

学号： 12177405211

兰州文理学院

大数据结课 论文

(2021 届本科)

题 目： 工具化分析国内 COVID-19 的基本
信息及发展趋势

学 院： 数字媒体学院

专 业： 软件工程

作者姓名： 蒲鹏飞

指导教师： 岳庆生 职称：

完成日期： 2020 年 6 月 20 日

二〇二〇年 六 月

目录

1 引言.....	1
1.1 摘要.....	2
1.2 关键词.....	2
2 数据采集.....	3
2.1 数据来源.....	3
3 数据分析.....	3
3.1 数据的基本处理.....	3
3.2 处理折叠问题.....	4
3.3 清洗数据 NAN.....	4
3.4 提取出其中的省/直辖市数据进行分析.....	6
4 COVID-19 的可视化分析.....	8
5 人流量对 COVID-19 传播的影响.....	15
5.1 COVID-19 湖北省各地级市死亡人数（雷达图）.....	15
5.2 小范围人口迁徙对疫情的影响.....	16
6 相较于 SARS, 大数据对 COVID-19 的积极影响.....	18
6.1 大数据追踪传播路径.....	18
6.2 大数据构建疫情发展模型.....	19
6.3 大数据助力资源配置.....	19
7 结语.....	20
8 参考文献.....	20
9 致谢.....	21

工具化分析国内 COVID-19 的基本信息及发展趋势

1 引言

2019 年末，一场突如其来的疫情改变了大部分人的生活，其传播速度之快，致死率高。严重程度丝毫不亚于 2003 年的非典型肺炎（简称 SARS）。加之遇到春节前后客乘流量高峰时期，给这场本就严重的疫情添加了更多的变数。天灾无情人有情，各项物资第一时间到达灾区，生产医疗物资的厂家日夜不停的为前线补给，每个人都在尽可能地为灾区捐去自己的一份爱心，现在国内疫情已经基本得到控制。从这次疫情来看，互联网技术已经发展地相当成熟，它在我们这次阻击疫情的过程中发挥出来了不可磨灭的力量。BAT 都在使用自己最擅长的尖端技术帮助人们度过这次漫长而又令人记忆深刻的冬天。首先，看到阿里巴巴利用全球购为需要救治的病人和我们的白衣天使求购医疗物资，在疫情结束后又为其他地区捐出大批的医疗物资，这其中资源整合的范围之大，速度之快，是完全依靠于当下日趋完善的互联网技术的。开学复课以后，又为上亿的学生提供钉钉在线课堂，并且服务器非常地稳定，深受老师和同学们的喜爱。百度为疫情提供实时监测数据和数据可视化，虽然每次看到离去的人都很心痛，但同时也不得不感叹于科技的伟大。腾讯微视则是为赋闲在家的人们带去欢乐的同时也让大家更好地了解防控疫情的知识要点，科学防控，把大家连成一个大家庭，互相监督。总之，我只介绍了 BAT，但是很多小企业，技术不是很出众，但他们也在使用者自己的方式，为抗疫贡献。本文我将使用数据分析和可视化，从两个方面分析疫情的基本信息以及疫情对人们生活的影响。



图 1 地图词云

1.1 摘要

COVID-19 自从在中国发生以后，无时无刻不在牵动着国人的心。无论何时，人们内心深处的恐惧更多的来源于未知。及时掌握第一手资料是非常重要的，大数据挖掘，追踪在当下成为了一种备受热捧的技术，虽然作为业内人士对于一些冷冰冰的数据可以产生浓厚的兴趣，但是对于其他领域的人来讲，想了解这个事物却有没兴趣去深入的了解它的想法，所以可视化提供了一个相当不错的桥梁，它结合数据，形象生动向人们解释着最核心的东西，再加上实时性，它完全就是模拟事物的发展，而且辅助人们建立模型，再次训练，优化，就可以创新出一种新的生产力工具，甚至可以总结规律，模拟事物接下来的发展趋势。

比如这次疫情，利用数据总结，建模，可视化分析，每个人都在极力地预测拐点的来临，我做这次报告为的是向更多没有去了解这次疫情的人普及疫情的情况，看到网上做类似于疫情可视化分析的也有很多，但是加上代码可能很多人就望而却步，我使用 Pycharm 和 Jupyter notebook 处理数据，

我得出最佳的分析数据后，然后使用 Echarts 和 灯果文化, FineBI 等数据制作网站，分析了一系列能够产生影响的数据，成了地图，雷达图，直方图等等适合不同数据的图形，深刻地展示出来了地缘对于疫情的影响，人员流动对于疫情的影响，封城决策以及各地政府宣传对于疫情的积极影响，最后提出大数据对这次疫情影响，包括挖掘，建模分析，资源整合。

1.2 关键词

COVID-19 数据可视化 Pycharm Jupyter notebook FineBI Github 灯果文化 百度慧眼

1.3 缩写术语

COVID-19: 新型冠状病毒肺炎

SARS: 非典型肺炎

Pycharm: 集成化开发环境

Jupyter notebook: 开源 web 应用程序

2 数据采集

Github 开源网站上有许多关于 COVID-19 的数据，不足之处是使用 git 上传以后数据就不是最新的，但是我为了获取实时数据采用了 python 爬虫获取中国境内的最新数据，每隔几分钟获取一次数据，以保证数据的最新性和完整性。（注：爬虫使用合理合法，来源于公开信息的网址）。

采用 Pycharm，每隔两分钟获取一次数据，以下就是爬虫采集数据展示：

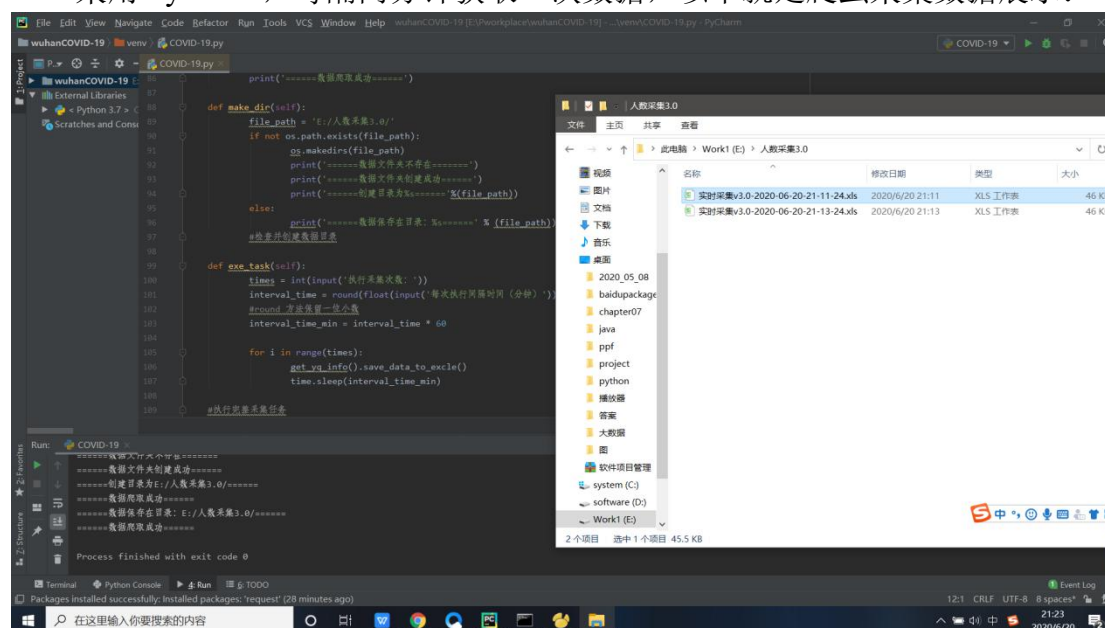


图2 python 爬取国家卫健委疫情公开数据

2.1 数据来源

所有数据均来自中国卫健委 2020-6-20 21:11:24 之前的数据。

3 数据分析

3.1 数据的基本处理

使用 `df=pd.read_excel（‘具体位置’）` 进行文件的读写。

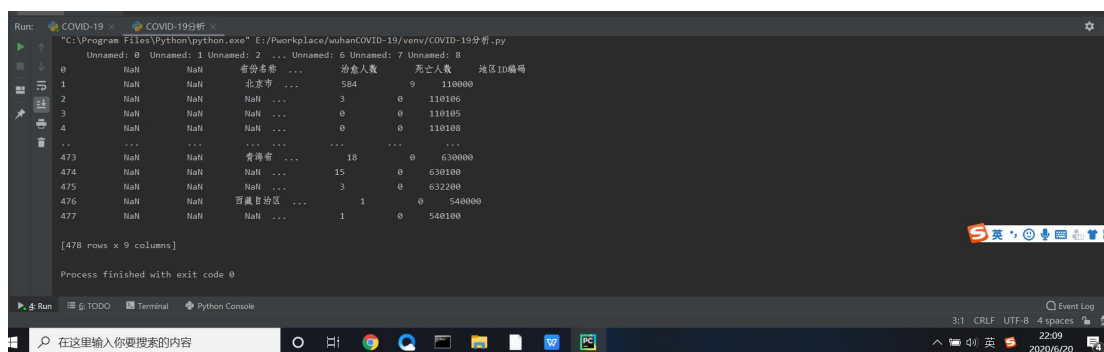


图 3 python 爬取数据读出

但是你会发现数据会发生一定程度的折叠和 NaN（Not A Number），现在首先要使得数据展开并且清洗 NaN。

3.2 处理折叠问题

`pd.set_option('display.max_rows',None)` //设置行最大化

`pd.set_option('display.max_columns',None)`//设置列最大化

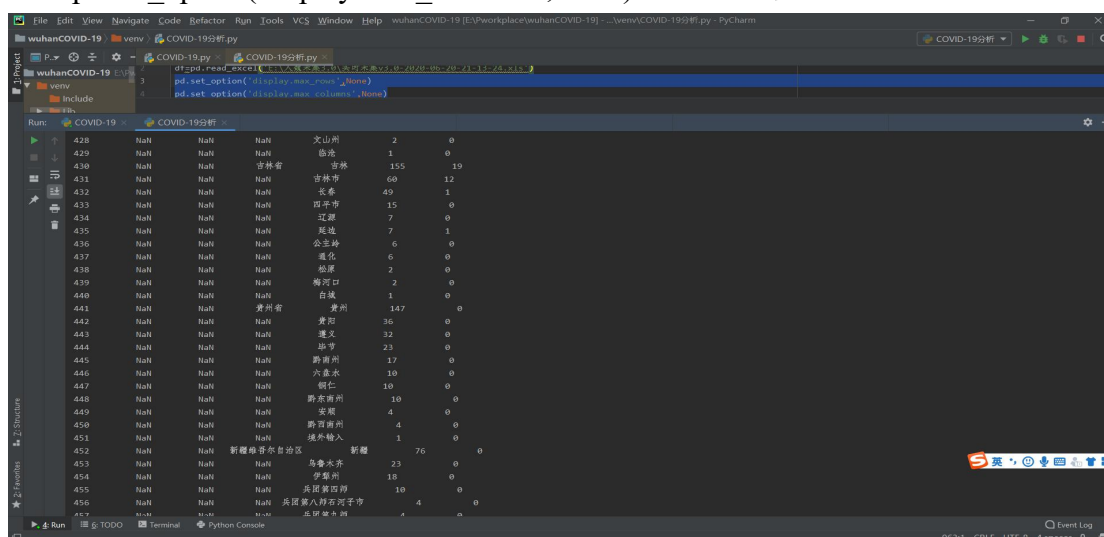
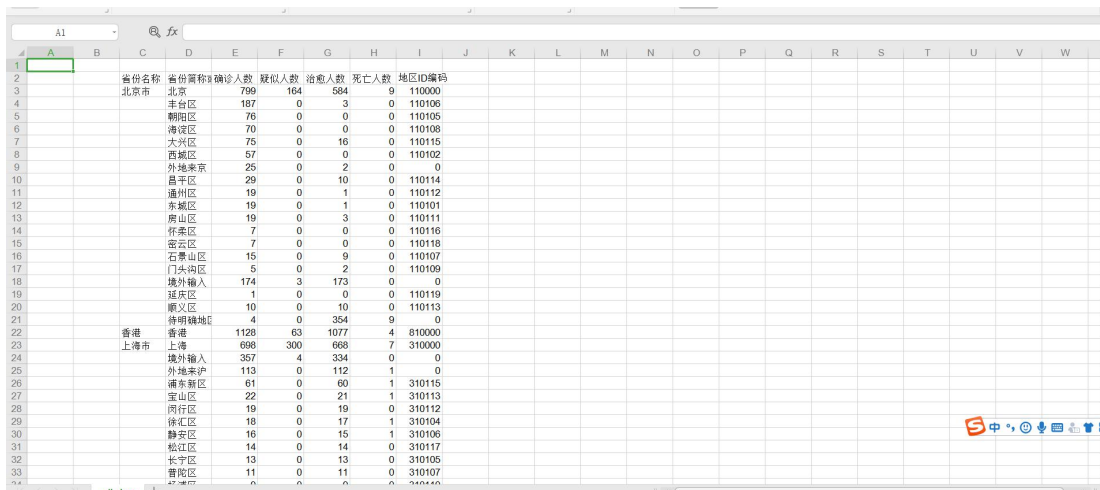


图 4 使折叠数据全部读出

3.3 清洗数据 NaN

通过看 xls 文档，发现 NaN 是由于空数据造成的，现在我们要对他进行清洗。

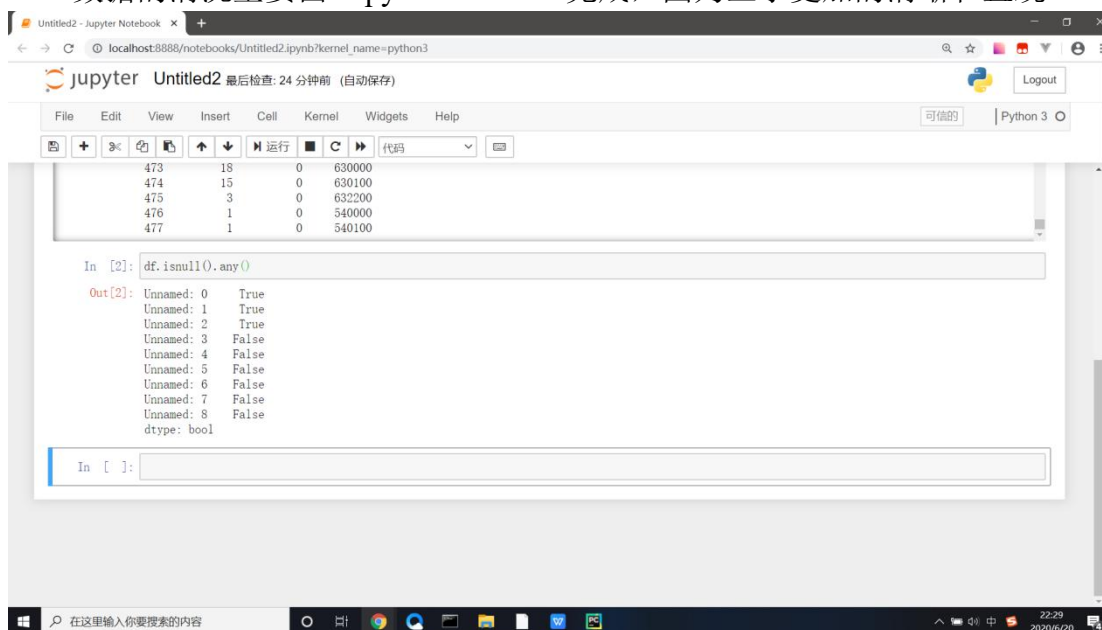


	省份名称	确诊病例数	疑似人数	治愈人数	死亡人数	地区ID编码
3	北京市	799	164	584	9	110000
4	丰台区	187	0	3	0	110106
5	朝阳区	76	0	0	0	110105
6	海淀区	70	0	0	0	110106
7	大兴区	75	0	16	0	110115
8	西城区	57	0	0	0	110102
9	外地来京	25	0	2	0	0
10	昌平区	29	0	10	0	110114
11	通州区	19	0	1	0	110112
12	东城区	19	0	1	0	110101
13	房山区	19	0	3	0	110111
14	怀柔区	7	0	0	0	110116
15	密云区	7	0	0	0	110118
16	石景山区	15	0	9	0	110107
17	门头沟区	5	0	2	0	110109
18	境外输入	174	3	173	0	0
19	延庆区	1	0	0	0	110119
20	顺义区	10	0	10	0	110113
21	崇明地区	4	0	354	9	0
22	香港	1128	63	1077	4	810000
23	上海市	698	300	668	7	310000
24	境外输入	357	4	334	0	0
25	外地来沪	113	0	112	1	0
26	浦东新区	81	0	60	1	310115
27	宝山区	22	0	21	1	310113
28	闵行区	19	0	19	0	310112
29	徐汇区	18	0	17	1	310104
30	静安区	16	0	15	1	310106
31	杨浦区	14	0	14	0	310117
32	长宁区	13	0	13	0	310105
33	普陀区	11	0	11	0	310107

图 5 读出数据显示在 EXCEL 表格中

这是 xls 文档，由此看出北京市与香港之间的空数据，全部为 NAN。

数据的清洗主要由 Jupyter notebook 完成，因为显示更加的清晰和直观。



```

In [2]: df.isnull().any()
Out[2]: Unnamed: 0    True
         Unnamed: 1    True
         Unnamed: 2    True
         Unnamed: 3    False
         Unnamed: 4    False
         Unnamed: 5    False
         Unnamed: 6    False
         Unnamed: 7    False
         Unnamed: 8    False
         dtype: bool

```

图 6 处理数据 NAN

使用 `df.isnull().any()` 发现具有 NAN 的列，很明显第 0, 1, 2 列数据需要进行清洗。这里有使用 boolean 填充的方法，但是个人选择使用直接删除 NAN。

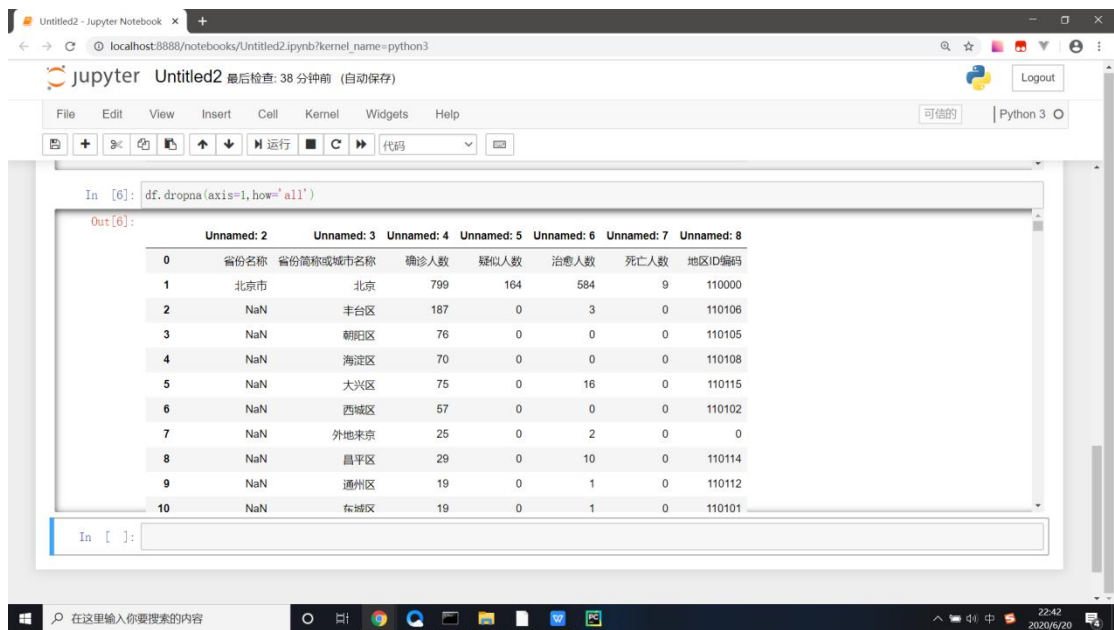


图 7 删除前两列“脏数据”

数据清洗完成后，发现第二行数据依然存在 NaN，实际上这行数据对我们将要进行的操作没有影响，为了便于观察，没有进行深度的处理。

使用 `df.drop(df.columns[0:2], axis=1, inplace=True)` 函数删除前两列。

3.4 提取出其中的省/直辖市数据进行分析

使用 `df[~df.isna().any(axis=1)]` 提取出不含 NaN 的行数据，也就是我们所需要的省/直辖市的数据。

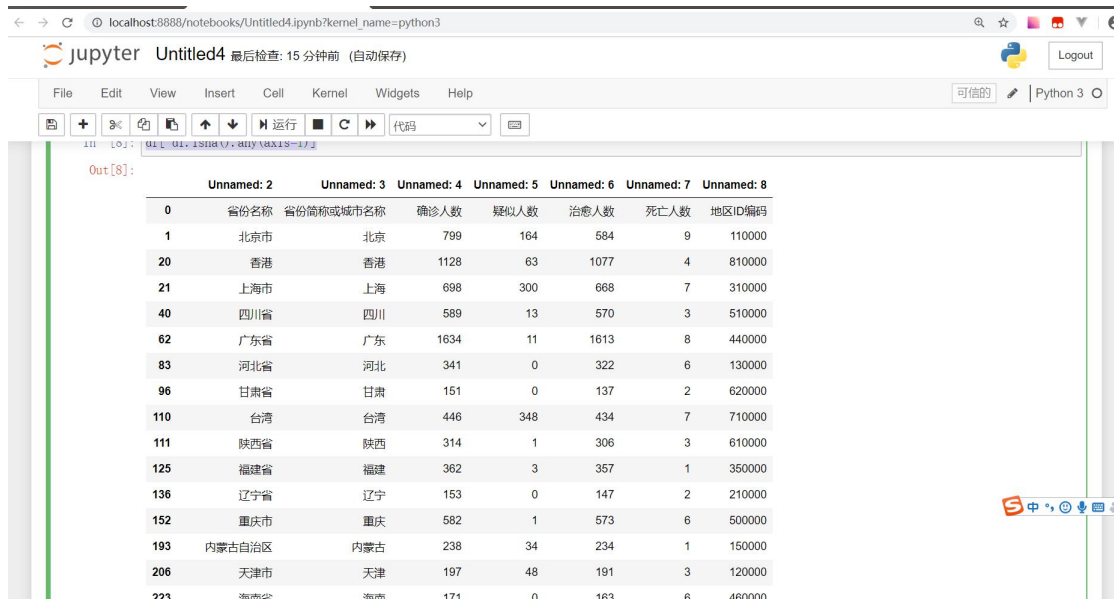


图 8 提取出数据中的省级数据

通过取出省/直辖市的数据，然后新建索引使用 lambda 匿名函数算出“治愈率”。

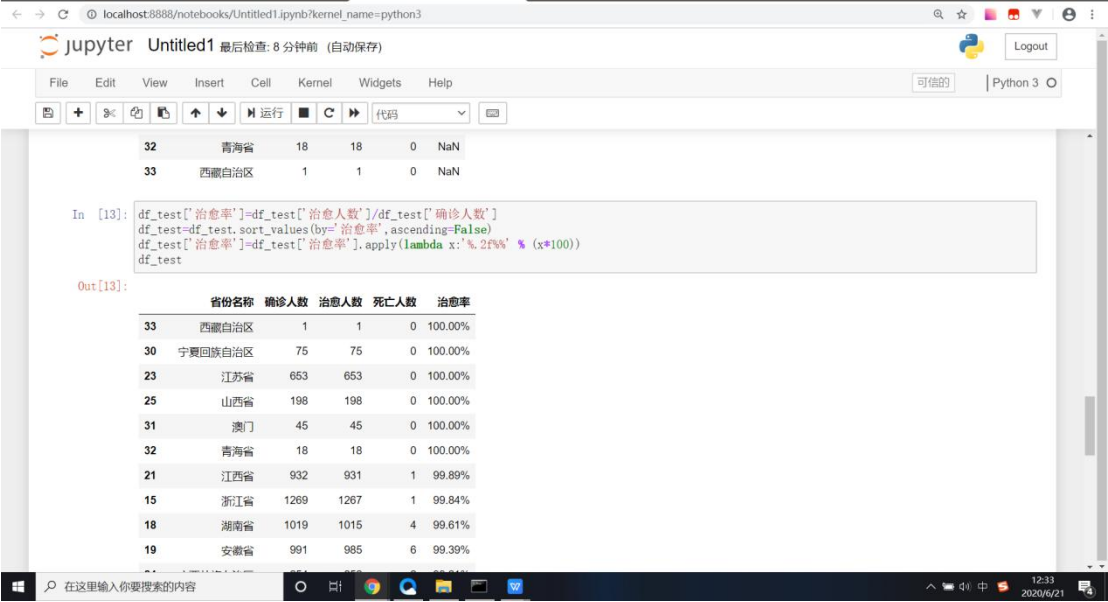


图 9 新建索引并计算“治愈率”

由以上数据来看，西藏、青海西部地区由于确诊人数少，因此治愈率很高，而宁夏由于防控到位，加之症状不像中心地带那样严重，也取得了不错的成效。

至于江苏，浙江地区的治愈率高，这个应该是当地的医疗水平要高于西北地区。

湖北的治愈率低，这跟它是疫区中心有很大的关系，医疗资源不够用，而甘肃由于外来人员入境，医疗水平低，导致疫情后开始了二次反弹。

通过计算得出每个省的致死率。

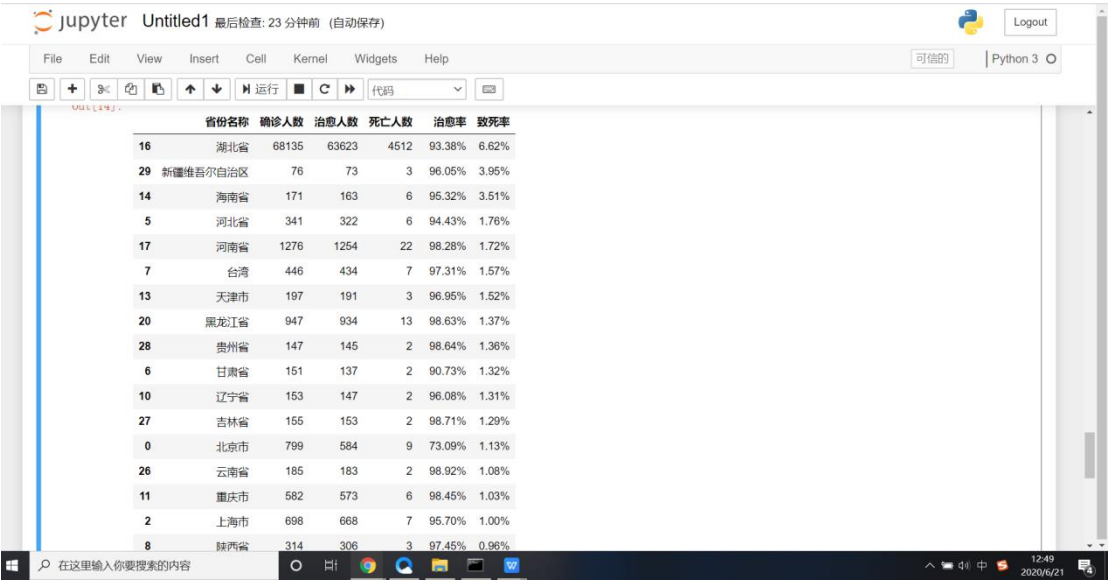


图 10 新建索引并计算“致死率”

不出所料，湖北的致死率依然是最高的，湖北的致死率高还是老原因，集中爆发加上医疗资源的调度不均匀造成的。新疆受制于医疗水平的原因，消息的及时性不够。

至于海南致死率高，很令人值得深思，由于海南是旅游城市，加之疫情爆发时间在春节时期，人流量大，对于疫情的传播起到了一定程度上的作用。

我觉得致死率分析的原因是多方面的，根据病人入院时的轻重程度，当地的医疗水平，比如澳门的致死率为 0.00%，都是有很大关系，浙江确诊人数很多，但是致死率仅仅为 0.08%。

治愈率和致死率虽然反应不出更详细的信息，但是可以从一定程度上反映出当地应急部门，疾控中心，宣传部门对于突发医疗事件的态度和响应程度。

从致死率来讲，情况没有 SARS 那样猛烈，但是它的潜伏时间长，难以发现，给予医护人员极大地挑战，并且这将是一个长期的挑战。

根据以上分析省级的规律和方法，计算出地级市的治愈率和致死率。

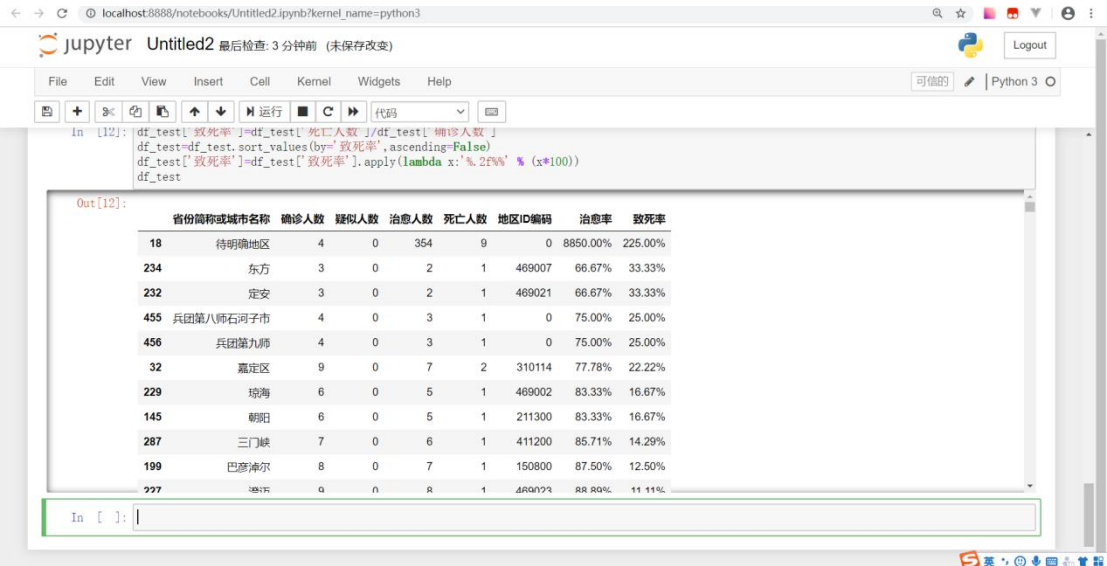


图 11 计算地级市的“治愈率”“死亡率”

通过分析地级市的治愈率和致死率发现出来了一个很极端化的情况，边远地区确诊基数不大，但是致死率高，灾区中心地带向外扩散，疫情有减缓的趋势，类似于同心圆，说明距离以及防控措施的收紧对于疫情有一定程度上的制约。

4 COVID-19 的可视化分析

根据省/直辖市的基本信息，我将用 MATPLOTLIB 完成以及配合一些其他的图形库。

此时我使用 Echarts 完成对疫情的直方图分析。

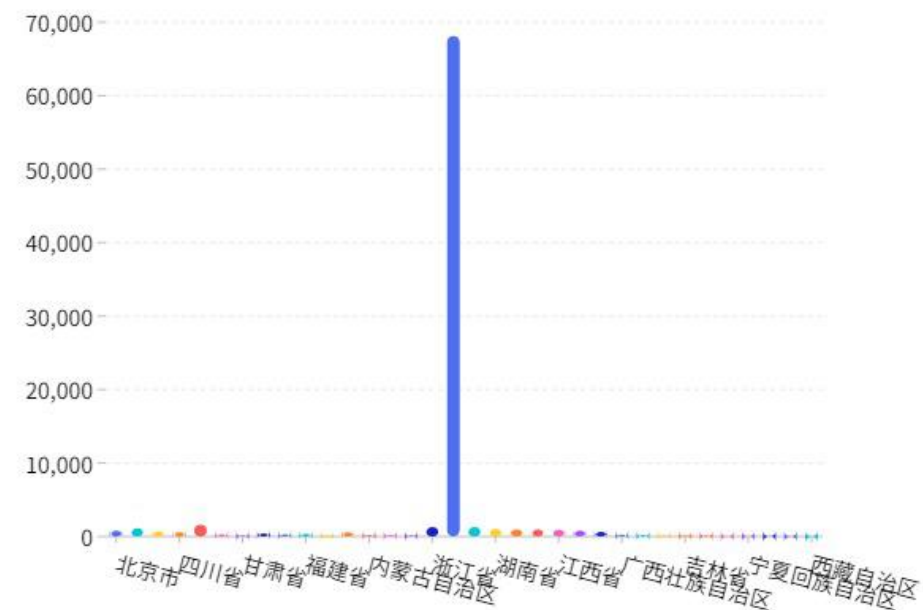
使用网站有花火数据、灯果可视化工具、FineBI 可视化软件，百度慧眼可视化等

等。

4.1 COVID-19 全国各省确诊人数（直方图）

COVID-19全国各省/直辖市确诊人数直方图

单位: 人



数据来源: 中国卫健委6.20

制作: 蒲鹏飞

图 12 COVID-19 全国各省确诊人数

由直方图得出，湖北的确诊人数要远远大于其他省份，这与灾情的严重程度十分相关，湖北武汉是最早发现此例病并且开始防控的城市。

4.2 COVID-19 各省治愈人数和死亡人数（蝴蝶图）

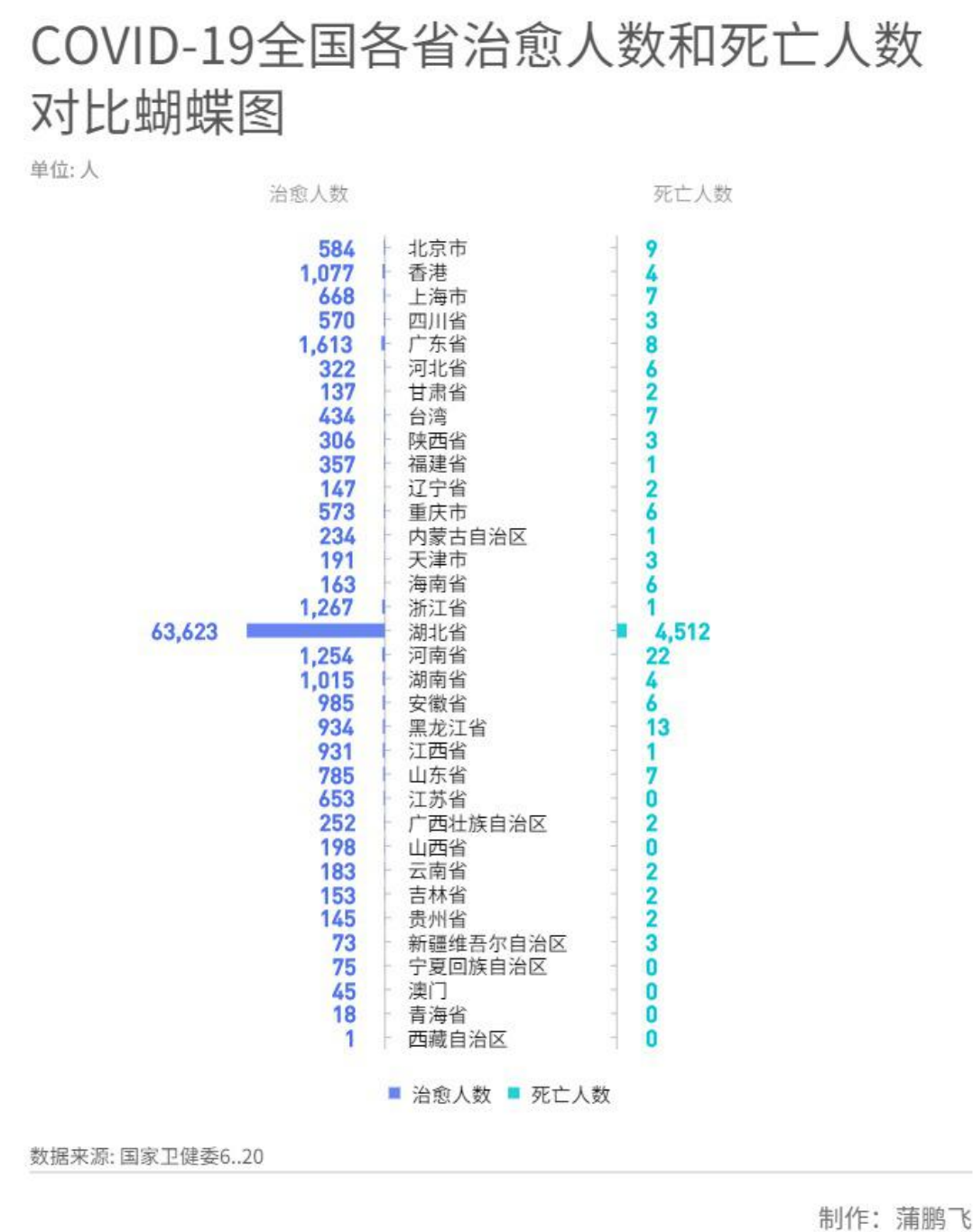


图 13 COVID-19 全国各省治愈人数和死亡人数

使用蝴蝶图能够清晰的反映出治愈人数和死亡人数，湖北省的治愈人数多，但是相对地它的死亡人数跨越治愈人数一个量级，反观甘肃，西藏，青海来讲，以现有的医疗水平能够将死亡人数控制在个位数，可见防控的毅力和决心。

4.4 COVID-19 全国各省确诊人数分析

COVID-19全国各省确诊人数分布地图

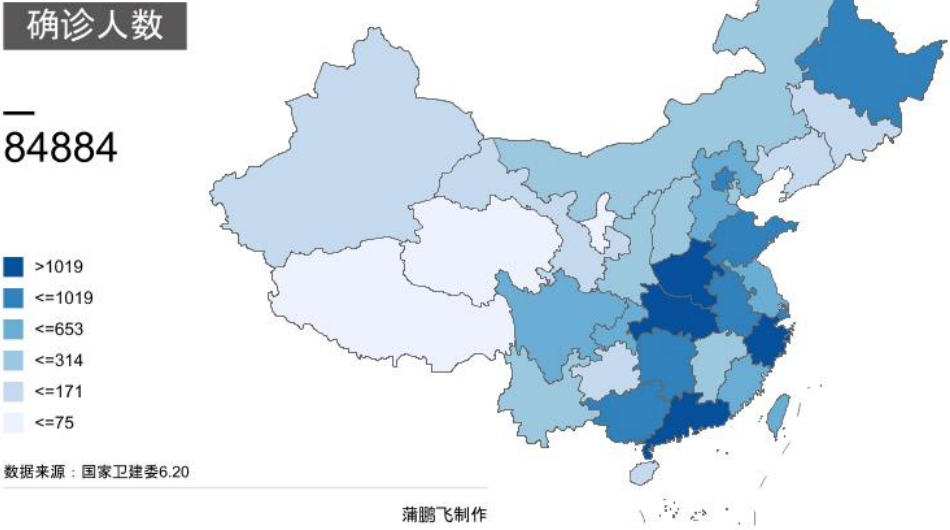


图 14 COVID-19 全国各省确诊人数地图

广州的疫情令我很意外，是境外输入造成的，还是国内疫情发展，还是与当地饮食习惯有关，这个我们从图中得不到想要的信息。

4.5 COVID-19 全国各省死亡人数分析

COVID-19全国各省死亡人数分布地图

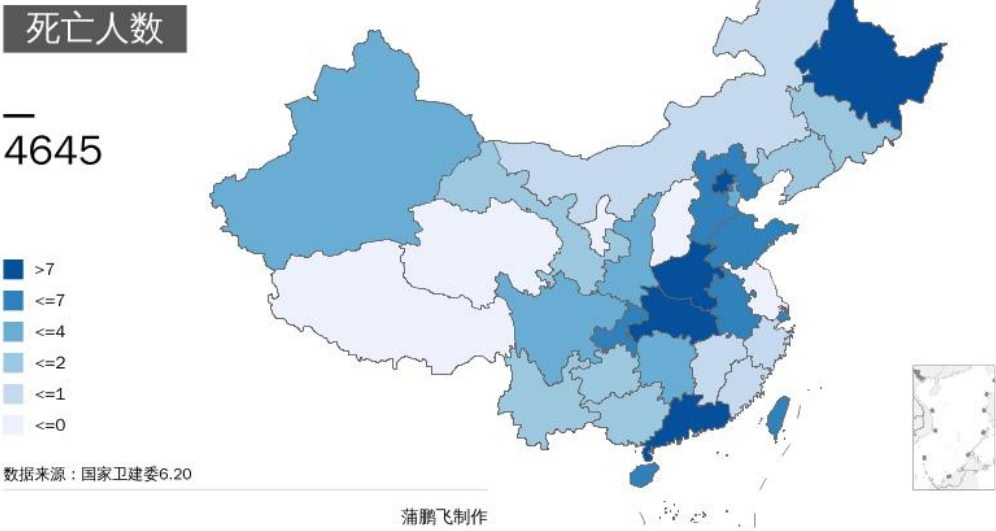


图 15 COVID-19 全国各省死亡人数地图

根据可视化地图综合数据得出的结论，湖北的灾情是及其严重的，甚至中心地带确诊人数达到了一个千人级的量级，这是非常可怕的，由于封城的原因，可以从图中明确的看出无论是确诊人数和死亡人数都和行政划分高度类似，如果没有采取封城的措施，那么疫情的扩散就是像光晕一样层层向外拓展，说明封城的措施是能够有效地遏制疫情的发展。

4.6 COVID-19 全国各省治愈率分析

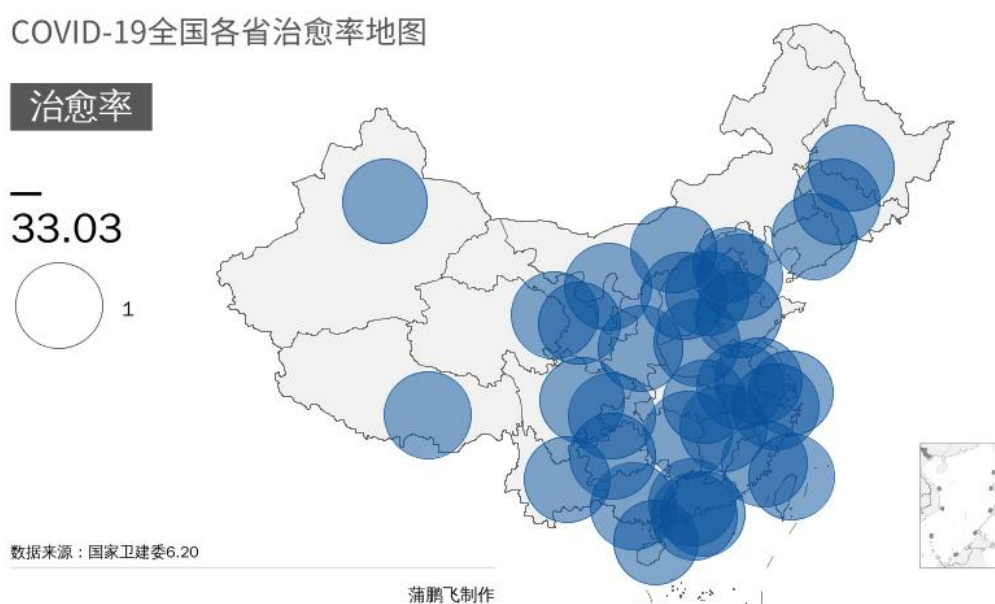


图 16 COVID-19 全国各省治愈率地图

很明显，西北地区的治愈率要远小于东南沿海地区，东南沿海地区的科技发达，医疗水平明显高于中西部地区，其次，消息也能够及时到达，掌握第一手的资料，无论是对于疫情的宣传还是控制，都会拥有先发制人的优势。

4. 7 COVID-19 全国各省致死率分析

COVID-19全国各省致死率地图

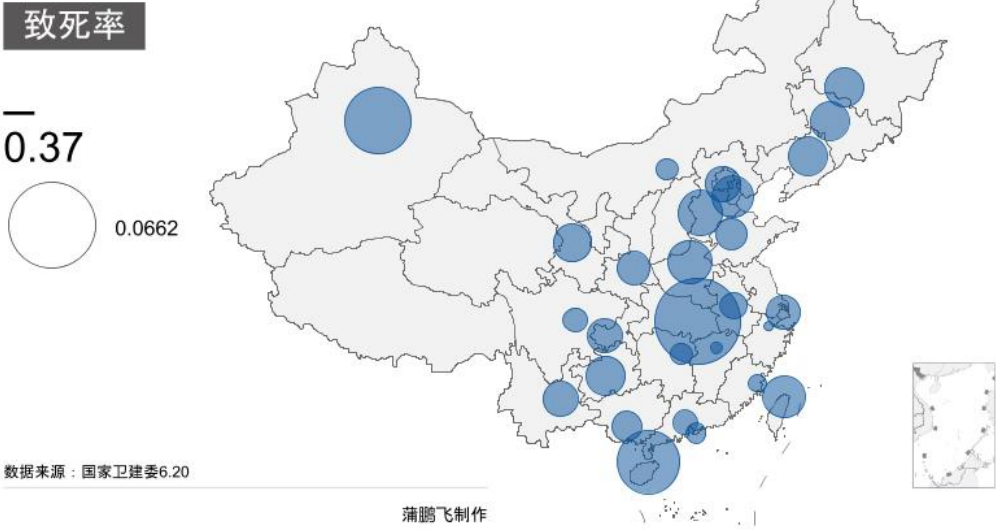


图 17 COVID-19 全国各省致死率地图

虽然西部地区也因为某些原因没有较大的确诊人数基数，但是致死率还是不容乐观，贫乏的医疗物资仍然是制约疫情防控的重要因素，不过，最重要的一点是，西北地区的商业活动以及人流量不如中部，南部活跃，如果能好好把控进出通道还是能有效地进行遏制疫情的扩散，而且地广人稀，极端复杂地天气都给病毒的传播造成了一定的阻碍。

4.8 COVID-19 确诊人数和致死率治愈率之间的相关性

COVID-19的确诊人数和致死率散点分布

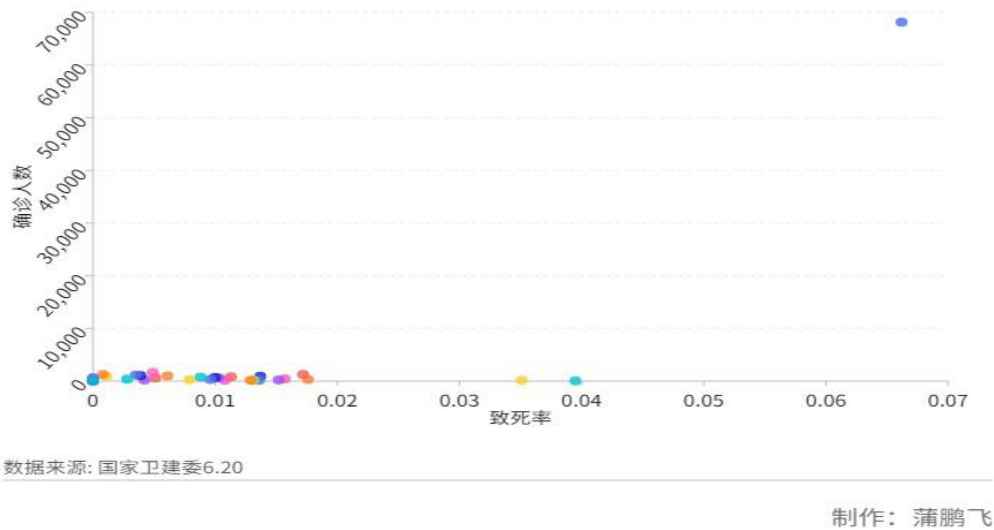


图 18 COVID-19 全国各省确诊人数和致死率散点分布

COVID-19的确诊人数和治愈率散点分布

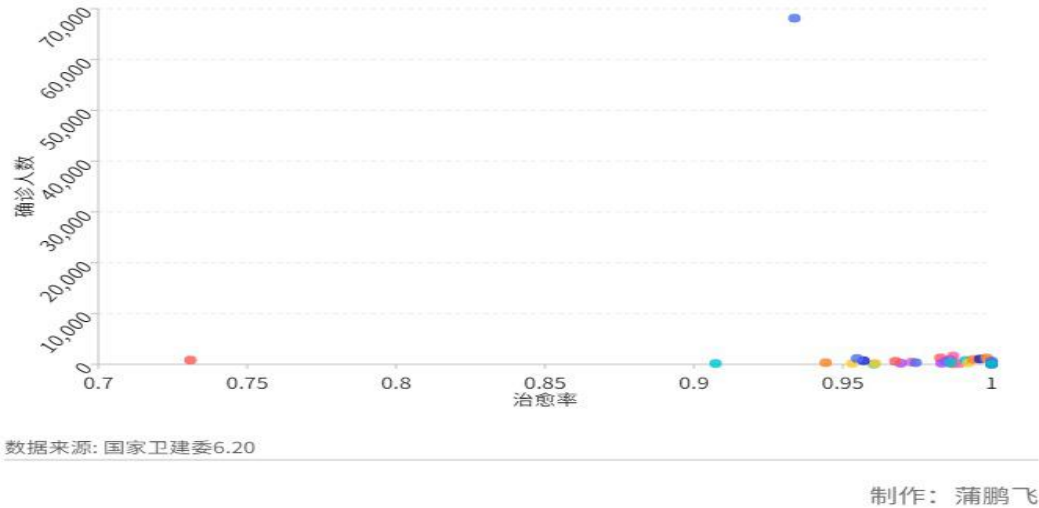


图 19 COVID-19 全国各省确诊人数和致死率散点分布

通过分析确诊人数的治愈率和致死率，以及全国各地疫情数据生成散点图后我们基本可以得出这么一个结论，截止数据收录以前，疫情发展到目前的特征趋势是高治疗率和低死亡率，虽然有人会说死亡人数很多，但是放在湖北人口基数面前还是可控

的，并且很有成效，可能很大程度上得益于第一时间各种行动的部署才显得卓有成效。

5 人流量对 COVID-19 传播的影响

5.1 COVID-19 湖北省各地级市死亡人数（雷达图）



图 20 COVID-19 湖北省各地级市死亡人数雷达图

由图可知孝感市的死亡人数最多，其次黄冈市，目前没有发现死亡人数和确诊人数有什么紧密的联系，但是随着可视化的进一步深入，再次调用更加详细的数据，配合使用热力图也许能够进一步发现其中所存在的一些联系。

5.2 小范围人口迁徙对疫情的影响

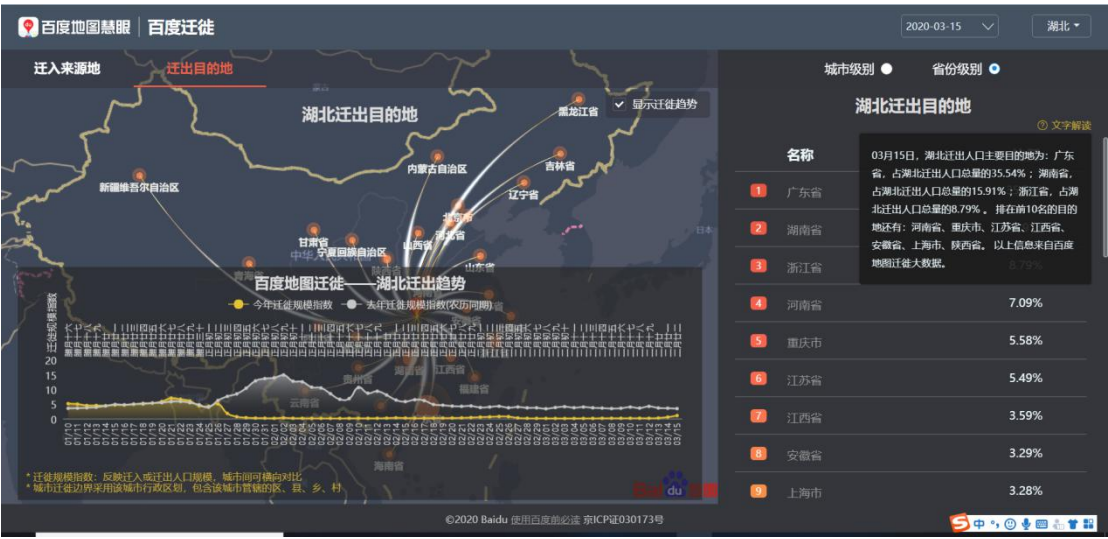


图 21 COVID-19 百度慧眼湖北迁出人流

根据百度地图慧眼数据可视化的显示湖北省在 1.21 日迁出人口达到了峰值，可能与春运有很大的关系。但是，对于疫情来讲，这个时间是封城前两天，由于一些捕风捉影的信息让人们产生了恐慌心理。

另外，上文中提到广东的疫情相较于其他地方较为反常，但是作为沿海地区，提供给人们大量的就业机会，并且，越来越多的“反向春运”成越来越多的人热衷的一种生活方式，还有外出东南亚旅游等等，给广东防疫带来了一定程度上的压力。

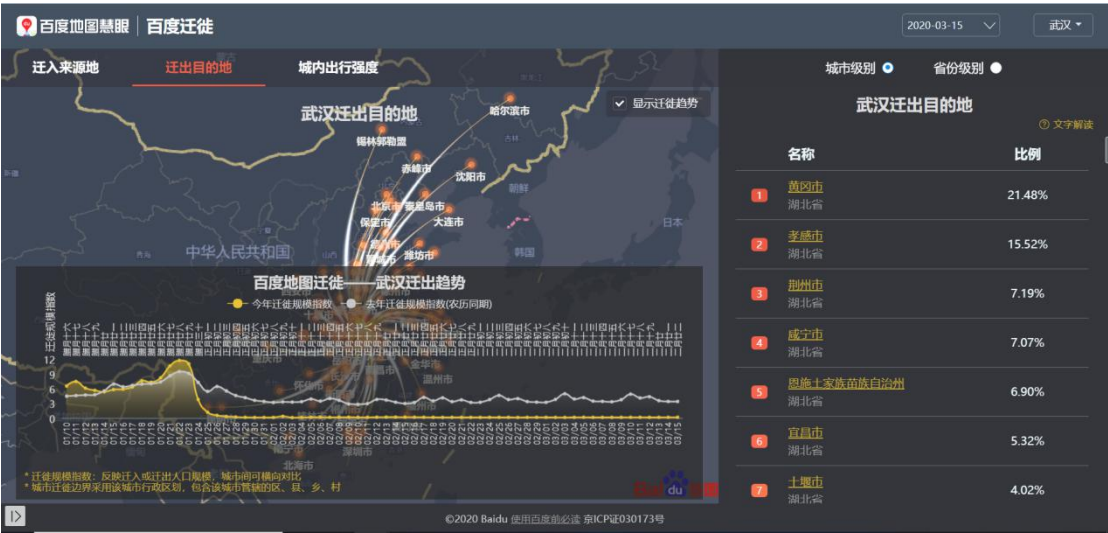


图 22 COVID-19 百度慧眼武汉迁出目的地

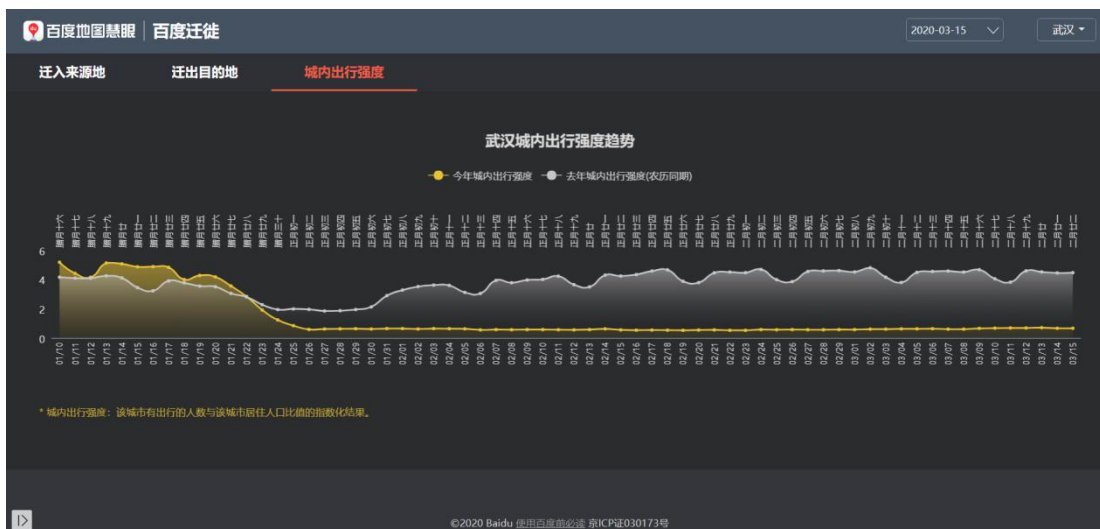


图 23 COVID-19 百度慧眼武汉城内活跃度

武汉是国内最早发现 COVID-19 通过分析武汉的人口出行，人们在封城前一天大量地涌出武汉，其中黄冈达到了 23.9%，孝感 11.2%，我们上面做的雷达图中清楚的反映出湖北各个地级市死亡人数，其中人数最多的是孝感，其次是黄冈，人数都是千级单位，这也给我们反应了防控工作中经验一定程度上不足，作为超大型城市，在封城前夕，应该是可以通过大数据实时监测平台发现这一特点。



图 24 COVID-19 百度慧眼孝感外来迁入人员信息

孝感迁入人数最多的城市是武汉，可以看出来武汉当时的情况已经达到了一种非常严重的地步，可以看出，封城绝对是非常正确的举措。对于决策者是一次极大地考验。

6 相较于 SARS, 大数据对 COVID-19 的积极影响

相较于 17 年前的 SARS，此次疫情表现出更强的传播性，感染人数曲线更为陡峭，对于疾病防控提出更高挑战。好在 17 年来，中国在疫情防控方面已建立了更加完备的制度体系、保障策略、应对措施，信息披露也更加及时透明，再加上大数据等创新科技的快速发展，在疫情防控工作中起到重要作用。

目前已有微信、360 等互联网平台上线“确诊患者交通工具同乘查询系统”、“疫情数据实时更新系统”、“发热门诊分布地图”等功能。

另外，支付宝以及微信小程序支持“健康码”的获取，为火车站、学校等高密集，人流量大的地方缓解了较大的压力

6.1 大数据追踪传播路径

追踪移动轨迹、建立关系图谱，在大数据技术日渐成熟的今天已不是新闻，在位置数据方面，除了航空、铁路、公路、轮渡等交通部门统计的出行数据外，在用户授权的前提下，中国移动、中国联通、中国电信三大运营商基于手机信令能够有效定位用户的手机位置，互联网企业也可以通过 APP 授权调用用户手机位置数据。

此外，地图、打车等 APP 提供的移动出行服务，电商、外卖平台等 APP 内的送货地址数据，以及移动支付位置数据等，也可以作为位置数据的有效补充。而关系图谱则可通过各类社交平台、通信网络、通话记录、转账记录等数据搭建。

将不同时间段的授权位置数据进行纵向串联，能够有效绘制出手机持有者的移动轨迹。这类个体数据，正如李兰娟院士提到的，可以用于追踪被感染者的疾病传播路径、定位感染源，配合关系图谱更可锁定被感染者曾经接触过的人群，以便及时采取隔离、治疗等防控措施，避免疫情更大范围扩散。为防控春运返程高峰时可能发生的传染事件提供有效工具。

而将这些个体数据集合形成的群体数据，则能够清晰显示重要疫区的人员流入及流出方向、动态及规模，如百度、腾讯等互联网企业均已基于授权数据制作此次春运期间的人口迁徙地图，可据此观察各城市的人口流入、流出状况，尤其是重点疫区人口流出方向。

这些数据有利于定位疫情输出的主要区域、预测地区疫情发展态势、预测地区潜在染病人群，为疾病防控部门及地区政府分类制定春运返程计划、有针对性地出台交

通管制措施等提供决策支撑。

除此之外，将同一时点不同个体的位置数据进行横向整合，还能够清晰展现出特定时间点曾经到过疫情高风险地区的人群，并可据此监测人群密度及动向。

如某大数据公司以疫情始发地为分析重点，利用位置数据定位自 2019 年 11 月起曾经去过疫情始发地的人，为潜在感染者的发现及自我隔离等提供信息参考。而这些人群密度地图、高染病区域地图、地区交通管制措施等数据信息还能个人规划返程路线提供有效参考。

6.2 大数据构建疫情发展模型

疫情还会传播多久？感染者还会大幅增加吗？哪里感染风险高？何时能够进入安全期？传染源都有哪些？

要解决这些问题，需要找出关键影响因素、分析疫情传播特征、搭建疫情发展模型，这其中大数据可发挥关键作用。

首先是优化数据采集。在大数据技术广泛应用之前，医疗数据采集具有明显的滞后性，这对在疫情传播早期阶段快速获取传播数据、分析疫情传播机理造成制约。而借助于医疗数据联网、各类智能设备数据归集渠道等，大数据时代的疫情传播数据采集更为及时、准确，可定位到个体、某一具体街区等，为疫情发展模型的搭建提供数据基础。

其次是丰富数据维度。除医疗数据外，疫情传播往往还涉及气候温湿度、地质、交通、社会行为、城市卫生等多维度因素影响，大数据技术的发展使得这些影响因素均可以数据形态展示，同时使得多维度、大规模的数据处理成为可能，可实现上万量级的影响因子建模，这极大地丰富了疫情发展模型的分析维度，对于定位疫情传播的关键影响因素，并据此提出针对性防治建议有重要作用。

最后是模型优化训练。海量数据基础为疫情发展模型提供丰富的优化、训练素材，模型的不断迭代对于优化模型参数、提升模型预测精准度有重要意义。

6.3 大数据助力资源配置

疫情在全国范围内的传播引发对医疗物资、生活物资等多维度资源的需求激增，而春节期间有限的生产供应能力难以在短时间内快速满足。基于此，提升物资调配效率，以有限资源保障医疗救助工作顺利开展，是当前疫情防控的重点。

现阶段各类资源需求信息的发布较为分散，以医疗物资保障为例，陆续有医院通过各自网站、媒体、社交平台等对外发布短缺物资清单。

但公布渠道的分散化，不利于防控机构统筹监测，也不利于捐赠者查询，还有可能出现因医院知名度不同而产生的物资获取差异，或重复捐赠等问题，不利于资源有效调配及使用。

基于此，已有志愿者基于公开需求数据爬取等方式建立资源对接平台，如“湖北医疗物资需求信息平台”等，将医疗资源需求按照城市、医院、类别等维度分类呈现，通过数据抓取等技术手段，展示需求物资名称、需求数量、联系方式及物资运输方式等信息，并支持信息查询，同时在后台统计整体需求数据，时时更新。

这有利于物资短缺信息的及时、有效展示，提升资源调配机构及捐赠者的信息获取速度，提高资源配置效率。而针对历史短缺数据的归集整理以及对资源对接时效的统计分析，也可帮助有关部门预测未来资源需求情况，科学筹划下阶段资源供应及调配。

7 结语

新冠肺炎疫情来势汹汹，全面考验国家及民众的危机应对能力，与 17 年前的 SARS 相比，中国在此次疫情防控工作中展现出了更高的医疗救治水平、更快的防疫反应速度、更透明的信息披露机制、更迅速的数据报送体系，同时将大数据等新一代创新科技，广泛应用于疫情追踪溯源、路径传播、发展模型预测、资源调配等领域。

8 参考文献

- [1] 中国计算机学会大数据专家委员会，中国大数据技术与产业发展报告(2013).[2018-1-5].<http://www.bigdataforum.org.cn/ccf/navigation?id=51&dtype=1>.
- [2] 中国计算机学会大数据专家委员会。CCF 大专委 2017 大数据发展趋势预测.[2018-1-5].<http://www.bigdataforum.org.cn/ccf/ccf/navigation?id=51&dtype=2>.
- [3] 陈为，沈则潜，陶煜波.数据可视化[M].北京：电子工业出版社，2012.
- [4] W3School 之中文文档.<http://www.w3school.com.cn/>.
- [5] 百度慧眼开放平台.<https://huiyan.baidu.com/>.
- [6] Python 爬虫 Requests 库.<http://www.python-requests.org/en/master/>.

- [7] Github 开放平台.<https://github.com/>
- [8] 国家卫健委开放平台.<http://www.nhc.gov.cn/>
- [9] 董付国.Python 程序设计基础[M].北京.清华大学出版社, 2015
- [10] Python 官方在线帮助文档.<https://docs.python.org/3/>.

9 致谢

首先,要感谢大环境,在互联网快速发展的今天,无论是找资料还是找工具,都是非常方便的,CSDN、Github 都提供给了我们非常便利的学习环境。

其次,非常感谢在我写论文时候帮助我的舍友,在我苦苦寻找湖北省的疫情前夕的人口出行数据的时候遇到了困难,官方没有公开的数据,因此没有办法爬取,但是舍友帮我找到了百度慧眼,才让我对疫情和人口迁徙的关系的分析得以继续。

谢谢对我论文帮助的每一个人!