ELSEVIER

PII: S0951-8320(97)00099-9

Reliability Engineering and System Safety **60** (1998) 13–28
© 1998 Elsevier Science Limited
All rights reserved. Printed in Northern Ireland
0951–8320/98/$19.00

TTT-based tests for trend in repairable systems data

Jan Terje Kvaløy & Bo Henry Lindqvist

Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7034 Trondheim, Norway

(Received 25 September 1996; revised 24 January 1997; accepted 15 July 1997)

A major aspect of analysis of failure data for repairable systems is the testing for a possible trend in interfailure times. This paper reviews some important and popular graphical methods and tests for the nonhomogeneous Poisson process model. In particular, the total time on test (TTT) plot is considered, and trend tests based on the TTT-statistic are motivated and derived. In particular, a test based on the Anderson–Darling statistic is suggested. The tests are evaluated and compared in a simulation study, both with respect to the achievement of correct significance level and rejection power. The considered alternatives to 'no trend' are the log-linear, power law and a class of bathtub-shaped intensity functions. The simulation study involves single systems, as well as the case where several independent systems of the same kind are observed. © 1998 Elsevier Science Limited.

1 INTRODUCTION

For maintained and repairable systems it is important to detect possible changes in the pattern of failures. For example, reliability growth corresponds to times between failures becoming longer as time goes, whereas various aging effects lead to shorter interfailure times.

In practice, decisions concerning the failure pattern have to be based on observed failure data and statistical methods. It is the purpose of this paper to study methods for *trend* testing in failure data from repairable systems.

Fig. 1 illustrates the failure process observed for a single repairable system put into operation at time $t = 0$. The successive failure times $T_1, T_2, \ldots$ are often called the *arrival times*, while the times between failures, $X_1, X_2, \ldots$ are called the *inter-arrival times*. In Fig. 1, the repair times are set to 0, as will be done throughout the paper. This is a reasonable assumption if the repair times are negligible compared to the inter-arrival times, and can in any case be justified if we let the time scale be *operating time*. The failures are assumed to be point events occurring in instants of time.

We say that there is a *trend* in the pattern of failures if the inter-arrival times tend to alter in some systematic way, which means that the inter-arrival times are not identically distributed. The question we wish to answer is whether such an alteration is statistically significant or not.

A trend in the pattern of failures can be either monotonic or non-monotonic. In the case of a monotonic trend the system is said to be *improving* if the inter-arrival times tend to get longer (*a decreasing trend*), and the system is said to be *deteriorating* if the inter-arrival times tend to get shorter (*increasing trend*). Various types of non-monotonic trend can be present, in particular we mention the cases of *cyclic* trend and *bathtub* trend. If the inter-arrival times tend to alter in some cyclic way between longer and shorter, we have a cyclic trend. A pattern of failures is said to have a *bathtub* trend if there is a decreasing trend in the beginning, then a period with no apparent trend, and finally an increasing trend at the end of the observation interval.

Perhaps the most common type of trend in the pattern of failures from a mechanical system is increasing or bathtub-shaped trend. A typical example of a system with decreasing trend is a software system, and systems exposed to seasonal or other cyclic varying stresses might have a cyclic trend.

The following argument explains why bathtub trend often is plausible: When a system is new there are often 'infant illnesses' present, and as these are weeded out we observe a decreasing trend. After the 'infant illness' phase the system reaches the 'useful life' phase characterized by no trend. Finally, as the system gets old the 'wear-out' phase with increasing trend occurs.

Ascher and Feingold[1] point out that since we only can observe a process during a limited time interval, it is difficult to know whether the trend we have observed propagates into the future or not. For example, assume that we have detected a significant increasing trend in our data. Then we should bear in mind that we really do not know whether the increasing trend continues, or whether we, for instance, have observed a portion of a slow oscillation.

Finally, one should remember the fact that the choice of time scale influences the pattern of failures. Using calendar time, operating time, mileage or cumulative repair cost as the time scale for a car will probably give quite different patterns of failures.

The most widely used model for repairable systems is the nonhomogeneous Poisson process (NHPP). Not only is this a flexible and mathematically tractable model, but it can also be given a theoretical justification in many applications ('minimal repair'). For a description of other models we refer to Ascher and Feingold[1].

In this paper, we restrict attention to NHPP models and study various properties of mainly three different trend tests. Among these are the *Laplace test* and the *Military Handbook test* which are believed to be the most popular trend tests. These are tests constructed for the alternative hypotheses of *monotone* trend (i.e. either decreasing or increasing trend). In order to be able to detect other kinds of trend, e.g. bathtub-shaped trend, we suggest in addition a new trend test based on the total time on test (TTT) plot for NHPPs[2], using the Anderson–Darling statistic[3]. Closely related to this test is a test based on the Cramér–von Mises statistic[4,5], which will also be briefly considered. In fact, this was our original choice, while the successful use of the Anderson–Darling statistic came up as a suggestion from a referee. These tests should, by their construction, be able to detect a variety of departures from the 'no trend' situation. Necessarily, such a good 'overall' property should imply less power than the Laplace and Military Handbook tests when used against monotonic alternatives. A simulation study has been performed in order to figure out how much one loses in these cases by using the new test. On the other hand, the simulation study is also able to show how much better the new test is in the bathtub case.

An earlier power study, considering various tests for trend in NHPPs, has been conducted by Bain et al.[6]. They studied the power properties of a number of tests, including the Laplace and the Military Handbook test, against the one-sided hypothesis of an increasing intensity function. They conclude that the Laplace test and the Military Handbook test are the best tests in their study. Cohen and Sackrowitz[7] gave a theoretical explanation of these findings and show that the Laplace test and the Military Handbook test have desirable properties against monotonic alternatives. None of the other tests considered in Bain et al.[6] has been included in our simulation study.

If more than one process is observed, we might want to perform a simultaneous trend test using all the processes together. We discuss generalizations of the Laplace test and the Military Handbook test to the case of more than one process. Properties of the standard generalizations of the tests are compared by simulation to the properties of an alternative generalization based on a total time on test concept. A comparison with the Anderson–Darling-based test is also presented.

It should be stressed that within the NHPP framework of the present study, *no trend* will correspond to an assumption of homogeneous Poisson process (HPP) of the failure process. In practice, however, no trend may instead mean that failures follow a *renewal* process. Well known tests for the null hypothesis of a renewal process are[1] the *Mann* and the *Lewis–Robinson* tests. These tests are not included in this paper, however. In fact, when attention is restricted to NHPP models and the null hypothesis of HPP, these tests are outperformed by the tests specially constructed for NHPP models. For a comparison of the Mann test and the Lewis–Robinson test with the Laplace test and the Military Handbook test, we refer to Lindqvist et al.[8]. A major message in that paper is that the use of tests like the Laplace test and the Military Handbook test in non-NHPP situations may be strongly misleading and give invalid conclusions.

2 IDENTIFICATION OF TREND

2.1 The nonhomogeneous Poisson Process (NHPP)

We refer again to Fig. 1. Let $N(t)$ be the number of events (failures) occurring in the time interval $[0,t]$. The counting process $\{N(t), t \geq 0\}$ is called a *nonhomogeneous Poisson process* (NHPP) with *intensity function* $\lambda(t)$ if (1) $N(0) = 0$, (2) the number of events (failures) in disjoint time intervals are stochastically independent, (3) $P(N(t + \Delta t) - N(t) = 1) = \lambda(t)\Delta t + o(\Delta t)$ *as* $\Delta t \rightarrow 0$, and (4) $P(N(t + \Delta t) - N(t) \geq 2) = o(\Delta t)$ as $\Delta t \rightarrow 0$. (The last assumption assures that two or more events cannot take place simultaneously.)

It is well known that the intensity function $\lambda(t)$ coincides with the ROCOF (Rate of Occurrence of Failures) associated with the repairable system[1].

Further, letting the *cumulative intensity* be given by

$$\Lambda(t) = \int_o^t \lambda(u)\mathrm{d}u,$$

the number of events in an interval $[t, t + v]$, i.e. $N(t + v) - N(t)$, has a *Poisson* distribution with mean $\Lambda(t + v) - \Lambda(t)$.

Popular parameterizations of $\lambda(t)$ in applications to repairable systems are the *power law* intensity

$$\lambda(t) = \alpha\beta t^{\beta - 1}, \quad \alpha, \beta > 0, \quad t \geq 0$$

and the *log-linear* intensity,

$$\lambda(t) = \mathrm{e}^{\alpha + \beta t}, \quad -\infty < \alpha, \beta < \infty, \quad t \geq 0$$

The NHPP with constant intensity $\lambda(t) \equiv \lambda$ is called a *homogeneous Poisson process* (HPP). The HPP is a process with no *trend*, while the NHPP permits the modeling of trend via the intensity function $\lambda(t)$.

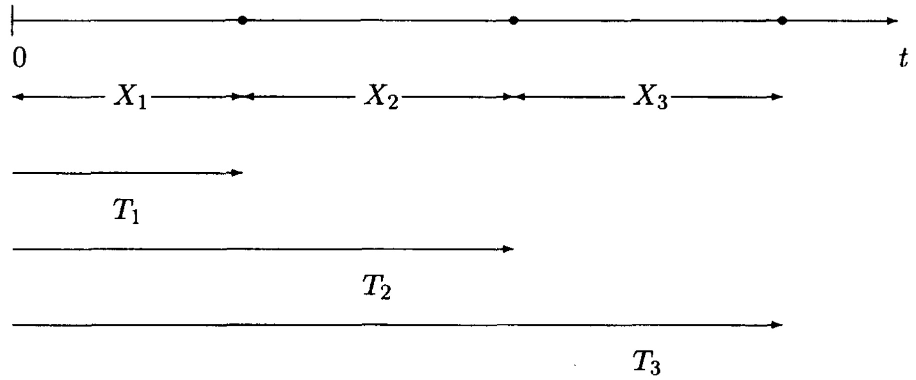

Fig. 1. Arrival times, T_i, and inter-arrival times, X_i.

2.2 The repairable system model

In the NHPP framework, the object is to decide whether a HPP or NHPP is the most relevant model. Both graphical and statistical methods are at hand. In this paper, we shall mainly pay attention to statistical methods, but some graphical methods will be considered first.

We shall assume that $m \geq 1$ independent systems, modeled by independent NHPPs with a common intensity function $\lambda(t)$, are observed. The ith system is observed in the time interval $(a_i,b_i]$ with n_i failures occurring at times T_{ij}, $j = 1,2,...,n_i$.

Note that the endpoints of the observation intervals $(a_i,b_i]$ may have different interpretations, according to the censoring schemes that are used. Two common censoring schemes are *time truncation* and*failure truncation*, defined in the following.

For time truncation the system is observed during a prespecified (operation) time. The observed number of failures is thus a random variable.

For failure truncation the system is observed until a prespecified number of failures has occurred. The length of the observation interval is now random.

Censoring strategies are important because data obtained by different censoring schemes are stochastically different. Hence, data must be treated differently depending on which censoring scheme is actually used.

2.3 Nelson–Aalen plot

A nonparametric estimate of the cumulative intensity function $\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$ is given by

$$\hat{\Lambda}(t) = \sum_{T_{ij} \leq t} \frac{1}{Y(T_{ij})}$$

where $Y(T_{ij})$ is the number of systems which are operating immediately before time T_{ij} and $\hat{\Lambda}(t) = 0$ for $t < \min_{ij} T_{ij}$. This estimator is studied, for example, in Andersen et al.[9].

The *Nelson–Aalen* plot is simply the plot of $\hat{\Lambda}(t)$ versus t, essentially a scatterplot of the points $(t_{ij}, \hat{\Lambda}(t_{ij}))$. If no trend is present, i.e. $\Lambda(t)$ is proportional to t, then the plot will tend to be nearly a straight line. Deviation from the straight line indicates some kind of trend.

If only one system is observed (m = 1), the Nelson–Aalen plot is simply a plot of cumulative number of failures versus operating time, which is the common way of plotting failure data from single repairable systems.

Note that the Nelson–Aalen plot may be misleading if all the a_i are greater than 0, or more generally if there are time intervals inside the interesting time domain with no processes under observation.

2.4 TTT plot

The TTT (Total Time on Test) plot is most well known as a graphical technique for data from nonrepairable systems[10]. A TTT plot for repairable systems data has been introduced by Barlow and Davis[2], based on the NHPP model. As above, assume that m independent NHPPs with common intensity function $\lambda(t)$ are observed, and assume that all observation intervals $(a_i,b_i]$ are contained in some time interval $(0,S]$. If n_i failures occurred in $(a_i, b_i]$, let $N = \Sigma_{i=1}^{m} n_i$. Let S_k denote the kth arrival time in the superposed process, i.e. S_k is an

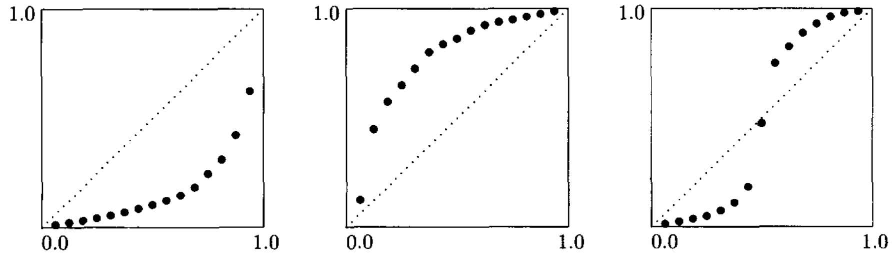

Fig. 2. Typical shape of TTT plot from NHPPs with decreasing, increasing and bathtub-shaped intensity function.

arrival time in one of the processes and $0 < S_1 \leq S_2 \leq \ldots \leq S_N \leq S$. Let $p(u)$ denote the number of processes under observation at time u and let $\mathcal{T}(t) = \int_0^t p(u)\mathrm{d}u$ denote the *total time on test* from time 0 to time t.

The (scaled) TTT plot for NHPPs is a plot of the scaled total time on test statistic

$$\frac{\mathcal{T}(S_k)}{\mathcal{T}(S)} = \frac{\int_0^{S_k} p(u)\mathrm{d}u}{\int_0^{S} p(u)\mathrm{d}u} \tag{1}$$

versus scaled failure number $k/N, k = 1, \ldots, N$. If $p(u) \equiv m$, then the scaled TTT plot is a scaled Nelson–Aalen plot with the axes interchanged. 'No trend' corresponds to a TTT plot located near the main diagonal. Various possible shapes of the TTT plot may give indications of the type of trend. Typical shapes of plots from NHPPs with decreasing, increasing and bathtub-shaped intensity functions are illustrated in Fig. 2.

The points $(k/N, \mathcal{T}(S_k)/\mathcal{T}(S))$ in the TTT plot are often connected with straight lines.

2.5 Laplace's test

Laplace's test for a single system is a test of the null hypothesis H_0: HPP, versus the alternative hypothesis H_1: NHPP with monotonic intensity function.

If a process is observed in the time interval $(a,b]$ and

$$\hat{n} = \begin{cases} n \text{ if the process is time truncated} \\ n-1 \text{ if the process is failure truncated} \end{cases}$$

then the test statistic

$$L = \frac{\Sigma_{j=1}^{\hat{n}} T_j - \frac{1}{2}\hat{n}(b+a)}{\sqrt{\frac{1}{12}\hat{n}(b-a)^2}} \tag{2}$$

is asymptotically standard normally distributed under the null hypothesis, i.e. when the underlying process is a HPP. The approximation with the normal distribution turns out to be very good, a rule of thumb says that $n \geq 3$ suffices. The test is optimal for the null hypothesis (HPP) when the alternative is a NHPP with log-linear intensity function, if the exact null distribution of L is used[6,11].

The intuitive idea behind this test is to compare the mean value of the failure times with the midpoint of the observation interval. Under the null hypothesis of a HPP, $T_1, \ldots, T_{\hat{n}}$ are the order statistic from a uniform distribution on $(a,b]$, and it follows that $\sum_{j=1}^{\hat{n}} T_j$ has expectation $\hat{n}(b+a)/2$ and variance $\hat{n}(b-a)^2/12$. This explains why eqn (2) is asymptotically standard normally distributed under the null hypothesis. If there is a monotone trend in the failure data, the mean value of the failure times will tend to deviate from the midpoint. The value of L indicates the direction of the trend. If $L < 0$ we have a decreasing trend and if $L > 0$ we have an increasing trend.

The generalization of the Laplace test to more than one process can be done in several ways. A straightforward generalization of eqn (2) if we have observations from m independent processes is

$$L_C = \frac{\Sigma_{i=1}^{m}\Sigma_{j=1}^{\hat{n}_i} T_{ij} - \Sigma_{i=1}^{m}\frac{1}{2}\hat{n}_i(b_i+a_i)}{\sqrt{\frac{1}{12}\Sigma_{i=1}^{m}\hat{n}_i(b_i-a_i)^2}} \tag{3}$$

The test based on this statistic is optimal for the null hypothesis of HPP, possibly with *different* intensities in each HPP, against the alternative of NHPPs with intensity $\lambda(t) = \mathrm{e}^{\alpha_i + \beta t}$ where β is common for all processes, while α_i is specific for each process[11]. We call this test the *combined* Laplace test.

2.5.1 TTT-based generalization of the Laplace test

A different way of generalizing the Laplace test to more than one process which we have not seen in the literature is to use the TTT-statistic [eqn (1)]. As in Section 2.4, we assume that we have observations from m independent NHPPs with *identical* intensity function $\lambda(t)$, and that each process has been observed in a subset of the time interval $(0,S]$. The superposed process has intensity function $\gamma(t) = \lambda(t)p(t)$, and under the null-hypothesis of no trend, i.e. $\lambda(t) \equiv \lambda$, it has cumulative intensity function $\Gamma(t) = \mathcal{T}(t)$. It follows from results on stochastic time changes [9], that the time transformed process $\Gamma(S_1), \ldots, \Gamma(S_N)$ is a HPP with intensity one. Consequently, the process $\mathcal{T}(S_1), \ldots, \mathcal{T}(S_N)$ is a HPP with intensity λ.

Thus, [12], if the process is failure truncated, $\mathcal{T}(S_1), \ldots, \mathcal{T}(S_{N-l})$ will have the same distribution as the order statistic corresponding to $N-1$ independent random variables uniformly distributed on the interval $(0, \mathcal{T}(S_N)]$. Similarly, if the process is time truncated, then conditional on the number of failures, N, $\mathcal{T}(S_1), \ldots, \mathcal{T}(S_N)$ will have the same distribution as the order statistic corresponding to N independent random variables uniformly distributed on the interval $(0, \mathcal{T}(S)]$. Notice that in the case of failure truncated processes $\mathcal{T}(S) = \mathcal{T}(S_N)$. Define $\hat{N}$ as

$$\hat{N} = \begin{cases} N \text{ if the processes are time truncated} \\ N-1 \text{ if the processes are failure truncated} \end{cases}$$

We conclude that (conditional on the total number of failures, N, in the case of time truncation) $(\mathcal{T}(S_k)/\mathcal{T}(S), k = 1, \ldots, \hat{N})$, is distributed as the order statistic of $\hat{N}$ uniform $(0,1)$ random variables. Hence,

$$L_T = \frac{\Sigma_{k=1}^{\hat{N}} \frac{\mathcal{T}(S_k)}{\mathcal{T}(S)} - \frac{1}{2}\hat{N}}{\sqrt{\frac{1}{12}\hat{N}}} \tag{4}$$

is asymptotically standard normally distributed under the null hypothesis that all processes are HPPs with *identical* intensities. We will call this the *TTT-based Laplace test.*

Notice that eqns (2)–(4) are identical in the case of only one process.

Recall that this is a test of a more restrictive null hypothesis than the combined Laplace test, which tests the null hypothesis of HPPs with possibly different intensities, while the TTT-based Laplace test tests the null hypothesis of HPPs with equal intensities. Hence, if the combined Laplace test rejects the null hypothesis we can conclude that we have a trend in our data, while if the TTT-based Laplace test rejects the null hypothesis we can only conclude that we do not have data from HPPs with identical intensities. Consequently, the TTT-based tests should be used only if we have reasons to believe that the systems are fairly homogeneous. These matters are further discussed in the simulation study.

2.6 Military Handbook test

This is another test constructed for the null hypothesis of a HPP versus the alternative of NHPP with monotone trend. The test statistic for a single system observed in the time interval $(a,b]$ is

$$M = 2\sum_{j=1}^{\hat{n}} \ln\left(\frac{b-a}{T_j - a}\right) \tag{5}$$

which, is (exactly) chi-square distributed with $2\hat{n}$ degrees of freedom under the null hypothesis[13].

The *one sided* Military Handbook test, which tests the null hypothesis of a HPP against the hypothesis of an *increasing* trend, is the optimal test when the alternative is a NHPP with increasing power law intensity function[6].

This test is based on the observation that if U is uniformly distributed on (0,1], then $-2\ln(U)$ will be chi-square distributed with two degrees of freedom. Thus, since $T_1,...,T_{\hat{n}}$ under the null hypothesis are distributed as the order statistic from a uniform distribution on $(a,b]$, this explains the null distribution of eqn (5). If we have a monotonically increasing trend, the test statistic M will become small compared to the null distribution, because then the failure times $T_1,...,T_{\hat{n}}$ will tend to be larger than the order statistics from the uniform distribution on $(a,b]$. Similarly, if we have a decreasing trend, M will be large compared to the null distribution.

The straightforward generalization of eqn (5) to more than one process is

$$M_C = 2\sum_{i=1}^{m}\sum_{j=1}^{\hat{n}_i} \ln\left(\frac{b_i - a_i}{T_{ij} - a_i}\right) \tag{6}$$

which is (exactly) chi-square distributed with $2q$ degrees of freedom, where $q = \Sigma_{i=1}^{m}\hat{n}_i$, under the null hypothesis of HPPs (possibly with different intensities). We call this the *combined* Military Handbook test.

Under the null hypothesis of independent HPPs with identical intensities, the test statistic

$$M_T = 2\sum_{k=1}^{\hat{N}} \ln\left(\frac{\mathcal{T}(S)}{\mathcal{T}(S_k)}\right) \tag{7}$$

is chi-square distributed with $2\hat{N}$ degrees of freedom, and we call this the *TTT-based Military Handbook test.*

3 A NEW TEST BASED ON THE TTT PLOT

The TTT plot was presented in Section 2 as an appropriate graphical method for visualizing trend in data from NHPPs. Moreover, we suggested TTT-based versions of the Laplace and Military Handbook tests. In this section, we shall demonstrate how a new statistical trend test based on the TTT plot can be obtained. As for the TTT-based versions of the Laplace test and the Military Handbook test, in the case of more than one process this is a test of the null hypothesis of identical HPPs.

One interesting feature of the test we shall derive is that it can be used to detect bathtub-shaped or other non-monotonic intensity functions.

Recall from Section 2.4 that in the 'no trend' case, the TTT plot tends to lie near the diagonal. Departures from this case, for example when the underlying intensity function is monotonically increasing, decreasing or bathtub shaped, will tend to increase the area between the TTT plot and the diagonal. This suggests that a test statistic for the null hypothesis of a HPP could be based on some function related to this area.

As in Section 2.5.1, we assume that we have N observations $S_1,...,S_N$ from $m \geq 1$ processes observed on $[0,S]$, where S_i is the ith arrival time in the superposed process and $S_N \leq S$ (see Section 2.4).

Recall the derivation of the TTT-based Laplace test where we show that $(\mathcal{T}(S_k)/\mathcal{T}(S), k = 1,...,\hat{N})$, has the same distribution as the order statistic based on $\hat{N}$ i.i.d. *uniform* (0,1) variables. We conclude that the empirical distribution function of $\mathcal{T}(S_k)/\mathcal{T}(S), k = 1,...,\hat{N}$, which is

$$F_N(v) = \frac{k-1}{\hat{N}}, \quad \frac{\mathcal{T}(S_{k-1})}{\mathrm{T}(S)} \leq v < \frac{\mathcal{T}(S_k)}{\mathrm{T}(S)},$$

approaches the cumulative distribution function of the *uniform* (0,1) distribution as N increases. Next define the process

$$C_N(v) = \sqrt{\hat{N}}(F_N(v) - v), \ 0 \leq v \leq 1$$

By its definition, C_N defines a measure of the distance between the TTT plot and the diagonal. Andersen et al.[9] suggest the signed area between the TTT plot and the diagonal, $\int_0^1 C_N(v)\mathrm{d}v$, as a measure of departure from the HPP assumption. They arrive at an easily evaluated asymptotically normal test statistic, but since they are using the signed area their test has the serious drawback of having very low power against non-monotonic trends that places area on both sides of the diagonal in the TTT plot, e.g. bathtub-shaped trend. Another test proposed by Andersen et al.[9] is to use the Kolmogorov statistic $\max_{v\epsilon[0,1]}|C_N(v)|$ as a test statistic. This test should be able to detect both monotonic and non-monotonic trend, but the convergence of the test statistic to its asymptotic distribution seems to be very slow. In our simulation studies this test achieved a far too low actual level for the moderate sample sizes considered (even with as many as 500 failures it did only achieve an actual level of about 4% when the nominal level was 5%). Thus, critical values of this test should be computed by other means than the asymptotic distribution. This test is not further discussed in the presentation of the simulation study.

In order to get a test statistic with fairly good power both against monotonic and non-monotonic trends, inspired by Aarset[14], we shall first propose the test statistic

$$W_N = \int_0^1 C_N^2(v)\mathrm{d}v \tag{8}$$

which (except for the squaring) can be viewed as a measure of the (unsigned) area between the TTT plot and the diagonal. Since $F(v) = v$ is the cumulative distribution function of the *uniform* (0,1) distribution and $F_N(v)$ is the empirical distribution function for data which under the null hypothesis are *uniform* (0,1) distributed, eqn (8) is a Cramér–von Mises statistic[4,5], and we call the test based on this statistic the *Cramér–von Mises test for trend.* A possible improvement of the Cramér–von Mises test is the Anderson–Darling test[3] given by the test statistic

$$A_N = \int_0^1 C_N^2(v)\frac{1}{v(1-v)}\mathrm{d}v \tag{9}$$

We shall call the test based on this statistic the *Anderson–Darling test for trend.* The difference between the two statistics (8) and (9) is the weight function $1/v(1-v)$ in the latter, which has the effect of giving greater importance to observations in the tails, counteracting the fact that $F_N(v) - v$ approaches zero in each tail. The asymptotic distributions of eqn (8) and (9) were derived by Anderson and Darling [3]. A nice practical review of these and related tests, containing percentage points for the asymptotic distributions, was given by Stephens[15]. An explicit expression for the limiting cumulative distribution function (9) was given by Anderson and Darling[16],

$$\begin{aligned}P(A \le a) = \frac{\sqrt{2}}{a}\sum_{j=0}^{\infty}&\frac{(-1)^j\Gamma(j+\frac{1}{2})(4j+1)}{j!}\\ &\times \mathrm{e}^{-[(4j+1)^2\pi^2]/(8a)}\\ &\times \int_0^{\infty}\mathrm{e}^{a/[8(w^2+1)]-[(4_j+1)^2\pi^2w^2]/(8a)}\mathrm{d}w.\end{aligned}$$

This is a good approximation of the exact distribution even for very small samples. Using the asymptotic distribution, on a 5% level the null hypothesis of no trend is rejected if $A_N \ge A_{0.05} = 2.492$.

For practical implementations, straightforward calculations from eqn (8) and (9) lead to the test statistics,

$$W_N = \sum_{i=1}^{\hat{N}}\left[\frac{\mathcal{T}(Si)}{\mathcal{T}(S)} - \frac{2i-1}{2\hat{N}}\right]^2 + \frac{1}{12\hat{N}}$$

for the Cramér–von Mises test for trend and

$$\begin{aligned}A_N = -\frac{1}{\hat{N}}[\sum_{i=1}^{\hat{N}}(2i-1)(\ln(\frac{\mathcal{T}(Si)}{\mathcal{T}(S)})\\ +\ln(1-\frac{\mathcal{T}(S_{\hat{N}+1-i})}{\mathcal{T}(S)}))] - \hat{N}\end{aligned} \tag{10}$$

for the Anderson–Darling test for trend.

In our simulation study, the Anderson–Darling trend test essentially seems to behave uniformly better than the Cramér–von Mises trend test, and thus with a few exceptions only, the Anderson–Darling trend test is the only one mentioned in the simulation study. More specifically, against monotonic trend there are only minor differences in power between the two tests, probably explained by a bit too low actual level for the Cramér–von Mises trend test on moderate sample sizes, while against bathtub-type trend the Anderson–Darling trend test represents a considerable improvement over the Cramér–von Mises trend test, as we would expect since the former puts more weight to observations near the endpoints.

The Anderson–Darling trend test should be used in close connection with visual inspection of the TTT plot or a Nelson–Aalen plot. Situations where the null hypothesis is rejected but the intensity function is neither increasing, decreasing nor bathtub shaped can be thought of; for instance cyclic trend or other non-monotonic trends. Although it may be difficult to classify the type of trend in certain situations, at least we know after rejection that if the NHPP assumption is valid, and only one or identically distributed processes are observed there is some kind of departure from the 'no trend' situation. The TTT plot or the Nelson–Aalen plot then gives a qualitative description of this departure. If more than one process is observed, individual plots for each process should be made as well. These plots can both visualize the individual trend in each process, and possibly help to distinguish situations with real trend from situations with heterogeneous HPPs.

4 SIMULATION STUDY

In order to compare the properties of the new tests to the commonly used Laplace test and the Military Handbook test, a simulation study has been carried out. The simulation code is written in C, and the C-function **random()** is used as random number generator, with the generation of seeds connected to the system clock as well as iteration number.

4.1 Single systems

We consider first the case $m = 1$ when one single system is observed. In this case, the TTT-based versions of the Laplace and Military Handbook test are not considered as they are exactly equal to the original tests. Thus, we consider the ordinary Laplace test, eqn (2), the ordinary

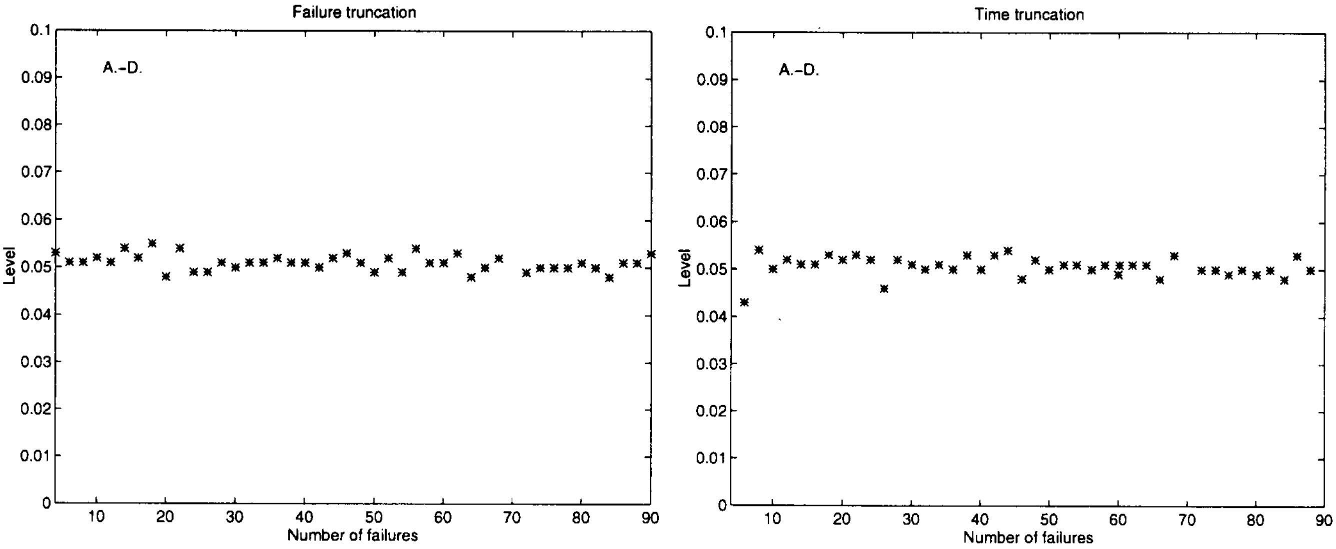

Fig. 3. Simulations of a HPP with intensity 1. Failure truncation with the number of failures ranging from four to 90, and time truncation with the expected number of failures ranging from four to 90. Twenty thousand replications for each number of failures.

Military Handbook test, eqn (5), and the Anderson–Darling trend test, eqn (10).

For each simulated process we consider the time from $t = 0$. When simulating a failure truncated process, the process is run until the prespecified number of failures has occurred. For time truncated processes, the truncation time is determined so that the *expected* number of failures equals the prespecified 'number of failures' (the actual number of failures will vary, and now and then it will even happen that *no* failures has occurred before the truncation time. This happened very rarely though, and at these few occasions we simply simulated a new process). For each run, a choice is made for the $\lambda(t)$-function.

For each given choice of truncation mechanism, number of failures and intensity function, the aim is to estimate the probability of rejection of the null hypothesis of 'no trend' (i.e. HPP) for each of the three tests. This is accomplished by simulating (usually) 10,000 processes with the same setup, and recording the relative number of rejections (absolute number divided by 10,000) for each test.

If the simulated process is a HPP, the estimated rejection probability is called the *actual level*, which we interpret as the true probability of incorrect rejection. This may be different from the *nominal level*, which is the desired level which is used for determining critical values. The difference between actual and nominal level results from the fact that in most cases critical values are computed from asymptotic *approximations* rather than the exact distributions. The nominal significance level has been set to 5% throughout the study.

If the simulated process is not a HPP, there is a kind of trend. In this case, the estimated rejection probability is called the *power* of the test. The power as a function of the trend parameter is called the *power function* of the test.

The estimated rejection probabilities of course become more accurate as the number of replications increase. Indeed, if we let $\hat{p}$ denote the estimated rejection probability, then the standard deviation is $\sqrt{\hat{p}(1-\hat{p})/n}$, which is bounded above by $1/(2\sqrt{n})$. Thus, if $n = 10{,}000$, the standard deviation is no larger than 0.005.

The following abbreviations are used in the graphs: *Laplace*—The Laplace test, eqn (2), *Mil-hbk*—The Military Handbook test, eqn (5), *A–D*—The Anderson–Darling trend test, eqn (10).

Note that the displayed curves, except for Fig. 3, are spline interpolations connecting the observation points.

4.1.1 Level properties

Fig. 3 displays the actual level of the Anderson–Darling trend test for a varying number of failures and data simulated from both a failure truncated and time truncated HPP.

The graphs show that the Anderson–Darling trend test achieves the correct actual level of 5% with satisfying accuracy even for very small sample sizes, and for both failure and time truncated processes.

As regards the Laplace and the Military Handbook tests, the latter keeps an exact level for HPPs, while the former is approximately exact even for as low as three failures.

4.1.2 Power properties

4.1.2.1 Log-linear intensity function. In this case $\lambda(t) = e^{\alpha+\beta t}$, and $\Lambda(t) = e^{\alpha}(e^{\beta t} - 1)/\beta$. Thus, if $\beta < 0$ we have a finite limit $\lim_{t\to\infty}\Lambda(t) = e^{\alpha}/|\beta|$. This means that with probability 1, only a finite number of failures will occur during infinite time. In order to avoid difficulties arising from this, we have simulated only the increasing case, $\beta > 0$.

Figure 4 presents estimated power functions with data simulated from failure and time truncated NHPPs, respectively, with log-linear intensity function and 15 (expected) failures. We observe that the Laplace test is the most powerful test, as expected, since it is the optimal test in this situation. The Anderson–Darling trend test has larger power than the Military Handbook and is only slightly less powerful than the Laplace test. We also observe that the power functions are quite similar in the failure truncated and time truncated cases, with slightly higher power for time truncated processes.

When 35 failures are simulated, see Fig. 5, the same picture as in Fig. 4 is seen, but the differences between the tests are smaller, and the power functions are of course steeper.

4.1.2.2 Power law intensity function. When data are simulated from a NHPP with power law intensity function, $\lambda(t) = \alpha\beta t^{\beta-1}$, both data with decreasing $(0<\beta<1)$ and increasing $(\beta > 1)$ trend can be simulated. Figure 6 displays graphs of estimated power functions with data simulated from NHPPs with power law intensity function and 15 failures (expected number for time truncated process). We observe that the Military Handbook test is the test with largest power, as expected, since this test is the optimal test in this case. The Laplace test is a bit stronger than the Anderson–Darling trend test against increasing trend, while the Anderson–Darling trend test is stronger than the Laplace test against decreasing trend. Again there are no big differences between the failure and time truncated cases.

When˙35 failures are simulated, see Fig. 7, the same effects are observed, the only difference being that the power functions are steeper and the Anderson–Darling trend test and Laplace test are almost identical against increasing trend.

4.1.2.3 Bathtub-shaped intensity function. A simple example of a bathtub-shaped intensity function is given in Fig. 8. The intensity function has been divided into three phases, I, II and III, which may be identified as the 'infant illness' phase, 'useful life' phase and 'wear out' phase, respectively.

Data have been simulated from NHPPs with 12 different bathtub-shaped intensity functions, described in Table 1 by specifying the expected number of failures in each phase and the slope of the intensity function in phase I and phase III. Note that the expected number of failures in each phase are easily found to be, in phase I, $\Lambda(t_1) = t_1(b+1)/2$, in phase II, $\Lambda(t_2) - \Lambda(t_1) = t_2 - t_1$ and in phase III, $\Lambda(\tau)-\Lambda(t_2)=(\tau-t_2)(c+1)/2$.

Both time and failure truncated processes are simulated. The time truncated processes are truncated at time τ, while the number of failures simulated in the failure truncated process equals the sum of expected number of failures in each phase. If the last simulated arrival time(s) are larger than τ in the case of a failure truncated process, the intensity function in phase III is extended beyond τ. Results of the simulations are given in Tables 2 and 3. For the sake of illustration of the difference between the Cramér–von Mises trend test and the Anderson–Darling trend test, the Cramér–von Mises trend test has also been included in these simulations.

The Anderson–Darling trend test is clearly the best test as it is the most powerful test in all of these situations. The Cramér–von Mises trend test is generally the second most powerful test, but the Anderson–Darling trend test is definitely better. The other tests are quite powerful in some of the considered situations, but have very low power in other situations, making them unsuitable as tests against bathtub trend.

4.2 Several processes

Now we proceed to consider the case $m > 1$. There is of course a huge number of situations to consider, with varying numbers of observed processes, various censoring schemes, different kinds of heterogeneities, etc. This is by no means a thorough study of the $m > 1$ case, only a few situations are considered to get a first feeling on how the various tests behave in this case.

For monotonic trends we have simulated data with two different designs. With the first design we have simulated data from $m = 5$ independent processes which are started at time $t = 0$, but are observed over different, but partially overlapping, time intervals. The observation interval for each process has been chosen such that the expected number of failures in each interval equals five, but with different starting points for each interval. Hence, the expected total number of failures is 25. A symbolic illustration is given in Fig. 9. The position and length of each interval on the *time axis* will of course depend on intensities in each process.

The second design used to simulate data with monotonic trend also simulate data from $m = 5$ independent processes, but this time we have observed all the processes from time $t=0$. In the first process we let the length of the observation interval vary, corresponding to a varying expected number of failures in this interval. In the other four processes the length of the observation intervals has been chosen such that the expected number of failures in each process equals three. A symbolic illustration for the case when the expected number of failures in the first interval, $n1$, equals 12 is given in Fig. 10.

4.2.1 No heterogeneities

First we consider the case where each process is simulated from NHPPs with identical intensity functions $\lambda(t)$ for $t \geq 0$. In the first example we have simulated data using the first design mentioned above, i.e. the ith process is observed in an interval $[a_i,b_i]$ where a_i and b_i are chosen such that the expected number of failures before and in the ith interval is as indicated in Fig. 9. Data are simulated from both a log-linear and a power law intensity function, and the results are shown in Fig. 11. The abbreviations used now are, *Laplace*—The combined Laplace test, eqn (3), *Laplace-TTT*—The TTT-based Laplace test, eqn (4), *Mil-hbk*—The combined Military Handbook test, eqn

(6), *Mil-hbk-TTT*—The TTT-based Military Handbook test, eqn (7), *A–D*—The Anderson–Darling trend test, eqn (10).

We realize that the TTT-based tests (the Anderson–Darling trend test, TTT-based Laplace test and TTT-based Military Handbook test), are far more powerful than the combined Laplace test and Military Handbook test. The relationships between the TTT-based tests in the power law case are the same as the relationships between the Anderson–Darling trend test, Laplace test and Military Handbook test shown in Figs 6 and 7 in the one process case, and for the log-linear case the relationship is the same as the relationship seen in Figs 4 and 5. Also notice that, in contradiction to the picture on Figs 6 and 7, the Laplace test has more power against increasing power law trend than the Military Handbook test in the studied case with five processes. Additional simulations indicate that this seems to happen when data are simulated from more than about three processes.

One possible explanation of the success of the TTT-based versions of the Laplace test and the Military Handbook test in the above example is that while the TTT-based tests superpose all the five processes to one process with a monotonic trend, the combined Laplace test and combined Military Handbook test search for trend within each single system (which they have to do since they allow for heterogeneities between the processes). It might also be argued that the design used is a bit artificial. The alternative design, Fig. 10, is probably closer to a typical practical situation, and if we are varying the expected number of failures in the first process we can get a picture of how the TTT-based tests behave compared to the other tests as the first process is more or less dominating. We have chosen two power-law intensities, $\lambda(t) = t^{\beta-1}$ with respectively $\beta = 1.5$ and $\beta = 0.75$, i.e. respectively, increasing and decreasing trend. We let the expected number of failures in the first process, $n1$(see Fig. 10), vary from three to 120, and rejection power as a function of this expected number is displayed in Fig. 12. We see that for $n1 = 3$, i.e. when all the five processes have been observed over the same interval, the TTT-based and combined versions of the Laplace test and the Military Handbook test coincide [which is easily seen from eqn (3) and (4), and (6) and (7)]. Otherwise, the TTT-based tests are stronger. Even when the expected number of failures in the first process is much larger than the total expected number of failures in the four other processes (12), the TTT-based tests are much stronger. The reason why the TTT-based tests are far better than the other tests is probably that these tests, making the stronger assumption of equal intensities, are using the information in the four processes observed over a short time interval more efficiently than the combined tests. As the combined tests have to allow for heterogeneities in the various processes they cannot extract the same amount of information from the four processes observed over a short time as the TTT-based test which superposes all observations into one process.

Data from several NHPPs with the same bathtub-shaped intensity functions are simulated as well. Five processes are simulated and the 12 intensity functions described in Table 1 are used. However, now the processes are not observed over the entire time interval $[0, \tau]$. One process is observed only in phase I (see Fig. 8), one observed only in phase II, one only in phase III, one observed in phases I and II, and the last process is observed in phases II and III. The results are presented in Table 4. Once again, the TTT-based tests are more powerful than the combined Laplace and Military Handbook tests, but the TTT-based Laplace test is not very much better than the combined Laplace test. The Anderson–Darling trend test is the most powerful test in all cases, while the TTT-based Military Handbook test is generally the second best test.

4.2.2 Heterogeneities

The results from the previous subsection seem to indicate that the combined Laplace and Military Handbook tests are completely outperformed by the TTT-based tests, but we should keep in mind that the latter tests are constructed for the more restrictive null hypothesis of HPPs with identical intensity functions, while the combined Laplace test and combined Military Handbook test allow heterogeneities in the HPPs under the null hypothesis of no trend. To study the effect of such heterogeneities, data are simulated from five processes with the observation interval for each process chosen according to the design in Fig. 9. NHPPs with intensity functions $\alpha_i\beta t^{\beta-1}$ and $e^{\alpha_i+\beta t}$ are simulated, where the α_is are *varying* from process to process, while the βs are common for all processes. The results and choice of parameter values are given in Fig. 13. Notice that with our choices of parameter values, the relative heterogeneities in the log-linear intensity functions are greater than the relative heterogeneities in the power law intensity functions (i.e. the range of intensity values in the no trend case is about two times greater with the choice of parameter values done in the log-linear case compared to the power law case).

The picture in Fig. 13 clearly displays the problem with the TTT-based tests; they are not constructed to allow heterogeneous intensity functions. In the power law case the tests achieve an actual level exceeding 0.1. In the log-linear case the TTT-based tests achieve a too *low* level. This difference in level behavior between the two cases is of course only explained by the difference in heterogeneities. The combined Laplace and Military Handbook tests behave reasonably well. The difference in level behavior of the TTT-based tests seen in the two cases, with less severe level properties in the most heterogeneous case, is somewhat unexpected. To investigate this further, some additional simulations of HPPs with unequal intensity functions were performed. The observation intervals indicated in Fig. 9 are still used. Hence, the expected number of failures in each simulation is 25, and the results are given in Table 5.

We see that the level behavior of the TTT-based tests are somewhat unpredictable, and often they achieve a far too

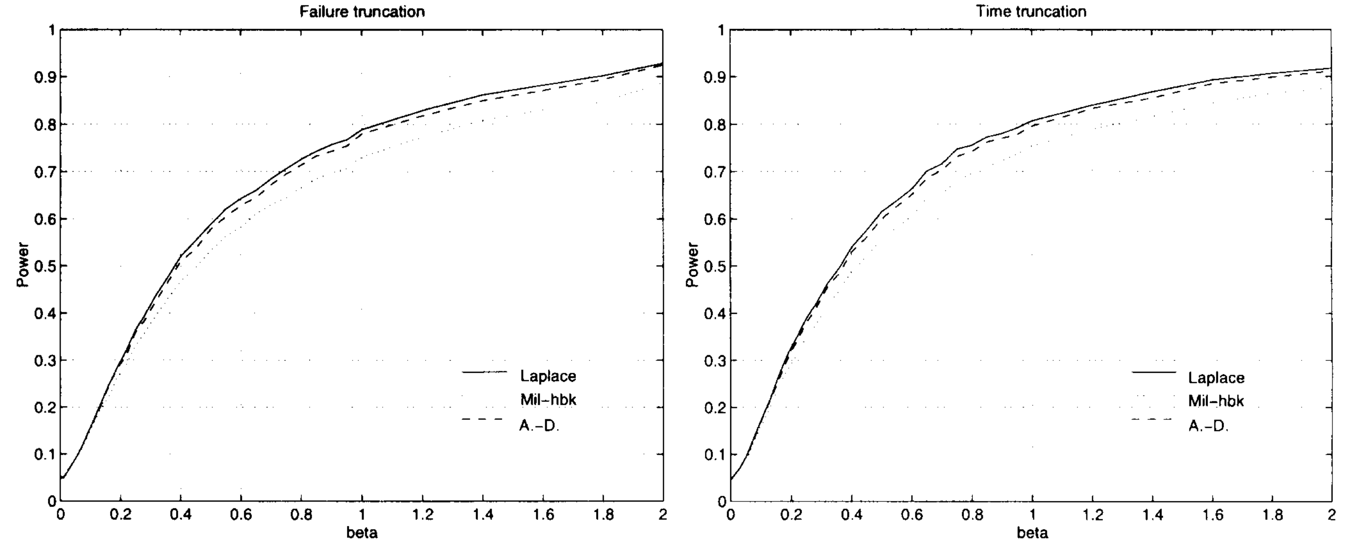

Fig. 4. Simulations of a NHPP with intensity function $\lambda(t) = e^{\beta t}$, where β = beta = {0,0.01,0.03,...0.40,0.45...,1,1.2,...,2.0}. Ten thousand replications for each beta-value. Failure truncation with 15 failures and time truncation with expected number of failures in each simulation equal to 15.

high actual level. Thus, if we suspect considerable heterogeneities in data from several processes, these tests should not be used. But we should also notice that in the case of only moderate heterogeneities, their level properties are tolerable. In fact, in these situations it *could* be favorable to use the TTT-based tests due to their far better power properties. In such cases the TTT-based Military Handbook test and Anderson-Darling trend test seem to have the best overall properties.

5 CASE STUDY

This is a simple example to demonstrate the use of the trend tests. A dataset presented by Barlow and Davis[2] is used. The data are failure truncated failure time data for 22 tractor engines. The pattern of failure times and a TTT plot for these failure times are displayed in Fig. 14.

The TTT plot clearly indicates an increasing trend in the data. Results of the statistical trend tests are presented in Table 6. All the trend tests reject the null hypothesis on a 5% level, hence we may safely conclude that there is an increasing failure trend in the tractor data. We noted a fairly large relative difference between the p-values of the TTT-based tests and the other tests, which could indicate a possible heterogeneity between tractors. However, even under possible heterogeneities among the tractors the HPP assumption is rejected since

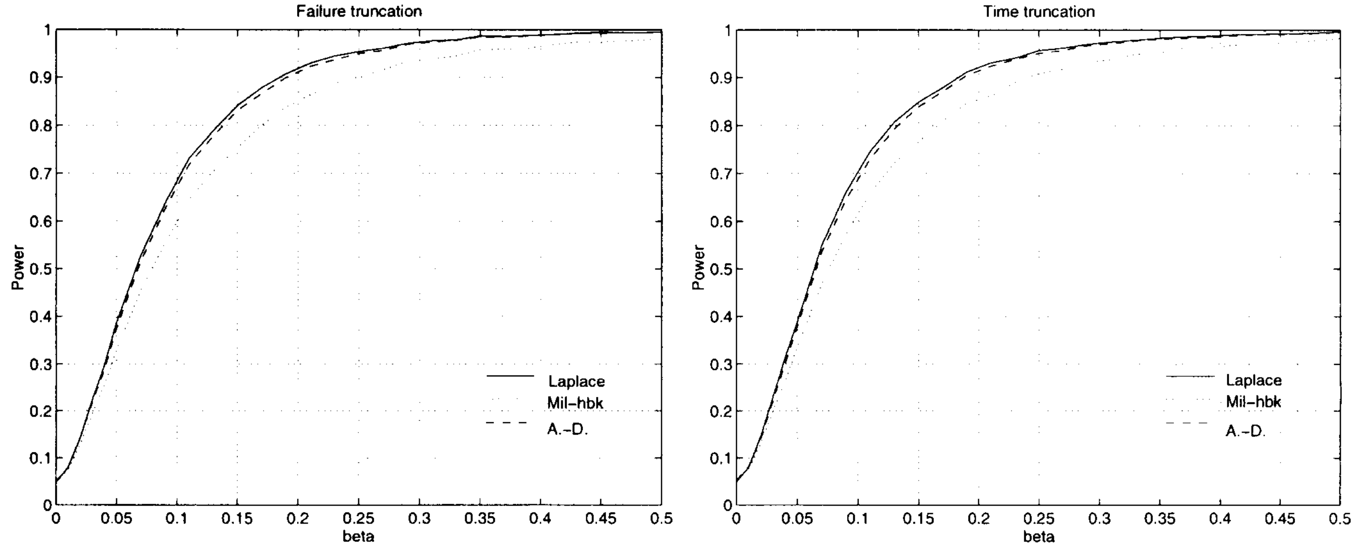

Fig. 5. Simulations of a NHPP with intensity function $\lambda(t) = e^{\beta t}$, where β = beta = {0,0.01,0.03,...,0.50}. Ten thousand replications for each beta-value. Failure truncation with 35 failures and time truncation with expected number of failures in each simulation 35.

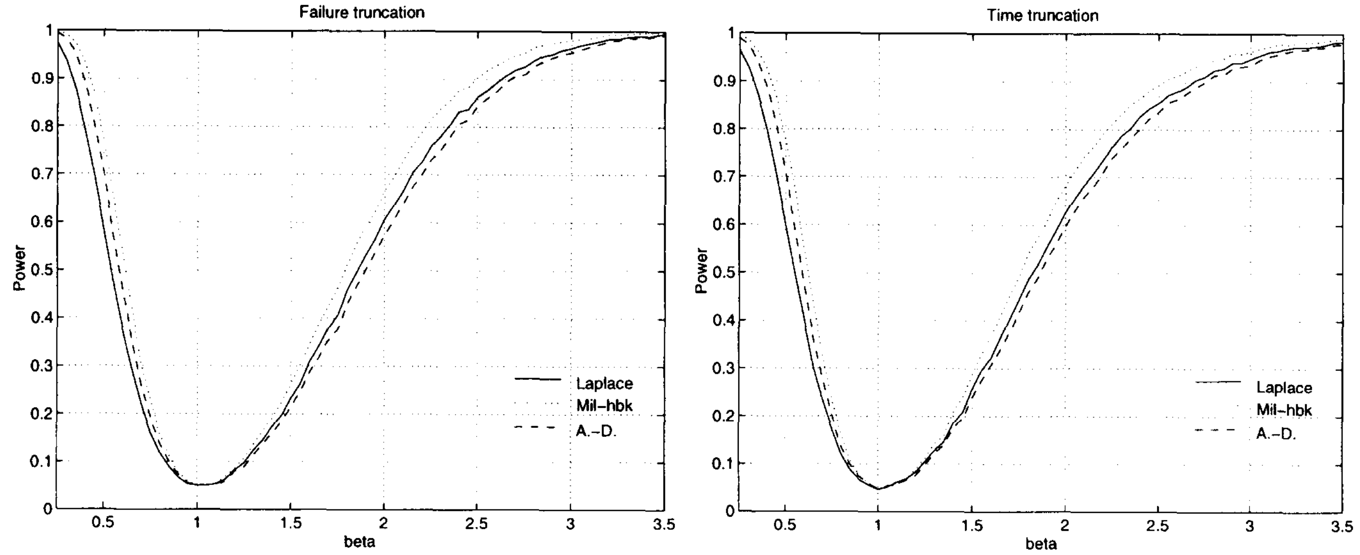

Fig. 6. Simulations of a NHPP with intensity function $\lambda(t) = t^{\beta-1}$, where β = beta = {0.25,0.30,...,3.5}. Ten thousand replications for each beta-value. Failure truncation with 15 failures and time truncation with expected number of failures in each simulation 15.

the combined Laplace test and the combined Military Handbook test reject the null hypothesis. Moreover, according to investigations made by Elvebakk[17], there seem to be no indications of heterogeneity in the tractor engine data.

With a maximal number of six observed failures for the single tractors, it is obviously difficult to analyze them individually. In fact, the null hypothesis of no trend would be rejected only for tractor number 14. But assuming a common trend for all the 22 tractor engines the conclusion is a significantly increasing trend.

6 CONCLUSIONS

6.1 Description of each test

The properties of each test are summarized below.

6.1.1 The Laplace test

The Laplace test is for single processes the most powerful test against NHPP with log-linear intensity function. In our study it was the least powerful test against NHPP with decreasing power law intensity and it is only slightly more

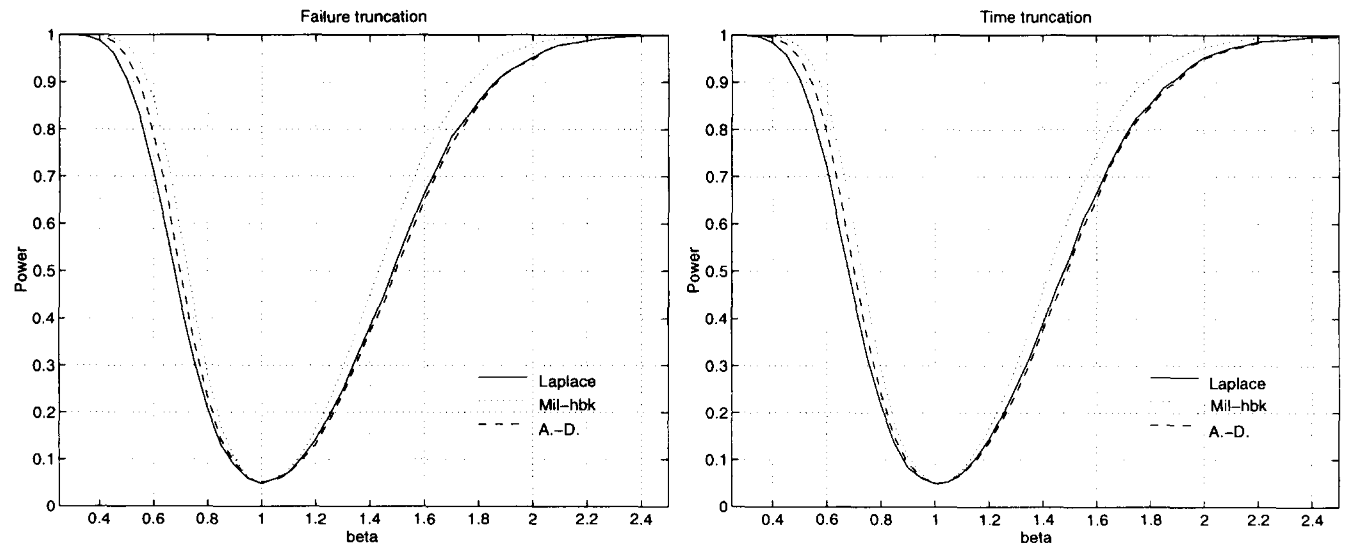

Fig. 7. Simulations of a NHPP($t^{\beta-1}$), where β = beta = {0.30,0.35,...,2.5}. Ten thousand replications for each beta-value. Failure truncation with 35 failures and time truncation with expected number of failures in each simulation 35.

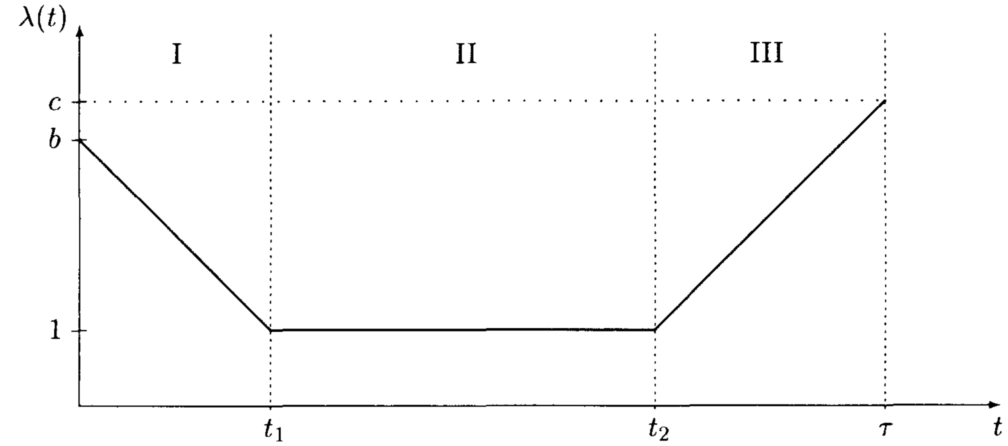

Fig. 8. Example of bathtub-shaped intensity function.

powerful than the Anderson-Darling trend test against increasing power law intensity. It has in general low power against NHPPs with bathtub-shaped intensity functions.

If more than one process is observed, the combined Laplace test is not particularly powerful against bathtub-shaped and decreasing power law intensity functions, but it seems to be more powerful than the combined Military Handbook test both against increasing power law and log-linear intensity functions if more than three processes are observed.

6.1.2 The TTT-based Laplace test

The TTT-based Laplace test generally has the same properties compared to the other TTT-based tests as the original test has in the one process case, but it has poor properties against bathtub-shaped intensity functions.

6.1.3 The Military Handbook test

The Military Handbook test is for single processes the most powerful test against NHPPs with power law intensity function. It is the least powerful test against NHPPs with log-linear intensity functions, and has generally low power against bathtub-shaped trend. If more than one process is observed it is generally the least powerful test against increasing trend. It is better than the Laplace test against decreasing and bathtub-shaped trend.

6.1.4 The TTT-based Military Handbook test

The TTT-based Military Handbook test has the same properties compared to the other TTT-based tests as the original test has in the one process case. It has good 'overall' properties.

6.1.5 The Anderson–Darling trend test

The Anderson–Darling trend test is by far the most powerful test against NHPPs with bathtub-shaped intensity functions. Against monotonic trends the Laplace test is slightly better against NHPP with increasing power law intensity functions, in all other situations that we considered the Anderson–Darling trend test is the second most powerful test against monotonic trend.

6.2 Final comments

It is obvious that no test is superior to the other tests in all situations. However, we feel that even for the single system case the Anderson–Darling trend test might be recommended as the best choice for general use. This is because the differences in power between the Anderson–Darling test

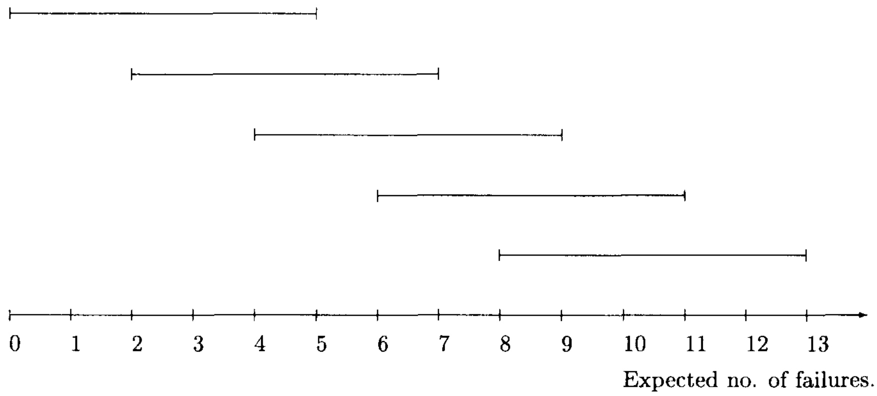

Fig. 9. Picture of expected number of failures in the processes. In case of a HPP the picture will be equal on the time axis, otherwise different.

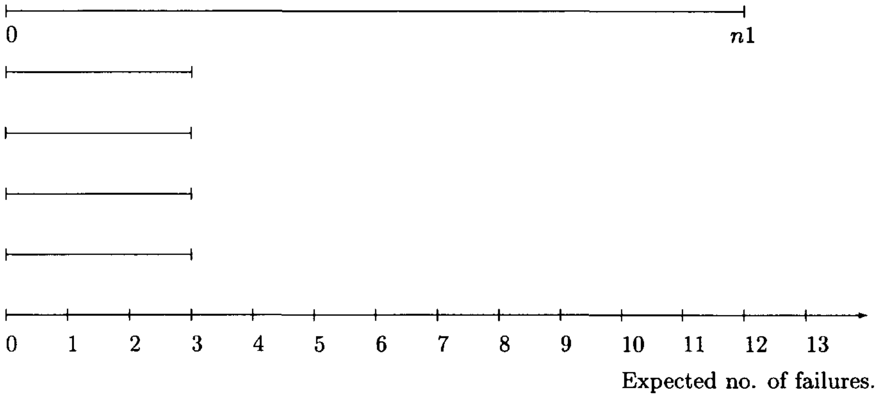

Fig. 10. Picture of expected number of failures in the processes. In this example the expected number of failures in the first process, $n1$, equals 12. In case of a HPP the picture will be equal on the time axis, otherwise different.

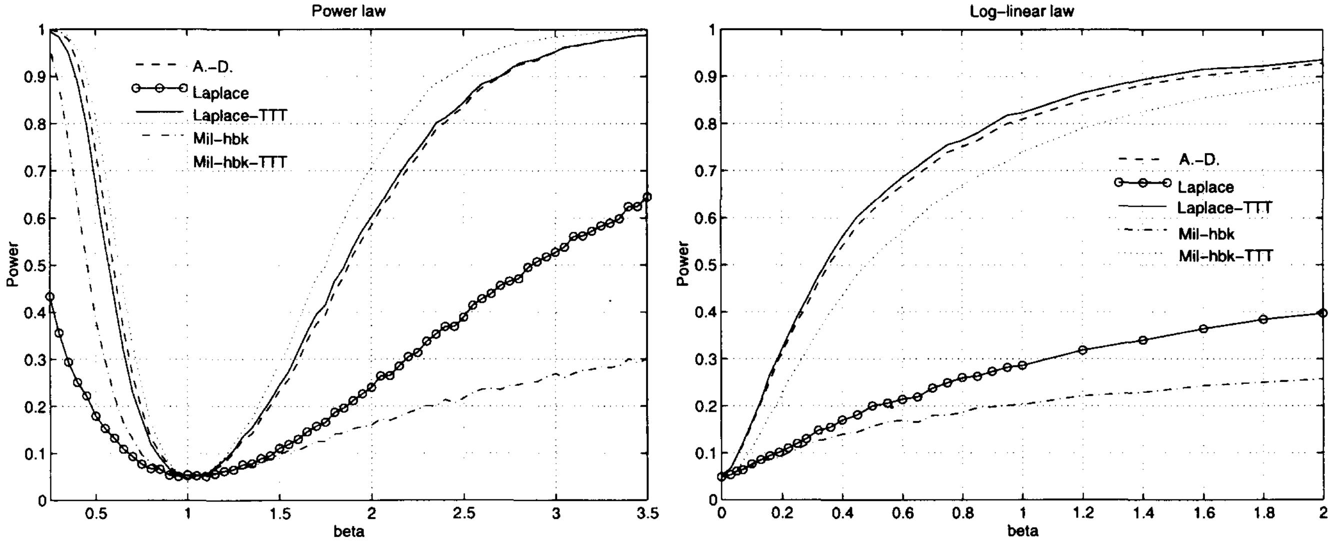

Fig. 11. Simulations of NHPPs with intensity function $\lambda(t) = t^{\beta-1}$, where β = beta = {0.25,0.30,...,3.5}, and from NHPPs with intensity function $\lambda(t) = e^{\beta t}$, where β = beta = {0,0.01,0.03,...,0.40,0.45,...,1,1.2,...,2.0}. Ten thousand replications for each beta-value. Data simulated from five processes using the simulation design in Fig. 9. Expected total number of failures is 25.

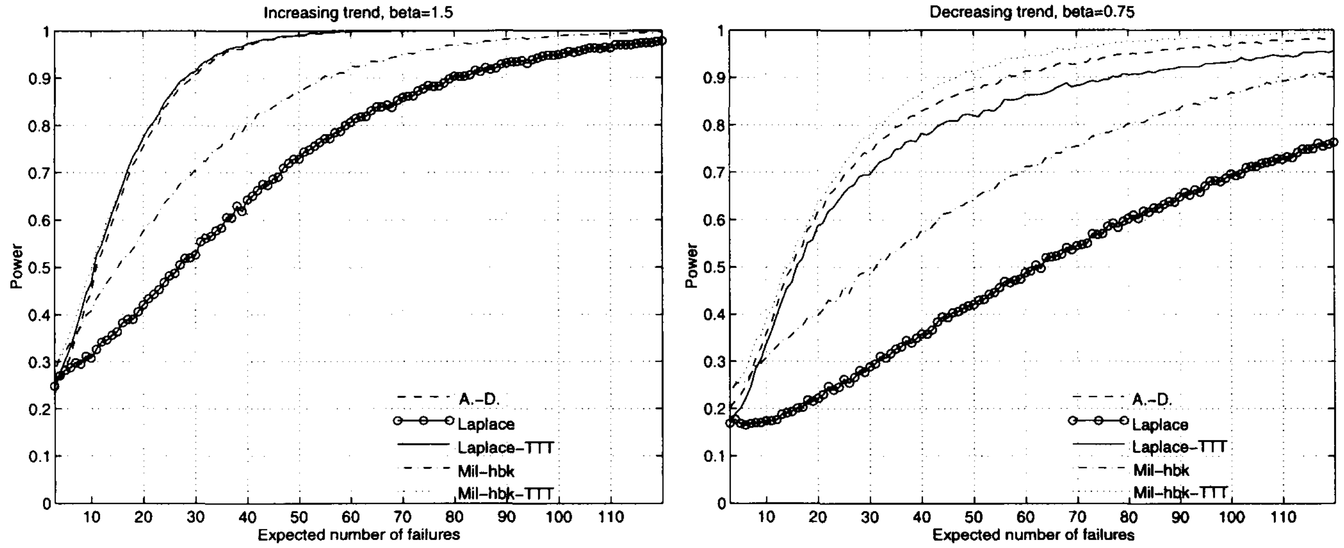

Fig. 12. Simulations of NHPPs with intensity function $\lambda(t) = t^{0.5}$, and from NHPPs with intensity function $\lambda(t) = t^{-0.25}$. Data simulated from five processes using the simulation design in Fig. 10. Expected total number of failures in the first process is varying from three to 120. Ten thousand replications for each expected number of failures in the first process. In all cases the expected number of failures in the four other processes equals three.

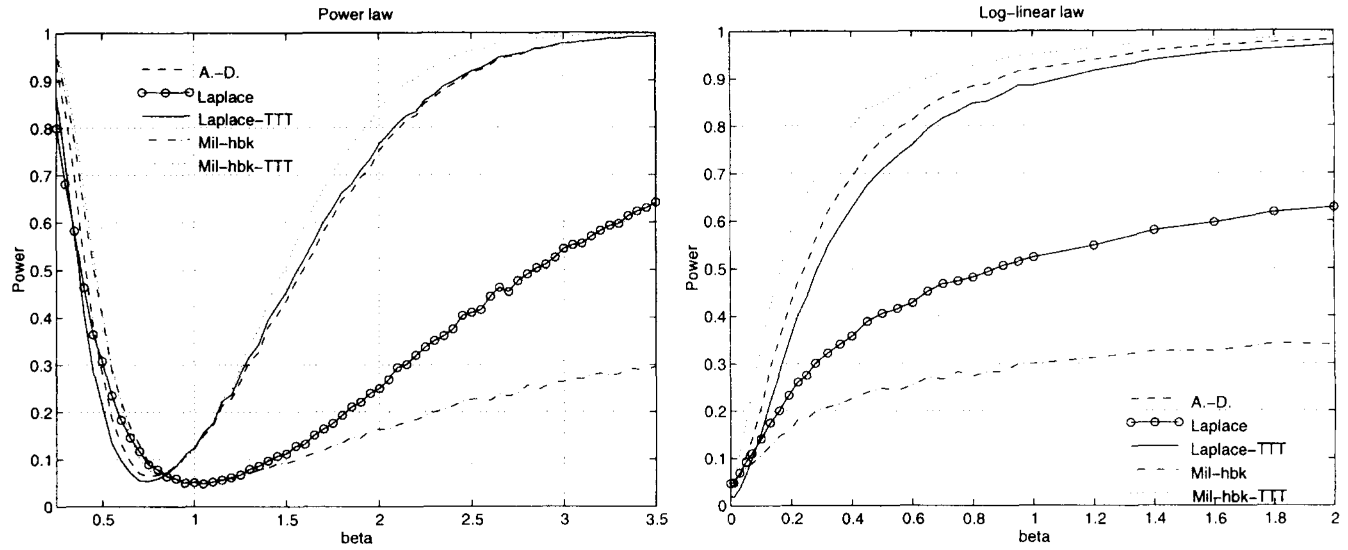

Fig. 13. Simulations of a NHPP with intensity function $\lambda(t)=\alpha_i t^{\beta-1}$, where $\alpha_1=0.5$, $\alpha_2=0.625$, $\alpha_3=0.75$, $\alpha_4=0.875$, $\alpha_5=1.0$, $\beta=\text{beta}=\{0.25,0.30,\ldots,3.5\}$, and from a NHPP with intensity function $\lambda(t)=e^{\alpha+\beta t}$, where $\alpha_1=-1.5$, $\alpha_2=-0.9$, $\alpha_3=-0.5$, $\alpha_4=-0.2$, $\alpha_5=0$, $\beta=\text{beta}=\{0,0.01,0.03,\ldots,0.40,0.45,\ldots,1,1.2,\ldots,2.0\}$. Ten thousand replications for each beta-value. Expected total number of failures is 25.

and the respective optimal test against monotonic alternatives, are small compared to the differences between them against nonmonotonic trends.

If more than one process is observed we must decide whether we want to test the null hypothesis of identical HPPs or if we want to allow for heterogeneous HPPs under the null hypothesis. In the former case, or with minor deviations from the former case, the Anderson–Darling trend test and the TTT-based Military Handbook test have the best 'all over' properties. In the latter case, the combined Laplace test or the combined Military Handbook test should be used. In fact, our simulations show that the TTT-based tests (including the Anderson–Darling test) may give misleading results in the presence of considerable heterogeneities.

We conjecture that in the case of identical intensity functions of the systems, the TTT-based tests are always at least as powerful as the combined ones. Intuitively this is so since by making stronger prior assumptions, one gets stronger inference results (as long as the assumptions hold). We have not found any situation which contradicts this conjecture, but by calculations and simulations not included in this paper we have found certain rather artificial situations where the combined tests were only slightly weaker than the TTT-based ones. This was in situations where the number $p(t)$ of observed processes increased strongly with time.

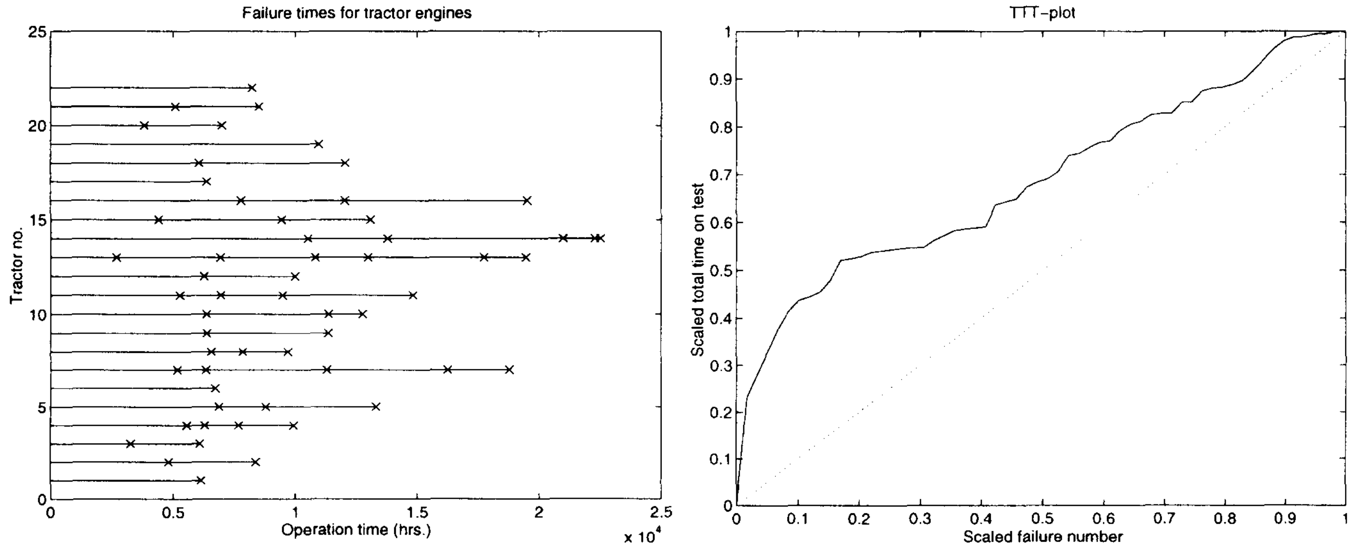

Fig. 14. Failure times and TTT plot for tractor data.

Table 1. Description of 12 bathtub-shaped intensity functions

	Slope		Expected number of failures		
Function	Phase I	Phase III	Phase I	Phase II	Phase III
1	2	2	8	8	8
2	1	1	8	8	8
3	1/2	1/2	8	8	8
4	2	1/2	8	8	8
5	1	1	5	5	5
6	1	1	15	15	15
7	1	1	10	5	10
8	1	1	5	10	5
9	1	1	10	0	10
10	0	1	0	10	10
11	2	1	4	8	10
12	1	1	10	8	4

Table 2. Simulated power with bathtub-shaped intensity function number 1–12

Test	Power											
	1	2	3	4	5	6	7	8	9	10	11	12
Laplace	0.25	0.22	0.17	0.24	0.16	0.26	0.20	0.17	0.14	0.49	0.27	0.36
Mil-hbk	0.36	0.27	0.19	0.44	0.17	0.48	0.26	0.21	0.14	0.32	0.18	0.45
C-vM	0.47	0.34	0.22	0.36	0.20	0.74	0.36	0.21	0.16	0.50	0.38	0.44
A–D	0.70	0.50	0.32	0.56	0.28	0.91	0.51	0.33	0.22	0.56	0.52	0.55

Ten thousand replications for each intensity function. Failure truncated processes.

Table 3. Simulated power with bathtub-shaped intensity function number 1–12

Test	Power											
	1	2	3	4	5	6	7	8	9	10	11	12
Laplace	0.13	0.11	0.09	0.15	0.10	0.12	0.11	0.10	0.08	0.53	0.25	0.29
Mil-hbk	0.31	0.22	0.16	0.42	0.14	0.43	0.21	0.16	0.10	0.33	0.14	0.47
C-vM	0.36	0.24	0.15	0.31	0.14	0.67	0.29	0.14	0.12	0.47	0.33	0.43
A–D	0.66	0.45	0.28	0.51	0.24	0.89	0.48	0.27	0.18	0.61	0.53	0.51

Ten thousand replications for each intensity function. Time truncated processed.

Table 4. Simulated power with bathtub-shaped intensity function number 1–12

Test	Power											
	1	2	3	4	5	6	7	8	9	10	11	12
Laplace	0.10	0.08	0.07	0.09	0.07	0.09	0.08	0.07	0.04	0.44	0.25	0.27
Laplace-TTT	0.13	0.11	0.09	0.18	0.09	0.12	0.12	0.09	0.09	0.76	0.37	0.44
Mil-hbk	0.15	0.11	0.08	0.21	0.08	0.57	0.09	0.09	0.05	0.29	0.10	0.31
Mil-hbk-TTT	0.54	0.38	0.26	0.68	0.23	0.38	0.38	0.28	0.14	0.46	0.14	0.77
A–D	0.95	0.80	0.54	0.84	0.54	1.00	0.87	0.46	0.30	0.86	0.84	0.82

Ten thousand replications for each intensity function. Five processes observed.

Table 5. Simulated actual level with various heterogeneous HPPs

	Intensities					
α_1	0.1	0.2	0.5	0.6	0.75	0.9
α_2	0.25	0.4	0.625	0.7	0.75	0.9
α_3	0.5	0.6	0.75	0.8	0.9	0.9
α_4	0.75	0.8	0.875	0.9	1	1
α_5	1	1	1	1	1	1
Test	Actual level					
Laplace	0.049	0.051	0.050	0.050	0.050	0.048
Laplace-TTT	0.530	0.029	0.130	0.108	0.076	0.055
Mil-hbk	0.051	0.049	0.051	0.051	0.050	0.050
Mil-hbk-TTT	0.095	0.015	0.087	0.073	0.058	0.053
A–D	0.596	0.071	0.129	0.106	0.074	0.056

Five processes. Twenty thousand replications.

Table 6. Analysis of tractor data

Test	Test statistic	P-value
Combined Laplace	2.08	0.0372
Laplace-TTT	5.03	0.0000
Combined Mil-hbk	42.6	0.0026
Mil-hbk-TTT	48.5	0.0000
A–D	13.3	0.0000

ACKNOWLEDGEMENTS

We gratefully acknowledge a number of important and suggestive comments from two referees, in our opinion leading to a significantly improved paper. We are especially indebted to the referee who suggested the use of the Anderson–Darling statistic as a basis for a trend test instead of the Cramér–von Mises statistic originally used by us.

Jan Terje Kvaløy is supported by a grant from the Norwegian Council of Research.

REFERENCES

1. Ascher, H. and Feingold, H., *Repairable Systems Reliability. Modeling, Inference, Misconceptions and Their Causes*, Marcel Dekker, New York, 1984.
2. Barlow, R. E. and Davis, B., Analysis of time between failures for repairable components. In *Nuclear Systems Reliability Engineering and Risk Assessment* (eds J. B. Fussell and G. R. Burdick), SIAM, Philadelphia, 1977, pp. 543–561.
3. Anderson, T. W. and Darling, D. A. Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *Ann. Math. Statist.*, 1952, **23,** 193–212.
4. Cramér, H., On the composition of elementary errors. *Skand. Aktuar.*, 1928, **11,** 13–74, 141–180.
5. von Mises, N., *Wahrscheinlichkeitsrechnung*, Deuticke, Leipzig, 1931.
6. Bain, L. J., Engelhardt, M. and Wright, F. T. Tests for an increasing trend in the intensity of a Poisson process: a power study. *J. Am. Statist. Assoc.*, 1985, **80,** 419–422.
7. Cohen, A. and Sackrowitz, H. B. Evaluating tests for increasing intensity of a Poisson process. *Technometrics*, 1993, **35,** 446–448.
8. Lindqvist, B., Kjønstad, G. A. and Meland, N., Testing for trend in repairable systems data. *Proceedings of ESREL '94*, La Baule, France, 30 May–3 June, 1994.
9. Andersen, P. K., Borgan, Ø., Gill, R. and Keiding, N., *Statistical Models Based on Counting Processes*, Springer, New York, 1992.
10. Høyland, A. and Rausand, M., *System Reliability Theory. Models and Statistical Methods*, Wiley, New York, 1994.
11. Cox, D. R. and Lewis, P. A. W., *The Statistical Analysis of Series of Events*, Methuen, London, 1966.
12. Ross, S. M., *Stochastic Processes*, John Wiley, New York, 1983.
13. MIL-HDBK-189, *Reliability Growth Management*, Headquarters, U.S. Army Communications Research and Development Command, ATTN: DRDCO-PT, Fort Monmouth, NJ 07702, 1981.
14. Aarset, M. V. How to identify a bathtub hazard rate. *IEEE Transactions on Reliability*, 1987, **R-36,** 106–108.
15. Stephens, M. A. EDF statistics for goodness of fit and some comparisons. *J. Am. Statist. Assoc.*, 1974, **69,** 730–737.
16. Anderson, T. W. and Darling, D. A. A test of goodness of fit. *J. Am. Statist. Assoc.*, 1954, **49,** 765–769.
17. Elvebakk, G., Personal communication.