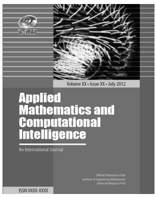
App. Math. and Comp. Intel., Vol. 2(2) (2013) 217–227
http://amci.unimap.edu.my
© 2013 Institute of Engineering Mathematics, UniMAP


Volume XX • Issue XX • July 2012

Applied Mathematics and Computational Intelligence

An International Journal

ISSN XXXX-XXXX

Bayesian Estimation for Poisson Process Models with Grouped Data and Covariate

J. Arasan^{a,*}, **L. Yue Fang**^{b}

[a] *Laboratory of Computational Statistics and Operations Research*
Institute for Mathematical Research, Universiti Putra Malaysia
43400 Serdang, Selangor, Malaysia

[b] *Department of Mathematics, Faculty of Science*
Universiti Putra Malaysia
43400 Serdang, Selangor, Malaysia

Received: Month, Day Year; Accepted: Month, Day Year

Abstract: *This paper looks into the Bayesian approach for analyzing and selecting the best Poisson process model for grouped failure data from a repairable system with covariate. The extended powerlaw model with a recurrence rate that incorporates both time and covariate effect is compared to the powerlaw, log-linear and HPP models. We propose the use of both informative and noninformative priors depending on the nature of the parameter. The MCMC techinque is utilized to obtain samples from the posterior distribution which was implemented via WinBUGS. We then apply the Bayesian Deviance Information Criteria (DIC) to select the best model for real data from ball bearing failures where information regarding previous failures are available. The credible interval is used to check the significance of the parameters of the selected model. We also used the posterior predictive distribution for model checking by comparing the observed and posterior predictive mean number of failures.*

Keywords: *NHPP; Repairable; Interval; DIC.*

PACS: *PACS1, PACS2, PACS3*

1 Introduction

A repairable system is a system that can withstand many failures and be brought back to functionality through some repair action. The “repair time” which is the duration of system downtime is often assumed to be negligible. Grouped data or interval failure data occurs when a component’s failure time falls within a certain interval $(t_{i-1}, t_i]$ where t_{i-1} is the lower inspection time and t_i is the upper inspection time of the i^{th} interval. These types of data usually arise when components are inspected periodically to carry out maintenance or

[*] Corresponding Author: jayanthi@math.upm.edu.my (J. Arasan)

repair actions.

The most widely used models for grouped failure data that are not independent are those based on the nonhomogenous Poisson process (NHPP) because it has a nonconstant recurrence rate, $\nu(t)$. The powerlaw and log-linear models are usually preferred to describe the rate of occurrence of failure (ROCOF) of an NHPP. The random variable of interest is the number of repairs or failures, also known as recurrences, over interval $(t_{i-1}, t_i]$, $N(t_{i-1}, t_i)$. The aim of this paper is to apply the Bayesian approach for analyzing data from a repairable system failure based on the extended powerlaw model with grouped data and covariate. Following that we would like to compare this model to the powerlaw, log-linear and HPP models and select the best model for the real data from ball bearing failures.

Many stochastic models have been developed to describe the failure rate of a NHPP such as the power law model proposed by [7] based on the ideas of [10]. Other popular models are the log-linear proposed by [6] and linear models discussed by [25] and [1]. Lawless and Thiagarajah [24] introduced an important repairable system model that incorporates both time trends and renewal type behavior. This is also known as a proportional intensity model. Guo et al. [14] later proposed a proportional intensity model that is based on the powerlaw model. Guo et al. [13] also developed a new general repair model based on the expected cumulative number of failures to capture the repair history.

Other literature on the repairable system models and recurrent events are by authors such as [3], [11], [17], [19], [18], [26] and [27]. Park et al. [22] presented some application of the log-linear and power log models for grouped failure data in water distribution systems. More details involving recurrent event models for grouped and interval failure data can also be found in [21]. Authors that have used the Bayesian approach in analyzing repairable system failures and lifetime data analysis include [12] and [2], ([16],[15]) and [20].

2 The data

The real data discussed here consists of the number of ball bearing failures in a conveyer belt of an automobile production as discussed by [23]. A total of 25 maintenance actions were performed by the inspection team at different inspection times (hours). There were several failures in a certain time interval for which repair action is then carried out. The time dependent covariate that is believed to affect the recurrence rate is the cumulative number of maintenance action. Information on several historical data is also available which creates opportunity to incorporate them into the current analysis.

Before selecting a suitable model for the data, it is best to do a graphical display of the cumulative number of failures, $N(t_i)$ versus t_i, which is the operating hours, see Fig. 1. This would help us look for trends in the data which then enables us to select a reasonable model. Because we are dealing with grouped data, the graph was drawn using the upper interval point. The plot suggests that the use of a NHPP model might be appropriate since the failure rate seems to be nonconstant.

3 Bayesian inference for Poisson process models with grouped data

In recent years the controversial Bayesian approach has become more popular as an alternative to classical methods. The Bayesian inference does not require the assumption of asymptotic normality as it is usually the case in frequentist statistics. This technique treats any parameter θ as a random variable which are characterized via a prior distribution, $p(\theta)$. Following that, the posterior distribution of θ given the observed data D can be obtained by

$$p(\theta|D) = \frac{p(\theta)L(\theta|D)}{\int L(\theta)L(\theta|D)d\theta}. \tag{1}$$

The Bayesian inference is very appealing because even in cases where there are no explicit solutions to the integral in the denominator, numerical algorithms, such as Markov chain Monte Carlo (MCMC) can be implemented via WinBUGS to draw samples from an arbitrary posterior distribution. The technique constructs a Markov chain where each sample depends on the previous one which eventually converges to the target distribution, also known as posterior distribution. Thus, the posterior distribution can almost always be obtained, even in complex problems.

Following that, the posterior means can be obtained by averaging the Markov chain samples. There are several MCMC algorithms but the most popular ones are the Metropolis-Hastings algorithm and the Gibbs sampling. In order for any Bayesian inference to be successful, the Markov chain has to converge. Several diagnostics for assessing convergence of the Markov chain are readily available in WinBUGS such as the trace plot, autocorrelation plot and MC error. Following that, parameter estimates can be obtained and inferences can be made.

Specification of the prior distribution is the most crucial factor in Bayesian estimation since it affects the posterior inference. An informative prior distribution should reflect some prior information or history about the parameter. These types of priors can be obtained from historical data. A noninformative prior have very little influence on the posterior distribution and are chosen when no prior information is available. Alternatively a low information or vague priors which are simply prior distributions with a large variances can be used.

3.1 Extended Powerlaw and powerlaw models

The powerlaw recurrence rate can be extended to accommodate the effect of covariates by describing the recurrence rate as $\nu(t) = abt^{b-1}e^{gx(t)}$, where $x(t)$ is a time dependent covariate that may affect system failure. The expected number of recurrence in $(t_{i-1}, t_i]$ is $\mu(t_{i-1}, t_i) = E[N(t_{i-1}, t_i)] = \int_{t_{i-1}}^{t_i} \nu(u)du = a[e^{gx(t_i)}(t_i^b - t_{i-1}^b)]$, where $i = 1, 2, \cdots, n$.

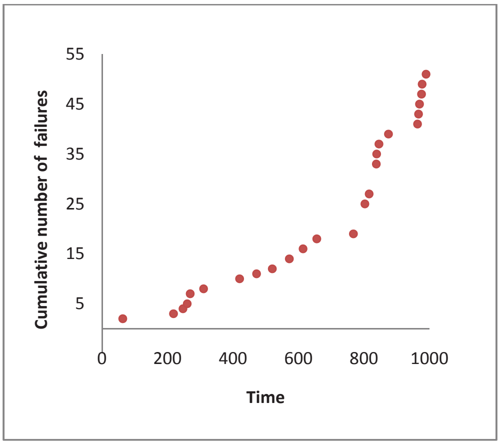


Figure 1: Cumulative number of failure vs. time

The extended powerlaw model allows us to analyze grouped failure data by incorporating the effect of time and covariates simultaneously. Sometimes, the effect of covariates are insignificant and the reduced form of the model may prove to be a better fit for the data and this can easily be obtained by setting $g = 0$, which gives us the powerlaw model. Similarly, when both covariate effect and time trend are insignificant, the model reduces to a HPP model, with a constant recurrence rate, a. The number of intervals is always less or equal to number of failures that we observed because there can be more than one failure in any time interval.

Let $\boldsymbol{d} = (d_1, d_2, \cdots, d_n)$ be the number of failures in the nonoverlapping intervals $\Big[(t_{i-1}, t_i], 1 \leq i \leq n\Big]$ and $x(t_i)$ is the value of covariate at time t_i. The likelihood function is,

$$L(a,b,g|\boldsymbol{d}) = \prod_{i=1}^{n} \frac{[ae^{gx(t_i)}(t_i^b - t_{i-1}^b)]^{d_i}}{d_i!} \times \\ \exp\Big[-\sum_{i=1}^{n} a[e^{gx(t_i)}(t_i^b - t_{i-1}^b)]\Big]. \qquad (2)$$

As discussed earlier, the specification of the prior distribution is very improtant and should be done with much caution. The model discussed above has 3 parameters, a, b and g. To establish the prior distribution for the parameter a, let us begin by considering the HPP model, with a constant recurrence rate, a which is a special case of the extended powerlaw model. For this model, the variable of interest, $N(t_{i-1}, t_i)$ has a Poisson distribution with parameter, $\mu(t_{i-1}, t_i) = a(t_i - t_{i-1})$. Using historical information on ball bearing failure data which in many instances follow a HPP with intensity rate, $a = 0.05$, we adopt a Gamma prior for a, $a \sim Gamma(25, 500)$.

The parameter b and g are easier to interpret compared to the parameter a. When $0 < b < 1$ ($b > 1$), we say the system is improving(deteriorating) which means that failure intensity is decreasing(increasing) with time. Similarly, $g < 0$ ($g > 0$) implies that the repair had a positive(negative) effect on the failure intensity. We adopt a uniform prior distribution for b, $b \sim Gamma(0, 1)$ to indicate that there is some belief of reliability growth based on previous ball bearing failure data but no information of what the value of b might be. For the parameter, g, we assign a noninformative normal prior with large variance, $g \sim N(10^{-3}, 10^{-3})$.

The extended powerlaw model reduces to the powerlaw model when the parameter $g = 0$, with recurrence rate, $\nu(t) = abt^{b-1}$. The expected number of recurrence, $\mu(t_{i-1}, t_i) = a[(t_i^b - t_{i-1}^b)]$, where $i = 1, 2, \cdots, n$. The likelihood for this model is given as,

$$L(a,b,g|\boldsymbol{d}) = \prod_{i=1}^{n} \frac{[a(t_i^b - t_{i-1}^b)]^{d_i}}{d_i!} \times \\ \exp\Big[-\sum_{i=1}^{n} a[(t_i^b - t_{i-1}^b)]\Big]. \qquad (3)$$

As in case of the extended powerlaw model, we assign a Gamma prior for a, $a \sim Gamma(25, 500)$ and a uniform prior distribution for b, $b \sim Gamma(0, 1)$.

3.2 Log-linear and HPP models

Another useful NHPP model that can be used to analyze grouped data without covariate is the log-linear model which has the recurrence rate, $\nu(t) = e^{a+bt}$, where a and b are the parameters of the model. The expected number of recurrence, $\mu(t_{i-1}, t_i) = \frac{e^{(a+bt_i)} - e^{(a+bt_{i-1})}}{b}$, where $i = 1, 2, \cdots, n$. The likelihood function for this model is,

$$L(a, b|\boldsymbol{d}) = \prod_{i=1}^{n} \frac{\left[\frac{e^{(a+bt_i)} - e^{(a+bt_{i-1})}}{b}\right]^{d_i}}{d_i!} \times$$
$$\exp\Big[-\sum_{i=1}^{n} \frac{e^{(a+bt_i)} - e^{(a+bt_{i-1})}}{b}\Big]. \qquad (4)$$

When $b < 0$ ($b > 0$), we say the system is improving(deteriorating) which means that failure intensity is decreasing(increasing) with time. It is hard to specify any informative prior distribution for both the parameters of this model due to the log-linear form of the recurrence rate. Thus, we decide to adopt a noninformative normal prior with large variance for both parameters, $a \sim N(10^{-3}, 10^{-3})$ and $b \sim N(10^{-3}, 10^{-3})$.

The NHPP models discussed above reduces to the HPP with a constant recurrence rate, a as a special case. The HPP model is suitable for a process that has independent increments without any time or covariate effect. The expected number of recurrence, $\mu(t_{i-1}, t_i) = a(t_i - t_{i-1})$, where $i = 1, 2, \cdots, n$. This model only involves 1 parameter, thus by specifying a suitable prior for the parameter, a, the posterior distribution can be obtained without the use of any MCMC sampling technique. The likelihood for this model is given as,

$$L(a|\boldsymbol{d}) = \prod_{i=1}^{n} \frac{\Big[a(t_i - t_{i-1})\Big]^{d_i}}{d_i!} \times$$
$$\exp\Big[-\sum_{i=1}^{n} a(t_i - t_{i-1})\Big] \qquad (5)$$
$$= \frac{a^{\sum d_i} \prod (t_i - t_{i-1})^{d_i} e^{-a\sum_{i=1}^{n}(t_i - t_{i-1})}}{\prod d_i!}. \qquad (6)$$

As in case of the extended powerlaw model, we assign a Gamma prior for a, $a \sim Gamma(\alpha, \beta)$, where $\alpha = 25$ and $\beta = 500$.

$$p(a) = \frac{\beta^{\alpha} a^{\alpha-1} e^{-\beta a}}{\Gamma(\alpha)}. \qquad (7)$$

The posterior distribution given the observed data, $\boldsymbol{d}$ is then,

$$p(a|\boldsymbol{d}) = \frac{a^{\sum d_i} \prod (t_i - t_{i-1})^{d_i} e^{-a\sum_{i=1}^{n}(t_i - t_{i-1})}}{\prod d_i!} \times$$
$$\frac{\beta^{\alpha} a^{\alpha-1} e^{-\beta a}}{\Gamma(\alpha)} \qquad (8)$$

$$\propto e^{-a(t_n+\beta)} a^{\sum d_i + \alpha - 1}. \qquad (9)$$

Thus, the posterior distribution is again gamma($\sum d_i + \alpha, t_n + \beta$). The posterior mean of a is then given by

$$E[a|\boldsymbol{d}] = \frac{\sum d_i + \alpha}{t_n + \beta}. \tag{10}$$

and the posterior variance is given by

$$V[a|\boldsymbol{d}] = \frac{\sum d_i + \alpha}{(t_n + \beta)^2}. \tag{11}$$

4 Model selection and results

4.1 Deviance information criteria

The ball bearing failure data is analyzed via the Bayesian approach using four models, namely the extended powerlaw, powerlaw, log-linear and the HPP. In order to compare and select the best model for the data, the Deviance Information Criterion (DIC) proposed by [4] can be used. This feature is easily available because it is built into WinBUGS/OpenBUGS and thus widely used. The DIC measures both the goodness of fit and the complexity of the model. The goodness of fit of the model is measured via Deviance where a smaller value indicates a better fit. Another appealing feature of DIC is that it can be used to compare models that are not nested.

The Deviance is defined as -2 times the loglikelihood, $D(\theta) = -2logL(D|\theta)$. In Bayesian statistics, the deviance can be calculated using the posterior mean of the deviance, $\overline{D(\theta)} = E_{\theta|D}[D]$, ([8], [9]). The deviance can also be evaluated at the posterior mean of the parameter, θ, $D(\bar{\theta}) = D\{E_{\theta|D}(\theta)\}$. The complexity is measured by estimating the effective number of parameters in the model using, $pD = E_{\theta|D}[D] - D\{E_{\theta|D}[\theta]\}$ or $pD = \overline{D(\theta)} - D(\bar{\theta})$. Thus, $DIC = \overline{D} + pD$.

As discussed earlier the DIC feature in built into WinBUGS where $\overline{D(\theta)}$ is the posterior mean of $-2logL(D|\theta)$ and $D(\bar{\theta})$ is $-2logL(D|\theta)$ at the posterior means. DIC can be monitored in WinBUGS from Inference/DIC menu where it indicates the model fit using $\bar{D}$ and $\hat{D}$ and the complexity of the model using pD. The model with the lowest DIC is thus preferred as the simplest model that fits the data best. The DIC can only be used to compare Bayesian models with complete data, however it can be modified to include the analysis on missing data as discussed by [5].

4.2 Results

The posterior estimates for all parameters were obtained via the MCMC using WinBUGS. We used 10,000 samples and discarded the first 1000 iterations. We computed the posterior means and the 95% credible interval for all parameters. Tables 1-4 give the posterior summary for the parameters of different models. To select the best model for the data , we observe the value of DIC given in table 5. The best model is the one with the lowest of DIC, which in this case is the extended powerlaw model.

The posterior mean of b for the extended powerlaw model implies that there is a reliability improvement. The posterior mean of g is positive which means that the maintenance action could not prevent the system from deteriorating with time. We know that if parameters g and b are significant then there is evidence of maintenance effect and time trend within the model.

The significance of the parameters, b and g can be checked using the 95% credible intervals. A credible interval is a Bayesian interval estimate which is an interval having a specified posterior probability. Both the credible intervals for b and g do not include the point zero which implies that both the parameters are significant. Fig. 2 and Fig. 3 show the marginal posterior distribution and the history plot for parameters, a, b and g.

4.3 Model checking using posterior predictive distribution

The main aim of any modeling in most cases is to carry out future predictions which can be very useful in saving costs, designing studies and checking model compatibility. The Bayesian inference allows future predictions to be carried out based on the posterior predictive distribution. Similarly, we can also compare the actual and the predicted number of failures using the posterior predictive distribution for any model. This would indicate if the model selected provides a good fit for the observed data.

Fig. 4 shows the observed cumulative number of failures and the posterior predicted mean cumulative number of failure using the extended powerlaw model. We can clearly see that the extended powerlaw model seems to show a good fit for the real data. Fig. 5 shows the 95% credible intervals for the posterior predicted mean cumulative number of failure.

5 Conclusion

In this paper, we used the Bayesian modeling approach to compare and select the best model for grouped ball bearing failure data with covariate. We found the extended powerlaw model to be the best model for this data. The model reduces to the power law and HPP as a special case, thus is very convenient and useful. This model also allows us to incorporate and analyze both time trend and covariate effects simultaneously. The results indicate that although there is a reliability improvement, the repair action could not prevent the system from deteriorating further.

Table 1. Posterior summary for extended powerlaw model

Parameter	Mean	sd	2.50%	97.50%
a	0.0500	0.0100	0.0323	0.0711
b	0.7910	0.0612	0.6707	0.9095
g	0.1309	0.0257	0.0816	0.1816

Table 2. Posterior summary for powerlaw model

Parameter	Mean	sd	2.50%	97.50%
a	0.0436	0.0091	0.0277	0.0631
b	1.0350	0.0358	0.9671	1.1070

Table 3. Posterior summary for Log-linear model

Parameter	Mean	sd	2.50%	97.50%
a	-4.6390	0.4266	-5.5380	-3.8580
b	0.0028	0.0006	0.0017	0.0039

Table 4. Posterior summary for HPP model

Parameter	Mean	sd	2.50%	97.50%
a	0.0510	0.0060	0.0398	0.0631

Table 5. DIC values for different models

Model	DIC	Dbar	pD
Ext.Powerlaw	500128	500126	2.006
Log Linear	500129	500127	1.961
Powerlaw	500151	500150	0.884
HPP	500152	500152	0.706

The Bayesian method is very appealing because it allows us to incorporate prior information regarding the system failure in our estimation process. The high capability of modern day computers makes the MCMC simulation very simple and practical allowing inferences even in complex models. The availability of the DIC measure also makes it very easy for model comparison and selection. In addition, the Bayesian posterior predictive distribution allows us to conduct model checking and also carry out future prediction. The credible intervals produced via the MCMC simulation provide a convenient way to check the significance of the model parameters. More research can be done by implementing the Bayesian methods discussed in this paper to other Poisson process models with different priors or the priors arising from historical data, also known as power priors Ibrahim and Chen [16].

References

[1] Atwood, C.L. Parametric estimation of time-dependent failure rates for probabilistic risk assessment. *Reliability Engineering and System Safety*, 37:181–194, 1992.

[2] Bar-Lev, S., Lavi, I. and Reiser, B. Bayesian inference for power law process. *Annals of the Institute of Statistical Mathematics*, 44(4):623–639, 1992.

[3] Brown, M. and Proschan, F. Imperfect repair. *Journal of Applied Probability*, 20: 851–859, 1983.

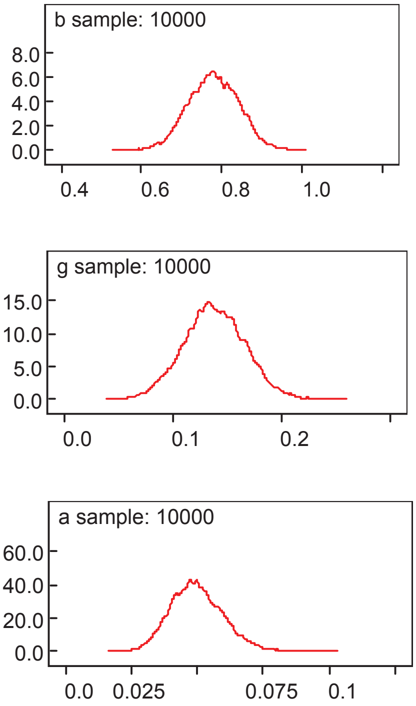


Figure 2: Posterior distribution of parameters of extended powerlaw model

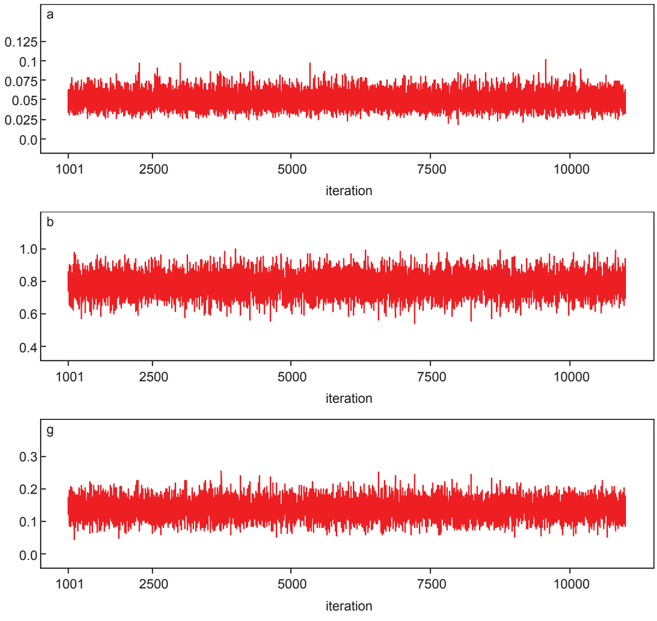


Figure 3: History plot of parameters of extended powerlaw model

[4] Spiegelhalter, D.L., Best, N.G., Carlin, B.P. and Van der Linde, A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64(4):583–1639, 2002.

[5] Celeux, G., Forbes, F., Robert, C.P. and Titterington, D.M. Deviance information criteria for missing data models. bayesian analysis. *Bayesian Analysis*, 1:651–674, 2006.

[6] Cox, D.R. and Lewis, P.A. *Statistical Analysis of Series of Events.* Methuen, London, 1966.

[7] Crow, L.H. Reliability analysis for complex, repairable systems in reliability and biometry, F.Proschan and R.J. Serfling, Eds, Philadelphia, PA. *SIAM*, pages 379–410, 1974.

[8] Dempster, A.P. The direct use of likelihood for significance testing. *In Proceedings of Conference on Foundational Questions in Statistical Inference,*(ed. O. Barndor-Nielsen, P. Blaesild, and G. Schou)pp. 335-354. University of Aarhus, 1973.

[9] Dempster, A.P. The direct use of likelihood for significance testing. *Statistics and Computing*, 7:247–252, 1997.

[10] Duanne, J.T. Learning curve approach to reliability monitoring. *IEEE Transactions on Aerospace*, 2:563–566, 1964.

[11] Gasmi, S., Love, C.E. and Kahle, W. A general repair, proportional-hazards, framework to model complex repairable systems. *IEEE Transactions on Reliability*, 52(1):26–32, 2003.

[12] Guida, M., Balabria, R. and Pulcini, G. Bayesian inference for nonhomogenous poisson process with power intensity law. *Annals of the Institute of Statistical Mathematics*, 44(4):623–639, 1992.

[13] Guo, H.R., Liao, H., Zhao, W. and Mettas, A. A new stochastic model for systems under general repairs. *IEEE Transactions on Reliability*, 56(1):40–47, 2007.

[14] Guo, H.R., Zhao, W. and Mettas, A. Practical methods for modeling repairable systems with time trends and repair effects. *Annual Reliability and Maintainability Symposium*, pages 182–188, 2006.

[15] Ibrahim, J.G. and Chen, M.H. Power prior distributions and bayesian computation for proportional hazards models. *Sankhya*, B 60:48–64, 1998.

[16] Ibrahim, J.G. and Chen, M.H. Power prior distributions for regression models. *Statistical Science*, 15:46–60, 2000.

[17] Kaminskiy, M. and Krivtsov, V. *"A Monte Carlo Approach to Repairable System Repairable Analysis", Probabilistic Safety Assessment and Management.* New York: Springer, 1998.

[18] Kijima, M. Some results for repairable systems with general repair. *Journal of Applied Probability*, 26:89–102, 1989.

[19] Kijima, M. and Sumita, N. A useful generalization of renewal theory: Counting process governed by non-negative markovian increments. *Journal of Applied Probability*, 23: 71–88, 1986.

[20] Kim, S.W and Ibrahim, J.G. On bayesian inference for proportional hazards models using noninformative priors. *Lifetime Data Analysis*, 6:331–341, 2001.

[21] Meeker, W.Q and Escobar, L.A. *Statistical Methods for Reliability Data.* Wiley, New York, 1998.

[22] Park, S., Jun, H., Kim, B.J. and Im, G.C. Modeling of water main failure rates using the log-linear rocof and the power law process. *Water Resource Management*, 22:1311–1324, 2008.

[23] Samira, E.F. *Modelling repairable system with covariate and interval failure data.* (Master's thesis), Universiti Putra Malaysia, 2010.

[24] Thiagarajah, K. and Lawless, J.F. A point-process model incorporating renewals and time trends, with application to repairable systems. *Technometrics*, 38:131–138, 1996.

[25] Vesely, W.E. *Estimating Common Cause Failure Probabilities in Reliability and Risk Analysis:Marshall-Olkin Specialization. In Nuclear System Reliability Engineering and Risk Assessment.* J.B. Fussell, and G.R. Burdick, eds. SIAM, Philadelphia, pp. 314-341, 1977.

[26] Wang, H. and Pham, N. Optimal age-dependent preventive maintenance policies with imperfect maintenance. *International Journal of Reliability, Quality and Safety Engineering*, 3:119–135, 1996.

[27] Yanez, M., Joglar, F. and Modarres, M. Generalized renewal process for analysis of repairable systems with limited failure experience. *International Journal of Reliability, Quality and Safety Engineering*, 77:1167–180, 2002.

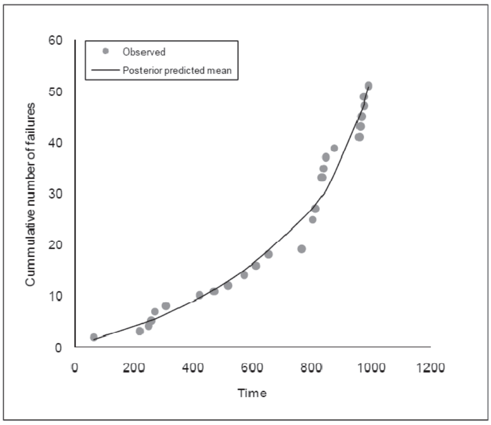


Figure 4: Observed vs. predicted cumulative number of failures

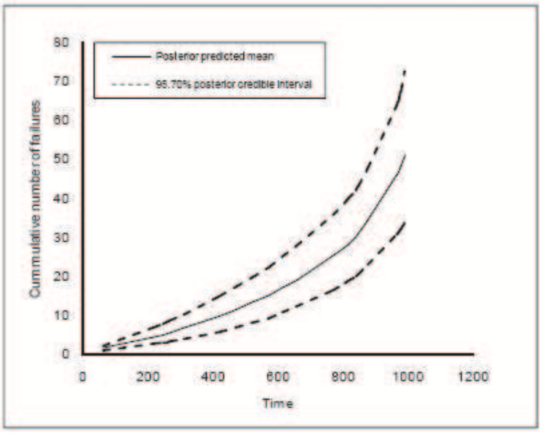


Figure 5: Observed vs. predicted cumulative number of failures