

应用拟合对早期新冠数据进行分析

周川博, 方娇红

中国地质大学(武汉), 数学与物理学院, 湖北 武汉

收稿日期: 2023年9月6日; 录用日期: 2023年10月1日; 发布日期: 2023年10月8日

摘要

传染病(如之前的COVID-19大流行病)对全球公共卫生构成重大威胁, 死亡率是衡量潜在治疗和干预措施有效性的关键指标。由于发现时间有限, 对新出现的流行病的研究也有限, 所以使用现有数据对死亡率和康复率进行分析, 对研究流行病的早期发病规律、发展趋势具有重要现实意义。我们运用了多项式拟合函数对三个地区(中国, 湖北, 武汉)进行了函数拟合, 并且给出了相关的函数曲线。通过对拟合结果和现实情况进行对比分析得出结论, 在中国大陆COVID-19大流行的早期阶段, 疾病严重程度与医疗资源可用性有较强相关性。

关键词

多项式拟合, COVID-19, 多项式分布, 死亡率

Application of Matching to Analyze Early COVID-19 Data

Chuanbo Zhou, Jiaohong Fang

School of Mathematics and Physics, China University of Geosciences (Wuhan), Wuhan Hubei

Received: Sep. 6th, 2023; accepted: Oct. 1st, 2023; published: Oct. 8th, 2023

Abstract

Infectious diseases (such as the previous COVID-19 pandemic) pose a major threat to global public health. Mortality is a key indicator to measure the effectiveness of latent treatments and interventions. Due to the limited time of discovery and the limited research on emerging epidemics, the use of existing data to analyze mortality and recovery rates is of great practical significance for discovering the early onset pattern and development trend of epidemics. We used the polynomial match function to perform function matching in three regions (China, Hubei, Wuhan), and gave the relevant function curves. By comparing the match results and the actual situation, it is concluded

that in the early stage of the COVID-19 pandemic in Chinese mainland, there is a certain correlation between disease severity and the availability of medical resources.

Keywords

Polynomial Matching, COVID-19, Polynomial Distribution, Mortality

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

Jones K. E [1]提出近几十年来, 全球新发传染病的发病率不断上升, 已成为全球公共卫生面临的最大威胁之一。根据世界卫生组织统计[2], 正在发生的冠状病毒大流行(COVID-19), 于 2019 年 12 月在中国武汉市大规模爆发, 影响到全球 217 个国家和地区, 到 2020 年底, 约 7900 万病例中的死亡人数将超过 170 万。COVID-19 危机已成为公共卫生紧急事件, 严重扰乱了我们生活、经济和社会的方方面面, 对于这种由新型病原体引起的致命传染病, 其死亡率被世界卫生组织用作衡量疾病严重程度的指标。

死亡率是衡量传染病毒性的最基本流行病学指标, 通常定义为所有确诊病例中的死亡比例, 是评估传染病严重程度和危害程度的重要指标之一。在目前的 COVID-19 大流行中, 卫生当局采用了死亡率作为严重性指标, 通过比较不同时间段或不同地区的死亡率, 可以评估应对策略的效果, 指导后续的防控措施。同理, 通过比较不同疾病的死亡率, 可以更好地了解不同疾病的危害程度, 指导公共卫生决策。

康复率是评估传染病治愈情况和防控措施效果的重要指标, 通常定义为感染者康复的比率, 反映疾病的治愈情况。康复率是评估疾病治愈情况的重要指标之一, 通过比较不同时间段或不同地区的康复率, 可以评估应对策略的效果, 指导后续的防控措施。同理, 通过比较不同疾病的康复率, 可以更好地了解不同疾病的危害程度, 帮助评估疾病的治愈情况和应对策略的有效性, 指导公共卫生决策, 提升公众信心。

Zhou Chuanbo [3]等人提出了用置换检验对实时死亡率进行分析, 但置换检验灵活度低, 实验过程更加繁琐, 可能会导致检验结果不准确。根据高尚[4]论文中提出的观点, 我们认识到多项式拟合具有灵活性高, 精确度高的优点, 所以我们采用多项式拟合对死亡率和康复率进行分析, 本文将在数据模拟部分比较不同地区的死亡率和康复率, 以证明政府部门在突发公共卫生事件期间实施的干预措施和医疗资源援助的有效性。

2. 基本原理

2.1. 统计量

我们将观察周期设置为 $[0, \tau]$, 希望在时间 τ 能做出可靠的决策。时间 0 可以设定为对特定群体实施某种干预、治疗或应对策略的那一天。我们将 $[0, \tau]$ 分为 H 个固定的时间间隔(自然以天或周为单位), 并在 h 区间结束时按顺序收集两个亚组的住院患者、死亡和康复人数的信息, $h=1, \dots, H$ 。分别用 $n_{k,D}(h)$ 和 $n_{k,R}(h)$ 表示 h 区间 k 组($k=1, 2$)的死亡和康复人数, 我们进一步用 $N_{k,D}(h)$ 和 $N_{k,R}(h)$ 分别表示 h 区间结束时 k 组的死亡和康复累计人数, 其中 $N_{k,D}(h) = \sum_{j=1}^h n_{k,D}(j)$, $N_{k,R}(h) = \sum_{j=1}^h n_{k,R}(j)$ 。设 $I_k(h-1)$ 是 k 组 h 区间开始前的住院人数, 我们假设在 h 区间, 每个 $I_k(h-1)$ 住院患者要么死亡, 要么康复, 要么继

续住院, 其概率分别为 $p_{k,D}(h)$, $p_{k,R}(h)$ 和 $1 - p_{k,D}(h) - p_{k,R}(h)$ 。以过去的信息为条件, 我们有

$$(n_{k,D}(h), n_{k,R}(h) | I_k(h-1)) \sim \text{Multinomial}(I_k(h-1); p_{k,D}(h), p_{k,R}(h)).$$

让 $\mathcal{F}_{k,h} = \{I_k(j), n_{k,D}(j), n_{k,R}(j), N_{k,D}(j-1), N_{k,R}(j-1), j \leq h\}$ 是由观测数据生成的滤波或历史, 它满足通常的正则性条件。Yip P. S. F [5] 和 Andersen P. K [6] 对于 k 的离散 RTFR 的 h 的离散 RTFR 定义为

$$\pi_k(h) = \frac{p_{k,D}(h)}{p_{k,D}(h) + p_{k,R}(h)}.$$

可视为以死亡或康复事件为条件的死亡概率, 的最大似然估计值(MLE), $\pi_k(h)$ 的最大似然估计值(MLE)可以很容易地证明为

$$\hat{\pi}_k(h) = \frac{\hat{p}_{k,D}(h)}{\hat{p}_{k,D}(h) + \hat{p}_{k,R}(h)} = \frac{n_{k,D}(h)}{n_{k,D}(h) + n_{k,R}(h)}.$$

2.2. 多项式回归模型

在多项式回归是一种回归分析方法, 它使用多项式函数来拟合数据。多项式函数是一种形式为 $y = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$ 的函数, 其中 x 是自变量, y 是因变量, $a_0, a_1, \dots, a_n$ 是多项式系数, n 是多项式的次数。多项式回归的基本思想是, 通过增加多项式的次数, 可以更好地拟合数据, 从而提高回归模型的预测精度。多项式回归可以使用最小二乘法来确定多项式系数。最小二乘法是一种常用的回归分析方法, 它通过最小化残差平方和来确定回归系数, 从而使得回归模型的预测误差最小化。假设我们有一个数据集 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, 我们要用一个三次多项式函数来拟合这个数据集, 即 $y = a_0 + a_1x + a_2x^2 + a_3x^3$ 。我们的目标是找到一组系数 a_0, a_1, a_2, a_3 , 使得拟合函数与数据集的残差平方和最小。

我们定义残差[7]为 $e_i = y_i - f(x_i)$, 其中 $f(x_i)$ 是拟合函数在 x_i 处的取值。残差平方和为 $\sum_{i=1}^n e_i^2$ 我们的目标是最小化 S

我们可以通过对 S 求导数, 令导数为零, 来求解系数 a_0, a_1, a_2, a_3 。具体来说, 我们需要求解下面的

$$\text{方程组: } \left(\frac{\partial S}{\partial a_0}, \frac{\partial S}{\partial a_1}, \frac{\partial S}{\partial a_2}, \frac{\partial S}{\partial a_3} \right) = 0$$

$$\text{对于 } \frac{\partial S}{\partial a_0} = 0 \text{ 我们有: } \frac{\partial S}{\partial a_0} = -2 \sum_{i=1}^n (y_i - a_0 - a_1x_i - a_2x_i^2 - a_3x_i^3) = 0.$$

$$\text{化简得到: } \sum_{i=1}^n y_i - na_0 - a_1 \sum_{i=1}^n x_i - a_2 \sum_{i=1}^n x_i^2 - a_3 \sum_{i=1}^n x_i^3 = 0.$$

$$\text{对于 } \frac{\partial S}{\partial a_1} = 0, \text{ 我们有: } \frac{\partial S}{\partial a_1} = -2 \sum_{i=1}^n x_i (y_i - a_0 - a_1x_i - a_2x_i^2 - a_3x_i^3) = 0.$$

$$\text{化简得到 } \sum_{i=1}^n x_i y_i - a_0 \sum_{i=1}^n x_i - a_1 \sum_{i=1}^n x_i^2 - a_2 \sum_{i=1}^n x_i^3 - a_3 \sum_{i=1}^n x_i^4 = 0.$$

$$\text{对于 } \frac{\partial S}{\partial a_2} = 0, \text{ 我们有: } \frac{\partial S}{\partial a_2} = -2 \sum_{i=1}^n x_i^2 (y_i - a_0 - a_1x_i - a_2x_i^2 - a_3x_i^3) = 0.$$

$$\text{化简得到: } \sum_{i=1}^n x_i^2 y_i - a_0 \sum_{i=1}^n x_i^2 - a_1 \sum_{i=1}^n x_i^3 - a_2 \sum_{i=1}^n x_i^4 - a_3 \sum_{i=1}^n x_i^5 = 0.$$

$$\text{对于 } \frac{\partial S}{\partial a_3} = 0, \text{ 我们有: } \frac{\partial S}{\partial a_3} = -2 \sum_{i=1}^n x_i^3 (y_i - a_0 - a_1x_i - a_2x_i^2 - a_3x_i^3) = 0.$$

$$\text{化简得到: } \sum_{i=1}^n x_i^3 y_i - a_0 \sum_{i=1}^n x_i^3 - a_1 \sum_{i=1}^n x_i^4 - a_2 \sum_{i=1}^n x_i^5 - a_3 \sum_{i=1}^n x_i^6 = 0.$$

通过解这个方程组, 我们可以得到系数 a_0, a_1, a_2, a_3 的解析表达式, 这些系数可以用来构建三次多项

式回归模型, 从而对数据集进行拟合。

根据上述方程组, 我们可以得到系数 a_0, a_1, a_2, a_3 的解析表达式。具体来说, 我们可以使用矩阵运算来求解系数, 如下所示:

$$\text{令 } X \text{ 为 } n \times 4 \text{ 的矩阵, 其中第 } i \text{ 行为 } (1, x_i, x_i^2, x_i^3), \text{ 即: } X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix}。$$

$$\text{令 } Y \text{ 为 } n \times 1 \text{ 的矩阵, 其中第 } i \text{ 行为 } y_i, \text{ 即: } Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}。$$

$$\text{则系数 } a_0, a_1, a_2, a_3, \text{ 可以通过以下公式求解: } \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = X^{-1}Y。$$

3. 数据模拟

3.1. 部分数据来源

根据中华人民共和国国家卫生健康委员会报道我们选取了武汉市、湖北省、中国 2020 年 3 月份部分时间的 COVID-19 每日死亡率和康复率数据为例, 见表 1~3。

Table 1. Some data of Wuhan in March

表 1. 武汉市 3 月份部分数据

地区	日期	死亡率	康复率	备注
武汉	2020.3.01	0.04516	0.429585319	
武汉	2020.3.02	0.04554	0.465969328	
武汉	2020.3.03	0.04606	0.502422285	
武汉	2020.3.04	0.04641	0.529806124	
武汉				

Table 2. Some data of Hubei Province in March

表 2. 湖北省 3 月份部分数据

地区	日期	死亡率	康复率	备注
湖北省	2020.3.01	0.04177	0.503062456	
湖北省	2020.3.02	0.04216	0.538063288	
湖北省	2020.3.03	0.04264	0.5726252	
湖北省	2020.3.04	0.04301	0.599991107	
湖北省				

Table 3. Some data from China in March**表 3.** 中国 3 月份部分数据

地区	日期	死亡率	康复率	备注
中国	2020.3.01	0.03639	0.55563192	
中国	2020.3.02	0.03672	0.588938379	
中国	2020.3.03	0.03714	0.621103775	
中国	2020.3.04	0.03746	0.647328035	
中国				

3.2. 武汉市数据模拟

我们根据 2020 年 3 月份的 COVID-19 数据计算出了每日死亡率和康复率, 运用三次多项式拟合给出了多项式拟合函数, 最后用 matlab 画出了拟合函数图像。

$$X = \begin{bmatrix} 31 & 496 & 10416 & 246016 \\ 496 & 10416 & 246016 & 6197520 \\ 10416 & 246016 & 6197520 & 16261657 \\ 246016 & 6197520 & 16261657 & 4388434896 \end{bmatrix}, Y = \begin{bmatrix} 23.0421 \\ 408.5504 \\ 8947.1380 \\ 215823.7912 \end{bmatrix}, \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0.4046 \\ 0.0313 \\ -0.00051 \\ 0.0000013 \end{bmatrix}$$

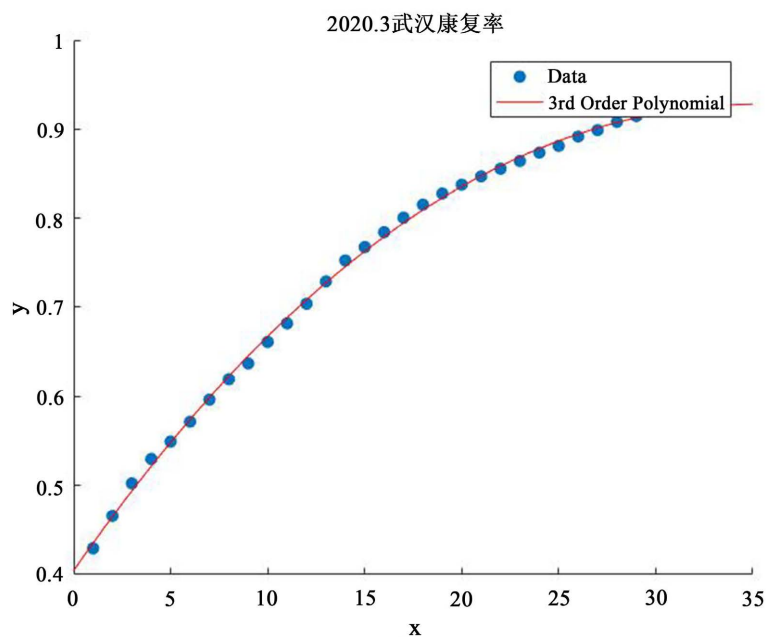
**Figure 1.** Recovery rate of COVID-19 patients in Wuhan**图 1.** 武汉市 COVID-19 患者康复率

图 1 是康复率拟合函数图像, 很明显地看出康复率以极快的速度上升, 并且到后期已经有趋近于 1 的形式。武汉疫情早期康复率较高, 结合实际情况, 我们做出以下推断, 疫情初期, 武汉市政府采取了及时、果断的措施, 对疫情进行了快速、有效的应对。例如, 对疫情严重的区域进行封锁、隔离, 对病患进行集中治疗等。这些措施[8]有助于控制疫情的传播, 提高患者的康复率。武汉市政府在疫情初期就建立了多家方舱医院, 为轻症患者提供了及时、有效的治疗, 这些方舱医院具有较好的隔离条件和

医疗设施,可以有效地减轻医院的压力,提高患者的康复率。同时国家从全国各地调集医疗团队、医疗资源,为武汉提供了大量的医疗支援,这些医疗队具有丰富的临床经验和专业知识,可以为患者提供及时、有效的治疗,提高患者的康复率。另外武汉市政府在疫情初期就开展了大规模的核酸检测,对疫情进行了全面、精准的监测,这有助于及时发现病例,对患者进行隔离和治疗,提高患者的康复率。

$$X = \begin{bmatrix} 31 & 496 & 10416 & 246016 \\ 496 & 10416 & 246016 & 6197520 \\ 10416 & 246016 & 6197520 & 16261657 \\ 246016 & 6197520 & 16261657 & 4388434896 \end{bmatrix}, Y = \begin{bmatrix} 1.5203 \\ 24.7810 \\ 524.3872 \\ 12433.14487 \end{bmatrix}, \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0.0447 \\ 0.0004955 \\ -0.0001444 \\ 0.000000165 \end{bmatrix}$$

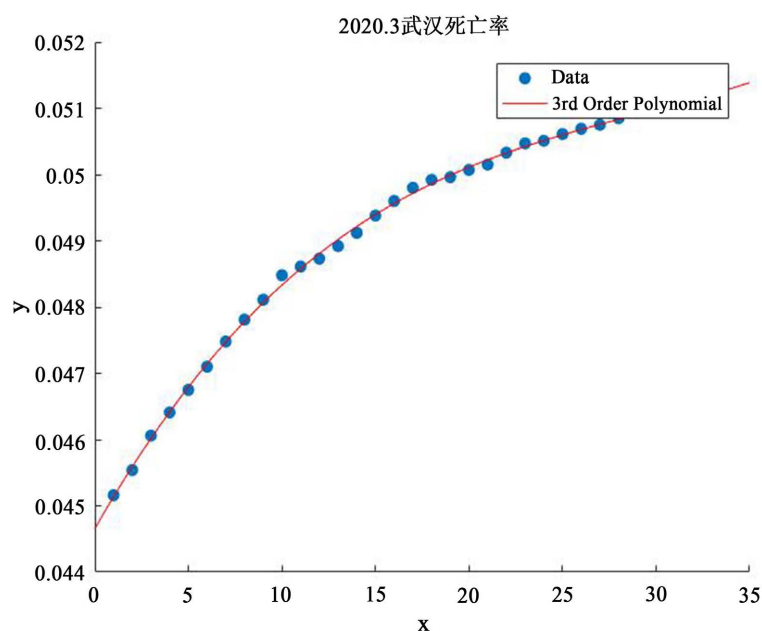


Figure 2. Mortality rate of COVID-19 patients in Wuhan
图 2. 武汉市 COVID-19 患者死亡率

图 2 为死亡率拟合函数图像,可以看出死亡率上升并不迅速,并且在后期已经有了平稳的趋势,说明死亡人数渐渐得到了控制。但早期死亡率还是比较高,结合实际情况,我们做出以下推断。疫情初期,对新冠病毒的认识不足,医疗资源紧张,医疗队伍不足等因素导致了医疗水平的不足。这些因素导致了患者的治疗难度增加,病情恶化,死亡率增加。疫情初期,武汉市政府面临着巨大的压力和挑战,对疫情的应对措施不够及时、有效。例如,疫情初期缺乏足够的医疗物资和设备,导致医疗资源的分配不均,一些患者无法及时得到救治,死亡率增加。疫情初期,新冠病毒的传染性较强,病情进展较快,对患者的身体状况和免疫系统的影响较大。这些因素导致了患者的病情恶化,死亡率增加。疫情初期,武汉市政府面临着信息不对称、谣言满天飞等问题,导致了公众的恐慌和不信任,一些患者不愿意前往医院就诊,错过了最佳的治疗时机,死亡率增加。

3.3. 湖北省和中国数据模拟

根据湖北省和中国 2020 年 3 月份的 COVID-19 数据计算出了每日死亡率和康复率,运用三次多项式拟合给出了多项式拟合函数,最后用 matlab 画出了拟合函数图像。

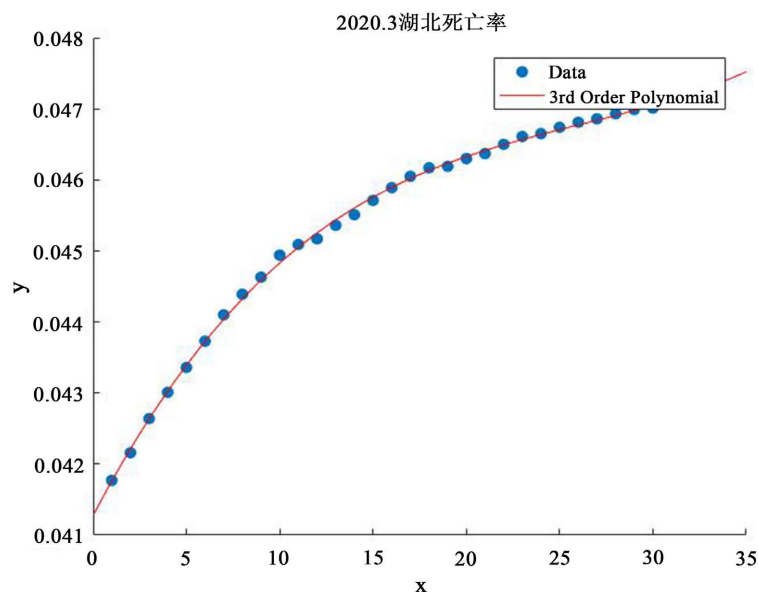


Figure 3. Mortality rate of COVID-19 patients in Hubei Province

图 3. 湖北省 COVID-19 患者死亡率

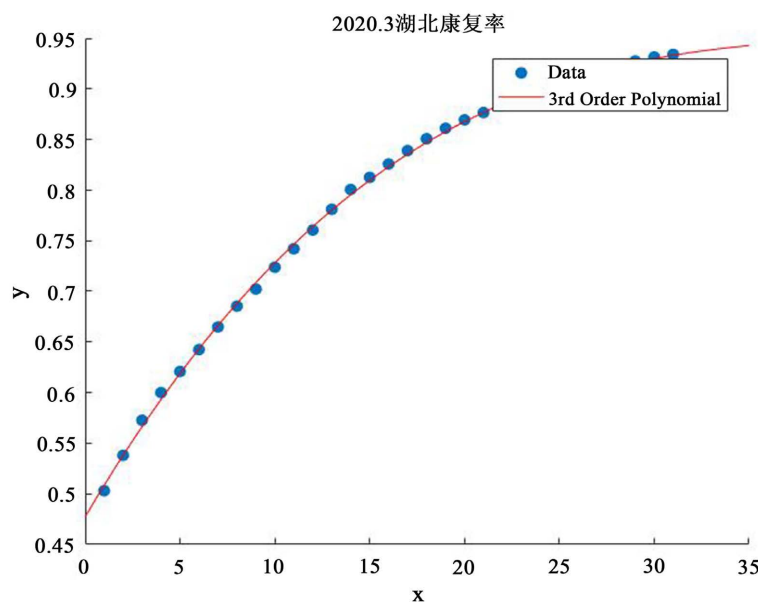


Figure 4. Recovery rate of COVID-19 patients in Hubei Province

图 4. 湖北省 COVID-19 患者康复率

通过拟合图像可以看出, 见图 3~6, 湖北省和全国范围 2020 年 3 月份的康复率和死亡率曲线基本一致, 随着时间推移走势趋于平缓。

3.4. 数据模拟分析

本文采用三次多项式拟合进行分析, 对比其他拟合方法有灵活性高、精度高等优点, 对比上述三个统计区域康复率和死亡率两项指标的拟合图像, 总体看趋势基本一致。但两项指标分别详细对比可以发现如下规律。

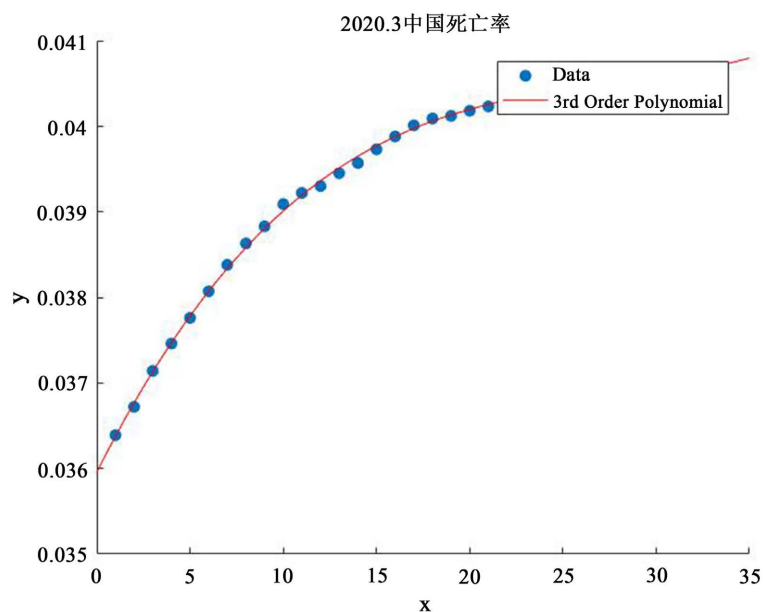


Figure 5. Mortality rate of COVID-19 patients in China

图 5. 中国 COVID-19 患者死亡率

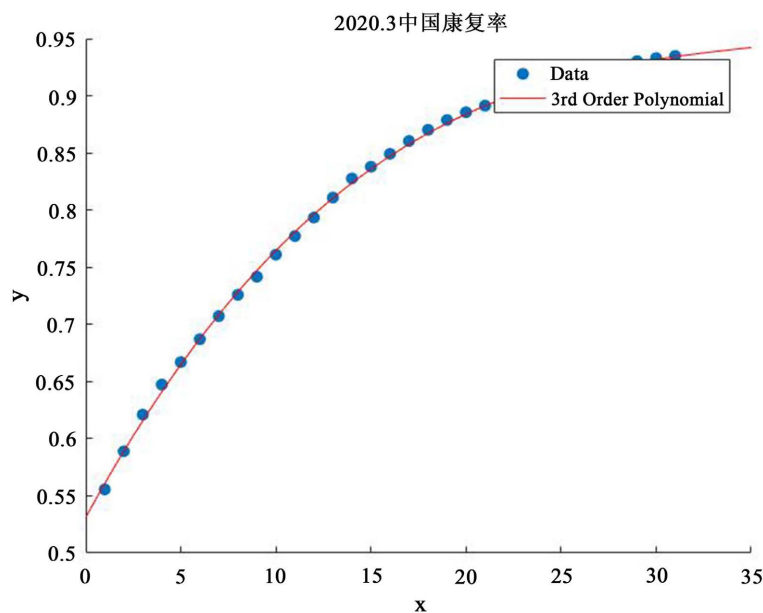


Figure 6. Recovery rate of COVID-19 patients in China

图 6. 中国 COVID-19 患者康复率

趋于稳定的康复率对比, 全国(0.095) $\approx$ 湖北(0.095) > 武汉(0.092), 抛开其他因素可以推测, 康复率与医疗资源的充裕程度有较强的相关性。原因是武汉病例集中, 医疗资源供给不足, 甚至对医疗资源形成挤兑, 导致患者没有得到及时的救治。

趋于稳定的死亡率对比, 全国(0.040) < 湖北(0.047) < 武汉(0.051), 死亡率与医疗资源的充裕程度也有较强的相关性, 湖北省及全国病例密度低, 医疗资源充足相对充足, 病人的到良好的救治, 死亡率自然就低。

总体来说, 通过数据分析与实际情况对比, 印证了医疗资源与上述两个统计指标的相关性。政府对医疗资源的投入、防控措施的实施等对疫情的早期干预, 使得新冠疫情得到快速控制, 同时避免了病毒在其他区域大规模扩散。

4. 拟合效果评估

4.1. 多项式拟合优点

- 1) 灵活性高: 多项式拟合可以适用于各种形状的数据, 可以拟合出曲线、折线、直线等各种形状的数据。
- 2) 精度高: 多项式拟合可以通过增加多项式的次数来提高拟合的精度, 可以拟合出非常复杂的数据。
- 3) 易于实现: 多项式拟合是一种简单易行的方法, 只需要进行简单的数学计算即可得到拟合结果。
- 4) 可解释性强: 多项式拟合的结果可以通过多项式的系数来解释, 可以帮助我们更好地理解数据的特征和规律。
- 5) 可扩展性强: 多项式拟合可以通过增加多项式的次数来适应更复杂的数据, 可以在一定程度上避免过拟合和欠拟合的问题。

4.2. 多项式拟合缺点

- 1) 容易受到噪声的影响: 多项式拟合对噪声比较敏感, 如果数据中存在噪声, 可能会导致拟合结果不准确。
- 2) 容易出现过拟合和欠拟合: 多项式拟合需要选择合适的多项式次数来避免过拟合和欠拟合的问题。如果选择的多项式次数过高, 可能会导致过拟合, 如果选择的多项式次数过低, 可能会导致欠拟合。
- 3) 计算复杂度较高: 多项式拟合需要进行复杂的数学计算, 计算复杂度较高, 对计算资源的要求较高。

5. 结论

本文运用多项式拟合的方法对三个地区的数据进行分析, 在新冠疫情的早期, 患者的死亡率和康复率都呈现出一个快速上升的趋势, 很快又逐渐趋于稳定, 说明政府对疫情的干预极其有效。也就意味着在 COVID-19 大流行的早期阶段, 疾病严重程度与医疗资源可用性有较强相关性, 并且对比其他研究方法, 本文多项式拟合适用性极高。

但在疾病爆发期间, 有些部门或者地区的监测数据不完整, 某些数据如性别、年龄段等难以获取, 尤其是在那些防护意识低下、医疗保健体系不全的地区更是如此。尽管目前已经进入后疫情时期, 有关新型冠状病毒肺炎的详细数据还是很难获得, 大多数地区只记录了每日新增的确诊病例数、死亡数和治愈数及其汇总数据。相关数据收集不足不利于系统分析, 但对于卫生部门来说, 要尽可能使用多项式拟合方法, 充分利用这些简单的数据进行多维度分析, 深入了解传染性疾病的传播、预防、治疗等特性, 以便迅速采取针对性的干预措施, 尽早抑制疾病的传播并降低死亡率。

致 谢

由衷地感谢中国地质大学(武汉)中央高校基本科研基金资助项目的资助。

参考文献

- [1] Jones, K.E., Patel, N.G., Levy, M.A., *et al.* (2008) Global Trends in Emerging Infectious Diseases. *Nature*, **451**, 990-993. <https://doi.org/10.1038/nature06536>

-
- [2] World Health Organization (WHO) (2020) Coronavirus Disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>
 - [3] Zhou, C.B., Fang, J.H. and Mao, M.Z. (2023) A New Analysis of Real-Time Fatality Rate in the Initial Stage of COVID-19. *Entropy*, **25**, Article 1028. <https://doi.org/10.3390/e25071028>
 - [4] 高尚. 多项式拟合的推广[J]. 大学数学, 2006, 22(3): 83-86.
 - [5] Yip, P.S.F., Lau, E.H.Y., Lam, K.F., *et al.* (2005) A Chain Multinomial Model for Estimating the Real-Time Fatality Rate of a Disease, with an Application to Severe Acute Respiratory Syndrome. *American Journal of Epidemiology*, **161**, 700-706. <https://doi.org/10.1093/aje/kwi088>
 - [6] Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (2012) *Statistical Models Based on Counting Processes*. Springer Science & Business Media, New York.
 - [7] 胡菊华. 基于残差分析的线性回归模型的诊断与修正[J]. 统计与决策, 2019, 35(24): 5-8. <https://doi.org/10.13546/j.cnki.tjyjc.2019.24.001>
 - [8] Chen, T., Wang, Y. and Hua, L. (2020) “Pairing Assistance”: The Effective Way to Solve the Breakdown of Health Services System Caused by COVID-19 Pandemic. *International Journal for Equity in Health*, **19**, Article No. 68. <https://doi.org/10.1186/s12939-020-01190-8>