Software Reliability Growth Model using Interval Domain Data

Geetha Rani Neppala
Associate Professor
Dept. of Computer Science
Abhinav Institute of Management & Technology
Singarayakonda-INDIA

Dr.R.Satya Prasad
Associate Professor
Dept. of Computer Science & Engineering
Acharya Nagarjuna University
Nagarjuna Nagar- INDIA

Prof.R.R.L.Kantam
Professor
Dept. of Statistics
Acharya Nagarjuna University
Nagarjuna Nagar- INDIA

ABSTRACT
Software reliability is one of the most important characteristics of software quality. Its measurement and management technologies employed during the software life cycle are essential for producing and maintaining quality/reliable software systems. Over the last several decades, many Software Reliability Growth Models (SRGMs) have been developed to greatly facilitate engineers and managers in tracking and measuring the growth of reliability as software is being improved. In this paper we proposed Pareto type II based software reliability growth model with interval domain data. The maximum likelihood (ML) estimation approach is used to estimate the unknown parameters of the model. This paper presents estimation procedures to access reliability of a software system using Pareto type II distribution, which is based on Non Homogenous Poisson Process (NHPP). We also present an analysis of two software failure data sets.

General Terms
Software failure data, Distribution function, Mean value function

Keywords
Software Reliability, NHPP, Pareto type II distribution, Parameter estimation, Interval domain data, ML estimation.

1. INTRODUCTION
Software reliability is the most dynamic quality characteristic which can measure and predict the operational quality of the software system during its intended life cycle. Software reliability is the probability of failure free operation of software in a specified environment during specified duration [Musa 1998]. For the past several decades, various statistical models have been proposed to access the software reliability for example Goel and Okumoto (1979), Musa(1980), Pham(2005), Wood(1996), Ramamurthy and Bastani(1982), Satya Prasad(2007) and Satyaprasad and Geetharani(2011). The most common approach to developing software reliability model is the probabilistic approach. The probabilistic model represents the failure occurrences and the fault removals as probabilistic events. They are classified into various groups, including error seeding models, failure rate models, curve fitting models, reliability growth models, Markov structure models and Non Homogenous Poisson Process (NHPP) models.

The NHPP based models are the most important models because of their simplicity, convenience and compatibility. The NHPP based software reliability growth models are proved quite successful in practical software reliability engineering [Musa et al., 1987].

The main issue in the NHPP model is to determine an appropriate mean value function to denote the expected number of failures experienced up to a certain time point. Model parameters can be estimated by using maximum likelihood estimate (MLE). Parameter values can be obtained using Newton Raphson Method. This paper presents Pareto type II model to analyze the reliability of a software system using interval domain data. The layout of the paper is as follows: Section 2 describes the formulation and interpretation of the model for the underlying NHPP. Section 3 describes the proposed Pareto type II software reliability growth model. Section 4 discusses parameter estimation of Pareto type II model based on interval domain data. Section 5 describes the techniques used for software failure data analysis for a live data and Section 6 contains conclusions.

2. MODEL FORMULATION
There are numerous software reliability growth models available for use according to probabilistic assumptions. The non homogenous Poisson process (NHPP) based software reliability growth models are proved quite successful in practical software reliability engineering. NHPP model formulation is described in the following lines.

A software system is subject to failures at random times caused by errors present in the system. Let $\{N(t), t>0\}$ be a counting process representing the cumulative number of failures by time t. Since there are no failures at t=0 we have

$$N(0) = 0$$

It is reasonable to assume that the number of software failures during non overlapping time intervals do not affect each other. In other words, for any finite collection of times $t1<t2<\ldots.<t_n$ the 'n' random variables $N(t_1)$, $\{N(t_2)-N(t_1)\}$, ….. $\{N(t_n) - N(t_{n-1})\}$ are independent. This implies that the counting process $\{N(t), t>0\}$ has independent increments.

Let m(t) represent the expected number of software failures by time 't'. The mean value function m(t) is finite valued, non decreasing, non negative and bounded with the boundary conditions.

$$m(t) = 0, \quad t = 0$$

$$= a, \quad t \rightarrow \infty$$

where a is the expected number of software errors to be eventually detected.

Suppose N(t) is known to have a Poisson probability mass function with parameters m(t) i.e.

$$P\{N(t) = n\} = \frac{[m(t)]^n . e^{-m(t)}}{n!}$$, n=0,1,2,…∞

then N(t) is called an NHPP. Thus the stochastic behavior of software failure phenomena can be described through the N(t) process. Various time domain models have appeared in the literature (Kantam and Subbarao, 2009) which describe the stochastic failure process by an NHPP which differ in the mean value functions m(t).

3. THE PROPOSED PARETO TYPE II SRGM

In this paper we consider m(t) as given by

$$m(t) = a\left[1 - \frac{c^b}{(t+c)^b}\right] \quad (3.1)$$

where [m(t)/a] is the cumulative distribution function of Pareto type II distribution (Johnson et al, 2004) for the present choice.

$$P\{N(t) = n\} = \frac{[m(t)]^n . e^{-m(t)}}{n!}$$

$$\lim_{n\to\infty} P\{N(t) = n\} = \frac{e^{-a}. a^n}{n!}$$

which is also a Poisson model with mean 'a'.

Let N(t) be the number of errors remaining in the system at time 't'

$$\begin{aligned} N(t) &= N(\infty) - N(t) \\ E[N(t)] &= E[N(\infty)] - E[N(t)] \\ &= a - m(t) \\ &= a - a\left[1 - \frac{c^b}{(t+c)^b}\right] \\ &= \frac{ac^b}{(t+c)^b} \end{aligned}$$

Let S_k be the time between (k-1)th and kth failure of the software product. Let X_k be the time up to the kth failure. Let us find out the probability that time between (k-1)th and kth failures, i.e. S_k exceeds a real number 's' given that the total time up to the (k-1)th failure is equal to x, i.e. P[S_k > s /X_{k-1} = x]

$$R\, S_k/X_{k-1}(s / x) = e^{-[m(x+s)-m(s)]} \quad (3.2)$$

This Expression is called Software Reliability.

4. PARAMETER ESTIMATION BASED ON INTERVAL DOMAIN DATA

In this section we develop expressions to estimate the parameters of the Pareto type II model based on interval domain data. Parameter estimation is of primary importance in software reliability prediction.

A set of failure data is usually collected in one of two common ways, time domain data and interval domain data. In this paper parameters are estimated from the interval domain data.

The mean value function of Pareto type II model is given by

$$m(t) = a\left[1 - \frac{c^b}{(t+c)^b}\right] \quad , \quad t \geq 0 \quad (4.1)$$

In order to have an assessment of the software reliability, a, b and c are to be known or they are to be estimated from software failure data. Expressions are now derived for estimating 'a', 'b' and 'c' for the Pareto type II model.

Assuming that the data are given for the cumulative number of detected errors n_i in a given time interval (0, t_i) where i=1,2,… n and 0 < t_1 < t_2 < … t_n , then the logarithmic likelihood function (LLF) for interval domain data [8] is given by

$$\text{Log L} = \sum_{i=1}^{k}(n_i - n_{i-1}).\log[m(t_i) - m(t_{i-1})] - m(t_k) \quad (4.2)$$

$$\text{Log L} = \sum_{i=1}^{k}(n_i - n_{i-1}) . \log\left[a\left[1 - \frac{c^b}{(t_i+c)^b}\right] - a\left[1 - \frac{c^b}{(t_{i-1}+c)^b}\right]\right] - a\left[1 - \frac{c^b}{(t_k+c)^b}\right] \quad (4.3)$$

$$\text{Log L} = \sum_{i=1}^{k}(n_i - n_{i-1})[\log a + b\log c + \log[(t_i + c)^b - (ti-1+c)b-b\,logti-1+c-b\log(ti+c) - a + acb(tk+c)b \quad (4.4)$$

Accordingly parameters 'a', 'b' and 'c' would be solutions of the equations

$$\frac{\partial \text{LogL}}{\partial a} = 0$$

$$a = \sum_{i=1}^{k}(n_i - n_{i-1})\frac{(t_k+c)^b}{(t_k+c)^b - c^b} \quad (4.5)$$

The parameter 'b' is estimated by iterative Newton Raphson Method using $b_{n+1} = b_n - \frac{g(b_n)}{g'(b_n)}$, where $g(b)$ and $g''(b)$ are expressed as follows.

$$g(b) = \frac{\partial \text{LogL}}{\partial b} = 0$$

$$g(b) = \sum_{i=1}^{k}(n_i - n_{i-1})\left[-\log(t_{i-1}+1) - \log(t_i+1) + (ti+1)b\log(ti+1) - (ti-1+1)b\log(ti-1+1)\,(ti+1)b - (ti-1+1)b\right]^{+}$$

$$\sum_{i=1}^{k}(n_i - n_{i-1}) \log\frac{1}{(t_k+1)} . \frac{1}{(t_k+1)^b - 1} = 0 \quad (4.6)$$

$$g'(b) = \frac{\partial^2 LogL}{\partial b^2} = 0$$

$$g'(b) = \sum_{i=1}^{k}(n_i - n_{i-1}) .$$

$$\frac{2(t_{i-1}+1)^b (t_i+1)^b \log(t_i+1) \log\frac{(t_{i-1}+1)}{(t_i+1)}}{[(t_i+1)^b - (t_{i-1}+1)^b]^2} +$$

$$\sum_{i=1}^{k}(n_i\text{-}n_{i-1}) \log (t_k + 1) \left[\frac{(t_k+1)^{b}\ log(t_k+1)}{[(t_k+1)^{b}-1]^2}\right]$$
(4.7)

The parameter 'c' is estimated by iterative Newton Raphson Method using $c_{n+1} = c_n - \frac{g(c_n)}{g'(c_n)}$, where g(c) and g'' (c) are expressed as follows.

$$g(c) = \frac{\partial \text{LogL}}{\partial c} = 0$$

$$g(c) = \sum_{i=1}^{k}(n_i - n_{i-1}) \left[\frac{1}{c} - \frac{1}{(t_{i-1}+c)} - \frac{1}{(t_i+c)}\right] + \sum_{i=1}^{k}(n_i - n_{i-1}) \frac{1}{(t_k+c)} = 0$$
(4.8)

$$g'(c) = \frac{\partial^2 LogL}{\partial c^2} = 0$$

$$g'(c) = \sum_{i=1}^{k}(n_i - n_{i-1}) \left[-\frac{1}{c^2} + \frac{1}{(t_{i-1}+c)^2} + \frac{1}{(t_i+c)^2}\right] - \sum_{i=1}^{k}(n_i - n_{i-1}) \frac{1}{(t_k+c)^2}$$
(4.9)

The values of 'b' and 'c' in the above equations can be obtained using Newton Raphson Method. Solving the above equations simultaneously, yields the point estimates of the parameters b and c. These equations are to be solved iteratively and their solutions in turn when substituted in equation (4.5) gives value of 'a'.

5. DATA ANALYSIS

In this section, we present the analysis of two software failure data sets. The set of software errors analyzed here is borrowed from a real software development project as published in Pham(2005), which in turn referred to (Pham(2005)) as Zhang et al., (2000). The data are named as phase 1 and phase 2 test data. The phase 1 test data are summarized in the below table.

Table 1. Phase 1 Test Data

Week Index	Exposure time (cum system test hours) (t_i)	Fault (f_i)	Cum. Fault (n_i)
1	356	1	1
2	712	0	1
3	1068	1	2
4	1424	1	3
5	1780	2	5
6	2136	0	5
7	2492	0	5
8	2848	3	8
9	3204	1	9
10	3560	2	11
11	3916	2	13
12	4272	2	15
13	4628	4	19
14	4984	0	19
15	5340	3	22
16	5696	0	22
17	6052	1	23
18	6408	1	24
19	6764	0	24
20	7120	0	24
21	7476	2	26

Solving equations in section 4 by Newton Raphson Method (N-R) method for the Phase 1 test data, the iterative solutions for MLEs of a, b and c are

$\hat{a}$ = 37.120867

$\hat{b}$ = 0.962019

$\hat{c}$ = 3396.758643

Hence, we may accept these three values as MLEs of a, b, c. The estimator of the reliability function from the equation (3.2) at any time x beyond 7476 hours is given by

$$R\, S_k/X_{k-1}\ (s / x) = e^{-[m(x+s)-m(s)]}$$

$$R\, S_{21}/X_{20}(7476/4272) = e^{-[m(4272+7476)-m(7476)]}$$

$$= 0.036560$$

The phase 2 test data are summarized in the below table.

Table 2. Phase 2 Test Data

Week Index	Exposure time (cum system test hours)(t_i)	Fault (f_i)	Cum. Fault (n_i)
1	416	3	3
2	832	1	4
3	1248	0	4
4	1664	3	7
5	2080	2	9
6	2496	0	9
7	2912	1	10
8	3328	3	13
9	3744	4	17
10	4160	2	19
11	4576	4	23
12	4992	2	25
13	5408	5	30
14	5824	2	32
15	6240	4	36
16	6656	1	37
17	7072	2	39
18	7488	0	39
19	7904	0	39
20	8320	3	42
21	8736	1	43

Solving equations in section 4 by Newton Raphson Method (N-R) method for the Phase 2 test data, the iterative solutions for MLEs of a, b and c are

$a^{\wedge} = 59.398002$

$b^{\wedge} = 0.961882$

$c^{\wedge} = 3969 .246055$

Hence, we may accept these three values as MLEs of a, b, c. The estimator of the reliability function from the equation (3.2) at any time x beyond 8736 hours is given by

$$R\,S_k/X_{k-1}\ (s / x) = e^{-[m(x+s)-m(s)]}$$

$$R\,S_{21}/X_{20}(8736/2080) = e^{-[m(2080+8736)-m(8736)]}$$

$$= 0.071912$$

6. CONCLUSION

Software reliability is an important quality measure that quantifies the operational profile of computer systems. In this paper we proposed Pareto type II software reliability growth model. This model is primarily useful in estimating and monitoring software reliability, viewed as a measure of software quality. Equations to obtain the maximum likelihood estimates of the parameters based on interval domain data are developed. Software reliability is accessed for a given two software failure data sets. This analysis shows that phase I data is less reliable compared to phase II data. It provides a plausible description of the software failure phenomenon. This is a simple method for model validation and is very convenient for practitioners of software reliability.

7. ACKNOWLEDGEMENTS

Our thanks to Department of Computer Science and Engineering; Department of Statistics, Acharya Nagarjuna University; Department of Computer Science, Abhinav Institute of Management & Technology, Singarayakonda for providing necessary facilities to carry out the research work.

8. REFERENCES

[1] Goel, A.L., Okumoto, K., 1979. Time- dependent error-detection rate model for software reliability and other performance measures. IEEE Trans. Reliab. R-28, 206-211.

[2] Musa J.D, Software Reliability Engineering MCGraw-Hill, 1998.

[3] Musa,J.D. (1980) "The Measurement and Management of Software Reliability", Proceeding of the IEEE vol.68, No.9, 1131-1142

[4] Musa J.D., Iannino, A., Okumoto, K., 1987. Software Reliability: Measurement Prediction Application. MC Graw Hill, New York.

[5] Pham. H (2005) "A Generalized Logistic Software Reliability Growth Model", Opsearch, Vol.42, No.4, 332-331.MC Graw Hill, New York.

[6] Ramamurthy, C.V., and Bastani, F.B.(1982). "Software Reliability Status and Perspectives", IEEE Transactions on Software Engineering, Vol.SE-8, 359-371.

[7] R.R.L.Kantam and R.Subbarao, 2009. "Pareto Distribution: A Software Reliability Growth Model". International Journal of Performability Engineering, Volume 5, Number 3, April 2009, Paper 9, PP: 275- 281.

[8] Satya Prasad, R and Geetha Rani, N (2011), "Pareto type II software reliability growth model", International Journal of Software Engineering, Volume 2, Issue(4) 81-86.

[9] Satya Prasad, R (2007) "Half logistic Software reliability growth model ", Ph.D Thesis of ANU, India.

[10] Wood, A(1996), "Predicting Software Reliability", IEEE Computer, 2253-2264

9. AUTHORS PROFILES

Mrs. N. Geetha Rani received M.C.A. and M.Tech degrees from Acharya Nagarjuna University. She is currently pursing Ph.D at Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India. She is currently working as Associate Professor and Head of the Department, in the Department of Computer Science, Abhinav Institute of Management and Technology, Singarayakonda, Andhra Pradesh. Her research include on Software Engineering.

Dr. R Satya Prasad received Ph.D. degree in Computer Science in the faculty of Engineering in 2007 from Acharya Nagarjuna University, Guntur, Andhra Pradesh, India. He have a satisfactory consistent academic track of record and received gold medal from Acharya Nagarjuna University for his outstanding performance in a first rank in Masters Degree. He is currently working as Associate Professor in the Department of Computer Science & Engineering, Acharya Nagarjuna University. He has occupied various academic responsibilities like practical examiner, project adjudicator, external member of board of examiners for various Universities and Colleges in and around in Andhra Pradesh. His current research is focused on Software Engineering, Image Processing & Database Management System. He has published several papers in National & International Journals.

R.R.L. Kantam is professor of statistics at Acharya Nagarjuna University, Guntur, India. He has 31 years of teaching experience in statistics at Under Graduate and Post Graduate programs. As researcher in Statistics, he has successfully guided many students for M.Phil and Ph.D in statistics. He has authored more than 60 research publications appeared various statistics and computer science journals published in India and other countries like US, UK, Germany, Pakistan, Srilanka and Bangladesh. He has been a referee for Journal of Applied Statistics (U.K.), METRON (Italy), Pakistan Journal of Statistics (Pakistan), IAPQR - Transactions (India), Assam Statistical Review (India) and Gujarat Statistical Review (India). He has been a special speaker in technical sessions of number of Seminars and Conferences. His areas of research interest are Statistical Inference, Reliability Studies, Quality Control Methods and Actuarial Statistics. As a teacher his present teaching areas are Probability Theory, Reliability and Actuarial Statistics. His earlier teaching topics include Statistical Inference, Mathematical Analysis, Operational Research, Econometrics, Statistical Quality Control, Measure Theory.