Reliability analyses of workstation failure data

Sara Brocklehurst, Karama Kanoun, Jean-Claude Laprie, Bev Littlewood, Sylvain Metge, Peter Mellor, A. Tanner

CENTRE NATIONAL DE LA
RECHERCHE SCIENTIFIQUE

LABORATOIRE D'ANALYSE ET
D'ARCHITECTURE DES SYSTEMES

RELIABILITY ANALYSES OF WORKSTATION FAILURE DATA

S. BROCKLEHURST, K. KANOUN, J.C. LAPRIE, B. LITTLEWOOD,
S. METGE, P. MELLOR, A. TANNER

LAAS REPORT 91172

MAY 1991

Paper to be presented at
ESPRIT Conference 1991 — Brussels, Belgium — 25-29 November 1991

Project Nr 3092 PDCS

RELIABILITY ANALYSES OF WORKSTATION FAILURE DATA

S. BROCKLEHURST*, K. KANOUN**, J.C. LAPRIE**, B. LITTLEWOOD*, S. METGE**, P. MELLOR* and A. TANNER*

* The City University
Northampton Square
London EC1V OHB
UK

** LAAS-CNRS
7, Avenue du Colonel Roche
31077 Toulouse Cedex
FRANCE

SUMMARY

Experience of applying reliability models in the past has shown that the relative predictive performance of the models depends entirely on the context. It has been found that there is no one model that performs well over all data sets. It has also been found that for some data sets all models are in error. In such cases two techniques for improving predictive accuracy have been shown to be beneficial: i) recalibrating the raw model predictions and ii) using the results of trend tests preliminary to applying the models. These two techniques may be used separately or in combination. This paper is mainly devoted to the first technique but we will also show the benefit to be gained by the application of the second technique. We apply a number of reliability models to some failure data collected from a workstation installed at City University, together with the recalibration technique, and assess the performance of the resulting prediction systems.

1. INTRODUCTION

There are now many software reliability models available for the assessment and prediction of the failure behaviour of a program (or system) undergoing debugging or in operation. Some are universally bad, while the performance of others varies from data set to data set, but *it is not possible to select a globally good model* which will perform well over even a particular class of systems. This has led to the development of techniques which assess the predictive performance and compare the relative merits of these models in a particular context (ie. when applied to the data set under investigation). These techniques allow us to apply a number of models to the data of interest and select a "best" model in this context.

One of the techniques used also allows us to assess one aspect of the error in each of the models (called raw models hereafter), a kind of "bias", and to recalibrate the raw model predictions with respect to this error. This recalibrated prediction system is truly predictive and we can therefore use the above analysis techniques to assess the improvement gained via recalibration every time it is applied. It has proved to be beneficial (5) and the computational effort required for this technique is negligible compared with the effort required to apply many of the initial model prediction systems. It is recommended, therefore, that it is applied as a matter of course to all raw model prediction systems. The approach we suggest is that, for each data set, we should apply a number of raw models (preferably as many as is practical), and the recalibration technique to each of the raw models, and then use the analysis techniques to compare predictive accuracy of all the resulting prediction systems.

When the trend in the data exhibits *changes*, another technique consists of applying reliability growth models on the periods of trend between such changes. Application of trend tests to the collected data allows these reliability growth, reliability decay and stable periods to be identified, and thus the data to be partitioned into subsections. Reliability growth models may then be applied to each subsection displaying trend in accordance with the model

assumptions, thus improving the accuracy of the results (12), (13), (2).

One of the major barriers facing research in the area of software reliability modelling is the lack of the data required in order to make predictions about the reliability of a program or system. Collecting the appropriate data can be a costly process and testing has to be conducted under a suitable operational profile (ie. a user profile) in order to get accurate reliability predictions. In this paper we shall analyse a new set of data which shows the failure behaviour of a single-user work station, installed at the City University in March 1985 and subsequently used over a period of four years.

2. RAW RELIABILITY MODELS

The data we are considering are times, $t_1,t_2,t_3,...$ (which can be C.P.U. execution time, calendar time, or any other applicable measure) between successive failures resulting from first occurrence of unique faults of a system. In the case of perfect debugging this will correspond to the times between successive failures arising from different faults. For the data presented in this paper corrections were (generally) not carried out; only the first failure arising from each fault was extracted in order to obtain the required inter-failure times. This effectively simulates what *would* happen under perfect debugging.

For simplicity we shall be considering one-step-ahead predictions. Using the previous inter-failure times, $t_1,t_2,...,t_{j-1}$ the raw models provide a prediction of the current (and as yet unobserved) inter-failure time, T_j, in the form of a *predictive cumulative distribution function* (*cdf*),

$$\hat{F}_j(t) = \hat{Pr}\,(T_j \leq t) \qquad .....(2.1)$$

From *(2.1)* we also have a *predictive probability density function* (*pdf*) for T_j,

$$\hat{f}_j(t) = \hat{F}'_j(t) \qquad .....(2.2)$$

These can be thought of as estimates of the true underlying *cdf* and *pdf*, $F_j(t)$ and $f_j(t)$, respectively.

The models considered in this paper are *parametric* models; that is, they assume a form for the *pdf* (*cdf*) which depends on some unknown parameter(s). Estimates for these parameters are made at each stage, j, by using the *previous* data and the method of maximum likelihood. These parameters are then substituted into the *pdf* (*cdf*) in order to make predictions about T_j. Note that performance of these models will depend not only on their precise mathematical structure, but on the maximum likelihood inference technique and the substitution rule for prediction. It should be noted that these last two approaches to statistical inference and prediction are chosen here for convenience: other techniques, such as Bayesian inference, tend to be computationally intensive.

The models which are applied in this paper are the Duane *(D)* (10), (11), Hyperexponential *(HE)* (18), (19), Keiller Littlewood *(KL)* (15), (16), Littlewood *(LM)*, (20) Littlewood non-homogeneous Poisson process *(LNHPP)* (1), Littlewood Verrall *(LV)* (21) and Musa Okumoto *(MO)* (24) models. Since most of these models are well known the details are omitted from this paper.

Many of the parametric models have the property that they tend towards simpler models as their parameters (or functions of their parameters) tend to extreme values. In particular the *HE, MO, LM* and *LNHPP* models tend to the stable reliability homogeneous Poisson process (*HPP*) when they are applied over regions of data which do not exhibit growth. There are other ways in which these models can tend to simpler solutions, details of which are summarised in (6).

All the models considered in this paper can model stable reliability or reliability growth (ie. constant failure rate or monotonically decreasing failure rate) while the *DU, LV* and *KL* models can also model reliability decay (ie. monotonically increasing failure rate). None of the models as they are applied in this paper, however, are able to model trend *changes* (for example the transition from stable reliability to reliability growth), since they all assume a smooth trend. However, even though each fitting of the model will be unable to represent a

change in the trend of the data, a sequence of predictions from the model may respond to such a change. That is, when applying a model over a succession of one-step-ahead predictions upon a data set containing changes in trend, it may be that the series of predictions *themselves* do not exhibit smooth trend.

3. ANALYSIS OF PREDICTIVE QUALITY AND RECALIBRATION

In this section the methods for assessing and comparing the performance of the various models are briefly described. They all depend on a comparison between the estimated *cdf* or *pdf* (see *(2.1)* and *(2.2)*) at stage j, and the (later observed) realisation, t_j, of the next inter-failure time, T_j. Suppose we have q inter-failure times altogether. Then we can apply the raw models summarized in section 2, to the data $t_1,...,t_{j-1}$, to obtain our prediction for T_j, for $j = s,...,q$, say, where s is a suitably large number. We then have what we shall refer to as our "prediction system" for each of the models, ie.

$$\{\hat{F}_j(t), \hat{f}_j(t) \; ; \; j = s,...,q\} \qquad .....(3.1)$$

The *u*-plot

Our first technique involves substituting the (later observed) t_j into the (earlier computed) predictive *cdf*:

$$u_j = \hat{F}_j(t_j) \qquad j = s,...,q \qquad .....(3.2)$$

If our prediction system is identical to the truth, ie. $\hat{F}_j \equiv F_j \; \forall j$, then the u's should behave as if they come from a $U(0,1)$ distribution. The u-plot is the sample *cdf* of the u's defined in *(3.2)* (1). Departures of the u-plot from the $45°$ line, which is the $U(0,1)$ *cdf*, indicate that our prediction system is inaccurate in some way. Informally, the u-plot is a powerful means of detecting various kinds of *consistent bias* in predictions, ie. situations where the prediction errors are in some sense *stationary*.

The *y*-plot

Let

$$x_j = log(1-u_j) \qquad j = s,...,q \qquad .....(3.3)$$

and

$$y_r = \frac{\sum_{j=s}^{r} x_j}{\sum_{j=s}^{q} x_j} \qquad r = s,...,q \qquad .....(3.4)$$

Then the y-plot is constructed similarly to the u-plot by drawing the sample *cdf* of the y's. Note, from *(3.3)* and *(3.4)*, that the y-plot preserves the order of occurrence of the u's over time whereas the u-plot does not. If our prediction system has captured the trend (eg. reliability growth) in the data then the y-plot should be close to the $45°$ line. Thus the y-plot is a means of detecting those cases where the prediction errors are *non-stationary*.

For a more detailed explanation of the u- and y-plots see (1), (9), (15) and (16).

The prequential likelihood ratio

It should be noted that it is possible for a prediction system to give good u- and y-plots and yet still be inaccurate; for example it could be very *noisy*, so that individual predictions emanating from it are very inaccurate even though on average there is no bias, and there is no evidence of non-stationarity in the errors of prediction. For this reason we use a further measure called the prequential likelihood ratio (*PLR*) which is intended as global comparison of goodness of prediction for one prediction system versus another.

Suppose we have two prediction systems, α and β, say. Then the *PLR* is defined to be

$$PLR = \frac{\prod_{j=s}^{q} \hat{f}_j^{\alpha}(t_j)}{\prod_{j=s}^{q} \hat{f}_j^{\beta}(t_j)} \qquad \ldots\ldots(3.5)$$

Notice that, unlike the *u*- and *y*-plots, this measure depends upon the *pdfs* rather than the *cdfs*.

If $PLR \to \infty$ as $q \to \infty$ then we would choose model α as being the better of the two models. Conversely if $PLR \to 0$ as $q \to \infty$, we would favour model β over model α. As we clearly never have $q \to \infty$ in practice, the best we can do is to look for steady increases and decreases in our *PLR* plots of one model versus another over the whole data set. In the analysis that follows we actually plot the *log(PLR)* at stage *j* against *j*. To see a more detailed explanation of the use of the *PLR* in the context of software reliability modelling see (1).

The recalibration technique

If we have captured the trend in the data, ie. the prediction errors are stationary, but the predictions are biased then the (approximately) consistent departure of the prediction system from the truth is represented by the departure of the *u*-plot from the $45°$ line. This is where the *recalibration technique* can use the *u*-plot to eliminate this bias from the raw prediction system and thus construct a new improved prediction system. This technique consists of using the joined up *u*-plot, G_i, based on *previous predictions* of $t_s,\ldots,t_{i-1}$, in order to adjust the current raw prediction,

$$\hat{F}_i^*(t) = G_i(\hat{F}_i(t)) \qquad \ldots\ldots(3.6)$$

For more details of this technique see (3), (5) and (17). We can repeat this recalibration procedure over $i = p,\ldots,q$, where p-s is suitably large and we then have a recalibrated prediction system,

$$\{\hat{F}_i^*(t),\hat{f}_i^*(t)\ ;\ i = p,\ldots,q\} \qquad \ldots\ldots(3.7)$$

We can then use the *u*- and *y*-plots and the *PLR* as outlined above in exactly the same manner as with the raw prediction system to assess our new recalibrated prediction system.

Since we are using the *PLR* to compare our prediction systems we will use the spline recalibration technique (see (7) and (8)) which consists of smoothing each G_i using least squares cubic splines before recalibration. This smoothing is necessary since discontinuity in the derivative of G_i causes the *PLR* to report badly about the resulting recalibrated predictions, although most predictions of interest will be altered little by smoothing (3).

All the raw prediction systems from the reliability models referred to in section 2 are spline-recalibrated. An *S* added onto the end of the model names will be used to denote the recalibrated prediction system (eg. *HES*, *MOS*, *LVS* etc.). This results in a number of different prediction systems for each data set, the performance of which we can compare using techniques for the analysis of predictive accuracy described above.

Preliminary data processing

A global method incorporating preliminary data processing has been investigated in (13) and (14). Such preliminary data processing makes it possible to identify outliers (doubtful data) and to test for changes in the trend in order that the reliability growth models can be applied to appropriate subsections of the data. Here we shall consider only this latter example of preliminary processing.

Many software reliability models assume that the inter-failure time data exhibits reliability growth. It is certainly the case that if we are truly debugging our system we would, on average, expect to see reliability growth. However, it has been observed that, particularly

early on in the operational life of a system, growth may not be present in spite of attempted bug-fixing, or even that there is reliability decay. One explanation of this phenomenon is that, early on in the life of a system, we are continually exploiting new parts of the system. Since we are not frequently reusing the same parts of the system, we might not expect to see significant growth, even though fixes are taking place. Other reasons include fault dependency, varying delays between identification and removal of faults, and even small specification changes.

As mentioned in section 2, some of the models considered in this paper can handle reliability decay as well as stable reliability or reliability growth. Change points in the trend in the data, however, for example from stable reliability to reliability growth, may present difficulties for all these models and result in inaccurate predictions. It is therefore important to be able to identify those subsections of the data within which there are no trend-change points. One method of determining whether such a subsection of data really does exhibit reliability growth (or decay, or stability) uses the Laplace statistic. For inter-failure time data $t_1,t_2,...,t_n$. this is

$$L(t_1,t_2,...,t_n) = \frac{\frac{\sum_{j=1}^{n-1} \tau_j}{(n-1)} - \frac{\tau_n}{2}}{\tau_n \sqrt{\frac{1}{12(n-1)}}} \qquad .....(3.8)$$

where $\tau_j = \sum_{r=1}^{j} t_r$.

Large negative values of this suggest there is reliability growth, large positive values decay. Since the statistic is approximately normally distributed (9), under the null hypothesis that the data come from a (trend-free) homogeneous Poisson process, very simple tests can be conducted for growth (or decay) in a particular data vector. More informally, it is often instructive to plot the changes in the statistic as the data vector increases in size, as we shall see later.

4. THE DATA SET

The data used in this paper comes from operational use of a single-user work station which was installed at the City University on the 18th March, 1985.

Data collected included real (or wall-clock) time of occurrence of each failure (recorded to the nearest minute), together with the identity of the particular fault which caused the failure (so that time to first occurrence of each unique fault can be recovered and hence the required inter-failure times). Additionally details of the type of usage and the version of the operating system in use at the time of each failure, the severity of failure, and type of the associated fault were recorded..

In order to assess software reliability by statistical analysis of time of occurrence of first failure due to each fault, it is necessary to have a measure of the amount of use to which the software has been subjected (22). Real, or wall-clock, time is rarely appropriate. In this case it was decided that "hands-on" time at the work station was suitable. This was because the failures were recorded as a result of observation by the user, and because many of them were "usability problems", ie., the encountering of features of the system which caused difficulty for the user, even though the system was behaving according to specification.

For reasons of space we shall analyse here only two subsets of this data. The largest, *USBAR*, consists of times between successive failures of certain unique faults (ie. ignoring failures of previously seen faults). Almost all failures are included here, including usability problems: we omit only those occurring during power-on, power-off and machine idle time (i.e. machine on but not being used). This data comprises *397* inter-failure times in all. From this data set we have extracted a subset, *TSW*, comprising only *software* failures; there are 129 of these. A more comprehensive analysis of this data, involving other subsets, can be found in (6).

760.	758.	303.	6.	22.	14.	42.	4.	84.	15.	221.	14.	15.	41.	1.	153.
409.	54.	24.	44.	180.	397.	19.	145.	36.	54.	1337.	163.	8.	1.	17.	16.
87.	19.	29.	0.5	300.	360.	10.	11.	100.	252.	460.	179.	3.	24.	253.	163.
54.	137.	328.	3.	9.	12.	18.	9.	75.	15.	366.	428.	212.	115.	264.	269.
276.	1.	999.	30.	495.	472.	344.	550.	131.	47.	92.	863.	991.	35.	9549.	249.
607.	83.	614.	352.	673.	4179.	111.	75.	407.	288.	894.	1314.	845.	55.	409.	36.
15.	1960.	60.	19.	20.	79.	24.	1737.	7984.	10.	20.	338.	250.	1682.	212.	287.
56.	4973.	3500.	59.	98.	2439.	1812.	6203.	385.	3500.	4892.	687.	62.	2796.	3268.	3845.
76.															

Table 1: Data set *TSW* (total number=*129*)

Table 1 shows the raw data for *TSW*. Space restrictions preclude our showing a similar table for *USBAR*, but Fig 1 shows this plotted as a cumulative number of failures against total elapsed time. It is clear that in each case there is *overall* reliability growth. The times in the later part of the table tend to be bigger on average than earlier ones; the slope of the plot in the figure (representing the rate of occurrence of failures) shows an overall decrease.

However, within this overall pattern of growth, there is some variation. This is clear even from the raw data of *USBAR* plotted in Fig 1. A more detailed analysis via the Laplace statistic is shown in Figs 2 and 3. In Fig 2 the Laplace statistic starts to show a consistent trend downwards about halfway through the data set; in Fig 3, there seems to be reliability growth after only the first few observations.

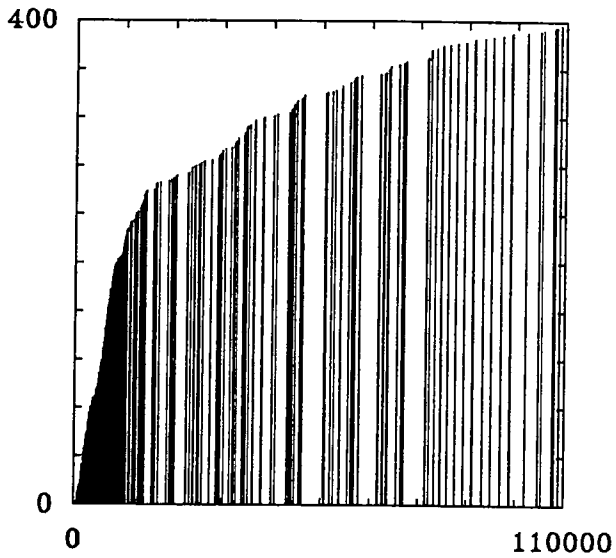

Fig. 1: Cumulative number of failures against total time for data set *USBAR*.

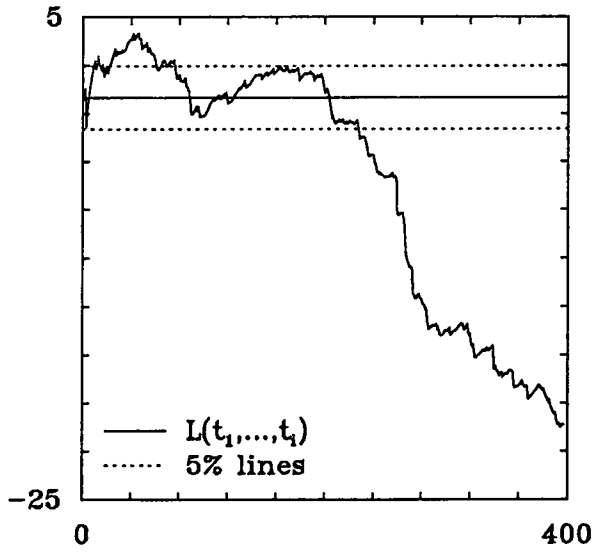

Fig. 2: Laplace statistic against failure number for data set *USBAR*.

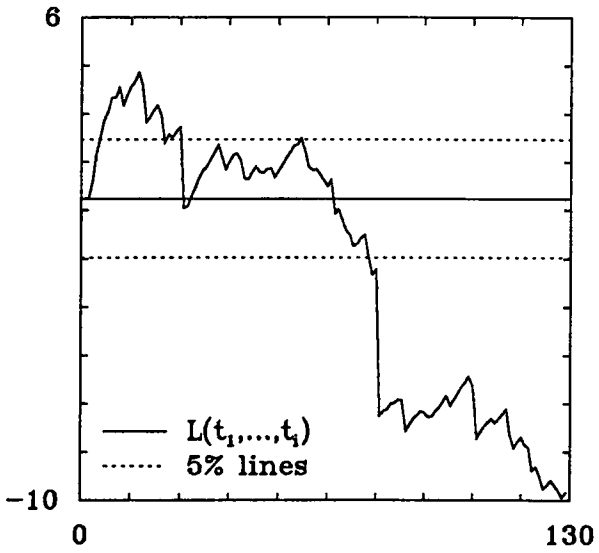

Fig. 3: Laplace statistic against failure number for data set *TSW*.

5. DATA ANALYSIS USING SEVERAL PREDICTION SYSTEMS

Due to space limitations and since we are mainly interested in model recalibration in this paper, all the models referred to in section 2 are first applied blindly (ignoring the Laplace statistic analysis on the various data sets) over the *whole* of the data available. However, in order to show the impact of taking into account this analysis, we shall apply one of the models, *HE*, over a period of time which takes account of the fact that in each of the data sets reliability growth does not start at the beginning of the data set. We shall use *HEc* to denote this additional prediction system: the model is applied over the range, $t_{c+1},t_{c+2},\ldots$. That is $t_{c+1},\ldots,t_{c+s}$ are used to make predictions about T_{c+s+1}, then $t_{c+1},\ldots,t_{c+s+1}$ are used to make predictions about T_{c+s+2}, and so on. An informal choice of c = 144 for *USBAR* and c = 14 for *TSW* was made from the preliminary trend analysis.

In this section we shall discuss application of the *7* parametric models referred to in section 2 (including the second application of the *HE* model, *HEc*), each both raw and spline-recalibrated, to data sets *USBAR* and *TSW*. The quality of the *16* resulting prediction systems, with respect to one-step-ahead predictions, will be discussed for each data set based on the *u*-plot, the *y*-plot and the prequential analysis techniques described in section 3.

In the application of the raw parametric models we have to decide on the number of inter-failure times which will be used in order to get the first prediction. If this number is too small then early predictions will be likely to be too noisy. On a purely informal basis we have chosen to use the first *20* inter-failure times to get the first raw prediction. We also have to decide at what point to begin recalibrating our raw predictions. If we recalibrate too early then the *u*-plot for recalibration will contain too few points and our estimate of the *cdf* of the current *u*, even in the presence of stationarity in the raw model errors, may be inaccurate, and hence recalibration may not be very efficient. We have, again informally, chosen to use the first *15* raw predictions in the *u*-plot to achieve the first recalibrated prediction. Thus, in the notation of section 3, $s = 21$ and $p = 36$. In the case of the second application of the *HE* model (ie. *HEc*) the resulting predictions (raw and recalibrated) will start later on in the data than those from all the other prediction systems.

Table 2 shows the significance levels (see (23)) for the *Kolmogorov-Smirnov* distances of the *u*- and *y*-plots from the *45°* line for the various prediction systems on the data sets. The Kolmogorov-Smirnov (*K-S*) distance is the maximum vertical distance of the plot from the *45°* line. Here we show in lower case the levels for the range of the *HEc* model predictions, ie. T_{c+37}, $T_{c+38},\ldots$, whilst the upper case refers to the *u*- and *y*-plots which include predictions of $T_{36},T_{37},\ldots$. The *u*- and *y*-plots for the shorter range of predictions are included since we wish to compare these models with *HEc*. Raw predictions prior to the stage of the first recalibrated predictions (ie. the first *15* raw predictions) are not included in the plots since we want to compare recalibrated and raw predictions over the same range of data.

	USBAR				*TSW*			
	c = 144				*c = 14*			
	RAW		*S*		*RAW*		*S*	
	y	*u*	*y*	*u*	*y*	*u*	*y*	*u*
HE	F	F	Fe	F	Db	F	A	Cd
HEc	f	f	b	f	b	f	a	d
MO	F	F	Df	Df	Bd	F	Df	A
DU	Fa	F	Fe	Df	A	F	Ac	A
LM	F	F	F	Dc	F	F	F	A
LNHPP	F	F	F	Dc	F	F	F	A
LV	Fa	F	Fa	Cf	A	F	A	A
KL	Ea	F	Cb	Af	A	F	A	A

A: insignificant at 20% B: 10-20% C: 5-10%
D: 2-5% E: 1-2% F: significant at 1%

The upper case letters represent the significance levels for $p = 36$ while the lower case letters represent the significance levels with $p = c + 37$ (if they differ).

Table 2: *u*- and *y*-plot significance levels for predictions of $T_p,T_{p+1},\ldots$.

Data set USBAR

From table 2 it can be seen that both the *u*- and *y*-plots for the various raw models on this data set are highly significant with the exception of the *y*-plots for the *DU, LV* and *KL* models over the latter half of the data set, where the *y*-plots are insignificant. The median predictions in Fig 4 show great disagreement between the models at the end of the data set.

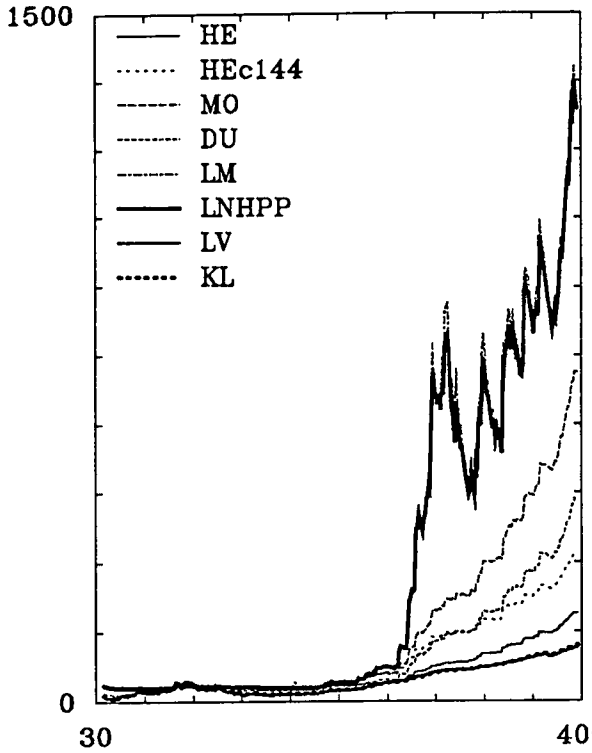

Fig. 4: Median predictions from the raw models for data set *USBAR*.

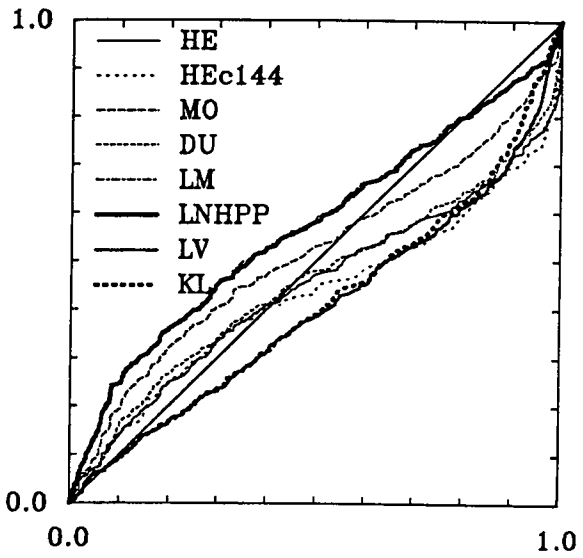

Fig. 5: u-plots from the raw models for data set *USBAR* for predictions of T_{181}, T_{182}, .. T_{397} for *HEc* and T_{36}, T_{37}, .. T_{397} for the remaining models.

The corresponding u-plots for these models (see Fig 5) indicate that the *LM* and *LNHPP* models are generally optimistic, while the *LV* and *KL* models are generally pessimistic. For the other models the bias in the distributions would appear to be more complicated than simple optimism and pessimism. Over most of the data prior to the 200^{th} failure the *HE*, *MO*, *LM* and *LNHPP* models tend to *HPP* predictions (notice, from Fig 2, that this coincides with the region where the Laplace statistic indicates absence of global reliability growth).

In the case of *HE* and *HEc*, the y-plot compared over the latter half of the data gives $K\text{-}S$ distances of 0.152 and 0.124, respectively, indicating that not considering the earlier data may indeed have enabled this model to better capture the trend in the later failure data. Indeed, Fig 6, which shows the *log(PLR)* plot for all these raw models against the *DU* model, as the data evolves, shows how dramatic this improvement is from $i \approx 270$. This point may correspond to a second change point (see change of slope of Laplace statistic plot in Fig 2) and from this point the *HE* model experiences great difficulties, whereas *HEc* does not. This plot also suggests that beyond this point the *MO*, *LM*, *LNHPP*, *LV* and *KL* models may be the better predictors, but it is clear, from the u-plots, that all these raw models are in error.

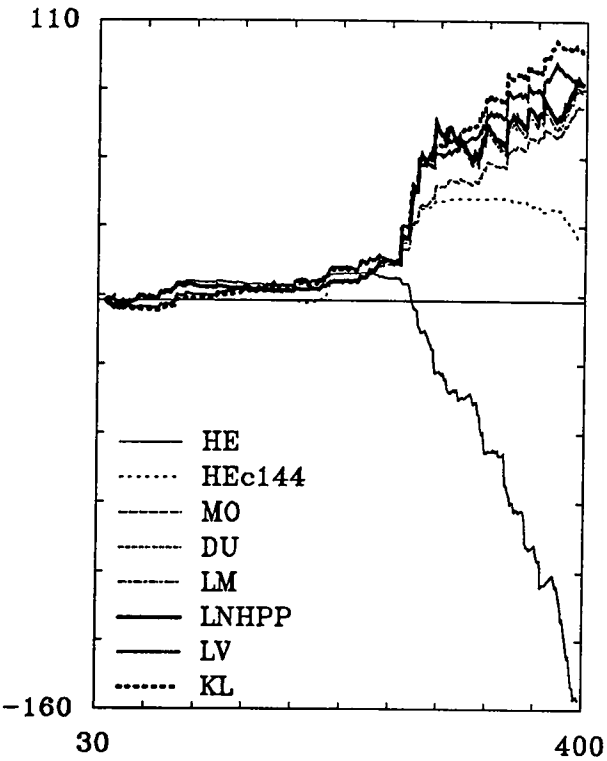

Fig. 6: *Log(PLR)* for the raw models versus *DU* for data set *USBAR*.

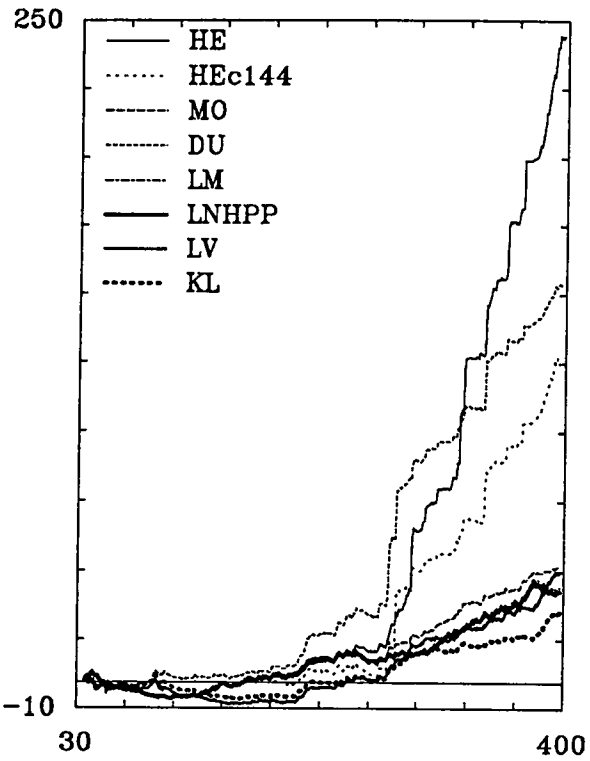

Fig. 7: *Log(PLR)* for the recalibrated versus raw prediction systems for data set *USBAR*.

Fig 7, which shows the *log(PLR)* plot of the recalibrated predictions versus the raw predictions, indicates that there is a huge improvement to be gained via recalibration in the latter part of the data set for those raw models which were initially performing particularly badly (ie. *HE* and *DU*) and for *HEc* there is also dramatic improvement, while for the remaining models there is slight improvement. Comparison of figures 5 and 8 show that there is dramatic improvement in the *u*-plots via recalibration, although in the case of the *HE* model it would seem that there is still pessimism in the recalibrated predictions. It is interesting to observe (see Fig 9[1]) that the recalibrated models which are now performing the best, according to the *PLR*, do not come from those raw models which were initially performing the best. In particular, the *DU* model, before recalibration, was steadily worse than the others (with the exception of *HE*), while after recalibration it

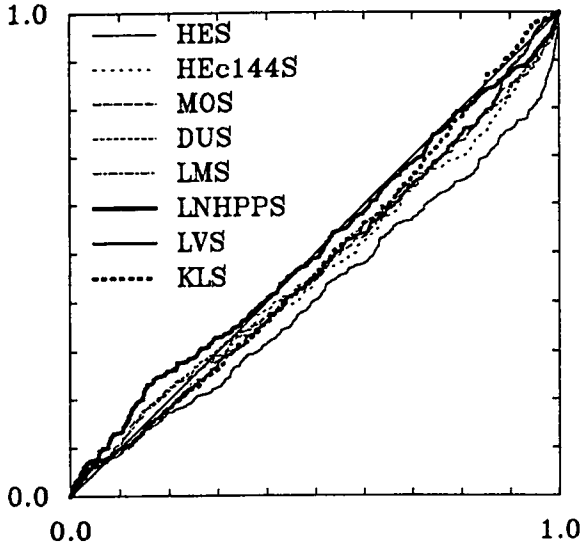

Fig. 8: *u*-plots from the spline-recalibrated models for data set *USBAR* for predictions of T_{181}, T_{182}, .. T_{397} for *HEcS* and T_{36}, T_{37}, .. T_{397} for the remaining models.

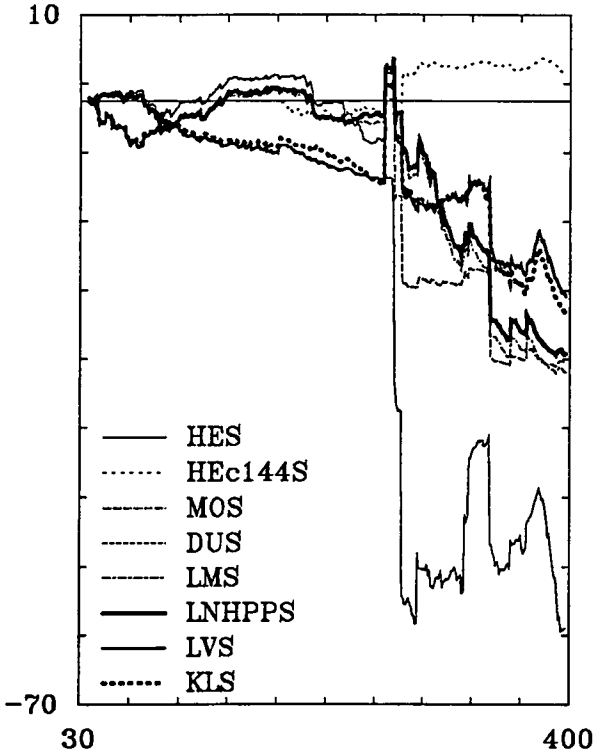

Fig. 9: *Log(PLR)* for the spline-recalibrated models versus *DUS* for data set *USBAR*.

[1] The large jumps downwards at $i = 267$ and $i = 273$, for *HES*, coincide with unusually large inter-failure times, t_{267} and t_{273}, and each *HES* predictive distribution clearly has a smaller right hand tail compared with the other recalibrated models. Comparison with Figs 6 and 7 suggest that this problem was present in the raw model predictions and is not an artefact of recalibration. Fig 7 suggests that this problem was also present for the raw *DU* model, but that it was eliminated by recalibration.

is one of the better predictors together with *HEc*. Fig 9 also shows the benefit from applying the *HE* model (*HEc*) on data from *145*; the improvement in the recalibrated *HEc* predictions over the recalibrated *HE* predictions indicates that the raw *HEc* model has indeed better captured the trend in the data. Comparison of figures 6 and 9 show that, in general, recalibration has had the effect of bringing the models into much closer agreement and the recalibrated median predictions in Fig 10 are indeed in closer agreement than the raw medians. Comparison of figures 4 and 10 shows that some of the model medians have been adjusted for optimism, while others have been adjusted for pessimism. Although these are closer in agreement than the raw predictions they still diverge as the data evolves.

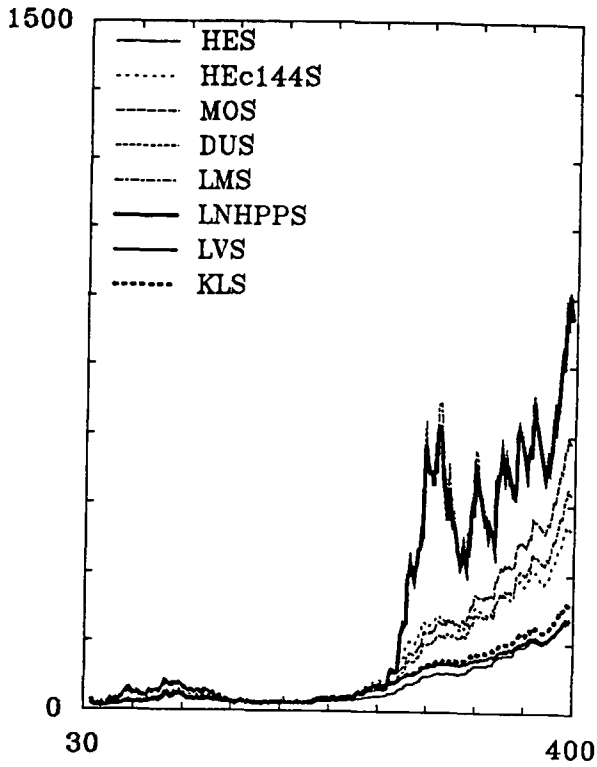

Fig. 10: Median predictions from the spline-recalibrated models for data set *USBAR*.

Data set TSW

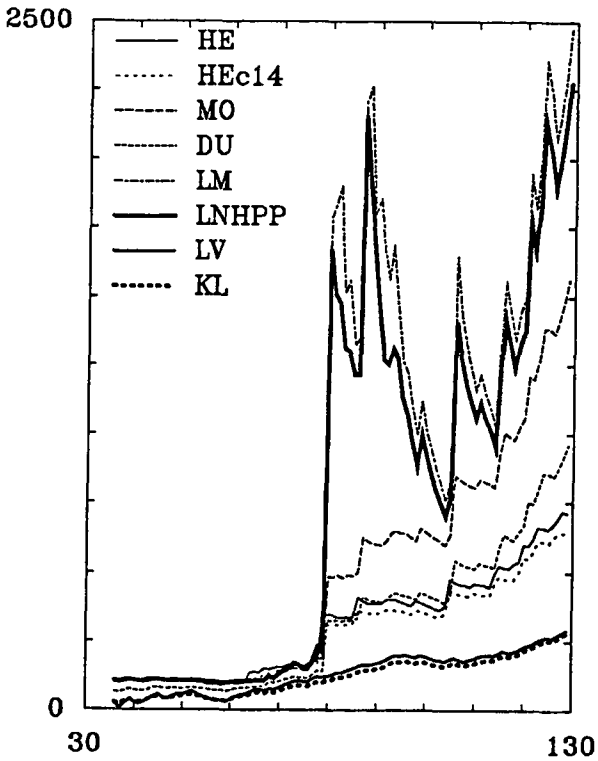

Fig. 11: Median predictions from the raw models for data set *TSW*.

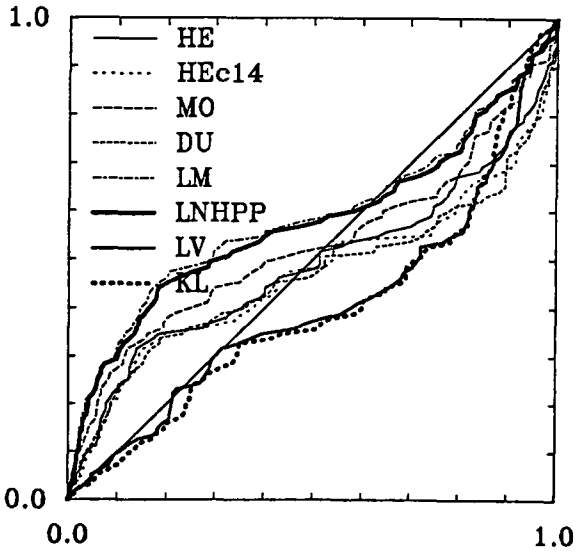

Fig. 12: *u*-plots from the raw models for data set *TSW* for predictions of T_{51}, T_{52}, .. T_{129} for HEc and T_{36}, T_{37}, .. T_{129} for the remaining models.

For this data set it can be seen from Fig 11 that the medians again diverge as the data evolves. Prior to the point at which these predictions diverge, the medians from some of the raw models appear to be coincident. This is because up to $i \approx 66$ the *HE*, *MO*, *LM* and *LNHPP* generally tend to an *HPP*; comparing this with Fig 3 reveals that this again coincides with absence of global reliability growth. From table 2 it can be seen that all the *u*-plots for the raw models are again highly significant and Fig 12 suggests that the *LM* and *LNHPP* median predictions are optimistic while the *LV* and *KL* median predictions are pessimistic. The *log(PLR)* plot in Fig 13 shows that there is little to choose between the various raw prediction systems over much of the data set with the exception of the large jump at $i = 79$[2].

For this data set we are again interested in the recalibrated predictions since the raw predictions are clearly in error. After recalibration we can see, from table 2 and Fig 14 that there is improvement in all the *u*-plots and most of them have become insignificant, while the recalibrated medians shown in Fig 15 are now in much closer agreement. It is surprising, at first, that the *HE* and *HEc* median predictions are so close to the others while the *u*-plot significance levels are so different from those of the other models. In fact, this is due to the *K-S* distances for these models being a result of pessimistic recalibrated predictions for *large* inter-failure times (ie. in the upper part of the distribution), while the *u*-plot near *0.5* is quite close to the *45°* line (see Fig 14).

Fig 16 shows the improvement gained by recalibration, for the various models, according to the *log(PLR)*; for the *DU* model, in particular, there seems to be a consistent improvement over the whole data set and for the other models there is improvement later on in the data while earlier on there is little to choose between the recalibrated and raw predictions. The *log(PLR)* plot in Fig 17 shows that, after recalibration, there is little to choose between the various prediction systems over much of the data set with the exception of the large jump at $i = 79$.

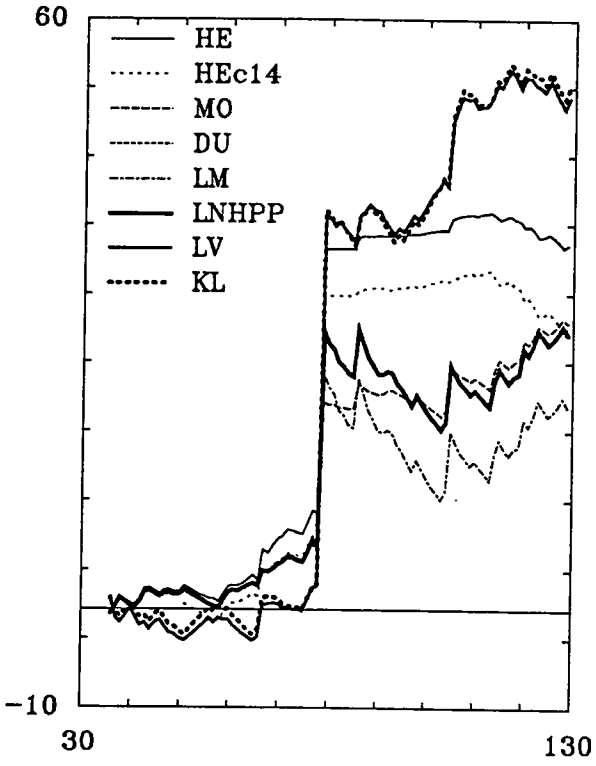

Fig. 13: *Log(PLR)* for the raw models versus *DU* for data set *TSW*.

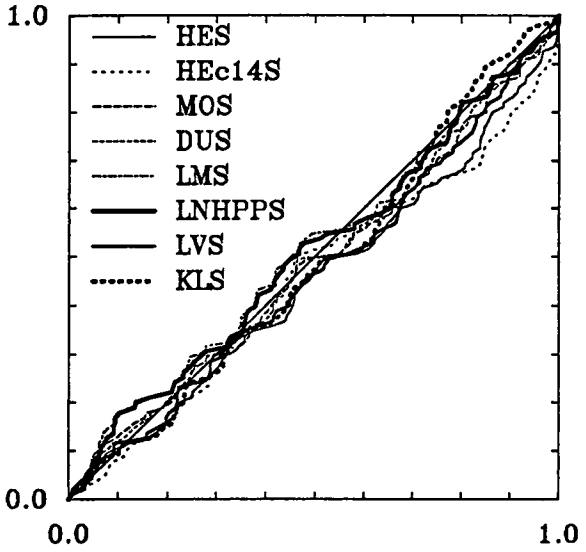

Fig. 14: *u*-plots from the spline-recalibrated models for data set *TSW* for predictions of $T_{51}, T_{52}, .. T_{129}$ for *HEcS* and $T_{36}, T_{37}, .. T_{129}$ for the remaining models.

2 The large jump upwards in this plot at $i = 79$ is again due to the occurrence of a particularly large inter-failure time ($t_{79} = 9549$). The differences in performance here for the different models are also present in the recalibrated predictions and so in this case recalibration has done little to help.

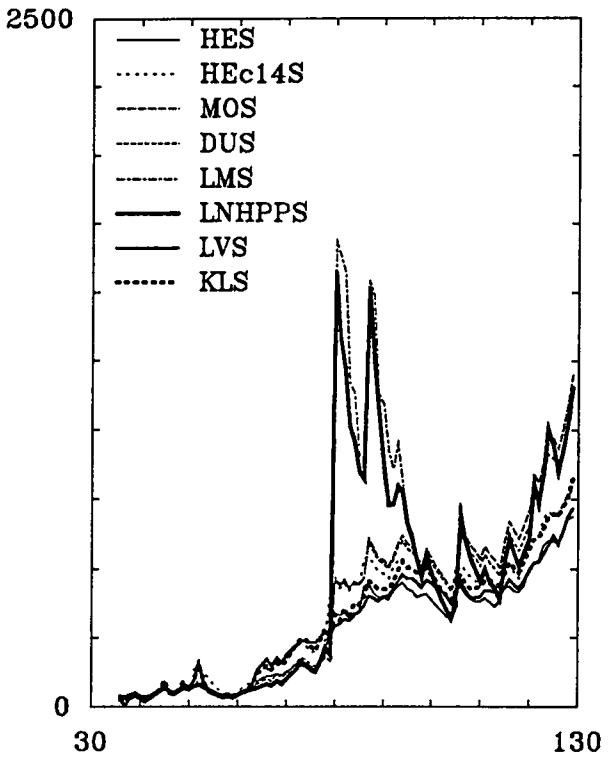

Fig. 15: Median predictions from the spline-recalibrated models for data set *TSW*.

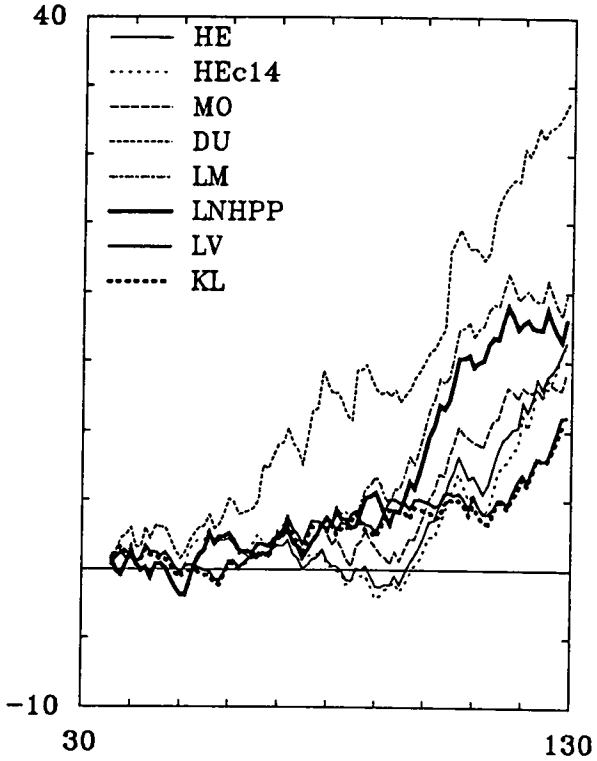

Fig. 16: *Log(PLR)* for the recalibrated versus raw prediction systems for data set *TSW*.

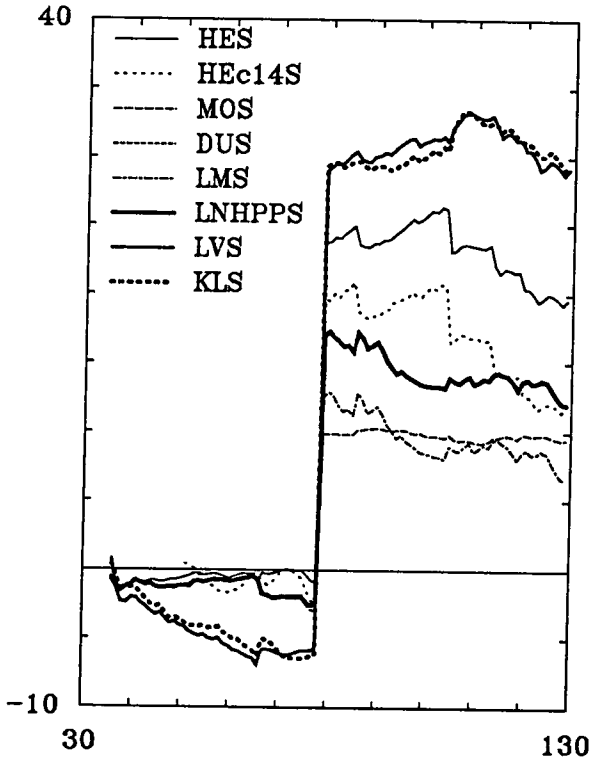

Fig. 17: *Log(PLR)* for the spline-recalibrated models versus *DUS* for data set *TSW*.

6. GENERAL COMMENTS

For the two data sets selected it is clear that recalibration is effective in improving on the raw model predictions, particularly later on in the data. It also seems that raw models which are initially comparatively bad, tend to be vastly improved by this technique so that, after recalibration, they are comparable with (and sometimes better than) the recalibrated versions of models which were initially better. This means that it may not always be necessary to apply a sophisticated model, when a cruder model, with recalibration, will do better. As stated earlier, the only requirement we wish in the initial raw model, is that its predictions (as opposed to how well it fits the data in retrospect) have captured the trend in the data. In this situation we can expect recalibration to be efficient. These results support those reported earlier in (5).

There is also evidence that the comparative performance of the prediction systems may change not only for different data sets, but for different ranges of predictions within one data set. This is particularly evident for the raw model predictions where performance for *USBAR* diverges at the end of the data while earlier in the data the various raw model predictions are closer in agreement. After recalibration, although performance of the resulting prediction systems is in closer agreement, there is still evidence that relative performance is different over different intervals of data.

For each of the data sets, there appears to be a change point before which the models are comparable in their performance, and after which their performance diverges. These points also coincide with the stage at which recalibration starts to give steady (and often dramatic) improvement for most of the models. It is clear that, for these data sets, the change point in the data creates difficulties for some of the raw models; similar situations have also been noticed in (12). It also seems likely that such change points may affect the efficiency of recalibration, since it may cause non-stationarity in the departure of the raw model predictions from the truth. For data set *USBAR*, for example, although the u-plots (and the predictions) are vastly improved by recalibration it can be seen that they are mostly still poor. In fact the significance levels vary depending on which range of data is included in the plots (see table 2). This indicates non-stationarity in the error in the recalibrated models which is likely to be the result of change points in the data. In such situations recalibration may be expected to improve the predictions, but still leave room for further improvement.

There are alternative ways that might further improve predictions. In (4; 6) the issues of applying raw reliability models and the recalibration technique using *moving windows* of fixed sizes across the data and the raw model predictions, respectively, are investigated. Alternatively the raw models and/or the recalibration technique could be applied using trend test results, in a similar manner to *HEc*. The latter technique is likely to be more efficient since using fixed moving windows may cause data to be thrown away unnecessarily, resulting in less bias in the predictions but at the expense of increased noise. For raw model application we wish to find windows of data which have stationary trend in order to optimise resulting predictive accuracy, whereas for recalibration we wish to find windows of raw model predictions for which the *prediction errors* are stationary.

For simplicity, in this paper, the analysis of the prediction systems resulting from the various models and from the recalibration technique, is based upon one-step ahead predictions. It is clear that there are many other types of predictions, further into the future, which may be of interest to the user and that the prediction system which gives the best one-step ahead predictions will not necessarily give the best predictions further into the future. This means that if we wish to use the techniques described in section 3, in order to assess the performance of such predictions, we ideally need to use predictions of the same *type* as the prediction of interest. It follows then, that we also need to use the same type of prediction in the u-plot for recalibration of the prediction of interest. On a single data set, it is not clear how to extend these techniques to predictions which are far into the future, simply because we are unlikely to see enough (if any) realisations of such predictions. Issues of how we achieve such predictions for the raw models, how we assess the accuracy of such predictions, and how we might recalibrate such predictions, need to be addressed in future work.

REFERENCES

(1) ABDEL GHALY, A.A., CHAN, P.Y., and LITTLEWOOD, B. (1986). "Evaluation of Competing Software Reliability Predictions", *IEEE Trans. on Software Engineering*, vol. SE-12, no.9, pp. 950-967.

(2) BASTOS MARTINI, M.R., KANOUN, K. and MOREIRA DE SOUZA, J. (1990). "Software Reliability Evaluation of the TROPICO-R Switching System", *IEEE Trans. on Reliability*, vol. 39, no. 3, pp. 369-379.

(3) BROCKLEHURST, S. (1987). "On the Effectiveness of Adaptive Software Reliability Modelling", C.S.R. Technical Report.

(4) BROCKLEHURST, S. (1989) "A Non-Parametric Approach to Software Reliability Modelling", C.S.R. Technical Report.

(5) BROCKLEHURST, S., CHAN, P.Y., LITTLEWOOD, B. and SNELL, J. (1990). "Recalibrating Software Reliability Models", *IEEE Trans. on Software Engineering*, vol. SE-16, no. 4, pp. 458-470.

(6) BROCKLEHURST, S., MELLOR, P. and TANNER, A. (1991). "A Multi-Modelling Approach to Software Reliability Prediction", C.S.R. Technical Report, in preparation.

(7) CHAN, P.Y. (1986). "Software Reliability Prediction", Ph.D. dissertation, City University.

(8) CHAN, P.Y., LITTLEWOOD, B. and SNELL, J. (1985). "Parametric Spline Approach to Adaptive Reliability Modelling", C.S.R. Technical Report.

(9) COX, D.R. and LEWIS, P.A.W. (1966). *Statistical Analysis of Series of Events*, Methuen, London.

(10) CROW, L.H. (1977). "Confidence Interval Procedures for Reliability Growth Analysis", U.S. Army Material Syst. Anal. Activity, Aberdeen, MD, Tech. Report 197.

(11) DUANE, J.T. (1964). "Learning Curve Approach to Reliability Monitoring", *IEEE Trans. Aerospace*, vol. 2, pp. 563-566.

(12) KANOUN, K. and SABOURIN, T. (1987). "Software Dependability of a Telephone Switching System", *Proc. 17th IEEE Int. Symp. on Fault-Tolerant Computing (FTCS-17)*, pp. 236-241, Pittsburgh, Pennsylvania, 6-8 July.

(13) KANOUN, K., LAPRIE, J.C. and SABOURIN, T. (1988). "A Method for Software Reliability Growth Analysis and Assessment", *Proc. of Int. Workshop on Software Engineering and its Applications*, pp. 859-878, Toulouse, France, 5-9 Dec.

(14) KANOUN, K. (1989) "Software Dependability Growth characterization, modeling and evaluation", Doctorat ès-Sciences thesis, Institut National polytechnique de Toulouse, LAAS report n° 89-320, Sept.; in French.

(15) KEILLER, P.A., LITTLEWOOD, B., MILLER, D.R. and SOFER, A. (1983). "On the Quality of Software Reliability Predictions", *Proc. of NATO ASI on Electronic Systems Effectiveness and Life Cycle Costing*, (ed. J. Skwirzinski), pp. 441-460, Springer-Verlag, Heidelberg, Germany.

(16) KEILLER, P.A., LITTLEWOOD, B., MILLER, D.R. and SOFER, A. (1983). "Comparison of Software Reliability Predictions", *Proc. 13th IEEE Int. Symp. on Fault-Tolerant Computing (FTCS-13)*, pp. 128-134.

(17) KEILLER, P.A. and LITTLEWOOD, B. (1984). "Adaptive Software Reliability Modelling", *Proc. 14th IEEE Int. Symp. on Fault-Tolerant Computing (FTCS-14)*, pp. 108-113.

(18) LAPRIE, J.C. (1984). "Dependability Modelling and Evaluation of Software and Hardware Systems", *Invited survey to the 2nd GI/NTG/GMR Conference on Fault-tolerant Computing*, Bonn, Germany, September 19-21.

(19) LAPRIE, J.C., KANOUN, K., BEOUNES, C. and KAANICHE, M. (1991). "The KAT (Knowledge-Action-Transformation) Approach to the Modeling and Evaluation of Reliability and Availability Growth", *IEEE Trans. on Software Engineering*, vol. SE-17, no. 4, pp. 370-382.

(20) LITTLEWOOD, B. (1981). "Stochastic Reliability Growth: A Model for Fault Removal in Computer Programs and Hardware Design", *IEEE Trans. on Reliability*, vol.R-30, no.4, pp. 313-320.
(21) LITTLEWOOD, B. and VERRALL, J.L. (1973). "A Bayesian Reliability Growth Model for Computer Software", *J. Royal Statist. Soc. C (Applied Statistics)*, vol.22, pp. 332-346.
(22) MELLOR, P. (1986). "Software Reliability Data Collection: Problems and Standards", *Software Reliability: A state of the Art Report*, Pergamon Infotech Ltd., London, pp. 165-181.
(23) MILLER, L.H. (1956). "Table of Percentage Points of Kolmogorov Statistics", *American Statistical Association Journal*, pp. 111-121, March.
(24) MUSA, J.D. and OKUMOTO, K. (1984). "A Logarithmic Poisson Execution Time Model for Software Reliability Measurement", *Proc. Seventh International Conference on Software Eng.*, IEEE Comp. Soc. New York, pp. 230-238.

Acknowledgments

This research was supported by the ESPRIT Basic Research Action PDCS (SB, KK, J-CL, BL, SM) while the data collection and extraction was supported by the Alvey SRM project SE-073 (PM, AT).