新冠病毒数据分析系列公益讲座 COVID-19 Data Analysis Webinars

第二讲中美疫情研究数据解读

主讲人：鲍曙明，中国数据研究所

主持人：关蔚禾，哈佛大学

2. Understanding the COVID-19 Data of the US and China

Speaker: Shuming Bao, China Data Institute

Chair: Wendy Guan, Harvard University

北京时间 2020年5月7日 上午9–10点

议题 Topics

1. 项目背景 Background

2. 数据来源 Data Sources

3. 数据解读 Data Reading

4. 数据共享 Data Sharing

5. 总结与讨论 Summary and Discussions

1 背景: 疫情数据研究共享存在的常见问题

(1) 数据序列断裂化：数据标准不断变化，缺少系统说明;

(2) 公开数据短期化：缺少历史积累;

(3) 数据存储零散化：不利长久保存;

(4) 数据来源差异化：不利数据比较;

(5) 数据内容离散化：不利发挥综合价值;

此外数据与方法分离，数据与工具分离，数据产权、数据私密性、敏感性，以及大数据与研究成果如何共享也是此类数据研究共享面临的一些核心问题。

1 背景: 研究项目Resources for COVID-19

项目背景： “新冠病毒疫情资源与全球研究”

https://projects.iq.harvard.edu/chinadatalab/resources-covid-19

项目目标: to provide an information infrastructure for the global study of novel coronavirus (COVID-19)

项目任务：Objectives:

- 疫情研究数据长久**保存**：To establish a **permanent collection** for the COVID-19 study at local, regional and global levels with information collected and integrated from different sources
- 提供方法与技术**支撑**：To facilitate the quantitative research on spatial spread and impacts of COVID-19 study with advanced **methodology and technology**
- 推动研究**合作**：To promote collaborative research in the COVID-19 study with the **cloud platform**
- 促进教学研究**应用**：To promote the use and sharing of data sources in **teaching and research** with DataVerse/WorldMap/CDL
- **团队**能力建设：To build the **capacity** for future collaborative projects

1 背景：研究团队/Research Teams

哈佛大学地理分析中心，美国

CGA

The Center for Geographical Analysis (CGA) at Harvard University. Its core mission is to support research and teaching in all disciplines across Harvard University with emerging geospatial technologies.

中国数据研究所，美国

CDI

The China Data Institute, a Michigan based not-for-profit organization. It aims to promote the use and sharing of China data; support quantitative research on China in social science, digital humanity and other research subjects.

武汉大学社会地理计算中心，中国

GCSS

The GeoComputation Center for Social Science at Wuhan University. It promotes the scientific research on the theory and method of spatial data in scientific research, personnel training, international cooperation and social practice.

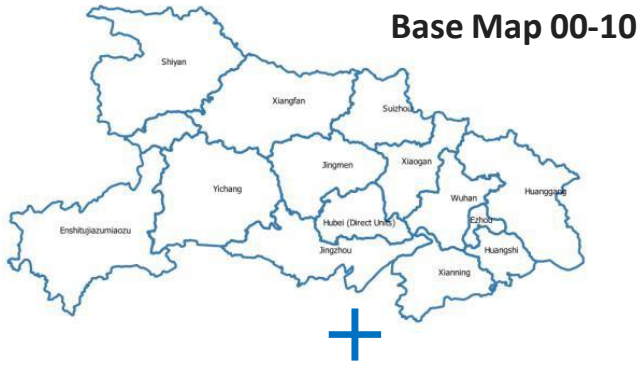
RMDS实验室，美国

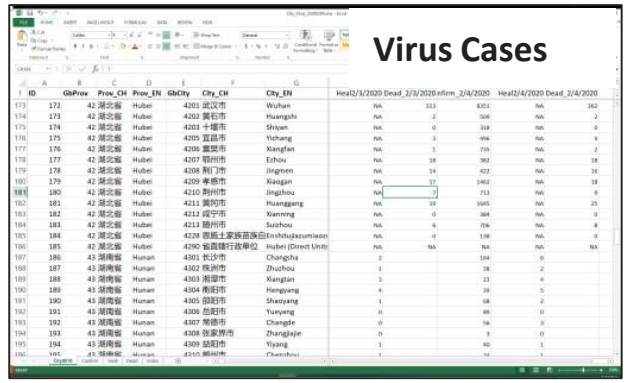
RMDS

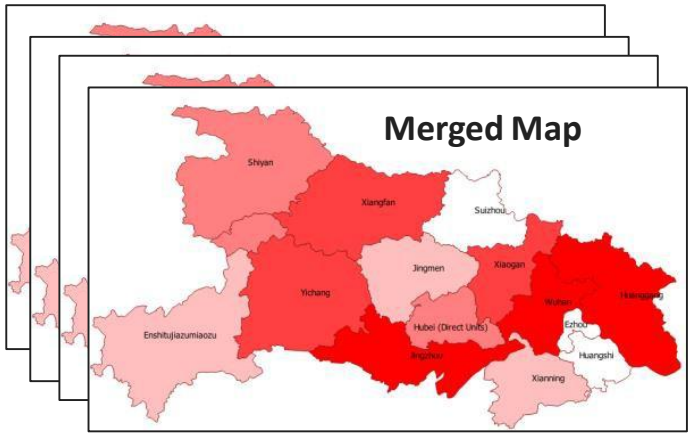
The RMDS Lab is designed to serve a global community of companies and data scientists, empowering them to reap the benefits of big data, artificial intelligence, and machine learning.

2 数据来源/Data Sources

- **基础数据 / Base data**
  - Base maps (province, prefecture and county)
  - Census data (population census and economic census)
  - Statistics (province, prefecture and county)
  - Others
- **病例数据 / Virus data**
  - Virus reports from gov't (country, province and city)
- **迁徙数据 / Migration data**
  - Daily floating population data (Baidu)
  - Migration data from census (NBS) 国家统计局
- **医疗机构数据 / Health facility data**
  - POI data (高德地图)
  - Economic Census (NBS)
- **社交媒体数据 / Social media data**
  - GDELT (Global Database of Events, Language, and Tone)
  - Twitter
- **论文 / Papers**
- **报告 / Reports**
- **政策法规 / Policies and regulations**







2 数据来源：基础数据

中国数据

Province
Prefecture
County
Township
ZIP
Sq Km Grid

- **Government Statistics 政府统计**
  - Provincial Statistics (1949 -)
  - City Statistics (1996 -)
  - County Statistics (1997 -)
- **Population Census 人口普查**
  - Census 1953
  - Census 1964
  - Census 1982
  - Census 1990
  - Census 2000/2010 (province, city, county, township, GRID)
- **Economic Census 经济普查**
  - Industrial Census 1995 (province, city, county, ZIP)
  - Basic Unit Census 2001 (province, city, county, ZIP)
  - Economic Census 2004/2008 (province, city, county, ZIP)
- **Establishments**
  - more than 7 millions organizations 基本单位
- **Geography and Environment 地理与环境**
  - Land Use data 土地利用
  - Night-Time data 夜光强度

美国数据

State
Metropolitan
County
CCD
TRACT
BLOCK

- **Population Census** (State, County, Tract, Block, Metropolitan, CCD, Place)
  - Census 1970
  - Census 1980
  - Census 1990
  - Census 2000
  - Census 2010
  - American Community Survey

Economic Census (State, Metropolitan, County)

- **Business Data** (State, Metropolitan, County, Tract, CCD, Place)

- **Establishments** (individual business and organizations)
  - Firm level records (more than 13 million records with 85 fields for each business)

3 数据解读：主要数据集 Data Collection

序号	数据项	Data Sets
2.1	疫情病例数据	Coronavirus cases data
2.2	人口迁徙数据	Population mobility data
2.3	医疗机构数据	Health facilities data
2.4	行为轨迹数据	Trace data
2.5	航线航班数据	Flight data
2.6	社交媒体数据	Social media data
2.7	新闻政策数据	News and Policy data
2.8	高铁班次数据	High-speed train data
2.9	气象气候数据	Meteorological data

3.1 疫情病例数据: 中国疫情数据

原始数据来源：国家卫健委、各省市卫健委、各省市区政府、港澳台官方渠道公开数据

加工数据来源：丁香园 https://github.com/BlankerL/DXY-COVID-19-Data

基础地图：China Data Institute, 2000-2010省市可比行政区划地图, https://chinadatacenter.net/Data/ServiceContent.aspx?id=1623

集成数据下载：哈佛 DataVerse 省市数据 "China COVID-19 Daily Cases with Basemap", https://doi.org/10.7910/DVN/MR5IJN

下载数据格式：压缩文件（rar, zip），表格文件（tab）

省市疫情数据指标：

- 新增病例
- 累积病例
- 新增死亡人数
- 累积死亡人数
- 新增治愈人数
- 累积治愈人数
- 境外输入（when？）

省市基础数据指标：

2000年城镇人口数	2000年户籍人口数
2010年城镇人口数	2010年户籍人口数
2000年乡村人口数	2000年户口登记地在外乡镇街道的总人口合计
2010年乡村人口数	2010年户口登记地在外乡镇街道的总人口合计
2000年总户数合计	2000年户口登记地在本县（市、区）的人口小计
2010年总户数合计	2010年户口登记地在本县（市、区）的人口小计
2000年属家庭户的户数	2000年户口登记地在本省其他县（市、区）的人口小计
2010年属家庭户的户数	2010年户口登记地在本省其他县（市、区）的人口小计
2000年属集体户的户数	2000年户口登记地在省外的人口小计
2010年属集体户的户数	2010年户口登记地在省外的人口小计
2000年总人口数合计	2000年农业户口人口小计
2010年总人口数合计	2010年农业户口人口小计
2000年男性人口数合计	2000年非农业户口人口小计
2010年男性人口数合计	2010年非农业户口人口小计
2000年女性人口数合计	城市地图边界长度
2010年女性人口数合计	城市地图面积

3.1 疫情病例数据: 美国疫情数据

原始数据来源：美国卫生疾控中心、各州、县、市、镇健康署公开数据

加工数据来源：The New York Times，https://github.com/nytimes/covid-19-data; https://usafacts.org/

集成数据下载：哈佛 DataVerse, 州县大都市区数据 "US COVID-19 Daily Cases with Basemap", https://doi.org/10.7910/DVN/HIDLTK

基础地图：China Data Institute, US Geo-Explorer, https://www.chinageoexplorer.com/

美国疫情数据指标：

- 确诊人数
- 治愈人数
- 死亡人数

美国基础数据指标：

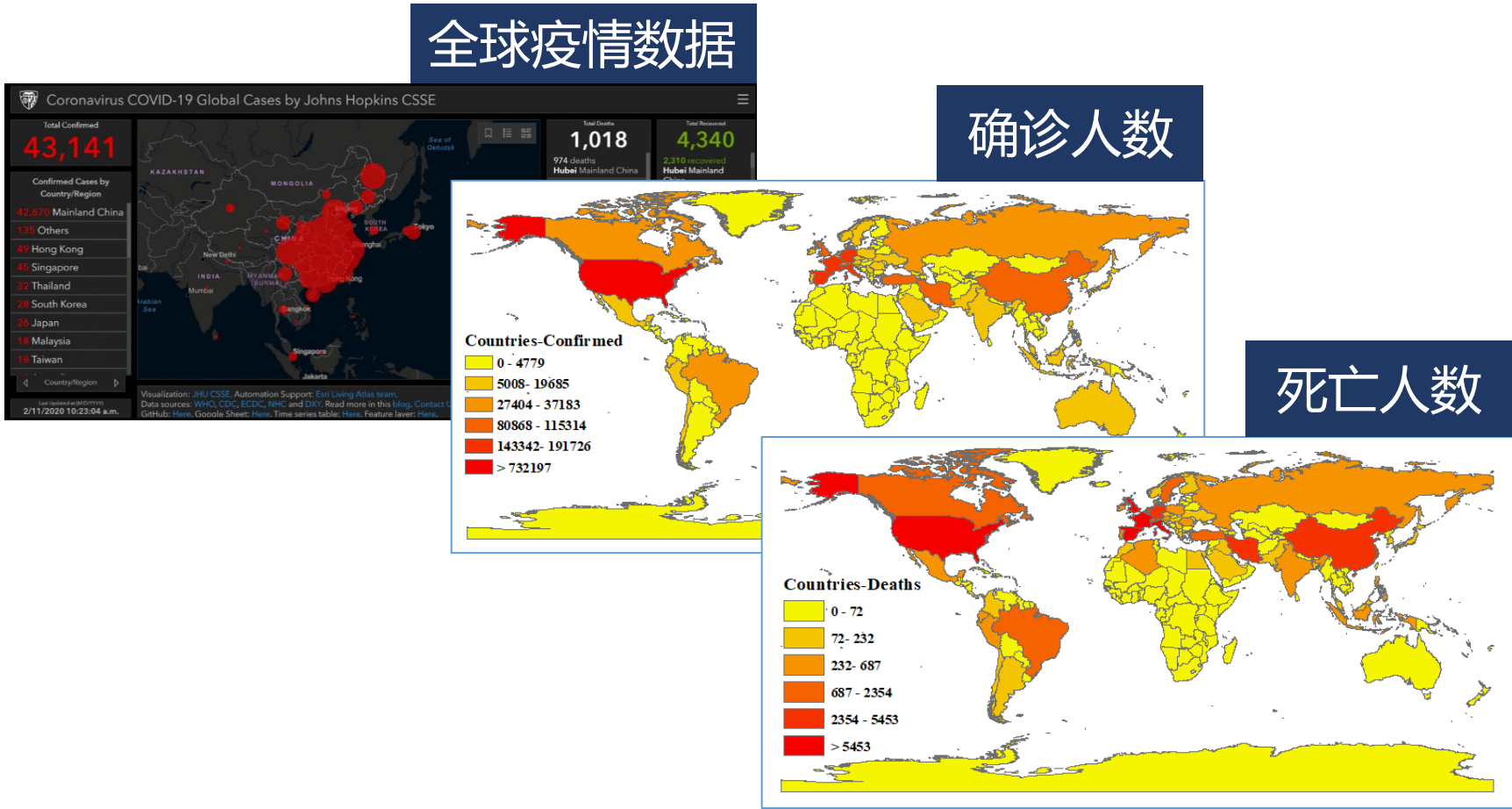
Population in 1970
Households in 1970
Population in 1980
Households in 1980
Population in 1990
Households in 1990
Population in 2000
Households in 2000
Population in 2010
Households in 2010

3.1 疫情病例数据: 全球疫情数据

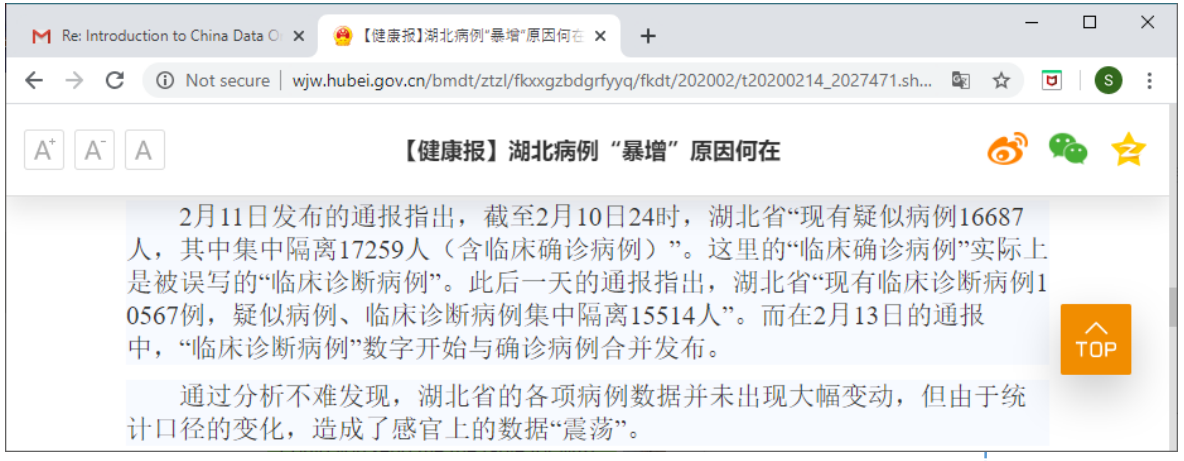
原始数据来源：WHO，CDC等多个来源公开数据，参见https://github.com/CSSEGISandData/COVID-19
加工数据来源：John Hopkins University，https://github.com/CSSEGISandData/COVID-19
集成数据下载：哈佛 DataVerse, 世界各国数据 https://github.com/CSSEGISandData/COVID-19
基础地图： ArcGIS Hub https://hub.arcgis.com/datasets/a21fdb46d23e4ef896f31475217cbb08_1

全球主要疫情数据指标：

- 确诊人数
- 治愈人数
- 死亡人数



3.1 疫情病例数据: 疫情时序数据变化（中国）



【健康报】湖北病例“暴增”原因何在

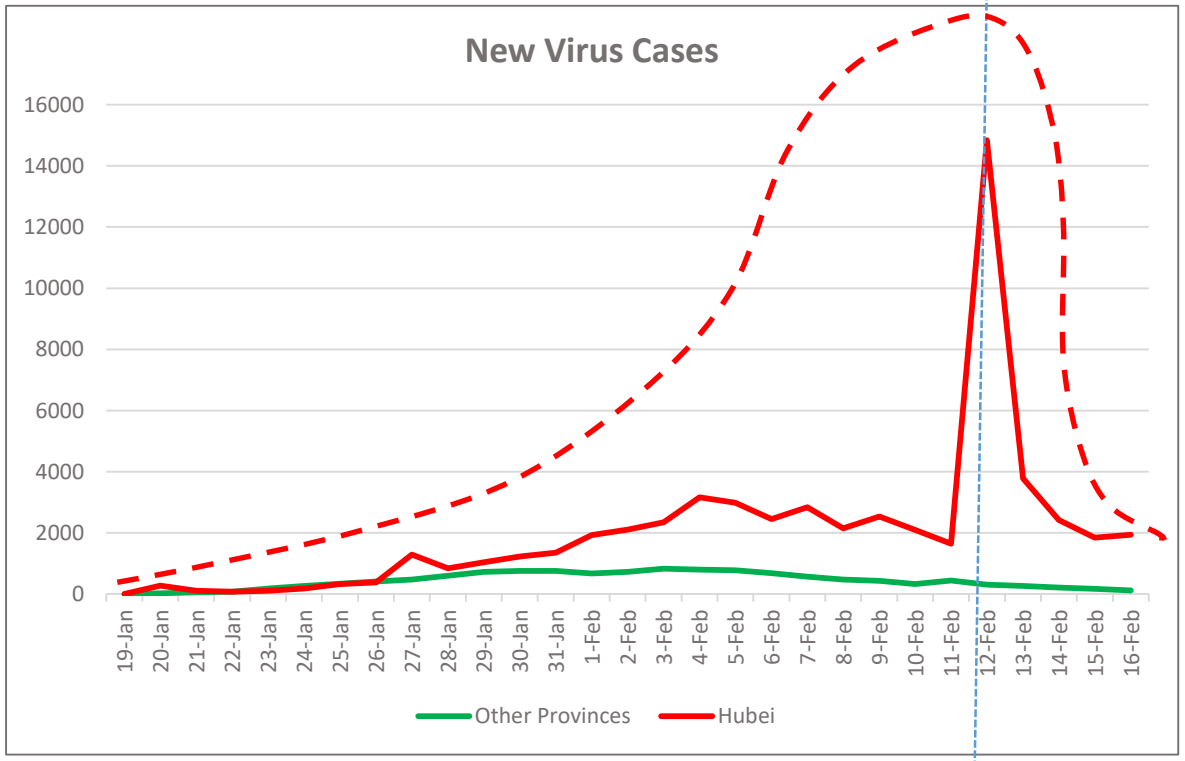
2月11日发布的通报指出，截至2月10日24时，湖北省“现有疑似病例16687人，其中集中隔离17259人（含临床确诊病例）”。这里的“临床确诊病例”实际上是被误写的“临床诊断病例”。此后一天的通报指出，湖北省“现有临床诊断病例10567例，疑似病例、临床诊断病例集中隔离15514人”。而在2月13日的通报中，“临床诊断病例”数字开始与确诊病例合并发布。

通过分析不难发现，湖北省的各项病例数据并未出现大幅变动，但由于统计口径的变化，造成了感官上的数据“震荡”。

湖北省委、武汉市委主要领导职务调整

发布时间：2020-02-13 12:30 来源：湖北日报 【字体：大 中 小】

2月13日，湖北省召开全省领导干部会议。中央组织部副部长吴玉良同志出席会议并宣布中央决定：应勇同志任湖北省委委员、常委、书记，蒋超良同志不再担任湖北省委书记、常委、委员职务；王忠林同志任湖北省委委员、常委和武汉市委书记，马国强同志不再担任湖北省委副书记、常委、委员和武汉市委书记



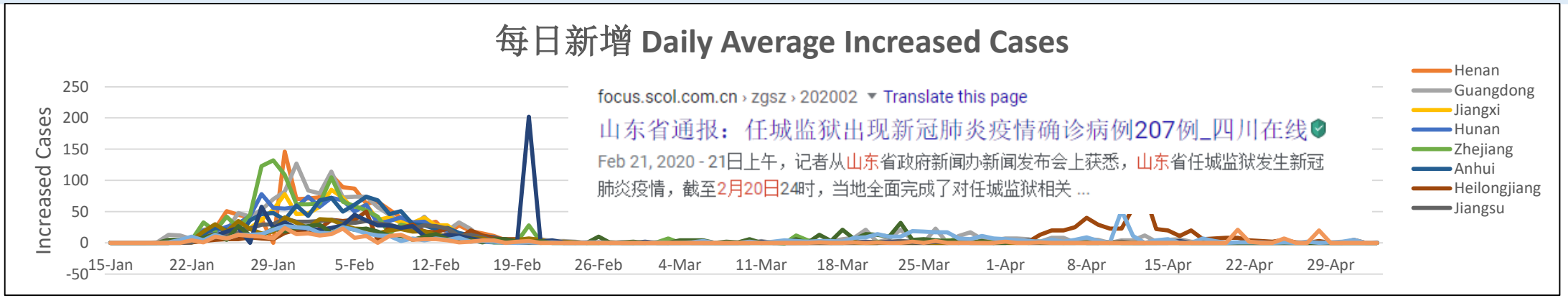
影响确诊人数时序指标变化主要因素：

- 检测标准
- 检测技术
- 防控强度
- 领导更替

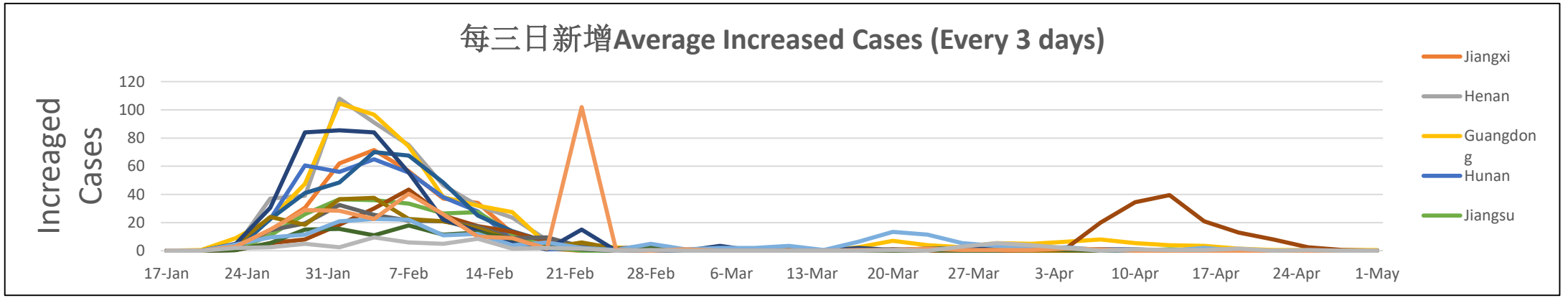
确诊人数与实际感染人数差异变化：

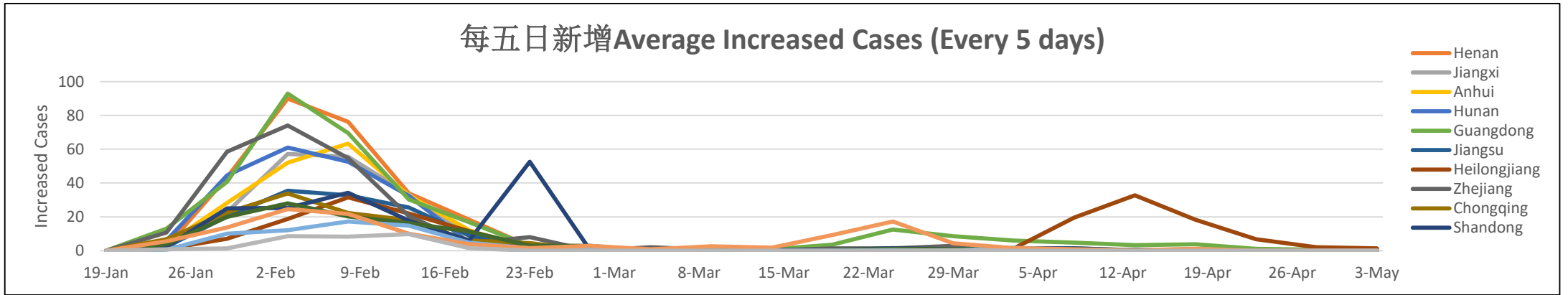
- 早期确诊人数远远小于实际感染人数
- 后期确诊人数趋于接近实际感染人数

3.1 疫情病例数据: 疫情时序数据变化（中国）



focus.scol.com.cn › zgsz › 202002 ▾ Translate this page
山东省通报：任城监狱出现新冠肺炎疫情确诊病例207例_四川在线
Feb 21, 2020 - 21日上午，记者从山东省政府新闻办新闻发布会上获悉，山东省任城监狱发生新冠肺炎疫情，截至2月20日24时，当地全面完成了对任城监狱相关 ...

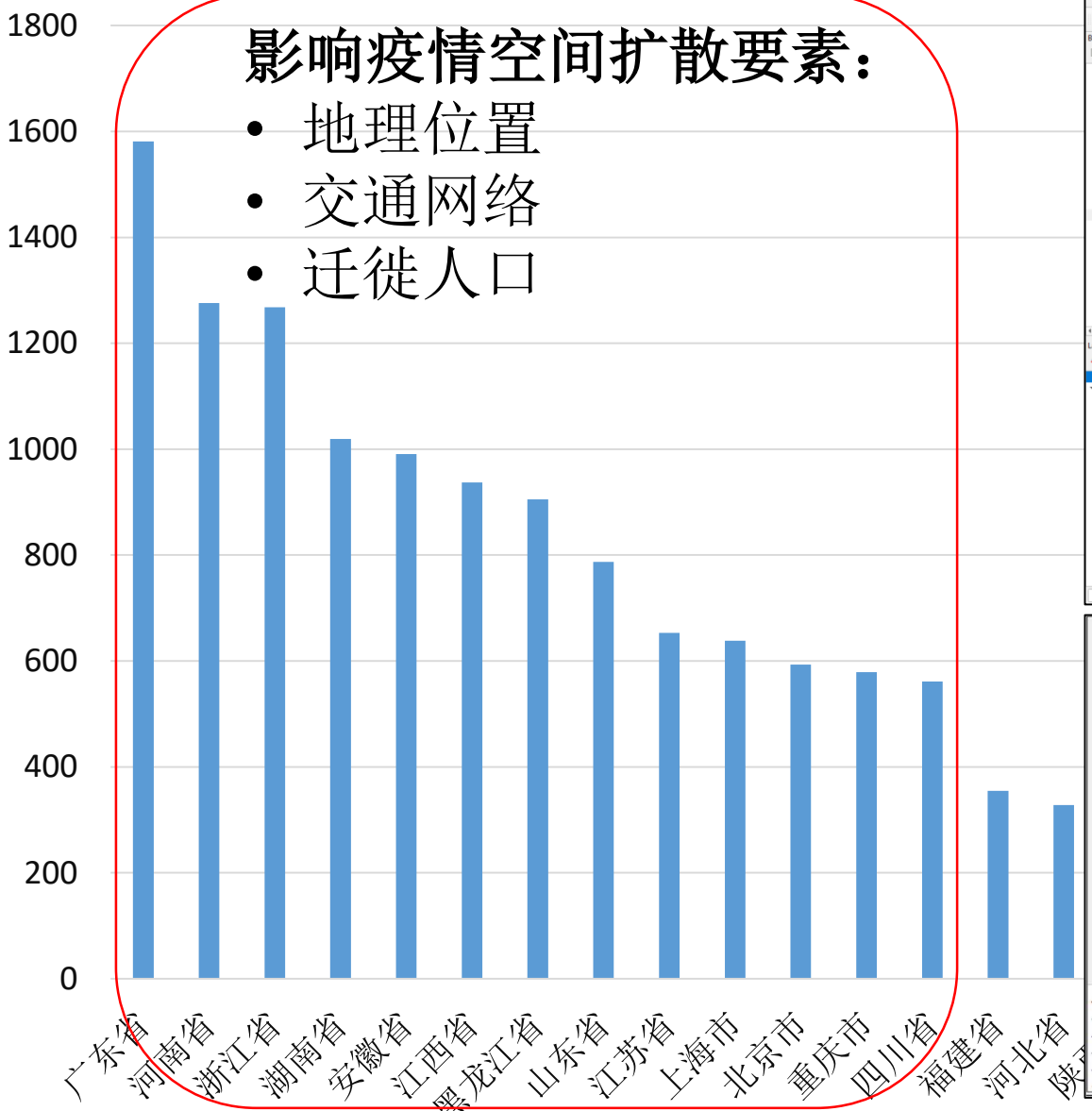


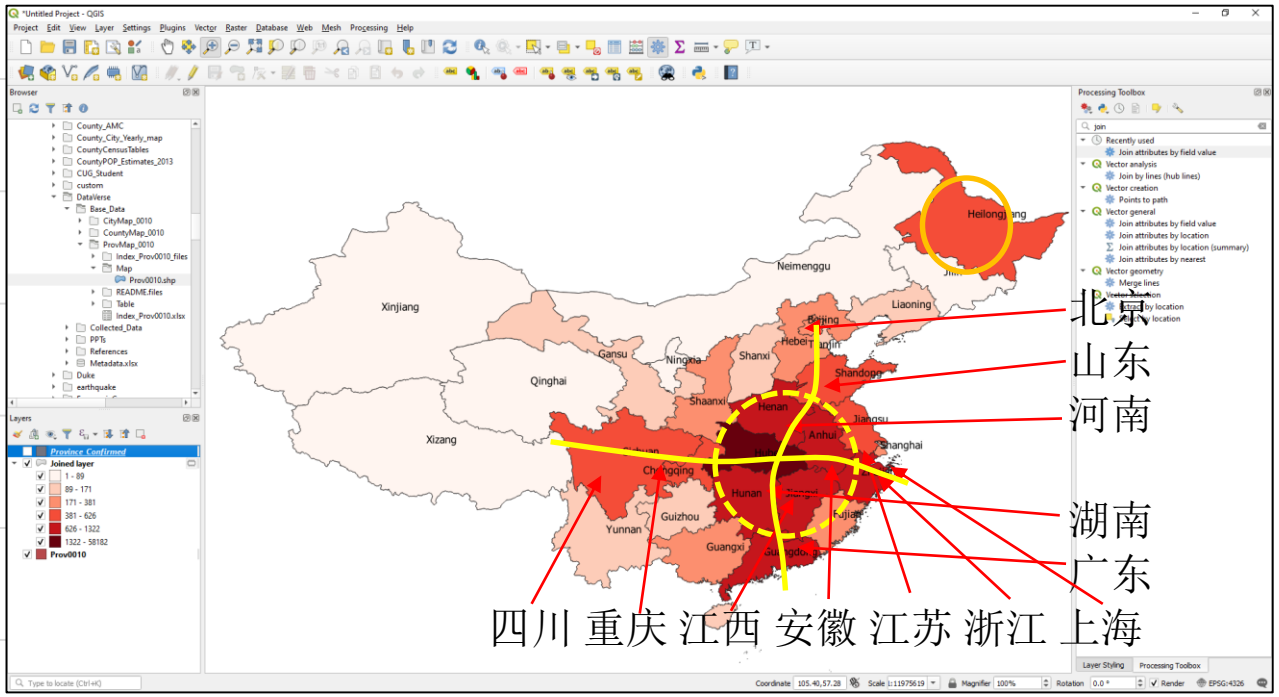


3.1 疫情病例数据: 疫情空间变化与交通网路

影响疫情空间扩散要素：

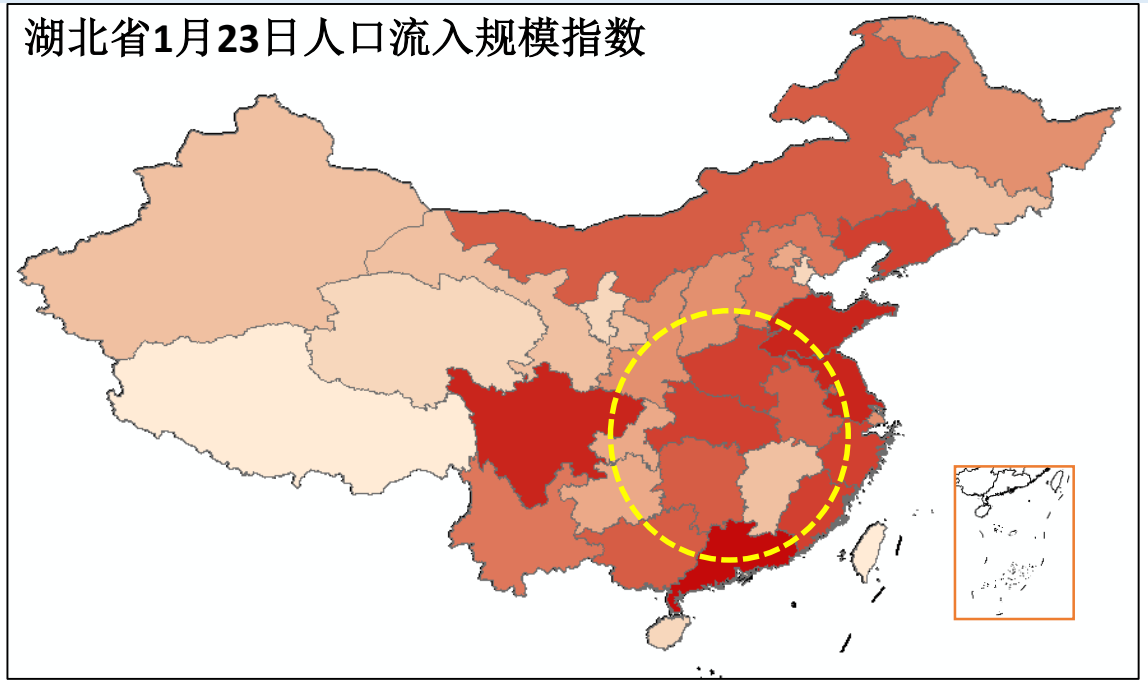
- 地理位置
- 交通网络
- 迁徙人口



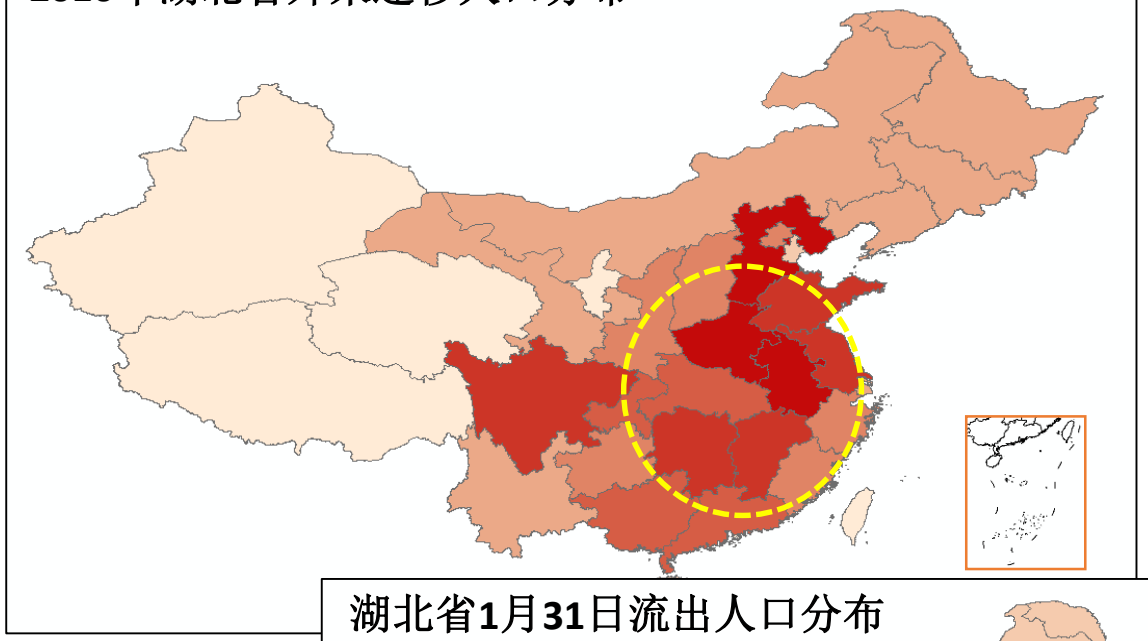


3.1 疫情病例数据: 疫情空间变化与人口迁徙

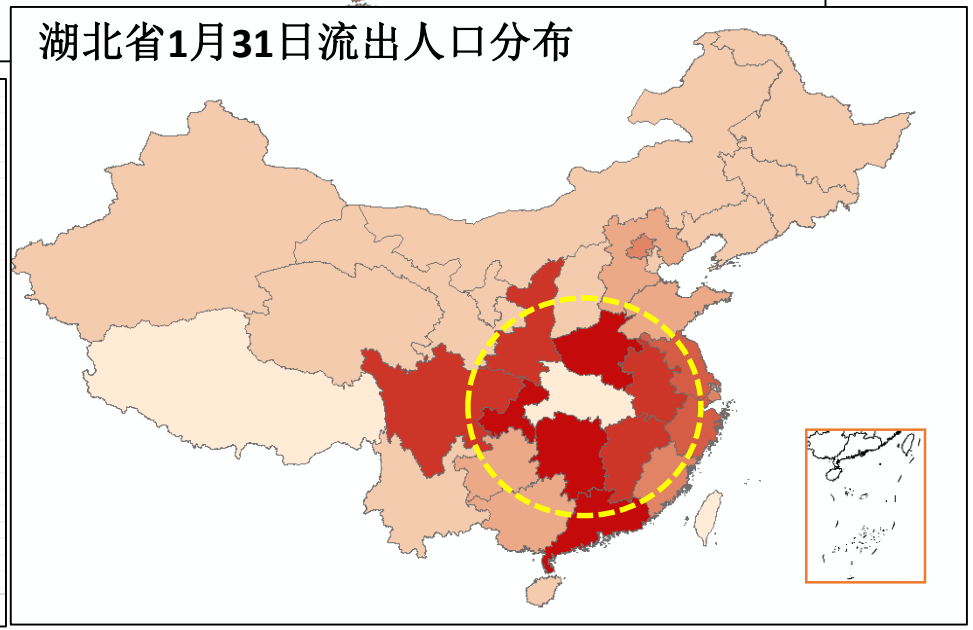
湖北省1月23日人口流入规模指数



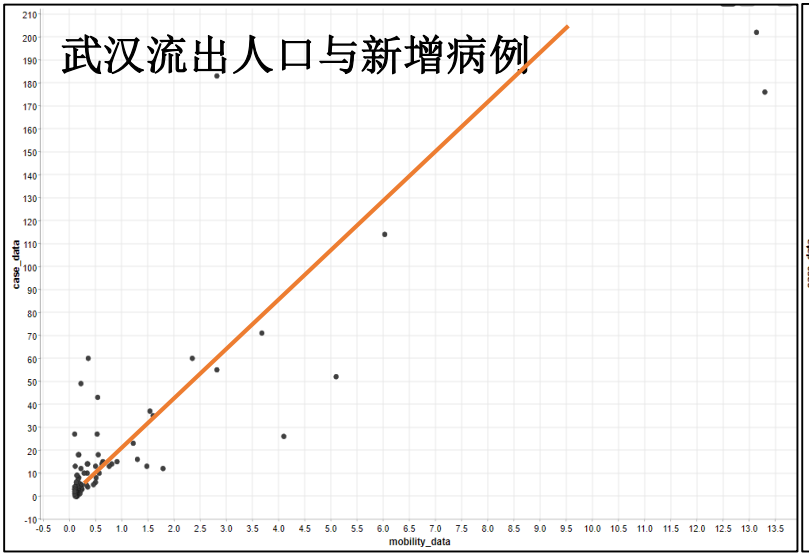
2010年湖北省外来迁移人口分布



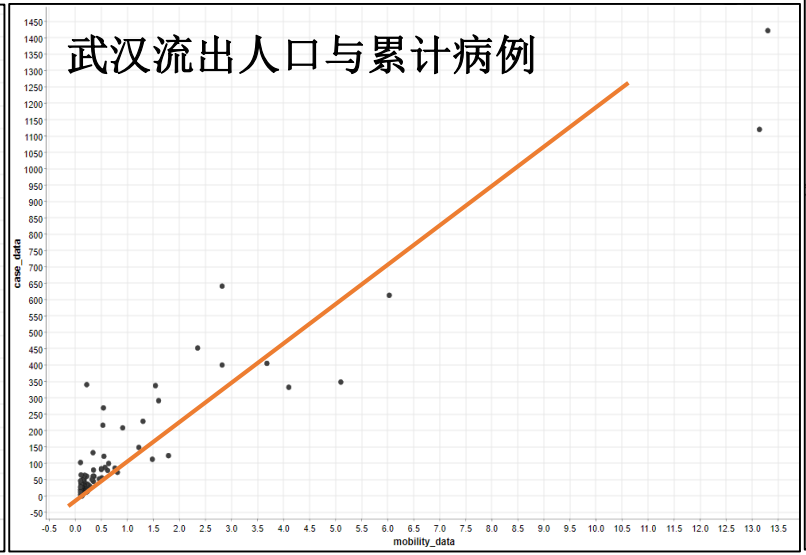
湖北省1月31日流出人口分布



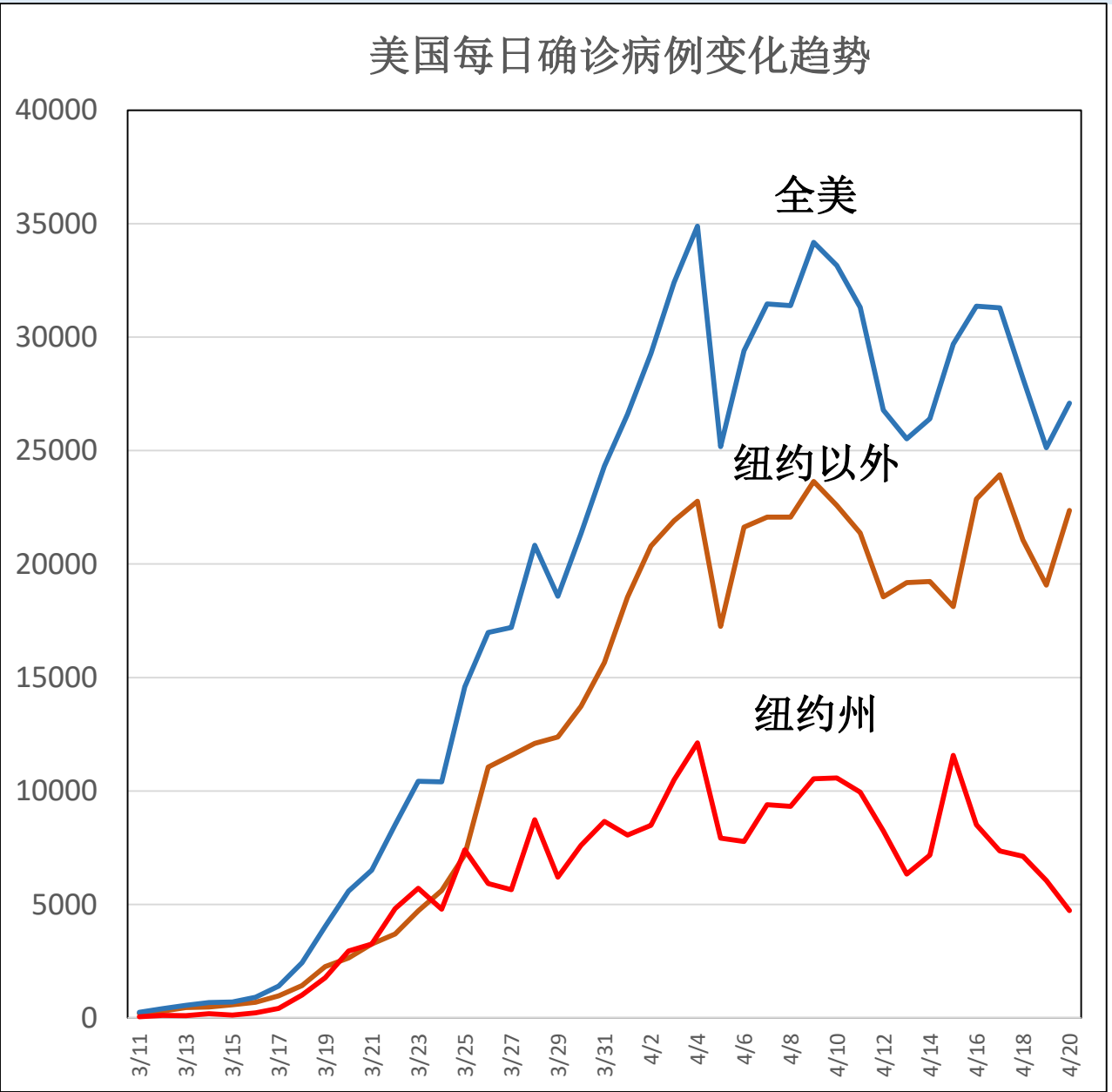
武汉流出人口与新增病例

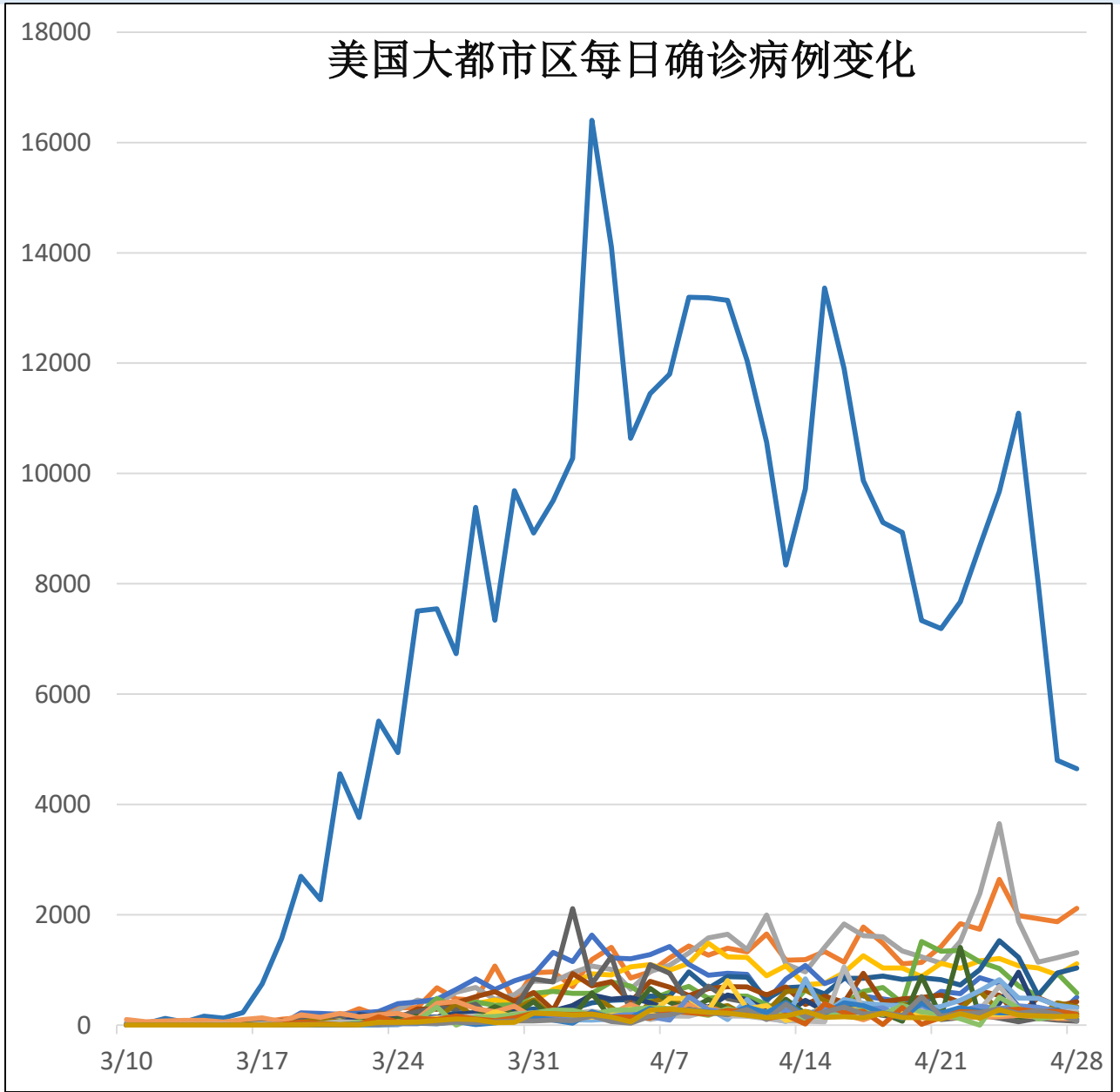


武汉流出人口与累计病例

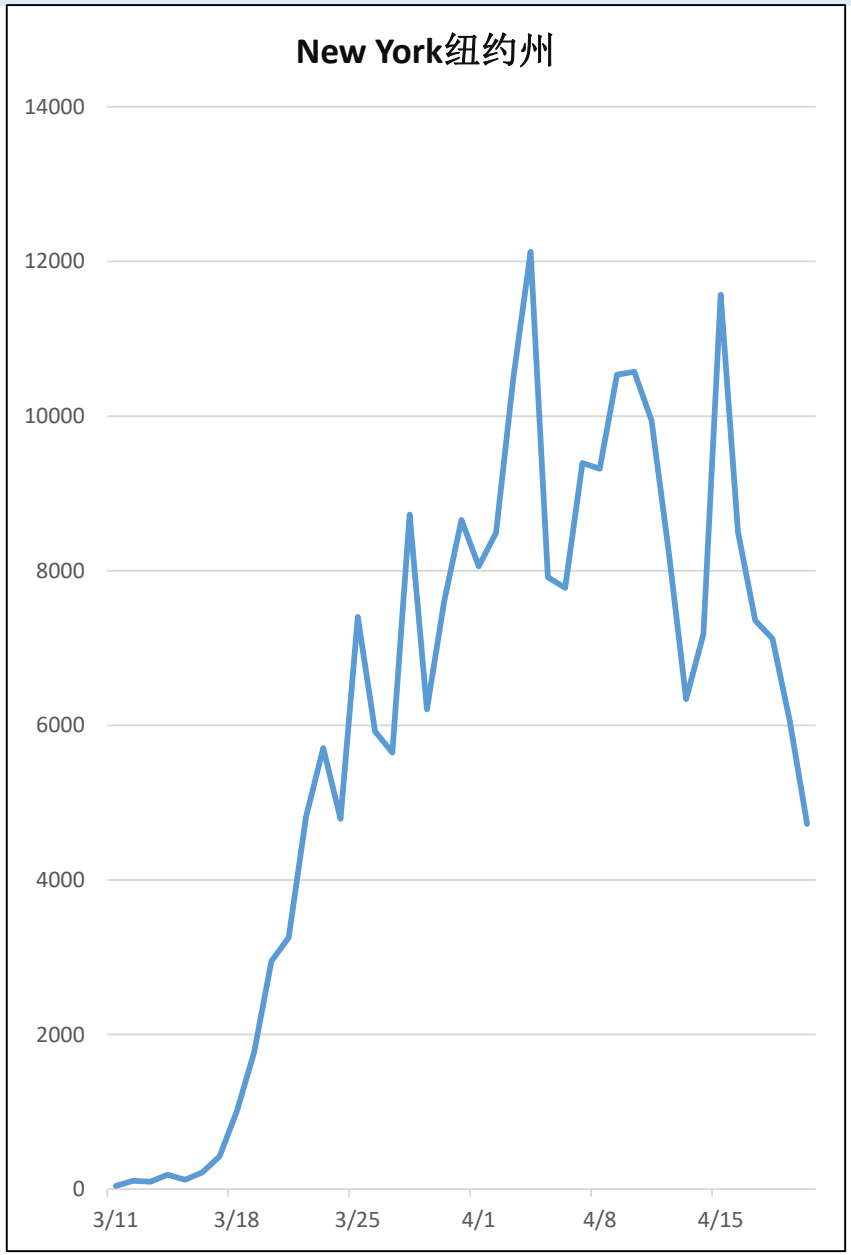


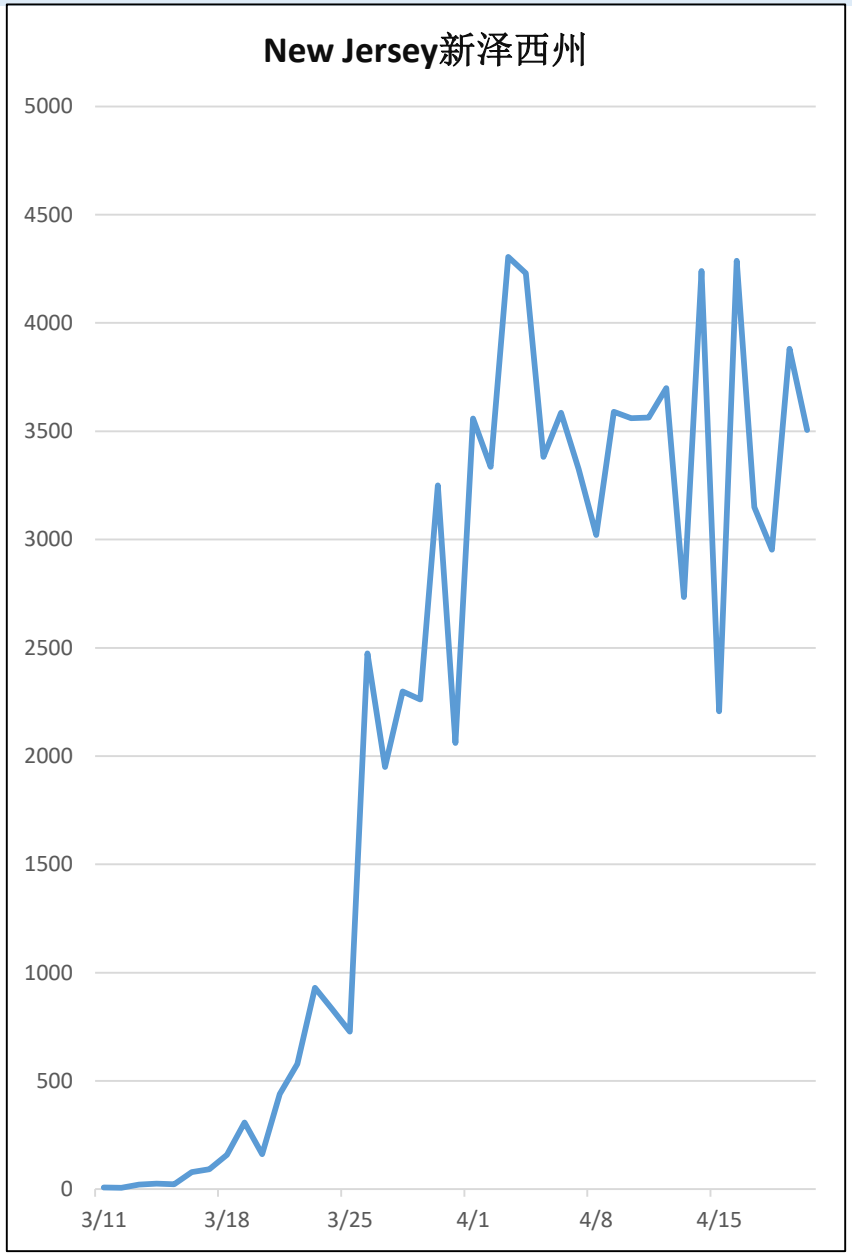
3.1 疫情病例数据: 疫情数据时间变化（美国主要州）

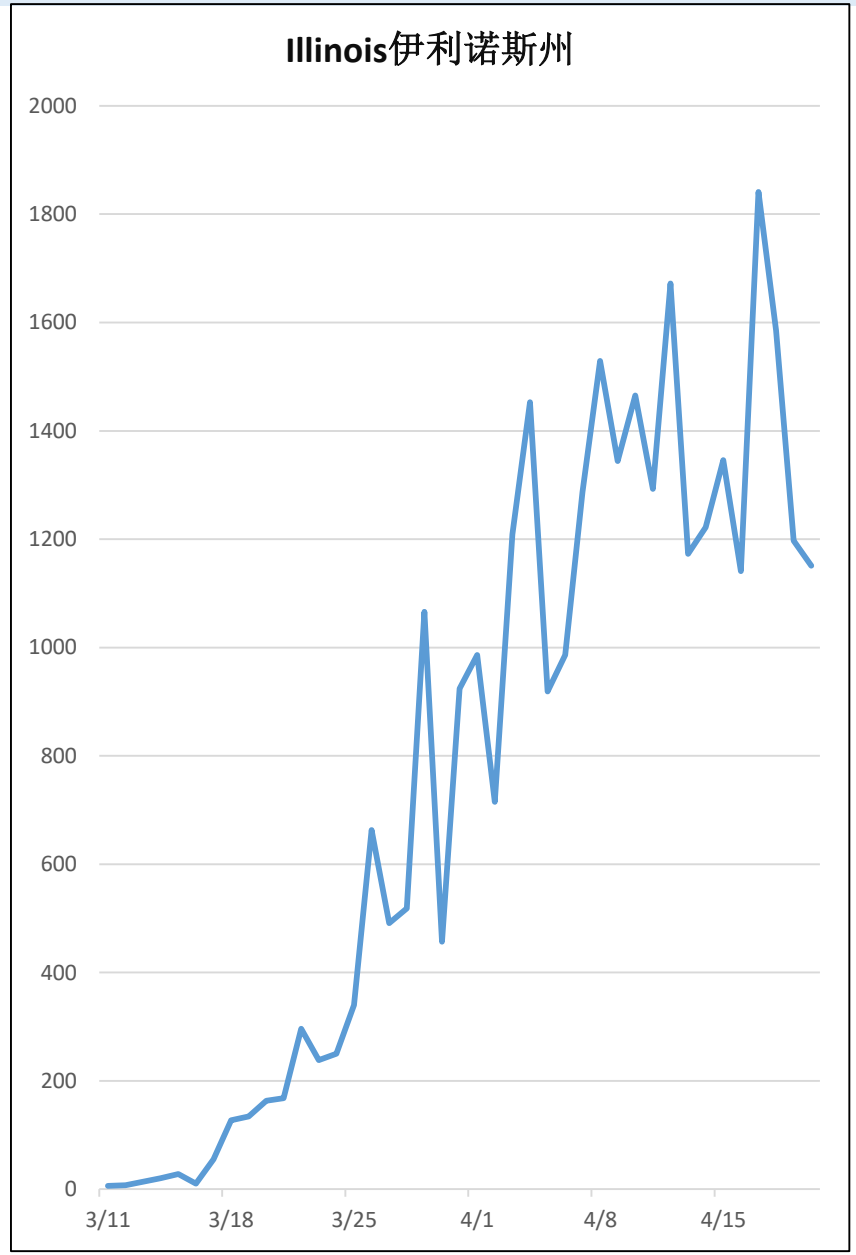




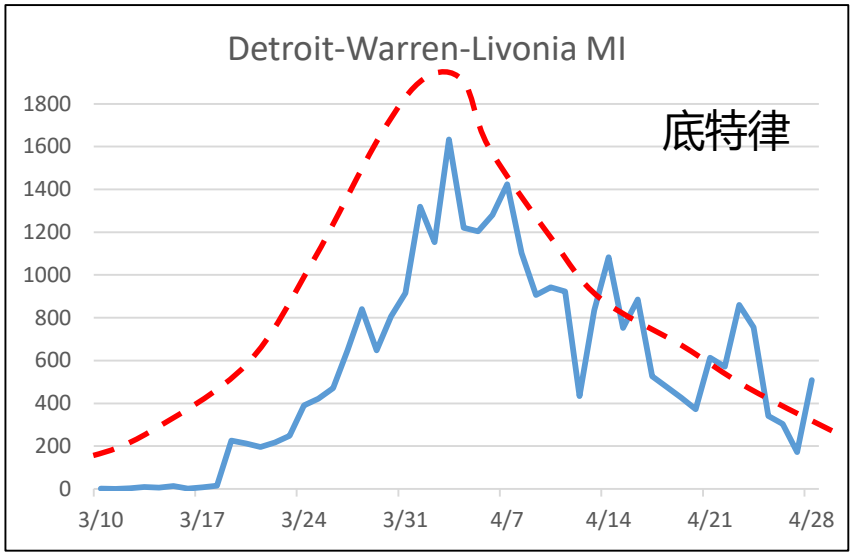
3.1 疫情病例数据: 疫情数据时间变化（美国主要州）

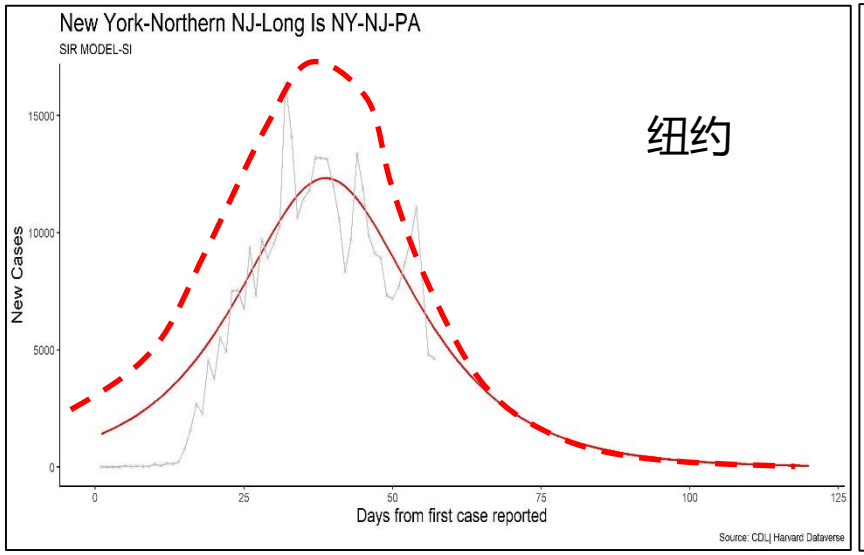


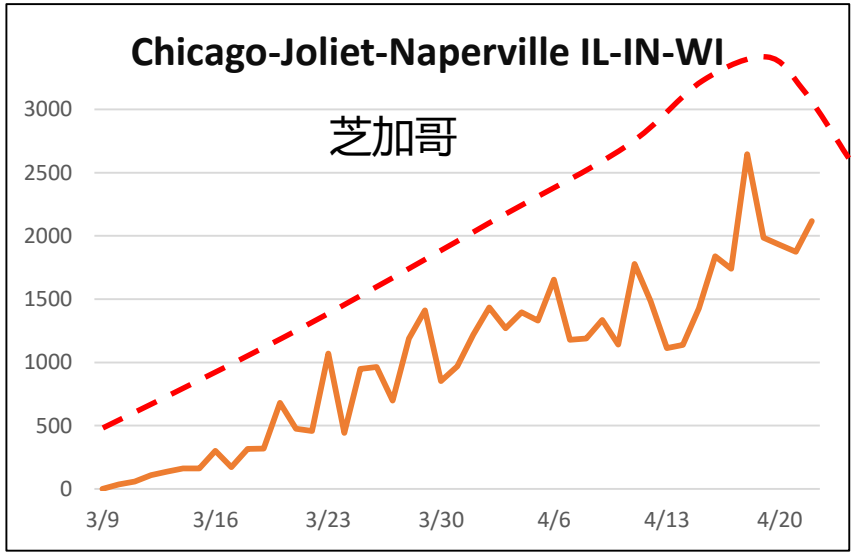


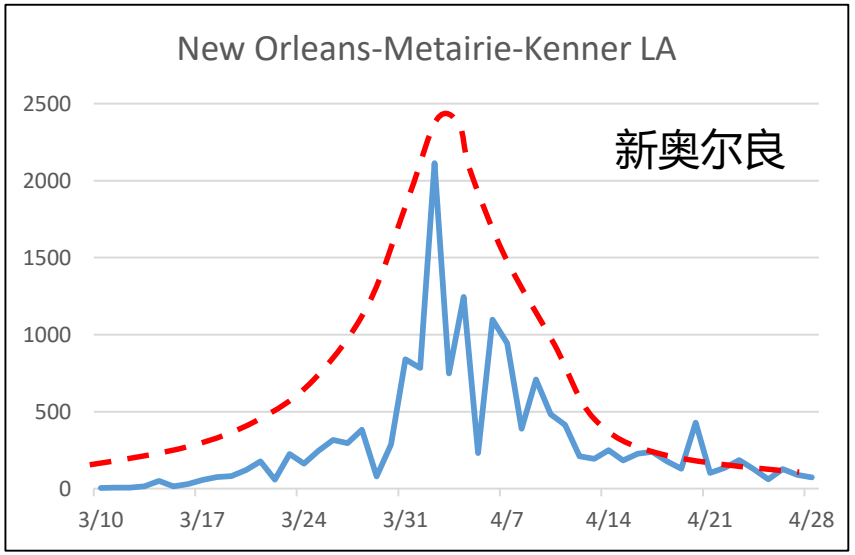


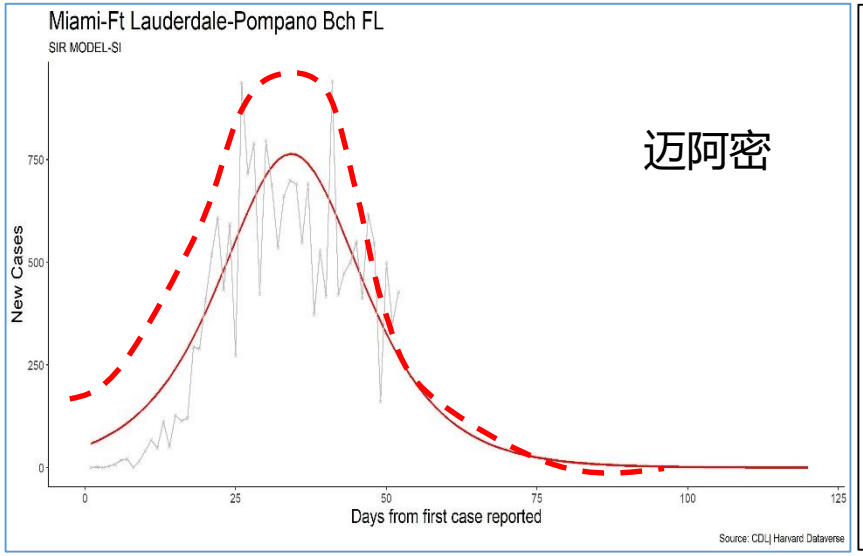
3.1 疫情病例数据: 疫情数据时间变化（美国大都市区）

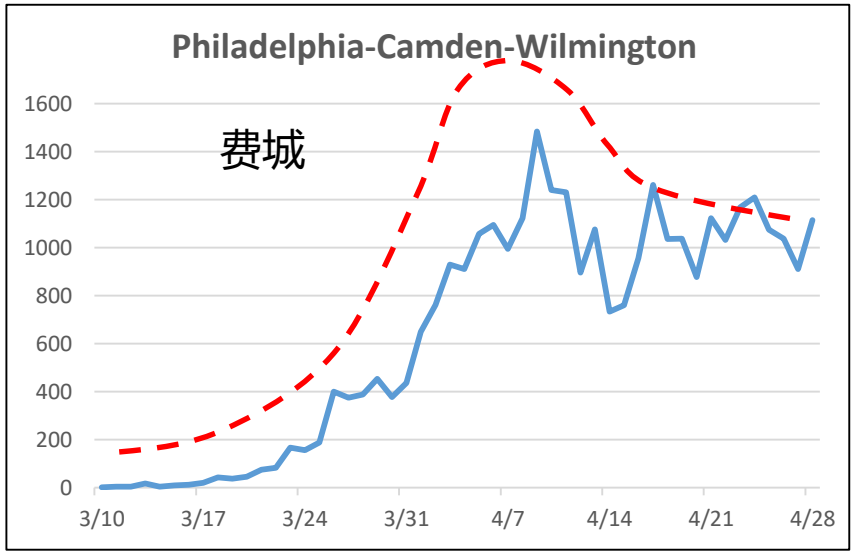






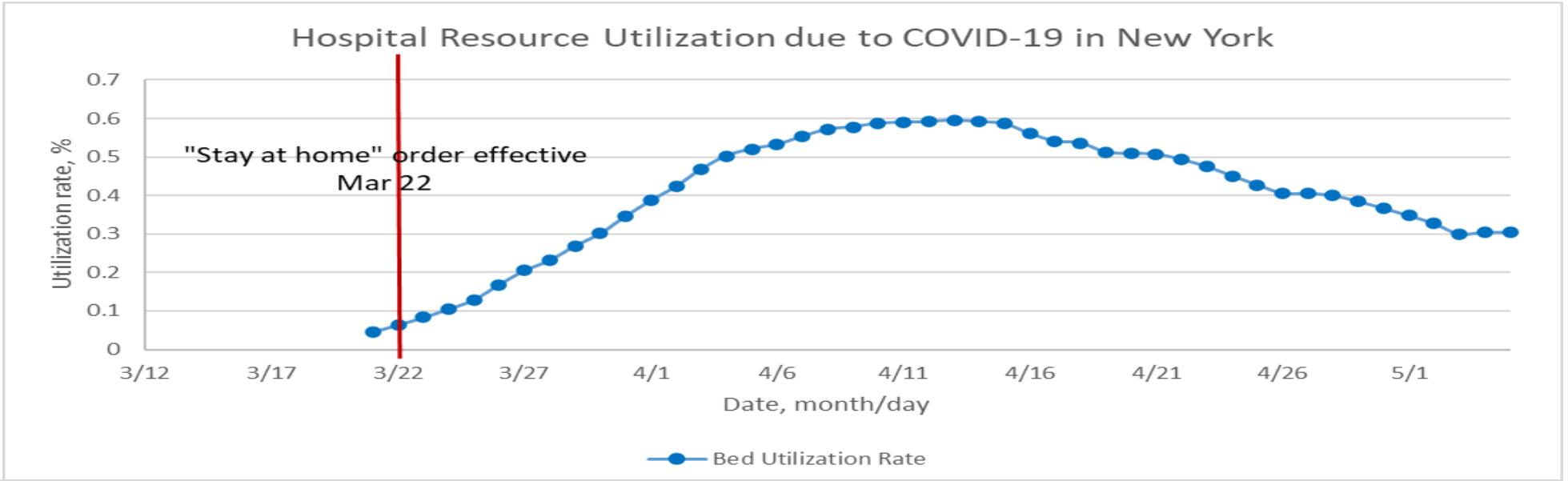


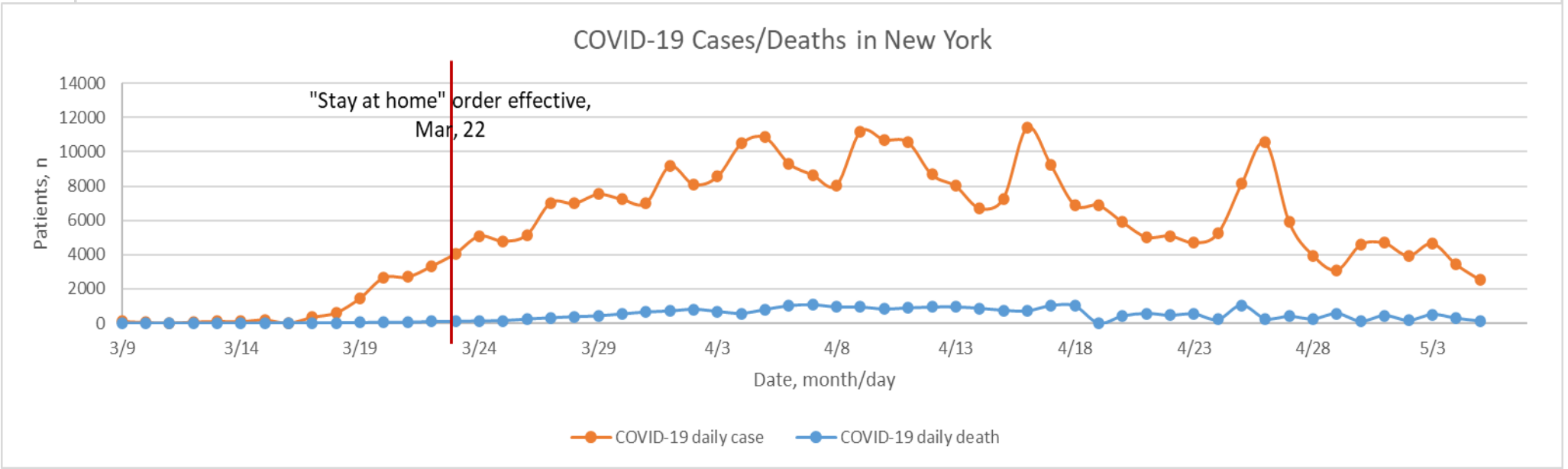




3.1 疫情病例数据: 纽约州疫情数据分析

数据来源：Total beds/ICU available for COVID-19,一亩三分地，https://coronavirus.1point3acres.com/zh/test

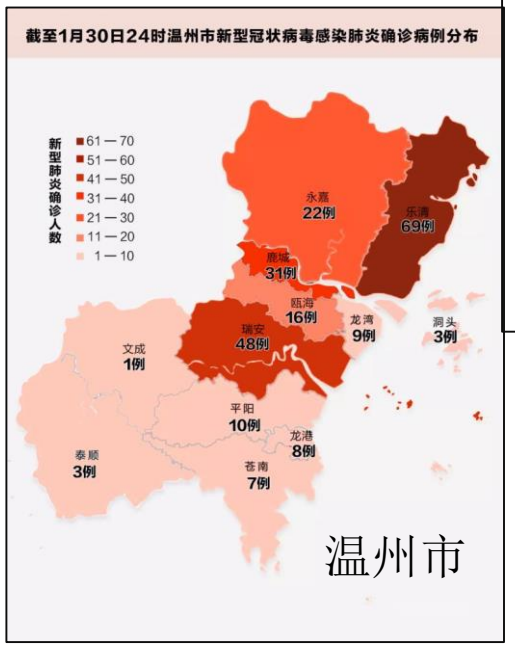


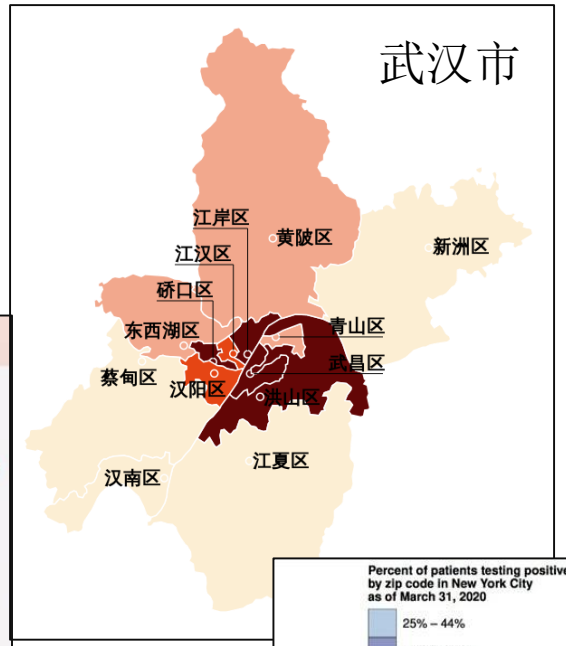


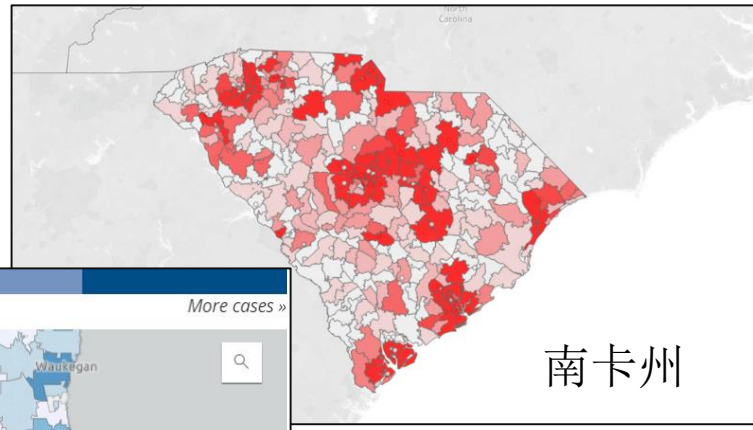
3.1 疫情病例数据: 中美疫情数据空间变化（地区）

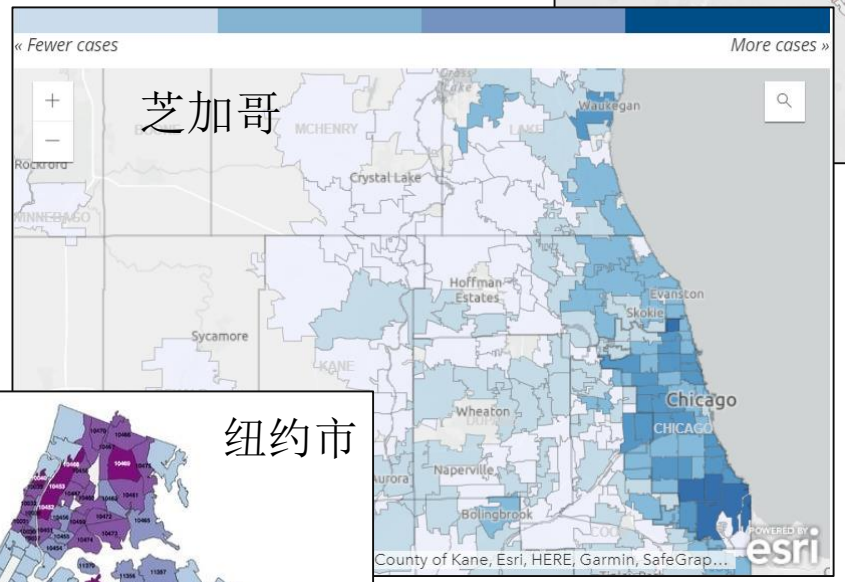
疫情空间扩散主要特征：

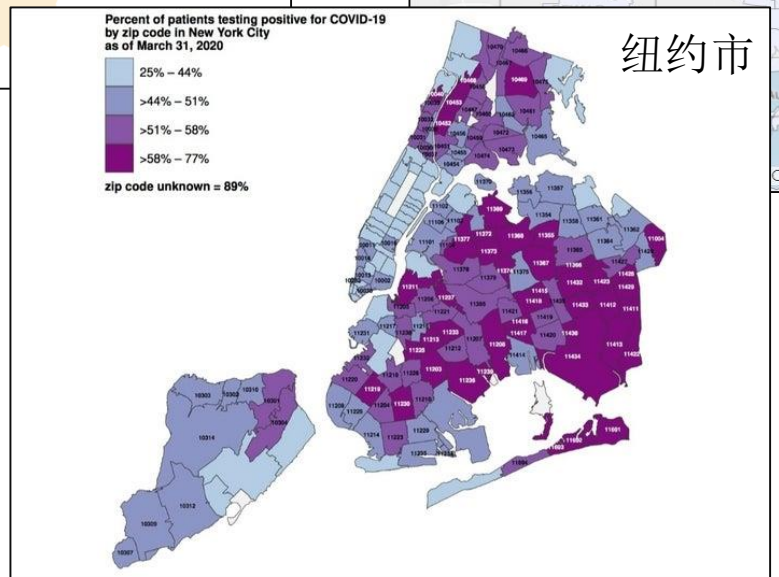
- 空间集聚
- 空间关联





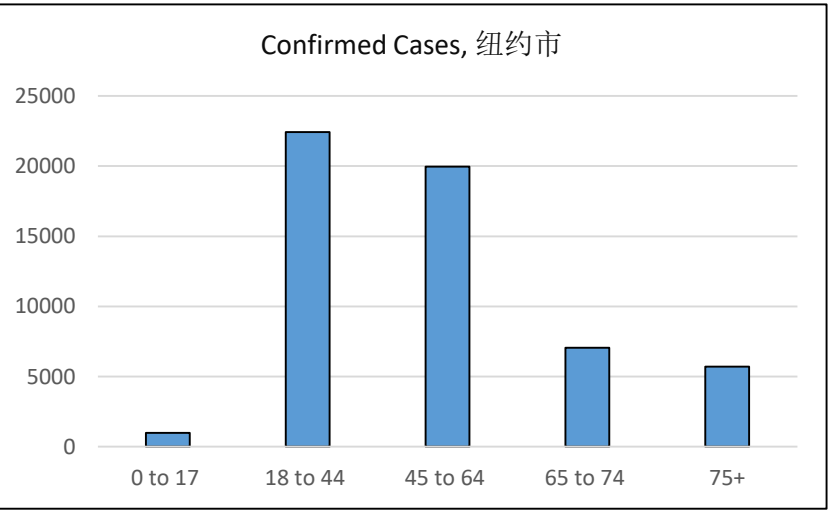


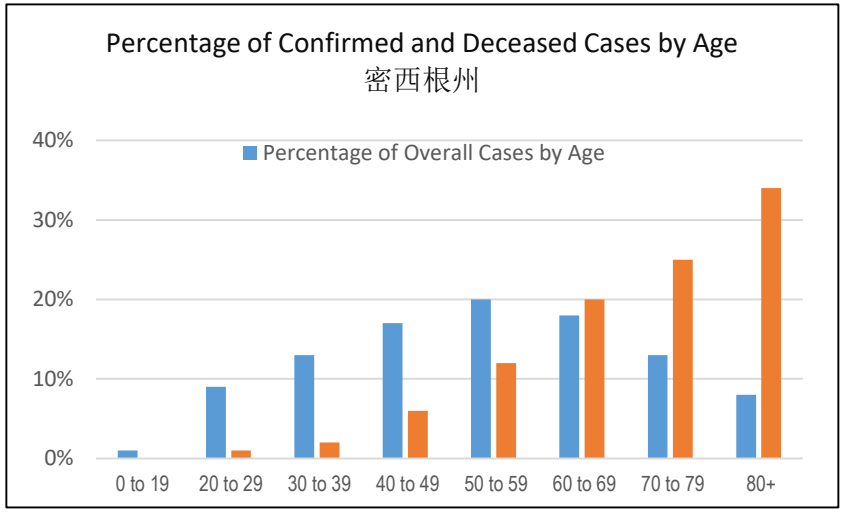


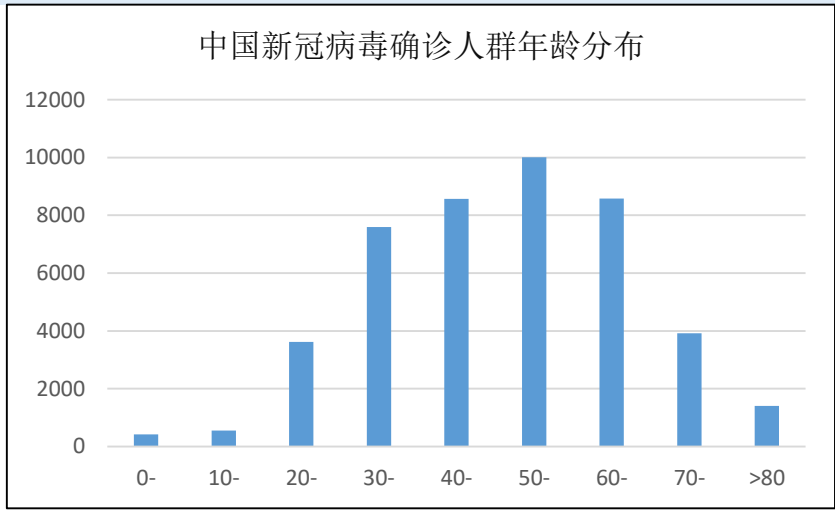


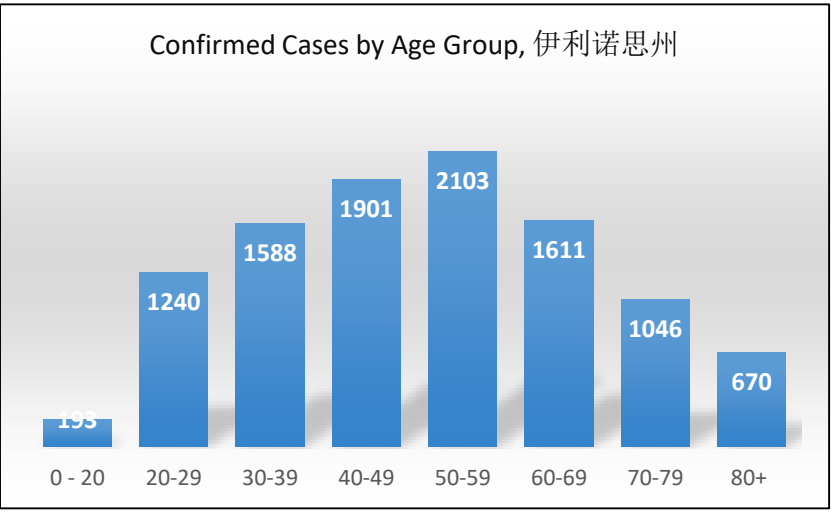
3.1 疫情病例数据：中美年龄影响

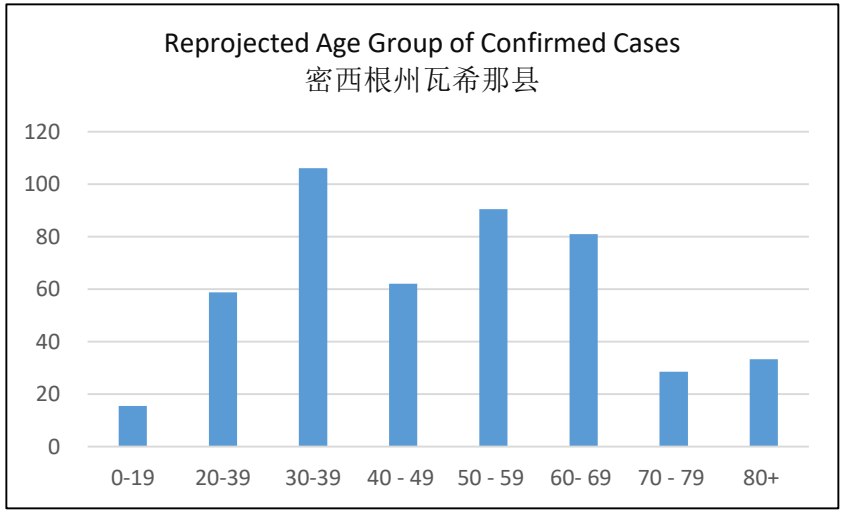
- 年青人是传播主体（20-40）
- 中老年人是受感染主体（40-60）
- 老年人是重创主体（60以上）

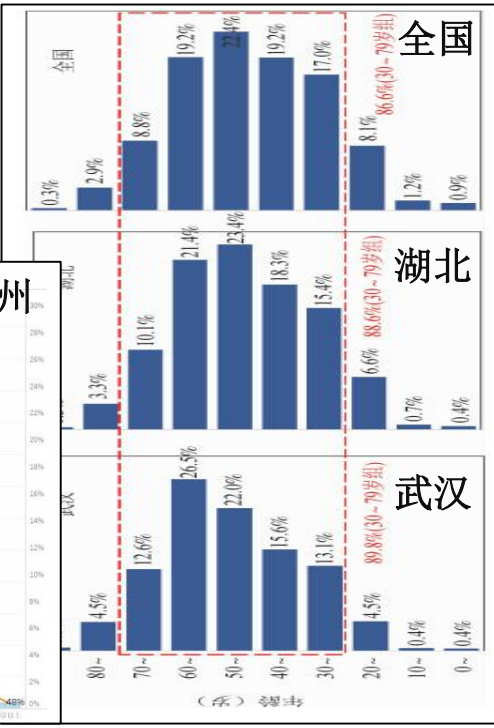


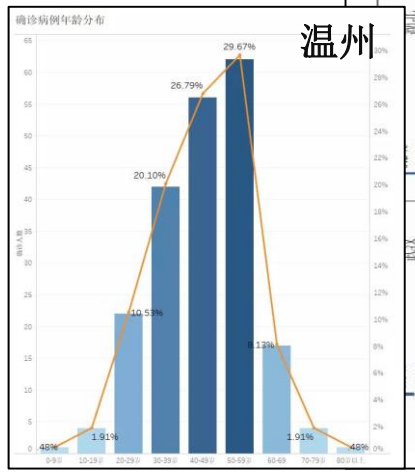






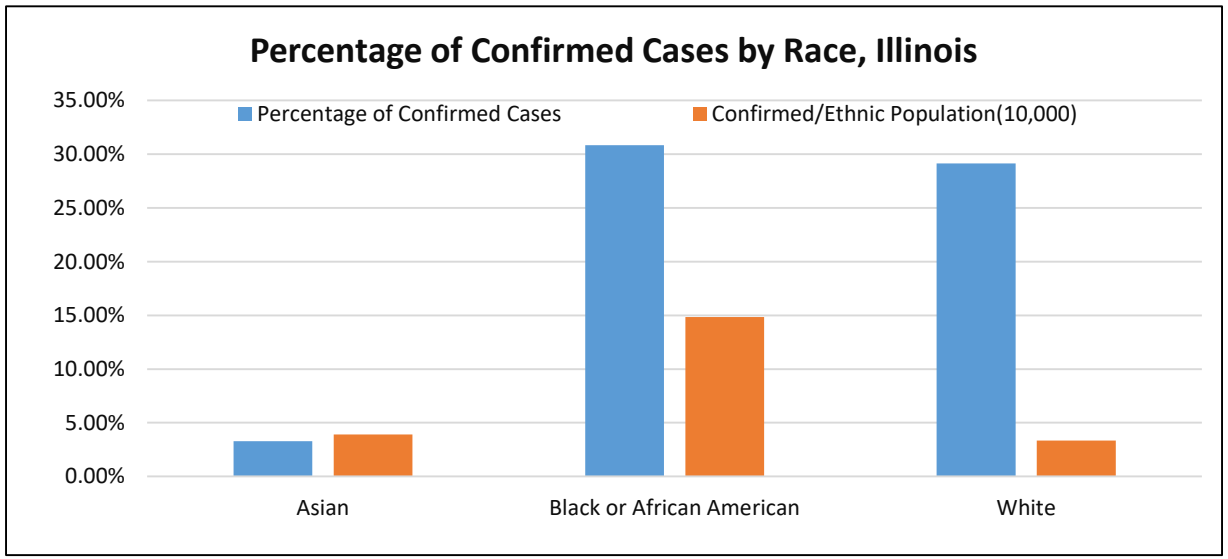


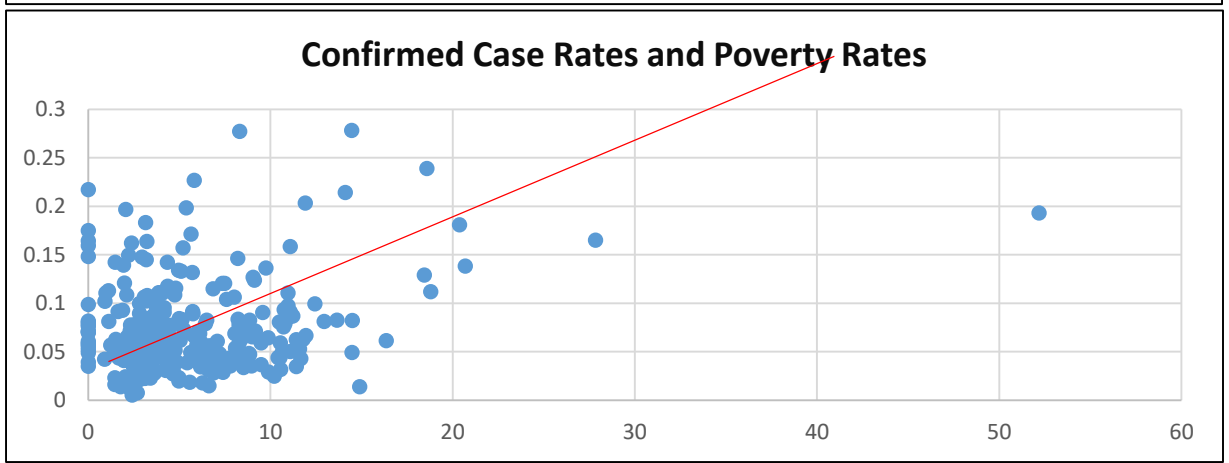


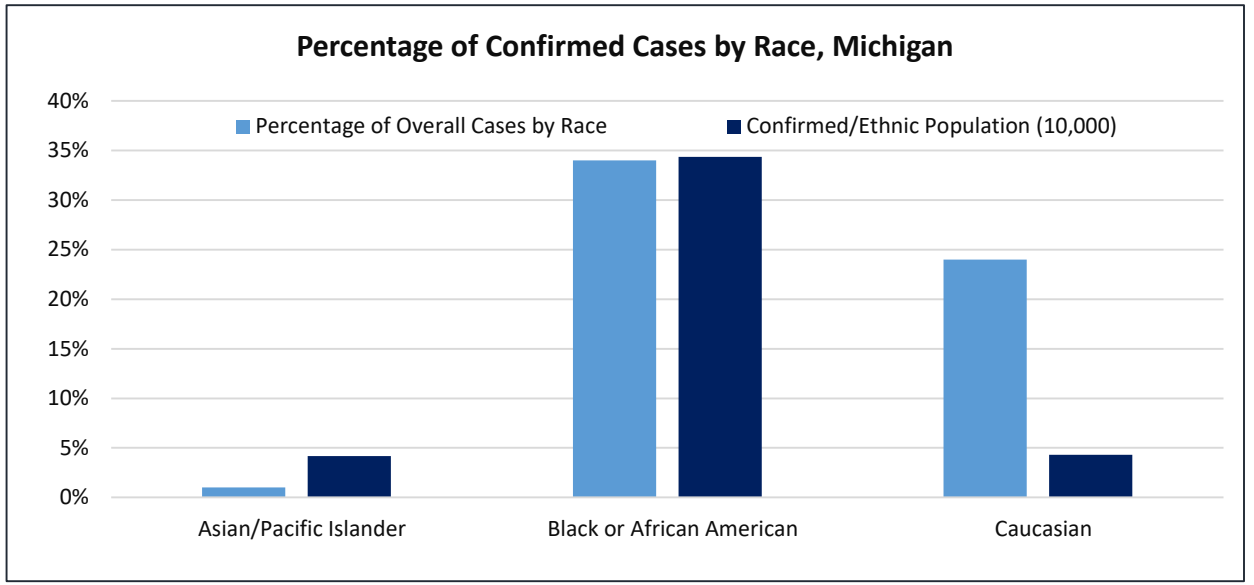


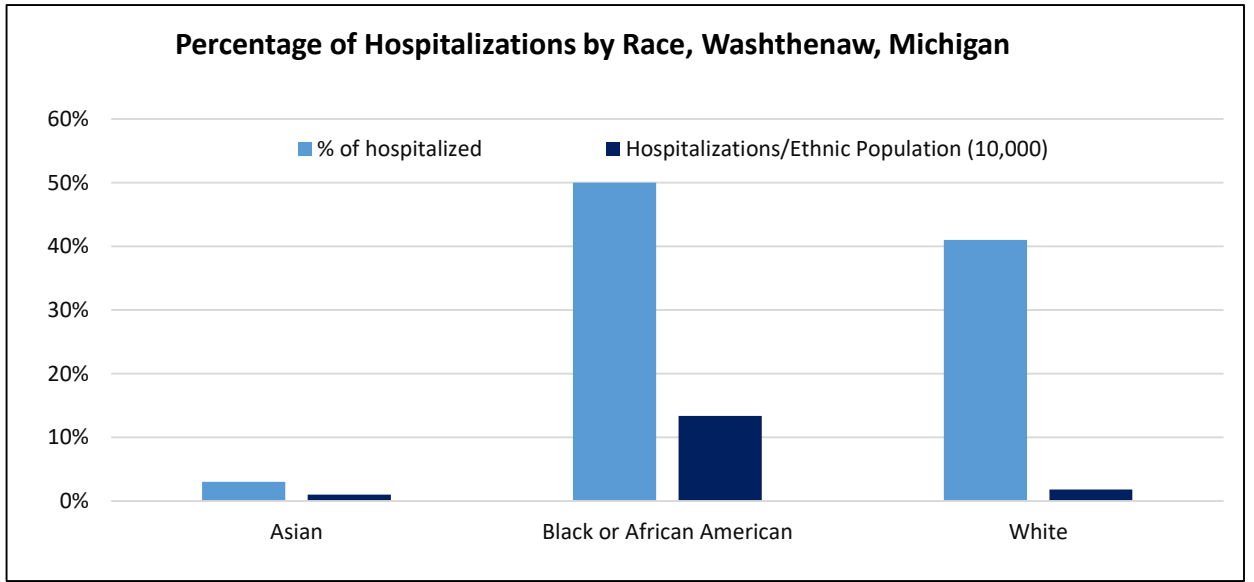
3.1 疫情病例数据：种族和收入影响

- 黑人和白人是感染主体
- 黑人相对感染率最高
- 白人和亚裔人口相对感染率最低
- 感染人口比例与贫困线以下人口比例成正比



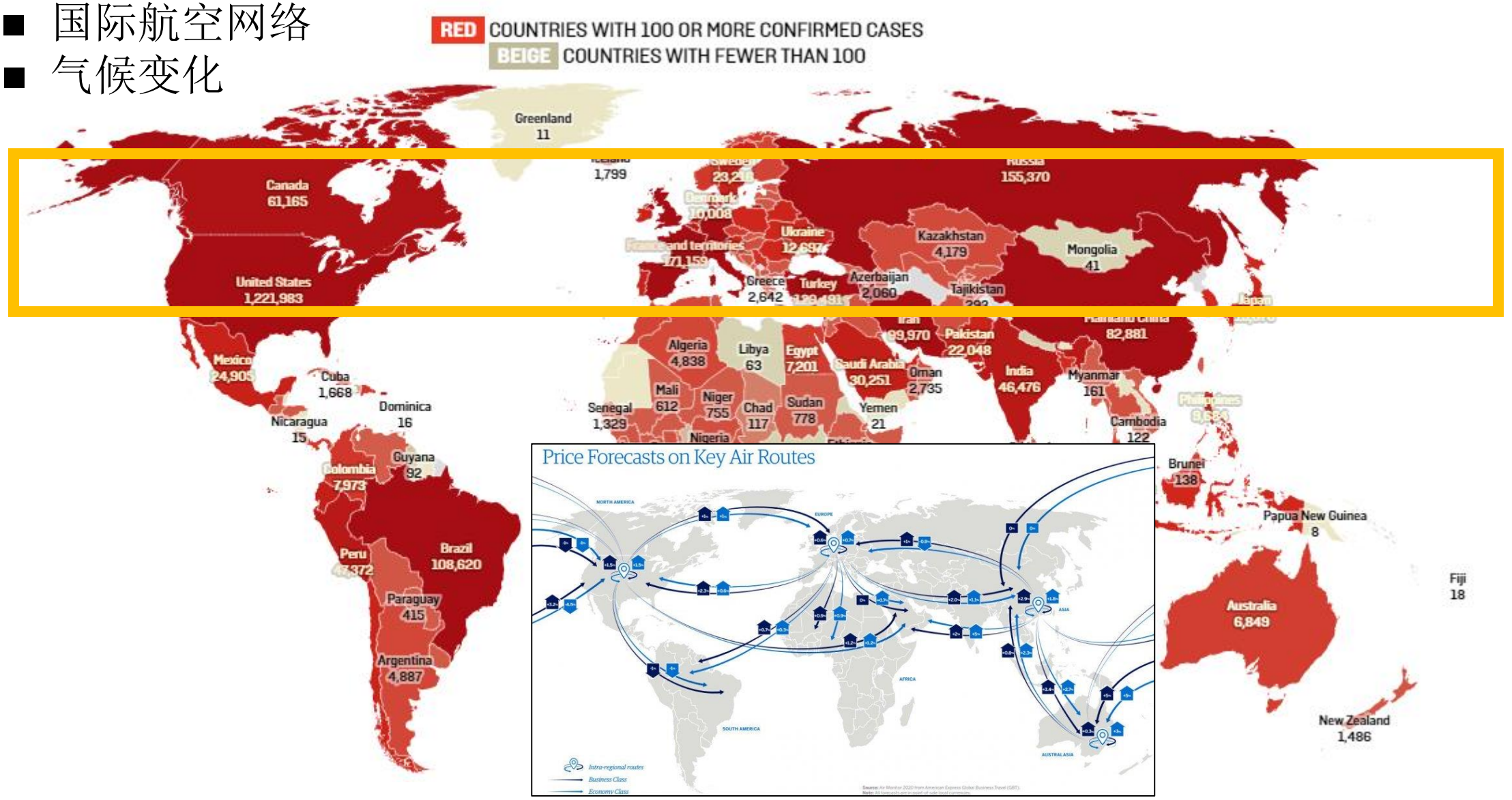






3.1 疫情病例数据：解读全球扩散影响

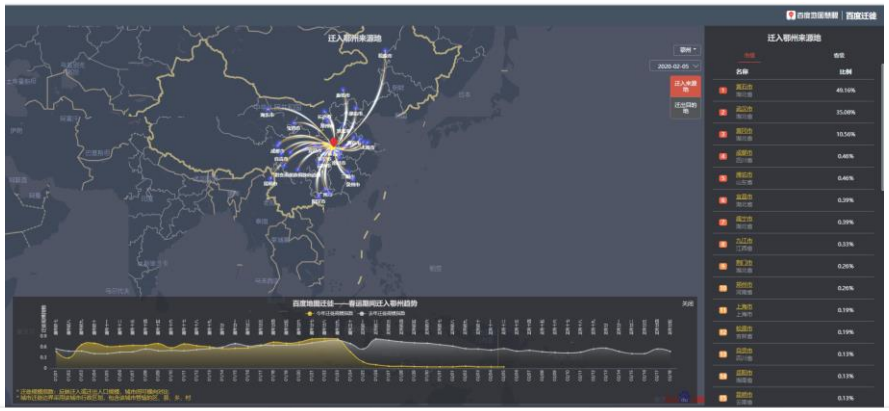
- 国际航空网络
- 气候变化

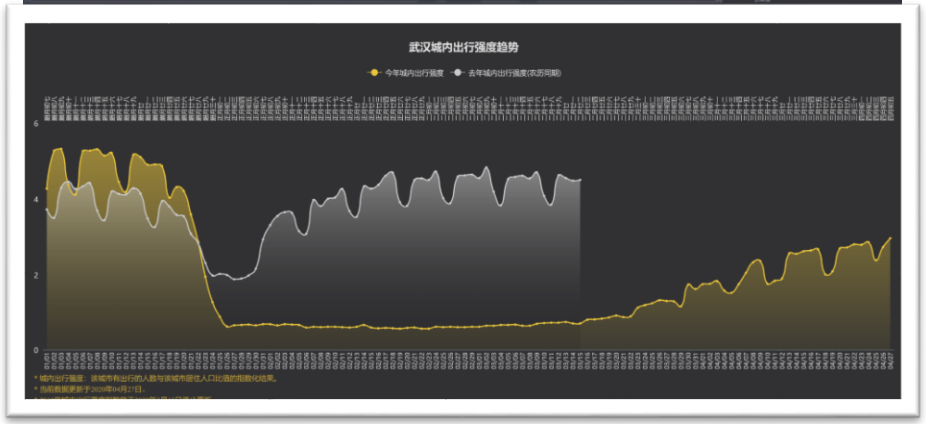


来源: Foreign Policies, https://foreignpolicy.com/2020/03/04/mapping-coronavirus-outbreak-infographic/

3.2 人口迁徙数据/Population mobility data

数据来源：百度人口迁徙数据, https://qianxi.baidu.com/
基础区划：China Data Institute, 2000-2010地市可比行政区划列表, https://chinadatacenter.net/Data/ServiceContent.aspx?id=1623
数据下载：哈佛 DataVerse 每日人口迁徙数据表 "Baidu Mobility Data", https://doi.org/10.7910/DVN/FAEZIO
数据格式：表格文件（tab）
数据指标：（1）迁入来源地比例：其他城市（省）（top100）一天迁徙到某城市（省）的人口比例
（2）迁出目的地比例：某个城市（省）一天迁徙到其他城市（省）（top100）的人口比例
（3）迁入迁出强度





mobility.xlsx - Excel

地区	北京市	天津市	河北	山西	内蒙古	辽宁	吉林	黑龙江	上海市	江苏	浙江	安徽	福建	江西	山东	河南
北京市		17.46	41.12	11.05	9.59	9.71	6.82	9.61	2.2	1.91	1.82	2.28	2.93	1.48	8.49	6.82
天津市	7.79		17.51	3.35	3.01	3.87	3.1	5.27	0.55	0.71	0.54	0.8	1.08	0.38	5.27	2.05
河北																
山西																
内蒙古																
辽宁																
吉林																
黑龙江																
上海市	1.62	0.83	0.58	1.41	0.75	2.3	1.79	2.18		28.36	22.4	9.73	5.41	4.22	3.08	3.78
江苏																
浙江																
安徽																
福建																
江西																
山东																
河南																
湖北																
湖南																
广东																
广西																

3.3 医疗机构数据/Health facilities data

数据来源：高德地图开放平台 API（https://lbs.amap.com）
基础地图：China Data Institute, 2000-2010省市县可比行政地图, https://chinadatacenter.net/Data/ServiceContent.aspx?id=1623
数据下载：省级医疗机构汇总， China Data Lab, 2020, "Health Facilities", https://doi.org/10.7910/DVN/KRSGT3
个体机构：全球疫情数据研究平台 https://chinadatalab.org
数据格式：表格文件（Excel）
机构分类：8个中类、25个小类（国家标准是20小类）。

省市汇总数据指标：

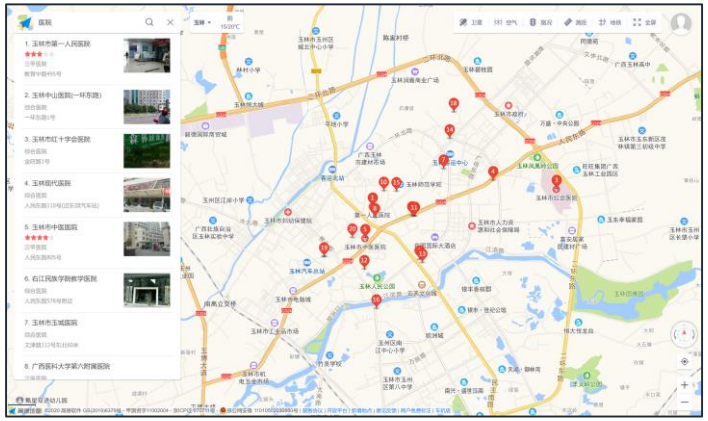
- 机构数量
- 各省机构比例
- 人均机构数量
- 基础人口指标

存在问题：

- 标准不一
- 机构重复
- 年代不详

中类	小类
医疗保健服务场所	医疗保健服务场所
综合医院	综合医院
综合医院	三级甲等医院
综合医院	卫生院
专科医院	专科医院
专科医院	整形美容
专科医院	口腔医院
专科医院	眼科医院
专科医院	耳鼻喉医院
专科医院	胸科医院
专科医院	骨科医院
专科医院	肿瘤医院
专科医院	脑科医院
专科医院	妇科医院
专科医院	精神病医院
专科医院	传染病医院
诊所	诊所
急救中心	急救中心
疾病预防机构	疾病预防
医药保健销售店	医药保健相关
医药保健销售店	药房
医药保健销售店	医疗保健用品
动物医疗场所	动物医疗场所
动物医疗场所	宠物诊所
动物医疗场所	兽医站

	0
0	华中科技大学同济医学院附属同济医院
1	华中科技大学同济医学院附属协和医院
2	武汉市儿童医院
3	武汉大学中南医院
4	武汉市第一医院
5	湖北省人民医院
6	武汉市中医医院
7	湖北省中医院
8	武汉市第三医院(光谷关山院区)
9	武汉市中心医院南京路院区
10	湖北省中西医结合医院
11	湖北省中医院(建设中)
12	湖北省第三人民医院
13	湖北省中医院凤凰门诊部
14	武汉市中西医结合医院(利济北路)
15	马应龙肛肠医院
16	武钢总医院(冶金大道)
17	华中科技大学同济医学院附属梨园医院
18	武汉市第五医院(显正街)
19	湖北省中医院省直分院
20	武汉市中心医院
21	天美乳腺医院爱乳俱乐部
22	湖北中医院附属医院制剂楼
23	武汉普爱医院东院区
24	武汉大学人民医院(汉南医院)
28	武汉大学人民医院东院
30	第三医院首义院区急诊中心
31	武汉普爱医院骨外科综合大楼
32	武汉市中医医院(二桥分院)
33	武汉市第三医院(首义院区)
34	湖北省中山医院-发热门诊
37	武警湖北省总队医院
39	武汉大学同仁医院
42	湖北省总队医院
44	武汉市第三医院
45	武汉市第六医院
47	武汉协和医院口腔科
49	长江航运总医院
52	武汉天美乳腺病专科医院
54	宏昇中医院(暂停营业)
60	湖北省新华医院中医部
62	湖北省中西医结合医院江北分部
63	武汉大学人民医院



3.4 确诊病患轨迹数据/Trace data

数据来源：腾讯活动轨迹数据（https://xw.qq.com/act/fytrace）
个体机构：全球疫情数据研究平台 https://chinadatalab.org
数据格式：表格文件（Excel）
数据指标：

省（自治区、直辖市）
市（自治州、盟）
县（区、市）
位置
发布时间
病患信息
其他信息
行为轨迹
信息来源
信息来源URL

数据价值：

- 年龄
- 性别
- 位置
- 轨迹

存在问题：

- 数据不完整
- 轨迹不精确

省（自治区、直辖市）	市	县（区、市）	位置	发布时间	病患信息	其它信息	行为轨迹	信息来源	信息来源url
四川省	达州市	开江县	淙城街道	2020/2/17	谷某，女	系确诊病例	1月24日,上	达州发布	https://mp.we
四川省	达州市	开江县	信合春天	2020/2/17	李某，女	系确诊病例	1月30日,中	达州发布	https://mp.we
四川省	南充市	阆中市	老观镇万	2020/2/18	杨某某，男，46岁		12月9日,夫	健康阆中	https://mp.we
四川省	南充市	阆中市	祁门县河	2020/2/18	冯某，男，37岁		1月20日,1	健康阆中	https://mp.we
四川省	南充市	顺庆区	嘉陵区嘉	2020/2/18	杨某，56岁	现住顺庆区	1月22日,患	南充播报	https://mp.we
四川省	宜宾市	叙州区	南岸街道	2020/2/18	翁某，男，	系2月11日	1月23日,凌	宜宾发布	https://mp.we

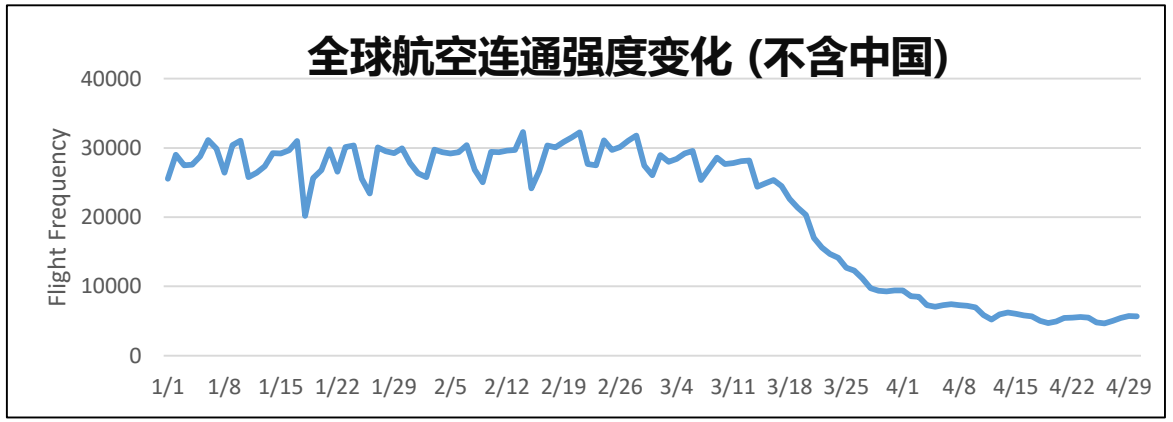
3.5 航空数据/Flight data

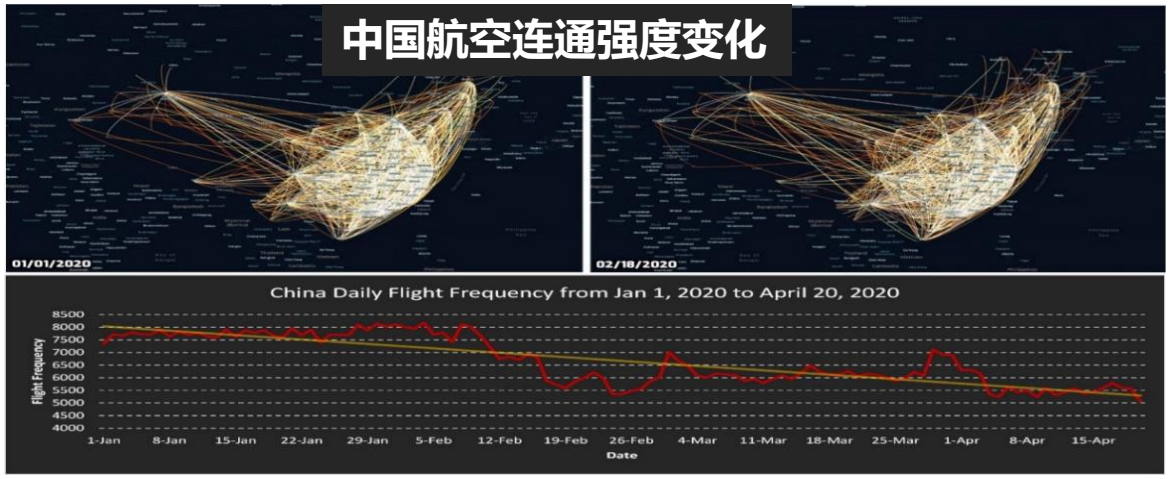
数据来源:

- 中国：VariFlight ADS-B (Automatic Dependent Surveillance - Broadcast) (https://flightadsb.variflight.com/)
- 国际：OpenFlights (https://openflights.org/data.html)

数据分享：全球疫情数据研究平台 https://chinadatalab.org

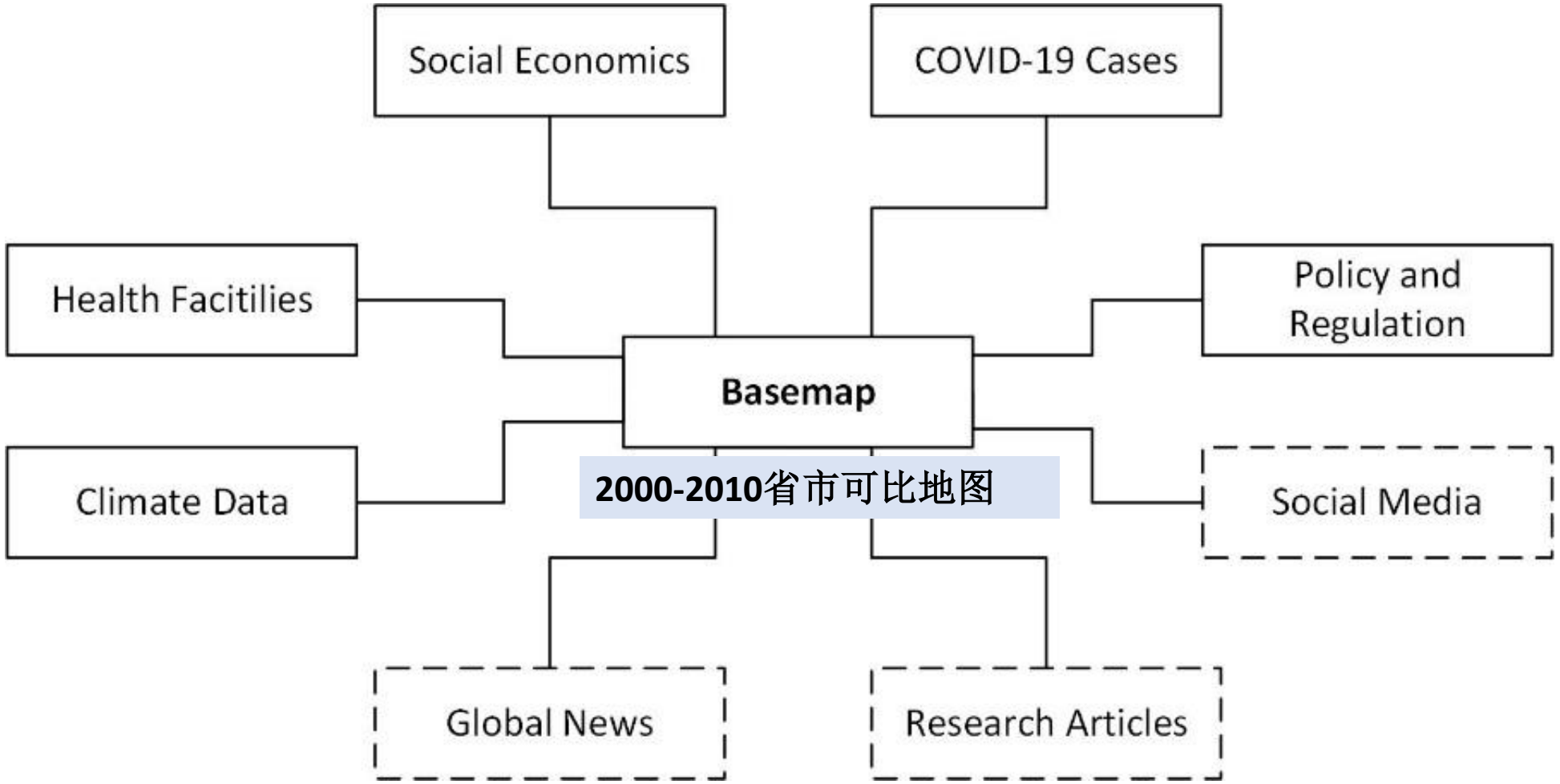
数据格式：表格文件（xlsx， csv）





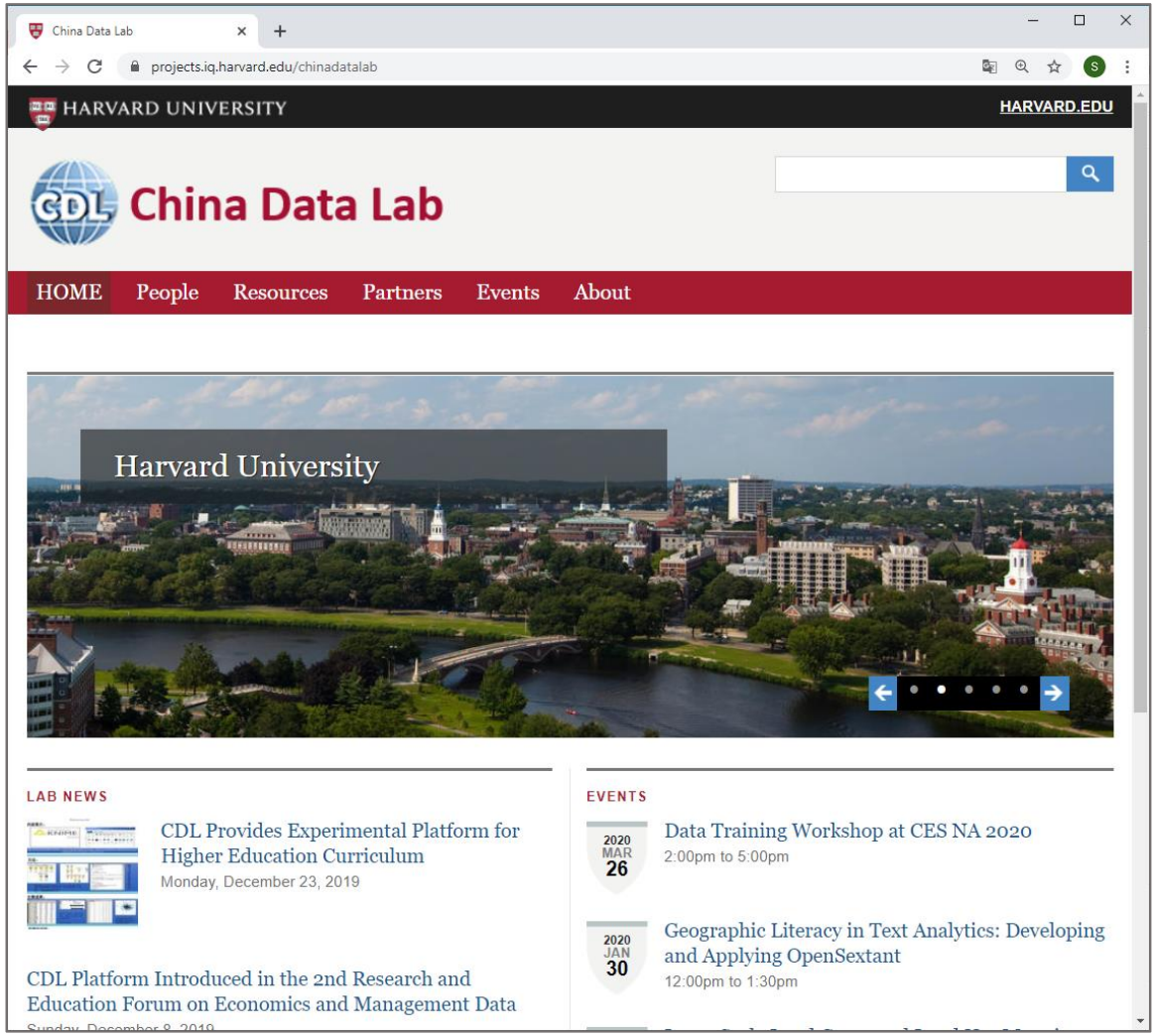
指标	指标
实际到达时间	计划起飞时间
实际起飞时间	出发机场名字
飞机代码	出发机场纬度(度)
到达地时区	出发机场经度(度)
预计到达时间	出发机场高程(英尺)
预计起飞时间	出发机场所在洲
到达机场三字代码	出发机场所属国
到达机场中文名	出发机场所属地区
到达机场英文名	出发机场所属市级行政区划
航班状态码	到达机场名字
航班号两字代码	到达机场纬度(度)
航班号三字代码	到达机场经度(度)
出发机场三字代码	到达机场高程(英尺)
出发机场中文名	到达机场所在洲
出发机场英文名	到达机场所属国
航班代码	到达机场所属地区
出发地时区	到达机场所属市级行政区划
计划到达时间	

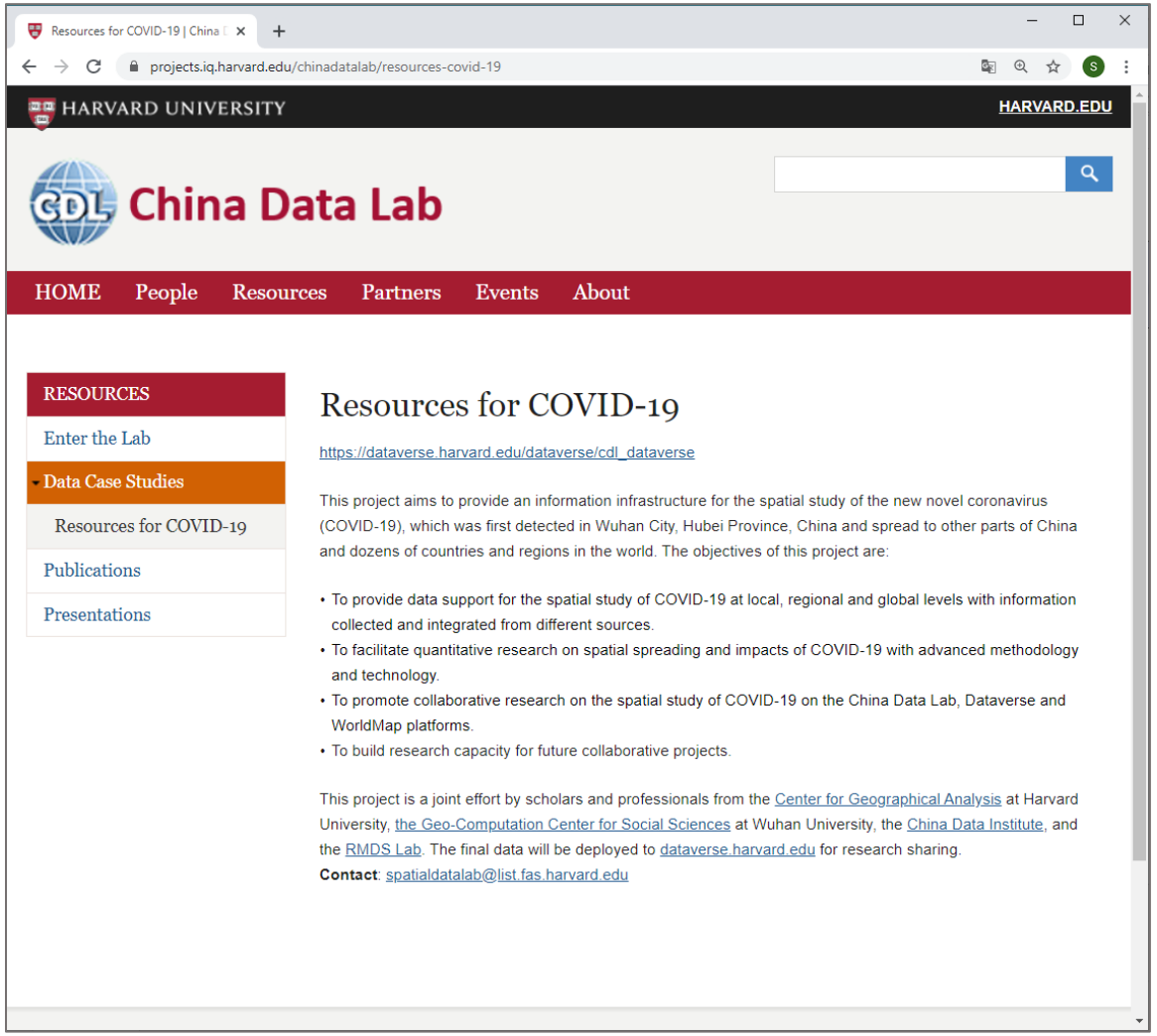
4 数据集成共享/Data Integration



4 资源共享：项目网站

项目网站: http://chinadatalab.net





4 数据下载： DataVerse

下载网址: https://dataverse.harvard.edu/dataverse/2019ncov

The Dataverse® Project

Open source research data repository software

Researchers

Enjoy full control over your data. Receive *web visibility*, *academic credit*, and *increased citation counts*. A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to the research community, and satisfies data management plans. Want to set up your personal dataverse?

Journals

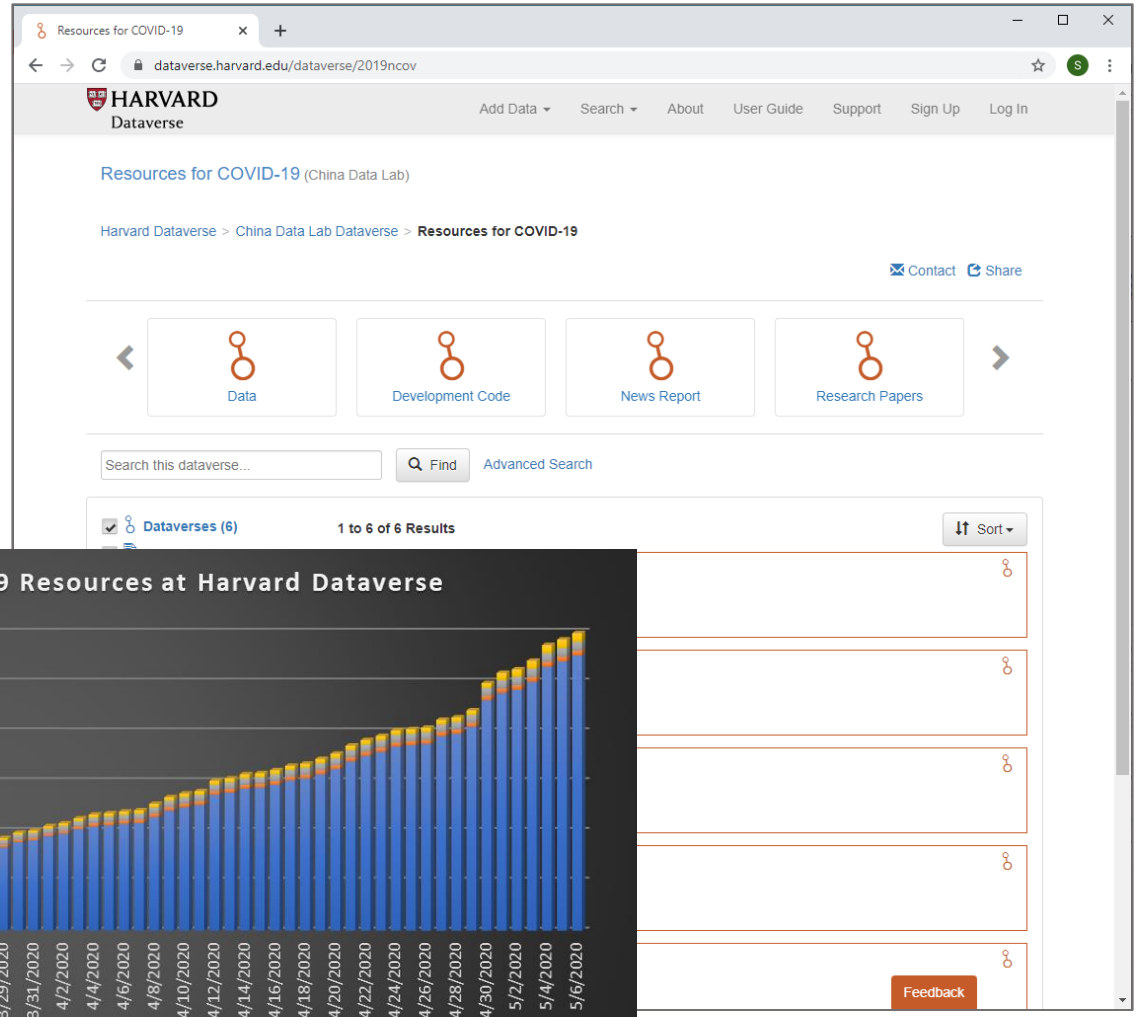
Seamlessly manage the submission, review, and publication of data associated with published articles. Establish an *unbreakable link* between *articles in your journal* and *associated data*. Participate in the open data mo
of repository recommendations

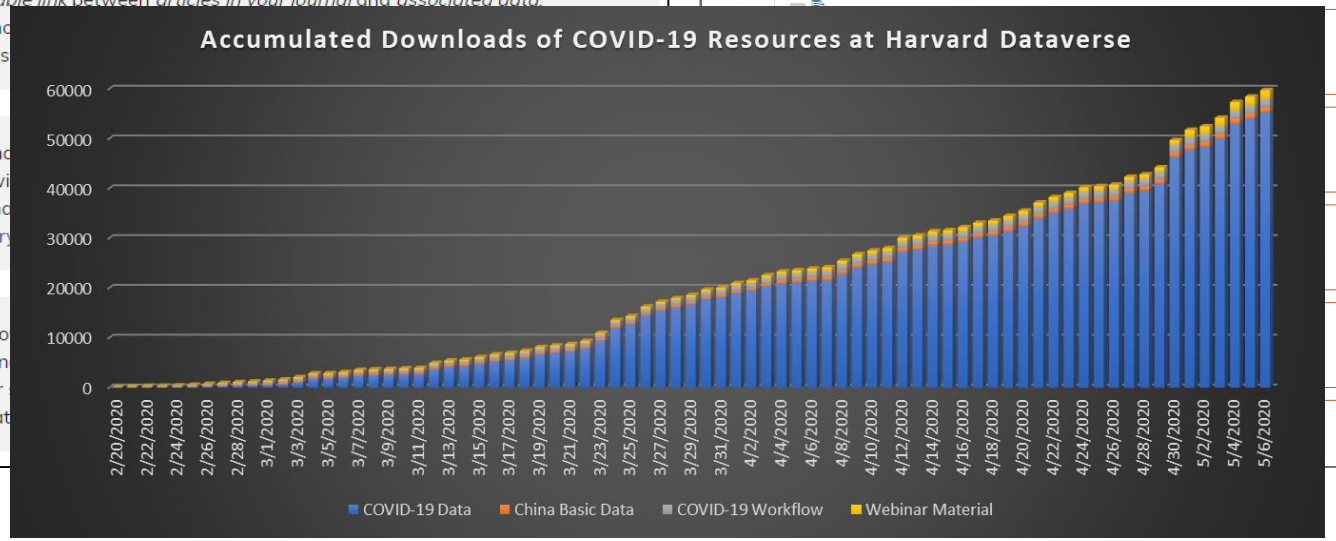
Institutions

Establish a research data manc
Dataverse repositories worldwi
in the drive to set norms for sha
to install a Dataverse repository

Developers

Participate in a vibrant and gro
preserving, citing, exploring, an
documentation, testing, and/or
tools, or other research and dat





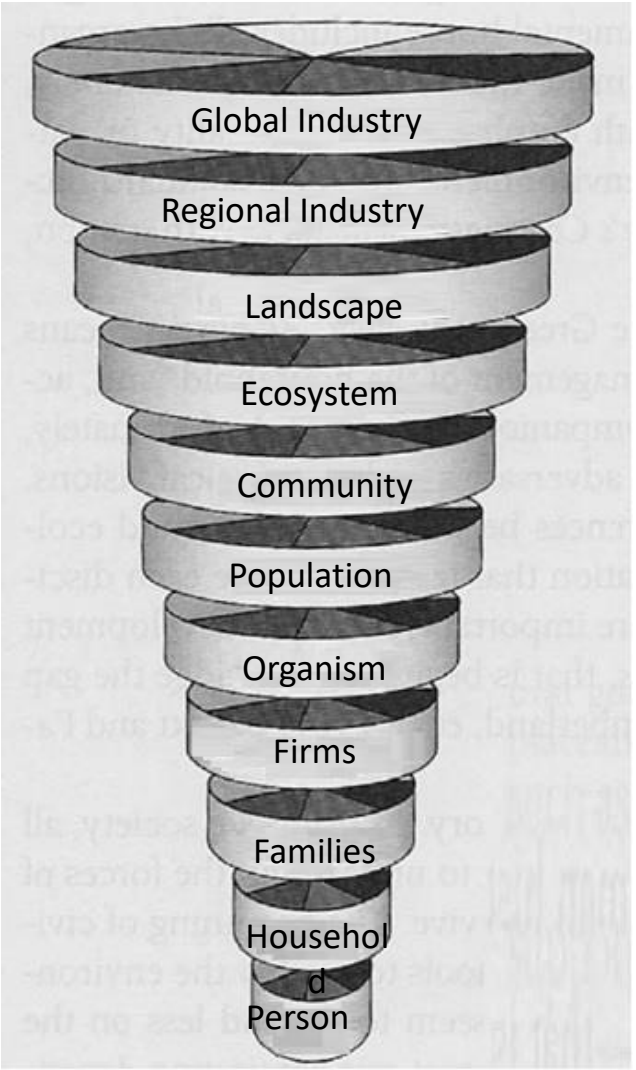
4 在线分析：Data Analysis on Cloud

http://chinadatalab.org

5. 总结与讨论

疫情数据影响因素：

- 人口密度
- 贫困人口
- 年龄分布
- 城乡位置
- 产业结构
- 交通网络
- 人口迁徙
- 检测技术
- 检测标准
- 防控意识
- 政府协调力与执行力
- 气候条件
- 。 。 。 。



5 未来计划

- 数据扩展
- 数据规范
- 数据应用
- 数据培训

参与人员/Participants

Tao Hu[1], Wendy Guan[1], Xinyan Zhu[2,3], Yuanzheng Shao[2,3], Shuming Bao[4], Lingbo Liu[5], Jing Du[2], Hongqiang Liu[6], Huan Zhou[7], Jialei Wang[7], Bing She[8], Luyao Zhang[9], Zhibin Li[1,10], Peixiao Wang[11], Yicheng Tang[12], Ruizhi Hou[9], Yun Li[13], Dexuan Sha[13], Yifan Yang[14], Ben Lewis[1], Devika Kakkar[1]

[1] 哈佛大学 Center for Geographic Analysis, Harvard University, Cambridge, MA 02138, USA;
[2] 武汉大学 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan, China;
[3] 武汉大学 Collaborative Innovation Center of Geospatial Technology, Wuhan University, 129 Luoyu Road, Wuhan, China;
[4] 中国数据研究所 China Data Institute, University of Michigan, Ann Arbor, MI, US;
[5] 武汉大学 School of Urban Design, Wuhan University, No.8 East lake South Road, Wuchang District, Wuhan, Hubei, China;
[6] 山东科技大学 College of Geomatics, Shandong University of Science and Technology, Qingdao, China;
[7] 武汉大学 School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Rd. Wuhan, Hubei, China;
[8] 密西根大学 University of Michigan, Ann Arbor, MI 48106;
[9] 华东师范大学 East China Normal University, 3663 Zhongshanbei Road, Putuo District, Shanghai, China;
[10] 北京大学 School of Government, Peking University, Beijing, China;
[11] 福州大学 The Academy of Digital China, Fuzhou University, Fuzhou, China;
[12] 合肥工业大学 School of Management, Hefei University of Technology, Hefei, China;
[13] 乔治梅森大学 Department of Geography and GeoInformation Science, George Mason University, Fairfax, VA, USA;
[14] 加州大学圣地亚哥分校 University of California, San Diego, 9500 Gilman Dr, La Jolla, CA, USA;

致谢/Acknowledgement

Geocomputation Center for Social Sciences, Wuhan University
武汉大学社会地理计算中心

Center for Geographic Analysis, Harvard University
哈佛大学地理分析中心

RMDS Lab 研究方法与数据科学实验室

China Data Institute 中国数据研究所

Amazon.com, Inc 亚马逊

MicroSoft China 微软中国

Vesystem Inc.和信创天科技有限公司

Knowledge Sharing Inc.百智享科技有限公司

All China Marketing Research Inc. 华通人市场研究有限公司

联系方式/Contact

Resources for Coronavirus Study

http://chinadatalab.net

The Cloud for COVID-19 Study

Http://chinadatalab.org

spatialdatalab@lists.fas.harvard.edu