学号：20123101
姓名：李昀哲
报告题目：基于矩阵代数工程应用的上海疫情发展分析

基于矩阵代数工程应用的上海疫情发展分析

1. 前言

上海疫情可谓是过去三个月间的焦点话题，至今病毒虽仍未消退，但正在逐渐向好发展。疫情肆虐期间，经济、民生受挫，上海各区先后破防；每日疫情通报的病例数牵动了全上海乃至全国人民的心。

本报告将结合矩阵代数所学知识，基于上海疫情数据、上海各区行政数据、经济数据等，研判影响上海各区疫情的主要因素；预测上海疫情发展（也可看作对目前发展的验证）；分析奥密克戎病毒特点。旨在利用矩阵代数所学通过建立函数、方程来确立各变量的关系，发挥本课程“应用”的特点，以报告形式将理论联系实际，培养分析问题、思考问题的能力。

2. 相关研究

早在四月中上旬，各大院校就对本轮上海疫情走向进行了预测和分析，如表 1 所示为预测较为合理的几项研究。结合研究成果中的预测数据和当下疫情现状，可以看出：上海交通大学对于拐点日的预测[1]最为准确，而西安交通大学在社会面清零日、感染总人数的预测[2]上和事实最为接近，拐点的预测偏差也不大，因此西安交通大学的研究大体上和事实吻合，值得作为参考、学习的资源。

表 1 上海疫情发展预测研究现状

研究团队	拐点	社会面清零	感染总人数
上海财经大学	4 月 9 日	4 月 16 日	170,000
上海交通大学	**4 月 13-15 日**	/	300,000
西安交通大学	4 月 12 日	**5 月 14-21 日**	**474,000**
南开大学	4 月 10-14 日	5 月 13-15 日	455,758
兰州大学	4 月 10 日	5 月 3 日	301,740
实际数据	**4 月 15 日**	**5 月 17 日**	**600,000+**

西安交通大学的这项研究由西安交通大学人工智能与机器人研究所新冠疫情趋势预测课题组完成，于 2022 年 4 月 9 日发表。基于国家卫健委员官方数据，利用 ISI 传染病模型进行预测。其中对于数据的处理、模型的选择、预测的方法并不难以理解，同时都具有很高的严谨性和统计素养，对本科阶段的学生而言，是个很值得研读的论文。

这些研究对于拐点、感染总人数等有了较为深入的研究，但对于引发如此大规模疫情爆发的缘由、上海各区爆发顺序的因素等却研究不够，网络也基本检索不到任何类似的研究。同时对于感染者的年龄分布、病毒传播特点、爆发街道（地区）疫苗接种率等等，都是值得关心的角度，对于未来病毒的防守和阻隔都起到关键作用。

3. 方法

3.1 数据集

由于附件中提供的“上海疫情数据.xlsx”已基本囊括了绝大部分数据，因此仅补充 4 月 8 日举行的上海第 147 场疫情防控工作新闻发布会上公布的：截止于 4 月 7 日 24 时，上海累计感染者 131,524 例的年龄结构分布，此分布的采样对象为 13 万上海感染者，数据量足够大，可以看作总体的分布。如图 1 所示，其中，年龄最小者仅为 10 天，年龄最大者 98 岁。

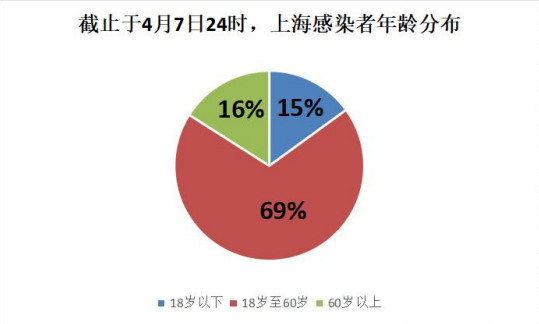


图 1 截止于 4 月 7 日，上海感染者年龄分布

3.2 数学模型

结合湖北武汉 2019 年新冠疫情爆发时的数学模型分析[3]，加之上海市对于疫情防控的策略调整，本次预测感染人数的模型选择高斯函数进行拟合。拟合函数为

$$y = ce^{-\frac{(x-b)^2}{a}} \tag{1}$$

数据上的选择以 3 月 28 日为起始点，起始点之后 14 天的数据作为初始拟合数据。得到的感染人数模型，将对之后三天的感染人数情况作短期预测。同时模型会将每一天过后的数据加入拟合数据集，对模型进行微调，从而对实际情况进行验证。

检索相关文献发现，还可以根据防疫政策等因素，通过 SIR 模型求解[4]，易感染者，感染者，移出者之和是个恒量即 N = S + I + R。考虑病人康复后具有免疫力，人与人之间有相同的接触率等因素得出微分方程求解，但考虑到当时武汉疫情防控和当下上海的差异以及个人的防护意识、群体免疫等因素，就暂且仅考虑较为一般的指数模型。

3.3 算法

3.3.1 研判上海各区数据与疫情发展关系

研判上海各区数据主要通过读入“上海疫情数据.xlsx”数据集中“各区信息”，将数据表转化为矩阵，算法的核心在于降维处理。降维在矩阵代数与应用中使用主成分分析是较好的方式，其中算法选择奇异值（SVD）分解，使用 pca()函数得到上海各区因素矩阵的特征向量、特征值和主成分系数。

3.3.2 拟合上海疫情发展模型

基于 3.2.1 中模型的数据集以及所要拟合的函数，拟合方法曾尝试使用最速下降法，利用矩阵微分确定梯度方向的方法。但由于步长的确定比较困难，步长制定过大就难以收敛，

太小又收敛速度过慢，虽然最终找到了较为合适的步长，但过程过于繁琐，因此仅作为一种尝试，最终还是使用最小二乘法。同时，在每一天过后对于数据集的变更，对拟合结果进行微调。

3.3.3 病毒特点

计算病亡和确诊、无症状感染者之间的比例关系得出一般结论。参考对比今年 5 月 *Nature* 杂志最新发表的两篇论文[5,6]，验证结论。

4. 数据实验

4.1.1 研判上海各区数据与疫情发展关系

主成分分析是常用的对数据进行降维处理的方式，可以帮助剔除冗余信息，找出影响最大的因素。本质上来看，就是将坐标系旋转，使得主要随机变量在旋转后坐标系下的起伏和波动具有代表性。

具体的实现可以从基本原理出发通过求期望、协方差矩阵、奇异值（SVD）分解（或特征值分解）和 *Hotelling* 变换几个步骤得到特征值、特征向量以及主成分系数；同时也可以使用 Matlab 中的 pca()函数进行直接计算，二者的结果是相同的。

读入数据集中数据后，得到一个 16x4 的矩阵。可将其看作影响因素矩阵，矩阵的每一行为一个区的“人口”、“占地面积”、“60 岁以上人口”和“GDP”。使用 pca()函数时，其中的“算法”参数选择奇异值（SVD）分解，原因为：在使用另一种算法（特征值分解）时，无法避免计算 X^TX，而在观测变量个数 n 非常大时，这一算法的劣势将被无限放大，协方差矩阵为 *n x n* 维；而采用奇异值（SVD）分解，只需要计算

$$\mathrm{T}_r = U_r \sum\nolimits_r \tag{2}$$

而 $\sum_{\mathrm{r}}$ 为对角阵，显然这一算法的计算量小很多。

主成分分析结果如图 2 所示，“人口”的主成分占比高达 98.14%，“占地面积”占比 1.828%，因此为次成分。剩下的“60 岁以上人口”以及“GDP”占比相对较低。

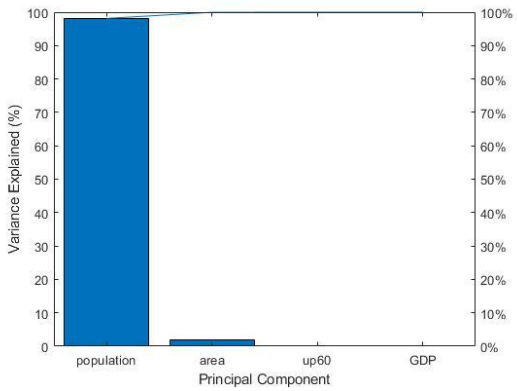


图 2 上海各区疫情影响因素主成分分析

4.1.2 拟合上海疫情发展模型

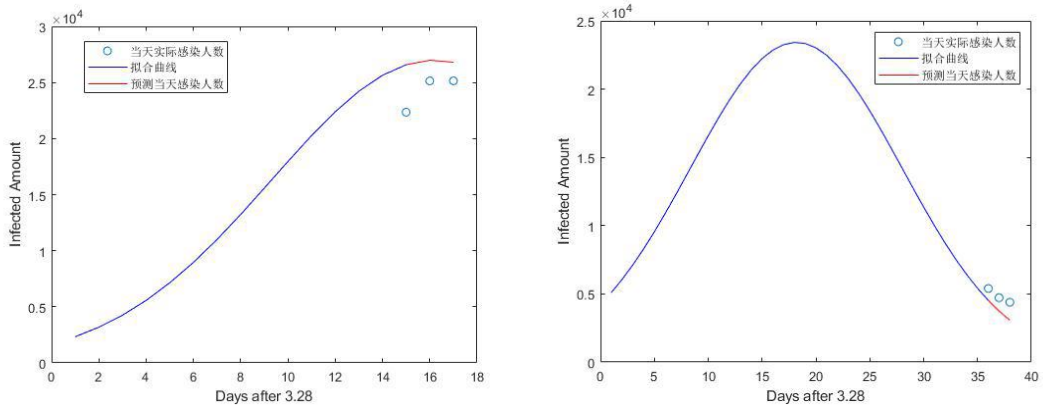
根据 3.2.1 的模型选择和数据集，首先进行以起始点后 14 天的函数拟合。拟合算法的选择见 3.3.2 所述，测试后选择最小二乘拟合，单独编写 createFit()函数进行拟合以适应每过一天加入当天数据再次微调拟合的情况。

数据实验基本思路为：

1. 采集自 3 月 28 日以来 14 天的每日新增感染者数据；
2. 进行最小二乘拟合，拟合 Gauss 函数；预测之后三天数据并计算拐点
3. 随后的每一天都将当天数据加入历史数据集用于更正曲线的拟合，同时展示后三天的预测数据。

其中添加“date_want_to_know”变量用以选择查看某一天预测的情况，可根据实际需要修改查看。基于 14 天数据拟合曲线后，计算得到拐点为 4 月 14 日。

短期预测的图例中，还添加了预测当天实际的感染人数的散点，可有直观的对比。如图 3(a)所示为以最初 14 天拟合的曲线，其中红色曲线为基于前 14 天数据预测结果，淡蓝色散点为实际真实结果，图 3(b)展示了收集了 35 天的数据后拟合的效果。不难发现，随着数据量的增多，拟合效果也逐渐变好。



(a) 初始 14 天数据拟合结果　　(b) 35 天时数据拟合结果

图 3 短期预测拟合效果

模型的量化标准为：拟合曲线各点和实际值曲线各点的方差，方差越小，拟合精度越高。如表 2 所示为各天数数据下拟合方差对比，不难发现，拟合精度基本和数据规模是正相关的，拟合数据规模越大，方差越小，精度越高。

表 2 模型拟合精度评估

数据集天数	14	20	25	30	35	40
方差	4227.1	4994.4	3825.1	2279.6	2220.6	2198.8

4.1.3 病毒特点

通过计算病亡者和无症状感染者及确诊者总数的比例关系，计算得出病亡的比率占全部确诊、感染者人数的 0.09%，因此我们认为本轮奥密克戎病毒的传播性虽强，但致死率并不高。同时根据上海发布给出的病亡者病因，并不完全是因为奥密克戎病毒导致。由图 4 可知，近 5 年上海的死亡率始终位于 5‰附近，为验证奥密克戎的致死率受其他各种疾病影响，只需在年末再次观测该数据，和近年进行对比。

根据 *Nature* 杂志 5 月 20 日新发表的两篇论文，更严谨的证明了奥密克戎病毒相较于之前毒株提高了免疫逃逸能力，但毒力有所下降。

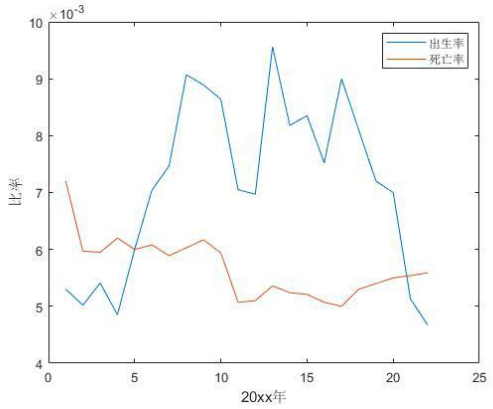


图 4 上海近 20 年出生死亡率曲线

5. 讨论

代数是用于研究数、数量、关系、结构之间联系的重要方式，矩阵、向量又是工程领域常用于描述事物量的载体，因此本报告基于矩阵、向量建立函数、方程来确立各量之间的关系，进行数据实验和分析讨论问题。

对于主成分分析实验而言，这里仅对可量化的因素进行了评估，但仍有很多不可量化却有极高评估价值的因素，如：行政区的基础设施水平（影响通风等），行政区群体免疫力（可通过疫苗接种率计算但数据难以获得）、行政区居民流动性（携带病毒可能）等等，都是值得关心的因素，因此在更严谨、更精确的研判中，都应建立更完整的模型进行评判；对于预测上海疫情发展而言，这里是站在当局者的角度，从疫情的末尾来观测问题，假设自己身处 4 月 10 日，用仅有的数据进行拐点、发展的预测，指数模型已经可以达到较为准确的效果，培养了一定的思考方法，但考虑因素并不像西安交大、南开大学那么完整，仍有提高空间；对于病毒特点而言，就更为抽象，不仅涉及数据方面的处理，还涉及传染病学等专业知识，这也从另一个角度应证了计算机始终不会是单一的、孤立的学科，而是需要应用、交叉的，这个实验中也得到了体现。

同时，如感染者年龄分布等关键数据，是考量病毒破防能力和制定治理方案的关键，本报告仅基于发布会中 13 万人的数据进行评判，虽然已经具有一定代表性，但后续为得到更细化的年龄分布，还可考虑编写代码获取官方 3 月 25 日前官方每日通报的数据进行分析。

矩阵永远不能只停留在课本，正是这样的实践，才能真正推进我们对于问题的理解。

6. 结论

综合上述分析和实验可得以下结论：

(1) 通过主成分分析，得出“人口”因素是第一主成分，“占地面积”因素是第二主成分；

(2) 疫情预测：通过 14 天的初始数据，预测拐点为 3 月 28 日后的 16 天，即 4 月 14 日，而实际数据为 4 月 15 日，大体接近；

(3) 病毒特点：传染性强、毒力一般。

参考文献

[1] 上海交通大学模型预测：https://new.qq.com/omn/20220429/20220429A0D2GP00.html

[2] 西安交通大学人工智能研究所研究结果：http://www.aiar.xjtu.edu.cn/info/1004/2371.htm

[3] 深圳大学管理学院新型冠状病毒感染人数模型：https://ma.szu.edu.cn/info/2529/5249.htm

[4] 2019-nC0V 新型冠状病毒传染分析：https://zhuanlan.zhihu.com/p/104376394?ivk_sa=1024320u

[5] Uraki, R., Kiso, M., Iida, S. *et al.* Characterization and antiviral susceptibility of SARS-CoV-2 Omicron/BA.2. *Nature* (2022).

[6] Suryawanshi, R.K., Chen, I.P., Ma, T. *et al.* Limited cross-variant immunity from SARS-CoV-2 Omicron without vaccination. *Nature* (2022).