An Area Based Test for Trend in Repairable Systems Data

Jan Terje Kvaløy* and Bo Lindqvist**

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

September, 1996

Abstract

A major aspect of analysis of failure data for repairable systems is the testing for a possible trend in interfailure times. The paper reviews some important and popular graphical methods and tests for the nonhomogeneous Poisson process model. In particular the TTT plot is considered, and a new trend test based on the area between this plot and the diagonal line corresponding to "no trend" is motivated and derived. Also, construction of trend tests by the likelihood ratio method is demonstrated. The tests are evaluated and compared in a simulation study, both with respect to the achievement of correct significance level and with respect to rejection power. The considered alternatives to "no trend" are the log-linear, the 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.

1 INTRODUCTION

For maintained and repairable systems it is important to be able 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.

Figure 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 Figure 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

*E–mail: `jtk@imf.unit.no`
**E–mail: `bo@imf.unit.no`

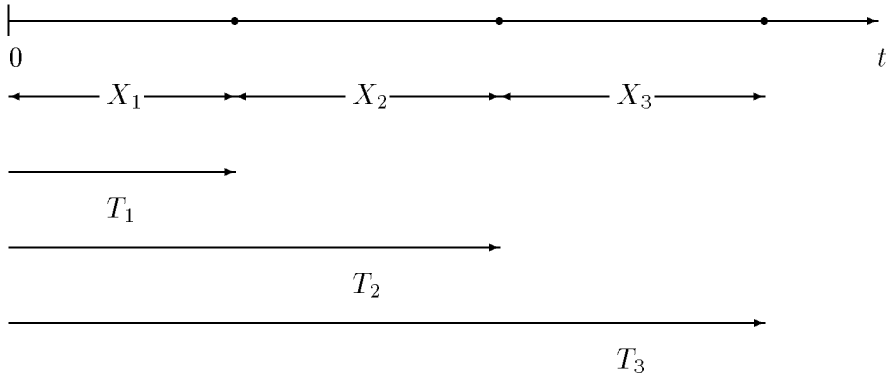

Figure 1: Arrival times, T_i, and inter-arrival times, X_i.

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 larger (a *decreasing* trend), and the system is said to be *deteriorating* if the inter-arrival times tend to get smaller (*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 larger and smaller, 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.

The perhaps 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 & Feingold [4] 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 [4], and for some possible fallacies by using NHPP models we also refer to Lindqvist et al. [15].

In this paper we restrict attention to NHPP models and study various properties of four 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 construct in addition a new trend test based on the TTT plot for NHPPs defined by Barlow and Davis [6]. This test should, by its 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.

In addition to the three mentioned tests, we demonstrate how to construct trend tests by the likelihood method. The test we derive is very similar to the Military Handbook test, in fact asymptotically equivalent, but has a bit different finite sample behavior.

If more than one process is observed, and the processes are assumed to be independent and identically distributed, we might want to perform a common trend test on all the processes together. We discuss how to generalize the Laplace test and the Military Handbook test to the case of more than one process. The properties of the standard generalizations of these tests are compared to the properties of a generalization based on a total time on test concept.

2 IDENTIFICATION OF TREND

2.1 The nonhomogeneous Poisson Process (NHPP)

We refer again to Figure 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 $\Delta t \to 0$, and (4) $P(N(t+\Delta t) - N(t) \geq 2) = o(\Delta t)$ as $\Delta t \to 0$. (The last assumption assures that two or more events can not 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 (see Ascher & Feingold [4]).

Further, letting the *cumulative intensity* be given by

$$\Lambda(t) = \int_0^t \lambda(u)du,$$

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

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

and the *log-linear* intensity,

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

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 $\lambda(t)$.

In Section 4 we shall need the property that if $T_1, T_2, \ldots$ are arrival times for a NHPP with cumulative intensity $\Lambda(t)$, then $\Lambda(T_1), \Lambda(T_2), \ldots$ is a HPP with intensity 1.

2.2 The repairable system model

In the NHPP framework, the object is to decide whether a HPP or a 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 NHPP's with a common intensity function $\lambda(t)$, are observed. The i'th system is observed in the time interval $(a_i, b_i]$ with n_i failures occurring at times T_{ij}, $j = 1, 2, \ldots, n_i$.

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 have 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 $\Lambda(t) = \int_0^t \lambda(u)du$ is given by

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

where $Y(T_{ij})$ is the number of processes 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. [2].

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. Deviations 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.

2.4 TTT Plot

The *TTT* (Total Time on Test) plot is most well known as a graphical technique for data from nonrepairable systems (see e.g. Høyland & Rausand [13]). A TTT plot for repairable systems data has been introduced by Barlow and Davis [6], based on the NHPP model. As above, assume that m independent NHPP's with common intensity function $\lambda(t)$ are observed, and assume that all observation intervals $(a_i, b_i]$ are contained in some time interval $[0, T]$. Let T_i denote the i'th arrival time in the superposed process, that is; T_i is an arrival time in one of the processes and $0 < T_1 \leq T_2 \leq \ldots \leq T_N \leq T$. Let $p(u)$ denote the number of processes under observation at time u and let $\mathcal{T}(t) = \int_0^t p(u)du$ denote the *total time on test* from time 0 to time t.

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

$$\frac{\mathcal{T}(T_i)}{\mathcal{T}(T_N)} = \frac{\int_0^{T_i} p(u)du}{\int_0^{T_N} p(u)du}$$

versus scaled failure number $\frac{i}{N}$. If $n(u) \equiv m$, 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 NHPP's with decreasing, increasing and bathtub shaped intensity functions are illustrated in Figure 2

The points $\left(\frac{i}{n}, \frac{\mathcal{T}(T_i)}{\mathcal{T}(T_N)}\right)$ in the TTT plot are often connected with straight lines.

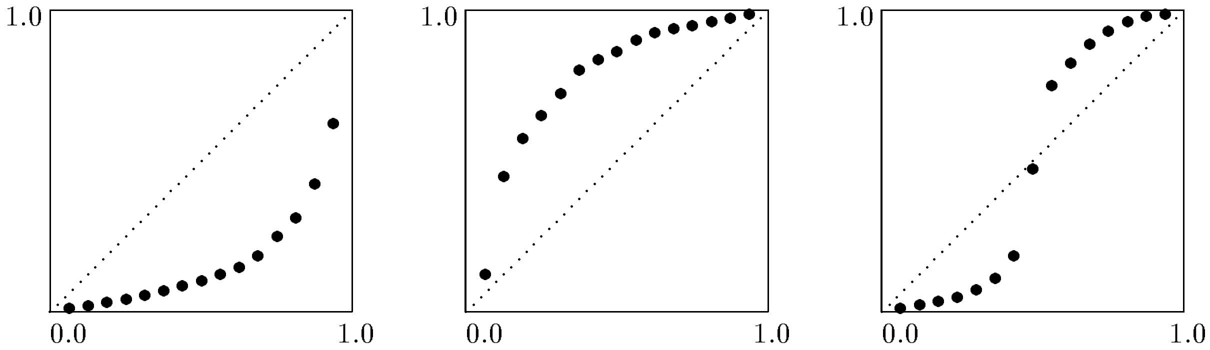

Figure 2: Typical shape of TTT plot from NHPP's with decreasing, increasing and bathtub shaped intensity function.

2.5 Laplace's Test

Laplace's test is a test of the null hypothesis H_0: HPP, vs. 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

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

is approximately standard normally distributed under the null hypothesis, i.e. when the underlying process is a HPP. The approximation turns out to be very good, a rule of thumb says that $n \geq 3$ suffices. The test is approximately optimal for the null hypothesis (HPP) when the alternative is a NHPP with log-linear intensity function (Cox & Lewis [9]).

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 to more than one process can be done in several ways. A straightforward generalization of (1) is

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

This test is approximately optimal for the null hypothesis of HPP, possibly with *different* intensities in each HPP, against the alternative of NHPP's with intensity $\lambda(t) = e^{\alpha_i + \beta t}$ where β is common for all processes while α_i is specific for each process (Cox & Lewis [9]). We call this test the *combined* Laplace test.

A different way of generalizing the test is to use the TTT-transform. Given the total number of failures N, $\frac{\mathcal{T}(T_k)}{\mathcal{T}(T_N)}$, $k = 1, \ldots, N$ is distributed as the order statistic of $N - 1$ *uniform*$(0, 1)$ random variables. (This is further discussed in Section 4.) Hence

$$L = \frac{\sum_{k=1}^{N-1} \frac{\mathcal{T}(T_k)}{\mathcal{T}(T_N)} - \frac{1}{2}(N-1)}{\sqrt{\frac{1}{12}(N-1)}} \tag{3}$$

is approximately standard normally distributed under the null hypothesis that each process is a HPP with *identical* intensities. We will call this the *TTT-based Laplace test.*

2.6 Military Handbook Test

This is another test constructed for the null hypothesis of a HPP vs. the alternative of NHPP with monotone trend. The test statistic is

$$M = 2 \sum_{j=1}^{\hat{n}} \ln \left(\frac{b - a}{T_j - a} \right) \tag{4}$$

which, is (exactly) χ-square distributed with $2n$ degrees of freedom under the null hypothesis.

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 (Bain et al. [5]).

The straightforward generalization of (4) to more than one process is

$$M = 2 \sum_{i=1}^{m} \sum_{j=1}^{\hat{n}_i} \ln \left(\frac{b_i - a_i}{T_{ij} - a_i} \right) \tag{5}$$

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

Under the null hypothesis of HPP's with identical intensities, the test statistic

$$M = 2 \sum_{k=1}^{N-1} \ln \left(\frac{\mathcal{T}(T_N)}{\mathcal{T}(T_k)} \right) \tag{6}$$

is χ-square distributed with $2(N-1)$ degrees of freedom, and we call this the *TTT-based Military Handbook test.*

3 A LIKELIHOOD-RATIO TEST

Various trend tests can be obtained by the likelihood-ratio recipe. The basic idea is to start from a model involving parameters that in some sense measure the degree of trend, and letting the null hypothesis correspond to values of these parameters giving no trend.

As an example, we shall consider the NHPP model with intensity function of the *power law* type, i.e.

$$\lambda(t) = \alpha\beta t^{\beta-1} \tag{7}$$

In this case "no trend" corresponds to $\beta = 1$, so we test

$$H_0 : \beta = 1 \qquad \text{vs.} \qquad H_1 : \beta \neq 1 \tag{8}$$

Suppose for simplicity that we have observed a single system in the time interval $(0, b]$ and recorded the arrival times $t_1, \ldots, t_n$. (A starting point different from 0 makes things a bit more complicated, so we assume that observation starts from time $t = 0$). The likelihood function for such data from a NHPP is particularly simple, given by (see e.g. Crowder et al. [10])

$$\exp\left\{-\int_0^b \lambda(u)\mathrm{d}u\right\}\left\{\prod_{j=1}^{n} \lambda(t_j)\right\}$$

Substituting (7) in this likelihood and taking the logarithm we get the log-likelihood,

$$l(\alpha, \beta) = n\log\alpha + n\log\beta + (\beta - 1)\sum_{j=1}^{n} \log t_j - \alpha b^{\beta}$$

The *likelihood ratio statistic* corresponding to the hypotheses (8) is

$$R = 2\{\max_{\alpha,\beta} l(\alpha, \beta) - \max_{\alpha} l(\alpha, 1)\}$$

In words this means that R is twice the difference between the maximum value of l in the full model and under the null hypothesis assumptions, respectively. The null hypothesis should be rejected for large values of R. Standard asymptotic theory (Crowder et al. [10]) states that this statistic is approximately χ^2-distributed with one degree of freedom when H_0 holds, and this enables us to find (approximate) critical values of the test.

It is an interesting fact that R can be written

$$R = 2n\log\frac{2n}{M} - 2n + M$$

where M is the Military Handbook test statistic (4) obtained when $m = 1$ and $a_1 = 0$. Thus, the derived test and the Military Handbook test are asymptotically equivalent.

However, when applied to samples of moderate size they have somewhat different properties.

Simulation studies indicate that for a failure truncated process this test tends to achieve an actual level larger than the nominal level. If the process is time truncated, however, the actual level achieved correctly equals the nominal level. One way to improve the finite sample accuracy of log-likelihood-ratio statistics is by the use of *Bartlett adjustments*, see e.g. Jensen [14]. If we want to improve the test statistic R by Bartlett adjustments the basic idea is to calculate

$$\tilde{R} = \frac{R}{\mathbf{E}(R)}$$

(where $\mathbf{E}$ denotes expectation). Usually $\mathbf{E}(R)$ can not be computed exactly, so one has to stick to approximations. Fortunately, for the special case considered here, $\mathbf{E}(R)$ *can* be calculated. Since we know that M is $\chi^2_{2\hat{n}}$-distributed, with $\hat{n}$ defined in (1), we get

$$\begin{aligned}
\mathbf{E}(R) &= 2n\ln(2n) - 2n\mathbf{E}(\ln(M)) - 2n + \mathbf{E}(M) \\
&= 2n\ln(2n) - 2n + 2\hat{n} \\
&\quad -2n\int_0^\infty \frac{1}{2^{\hat{n}}\Gamma(\hat{n})}\ln(x)x^{\hat{n}-1}e^{-\frac{x}{2}}\,dx \\
&= 2n\ln(2n) - 2n + 2\hat{n} - 2n(\ln 2 + \Psi(\hat{n})) \\
&= 2n\ln n - 2(n-\hat{n}) - 2n\Psi(\hat{n})
\end{aligned}$$

The function $\Psi(x) = \frac{\frac{d}{dx}\Gamma(x)}{\Gamma x}$ is called the *digamma* function (Abramowitz & Stegun [1]). The integral may be evaluated e.g. with aid of the software system *Mathematica [17]*. If the process is time truncated the adjusted test statistic is

$$\tilde{R} = \frac{R}{2n(\ln n - \Psi(n))} \tag{9}$$

and if the process is failure truncated:

$$\tilde{R} = \frac{R}{2(n\ln n - 1 - n\Psi(n-1))} \tag{10}$$

According to Barndorff-Nielsen & Cox [7] the null distribution of $\tilde{R}$ differs from the χ^2_1-distribution by a term of order $\mathcal{O}(n^{-\frac{3}{2}})$.

If we observe m independent systems, each from time $t = 0$, but to possibly different endpoints b_i, we redefine the $l(\alpha, \beta)$ to be the *sum* of the log likelihoods for each system and then define R in the same way by maximizations of l.

However this maximization must be done by numerical methods, and the use of the total time on test recipe is not as straightforward as with the Laplace test and the Military Handbook test. As a consequence, in the simulation study the likelihood ratio test is only considered in the single process case.

4 THE AREA BASED TEST

The TTT plot was presented in Section 2 as an appropriate graphical method visualizing trend in data from NHPP's. We shall in this section demonstrate how a statistical trend test based on the TTT plot can be obtained.

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.

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 this area.

As in Section 2.4 we shall assume that we have observations from m independent NHPP's with common intensity function $\lambda(t)$, and that each process has been observed in a subset of the time interval $[0, T]$. Recall that the scaled TTT plot is the plot of the points $\left(\frac{i}{n}, \frac{\mathcal{T}(T_i)}{\mathcal{T}(T_n)}\right)$, where $\mathcal{T}(u)$ is the total time on test statistic ($\mathcal{T}(u) = \int_0^u p(v)dv$ where $p(v)$ is the number of processes under observation at time v) in the superposed process with arrival times $0 < T_1 \leq T_2 \leq \ldots \leq T_n \leq T$.

The superposed process has intensity function $\gamma(t) = \lambda(t)p(t)$, and under the null-hypothesis of no trend, $\lambda(t) \equiv \lambda$, it has cumulative intensity function $\Gamma(t) = \lambda\mathcal{T}(t)$. Hence (see Section 2.1) the time transformed process $\Gamma(T_1), \ldots, \Gamma(T_n)$ is a HPP with intensity one, and it follows that the process $\mathcal{T}(T_1), \ldots, \mathcal{T}(T_n)$ is a HPP with intensity λ. If this process, with probability one, is stopped at the n'th event (for instance if it is failure truncated) it is well known (see e.g. Ross [16]) that $\frac{\mathcal{T}(T_k)}{\mathcal{T}(T_n)}$ $k = 1, \ldots, n-1$ will have the same distribution as the order statistic based on $n-1$ i.i.d. *uniform*$(0,1)$ random variables. Gill [12] shows that under rather mild regularity conditions this still holds even if it is not the case that at least a prespecified number of observations will occur. For example it still holds if the processes are time truncated with n equal to the now random number $N(T)$ and T_n thus is not a stopping time. We conclude that the empirical distribution function of the $\left\{\frac{\mathcal{T}(T_k)}{\mathcal{T}(T_n)}, k = 1, \ldots, n-1\right\}$,

$$F_n(v) = \frac{k-1}{n}, \qquad \frac{\mathcal{T}(T_{k-1})}{\mathcal{T}(T_n)} \leq v < \frac{\mathcal{T}(T_k)}{\mathcal{T}(T_n)}$$

approximately equals the cumulative distribution function of the *uniform*$(0,1)$ distribution.

The step function $F_n(v)$ may be replaced by the *continuous* function

$$G_n(v) = \frac{1}{n}\frac{1}{\frac{\mathcal{T}(T_k)-\mathcal{T}(T_{k-1})}{\mathcal{T}(T_n)}}\left(v - \frac{\mathcal{T}(T_{k-1})}{\mathcal{T}(T_n)}\right) + \frac{k-1}{n}, \qquad \frac{\mathcal{T}(T_{k-1})}{\mathcal{T}(T_n)} \leq v < \frac{\mathcal{T}(T_k)}{\mathcal{T}(T_n)}$$

which connects the points $(\frac{\mathcal{T}(T_i)}{\mathcal{T}(T_n)}, \frac{i}{n})$ by straight lines, and which has the same asymptotic properties as $F_n(v)$. Now a plot of $(G_n(v), v)$ for $0 \leq v \leq 1$ is simply the TTT plot with the points $(\frac{i}{n}, \frac{\mathcal{T}(T_i)}{\mathcal{T}(T_n)})$ connected by straight lines.

Define now the process

$$C_n(v) = \sqrt{n}(G_n(v) - v)$$

It follows from standard convergence theory (Billingsley [8]) that C_n converges in distribution to the *Brownian Bridge* W on $(0,1)$.

By its definition, C_n defines a measure of the distance between the TTT plot and the diagonal. Andersen et al. [2] uses the signed area between the TTT-plot and the diagonal as a measure of departures from the HPP assumption. They arrive at an easily evaluated asymptotically normal test statistic, but their test have the serious drawback of having no power against non-monotonic trends that places equal portions of area on both sides of the diagonal in the TTT-plot, e.g. bathtub shaped trend. Another test proposed by Andersen et al. [2] is to use $\max_{v\in[0,1]} C_n(v)$ 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%), and hence has very low power. As a consequence of this it 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 [18], we propose the test statistic

$$A_n = \int_0^1 C_n^{\prime 2}(v)dv$$

which can be viewed as a measure of the (unsigned) area between the TTT plot and the diagonal.

According to the invariance principle (Billingsley [8]), $A_n \xrightarrow{D} A$ where:

$$A = \int_0^1 W^2(v)dv$$

The cumulative distribution function of A is (Durbin [11]),

$$P(A \leq a) = 1 - \frac{1}{\pi}\sum_{j=1}^{\infty}(-1)^{j-1}\int_{(2j-1)^2\pi^2}^{(2j)^2\pi^2}\frac{1}{y}\left(\frac{-\sqrt{y}}{\sin\sqrt{y}}\right)^{\frac{1}{2}}e^{-\frac{ay}{2}}dy \quad , a \geq 0$$

which is tabulated in Anderson & Darling [3]. For instance $P(A > 0.461) = 0.05$ and $P(A > 0.743) = 0.01$

It only remains to calculate A_n. The test statistic is given as

$$A_n = n\int_0^1 (G_n(v) - v)^2 dv$$

where in general $n = N(T)$ with $[0, T]$ being the interval of observation. Tedious but straightforward calculations give the test statistic:

$$\begin{aligned} A_n \;&=\; \sum_{k=1}^{n}\Big[\frac{k^2 - k + \frac{1}{3}}{n}\frac{\mathcal{T}(T_k) - \mathcal{T}(T_{k-1})}{\mathcal{T}(T_n)} - (k - \frac{1}{3})\frac{\mathcal{T}(T_k)^2}{\mathcal{T}(T_n)^2} + (k - \frac{2}{3})\frac{\mathcal{T}(T_{k-1})^2}{\mathcal{T}(T_n)^2} \\ &\quad + \frac{1}{3}\frac{\mathcal{T}(T_k)\mathcal{T}(T_{k-1})}{\mathcal{T}(T_n)^2}\Big] + \frac{n}{3} \end{aligned} \tag{11}$$

On a 5% (nominal) level the null hypothesis is rejected if $A_n > 0.461$.

In the special case when the superposed process is failure truncated after n failures, and all the processes have been observed during the entire interval $[0, T_n]$ the test statistic becomes:

$$A_n = \sum_{k=1}^{n}\left[\frac{k^2 - k + \frac{1}{3}}{n}\frac{T_k - T_{k-1}}{T_n} - (k - \frac{1}{3})\frac{T_k^2}{T_n^2} + (k - \frac{2}{3})\frac{T_{k-1}^2}{T_n^2} + \frac{1}{3}\frac{T_k T_{k-1}}{T_n^2}\right] + \frac{n}{3} \tag{12}$$

The area based 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, if the NHPP assumption is valid, that there is a departure from the "no trend" situation. The TTT plot or the Nelson-Aalen plot then gives a qualitative description of this departure.

5 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.

5.1 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 if the processes are failure truncated, and approximately equal if the processes are time truncated.

For each simulated process we consider the time from $t = 0$. When simulating a failure truncated processes, the process is run until this 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 may vary). 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 four tests. This is accomplished by simulating (usually) 10000 processes with the same setup, and recording the relative number of rejections (absolute number divided by 10000) 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 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 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 approximately $\sqrt{\hat{p}(1-\hat{p})/n}$, which is bounded above by $1/(2\sqrt{n})$. Thus, if $n = 10000$, the standard deviation is no larger than 0.005.

The following abbreviations are used in the graphs: *Laplace* – The Laplace test (1), *Mil-hbk* – The Military Handbook test (4), *Lik-ratio* – Likelihood-ratio test (9,10), *Area* – Area-based test (11).

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

5.1.1 Level Properties

Figure 3 displays the actual levels of the two new tests (Likelihood ratio and area tests, respectively), for a varying number of failures and data simulated from a failure truncated HPP. Corresponding simulations for the time truncated case show basically the same picture and are therefore not included.

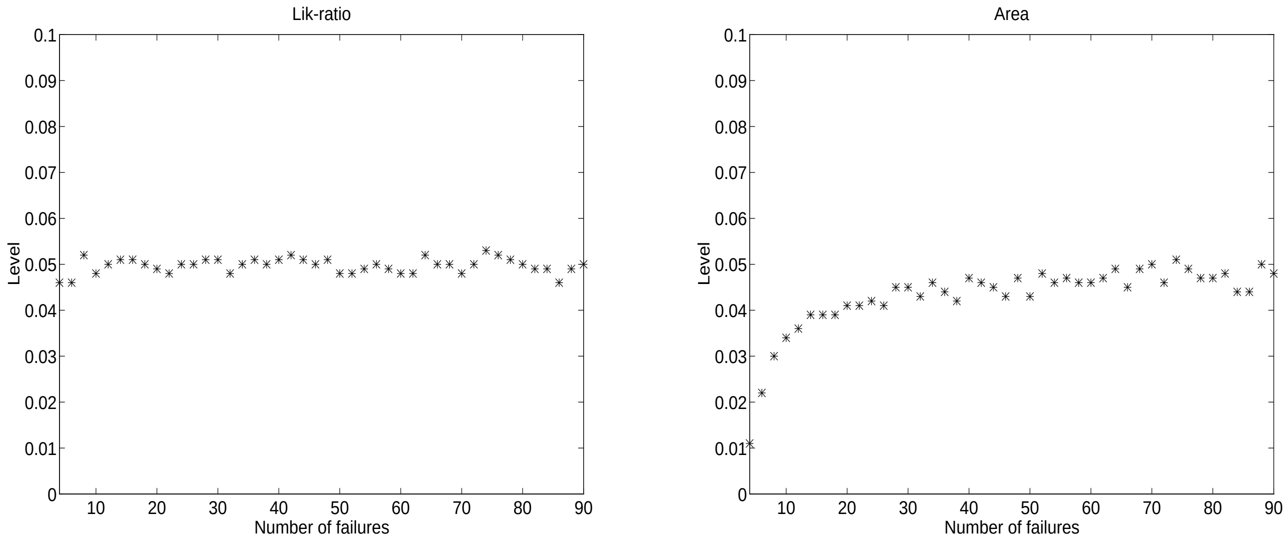

Figure 3: Simulations of a HPP with intensity 1. Failure truncation with the number of failures ranging from 4 to 90. 20000 replications for each number of failures.

The graphs show that the likelihood ratio test achieves the correct actual level of 5% with satisfying accuracy. The area test, however, achieves the 5% actual level only for large samples. This test has actual level about 4% on samples of size less than about 30, and even lower for samples of size less than 14.

As regards the Laplace- and the Military Handbook tests, the latter keeps an exact level for HPP's, while the former is approximately exact even for as low as 3 failures.

5.1.2 Power Properties

Log-Linear Intensity Function. In this case $\lambda(t) = e^{\alpha+\beta t}$, and $\Lambda(t) = \frac{1}{\beta}e^{\alpha}(e^{\beta t} - 1)$. Thus, if $\beta < 0$ we have a finite limit $\lim_{t\to\infty} \Lambda(t) = \frac{e^{\alpha}}{|\beta|}$. This means that only a finite number of failures are expected to occur during infinite time or that an inter-arrival time with positive probability might be infinite if $\beta < 0$. In order to avoid difficulties arising from this, we have simulated only the increasing case, $\beta > 0$.

Estimated power functions with data simulated from a failure truncated NHPP with log-linear intensity function and 15 failures and from a time-truncated process where the truncation times are chosen so that the expected number of failures in each simulation is 15 are presented in Figure 4.

From Figure 4 we observe that in the case of a failure truncated process with 15 failures, the likelihood-ratio test is actually the test with largest power if $\beta < 1.1$, while the Laplace test which asymptotically is the optimal test in this case is the most powerful test if $\beta > 1.1$. The area test has larger power than the Military Handbook test if $\beta > 0.4$, but is slightly less powerful when $\beta < 0.4$. This is probably due to the the low actual level of the area test (about 4%) in this case.

From Figure 4 we also observe the interesting fact that in the case of a time truncated process with expected number of failures equal to 15, the Laplace test is the most powerful test for all β-values. The area test suffers even more from its low actual level in the case of

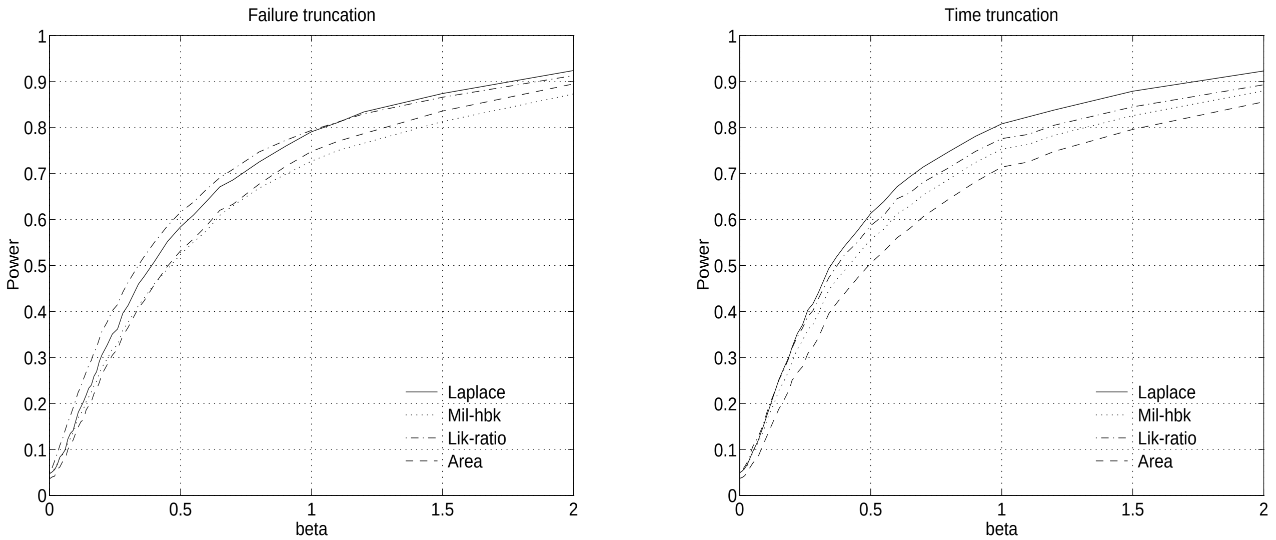

Figure 4: Simulations of a NHPP with intensity function $\lambda(t) = e^{\beta t}$, where β=beta=$\{0,0.01,0.03,\ldots,0.40,0.45,\ldots,1,1.2,\ldots,2.0\}$.
10000 replications for each beta-value. Failure truncation with 15 failures and time truncation with expected number of failures in each simulation 15.

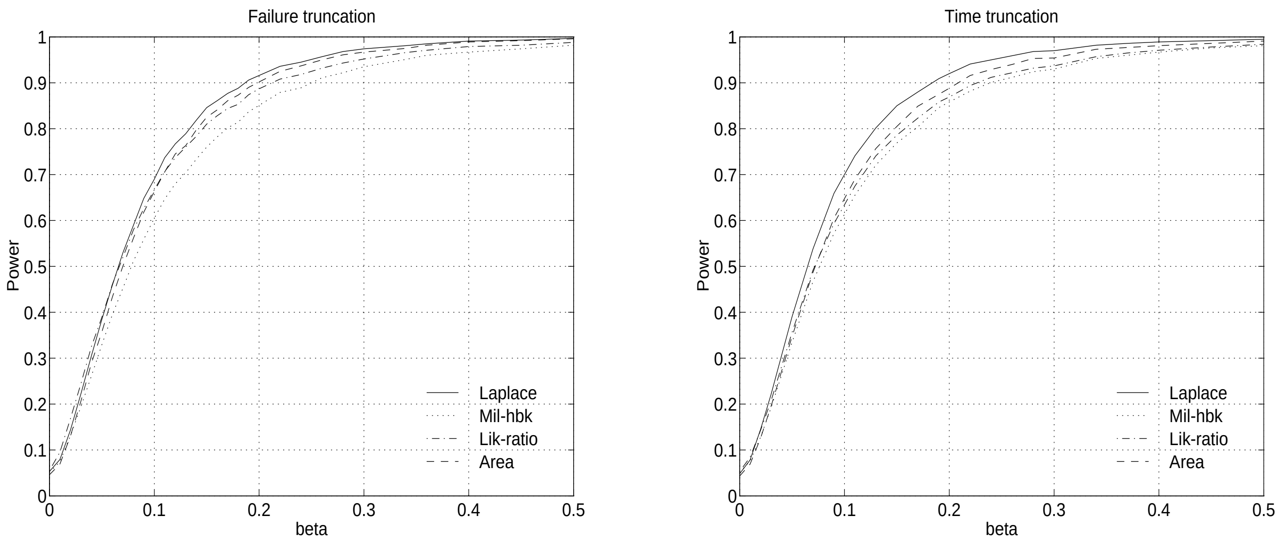

Figure 5: Simulations of a NHPP with intensity function $\lambda(t) = e^{\beta t}$, where β=beta=$\{0,0.01,0.03,\ldots,0.50\ \}$.
10000 replications for each beta-value. Failure truncation with 35 failures and time truncation with expected number of failures in each simulation 35.

time truncation and has less power than the Military Handbook test for all β-values.

While the Laplace test and the Military Handbook test have nearly the same power in the failure truncated and the time truncated process, the likelihood-ratio test and the area test, have less power in the time truncated process.

When samples of size 35 are simulated, see Figure 5, we recognize that for small values of β in the failure truncated process the likelihood-ratio test still have slightly larger power than the Laplace test, but elsewhere the Laplace test is the most powerful test. The area test has now larger power than the Military Handbook test everywhere, and for $\beta > 0.1$ it is also more powerful than the likelihood-ratio test, both in the time and in the failure truncated process.

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 a failure truncated NHPP with power law intensity function and 15 failures and estimated power functions with data simulated from a time truncated process with expected number of failures in each simulation equal to 15.

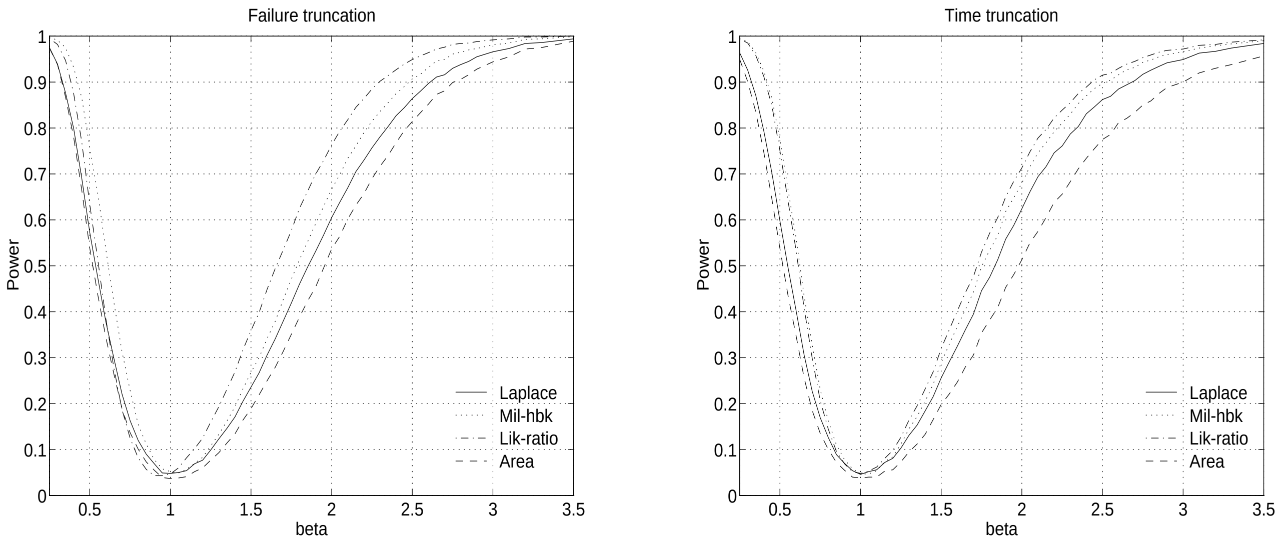

Figure 6: Simulations of a NHPP with intensity function $\lambda(t) = t^{\beta-1}$, where β=beta=$\{0.25,0.30,\ldots,3.5\}$.
10000 replications for each beta-value. Failure truncation with 15 failures and time truncation with expected number of failures in each simulation 15.

From Figure 6 we observe that in the case of a failure truncated process with 15 failures, the Military Handbook test is the test with largest power on data with decreasing trend, while the likelihood-ratio test definitely is the most powerful test on data with increasing trend. The likelihood-ratio test is a bit "skew". It has low power against weakly decreasing trend, if $\beta \in \{0.75, 0.95\}$ it is the least powerful of the evaluated tests. The Laplace test is less powerful than the Military Handbook test, and the area test is less powerful than the Laplace test.

We realize that in the case of a time truncated process the adjusted likelihood-ratio test is more "symmetric", but though closer to the Military Handbook test it is still the test with largest power against increasing trend, and it has nearly the same power against decreasing trend as the Military Handbook test.

Just as in the case of log-linear intensity function the Laplace test and the Military Handbook test have the same power function both in the time and in the failure truncated processes, while the area test has less power when the process is time truncated. The reason is probably that the area test does not use the information which the time from the last failure until the truncation time gives.

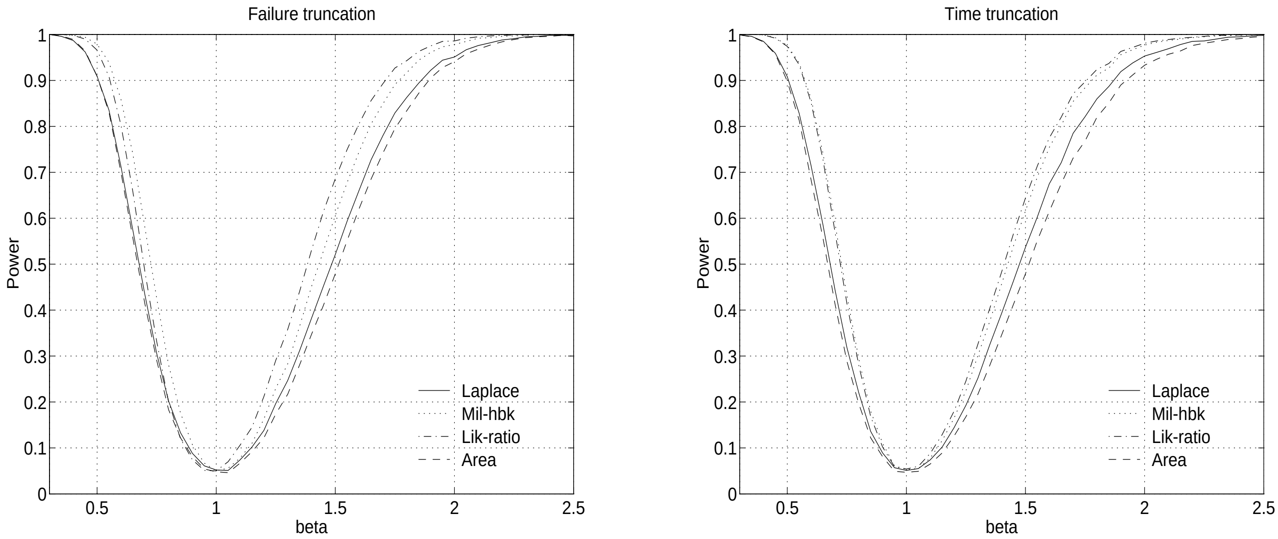

Figure 7: Simulations of a NHPP($t^{\beta-1}$), where β=beta=$\{0.30, 0.35, \ldots, 2.5\}$.
10000 replications for each beta-value. Failure truncation with 35 failures and time truncation with expected number of failures in each simulation 35.

When samples of size 35 or expected size 35 are simulated, see Figure 7, exactly the same effects as in the case of size 15 are observed, albeit the differences between the tests are certainly smaller.

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

Data have been simulated from NHPP's 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, $\Lambda(t_1) = \frac{t_1}{2}(b+1)$, in phase II, $\Lambda(t_2) - \Lambda(t_1) = t_2 - t_1$ and in phase III, $\Lambda(t_2) - \Lambda(t_1) = \frac{\tau - t_2}{2}(c+1)$.

Both time and failure truncated processes are simulated, the time truncated processes are truncated at time τ and 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

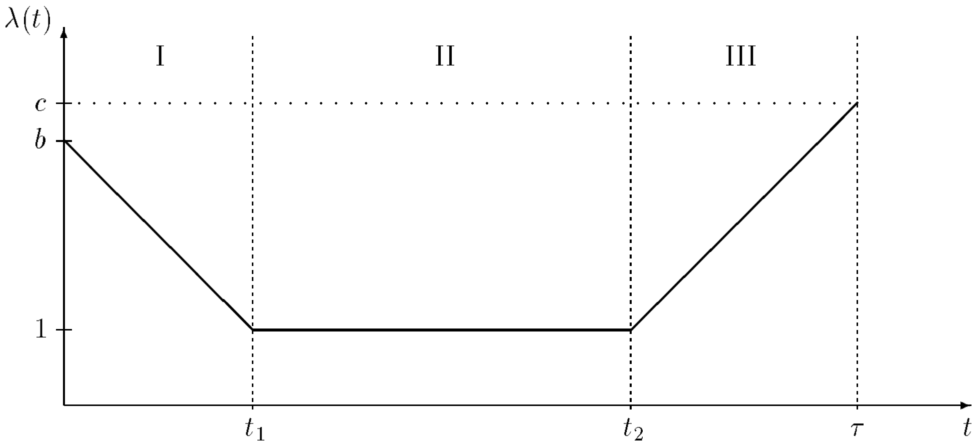

Figure 8: Example of bathtub shaped intensity function

Function	Slope		Expected number of failures		
	Phase I	Phase III	Phase I	Phase II	Phase III
1	2	2	8	8	8
2	1	1	8	8	8
3	$\frac{1}{2}$	$\frac{1}{2}$	8	8	8
4	2	$\frac{1}{2}$	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 1: Description of twelve bathtub shaped intensity functions

time(s) are larger than τ in the case of a failure truncated process, the intensity function in phase III is used for $t > \tau$ as well. Results of the simulations are given in Table 2 and Table 3

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	0.30	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
Lik-ratio	0.28	0.20	0.13	0.34	0.11	0.42	0.19	0.14	0.10	0.39	0.16	0.33
Area	0.50	0.36	0.23	0.38	0.19	0.77	0.38	0.21	0.17	0.49	0.42	0.45

Table 2: Simulated power with bathtub shaped intensity function number 1 to 12.
10000 replications for each intensity function. Failure truncated processes.

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	0.39	0.29
Mil-hbk	0.31	0.22	0.16	0.42	0.14	0.43	0.21	0.16	0.10	0.33	0.15	0.47
Lik-ratio	0.29	0.20	0.14	0.40	0.13	0.42	0.20	0.15	0.10	0.36	0.16	0.44
Area	0.40	0.26	0.16	0.33	0.13	0.71	0.31	0.14	0.13	0.46	0.50	0.44

Table 3: Simulated power with bathtub shaped intensity function number 1 to 12.
10000 replications for each intensity function. Time truncated processes.

The area test is clearly the test which generally is most powerful in these situations. The other tests are rather powerful in some of the considered situations, but have very low power in other situations, making them unsuitable as tests against bathtub trend. Except for the likelihood-ratio test all the tests are generally less powerful when the processes are time truncated.

5.2 Several processes

Now we proceed to consider the case $m > 1$. There is of course a huge number of situations to consider, with varying number 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.

We have simulated data from $m = 5$ independent processes which 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 the intervals equals 5, but with different starting points for each interval. Hence, the expected total number of failures is 25, a symbolic illustration is given in Figure 9. The position and length of each interval on the *time axis* will of course depend on intensities in each process.

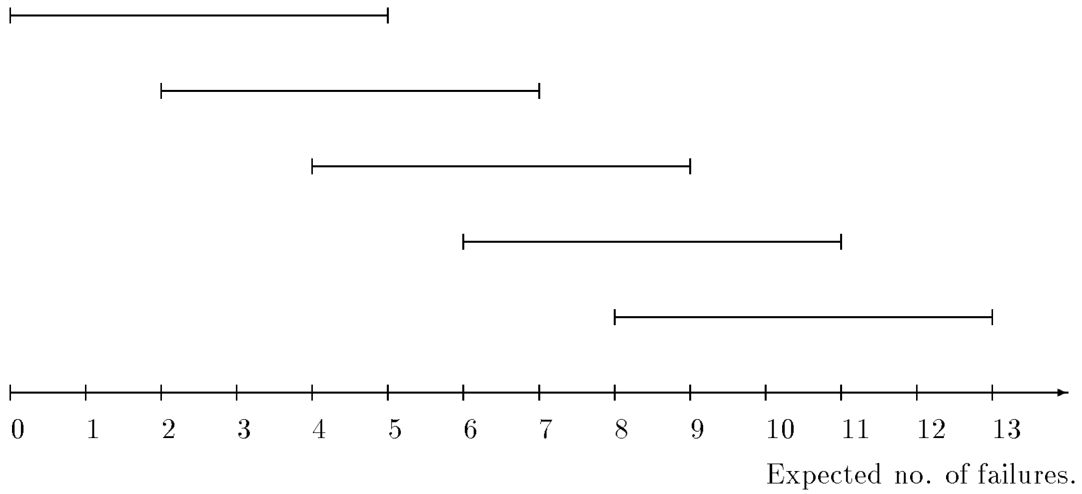

Figure 9: Picture of expected number of failures in the processes. In case of a HPP the picture will be equal on the time axis, otherwise different.

5.2.1 No heterogeneities.

First we consider the case where each process is simulated from NHPP's with identical intensity functions $\lambda(t)$ for $t \geq 0$. The i't 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 i't interval is as indicated in Figure 9. Data are simulated from both a log-linear and a power law intensity function, and the results are shown in Figure 10. The abbreviations used now are, *Laplace* – The combined Laplace test (2), *Laplace-TTT* – The TTT-based Laplace test (3), *Mil-hbk* – The combined Military Handbook test (5), *Mil-hbk-TTT* – The TTT-based Military Handbook test (6), *Area* – Area-based test (11).

We realize that the TTT-based tests (the area test, the TTT-based Laplace test and the TTT-based Military Handbook test), are far more powerful than the combined Laplace test and the combined Military Handbook test. The relationship between the TTT-based tests in the power law case are exactly the same as the relationship between the area test, Laplace test and Military Handbook test shown in Figures 6 and 7 in the one process case. For the log-linear case the relationship is the same as seen in the time truncated case on Figure 4. Also notice that, in contradiction to the picture on Figures 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.

Data from several NHPP's with the same bathtub shaped intensity functions are simulated as well. Five processes are simulated, and all the twelve 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 Figure 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 far more powerful than the combined Laplace and Military Handbook tests. The area test is the best test, but the TTT-based Military Handbook test is fairly powerful as well.

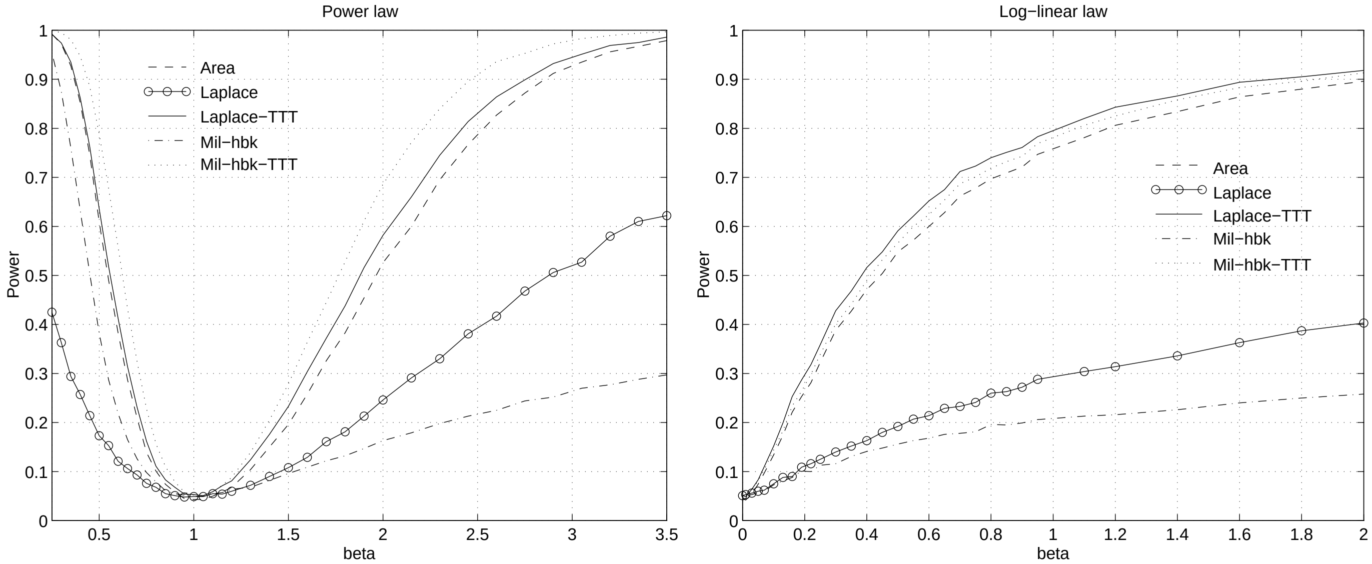

Figure 10: Simulations of a NHPP with intensity function $\lambda(t) = t^{\beta-1}$, where β=beta=$\{0.25, 0.30, \ldots, 3.5\}$, and from a NHPP with intensity function $\lambda(t) = e^{\beta t}$, where β=beta=$\{0, 0.01, 0.03, \ldots, 0.40, 0.45, \ldots, 1, 1.2, \ldots, 2.0\}$
10000 replications for each beta-value. Expected total number of failures is 25.

The TTT-based Laplace test on the other hand is not very much better than the combined Laplace test.

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	0.12	0.11	0.10	0.21	0.10	0.12	0.12	0.09	0.09	0.71	0.33	0.49
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	0.57	0.42	0.28	0.71	0.25	0.40	0.41	0.29	0.16	0.43	0.15	0.80
Area	0.77	0.54	0.31	0.61	0.24	0.98	0.72	0.21	0.21	0.76	0.64	0.70

Table 4: Simulated power with bathtub shaped intensity function number 1 to 12.
10000 replications for each intensity function. Five processes observed.

5.2.2 Heterogeneities.

The results from the previous section 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 HPP's with equal intensity functions, while the combined Laplace test and the combined Military Handbook test allow heterogeneities in the HPP's under the null hypothesis of no trend. To study the effect of such heterogeneities, data are simulated from five processes with observation interval for each process chosen as indicated in Figure 9. NHPP's with intensity functions $\alpha_i \beta t^{\beta-1}$ and $e^{\alpha_i + \beta t}$ are simulated, where the α_i's are varying from process to process, while the β's are common for all processes. The results and choice of parameter values are given in Figure 11. Notice that with our choices of parameter values, the heterogeneities in the log-linear intensity functions are more than twice as big as the heterogeneities in the power law intensity functions.

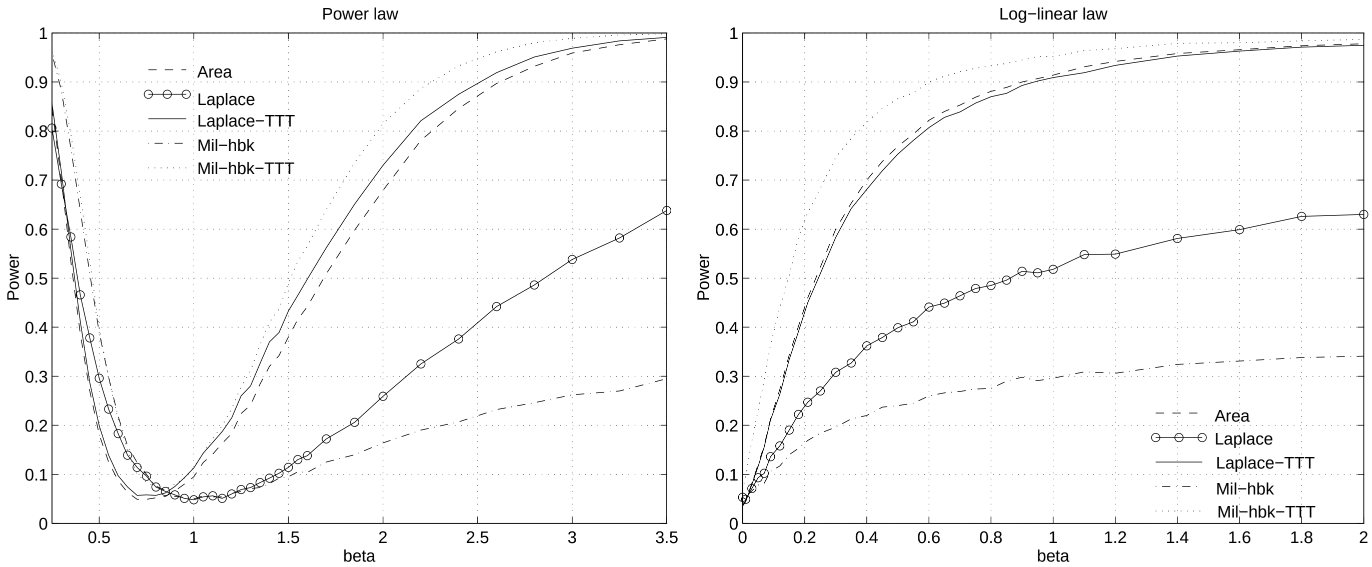

Figure 11: Simulations of a NHPP with intensity function $\lambda(t) = \alpha_i t^{\beta-1}$, where $\alpha_1 = 0.5$, $\alpha_2 = 0.625$, $\alpha_3 = 0.75$, $\alpha_4 = 0.875$, $\alpha_5 = 1.0$, β=beta={0.25,0.30,...,3.5}, and from a NHPP with intensity function $\lambda(t) = e^{\alpha+\beta t}$, where $\alpha_1 = -1.5$, $\alpha_2 = -0.9$, $\alpha_3 = -0.5$, $\alpha_4 = -0.2$, $\alpha_5 = 0$, β=beta={0,0.01,0.03,...,0.40,0.45, ...,1,1.2,...,2.0}
10000 replications for each beta-value. Expected total number of failures is 25.

The picture in Figure 11 clearly displays the problem with the TTT-based tests; they are not constructed to allow heterogeneous intensity functions. In the power law case the power functions are skewed to the left and the tests achieve an actual level exceeding 0.1. In the log-linear case the TTT-based Military Handbook test achieves a too *high* level while the area test and the TTT-based Laplace test achieve a too *low* level. 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 HPP's with unequal intensity functions were performed. The observation intervals indicated in Figure 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 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 the area test seems to have the best overall properties.

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	0.048	0.051	0.052	0.048	0.041	0.049
Laplace-TTT	0.355	0.031	0.116	0.100	0.070	0.053
Mil-hbk	0.052	0.049	0.050	0.048	0.047	0.052
Mil-hbk-TTT	0.030	0.054	0.114	0.092	0.071	0.054
Area	0.431	0.037	0.100	0.087	0.062	0.045

Table 5: Simulated actual level with various heterogeneous HPP's. Five processes. 10000 replications.

6 Case Study

This is a simple example to demonstrate the use of the trend-tests. A dataset presented by Barlow & Davis [6] 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 Figure 12. The TTT-plot clearly indicates an increasing trend in the data. Results of the

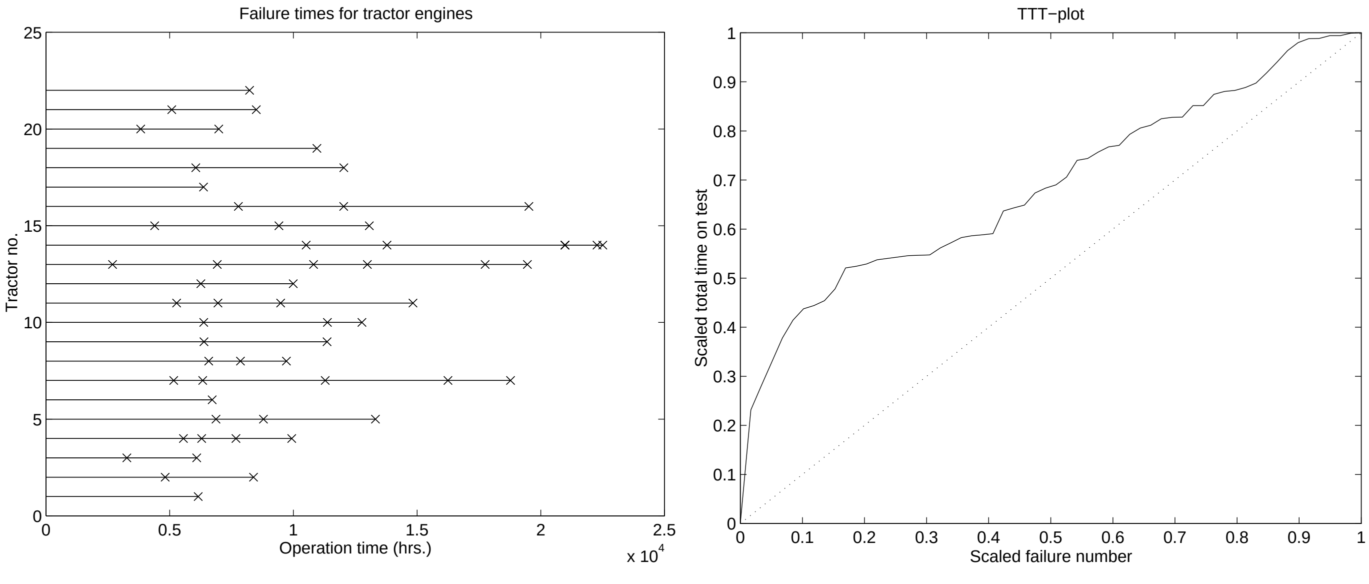

Figure 12: Failure times and TTT-plot for tractor data.

statistical trend tests are presented in Table 6. All the trend tests reject the null hypothesis on a 5% level, hence we safely conclude that there is an increasing failure trend in the tractor data. Even if there are heterogeneities the HPP assumption is rejected since the combined Laplace test and the combined Military Handbook reject the null hypothesis.

With a maximal number of 6 observed failures it is obviously difficult to analyze the

Test	Test statistic	p-value
Combined Laplace	2.056	0.0398
Laplace-TTT	5.008	0.0000
Combined Mil-hbk	42.80	0.0027
Mil-hbk-TTT	48.69	0.0000
Area	2.567	0.0000

Table 6: Analysis of tractor data.

tractors individually. If each tractor is analyzed individually, the Laplace test and the Military Handbook test reject the null hypothesis only for tractor number 14 and the area test never rejects the HPP assumption. But assuming a common trend for all the 22 tractor engines the conclusion is a significantly increasing trend.

7 CONCLUSIONS

7.1 Description of each test

The properties of each test is summarized below.

The Laplace test

The Laplace test is for single processes in general the most powerful test against NHPP with log-linear intensity function. In our study it was only the third most powerful test against NHPP with power law intensity, and it has in general low power against NHPP's with bathtub shaped intensity functions.

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

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 intensity functions.

The Military Handbook test

The Military Handbook test is for single processes the most powerful test against NHPP's with decreasing power law intensity function, and the second most powerful test against increasing power law intensity function. It was among the two least powerful tests against NHPP's 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.

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.

The likelihood-ratio test

The likelihood-ratio test is the most powerful test against NHPP's with increasing power law intensity functions. If the process is failure truncated, the test is "skew" with low power against weakly decreasing trend. Against NHPP's with log-linear intensity function and 15 failures, the likelihood-ratio test is the most powerful test in the case of a failure truncated process with not too strong trend, and it is always more powerful than the Military Handbook test against the considered NHPP's with log-linear intensity function. The likelihood-ratio test has generally low power against NHPP's with bathtub shaped intensity functions.

The area test

The area test is generally the most powerful test against NHPP's with bathtub shaped intensity functions. The actual level equals the nominal level only for large samples. For sufficiently small samples, the actual level of the area test drops below 4%. Low actual level means a "defensive" test and this necessarily leads to a too low power, too. In the single process case it is the least powerful test against power law trend, but only the Laplace test is more powerful in all cases of log-linear trend, and the area test is the most powerful test against bathtub shaped trend. In the case of several processes, the area test compares in a similar way to the TTT-based versions of the other tests.

7.2 Final comments

It is obvious that no test is superior to the other tests in all situations. Consider the single process case. Against a NHPP with log-linear intensity function, the Laplace test is the best test. Against a NHPP with increasing power law intensity function the likelihood-ratio test is the most powerful test. The Military Handbook test is the most powerful test against NHPP's with decreasing power law intensity function and the area test has the largest power against NHPP's with bathtub shaped intensity functions. If we want power against increasing trend the likelihood-ratio test should be used, and if we want power against non-monotonic trend the area test should be used.

If more than one process is observed we must decide whether we want to test the null hypothesis of equal HPP's or if we want to allow for heterogeneous HPP's. In the first case, or with minor deviations from the first case, the area test and the TTT-based Military Handbook test have the best "all over" properties. In the last case the combined Laplace test or the combined Military Handbook test has to be used.

References

[1] Abramowitz, M & Stegun, I. A. *Handbook of Mathematical Functions.* Dover, New York, 1970.

[2] Andersen, P. K., Borgan, Ø., Gill, R. & Keiding, N. *Statistical Models Based on Counting Processes.* Springer Verlag, New York, 1992.

[3] Anderson, T. W. & Darling, D. A., Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *Ann. Math. Statist.*, **23** (1952), 193-212.

[4] Ascher, H. & Feingold, H. *Repairable Systems Reliability. Modeling, Inference, Misconceptions and Their Causes.* Marcel Dekker, Inc., New York, 1984.

[5] Bain, L. J., Engelhardt, M., & Wright, F. T. Tests for an Increasing Trend in the Intensity of a Poisson Process: A Power Study. *J. Am Statist. Ass.*, **80** (1985) 419-422.

[6] Barlow, R. E. & Davis, B., Analysis of Time Between Failures for Repairable Components. In: *Nuclear Systems Reliability Engineering and Risk assessment.* (J. B. Fussell & G. R. Burdick, eds.) SIAM, Philadelphia, 1977, 543-561.

[7] Barndorff-Nielsen, O. E. & Cox, D. R., Bartlett Adjustments to the Likelihood Ratio Statistic and the Distribution of the Maximum Likelihood Estimator. *J. Roy. Statist. Soc.*, **B 46** (1984) 483-495.

[8] Billingsley, P. *Convergence of Probability Measures.* John Wiley & Sons, New York, 1968.

[9] Cox, D. R. & Lewis, P. A. W. *The Statistical Analysis of Series of Events.* Methuen, London, 1966.

[10] Crowder, M. J., Kimber, A. C., Smith, R. L. & Sweeting, T. J. *Statistical Analysis of Reliability Data.* Chapman & Hall, 1991.

[11] Durbin, J. Distribution theory of tests based on the sample distribution function. *Reg. Conf. Series in Appl. Math.*, 1973, p. 32

[12] Gill, R. D., The Total Time on Test Plot and the Cumulative Total Time on Test Statistic for a Counting Process. *Ann. Statist.*, **14** (1986) 1234-1239

[13] Høyland, A. & Rausand, M., *System Reliability Theory. Models and Statistical Methods.* Wiley, New York, 1994.

[14] Jensen, J. L., A Historical Sketch and Some New Results on the Improved Log Likelihood Ratio Statistic. *Scand. J. Statist.*, **20** (1993) 1-15.

[15] Lindqvist, B., Kjønstad, G. A. & Meland, N., Testing for trend i repairable systems data. *Proceedings of ESREL '94.* La Baule, France, May 30. - June 3. 1994.

[16] Ross, S. M. *Stochastic Processes* John Wiley & Sons, New York, 1983.

[17] Wolfram, S. *Mathematica. A System for Doing Mathematics by Computer.* Addison-Wesley, Reading, MA, 1991.

[18] Aarset, M. V., How to Indentify a Bathtub Hazard Rate. *IEEE Transactions on Reliability.*, **R-36** (1987) 106-108.