Modeling and Simulation 建模与仿真, 2023, 12(6), 5156-5164
Published Online November 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.126469

Hans汉斯

基于光滑样条回归的软件可靠性模型

仇正霞[1]，杨剑锋[1,2*]，胡文生[3]，黄嘉悦[1]

[1]贵州大学，数学与统计学院，贵州 贵阳
[2]贵州理工学院，大数据学院，贵州 贵阳
[3]贵州交通职业技术学院，信息工程系，贵州 贵阳

收稿日期：2023年9月18日；录用日期：2023年11月2日；发布日期：2023年11月9日

摘　要

随着软件产品在各行各业的广泛使用，其高可靠、高安全成为衡量一个软件质量的重要属性。本文引入了一种基于光滑样条回归的软件可靠性模型，并将其与传统的软件可靠性模型进行比较。此外，使用了最小二乘估计方法来估计模型中的参数。最后，基于开源软件Tomcat3-11服务器的真实失效数据，利用R软件对这4类可靠性模型进行性能对比分析，结果表明光滑样条回归模型的拟合与预测效果较好。

关键词

光滑样条回归，软件可靠性模型，开源软件，失效数据

Software Reliability Model Based on Smooth Splines Regression

Zhengxia Qiu[1], Jianfeng Yang[1,2*], Wensheng Hu[3], Jiayue Huang[1]

[1]School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou
[2]School of Data Science, Guizhou Institute of Technology, Guiyang Guizhou
[3]Department of Information Engineering, Guizhou Communications Polytechnic, Guiyang Guizhou

Received: Sep. 18th, 2023; accepted: Nov. 2nd, 2023; published: Nov. 9th, 2023

Abstract

With the widespread use of software products in various industries, their high reliability and se-

*通讯作者。

文章引用：仇正霞，杨剑锋，胡文生，黄嘉悦．基于光滑样条回归的软件可靠性模型[J]．建模与仿真，2023，12(6): 5156-5164. DOI: 10.12677/mos.2023.126469

curity have become important attributes for assessing software quality. This paper introduces a software reliability model based on smooth spline regression and compares it with traditional software reliability models. In addition, the least squares estimation method is used to estimate the parameters in the model. Finally, real failure data from the open-source Tomcat 3-11 server is used to perform a performance comparison analysis of these four types of reliability models using the R software. The results show that the smooth spline regression model provides better fitting and predictive performance.

Keywords

Smooth Spline Regression, Software Reliability Model, Open-Source Software, Failure Data

1. 引言

近年来，计算机软件已经渗透到几乎所有领域，成为现代社会的重要组成部分，深刻地改变了人们的生活方式、工作方式和社会互动。但是，一旦计算机软件发生故障将会导致严重后果，从数据丢失和业务中断到安全泄露和法律责任。例如：2020 年 8 月 9 日，美国国土安全部的一些系统经历了持续数小时的故障。这次故障影响了海关和边境保护局的运营，造成长时间的边境延误。2021 年 3 月 2 日，微软 Exchange 服务器发生了漏洞，导致数千个组织的电子邮件数据泄露和被攻击。这些软件失效案例不胜枚举，因此，研究软件可靠性保障软件系统正常运行、提高软件质量是现代软件开发和维护过程中不可或缺的一部分。

软件可靠性增长模型(software reliability and growth model, SRGM) [1] [2] [3]是软件工程领域的关键工具，用于建立数学模型描述和预测软件系统可靠性的变化趋势。此类模型在软件可靠性的评测、保证、测试资源优化和发布策略[4] [5]研究中具有重要作用。当前最常用的参数模型是非齐次泊松过程(non-homogeneous Poisson process, NHPP) [6] [7]类软件可靠性增长模型。最先提出的 NHPP 类模型是 G-O 模型[8]，其后改进得到了 Delayed S-shaped 和 Inflection S-shaped 等经典模型。建立 NHPP 类软件可靠性增长模型首先需要提出基本假设，建立数学模型，然后基于具体数据利用最小二乘估计或最大似然估计求解模型中的参数，得到具体的拟合函数。传统的参数化软件可靠性模型存在许多缺点，这些缺点与其不切实际的假设、环境依赖的适用性和可疑的预测性直接相关。

非参数模型无需预先假设函数的具体形式，而是基于数据集进行估计。与参数化软件可靠性增长模型相比，非参数模型有着更强的适应性且拟合精度较高。Hao [9]等引入一种对数风险函数的惩罚非参数最大似然估计，用于分析右删失数据，并使用光滑样条平滑估计。Yu [10]等通过建立分类回归模型来表征交通流与不同时间点的关系，并通过具有光滑样条的负二项式模型识别不同的交通流模式。Suk [11]等首先提出分段线性模型，其中数据的时域被划分为连续的阶段，并且在每个阶段拟合一条单独的线性回归线，其中惩罚样条模型通过引入惩罚项来实现拟合度和平滑度之间的平衡。Liu [12]等提出了一种贝叶斯时变系数模型来评估多类型循环事件的强度的时间分布，该模型使用贝叶斯惩罚样条获得时变系数和基线强度的平滑估计值。Dohi [13]等考虑了数据驱动的软件可靠性评估方法，可以在不完全的故障计数分布知识下为软件可靠性预测提供有用的概率信息。Choudhary [14]等进行了软件可靠性预测建模：参

数化和非参数化建模的比较，评估并比较了 2 个参数和 2 个非参数软件可靠性增长模型在 3 个真实数据集上软件故障的准确性。Dharmasena [15]等采用基于具有核平滑的局部多项式建模的非参数方法来进行 SRGM 建模，并提供数值示例对模型进行评估比较。

综上所述，关于软件可靠性模型的研究主要分为参数模型和非参数模型。本文利用非参数模型中的光滑样条回归模型对开源软件累计故障数进行研究。其余部分主要内容如下：第二节主要介绍了基于光滑样条回归的软件可靠性模型；第三节针对真实数据集进行案例分析；第四节根据分析结果得出结论。

2. 基于光滑样条回归的软件可靠性模型

2.1. 非参数回归模型

非参数回归[16] [17] [18]是统计学研究的一个热门方向，有着广泛的应用前景，受到国内外学者的广泛关注。由于失效数据没有固定的分布以及可靠性模型没有具体的函数形式，因此本文选择非参数回归模型进行累计失效数的拟合和预测。

给定一组累计失效数$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，若想研究软件失效数 Y 与时间变量 X 之间的关系，可将其表示为非参数回归模型的形式：

$$y_i = f(x_i)+\varepsilon_i, i=1,2,\cdots,n \tag{1}$$

其中 $f(\bullet)=E(y\mid x)$ 为未知光滑函数，假定随机误差 $\varepsilon_i \sim N(0,\sigma^2)$。

令 $y=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$，$f(x)=(f(x_1),f(x_2),\cdots,f(x_n))^{\mathrm{T}}$，$\varepsilon=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)$，则模型(1)可写成：

$$y=f(x)+\varepsilon, \varepsilon \sim N(0,\sigma^2 I_n) \tag{2}$$

2.2. 光滑样条回归模型

光滑样条回归[19]是一种常见的非参数回归模型，可以用于拟合开源软件以月为单位的累计失效数。由于光滑样条回归软件可靠性模型是由数据驱动模型，因此具有较强的稳健性和适应性。其在软件可靠性回归拟合中的思想是使用样条在每个时间段内拟合一个方程，然后在时间点处将这些分段曲线光滑的连接起来。光滑样条回归的目的是在保持模型光滑性的同时，尽量减小观测数据与模型之间的残差。即需要满足残差平方和最小准则：

$$\min Q(f)=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-f(x_i))^2 \tag{3}$$

其中：x_i 为时间，y_i 为累计失效数，$\hat{y}_i$ 为回归方程的拟合值。由于光滑样条的节点数较多，会导致模型出现过拟合现象，因此需要在(3)式的基础上引入一个惩罚项，故判断准则为：

$$\min Q(f)=\sum_{i=1}^{n}[y_i-f(x_i)]^2+\lambda\int[f''(x)]^2\,\mathrm{d}x \tag{4}$$

其中，λ 是一个非负的调节参数，$\lambda\geq 0$，主要是为了控制回归函数的拟合优度和光滑程度之间的平衡，其值过小会导致过拟合，过大会出现有偏估计。$\lambda\int[f''(x)]^2\,\mathrm{d}x$ 为惩罚项，其目的是为了提高拟合曲线的光滑程度。$f''(x)$ 是 $f(x)$ 的二阶导数，表示拟合函数 $f(x)$ 的弯曲程度，$f''(x)$ 值越大，曲线波动越厉害，曲线也就越粗糙。接下来本文将以自然三次样条为例，展示光滑样条回归模型对累计失效数估计值的显示表达式。

2.3. 自然三次样条

在$[a,b]$上有一组软件失效时间$x_1,x_2,\cdots,x_n$，$a=x_0<x_1<x_2<\cdots<x_n<x_{n+1}<b$，如果$f$满足以下两个条件：

1) f在区间$(x_{i-1},x_i](i=1,2,\cdots,n)$上都是三次多项式；

2) f在时间节点x_i处的一阶、二阶导数都连续，且在$[a,b]$上也连续。

那么称f为三次样条函数。此外，如果满足$f''(a)=f''(b)=f'''(a)=f'''(b)=0$，那么$f$为自然三次样条，即$f$在$[a,x_1],[x_n,b]$是线性的。自然三次样条的形式为：

$$f(x)=d_i(x-x_i)^3+c_i(x-x_i)^2+b_i(x-x_i)+a_i,\ x_i\le x\le x_{i+1} \tag{5}$$

令$f_i=f(x_i)$，$\gamma_i=f''(x_i)$，根据自然三次样条的边界条件可得$\gamma_1=\gamma_n=0$。接下来先引入两个带状矩阵Q、R：

$$Q=(q_{ij})_{n\times(n-2)},\ i=1,2,\cdots,n;\ j=2,3,\cdots,n-1 \tag{6}$$

其中元素$q_{j-1,j}=h_{j-1}^{-1}$，$q_{j,j}=-h_j^{-1}$，$q_{j+1,j}=h_j^{-1}$，$h_i=x_{i+1}-x_i$，当$|i-j|\ge 2$时，$q_{ij}=0$。

$$R=(r_{ij})_{(n-2)\times(n-2)},\ i,j=2,3,\cdots,n-1 \tag{7}$$

其中元素$\gamma_{i,i}=\frac{1}{3}(h_{i-1}+h_i)$，$\gamma_{i,i+1}=\gamma_{i+1,i}=\frac{1}{6}h_i$，其中当$|i-j|\ge 2$时，$\gamma_{ij}=0$。

易知矩阵R为严格正定矩阵，令$K=QR^{-1}Q^{\mathrm{T}}$，由f和γ确定自然三次样条的充要条件是$Q^{\mathrm{T}}f=R\gamma$，则

$$\int_a^b[f''(x)]^2\,\mathrm{d}x=\gamma^{\mathrm{T}}R\gamma=f^{\mathrm{T}}Kf \tag{8}$$

因此，(4)式可表示为矩阵形式为

$$(y-f)^{\mathrm{T}}(y-f)+\lambda f^{\mathrm{T}}Kf \tag{9}$$

对(9)式关于f求导，令其导数为零可使得上式最小化，故得f的估计为

$$\hat{f}=(I+\lambda K)^{-1}y \tag{10}$$

回归函数f对累计故障数的拟合程度可归结于参数λ的取值。

3. 案例分析

为了验证与比较模型的性能，本文基于真实数据集对 GO 模型、DSS 模型、ISS 模型以及光滑样条回归模型进行对比分析。

3.1. 失效数据

Tomcat 服务器是一个由 Apache 软件基金会开发和维护的轻量级、开源的 Web 应用服务器，通常在中小型系统以及访问并发较低的场景中广泛使用。Tomcat 服务器通常与其他组件和工具一起使用，如 Apache HTTP 服务器、数据库、应用程序框架等，以构建完整的 Web 应用程序堆栈。本文所需的故障数据来源于 Tomcat 服务器版本 3-11 的用户缺陷跟踪系统(http://bz.apache.org/bugzilla/)。

Tomcat 的主要数据字段有 Auth、Catalina、Cluster、EL、Jasper 和 Manager 等。本文以月为单位进行数据整合，提取了从 2010 年 1 月到 2023 年 8 月期间检测到的故障数(共 164 组数据)，失效数据见表 1。

Table 1. Tomcat3-11 failure data
表 1. Tomcat3-11 失效数据

时间/月	失效数	时间/月	失效数	时间/月	失效数	时间/月	失效数	时间/月	失效数	时间/月	失效数
1	33	29	18	57	18	85	10	113	3	141	8
2	29	30	25	58	24	86	15	114	3	142	1
3	32	31	20	59	22	87	18	115	12	143	3
4	59	32	23	60	17	88	5	116	14	144	6
5	13	33	10	61	16	89	12	117	12	145	5
6	30	34	20	62	23	90	22	118	22	146	3
7	21	35	17	63	28	91	10	119	15	147	5
8	22	36	15	64	18	92	12	120	15	148	6
9	43	37	27	65	12	93	9	121	6	149	7
10	31	38	14	66	9	94	16	122	8	150	4
11	28	39	33	67	15	95	16	123	16	151	4
12	17	40	16	68	9	96	9	124	12	152	12
13	33	41	22	69	49	97	12	125	8	153	4
14	29	42	24	70	18	98	7	126	19	154	5
15	32	43	58	71	22	99	8	127	4	155	6
16	59	44	23	72	9	100	9	128	3	156	8
17	13	45	12	73	9	101	10	129	11	157	6
18	30	46	16	74	11	102	10	130	9	158	4
19	21	47	17	75	25	103	8	131	10	159	14
20	22	48	16	76	13	104	11	132	7	160	3
21	43	49	35	77	14	105	18	133	4	161	9
22	31	50	13	78	7	106	12	134	6	162	5
23	28	51	22	79	19	107	6	135	6	163	3
24	17	52	29	80	16	108	13	136	8	164	4
25	18	53	20	81	17	109	6	137	9		
26	32	54	23	82	12	110	10	138	17		
27	27	55	21	83	24	111	13	139	6		
28	24	56	16	84	14	112	10	140	3		

3.2. 模型评估准则

为了比较模型的性能，本文选择以下指标进行衡量：MSE，AIC。

1) 均方误差(Mean Squared Error，简称 MSE)是一种常用的评估指标，用于衡量估计值与实际观测值之间的差异，从而量化模型的拟合精度。数学上，MSE 可以表示为：

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \tag{11}$$

2) 赤池信息量准则(AIC) [20]是一种模型选择的统计指标。它用于平衡模型的拟合优度和复杂性，以便选择最适合的模型。在不同模型选择时，优先使用具有较小 AIC 值的模型。AIC 的计算公式为：

$$\mathrm{AIC} = -2\ln(L) + 2k \tag{12}$$

其中：L 是模型的似然函数值，k 是模型中的参数数量，也称模型的复杂度。当误差服从正态分布时，AIC 可以表示为：

$$\mathrm{AIC} = m\ln\left(\frac{\mathrm{RSS}}{m}\right) + 2k \tag{13}$$

其中：m 是样本容量，RSS 是残差平方和。

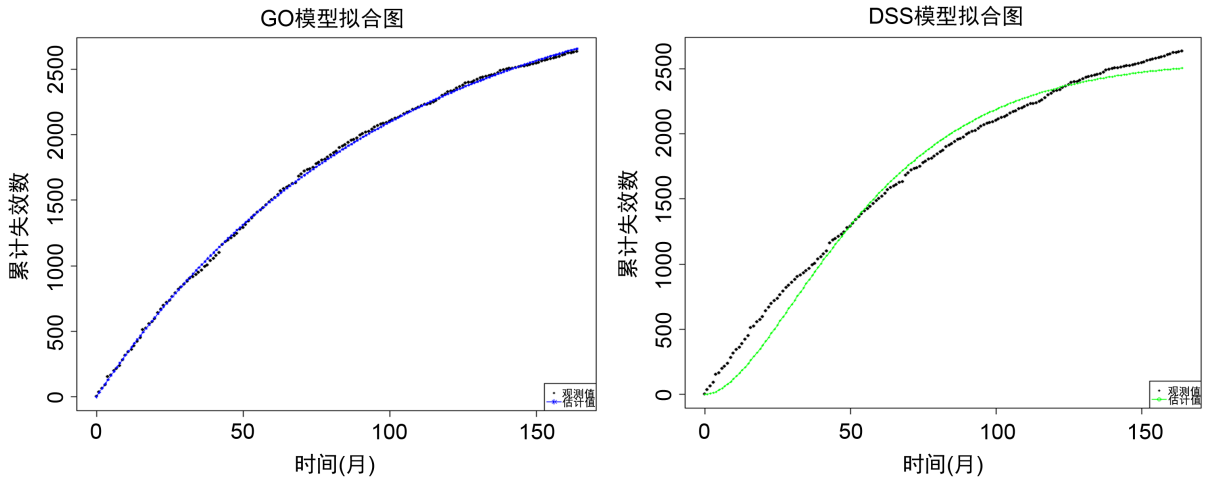
3.3. 模型性能对比分析

本文基于表 1 中 Tomcat 的真实数据集，利用最小二乘估计求解出了 NHPP 类软件可靠性增长模型的参数估计结果，进而得到模型的具体表达式。模型的参数估计结果以及拟合优度结果如表 2 所示。从表中可以看出光滑样条回归模型的 MSE (10.6977)和 AIC (397.0549)这两个拟合度评估指标比其他 3 个可靠性模型的值都要小。从以上评价指标的数值可以看出，对于本文的失效数据，光滑样条回归的拟合效果最好，而 DSS 模型的拟合效果最差。

Table 2. Least squares estimation of parameters and model comparison results
表 2. 参数的最小二乘估计和模型比较结果

模型名称	参数估计结果			MSE	AIC
	a	b	c		
GO 模型	3256.275	0.0103		300.0483	945.1506
DSS 模型	2569.615	0.0338		10637.98	1533.911
ISS 模型	3256.821	0.0103	-0.0039	282.112	936.9802
光滑样条回归				10.6977	397.0549

各模型的失效拟合图如图 1 所示。从图中可以看出，光滑样条回归模型对累计故障数拟合效果较好。



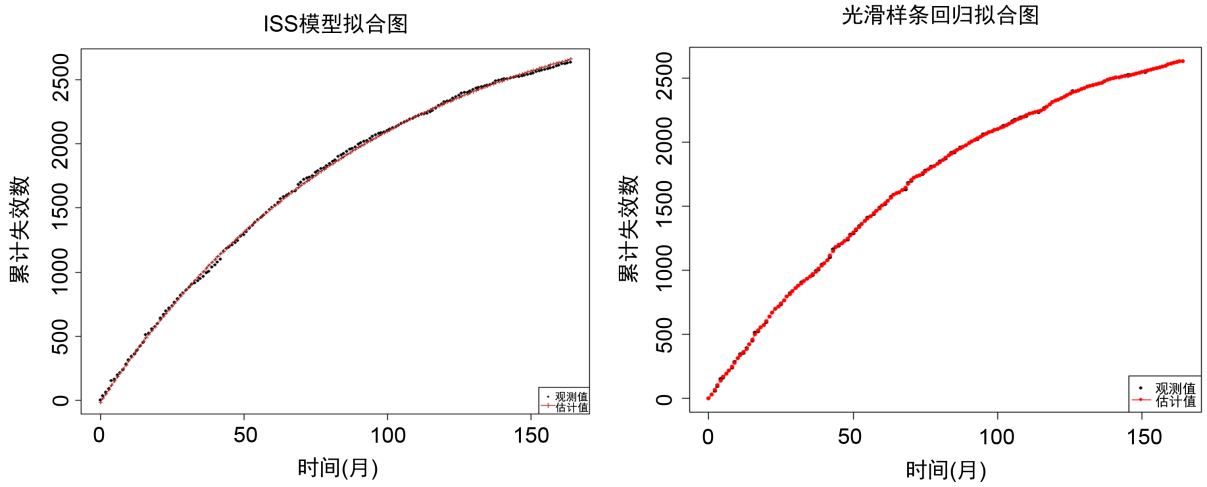


Figure 1. Cumulative failure count fits for each model
图 1. 各模型的累计失效数拟合图

图 2 为四种模型的拟合对比图，整体分析可知，光滑样条回归模型的拟合效果要优于 NHPP 类软件可靠性增长模型。这也说明由于非参数回归不需要前提假设且对数据集进行分段拟合，因此非参数回归相比于参数回归有着更强的稳健性和适应性。

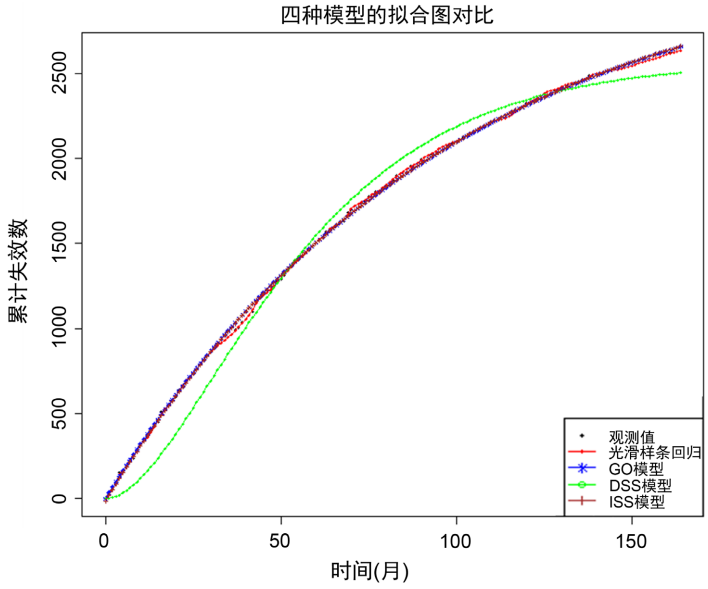


Figure 2. Comparison of fits for four different models
图 2. 四种模型的拟合对比

3.4. 预测性能对比分析

为了分析模型的预测性能，本文绘制了四种模型的相对误差(RE)曲线。其中 RE 曲线越趋近于 0，预测性能越好；大于 0 是正向预测；小于 0 是负向预测。

使用 Tomcat 数据集进行的预测揭示了模型对未来测试性能的描述和对未来累计故障数的检测能力。从图 3 所示的相对误差曲线和数据分析可知：(1) 整体上，除了 DSS 模型出现了一定的预测偏差之外，其余三个模型的预测曲线随着时间的增大都逐渐趋向于 0，即表明预测效果较好；(2) 在测试的初始阶段，

RE 曲线的起伏表明模型正在对数据进行拟合适应，其中光滑样条回归模型的波动程度较小，表明该模型对数据的拟合适应性较强。

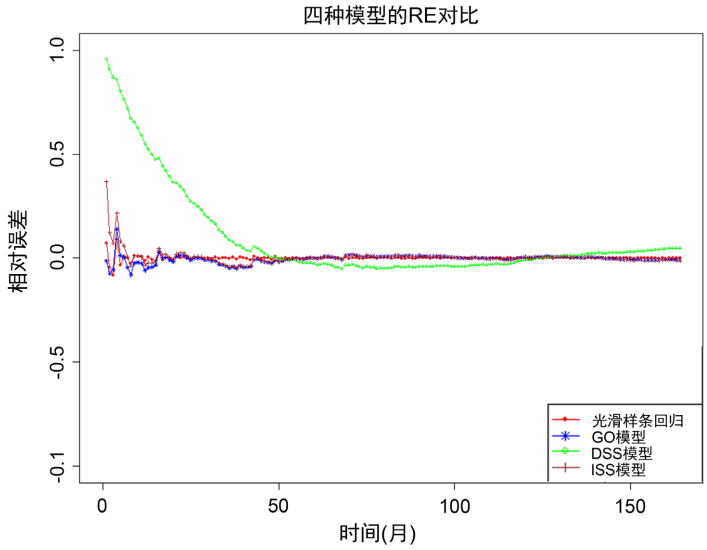


Figure 3. Comparison of relative error curves for four different models
图 3. 四种模型的相对误差曲线对比

4. 结论

在软件可靠性研究中，单位时间内累计发生失效次数对软件可靠性起决定作用。本文以 Tomcat 服务器累计失效数为研究对象，提出一种非参数可靠性模型——光滑样条回归模型，将该模型与传统的 NHPP 类可靠性增长模型进行对比分析。通过绘制拟合图、预测图以及计算评估指标得出具体结论，光滑样条回归软件可靠性模型的拟合和预测效果更好，且对数据具有较强的稳健性和适应性。

基金项目

国家自然科学基金项目(No.72361008),贵州省交通运输厅科技项目(No.2022-321-013)。

参考文献

[1] Luo, H., Xu, L.J., He, L., Jiang, L.D. and Long, T. (2023) A Novel Software Reliability Growth Model Based on Generalized Imperfect Debugging NHPP Framework. *IEEE Access*, **11**, 71573-71593. https://doi.org/10.1109/ACCESS.2023.3292301

[2] Li, Q.Y. and Pham, H. (2019) A Generalized Software Reliability Growth Model with Consideration of the Uncertainty of Operating Environments. *IEEE Access*, **7**, 84253-84267. https://doi.org/10.1109/ACCESS.2019.2924084

[3] Zhu, M.M. and Pham, H. (2018) A Multi-Release Software Reliability Modeling for Open Source Software Incorporating Dependent Fault Detection Process. *Annals of Operations Research*, **269**, 773-790. https://doi.org/10.1007/s10479-017-2556-6

[4] Huang, Y.S., Fang, C.C., Chou, C.H., et al. (2023) A Study on Optimal Release Schedule for Multiversion Software. *Informs Journal on Computing*. https://doi.org/10.1287/ijoc.2021.0141

[5] Xie, M., Li, X. and Ng, S.H. (2011) Risk-Based Software Release Policy under Parameter Uncertainty. *Proceedings of the Institution of Mechanical Engineers Part O*: *Journal of Risk and Reliability*, **225**, 42-49. https://doi.org/10.1177/1748006XJRR286

[6] Liu, X.M. and Xie, N.M. (2022) Grey-Based Approach for Estimating Software Reliability under Nonhomogeneous Poisson Process. *Journal of Systems Engineering and Electronics*, **33**, 360-369. https://doi.org/10.23919/JSEE.2022.000038

[7] Okamura, H. and Dohi, T. (2021) Application of EM Algorithm to NHPP-Based Software Reliability Assessment with Generalized Failure Count Data. *Mathematics*, **9**, Article 985. https://doi.org/10.3390/math9090985

[8] Wang, Y.Z., Liu, H.T., Yuan, H.J. and Zhang, Z.H. (2023) Comprehensive Evaluation of Software System Reliability Based on Component-Based Generalized G-O Models. *PeerJ Computer Science*, **9**, e1247. https://doi.org/10.7717/peerj-cs.1247

[9] Hao, M.L., Lin, Y.Y. and Zhao, X.Q. (2020) Nonparametric Inference for Right-Censored Data Using Smoothing Splines. *Statistica Sinica*, **30**, 153-173.

[10] Yu, Z., Yang, J. and Huang, H.H. (2023) Smoothing Regression and Impact Measures for Accidents of Traffic Flows. *Journal of Applied Statistics*. https://doi.org/10.2139/ssrn.4103425

[11] Suk, H.W., West, S.G., Fine, K.L., *et al*. (2019) Nonlinear Growth Curve Modeling Using Penalized Spline Models: A Gentle Introduction. *Psychological Methods*, **24**, 269-290. https://doi.org/10.1037/met0000193

[12] Liu, Y. and Guo, F. (2020) A Bayesian Time-Varying Coefficient Model for Multitype Recurrent Events. *Journal of Computational and Graphical Statistics*, **29**, 383-395. https://doi.org/10.1080/10618600.2019.1686988

[13] Dohi, T., Zheng, J.J. and Okamura, H. (2020) Data-Driven Software Reliability Evaluation under Incomplete Knowledge on Fault Count Distribution. *Quality Engineering*, **32**, 421-433. https://doi.org/10.1080/08982112.2020.1757705

[14] Choudhary, A., Baghel, A.S. and Sangwan, O.P. (2016) Software Reliability Prediction Modeling: A Comparison of Parametric and Non-Parametric Modeling. 2016 6*th International Conference on Cloud System and Big Data Engineering* (*Confluence*), Noida, 14-15 January 2016, 649-653. https://doi.org/10.1109/CONFLUENCE.2016.7508198

[15] Dharmasena, L.S., Zeephongsekul, P. and Jayasinghe, C.L. (2011) Software Reliability Growth Models Based on Local Polynomial Modeling with Kernel Smoothing. 2011 22*nd IEEE International Symposium on Software Reliability Engineering* (*ISSRE*), Hiroshima, 29 November-2 December 2011, 220-229. https://doi.org/10.1109/ISSRE.2011.10

[16] Ding, J.H. and Zhang, Z.Q. (2021) Statistical Inference on Uncertain Nonparametric Regression Model. *Fuzzy Optimization and Decision Making*, **20**, 451-469. https://doi.org/10.1007/s10700-021-09353-0

[17] Dong, H., Otsu, T. and Taylor, L. (2023) Bandwidth Selection for Nonparametric Regression with Errors-in-Variables. *Econometric Reviews*, **42**, 393-419. https://doi.org/10.1080/07474938.2023.2191105

[18] Li, J.J. and Zhao, L.F. (2020) Hydropower Price Prediction with the Nonparametric Statistics Regression Model. *Journal of Coastal Research*, **104**, 402-405. https://doi.org/10.2112/JCR-SI104-072.1

[19] Xu, L.W. and Zhou, J.B. (2019) A Model-Averaging Approach for Smoothing Spline Regression. *Communications in Statistics-Simulation and Computation*, **48**, 2438-2451. https://doi.org/10.1080/03610918.2018.1457694

[20] Cavanaugh, J.E. and Neath, A.A. (2019) The Akaike Information Criterion: Background, Derivation, Properties, Application, Interpretation, and Refinements. *WIREs Computational Statistics*, **11**, e1460. https://doi.org/10.1002/wics.1460