Advances in Applied Mathematics 应用数学进展, 2022, 11(7), 4219-4227
Published Online July 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.117449

Hans汉斯

基于Logistic增长模型和相关性分析的疫情预测
——以英国为例

李　乾

华北电力大学，北京

收稿日期：2022年6月4日；录用日期：2022年6月28日；发布日期：2022年7月5日

摘　要

目前，新冠肺炎病毒已经在世界各国传播。随着环境的变化和感染者人数的上升，病毒开始不断发生突变和进化，至今已经产生了11种变异毒株。由于发现不及时，各国在进行基因测序统计时，变种病毒往往已经开始传播了一段时间，缺乏初期传播的感染人数数据。本文以英国为例，根据病毒传播的特征，我们选择构造Logistic增长模型来模拟病毒感染过程，求得了各病毒的感染总人数曲线表达式，计算出了传播开始时间以及持续时间。此外，通过计算Pearson相关系数，对各类气候因素与病毒增长人数进行相关性分析，发现每日增长感染人数与温度和湿度相关性最大，且与温度呈负相关、湿度呈正相关。

关键词

疫情预测，Logistic增长模型，相关性分析，Shapiro-Wilk检验，Pearson相关系数

Epidemic Prediction Based on Logistic Growth Model and Correlation Analysis
—Taking the UK as an Example

Qian Li

North China Electric Power University, Beijing

Received: Jun. 4th, 2022; accepted: Jun. 28th, 2022; published: Jul. 5th, 2022

Abstract

Nowadays, the COVID-19 virus has spread all around the world. As the environment changed and the number of infected people rose, the virus began to mutate and evolve. So far, 11 mutated strains have emerged. Because of the delay in detection, the mutated strains have already started

文章引用: 李乾. 基于 Logistic 增长模型和相关性分析的疫情预测[J]. 应用数学进展, 2022, 11(7): 4219-4227.
DOI: 10.12677/aam.2022.117449

to spread for some time, giving rise to the lack of data on initial stage. Without losing generality, we shall discuss the case in the UK in this paper. According to the characteristics of the virus transmission, we build Logistic growth models to simulate the process of the spread of virus and obtain the expression of the number of being infected in pace with time. As a result, we obtain the start time and duration of the virus. Additionally, we analyze the correlation between daily increase number and climatic factors in terms of the Pearson correlation coefficient and we find that the daily increase number is most correlated with temperature and humidity, specifically negative affected by temperature and positive affected by humidity.

Keywords

Epidemic Forecast, Logistic Growth Model, Correlation Analysis, Shapiro-Wilk Test, Pearson Correlation Coefficient

1. 引言

从新冠病毒在2019年底爆发以来，新冠病毒至今已经产生了11种变异毒株，包括我们所熟知的Alpha、Beta、Gamma、Delta等。经过各国以及各地区的基因测序抽检，发现Alpha、Delta、Delta变种(VUI-21OCT-01)变种病毒相对其余变种病毒而言，感染人数增长更加明显，感染人数总量更多。因此，我们可以通过这些病毒感染数据来计算出它们的传播速度和持续时长，从而划分疫情发展阶段。由于基因测序工作的延迟性，往往病毒已经开始传播一段时间才进行测序，所以经常无法确定病毒的确切开始时间，需要根据现有病毒传播特性以及现有数据进行反向推理。除此之外，温度、湿度与大气压等气候因素会对病毒的传播产生影响，因此，研究这些因素与新增感染人数的相关性，对预测和控制疫情发展具有很大的作用。

对于疫情感染人数的预测一直以来都是控制疫情传播的重点工作。国内外学者已经建立了许多相关预测模型，包括适用于小数据量的灰色系统预测模型[1]、适用于大数据量的ARIMA模型[2]、指数平滑模型[3]以及基于传染病动力学的SI、SIR、SIRS、SEIR模型[4]等。灰色系统预测、ARIMA模型和指数平滑模型都是先探索原始数据本身所具有的规律性，从而对未来的发展趋势进行预测，忽视了病毒本身所具有的传播规律。而SEIR模型等则是根据传染病传播和染病的过程将人群分类，从而构建相应的微分方程进行模拟，需要知道病毒实际的治愈率、致死率等数据，要求较高。

通过分析本文数据，发现已知数据类型仅有病毒感染总人数一种数据，所以传染病动力学模型并不适用。除此之外，已知的感染数据大多数都是病毒传播的中后期数据，感染总人数增长逐渐停止，因此使用灰色预测或ARIMA模型向后预测意义不大。考虑实际情况中，病毒传播的前期没有及时进行干预，病毒增长不受控制，增速较快。而在中后期，感染者中的部分被治愈而获得抗体，加上政府开始实施防疫措施，病毒的增速逐渐下降直至停滞。所以，我们可以类比人口增长[5]和种群增长[6]，在已知病毒的传播特性和数据存在缺失且类型单一的情况下，采用Logistic增长模型[7]对病毒传播过程进行模拟，从而预测病毒的开始时间和初期传播情况。

2. 数据来源

本文通过英国新冠疫情官方统计网站，英国政府官方网站公开发布的信息，获取了英国全国以及379

个地区 2020 年 2 月 1 日至 2022 年 3 月 31 日每日新增病例数据以及这 379 个地区在这段时间内每日的温度、露点(一个气象学概念)、湿度、风速、压强数据。通过英国政府发布的每日基因测序抽检结果，得到了自 2021 年 5 月 20 日开始统计至 2022 年 3 月 31 日期间每周的 Alpha、Delta、Delta 变种(VUI-21OCT-01)的病毒感染数量增长数据。获得的数据时间跨度广，总体质量高，可靠性强。

3. 模型假设

3.1. 模型的假设

- 病毒感染的总人数数据连续；
- 不同病毒之间存在竞争关系；
- 英国人口总数维持基本恒定，即出生率等于死亡率；
- 新冠病毒及其变种的感染速度受环境因素和已感染人数影响，且存在上限；
- 某种病毒感染人数至上限时，未来将不会继续传播。

3.2. 符号说明

t ——时间；

r_1, r_2, r_3 ——Alpha 病毒、Delta 病毒、Delta 变种病毒的固有增长率；

$x_1(t), x_2(t), x_3(t)$ ——Alpha 病毒、Delta 病毒、Delta 变种病毒感染总人数；

$r_1(x), r_2(x), r_3(x)$ ——Alpha 病毒、Delta 病毒、Delta 变种病毒感染增长率；

x_{m1}, x_{m2}, x_{m3} ——Alpha 病毒、Delta 病毒、Delta 变种病毒最大感染人数；

t_{01}, t_{02}, t_{03} ——Alpha 病毒、Delta 病毒、Delta 变种病毒感染初始时间；

x_{01}, x_{02}, x_{03} ——Alpha 病毒、Delta 病毒、Delta 变种病毒感染初始感染数。

4. 模型的建立、求解和分析

4.1. Logistic 增长模型

由于英国政府在疫情持续了一段时间之后才开始对大众进行基因测序抽检，所以目前我们仅能掌握每种病毒一段时间内的感染人数数据，无法判断病毒开始的传播的时间、持续时间以及目前病毒传播的阶段。考虑到病毒的增长受到环境中人口数量、人口密度、气候等因素的影响，随着感染数量的增加，感染速度将逐渐缓慢，直到感染数量到达上限时，增长完全停止。根据 Logistic 增长模型[7]，病毒感染率 $r_i(x)$ 会随着人口数量 $x_i(t)$ 的增长而下降。我们不妨设 $r_i(x)$ 为 $x_i(t)$ 的线性减函数，即：

$$r_i(x) = r_i - s_i x_i(t) \tag{1}$$

r_i 为病毒的固有增长率，表示感染初期($t=0$ 时)感染人数很少时的增长率。根据假设，当达到最大感染数量 x_{mi} 时，病毒不会继续传播，则 $r_i(x_{mi}) = 0$， $s_i = r_i / x_{mi}$，从而：

$$r_i(x) = r_i - \frac{r_i}{x_{mi}} x_i(t) \tag{2}$$

建立感染人数 Logistic 增长模型的微分方程组：

$$\begin{cases} \dfrac{\mathrm{d}x_i(t)}{\mathrm{d}t} = r\left(1 - \dfrac{x_i(t)}{x_{mi}}\right) x_i(t) \\ x_i(t_{0i}) = x_{0i} \end{cases} \tag{3}$$

对其分离变量进行求解：

$$\frac{1}{x_i(t)}dx_i(t)-\frac{1}{x_{mi}-x_i(t)}d\left(x_{mi}-x_i(t)\right)=r_i dt$$

$$\ln x_i(t)-\ln\left(x_{mi}-x_i(t)\right)=r_i t+C$$

$$\ln\frac{x_i(t)}{x_{mi}-x_i(t)}=r_i t+C$$

$$x_i(t)=\frac{x_{mi}}{1+\mathrm{e}^{-r_i t}/C}$$

代入初值条件，得到最终感染人数随时间的表达式：

$$C=\frac{x_{0i}}{x_{mi}-x_{0i}}\mathrm{e}^{-r_i t_{0i}}$$

$$x_i(t)=\frac{x_{mi}}{1+\left(\frac{x_{mi}}{x_{0i}}-1\right)\mathrm{e}^{-r_i(t-t_{0i})}} \tag{4}$$

根据上述结果表达式，我们可以依次求出原始病毒、Alpha 病毒、Delta 病毒的阻滞增长拟合曲线。由于英国全国病毒感染总人数的数据是从 2020 年 1 月 30 日开始的，我们不妨设这一天为 $t=1$。

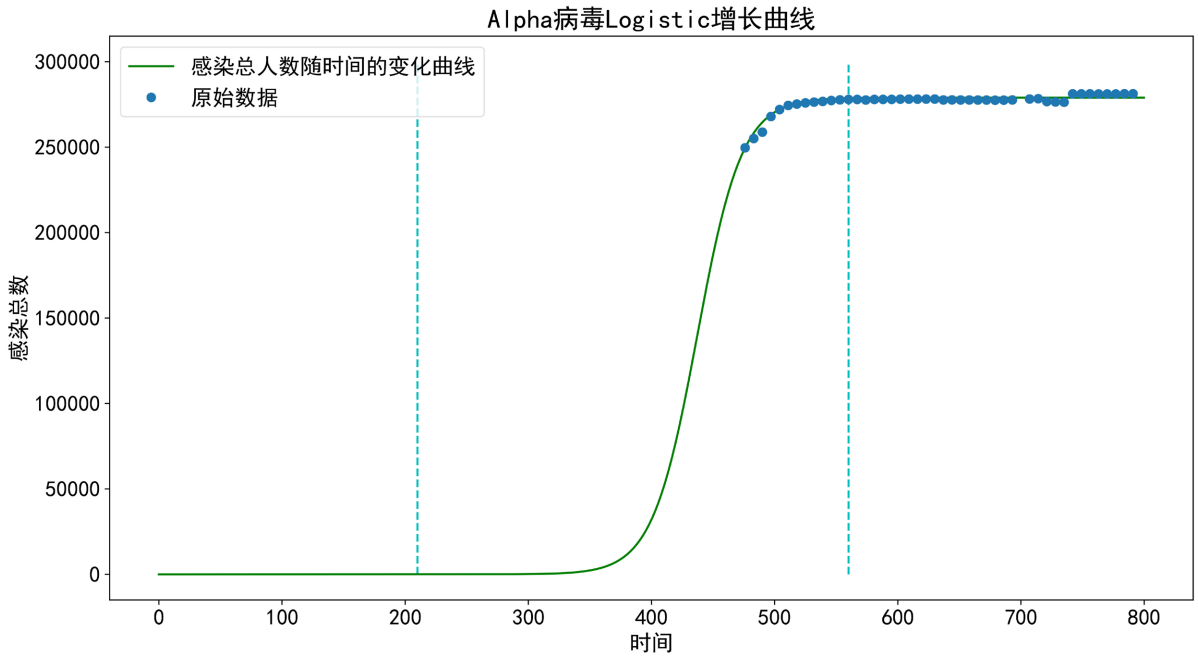


Figure 1. The Logistic growth curve of Alpha virus
图 1. Alpha 病毒 Logistic 增长曲线

首先考虑 Alpha 病毒。从已知数据来看 Alpha 病毒感染人数具体统计数据是从 2021 年 5 月 20 日开始记录的，当日总感染人数为 249,637 人，则 $t_{01}=476$，$x_{01}=x_1(t_{01})=249,637$。为了判断 Alpha 病毒的

开始时间，我们使用公式(4)进行拟合，编写代码对上述公式中的未知数进行最优拟合求解，最终得 $x_{m1}=279{,}000$，$r_1=0.055$，拟合曲线如图 1 所示。通过求得的感染人数表达式，代入已知数据的时间点，求出预测数据与实际数据进行对比，求得可决系数 $R^2=0.924$，说明拟合优度良好。

通过图 1 可以看出，目前已知 Alpha 病毒数据属于传染后期数据。为了求得病毒开始时间，只需令 Alpha 病毒的 Logistic 增长曲线表达式 $x_1(t)=1$，计算得出 $t=210$，推断 Alpha 病毒的实际开始时间为 2020 年 8 月 28 日左右，与新闻报道的英国首例 Alpha 病毒发现时间基本重合。从图 1 观察到，$t=560$ 时，感染人数不再继续上升，即可视为 Alpha 病毒到达感染上限。所以 Alpha 病毒在英国的持续时长约为 350 天，基础增长率为 0.055。

接下来考虑 Delta 病毒，通过已知数据得到初值条件 $t_{02}=469$，$x_{02}=x_2(t_{02})=1213$。同上述步骤，编写程序求解，得 $x_{m2}=1{,}740{,}000$，$r_2=0.05$，拟合曲线如图 2 所示。和实际数据比较，$R^2=0.928$，说明拟合程度良好。

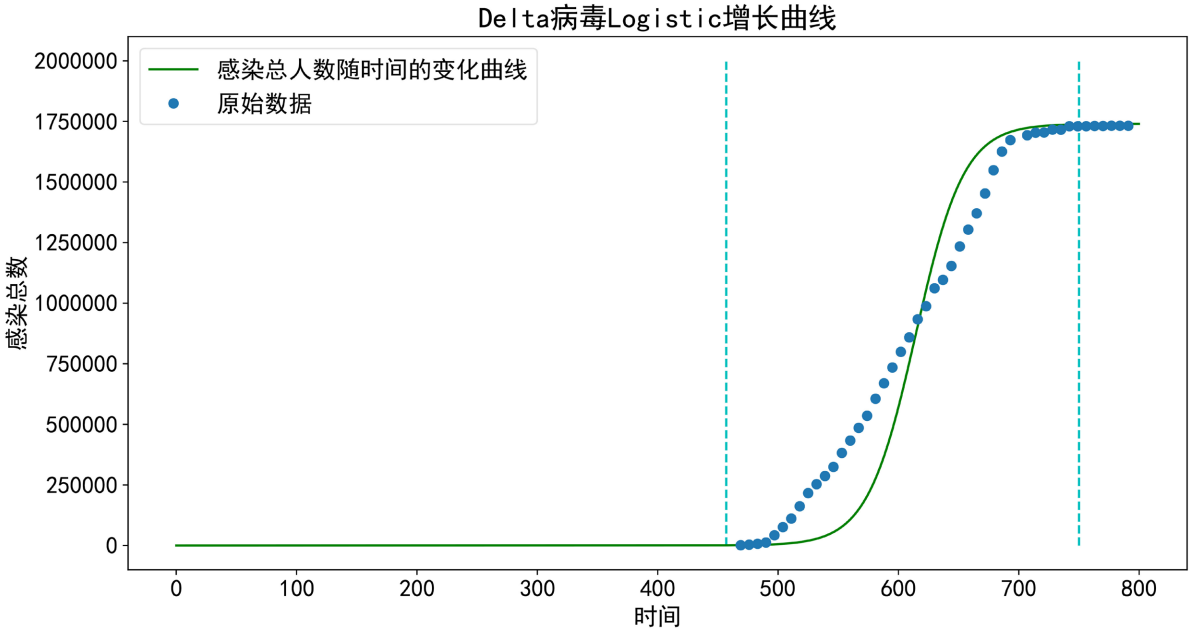


Figure 2. The Logistic growth curve of Delta virus
图 2. Delta 病毒 Logistic 增长曲线

由于 Delta 病毒在英国为境外输入型病毒，所以病毒的开始时间不能直接用求得的曲线表达式求出。通过查阅资料，Delta 在英国的开始时间约为 2021 年 5 月 1 日，即 $t=457$，通过图像观察 Delta 感染达到上限时间为 $t=750$。所以 Delta 病毒在英国的持续时长约为 293 天，基础增长率为 0.05。

最后考虑 Delta 变种病毒，通过已知数据得到初值条件 $t_{03}=637$，$x_{03}=x_3(t_{03})=21{,}146$。同上述步骤，编写程序求解，得 $x_{m3}=123{,}500$，$r_3=0.075$，拟合曲线如图 3 所示。和实际数据比较，$R^2=0.919$，说明拟合程度良好。

根据所求得的公式，令 $x_3(t)=1$，求得 $t=502$，推断 Delta 变种病毒在英国的实际开始时间为 2021 年 6 月 15 日左右。考虑到病毒的潜伏期，所以拟合结果基本与新闻报道的日期相符。通过观察图 3，可以看出所记录的 Delta 变种病毒感染数据为中后期数据，感染人数到达上限时间为 $t=780$。所以 Delta 变种病毒在英国的持续时长约为 378 天，基础增长率为 0.075。

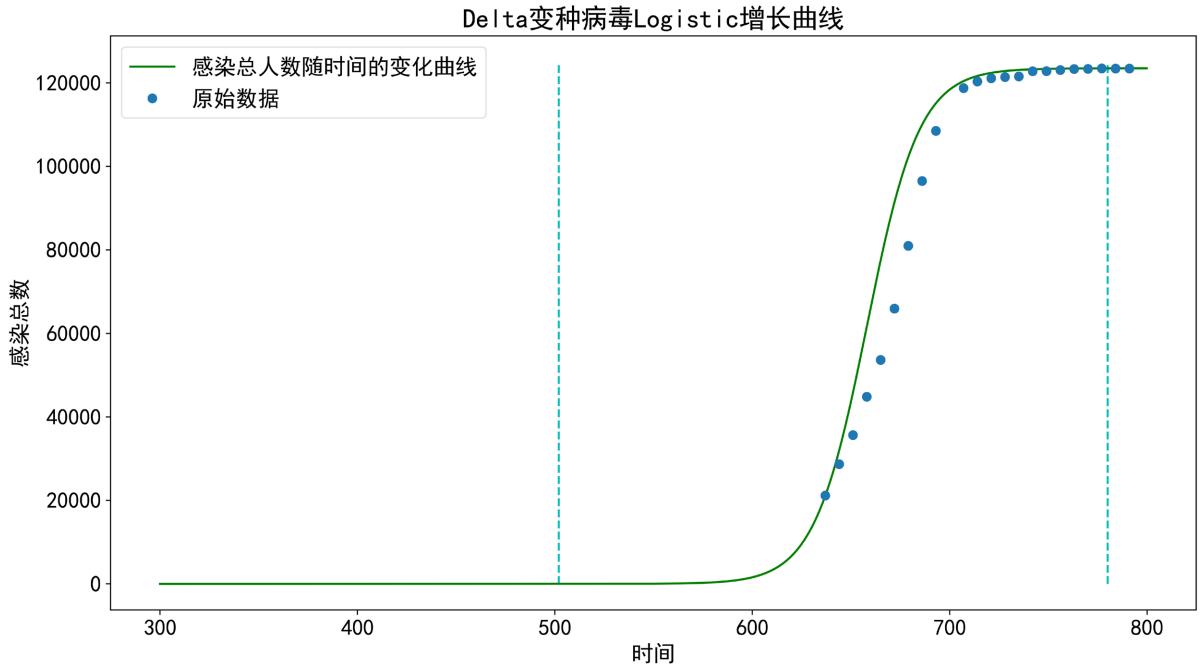


Figure 3. The Logistic growth curve of mutated Delta virus
图 3. Delta 变种病毒 Logistic 增长曲线

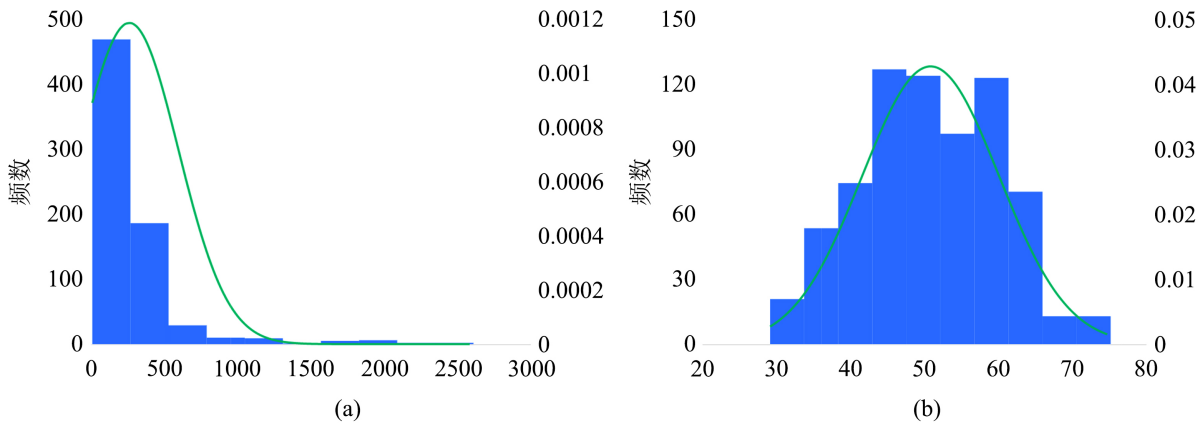
4.2. 相关性分析

接下来考虑气候因素对病毒传播速度的影响。我们不妨选择位于英格兰的曼彻斯特作为特例进行研究。根据英国官方网站的统计数据，我们获得了曼彻斯特 2020 年 4 月 1 日至 2022 年 3 月 31 日每日的新增感染人数、温度、露点、湿度、风速、压强数据。对这些变量进行 Pearson 相关性分析。

4.2.1. 数据的正态性检验

首先对数据的进行正态性检验。由于各变量的数据数量未超过 5000，即样本为小数据样本，所以我们采用 Shapiro-Wilk 检验。使用 SPSS 进行分析[8]，绘制正态性检验直方图，见图 4。

通过正态性检验结果直方图来看，各因素的图像基本上呈现出正态分布的钟形(中间高，两端低)，说明虽然数据不是绝对正态，但是基本可以接受为正态分布，因此可以选择构造 Pearson 相关系数对各变量进行相关性分析。



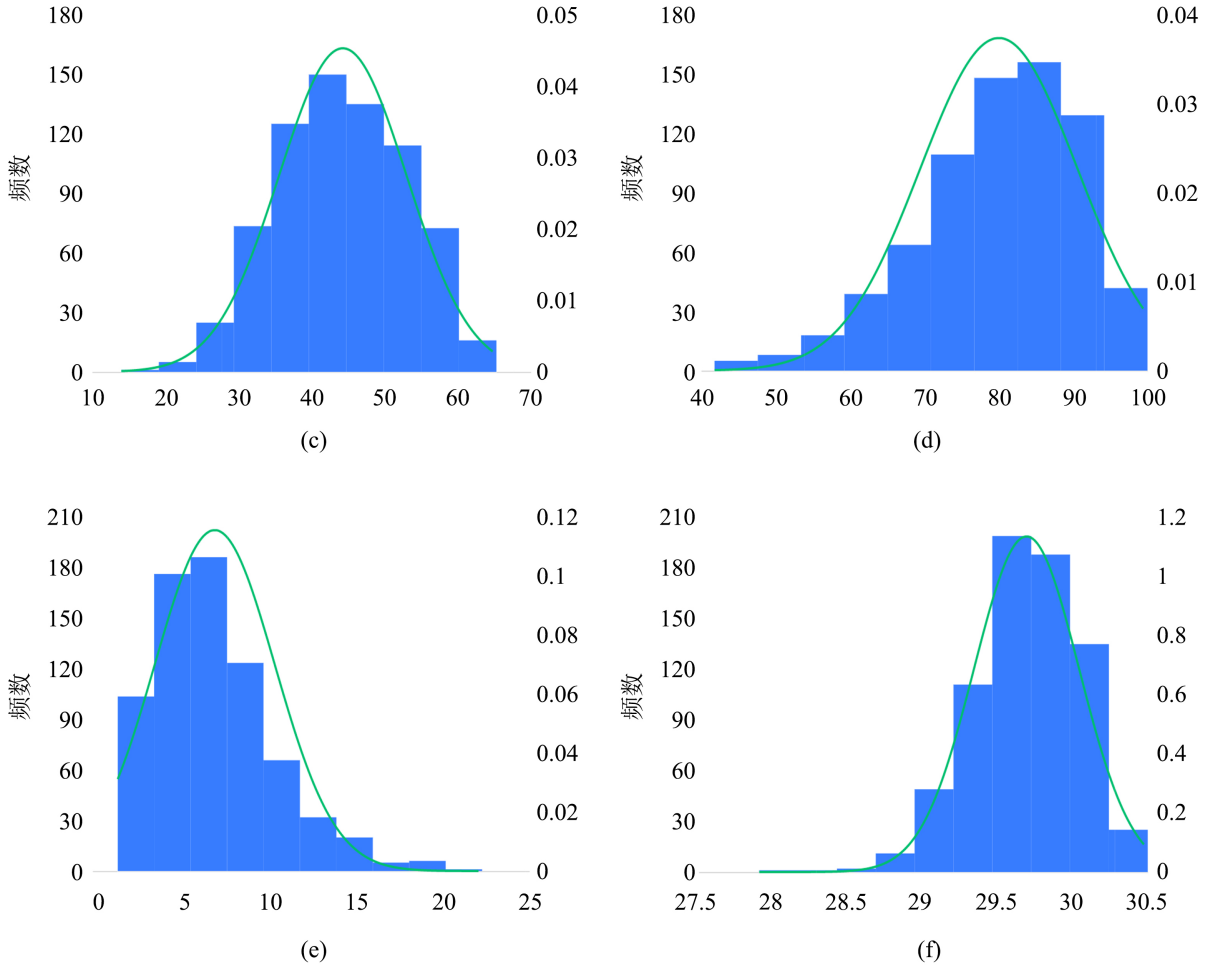


Figure 4. The histograms of normality test: (a) the number of daily new cases; (b) temperature; (c) dew-point temperature; (d) humidity; (e) wind speed; (f) atmospheric pressure
图 4. 正态性检验直方图：(a) 新增感染人数；(b) 温度；(c) 露点；(d) 湿度；(e) 风速；(f) 压强

4.2.2. 相关性分析

基于正态分布检验结果，我们对各变量构造 Pearson 相关系数。设任意两个变量之间的样本点为 (X_i, Y_i)，两变量的样本均值分别为 $\overline{X}, \overline{Y}$，则它们之间的 Pearson 相关系数可以定义为[9]：

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{5}$$

使用 SPSS 代入数据进行求解，求得两两变量之间的显著性 P 值以及 Pearson 相关系数，如下图 5 所示。

从表 1 和图 5 上可以看出，每日新增病例数量与温度和湿度的显著性 P 值均小于 1%，即都具有显著性，说明两变量之间存在相关性。通过相关系数热力图可以看出，新增病例数量与温度和湿度的 Pearson 相关系数的绝对值相对其他变量较大，分别为−0.218 和 0.293，说明新增数量主要受这两个因素的影响，且与温度呈负相关，与湿度呈正相关。

Table 1. P values of each variable
表 1. 各变量显著性 P 值

	新增	温度	露点	湿度	风速	压强
新增	0.000***	0.000***	0.014**	0.000***	0.017**	0.413
温度	0.000***	0.000***	0.000***	0.000***	0.054*	0.459
露点	0.014**	0.000***	0.000***	0.000***	0.029**	0.002***
湿度	0.000***	0.000***	0.000***	0.000***	0.096*	0.000***
风速	0.017**	0.054*	0.029**	0.096*	0.000***	0.000***
压强	0.413	0.459	0.002***	0.000***	0.000***	0.000***

注：***、**、*分别代表 1%、5%、10%的显著性水平。

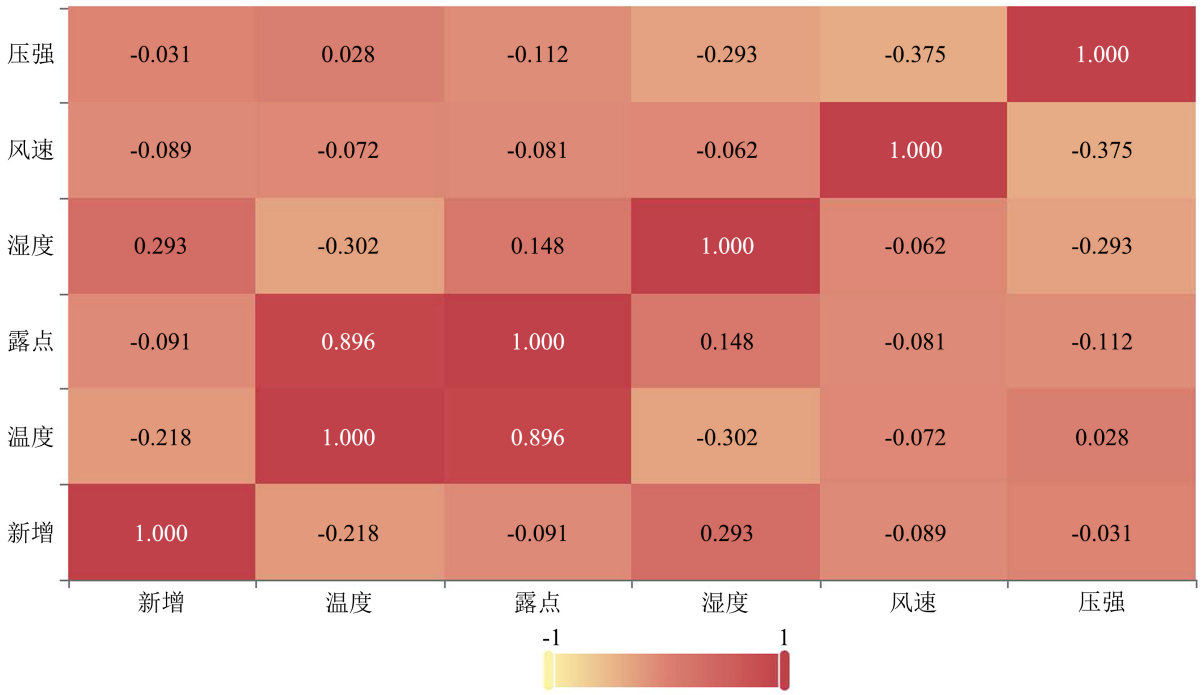


Figure 5. The heat map of Pearson correlation coefficients between each variable
图 5. 各变量之间的 Pearson 相关系数热力图

5. 结论

本文在英国政府官网上公布的病毒感染数据基础上，通过构造 Logistic 增长曲线，我们求得了 Alpha 病毒、Delta 病毒、Delta 变种病毒在英国的传染总人数曲线图，分别计算出了三种病毒的基础增长率，求出了 Alpha 病毒、Delta 变种病毒的开始时间，和新闻报道的第一例出现时间基本吻合，在此基础上求出了三种病毒各自的持续时间。选择曼彻斯特作为研究对象，通过正态性检验和构造 Pearson 相关系数分析温度、露点、湿度、风速、压强等气候因素与每日新增感染人数的相关性。求得每日新增人数受温度和湿度影响最大，且新增人数与温度呈负相关，与湿度呈正相关。

参考文献

[1] 邓聚龙. 灰色系统理论教程 [M]. 武汉: 华中理工大学出版社, 1990.

[2] 杨亚柳. 基于 ARIMA 模型和 ARDL 模型对 COVID-19 疫情的研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2021.

[3] 王旭艳, 喻勇, 胡樱, 宇传华. 基于指数平滑模型的湖北省新冠肺炎疫情预测分析[J]. 公共卫生与预防医学, 2020, 31(1): 1-4.

[4] 薛明劲, 黄钊慰, 胡雨迪, 杜进林, 黄志刚. 传染病动力学模型研究进展[J]. 预防医学, 2022, 34(1): 53-57.

[5] 胡朝晖. 美国人口数据的阻滞增长模型拟合分析[J]. 科技广场, 2007(4): 15-17.

[6] 马知恩. 种群生态学数学建模与研究[M]. 合肥: 安徽教育出版社, 1996.

[7] 姜启源. 数学模型[M]. 第 3 版. 北京: 高等教育出版社, 2003.

[8] 薛薇. 统计分析与 SPSS 的应用[M]. 北京: 中国人民大学出版社, 2014.

[9] Lee Rodgers, J. and Nicewander, W.A. (1988) Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, **42**, 59-66. https://doi.org/10.2307/2685263