新型冠状病毒肺炎全球扩散情况及疫情输入风险分析—基于 Python 的可视化实现

学校：广州市海珠区实验小学

撰 写 人： 刘子悠

研究小组： 刘子悠 （组长） 刘春杉 卢映玲

指导老师： 郑贤 梁戈

编写日期： 2020 年 3 月

一、项目来源

2 月 29 号，科学课梁老师为我们介绍了“战疫情，向未来”主题研学周活动的情况，同日班主任林老师下发了《广州市海珠区实验小学‘宅家研学’基于项目学习任务单》，要求以学生为中心，以家庭为单位开展项目学习。以新冠肺炎疫情为主题，提出问题，设定任务内容，通过学习研究，解决实际问题，从而让我们学习整合新旧知识，提高我们解决问题的能力。

3 月伊始，我们经家庭讨论，成立了“星辰研学小组”，由我担任研学项目组长，我的爸爸、妈妈为研学项目组成员。考虑到爸爸的特长是地理和地图，而我对计算机操作比较熟练，也学过一点编程，我们决定研学任务是把疫情和地图结合起来，通过地图软件绘制不同时间疫情地图，展示疫情扩散的过程，并建立某种模型，分析疫情的风险。定下研学目标后，我们讨论了每个人的分工：爸爸负责编程和建模，我参与部分代码编写；我负责各国疫情数据录入，爸爸妈妈审核；请郑贤老师和梁戈老师担任研学项目组的指导老师。

二、研学过程

按照最初设想，我们的研学周期为 29 天，分为三个阶段：

表 1 研学计划

阶段	时间	任务内容	阶段成果
第一阶段	3.1-3.7	确定研究重点，筛选地图平台。筛选数据来源	提交研学任务单。
第二阶段	3.8-3.22	收集整理各国疫情数据，绘制地图、表格、专题统计图等图表。	疫情数据、专题地图和统计表。
第三阶段	3.23-3.29	风险模型建立、页面布局和配色设计、成果整理。	风险模型、疫情数据库、基于专题地图和统计表成果页面、程序源代码、研学报告。

第一阶段：

1. 确定研究重点

从过年前开始，通过电视和网络，我们对疫情有了比较充分的了解。根据观察，进入二月中下旬以后，随着全国各地疫情防控措施加强，全国支援湖北，武汉方舱医院投入使用，中国的疫情较快地稳定了下来，而国外的疫情却逐渐失控，呈爆发的趋势。2 月 15 日，中国的确诊人数为 68431 例，国外仅 599 例，而 20 天以后的 3 月 6 日，全球病例首次突破 10 万例，其中中国 80735 例，增加 18%，国外 21049 例，增加 35 倍（如图 1 所示）。

我们判断，随着国内、国外疫情发展趋势的反转分化，疫情的关注焦点将由国内转移到国外，而国内面临国外疫情回流输入的风险也将越来越高，因此我们决定把研究重点放在国外，通过整合数据来反映全球疫情的扩散趋势，并借此分析全球疫情对国内的影响。

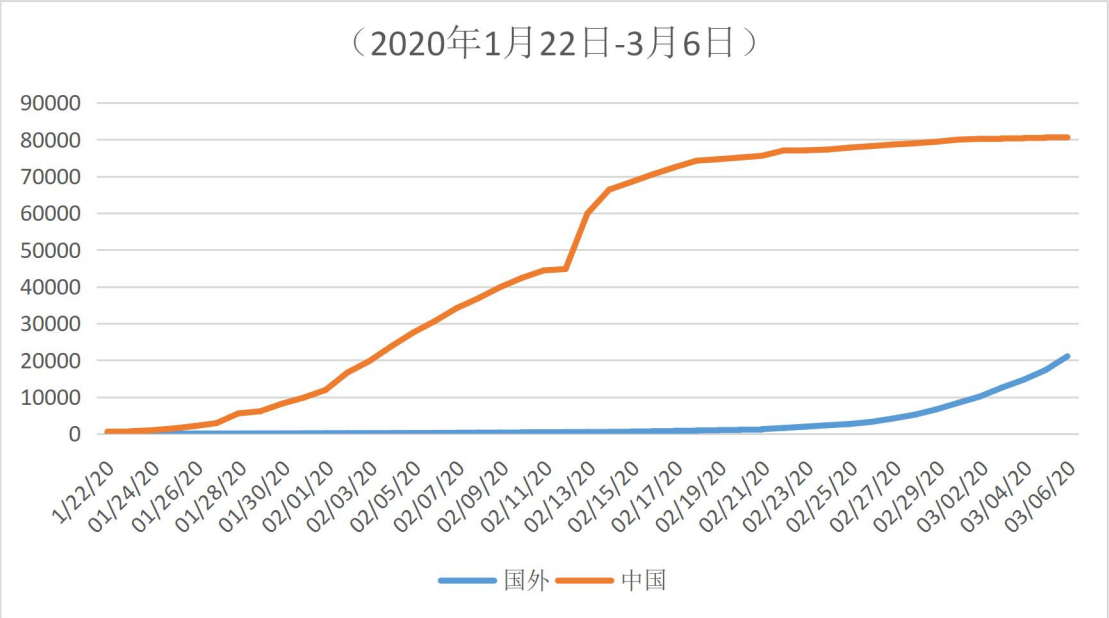

图 1　国内外新冠疫情增长曲线对比

（数据整理自 Johns Hopkins Coronavirus Resource Center）

2. 筛选地图平台

我们对比了多门编程语言和制图软件（MapInfo/Flash/Python），MapInfo 是主流地图软件，但缺少可动态展示能力；Flash 曾是知名网络绘图软件，但由于需要安装插件，目前支持平台越来越少；Python 是一个扩展性很强的编程语言，可以利用现有的地图库快速制作网络动态地图，但我们刚刚接触，了解不多。经过讨论，我们认为使用 MapInfo 或者 Flash 虽然可以很快把专题地图绘制出来，但前者表现力会差点，后者分享难度更大。虽然我们对 Python 了解不多，但从接触的范例来看，Python 的表现形式和扩展性会更好一些，而且使用 Python 可以跟学校信息技术课相结合，因此，我们决定采用 Python 来实现全球的疫情扩散的趋势分析以及各个国家对中国的疫情输入性风险评估。

确定了软件平台后，我们按照梁戈老师授课和课后指导，先后完成了 Python 的安装和 pyecharts 的导入。

3. 筛选数据来源

在多个国内网站比较后，我们认为网易的国内疫情统计方面做得很直观，初步选取网易的新冠疫情栏目作为疫情数据来源。

3 月 7 日，我们向学校提交了研学任务单（见附件 1）。

第二阶段：

1. 数据整理录入

我们收集全球各个国家的的名称和人口，利用 EXCEL 表格建立了疫情数据库，并设计了大洲、国家、人口、每日疫情等列，将每日收集的各国疫情数据整理录入到数据库中，包括每天的新增病例、累计确诊病例、累计死亡病例和累计治愈病例。后期发现新增病例与前后天累计确诊病例数据对应不上，新增病例不再录入（见图 2、表 2）。

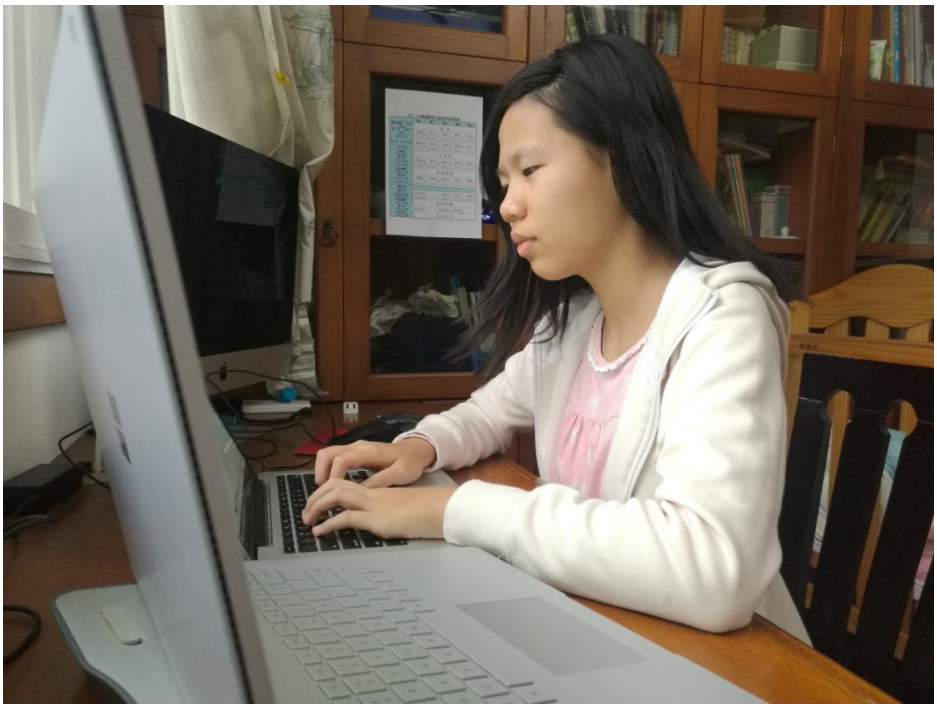

图 2 我在录入数据

表 2 疫情数据库节选（3 月 15 日）

人口单位：千人；疫情单位：人

序号	大洲	COUNTRY	国家	人口	0315 累计确诊	0315 累计治愈	0315 累计死亡
1	亚洲	CHN	中国	1432000	81099	67923	3218
71	欧洲	ITA	意大利	60431	24938	3086	2368
3	亚洲	IRN	伊朗	81800	14991	4996	853
73	欧洲	ESP	西班牙	46724	9191	571	309
38	亚洲	KOR	韩国	51635	8236	1137	77
75	欧洲	DEU	德国	82928	5917	49	13
55	欧洲	FRA	法国	66987	5423	31	127
161	北美洲	USA	美国	327167	3774	56	69
77	欧洲	CHE	瑞士	8517	2220	4	13
57	欧洲	GBR	英国	66489	1543	19	35
60	欧洲	NLD	荷兰	17231	1413	-	24
59	欧洲	NOR	挪威	5314	1256	1	3
61	欧洲	BEL	比利时	11422	1085	1	4

在数据录入过程中，我们发现了百度的全球数据更完整，可分洲统计，因此改用百度疫情专栏作为数据源。由于时差问题，百度每天都在晚上 11 点至凌晨 1 点才更新数据，随着疫情的发展，受感染的国家越来越多，每天都要对应不同国家的排序，整理疫情数据。在录入过程中，我又发现欧美各国数据百度更新仍存在不及时、不同步的问题，新增数据与每日数据经常对不上。为提高数据的准确性，爸爸后来又去世界卫生组织（WHO）去下载每日报告（数据一般会延后一至两日）给我进行数据核对。世卫的每日报告是英文版的，这对我也是个锻炼。

2. 统计地图与表格编制

爸爸利用下班时间开展疫情统计地图和表格的编制。利用 Python IDLE 作为编制平台，引入 pandas 库调用疫情数据库；引入 pyecharts

库中的 MAP 对象绘制统计地图、TABLE 对象绘制表格、TIMELINE 对象实现多期地图动态展示，最后利用 PAGE 对象发布成果页面。经讨论，我们确定除了展示各国累计确诊量以外，还要展示各国的现存确诊量和感染指数（每百万人中感染病毒人数）。其中：

现存确诊量 = 累计确诊量 - 累计治愈病例 - 累计死亡病例

感染指数 = 10000000 * 累计确诊量/人口数

为了让我也能参与编程，爸爸将代码加上注释行，每一行代码都和我详细解释（见图 3），并设计了统计地图的代码模板。

图 3　爸爸教我编程

该代码模板包括三个步骤，第一步是提取所需要的数据（举例：把大象拿过来），第二步是指定地图的模板（举例：把冰箱门打开），第三步是把数据灌入地图（举例：把大象塞进冰箱），通过举例我基本搞清楚了如何编写地图代码，并按照模板，依葫芦画瓢完成了累计

确诊量和感染指数两个地图代码的编写。（见图 4）。在此过程中，爸爸的学习能力、解决问题的能力，以及跟梁戈老师的互动配合，给了我很大的启发和触动。

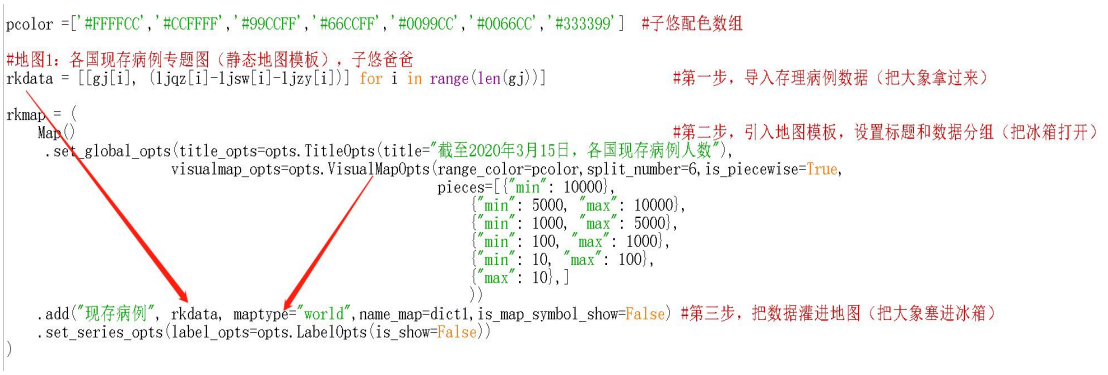

图 4 地图代码模板截图

第三阶段

截至 3 月 15 日，全球各国的累计病例和死亡病例都已经超过了中国，研学任务单中的预测已经变成了现实，国际输入性风险已经成为我们国家疫情防控最大风险源。此时我们已经实现了地图展示功能，我和爸爸认为应该提前结束第二阶段工作，尽快建立模型，通过数据分析判断出哪些国家输入性的更大。因此，从 3 月 16 日开始，我们转入第三阶段工作。

1. 输入性风险分析模型建立

经讨论，我们认为，国外输入性风险的影响因素包括该国确诊人数、传染速度和华侨人数。具体来说，确诊人数越多，传染速度越快的国家传染风险越高，而当地疫情一旦失控，当地华侨就有可能大规模回国，回国华侨人数越多，带病毒回到中国可能性就越大。

我们选取了 3 月 15 日国外病例超 1000 人的国家，通过网络查询了这些国家的华侨人数；通过 3 月 15 日与 3 月 8 日各国确诊的人数

差比，计算传染速度。爸爸根据我们讨论的结果建立了输入性风险评估模型，并通过 EXCEL 公式功能计算风险指数，再通过 pyecharts 库的 BAR 对象绘制了不同国家的输入风险指数条形统计图。

国外输入性风险评估模型：

设：单一国家输入性风险指数：X_i ， 单一国家 3 月 15 日累计确诊人数：a_i
单一国家 3 月 8 日累计确诊人数：b_i ,单一国家华侨人数：h_i
单一国家传染速度为：S_i ,各国 3 月 15 日累计确诊人数数组：An
各国华侨人数数组：Hn ,各国疫情传染速度数组：Sn
传染速度权重，华侨人数权重，累计确诊权重：Cs,Ch,Ca
数组中最大值：max[数组]

$$S_i = (a_i - b_i) \div a_i$$

$$Cs + Ch + Ca = 10$$

$$X_i = Cs \times (S_i \div \max[Sn]) + Ch \times (h_i \div \max[Hn]) + Ca \times (a_i \div \max[An])$$

输入性风险指数理论最大值为 10，理论最小值为 0，值越大代表风险越高，经我和爸爸反复讨论，我们给 Cs、Ch、Ca 分别赋值为 3，2，5。根据我们的分析结论，意大利、西班牙和美国对我国的输入性风险最高，虽然截至 3 月 15 日，美国的确诊病例并不多，但美国的传染速度快，在美华侨多，其输入性风险小不可小觑，在其后的一周时间里，充分证明了这一点。

2. 页面布局与地图配色

我们讨论了地图展示页面布局和地图配色。我们一共设计了三幅地图（其中两幅是我编写的代码）、两个表格和一个条形统计图；最终页面采用了我提出的“三明治”布局，即图、表从上往下相间排布；地图配色也从红色系（爸爸原来采用百度地图的配色方案）改为我中意的蓝色系。我一共挑选了由浅到深 6 个蓝色，并用这 6 个颜色编码

建立了配色数组。

3 月 17 日，我代表星辰研学项目组，正式向学校提交了研学成果，包括疫情数据库（EXCEL 格式）、源代码（Python 格式）和成果页面，成果页面发布在：

http://www.zxgzs.net/lzy/lzy_covid19_maps.html

三、研学收获

本次研学任务中，我们使用 Python 语言，绘制了 2020 年 3 月 8 日至 15 日期间，全球新冠疫情累计确诊病例动态地图（见图 5）、现存确诊病例地图和感染指数动态地图，统计了截至 3 月 15 日累计确诊排名前十的国家和感染指数排名前十的国家，建立了国外输入性风险评估模型，并以条形图的方式分析了 13 个国家（累计确诊病例超过 1000 例）的输入性风险指数（见图 6）。

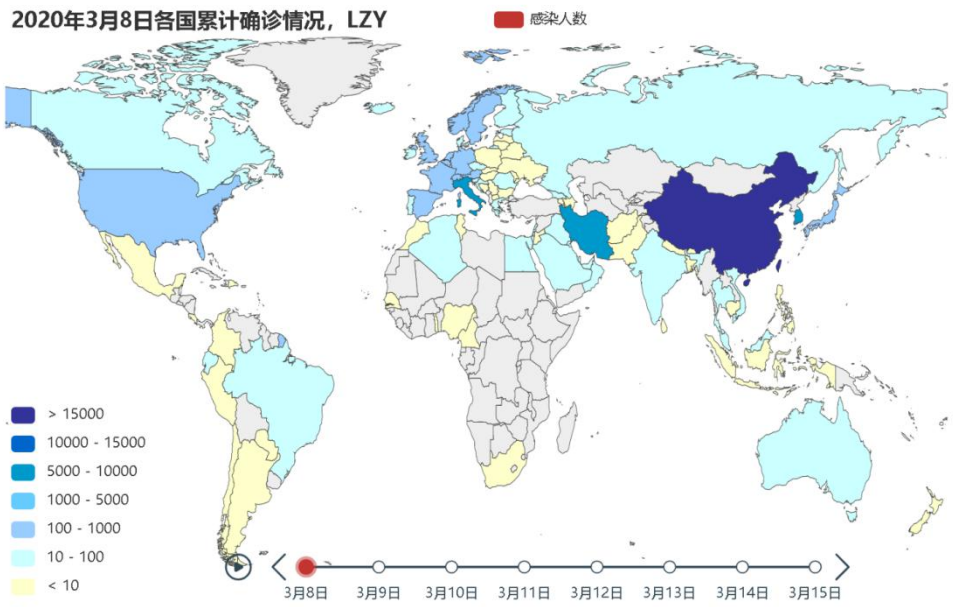

图 5 全球新冠疫情累计确诊病例动态地图截图

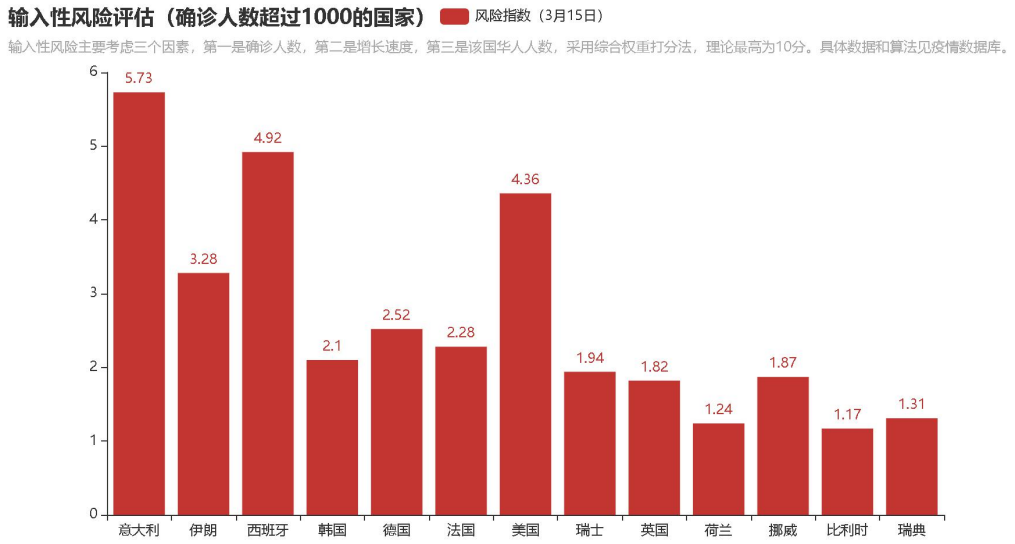

图 6　输入性风险评估指数统计图

通过研学任务，我初步掌握了 Python 代码编程的方法，提高了 EXCEL 数据表的应用能力，了解了更多的国家情况，对于地图化数据分析和数学建模也有了粗浅的认识。除此之外，我还有三点感受：

第一，付出努力必有回报。由于时差问题，我每天都必须深夜去录入数据，要去百度和世界卫生组织网站上去查找不同国家的疫情，录入过程还要反复核对数字。过程确实比较辛苦，但通过自己收集录入数据，对全世界的疫情有的更深入的了解，在后期与爸爸讨论地图设计和模型建立时都能提出自己的看法。

第二，疫情的发展与控制，证明了我们国家制度的优越性。通过对全世界各个国家疫情数据的趋势分析，我发现国外尤其是欧美的病例增长速度比我们国家更快，甚至比武汉还快，而且拐点迟迟未到来，这充分证明了我们国家的防控措施比国外做得更坚决有效。而这种有效的防控措施又进一步说明，我们的科学家更早认识到病毒的危害和规律，我们的人民更有牺牲精神和自律性，我们的政府执行力也更强。

第三，从疫情数据看，全世界疫情还远没有结束，国外疫情回流

风险越来越大，我们还应该继续保持警惕。我们还应认识到，欧美的新冠病例数和病死数字的高企，是建立在发达的医疗体系和较高的检测率之上。虽然目前南亚、非洲等地新冠病例数相对较少，但这种相对少的数字是建立不发达的医疗体系和较低的检测率之上，以新冠病毒的传染性来看，除了个别偏远孤岛外，没有哪个国家和地区能够逃离病毒的攻击。全球疫情很可能仍会有第三波、第四波，我们仍应当保持警惕，做好自身的防护，只有每个人都努力，团结一致，我们才能最终战胜病毒。

四、后续设想

属于我的研学战役告一段落，属于全球的战“疫”却是激战正酣，我和爸爸仍在高度关注全球疫情发展。进入三月中下旬，美国成为了全球新冠病毒新的爆发点。为了推卸前期防疫工作失误的责任，美国政府试图甩锅给中国，既不体面，也不科学。

我们希望接下来能有机会，用客观的数据进行分析，来反驳美国政府针对中国的甩锅行径。我们初步设想：录入完整的全球疫情数据记录，分析从 1 月 23 日以来的全球疫情扩散情况（见图 7）；然后进一步通过疫情扩散数据和病毒潜伏周期，结合各国不同时期的防疫政策，分析防疫政策对疫情发展的影响，提出我们的建议。

最后，感谢我的妈妈，她虽然没有参与到我的研学任务中，但多亏她管住了我调皮捣蛋的小弟弟，使我们有了安宁的环境来开展研学；更要感谢千千万万医护人员和逆行者，没有她（他）们，就不可

能有我们现在相对安宁的环境，也不可能有以上的研学报告。

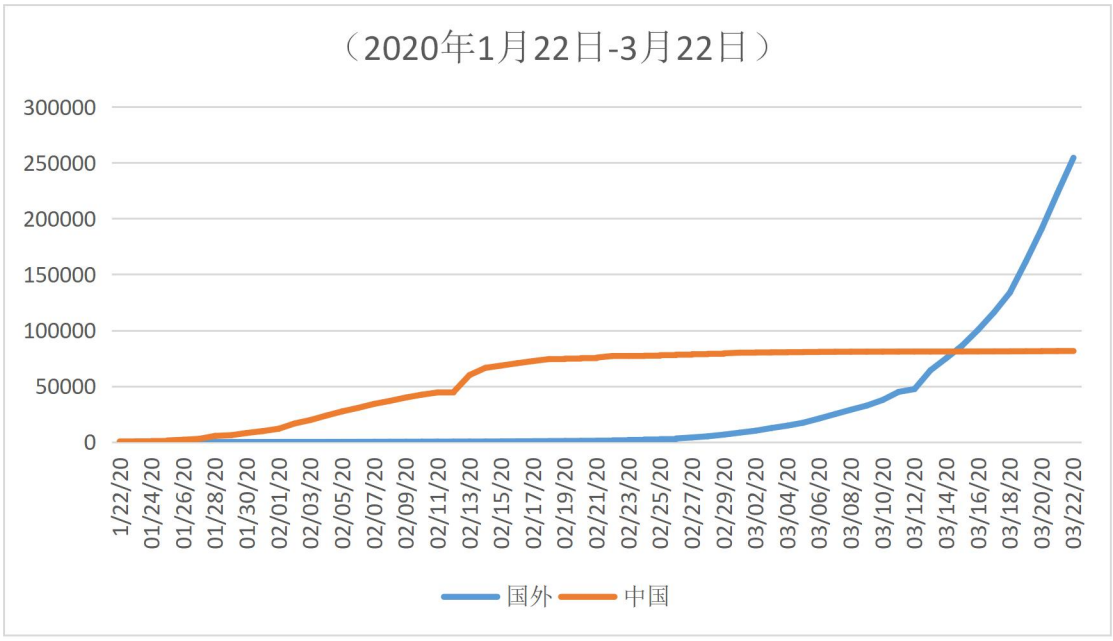

图 7 国内外两个月疫情发展趋势