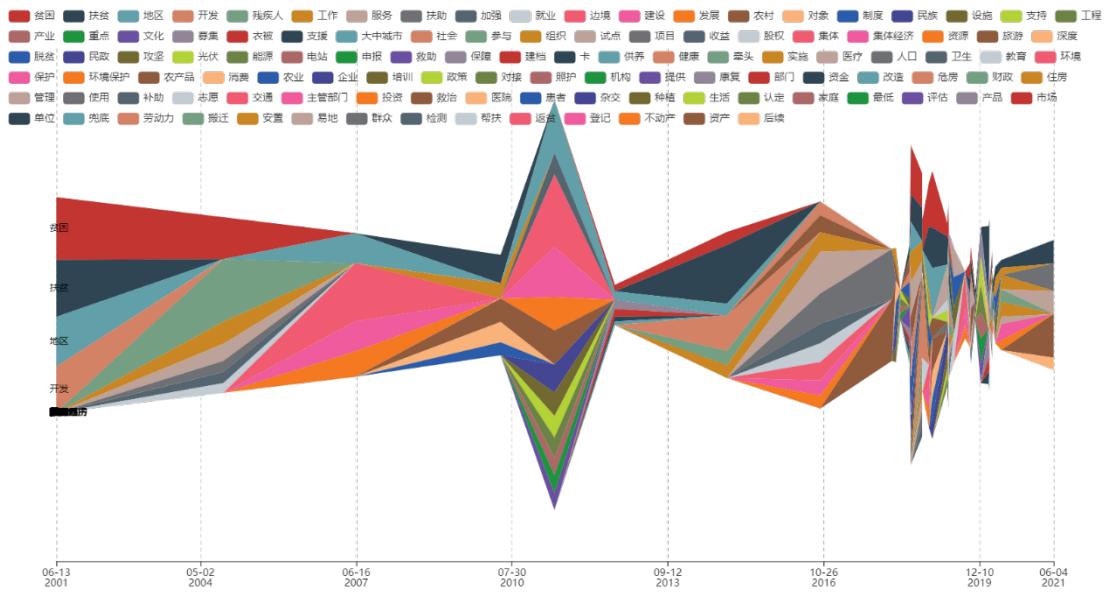
四、数据分析与可视化应用

1. 形势与政策

首先对爬取的政策文件进行分析，根据一定的阈值进行筛选词频后得到以下图像。从词云中可以分析得出，我国扶贫工作的重心在于地区性的措施和偏远地区的发展，具体措施是通过各种脱贫项目包括农副产品消费、文旅开发项目等带动地区经济发展，同时兼顾卫生健康和生态保护。



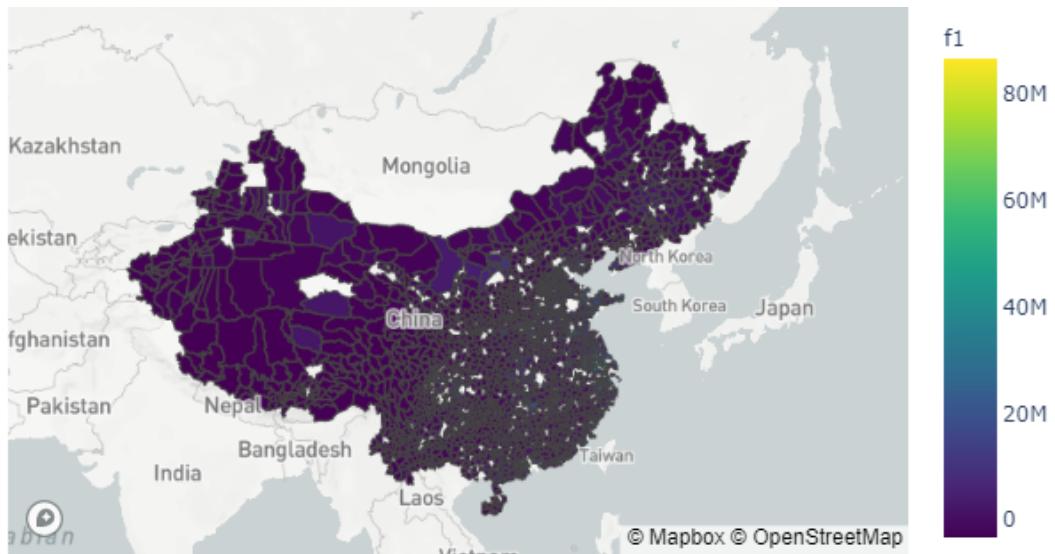
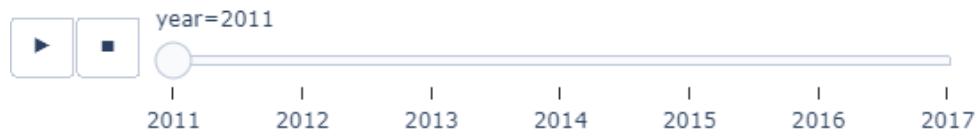
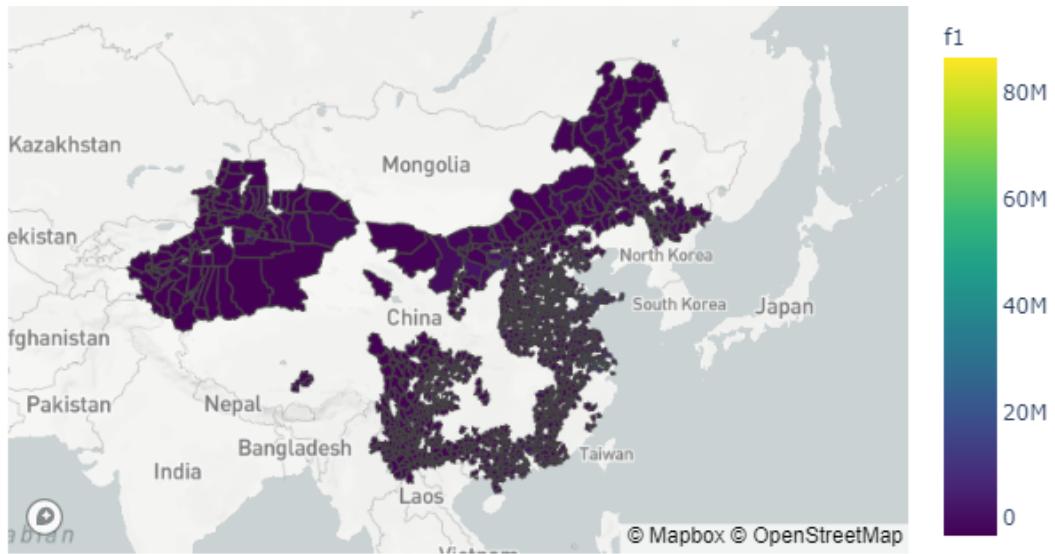
通过主题河流的嬗变可以分析得到我国政策在不同时间中的不变——扶贫的初心不变、注重边远地区不变、重视弱势群体不变，这体现我国的扶贫工作宗旨；而变化——由矛盾的提出到问题的具体解决方案再到完成后的稳固工作，这无一不体现出我国扶贫事业的发展。其中，尤其以 2014 年，习近平总书记提出了“精准扶贫”的概念以后，我国扶贫事业有了新的面貌，在更广更深更细致的层面全面布局施策。以下分别是全部词语和出现次数在 60 次以上的词语可视化得到的主题河流。



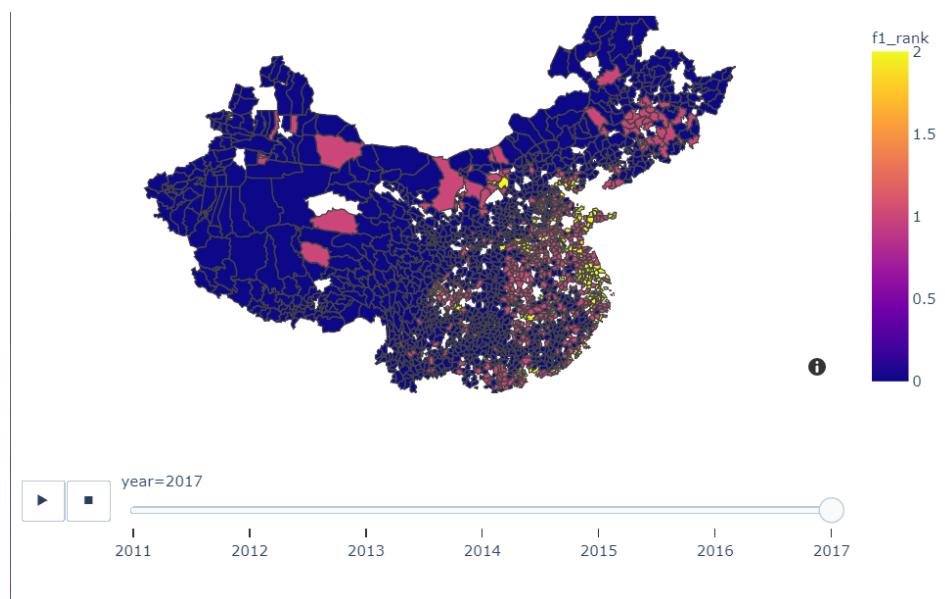
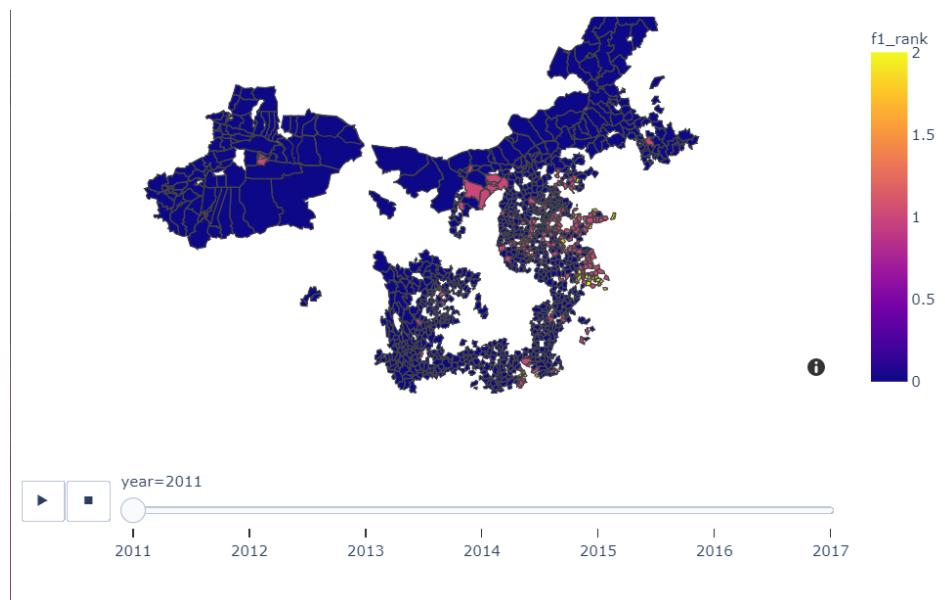
2. 多维指标时空格局

将建立的指标演变呈现在地图格局上可以直观地表现区域发展的状况，这里以 F_1 为例，在附录另有 F_2 、IMPI以及人口密度的变化。首先是对数值指标直接进行可视化，效果并不好，色块间没有很好的区分度。如下是 F_1 指标在2011年和2017年普查数据可视化的对比。

需要说明的是，图中有些县市的块是缺失的，这是因为：县区命名由地方管理，每年会发生变更，本次项目采用的是民政部2019年更新文件中的命名方法，而相比往年县市统计年鉴中存在更名、设立、撤销的情况。



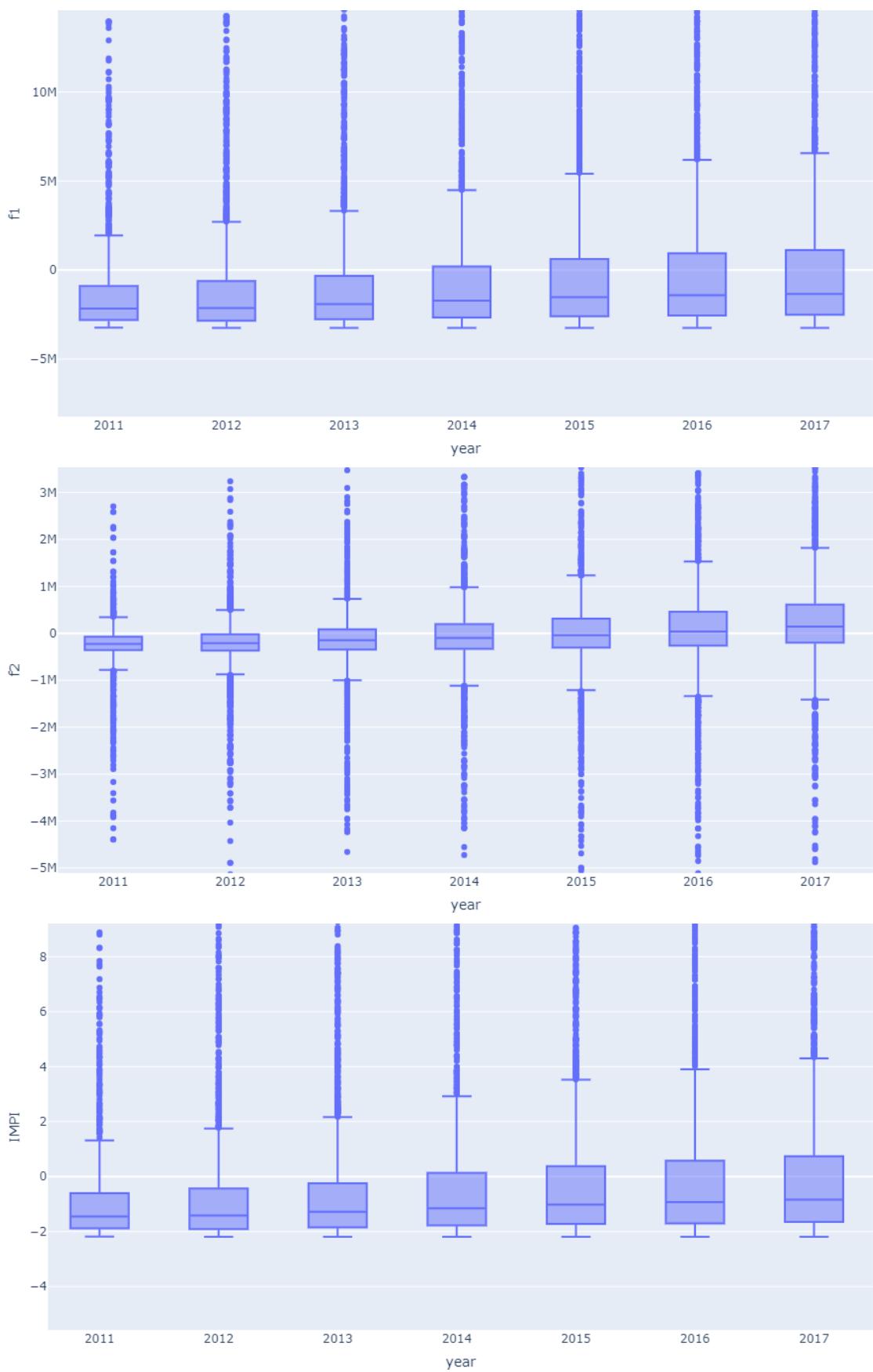
因此，考虑使用分箱操作，将各数值映射到“low”、“medium”、“high”三个等级，则上示图样成为如下具有较好信息表达能力的图样。

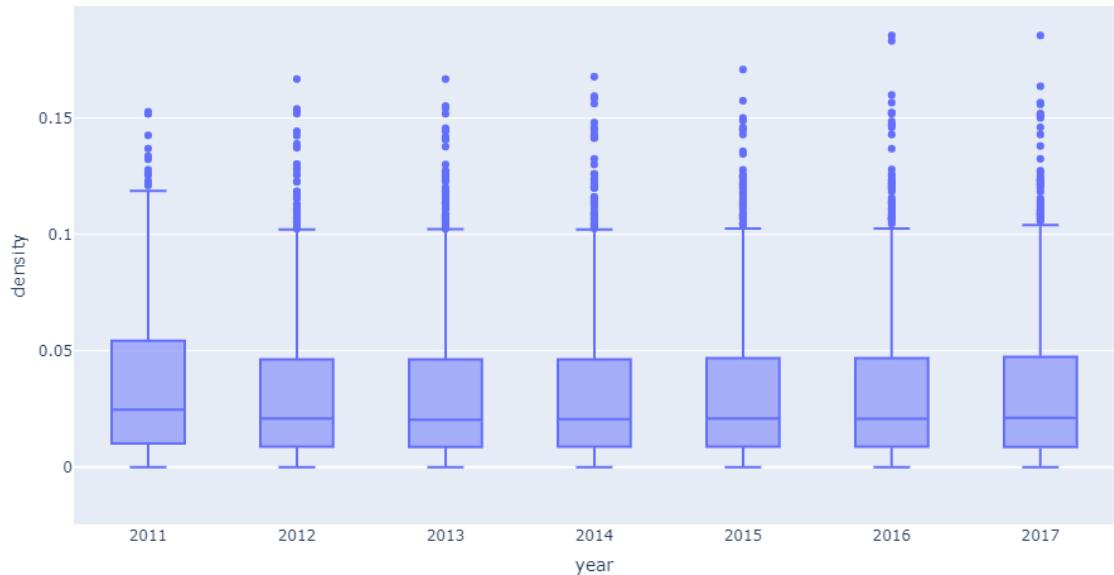


可见，经过分箱后数地区间的相异程度有很大提高，从图中的分布也不难看出。发展状况中等及以上的县区大多集中在黑河-腾冲县以东南地区，发展较好的县区在长三角地区存在聚集，这符合我们的认知，且随着时间的变迁，越来越多贫困县加入到红色（“medium”）的行列中。人口密度的分布则变化不大，河南、山东部分县市人口密度较大。

3. 单维指标分析

针对多维指标中的 F_1 、 F_2 值以及综合指数IMPI在时间轴上的分布变化有助于我们从整体把握产业和经济发展状况，另外还对人口密度的变化进行可视化，这里采用箱线图展现这种变化。





从箱线图中可以很清楚地分析出以下结论：我国工业和经济水平逐年上升，底部数据向上靠拢，这体现出扶贫的兜底作用；但在空间分布上呈现向两级分化的趋势，这反映了我国目前国情与矛盾痛点，即不充分不平衡的发展；最后是人口密度的变化，总体上平稳，细节上存在略微下降，而总人口是逐年增长的，这说明了我国人口流动中趋于平衡的分布。

4. 贫困县摘帽时空演变

将贫困县的经纬度标注在地图上按年份分组得到如下图像，根据地理学第一定律：所有的事物都与其他事物相关联，且距离越近关联性越强，这种空间关系成为“空间自相关”，即位置相近的区域具有相似的变量取值。在图中的体现就是贫困县分布往往成片连区具有这种空间相关性。

而站在时间维度上看，可见扶贫后期关键的脱贫任务集中在云、贵、川三省，这部分地区发展状况长足欠发达，属于扶贫任务中的硬骨头，具体分年分布图像见附图 2。



五、总结与展望

1. 本文总结

本次项目主要完成了一个简单的以县域为粒度的地理信息系统，以社会剥夺理论为基点构建了多维贫困指数，在时空格局上展现了全国各县区的统计数据及发展水平。此外对国务院扶贫相关政策文本也进行了一定的可视化分析，从管理层看清扶贫形势。最后整理了贫困县摘帽数据展现了我国打赢这场脱贫攻坚战的伟大胜利，明确了任务中的重难点，对后续政策的制定有一定参考意义。

总体而言，中国社会剥夺格局区域差异明显，中西部地区的剥夺水平较高，社会整体发展水平较差，而东部沿海城市社会发展水平较好，意味着我国正在经历“社会分化”。在新的发展契机下，应该对这样的区域不平衡现象加以重视，制定致力于推进协调发展的区域政策。社会剥夺的时空格局揭示了多维度的社会剥夺特征，可以识别下个发展阶段中城市需要重点解决的社会问题，成为重要的政策参考依据。

2. 工作展望

首先是本次项目未来的工作，在简单的时空格局可视化之后可以对这些区域进行聚类，发掘出一些模式关联；使用具体的空间自相关指标如 Moran' I 等进行更精密的、肉眼之外的判断。

另外则是此次项目中遇到的困难，在县域普查年鉴数据获取与清洗方面存在极大问题。首先是数据来源不易，更重要的是数据格式不符合数据库关系表的基本范式，而且数据中存在很多脏点，这些问题的解决耗费了极大的工作时间。基于此提出的展望是：我国可以效仿一些国家，对一些政务数据进行开源和规范化，这有利于学术界对我国发展工作做出有意义的成果，在制定相关决策时能广泛吸取科学的意见。目前，深圳市数据共享计划在中国各大城市中进行得较好，开发者可以很便利地获取市政相关数据进行相应分析，希望其他地区政府能跟进，这也正符合十四五规划中关于“加快数字化发展 建设数字中国”的发展远景。

参考文献

- [1] Li, G., Cai, Z., Liu, J. et al. Multidimensional Poverty in Rural China: Indicators, Spatiotemporal Patterns and Applications. *Soc Indic Res* 144, 1099 – 1134 (2019). <https://doi.org/10.1007/s11205-019-02072-5>
- [2] Chen Wan , Shiliang Su. et al. China's social deprivation: Measurement, spatiotemporal pattern and urban applications. *Habitat International* 62, 22 – 42 (2017). <https://doi.org/10.1007/s11205-019-02072-5>

附录

由于图表过多，在正文中仅列举了部分图表内容，附录中是被省略的部分。

表格

表 1 普查指标变化表

2013 及以前	2014 及以后
一、基本情况	一、基本情况
行政区域土地面积	行政区域面积
乡(镇)个数	乡个数
村民委员会个数	镇个数
年末总户数	街道办事处个数
其中：乡村户数	户籍人口
年末总人口	第二产业从业人员
乡村人口	第三产业从业人员
年末单位从业人员数	固定电话用户
乡村从业人员数	二、综合经济
其中：农林牧渔业	地区生产总值
农业机械总动力	第一产业增加值
固定电话用户	农业增加值
二、综合经济	牧业增加值
第一产业增加值	第二产业增加值
第二产业增加值	公共财政收入
地方财政一般预算收入	各项税收
地方财政一般预算支出	公共财政支出
居民储蓄存款余额	居民储蓄存款余额
年末金融机构各项贷款余额	年末金融机构各项贷款余额
三、农业、工业及投资	三、农业、工业及投资
粮食产量	农业机械总动力
棉花产量	机收面积
油料产量	设施农业占地面积
肉类总产量	粮食总产量
规模以上工业企业个数	棉花产量
规模以上工业总产值(现价)	油料产量
固定资产投资(不含农户)	肉类总产量
四、教育、卫生和社会保障	规模以上工业企业单位数
普通中学在校学生数	规模以上工业总产值
小学在校学生数	固定资产投资
医院、卫生院床位数	四、教育、卫生和社会保障

各种社会福利收养性单位数	普通中学在校学生数
各种社会福利收养性单位床位数	中等职业教育学校在校学生数
	小学在校学生数
	医疗卫生机构床位数
	各种社会福利收养性单位数
	各种社会福利收养性单位床位数

表 2 指标映射对照表

real	2013	2014
a1	行政区域土地面积	行政区域面积
a2	乡(镇)个数	乡个数+镇个数
a3	村民委员会个数	街道办事处个数
a4	年末总人口	户籍人口
a5	固定电话用户	固定电话用户
a6	第一产业增加值	第一产业增加值
a7	第二产业增加值	第二产业增加值
a8	地方财政一般预算收入	公共财政收入
a9	地方财政一般预算支出	公共财政支出
a10	居民储蓄存款余额	居民储蓄存款余额
a11	年末金融机构各项贷款余额	年末金融机构各项贷款余额
a12	农业机械总动力	农业机械总动力
a13	粮食产量	粮食产量
a14	棉花产量	棉花产量
a15	油料产量	油料产量
a16	肉类总产量	肉类总产量
a17	规模以上工业企业个数	规模以上工业企业单位数
a18	规模以上工业总产值(现价)	规模以上工业总产值
a19	固定资产投资(不含农户)	固定资产投资
a20	普通中学在校学生数	普通中学在校学生数+中等职业教育学校在校学生数
a21	小学在校学生数	小学在校学生数
a22	医院、卫生院床位数	医疗卫生机构床位数
a23	各种社会福利收养性单	各种社会福利收养性单

	单位数	位数
a24	各种社会福利收养性 单位床位数	各种社会福利收养性单 位床位数

表 3 成分因子贡献表

指标	成分		
	F1	F2	F3
行政区域土地面积	-0. 0002	#####	0. 000761
居民储蓄存款余额	0. 184386	0. 484927	-0. 03379
年末金融机构各项贷款余 额	0. 285237	0. 768518	0. 269521
农业机械总动力	1. 97E-06	#####	#####
粮食产量	0. 009649	-0. 01074	-0. 09888
棉花产量	#####	#####	#####
油料产量	0. 000514	-0. 00041	-0. 00619
肉类总产量	0. 00184	-0. 00078	-0. 01653
规模以上工业企业个数	3. 02E-05	2. 15E-05	7. 19E-06
规模以上工业总产值(现 价)	0. 896819	-0. 3875	0. 152189
固定资产投资(不含农户)	0. 181217	0. 127998	-0. 92398
乡(镇)个数	#####	3. 81E-07	#####
普通中学在校学生数	0. 001485	0. 002627	-0. 00579
小学在校学生数	0. 001844	0. 00304	-0. 00652
医院、卫生院床位数	0. 000137	0. 000254	-0. 00043
各种社会福利收养性单位 数	7. 63E-07	1. 45E-06	#####
各种社会福利收养性单位 床位数	0. 000155	0. 000212	-0. 00038
村民委员会个数	7. 46E-07	#####	2. 88E-05
年末总人口	2. 59E-06	3. 38E-06	#####
固定电话用户	0. 009735	0. 017335	0. 002296
第一产业增加值	0. 013109	0. 006064	-0. 08765
第二产业增加值	0. 213581	0. 062437	-0. 16441
地方财政一般预算收入	0. 028122	0. 032407	-0. 0287
地方财政一般预算支出	0. 027236	0. 047641	-0. 06003

图片

图 1 多维指标时空格局($F_1 - Rank$)

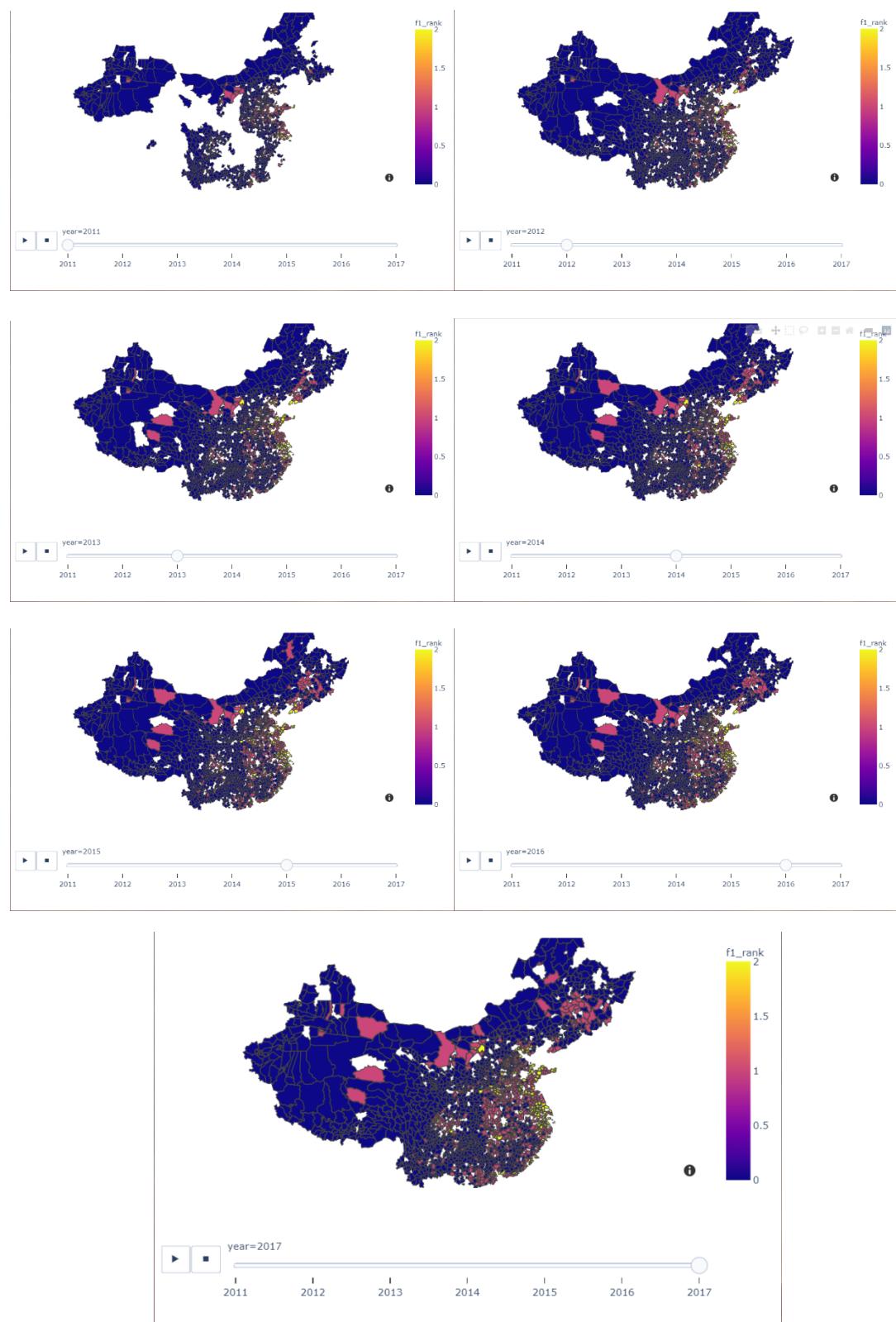


图 2 多维指标时空格局($F_2 - Rank$)

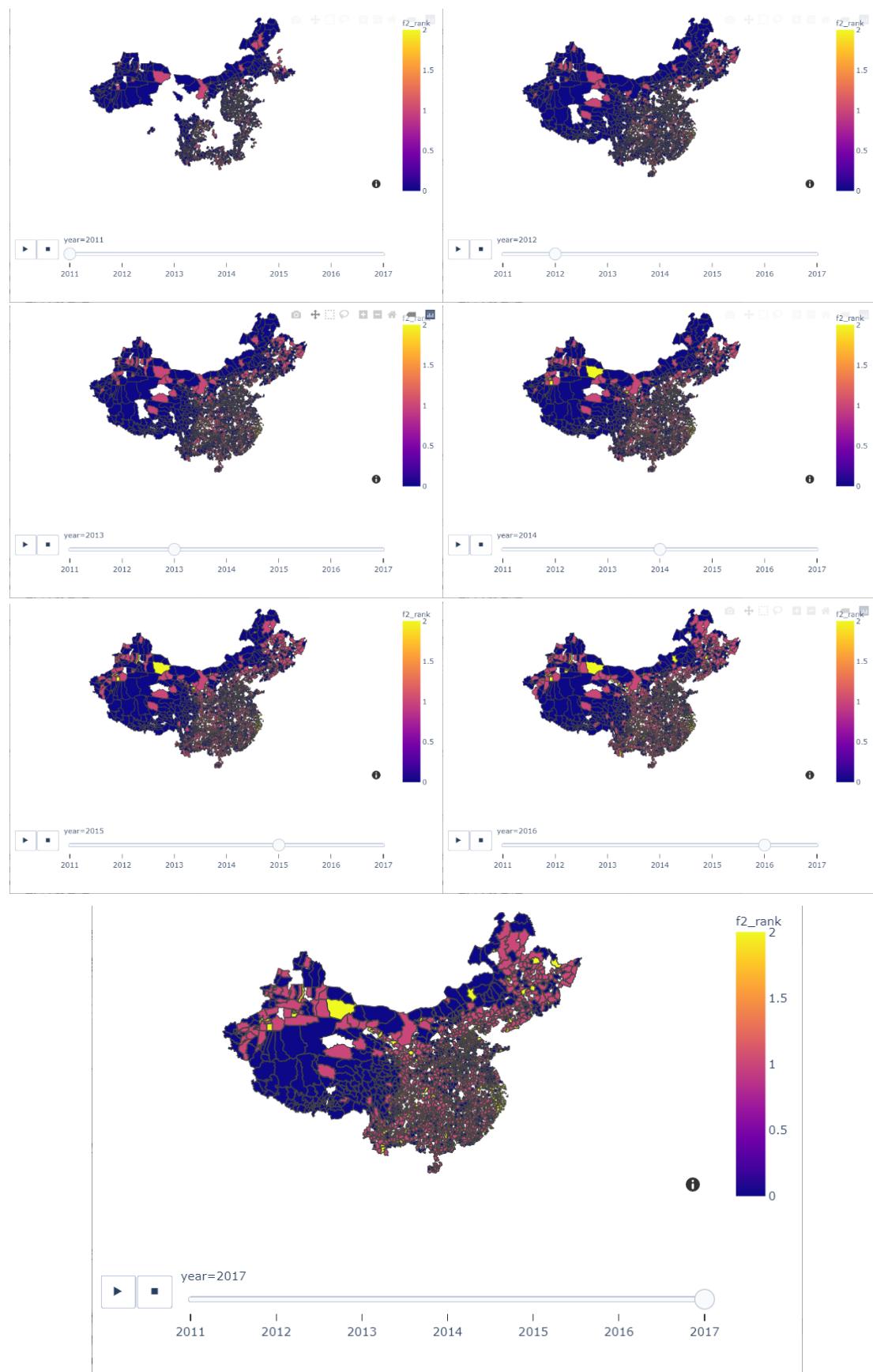


图3 多维指标时空格局(*IMPI*)

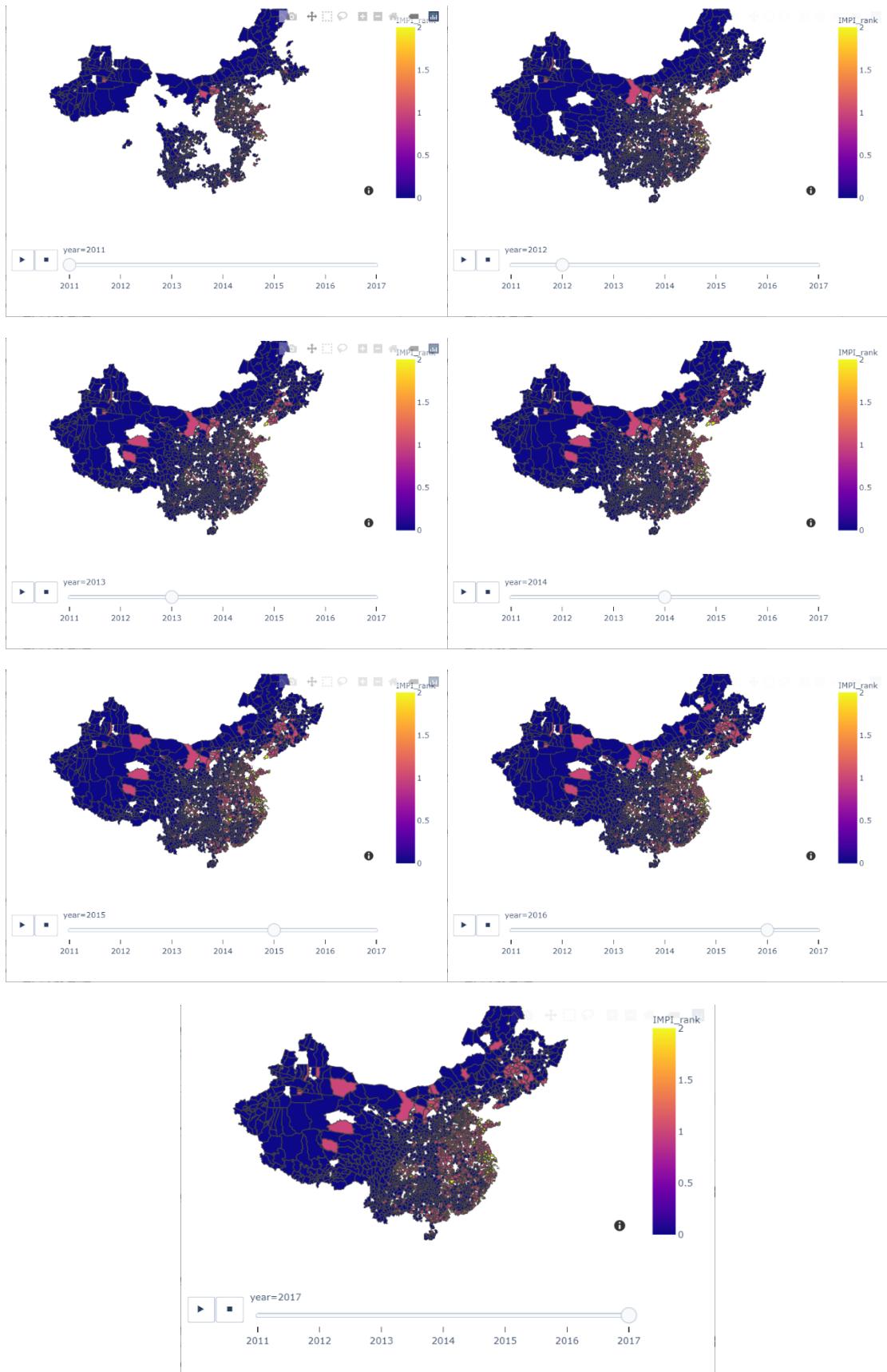


图4 多维指标时空格局(人口密度)

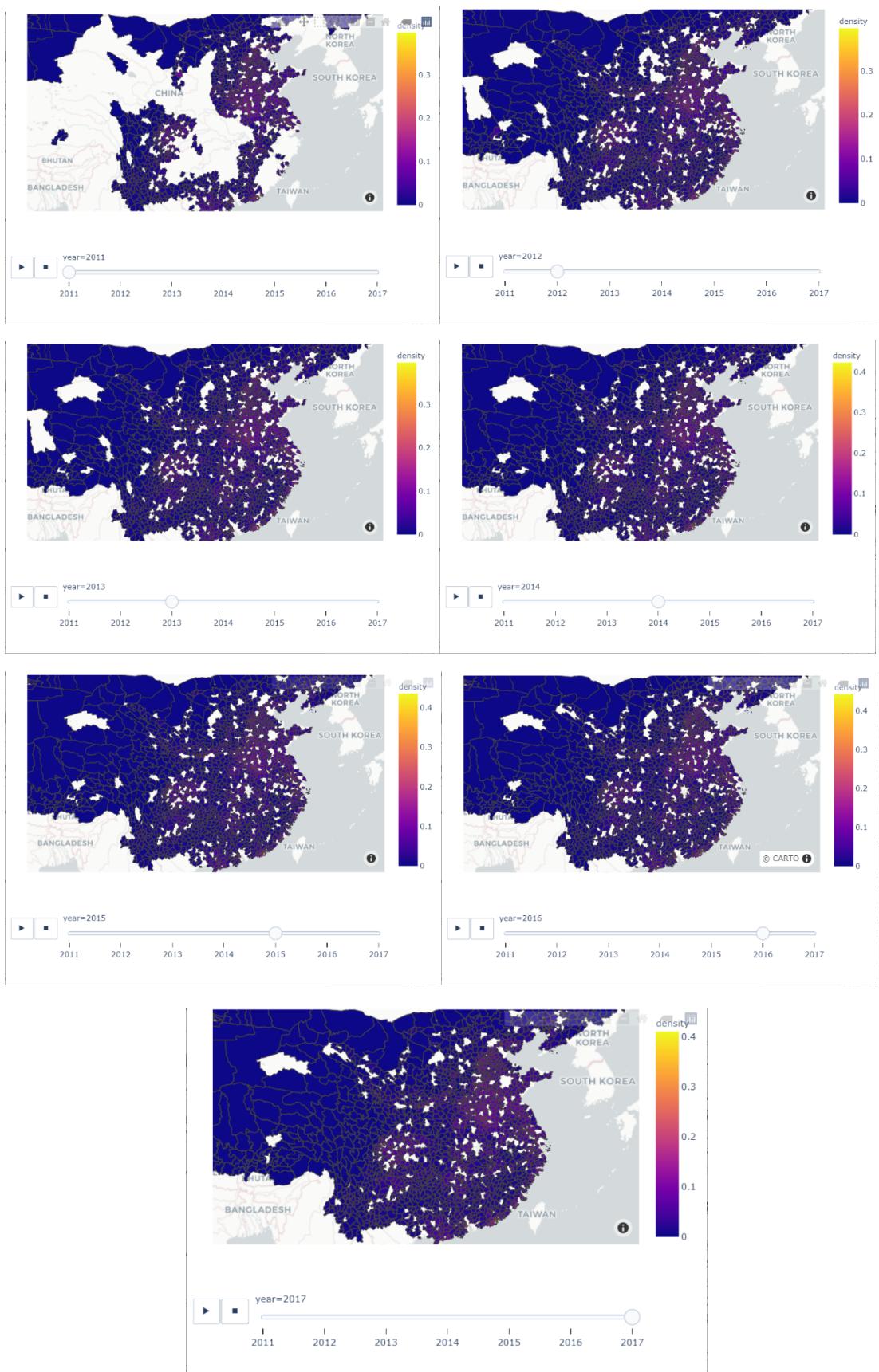


图 5 贫困县摘帽逐年分布

