Shahid Bahonar University of Kerman
Journal of Mahani Mathematical Research
JMMR 2012
Print ISSN: 2251-7952 Online ISSN: 2645-4505

GOODNESS-OF-FIT TESTS FOR IMPERFECT MAINTENANCE MODELS BASED ON MARTINGALE RESIDUALS, VARENTROPY, AND PROBABILITY INTEGRAL TRANSFORM

F. Nezampour, H. Alizadeh Noughabi ✉, and M. Chahkandi

Article type: Research Article

(Received: 14 December 2024, Received in revised form 05 May 2025)

(Accepted: 01 July 2025, Published Online: 01 July 2025)

Abstract. In recent years, various goodness-of-fit tests have been developed to identify the underlying distribution of failure data. In this paper, we extend the application of such tests to evaluate the adequacy of imperfect maintenance models for engineering systems. Specifically, we investigate and compare three types of test statistics: those based on martingale residuals, the probability integral transform, and varentropy—a concept derived from information theory. The null hypothesis assumes that the failure times follow the ARA_∞ model with a power law process (PLP) as the initial hazard rate. To evaluate the performance of the proposed tests, we conduct extensive simulation studies under different alternative maintenance models (e.g., ARA_1, ARA_∞–Log Linear Process(LLP)) and varying parameter settings. Our findings show that the power of the tests varies depending on the nature of the alternatives, and varentropy-based statistics outperform others under certain conditions. Finally, we apply the proposed methods to a real dataset (Ambassador vehicle failure times) to assess their practical relevance. The results confirm the validity of the fitted model and demonstrate the usefulness of varentropy-based approaches for detecting subtle deviations in maintenance patterns.
Keywords: Bootstrap, Goodness-of-fit test, Imperfect maintenance, Repairable systems, Varentropy.
2020 MSC: 62G10, 62N05.

1. Introduction

Reliability studies often involve systems with high replacement costs, making maintenance essential for their continuous operation. Maintenance significantly affects a system's reliability, and understanding its impact is crucial. Two primary maintenance models are commonly used: minimal repair and complete repair.

Minimal repair assumes that after maintenance, the system's hazard rate (the probability of failure) remains unchanged, effectively leaving it "as bad as old." In contrast, complete repair restores the system to an "as good as

✉ alizadehhadi@birjand.ac.ir, ORCID: 0000-0002-7515-1896
https://doi.org/10.22103/jmmr.2025.24517.1733

Publisher: Shahid Bahonar University of Kerman
How to cite: F. Nezampour, H. Alizadeh Noughabi, M. Chahkandi, *Goodness-of-fit tests for imperfect maintenance models based on Martingale residuals, varentropy, and probability integral transform*, J. Mahani Math. Res. 2026; 15(1): 71-87.

new" condition, resetting its hazard rate. However, reality often lies between these extremes. Imperfect maintenance acknowledges that while maintenance improves a system's performance, it may not fully restore it to its original state. The system's hazard rate after imperfect maintenance falls between the "as bad as old" and "as good as new" states.

In this context, understanding and assessing the effectiveness of maintenance models becomes essential. This research focuses on a repairable system under imperfect repair, emphasizing corrective maintenance, where repairs are conducted after a system failure, followed by restarting the system. The study specifically investigates the sequence of failures or corrective maintenance times within a system and evaluates the goodness-of-fit for several maintenance models.

This article examines several maintenance models, such as ARA_1 (Arithmetic Reduction of Age 1), ARA_∞ (Arithmetic Reduction of Age Infinity), QR (Quasi-Renewal), EGP (Extended Geometric Process) and BP (Brown–Proschan).
The incomplete maintenance model comprises two components:

- Initial Hazard Rate: Describes the inherent wear of the system before any maintenance.
- Effect of Maintenance: If the system's hazard rate remains unchanged after maintenance, the model is considered 'as bad as old,' and the counting process follows a non-homogeneous Poisson process with hazard rate $\lambda_t = l(t)$ for $t \geq 0$, where $l(t)$ is the system's initial hazard rate.

Typically, the first failure time is assumed to follow a Weibull distribution, resulting in the initial intensity being that of a Power Law Process (PLP), given by $l(t) = abt^{b-1}$ where $a > 0$ and $b > 0$, for $t \geq 0$. In practice, the intensity of a Log-Linear Process (LLP) is also considered, expressed as $l(t) = \exp(a + bt)$ where $a, b \in \mathbb{R}$, for $t \geq 0$. Maintenance can also restore the system to an "as good as new" state, in which case the times between successive failures are identically distributed and independent, and the counting process is a renewal process with hazard rate $\lambda_t = l(t - T_{N_{t^-}})$ for $t \geq 0$. Now, we present an overview of some known imperfect repair models, where their hazard functions of the proposed imperfect repair after the repair actions are in Table 1.

- ABAO (As Bad As Old [12]): The system's hazard rate remains unchanged after maintenance.
- AGAN (As Good As New [30]): The system's hazard rate is reset to its initial state after maintenance. The $T_{N_{t^-}}$ is the last time before the most recent repair time.
- BP (Brown–Proschan [7]): After each failure, with probability p, a perfect repair is performed, and with probability $1 - p$, a minimal repair is performed. In Table 1, C_k's are independent and identically distributed random variables with a Bernoulli distribution parameter

p such that $C_r = 1$ if the r-th repair is complete and $C_r = 0$ if the r-th repair is minimal.

- QR (Quasi-Renewal [30]): The times between failures follow a geometric process influenced by the system's aging or rejuvenation. $q > 0$ is a parameter that determines the effect of the repair.
- EGP (Extended Geometric Process [6]): A generalized version of the QR model with varying intervals between failures, where $\{g_i\}_{\{i \in \mathbb{N}\}}$ is an increasing sequence of positive real numbers such that $g_1 = 0$ and $\lim_{i \to \infty} g_i = \infty$.
- ARA_∞ (Arithmetic Reduction of Age Infinity [15]): The virtual age model with infinite memory is multiplied from the beginning to time t. Maintenance reduces the effective age by a repair factor $\rho \leq 1$. The repair factor, or ρ, is a parameter that determines the efficiency of maintenance.
- ARA_1 (Arithmetic Reduction of Age 1 [15]): The virtual age model with memory one is multiplied by a factor in each interval. Maintenance reduces the effective age by a fixed proportion ρ. The repair factor, or ρ, is a parameter that determines the efficiency of maintenance.

Table 1. Hazard rate functions for different imperfect repair models.

Model	**Hazard Rate Function**
ABAO	$\lambda_t = l(t)$
AGAN	$\lambda_t = l(t - T_{N_{t^-}})$
BP	$\lambda_t = l(t - T_{N_{t^-}} + \sum_{j=1}^{N_{t^-}} (\prod_{k=j}^{N_{t^-}} [1 - C_k])(T_j - T_{j-1}))$
QR	$\lambda_t = q^{-N_{t^-}} l(q^{-N_{t^-}}(t - T_{N_{t^-}}))$
EGP	$\lambda_t = q^{-g_i} l(q^{-g_i}(t - T_{N_{t^-}}))$
ARA_∞	$\lambda_t = l(t - \rho \sum_{j=0}^{N_{t^-}-1} (1-\rho)^j T_{N_{t^-}-j})$
ARA_1	$\lambda_t = l(t - \rho T_{N_{t^-}})$

Figures 2 and 1 illustrate the intensity plots for models ARA_∞ and ARA_1, respectively. Both figures use an initial Weibull distribution with parameters (1, 3) and a repair factor of 0.2. The models are plotted for 7 failure times. A critical aspect of reliability analysis is evaluating whether observed data aligns with a specific model. Goodness-of-fit tests play a crucial role in this process, particularly in the context of imperfect maintenance models. These tests help check if the chosen model fits the observed data well. Traditional goodness-of-fit tests, such as those described by D'Agostino and Stephens [13], aim to transform data into a state where it is identically distributed and independent. Techniques like conditional probability integral transforms (Gaudoin [18]) and

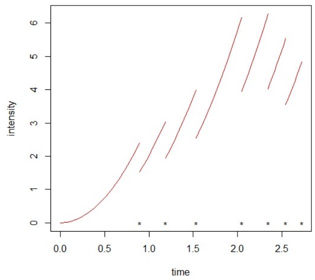

FIGURE 1. Plot of the failure intensity for ARA_∞ model.

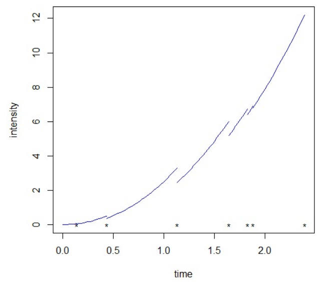

FIGURE 2. Plot of the failure intensity for ARA_1 model.

sequential transforms (Crétois et al. [11]) fall into this category. Other approaches have been proposed by Park and Kim [24], Zhao and Wang [31], and Lindqvist and Rannestad [22].

Lindqvist et al. [21] and Liu et al. [23] developed methods for goodness-of-fit tests in imperfect maintenance models, focusing on transforming failure times into uniform random variables. However, these studies lacked simulation-based performance evaluations. Chauvel et al. [8] introduced goodness-of-fit tests using parametric bootstrap for examining imperfect maintenance models.

Incorporating recent advancements, Varentropy has demonstrated growing significance across diverse research fields. For instance, Saha and Kayal [26] proposed a weighted (residual) varentropy framework and showcased its effectiveness in various applied contexts, while Leonenko et al. [20] developed a novel estimation method for varentropy based on nearest neighbor graphs, enhancing its applicability in high-dimensional settings. Furthermore, Alizadeh Noughabi and Shafaei Noghabi [4] applied varentropy estimators to develop goodness-of-fit tests for the Gumbel distribution, demonstrating favorable power properties. Additionally, Alizadeh Noughabi [3] introduced varentropy estimators for constructing tests of uniformity, which performed well against various alternatives.

These studies collectively highlight the increasing interest and utility of Varentropy in enhancing the robustness and applicability of both predictive models and goodness-of-fit tests across diverse research areas.

This paper aims to demonstrate that test statistics based on martingale residuals, probability integral transform, and varentropy outperform those proposed by Chauvel et al. [8] in terms of power. This is evidenced through a detailed comparison and analysis of various test results. The findings suggest that these statistics are particularly effective at detecting patterns and changes in data.

The research employs simulations to generate synthetic data from different models and evaluate the performance of the statistics. The simulation results underscore the superior ability of these statistics to detect changes and patterns under various conditions.

In the second section, the importance of goodness-of-fit tests in reliability analysis was discussed and various methods were introduced. The third section detailed the simulation methodology used to evaluate the performance of proposed statistics. The fourth section presented the results, showing the superior power of the proposed methods. Finally, the fifth section concluded that the proposed statistics are effective tools for detecting patterns and changes in data.

2. **Parametric Bootstrap Goodness-of-Fit Analysis**

The imperfect maintenance model is characterized by its hazard rate function, defined as:

$$P = \{\lambda(\theta) \mid \theta \in \Theta \subset \mathbb{R}^d\}$$

where θ represents the model parameter. Our goal is to assess whether P is an appropriate model for the observed data $T_1, \ldots, T_n$. The goodness-of-fit test is formulated as:

$$H_0 : \lambda \in P \quad \text{versus} \quad H_1 : \lambda \notin P.$$

We aim to reject the null hypothesis when the test statistic falls within the critical region, determined using the exact or asymptotic distribution quantiles of the statistic under H_0. Therefore, we seek a test statistic that quantifies the discrepancy between the data and the model, allowing us to determine the distribution of the statistic under H_0.

Bootstrap Methods: Chauvel et al. [8] proposed two families of goodness-of-fit tests based on martingale residuals and probability integral transforms. For each test, the quantiles of the statistics under H_0 are calculated using parametric bootstrap methods.

Bootstrap methods, introduced by Efron [17], are part of a larger class of resampling methods. The general idea behind these methods is that observations contain all the information about their distribution without making any additional assumptions. For more information on bootstrap, refer to Tibshirani and Efron [28] and Davison and Hinkley [14]. Chauvel et al. [8] fitted a parametric model to the data and used parametric bootstrap goodness-of-fit tests derived from the method developed by Stute et al. [27] for identically distributed and independent random variables.

Suppose $W(\widehat{\theta})$ represents the test statistic. Under the null hypothesis H_0, $W(\widehat{\theta})$ is derived from the dataset $T_1, \ldots, T_n$, which originates from a point process characterized by the hazard rate $\lambda(\theta)$. A point process describes data as a collection of irregular and random points within a specified region (Vere-Jones and Daley, [29]). Our objective is to obtain times that are identically distributed and independent of $W(\widehat{\theta})$ to compute empirical quantiles. Since the parameter θ is unknown, we estimate it with $\widehat{\theta}$ and simulate identically distributed and independent times $T_1^*, \ldots, T_n^*$ from a point process with hazard rate $\lambda(\widehat{\theta})$. For each sample of times, the maximum likelihood estimator θ^* and the test statistic $W^*(\widehat{\theta})$ can be computed. The similarity between θ and $\widehat{\theta}$ implies that we expect minimal differences between the empirical quantiles of $W(\widehat{\theta})$ and $W^*(\widehat{\theta})$.

General Bootstrap Method: The general bootstrap method for applying the test is outlined in the following steps:

1. Calculate the Estimate: Determine the maximum likelihood estimate $\widehat{\theta}$ within the class of models P and compute the statistic $W(\widehat{\theta})$ using the dataset $T_1, \ldots, T_n$.

2. Bootstrap Sampling:

- For $i = 1$ to L:
- Generate bootstrap samples $T_{n,1}^*, \ldots, T_{n,i}^*$ under the model with hazard rate $\lambda(\widehat{\theta}) \in P$.
- Calculate the maximum likelihood estimate $\widehat{\theta_i}^*$ from the bootstrap samples $T_{n,1}^*, \ldots, T_{n,i}^*$ within the model P.
- Compute the bootstrap test statistic $W_i = W_i(\widehat{\theta_i^*})$ from the bootstrap samples and the estimate $\widehat{\theta_i}^*$.

3. Hypothesis Testing: Compare the observed test statistic $W(\widehat{\theta})$ with the empirical quantiles of order $1-\alpha$ derived from the bootstrap statistics $W_1^*, \ldots, W_L^*$. Based on this comparison, decide whether to reject the null hypothesis H_0 at the significance level α.

All simulations and analyses have been conducted using the R programming software.

2.1. **Tests Based on Martingale Residuals.** Consider the residuals $\widehat{M} = d - \widehat{e}$, which calculate the difference between what we observed and what we expected during the period the system was at risk. Positive values mean the system failed earlier than expected, and negative values mean the system lasted longer than expected. The stochastic process $M = (M_t)_{t\geq 0}$ where

$$E(M(t)) = 0, \forall t,$$
$$E[M(t)|M(s)] = M(s), \forall s < t.$$

is known in statistics as a martingale (Andersen et al. [5]). Now, suppose $\Lambda = (\Lambda_t)_{t\geq 0}$ represents the cumulative hazard rate of the process N, such that $\Lambda_t = \int_0^t \lambda_s ds$ for $t \geq 0$. The process $M = (M_t)_{t\geq 0}$ defined as $M = N - \Lambda$ is a zero-mean martingale. Thus, N is close to Λ in the sense that we expect their difference to be zero. In the hypotheses, the hazard rate has a parametric form and is represented as $\lambda(\theta) = (\lambda_t(\theta))_{t\geq 0}$ for $\theta \in \Theta \subset R^d$. The cumulative hazard rate is $\Lambda(\theta)$ and the corresponding martingale is $M(\theta) = N - \Lambda(\theta)$. In practice, the parameter θ is unknown and the cumulative hazard rate is estimated from $T_1, ..., T_n$. Suppose $\widehat{\theta}$ is the maximum likelihood estimator of θ. The random variables $\widehat{M_1}, ..., \widehat{M_n}$ are defined as:

$$\widehat{M_i} = N_{T_i} - \Lambda_{T_i}(\widehat{\theta}), i \in \{1, ..., n\},$$

which are referred to as martingale residuals (Cook and Lawless [9]). It is anticipated that N closely approximates $\Lambda(\theta)$.
The first group of goodness-of-fit tests is based on the discrepancy between N and $\Lambda(\theta)$. These tests reject the validity of the model if the two processes diverge significantly. Chauvel et al. [8] introduced three test statistics based on martingale residuals. The first is a Kolmogorov-Smirnov type statistic:

$$(1) \qquad KS_m(\widehat{\theta}) = \sup_{i=1,...,n} |\widehat{M_i}| = \sup_{i=1,...,n} |i - \Lambda_{T_i}(\widehat{\theta})|.$$

The second is a Cramér-von Mises type statistic as:

$$(2) \qquad CvM_m(\widehat{\theta}) = -\frac{1}{3}\sum_{i=1}^{n}\{(i-1-\Lambda_{T_i}(\widehat{\theta}))^3 - (i-1-\Lambda_{T_{i-1}}(\widehat{\theta}))^3\}.$$

The third is an Anderson-Darling type statistic:

$$(3) \qquad AD_m(\widehat{\theta}) = \frac{1}{(n+1)}\sum_{i=2}^{n}\{(i-1)^2 \log(\frac{\Lambda_{T_i}(\widehat{\theta})}{\Lambda_{T_{i-1}}(\widehat{\theta})}) - (n+2-i)^2 \log(\frac{n+1-\Lambda_{T_i}(\widehat{\theta})}{n+1-\Lambda_{T_{i-1}}(\widehat{\theta})})\} + (n+1)\log(1-\frac{\Lambda_{T_1}(\widehat{\theta})}{n+1}) - n.$$

The distribution of test statistics under the null hypothesis does not follow standard distributions and may be parameter-dependent. Consequently, their quantiles need to be determined using parametric bootstrap methods. If the values of the statistics in (1) to (3) exceed their parametric bootstrap quantiles, the null hypothesis is rejected.

2.2. **Tests Based on Probability Integral Transform.** The second category of tests relies on the random variables $\Lambda_{T_{i+1}}(\theta)-\Lambda_{T_i}(\theta)$ for $i=0,\ldots,n-1$. Under the null hypothesis H_0, these variables are identically distributed and independent, following a standard exponential distribution (Cook and Lawless [9]). For $i=0,\ldots,n-1$, let $\overline{F}(\cdot|T_i;\theta)$ be the reliability function of the failure time $T_{i+1}-T_i$ conditional on $T_i=(T_1,T_2,\ldots,T_i)$, then

$$\overline{F}(a|T_i;\theta)=P(T_{i+1}-T_i>a|T_i;\theta)=\exp(-\Lambda_{T_i+a}(\theta)+\Lambda_{T_i}(\theta)),\quad a\geq 0.$$

We introduce the variables $U_i(\theta)=\overline{F}(T_{i+1}-T_i|T_i;\theta)$ for $i=0,\ldots,n-1$. Under the null hypothesis $H_0:\lambda\in V$, the U_i variables are identically distributed and independent, following a standard uniform distribution. This transformation of failure times is commonly known as the probability integral transform. It involves applying a cumulative distribution function to a random variable (D'Agostino and Stephens [13]). In probability theory, this transformation demonstrates that data values modeled as random variables from any continuous distribution can be converted into random variables with a standard uniform distribution. When the cumulative distribution function is conditional on past values, it is referred to as the Rosenblatt transform (Rosenblatt [25]), also known as the conditional probability integral transform.
The second type of goodness-of-fit tests relies on the conditional probability integral transform of the inter-failure times. If there is a lack of fit, the uniformity assumption will no longer hold. In practice, θ is estimated, and the Kolmogorov-Smirnov, Cramér-von Mises, and Anderson-Darling test statistics are used to test the uniformity of $U_0(\widehat{\theta}),\ldots,U_{n-1}(\widehat{\theta})$ (Chauvel et al. [8]). The test statistics are as follows:

$$(4)\quad KS_u(\widehat{\theta})=\sqrt{n}\max\{\max_{i=1,\ldots,n}(\frac{i}{n}-U_{(i-1)}(\widehat{\theta})),\max_{i=1,\ldots,n}(U_{(i-1)}(\widehat{\theta})-\frac{i-1}{n})\}.$$

$$(5)\quad CvM_u(\widehat{\theta})=\sum_{i=1}^{n}(U_{(i-1)}(\widehat{\theta})-\frac{2i-1}{2n})^2+\frac{1}{12n}.$$

$$(6)\quad AD_u(\widehat{\theta})=-n-\frac{1}{n}\sum_{i=1}^{n}(2i-1)\{log(U_{(i-1)}(\widehat{\theta}))+log(1-U_{(i-1)}(\widehat{\theta}))\}.$$

where $U_{(0)}(\widehat{\theta}),\ldots,U_{(n-1)}(\widehat{\theta})$ are the order statistics of $U_i(\widehat{\theta})$. If the values of the statistics in (4) to (6) exceed their parametric bootstrap quantiles, the null hypothesis is rejected.

2.3. **Proposed Tests Based on Varentropy.** Previous studies have utilized varentropy as a tool for testing the goodness-of-fit of data to specific probability distributions. In this work, we extend the application of varentropy-based test statistics to assess the suitability of imperfect maintenance models. Specifically, we explore the potential of these statistics in identifying whether the transformed data follow a uniform distribution, which is a key step in model validation via the probability integral transform. Varentropy is a measure of the variability or dispersion in the informational content of a random variable. It provides deeper insights into the uncertainty structure of data beyond traditional entropy measures. For a continuous random variable T with a probability density function f, the varentropy of T is given by:

$$\text{Var}[\log(f(t))] = \int_{-\infty}^{+\infty} f(t)(\log(f(t)))^2 dt - \left(\int_{-\infty}^{+\infty} f(t)\log(f(t))dt\right)^2$$

Alizadeh Noughabi and Shafaei Noughabi [4] proposed the following test statistics based on varentropy:

$$(7)\quad VV_{mn} = \frac{1}{n}\sum_{i=1}^{n}\left(\log^2(U_{(i+m)} - U_{(i-m)})\right) - \left(\frac{1}{n}\sum_{i=1}^{n}\log(U_{(i+m)} - U_{(i-m)})\right)^2$$

and
(8)

$$VE_{mn} = \frac{1}{n}\sum_{i=1}^{n}\left(\log^2\left(\frac{r_i m/n}{U_{(i+m)} - U_{(i-m)}}\right)\right) - \left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{r_i m}{n}\frac{1}{U_{(i+m)} - U_{(i-m)}}\right)\right)^2$$

Ebrahimi et al. [16] introduced the coefficient r_i as follows:

$$r_i = \begin{cases} 1 + \frac{i-1}{m} & 1 \le i \le m, \\ 2 & m+1 \le i \le n-m, \\ 1 + \frac{n-i}{m} & n-m+1 \le i \le n. \end{cases}$$

In these statistics, $m = [\sqrt{n} + 0.5]$ and if $i + m > n$, then $i + m = n$. Also, if $i - m < 1$, then $i - m = 1$
Here, $U_{(i)}$ denotes the i-th order statistic of the transformed data U_i, obtained through the probability integral transform. The parameter m defines a window size around each data point and is chosen to balance bias and variance in the estimation process.
To avoid invalid indexing when i is near the dataset boundaries, the conditions $i + m > n$ and $i - m < 1$ are handled by setting $i + m = n$ and $i - m = 1$, respectively. This ensures that the window remains within valid bounds and the statistics are well-defined throughout the data range.

3. Simulation Methodology

To assess the performance of the test statistics, we applied them to a large number of simulated datasets. The power of a test is estimated by the percentage of times the null hypothesis is rejected. First, we need to ensure that the significance level of the tests is well-maintained. The empirical level is the percentage of times H_0 is rejected when data are simulated under H_0. The empirical level should be close to the theoretical level α. When data are simulated under a different model, the rejection percentage should be as high as possible.

We assumed the significance level of the test to be $\alpha = 0.05$. The null hypothesis corresponds to the imperfect maintenance model ARA_∞ (Kijima Type II model) with an initial hazard rate PLP. This model is denoted by $ARA_\infty - PLP$. More precisely, we have:

$$H_0 : \lambda \in P = \left\{ ab(t - \rho \sum_{j=0}^{N_{t^-}-1} (1-\rho)^j T_{N_{t^-}-j})^{b-1},\ t \geq 0,\ a > 0,\ b \geq 1,\ 0 \leq \rho \leq 1 \right\},$$

$H_1 : \lambda \notin P$.

The power of the test is calculated by simulating data from models that are not described by the null hypothesis H_0. For assessing the maintenance effect, we selected the following models: ARA_1, ARA_∞, BP, QR, and EGP. For the ARA models, we used $\rho \in \{0.2, 0.8\}$, representing low and strong maintenance effects, respectively. For the BP model, we selected $p \in \{0.2, 0.8\}$, and for the QR model, $q \in \{0.8, 0.9, 0.95\}$. When q approaches 1, the counting process behaves like a renewal process. For the EGP model, $q \in \{0.8, 0.9, 0.95\}$ and $g_i = i - 1$ or $g_i = \sqrt{i-1}$ for $i \in \{1, \ldots, n\}$.

For the initial hazard rate, we first selected a PLP with a scale parameter $a = 0.05$ and a shape parameter $b \in \{1.5, 2, 2.5, 3\}$. We also considered the initial hazard rate LLP with $a = -5$ and $b \in \{0.005, 0.01, 0.05, 0.1\}$. To generate data from the ARA_1, ARA_∞, QR, and EGP models, we used the VAM package developed at the LJK laboratory in Grenoble, available in the R programming language. For generating data from the BP model, we employed the algorithm introduced by Augustin and Pena [1].

For each model, we simulated $M = 1000$ datasets, each consisting of $n = 30$ time points. The parameters were estimated using the maximum likelihood method under the constraints $b \geq 1$ and $\rho \in [0, 1]$. These constraints ensure that the system is deteriorating and maintenance is effective. To evaluate the quantiles of the statistical distributions, we generated $L = 1000$ bootstrap samples.

4. Results and Discussion

For examining the Type 2 error, we first considered the hypothesis $ARA_\infty - PLP$ against the hypothesis $ARA_\infty - PLP$. Table 2 presents the empirical levels, which are close to the theoretical level $\alpha = 0.05$.

When data are simulated under the $ARA_1 - PLP$ model, the test powers are generally low. According to Table 3, for $\rho = 0.2$ and $b = \{1.5, 2\}$, the highest power is achieved by the test statistic VE_{mn}. When $\rho = 0.2$ and $b = 2.5$, the highest power is achieved by the test statistics VE_{mn} and VV_{mn}. When $\rho = 0.2$ and $b = 3$, the highest power is achieved by the test statistics VE_{mn} and KS_v. For $\rho = 0.8$ and $b = 1.5$, the highest power is achieved by the test statistic CvM_m. When $\rho = 0.8$ and $b = 2$, the highest power is achieved by the test statistic AD_m. When $\rho = 0.8$ and $b = 2.5$, the highest power is achieved

Table 2. Actual error rates of the tests for $ARA_\infty - PLP$ model.

ρ	0.2				0.8			
b	1.5	2	2.5	3	1.5	2	2.5	3
KS_m	0.059	0.057	0.061	0.049	0.056	0.055	0.051	0.051
CvM_m	0.050	0.035	0.052	0.046	0.064	0.056	0.049	0.065
AD_m	0.050	0.038	0.039	0.047	0.060	0.056	0.057	0.056
KS_v	0.061	0.066	0.055	0.067	0.053	0.048	0.045	0.052
Cvm_v	0.054	0.065	0.060	0.060	0.042	0.048	0.041	0.061
AD_v	0.049	0.071	0.059	0.062	0.042	0.042	0.043	0.052
VV_{mn}	0.055	0.059	0.050	0.051	0.062	0.051	.053	0.046
VE_{mn}	0.041	0.054	0.053	0.051	0.054	0.049	0.048	0.053

Table 3. Power of the tests for the $ARA_\infty - PLP$ model against the $ARA_1 - PLP$ Model.

ρ	0.2				0.8			
b	1.5	2	2.5	3	1.5	2	2.5	3
KS_m	0.035	0.026	0.026	0.019	0.128	0.086	0.013	0.040
CvM_m	0.033	0.031	0.030	0.019	**0.158**	0.113	0.020	0.040
AD_m	0.049	0.052	0.053	0.050	0.126	**0.114**	0.073	0.031
KS_v	0.064	0.057	0.044	**0.065**	0.041	0.054	0.092	0.130
Cvm_v	0.065	0.062	0.062	0.058	0.034	0.049	0.103	**0.153**
AD_v	0.065	0.055	0.059	0.048	0.041	0.038	0.077	0.109
VV_{mn}	0.067	0.064	**0.065**	0.064	0.060	0.061	0.106	0.064
VE_{mn}	**0.069**	**0.066**	**0.065**	**0.065**	0.061	0.063	**0.107**	0.061

by the test statistics VE_{mn}. When $\rho = 0.8$ and $b = 3$, the highest power is achieved by the test statistics CvM_v. Based on the stated results, it can be said that the test statistics based on varentropy perform better when the alternative hypothesis is $ARA_1 - PLP$.

When data are simulated under the $ARA_\infty - LLP$ model, the test powers are generally low. According to Table 4, for $\rho = 0.2$ and $b = 0.005$, the highest power is achieved by the test statistics VE_{mn}, VV_{mn}, AD_v, KS_v, and AD_m. When $\rho = 0.2$ and $b = \{0.01, 0.05\}$, the highest power is achieved by the test statistic AD_m. When $\rho = 0.2$ and $b = 0.1$, the highest power is achieved by the test statistics VE_{mn} and VV_{mn}. For $\rho = 0.8$ and $b = \{0.005, 0.01, 0.05\}$, the highest power is achieved by the test statistic VE_{mn}. When $\rho = 0.8$ and $b = 0.1$, the highest power is achieved by the test statistic AD_v. Based on the stated results, it can be said that the test statistics based on varentropy perform better when the alternative hypothesis is $ARA_\infty - LLP$.

TABLE 4. Power of the tests for the $ARA_\infty - PLP$ model against the $ARA_\infty - LLP$ model. ($a = -5$).

ρ	0.2				0.8			
b	0.005	0.01	0.05	0.1	0.005	0.01	0.05	0.1
KS_m	0.050	0.031	0.035	0.049	0.039	0.058	0.035	0.049
CvM_m	0.049	0.043	0.044	0.042	0.039	0.046	0.048	0.052
AD_m	**0.054**	**0.081**	**0.075**	0.065	0.039	0.068	0.055	0.048
KS_v	**0.054**	0.059	0.057	0.048	0.065	0.056	0.076	0.078
Cvm_v	0.052	0.051	0.061	0.051	0.091	0.060	0.084	0.092
AD_v	**0.054**	0.058	0.061	0.052	0.078	0.059	0.098	**0.109**
VV_{mn}	**0.054**	0.063	0.059	**0.068**	0.093	0.070	0.098	0.093
VE_{mn}	**0.054**	0.069	0.061	**0.068**	**0.095**	**0.072**	**0.099**	0.099

When data are simulated under the $QR - PLP$ model, the power of the tests

TABLE 5. Power of the tests for the $ARA_\infty - PLP$ model against the QR model.

b	1.5			2			2.5		
q	0.8	0.9	0.95	0.8	0.9	0.95	0.8	0.9	0.95
KS_m	0.582	0.963	0.556	0.854	**1**	0.637	0.952	**1**	0.672
CvM_m	0.504	0.975	0.435	0.864	**1**	0.641	0.963	**1**	0.658
AD_m	**0.691**	**0.994**	**0.634**	**0.939**	**1**	**0.741**	**0.989**	**1**	**0.789**
KS_v	0.039	0.043	0.027	0.045	0.030	0.027	0.050	0.029	0.019
Cvm_v	0.036	0.039	0.025	0.047	0.033	0.026	0.065	0.021	0.016
AD_v	0.027	0.027	0.025	0.043	0.035	0.034	0.052	0.024	0.017
VV_{mn}	0.056	0.063	0.053	0.094	0.065	0.064	0.098	0.065	0.066
VE_{mn}	0.054	0.062	0.055	0.096	0.066	0.064	0.095	0.065	0.064

generally increases as b increases, while keeping q constant. As q approaches 1, the model becomes more similar to the renewal process model, resulting in a decrease in the power of all tests. Tests based on martingale residuals are significantly more powerful than those based on entropy and the probability integral transform. Referring to Table 5, for $q = \{0.8, 0.9, 0.95\}$ and $b = \{1.5, 2, 2.5\}$, the test statistic AD_m, which is based on martingale residuals, demonstrates the highest power among the test statistics.

When data are simulated under the $EGP - PLP$ model, the power of the tests generally increases as b increases, while keeping q constant. As q approaches 1, the model becomes more similar to the renewal process model, leading to a decrease in the power of all tests. Tests based on martingale residuals are significantly more powerful than those based on entropy and the probability integral transform. Referring to Table 6, for $q = \{0.8, 0.9, 0.95\}$ and $b = \{1.5, 2, 2.5\}$,

TABLE 6. Power of the tests for the $ARA_\infty - PLP$ model against the EGP model.

b	1.5			2			2.5		
q	0.8	0.9	0.95	0.8	0.9	0.95	0.8	0.9	0.95
KS_m	0.265	0.107	0.069	0.428	0.115	0.075	0.537	0.161	0.080
CvM_m	0.237	0.064	0.049	0.401	0.073	0.049	0.520	0.116	0.047
AD_m	**0.421**	**0.114**	**0.072**	**0.691**	**0.181**	**0.089**	**0.855**	**0.266**	**0.082**
KS_v	0.045	0.048	0.061	0.059	0.043	0.048	0.047	0.038	0.042
Cvm_v	0.041	0.040	0.057	0.056	0.035	0.059	0.041	0.040	0.041
AD_v	0.040	0.041	0.050	0.047	0.032	0.052	0.043	0.043	0.042
VV_{mn}	0.065	0.065	0.066	0.056	0.054	0.060	0.077	0.062	0.059
VE_{mn}	0.067	0.062	0.065	0.055	0.054	0.061	0.081	0.064	0.060

the test statistic AD_m, which is based on martingale residuals, demonstrates the highest power among the test statistics.

When data are generated from the $BP - PLP$ model, we have the following

TABLE 7. Power of the tests for the $ARA_\infty - PLP$ model against the BP model.

p	0.2				0.8			
b	1.5	2	2.5	3	1.5	2	2.5	3
KS_m	0.035	0.018	0.031	0.029	0.030	0.040	0.038	0.032
CvM_m	0.033	0.027	0.026	0.031	0.031	0.045	0.034	0.034
AD_m	0.052	0.042	0.038	0.042	0.052	0.046	**0.040**	0.039
KS_v	0.054	0.059	0.059	0.070	**0.065**	0.061	0.034	0.042
Cvm_v	0.049	0.055	0.055	**0.076**	0.064	**0.063**	**0.040**	**0.063**
AD_v	**0.056**	0.058	0.059	0.071	0.064	0.059	0.036	0.060
VV_{mn}	**0.056**	0.058	0.059	0.071	0.064	0.059	0.036	0.060
VE_{mn}	0.054	**0.062**	**0.065**	0.068	0.065	**0.063**	0.038	0.061

results: If $\rho = 0.2$ and $b = 1.5$, the test statistics AD_v (based on the probability integral transform) and VV_{mn} (based on varentropy) have the highest power. If $\rho = 0.2$ and $b = \{2, 2.5\}$, the test statistic VE_{mn} has the highest power. If $\rho = 0.2$ and $b = 3$, the test statistic Cvm_v (based on the probability integral transform) has the highest power. If $\rho = 0.8$ and $b = 1.5$, the test statistic KS_v (based on the probability integral transform) has the highest power. If $\rho = 0.8$ and $b = 2$, the test statistics Cvm_v (based on the probability integral transform) and VE_{mn} (based on varentropy) have the highest power. If $\rho = 0.8$ and $b = 2.5$, the test statistics AD_m and Cvm_v have the highest power. If $\rho = 0.8$ and $b = 3$, the test statistic Cvm_v has the highest power. Based on the results and conditions presented, it appears that if data are generated from the $BP - PLP$ model, the tests based on the probability integral transform generally perform better in most conditions.

5. Application to Real-World Data

To validate the effectiveness of our proposed goodness-of-fit tests, we applied them to the AMC dataset, which includes 18 failure times for Ambassador vehicles used by the Ohio state government. This dataset has been widely used in reliability studies, particularly in Bayesian analysis, making it a suitable benchmark for evaluating the performance of our tests.
Sources of the data: These failure times are based on a dataset that has been studied in various reliability research articles, including the work of Ahn, Chae, and Clark [2], and more recently in Bayesian analyses by Guilda and Pulcini [19] and Corset et al. [10]. This set of data has been utilized as a standard example in numerous reliability modeling studies.
We examined the $ARA_\infty - PLP$ model using three distinct approaches: Martingale residuals, integral probability transforms, and varentropy-based test statistics. A total of eight different test statistics were applied to assess the model's goodness-of-fit. The estimated parameter values and corresponding p-values for each test statistic are presented in Table 8.

TABLE 8. Results of Varentropy-Based Tests on AMC Dataset

Parameter	**Estimated Value**
$\hat{a}$	2.11×10^{-11}
$\hat{b}$	3.58
$\hat{\rho}$	0.25
Test Statistic	**P-Value**
KS_m	0.264
CvM_m	0.145
AD_m	0.174
KS_v	0.198
Cvm_v	0.197
AD_v	0.187
VV_{mn}	0.129
VE_{mn}	0.134

Considering the results presented in Table 8, the p-values associated with all test statistics unequivocally support the acceptance of the null hypothesis, thereby lending strong credence to Model $ARA_\infty - PLP$ as the superior representation. Specifically, the absence of statistically significant evidence to reject the null hypothesis across all tests suggests that Model $ARA_\infty - PLP$ provides a more accurate and robust characterization of the underlying phenomenon compared to alternative formulations.

6. Conclusion

In this study, we evaluated the goodness-of-fit for the $ARA_\infty - PLP$ model using several alternative hypotheses, including ARA_1, $ARA_\infty - LLP$, QR, EGP, and BP. The null hypothesis was the $ARA_\infty - PLP$ model. Our key contribution lies in the introduction and application of varentropy-based test statistic (VE_{mn}) to detect deviations from imperfect maintenance models a methodological direction not previously addressed in the literature. Simulation results demonstrate that while VE_{mn} shows high sensitivity under specific alternatives such as ARA_1 and ARA_∞–LLP, it does not dominate across all scenarios. In contrast, the martingale-based AD_m test outperforms in settings like QR and EGP, and Cvm_v is optimal for the BP model. These findings emphasize the importance of tailoring the choice of test statistic to the expected pattern of deviation. In cases where no specific alternative is hypothesized, a combination of VE_{mn} and AD_m is recommended. The findings from this study highlight that selecting different tests based on the type of alternative hypothesis can significantly improve the detection of patterns and changes in data, thereby enhancing the performance of reliability analysis.

7. Aknowledgement

The authors are grateful to three anonymous referees and the associate editorfor providing some useful comments on an earlier version of this manuscript.

References

[1] Agustin, M. Z. N., & Peña, E. A. (1999). Order statistic properties, random generation, and goodness-of-fit testing for a minimal repair model. *Journal of the American Statistical Association*, **94**(445), 266–272. DOI: `https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10473842`

[2] Ahn, C. W., Chae, K. C., & Clark, G. M. (1998). Estimating parameters of the power-law process with two measures of time. *Journal of Quality Technology*, **30**(2), 127–132.`https://doi.org/10.1080/00224065.1998.11979831`

[3] Alizadeh, H., & Shafaei, M. (2024). Varentropy estimators applied to goodness of fit tests for the Gumbel distribution. *São Paulo Journal of Mathematical Sciences*, **2**, 1944–1962. `https://link.springer.com/article/10.1007/s40863-024-00456-1`

[4] Alizadeh Noughabi, H., & Shafaei Noughabi, M. (2023). Varentropy estimators with applications in testing uniformity. *Journal of Statistical Computation and Simulation*, **93**(15), 2582–2599. DOI: `https://doi.org/10.1080/00949655.2023.2196627`

[5] Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1993). Statistical models based on counting processes. New York, NY, USA: Springer-Verlag.

[6] Bordes, L., & Mercier, S. (2013). Extended geometric processes: Semiparametric estimation and application to reliability. *Journal of the Iranian Statistical Society*, **12**(1), 1–34.DOI: `https://hal.science/hal-00867031/`

[7] Brown, M., & Proschan, F. (1983). Imperfect repair. *Journal of Applied Probability*, **20**(4), 851–859. DOI: `https://doi.org/10.2307/3213596`

[8] Chauvel, C., Dauxois, J. Y., Doyen, L., & Gaudoin, O. (2016). Parametric bootstrap goodness-of-fit tests for imperfect maintenance models. *IEEE Transactions on Reliability*, **65**(3), 1343–1359. DOI: `https://doi.org/10.1109/TR.2016.2578938`

[9] Cook, R. J., & Lawless, J. F. (2007). The statistical analysis of recurrent events.DOI: https://link.springer.com/book/10.1007/978-0-387-69810-6

[10] Corset, F., Doyen, L., & Gaudoin, O. (2012). Bayesian analysis of ARA imperfect repair models. *Communications in Statistics - Theory and Methods*, **41**(21), 3915–3941. DOI: https://doi.org/10.1080/03610926.2012.698688

[11] Crétois, E., Gaudoin, O., & El Aroui, M. A. (1999). U-plot method for testing the goodness-of-fit of the power-law process. *Communications in Statistics - Theory and Methods*, **28**(7), 1731–1747. DOI: https://doi.org/10.1080/03610929908832382

[12] Crow, L. H. (1974). Reliability analysis for complex repairable systems. *Reliability and Biometry*, **13**(6), 379–410. DOI: 10.1007/978-94-011-3482-8-4

[13] D'Agostino, R., & Stephens, M. (1986). *Goodness-of-Fit Techniques.* Boca Raton, FL, USA: CRC Press.

[14] Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge, UK: Cambridge University Press.

[15] Doyen, L., & Gaudoin, O. (2004). Classes of imperfect repair models based on reduction of failure intensity or virtual age. *Reliability Engineering & System Safety*, **84**(1), 45–56. DOI: https://doi.org/10.1016/S0951-8320(03)00173-X

[16] Ebrahimi, N., Pflughoeft, K., & Soofi, E. S. (1994). Two measures of sample entropy. *Statistics & Probability Letters*, **20**(3), 225–234. DOI: https://doi.org/10.1016/0167-7152(94)90046-9

[17] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**(1), 1–26. DOI:https://link.springer.com/chapter/10.1007/978-1-4612-4380-9-41

[18] Gaudoin, O. (1998). CPIT goodness-of-fit tests for the power-law process. *Communications in Statistics - Theory and Methods*, **27**(1), 165–180. DOI: https://doi.org/10.1080/03610929808832658

[19] Guilda, M., & Pulcini, G. (2006). Bayesian analysis of repairable systems showing a bounded failure intensity. *Reliability Engineering & System Safety*, **91**, 828–838. DOI: 10.1016/j.ress.2005.08.008

[20] Leonenko, N., Sun, Y., & Taufer, E. (2024). Varentropy Estimation via Nearest Neighbor Graphs. *arXiv preprint*, arXiv:2402.09374. Available at: https://arxiv.org/abs/2402.09374

[21] Lindqvist, B. H., Elvebakk, G., & Heggland, K. (2003). The trend-renewal process for statistical analysis of repairable systems. *Technometrics*, **45**(1), 31–44. DOI: https://doi.org/10.1198/004017002188618671

[22] Lindqvist, B. H., & Rannestad, B. (2011). Monte Carlo exact goodness-of-fit tests for nonhomogeneous Poisson processes. *Applied Stochastic Models in Business and Industry*, **27**(3), 329–341. DOI: https://doi.org/10.1002/asmb.841

[23] Liu, Y., Huang, H. Z., & Zhang, X. (2011). A data-driven approach to selecting imperfect maintenance models. *IEEE Transactions on Reliability*, **61**(1), 101–112. DOI: https://doi.org/10.1109/TR.2011.2170252

[24] Park, W. J., & Kim, Y. G. (1992). Goodness-of-fit tests for the power-law process. *IEEE Transactions on Reliability*, **41**(1), 107–111. DOI: https://doi.org/10.1109/24.121461

[25] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3), 470-472. DOI: https://www.jstor.org/stable/2236692

[26] Saha, S., & Kayal, S. (2023). Weighted (residual) varentropy and its applications. *Journal of Computational and Applied Mathematics*, **442**, 115710. DOI: https://doi.org/10.1016/j.cam.2023.115710

[27] Stute, W., Manteiga, W. G., & Quindimil, M. P. (1993). Bootstrap-based goodness-of-fit tests. *Metrika*, **40**(1), 243–256. DOI: https://link.springer.com/article/10.1007/BF02613687

[28] Tibshirani, R. J., & Efron, B. (1993). *An introduction to the bootstrap*. Monographs on Statistics and Applied Probability, **57**(1), 1–436. DOI: 10.1007/978-1-4899-4541-9

[29] Vere-Jones, D., & Daley, D. J. (2008). *An Introduction to the Theory of Point Processes*. New York, NY, USA: Springer. DOI: 10.1007/978-0-387-49835-5

[30] Wang, H., & Pham, H. (1996). A quasi-renewal process and its applications in imperfect maintenance. *International Journal of Systems Science*, **27**(10), 1055–1062. DOI: https://doi.org/10.1080/00207729608929311

[31] Zhao, J., & Wang, J. (2005). A new goodness-of-fit test based on the Laplace statistic for a large class of NHPP models. *Communications in Statistics - Simulation and Computation*, **34**(3), 725–736. DOI: https://doi.org/10.1081/SAC-200068389

Fattaneh Nezampour
Orcid number: 0009-0008-5056-4409
Department of Statistics
Faculty of Mathematical Sciences and Statistics
University of Birjand
Birjand, Iran
Email address: fattaneh.nezmpour2021@birjand.ac.ir

Hadi Alizadeh Noughabi
Orcid number: 0000-0002-7515-1896
Department of Statistics
Faculty of Mathematical Sciences and Statistics
University of Birjand
Birjand, Iran
Email address: alizadehhadi@birjand.ac.ir

Majid Chahkandi
Orcid number: 0000-0003-0069-5407
Department of Statistics
Faculty of Mathematical Sciences and Statistics
University of Birjand
Birjand, Iran
Email address: mchahkandi@birjand.ac.ir