Robust Identification of Reliability Models

Pavel Kovanic and Petr Volf[1]

[1]The Institute of Information Theory and Automation, Czechoslovak Academy of Sciences, Pod vodárenskou věží 4, Prague 8, 182 08 Czechoslovakia
Fax: (42)(2)847452 Phone: (42)(2)8152230 Telex: 122018 atom c

Robust Identification of Reliability Models

Summary

A new technique for data–based modelling of reliability has been developed including significant innovations. No a priori statistical model of random data components is assumed; the distribution function estimate results from the analysis together with estimates of reliability model parameters. All results are robust not only with respect to a priori statistical model (since there is none) but also to outliers and peripheral data clusters. Both uncensored and censored data are taken in account for estimation. The method has been successfully implemented to testing reliability of truck components but its applicability to another survival problems can be expected.

KEY WORDS: Robust modelling Distribution free reliability model Nonstatistical uncertainty model Gnostical method

Introduction

Problems of quality closely connected with reliability and surviving have recently become the object of top interest in many fields of science, health care and technology. Well–known statistical software packages and systems (e. g. BMDP, S plus and others) keep step with the development of new methodologies. As a rule, they incorporate new results in data analysis with only a small delay. However, the contemporary survival analysis is prevailingly devoted to the biostatistics, reflecting its needs and specific features (cf. the series of procedures starting from the Cox's regression model, development in the field of intensity-based models, etc.). The progress is slower when it comes to new methods penetrating into the field of reliability testing. Moreover, the rapidly expanding application fields enable the practical verification of existing methods and the evaluation of both their merits and drawbacks.

We wish to model (and to estimate) the distribution of time to failure (survival or life–time) for some particular device. The analysis is based on statistical data, i. e. on results of repeated experiments. The aim of such a modelling is the evaluation and prediction of the reliability of devices in dependence on various conditions or variables (covariates), e. g. on the load, production technology etc. The task involves thus the estimation of the model for this dependence. From the statistical point of view, the task is formulated as a problem of nonlinear regression analysis. We suppose that our task possess some standard features of nonlinear regression, namely:

1. It has the structure of a positively valued regression function $t(z, \theta)$, $(z > 0, \theta \in R_M)$ of a covariate z (load) and of an unknown (vector) parameter θ that has to be estimated.

2. The $k-$th observed value of the time to failure for the load Z_k is supposed to have the multiplicative model $T_k = t(Z_k, \theta) \cdot \rho_k$, where the ρ_k (the residual) states for the $k-$th

realization of the (scalar, positive) random variable modelling the data uncertainty. The model may also be considered in its logarithmized form, i. e. $Y_k = y(Z_k, \theta) + \varepsilon_k$.

3. The analyst has also to consider the presence of censored data (characterized by an indicator I_k). It means that some experiment has been interrupted before the failure occurs. The $k-th$ result then yields the information 'time to failure is greater than T_k' ($I_k = 0$) instead of 'time to failure equals T_k' ($I_k = 1.$). Even such a limited information is valuable and the method has to incorporate it adequately into the computing procedure. Technically, we consider the scheme of random censoring from the right side.

Simultaneously, our further assumptions try to reflect some nonstandard, but realistic conditions:

4. The distribution law of the variable ρ is not specified, only the existence of its continuous density function is assumed. We cannot expect any 'popular' form of distribution function (d.f.), neither symmetry nor similar simplifying property.

5. The unknown distribution function is not necessarily defined on an infinite domain. Unknown bounds of their support may also be object of the estimation.

6. There may be restrictions on values of some or all estimated parameters dictated by the nature of the process under consideration.

7. The reliability experiments are often time-consuming and expensive. That is why the number of data may be rather small. No asymptotical results of statistical theory can thus be utilized for our inference.

8. Robustness with respect to a priori statistical assumptions is attained by not using such assumptions at all. However, results of reliability experiments are as a rule widely spread. It is therefore necessary to apply procedures robust also with respect to outliers.

There exists large and developed methodology for the nonlinear regression analysis (as an example, consider [2], or the works concerning the robust regression, e. g. [7]). The geometric interpretation is presented, the possibilities of approximation are discussed. Nevertheless, essentially the methods have a few sources: maximal likelihood, least squares approach, or the minimalization of another divergence of results from regression curve.

From the claims and assumptions summarized above it follows immediately that the commonly used and standard statistical approaches are not convenient for solution of the formulated task. The maximum likelihood method (or the method of moments, as well as the Bayesian approach) are too closely connected with assumptions about the type (or a family) of fixed d.f.s of residuals (i. e. of the variable ρ). However, a theoretical basis for the choice of a particular distribution function is not always available for reliability studies.

Most of the standard procedures for regression analysis are based on the least squares method. They result in atlases of families of regression functions and ensure the computing ability of procedures and optimality of iterations. Unfortunately, a successful utilization of

these procedures needs at least the local symmetry of distribution of residuals. Moreover, these methods are nonrobust to outlying data.

Some contemporary methods of robust statistical analysis can be now found in modern software systems. However, even their usage is limited because the necessity to accept some a priori assumptions is not removed completely and validity of results is conditioned by the validity of these assumptions. Such methods are often based on the assumption of a 'main' distribution having a given form that is contaminated by another one. Such a model is ordinarly not applicable to reliability problems.

Main ideas of the new approach

To satisfy mentioned requirements, it was necessary to develop a new method. As a nonstatistical alternative, the *gnostical* estimates of distribution functions (cf. [5] and [1]) have been applied. These d.f. are parametrized only by data (residuals R are data in our case) and by a positive scale parameter s that is not necessarly constant. In the considered case, it was supposed $s = S_0 * exp(S_1 * R)$ with some unknown constants S_0 and S_1. The real domain of the d.f. is supposed to have a general form $\mathcal{D} := (R_L, R_U)$ with some unknown – finite or infinite – bounds R_L and R_U. The parameter vector to be estimated was therefore $\Theta = \langle S_0, S_1, R_L, R_U, A, B, C, D, E, \rangle$, where $A, .., E$ were parameters of the regression curve. The process included following steps:

1. Setting or calculating initial conditions as first – rough – approximation to Θ.
2. Evaluation of model residuals using uncensored data.
3. Calculating of point estimates of the empirical d.f. (PEDF) of residuals of these data.
4. Correcting the PEDF with respect to censored data.
5. Evaluation of next approximation to Θ using an optimization procedure and the condition of improving fit of the PEDF by the gnostical d.f. This step thus modifies both surviving model and the d.f. of residuals.
6. Repetition from 2) till stabilization of results.
7. Substitution of resulting Θ into procedures for estimation of quantiles and of probability of life–times.

A natural question may arise why such a procedure should converge to a reasonable solution. The explanation is connected with a unique feature of the gnostical d.f. of *global* type: This d.f. is robust with respect to outlying data and even to peripheral data subclusters. The better Θ, the better fit, the lower value of the criterion function that evaluates fitting errors with dominating weights of 'central' multiplicative residuals close to unit.

Let us consider these points in more detail.

Steps of the solution

Initial estimation of parameters

As a first approximation, the experience from previous runs of programs can be used. If there is no such experience, then the standard solution of the nonlinear regression analysis problem is of help based on the minimization of the sum of least squares of residuals in the additive form of the model. At first, all data are considered as uncensored and the least squares estimate is obtained by iterations necessary because of the nonlinear form of regression function y. To take into account the censored data, the EM algorithm can be applied. This iterative procedure alters two steps. The E step reconstructs the censored datum (i.e. for $I_i = 0$) in the sense that $\hat{Y}_i = E\left\{Y \,|\, Y \geq Y_i;\, Z_i, \hat{\Theta}\right\}$, where the value $\hat{\Theta}$ has been estimated in preceeding M step. The following M step computes new estimate of Θ from just reconstructed data, again by least squares method. The repetion of E and M steps leads to convergence (fixed point) of the iteration as shown rather generally in [3].

Point estimate of empirical d.f. of residuals

The empirical d.f. is a primary characteristic of a data sample having the form of an irregular staircase. For our case of an $N-$couple $\mathcal{R}$ of ordered residuals $R_{[i]}$ (where $R_{[i]} \leq R_{[i+1]},\ i = 1, .., N-1$) it corresponds to the step function F_S having a jump of height of $1/N$ at each $R_{[j]},\ j = 1, .., N$. This d.f. can be used in Kolmogorov–Smirnov's goodness–of–fit test: the hypothesis that the population from which the data $\mathcal{R}$ were sampled has a continuous d.f. F_t is rejected on the significance level α if

$$K_{KS} := \max_i(|F_t(R_{[i]}) - \frac{i-1}{N}|, |F_t(R_{[i]}) - \frac{i}{N}|) > \kappa_\alpha, \quad (i = 1, .., N) \tag{1}$$

holds, where κ_α is the critical value of the Kolmogorov–Smirnov statistics for the significance α. In principle, it would be possible to estimate some unknown parameters ϕ using the relation

$$\phi = argmin\,\{K_{KS}(\phi)\} \tag{2}$$

to get the best fit of data. However, this is a non–smooth problem involving difficulties especially when ϕ is not a scalar. This is why a simple point estimate of empirical d.f. (PEDF) was choosen having the form

$$F_p(R_{[i]}) := \frac{i-0.5}{N}, \quad (i = 1, .., N) \tag{3}$$

for uncensored data. To support this choice one should notice that $|F_t(R_{[i]}) - F_p(R_{[i]})|$ differs from the K_{KS} determined by (1) only by the constant $\frac{1}{2N}$. (There is no place here to discuss other motivations for this choice.) Using a procedure similar to the EM algorithm, censored data were then used to get the modification F_p' representing the whole set R of residuals.

Gnostical estimates

The gnostical theory of uncertain data (Kovanic 1986) is an alternative to statistics developed for practical applications which do not have a statistical model or for which such a model would be an inadequate instrument. Data may be strongly disturbed, observed objects far from steady-state and there may be insufficient data to develop a statistical model. Gnostical programs have an inherent robustness with respect to outlying data. The range of possible applications includes "heavy-duty" technological systems as well as economical and other analyses.

Consider the i-th real-valued observation A_i (an 'additive' datum) together with its 'multiplicative' equivalent

$$Z_i = exp(A_i) \tag{4}$$

having a strictly positive value. For a positive real scale parameter s, a real variable $z > 0$ and for a sample of N data, define the auxiliary quantities

$$q_i(z,s) = (Z_i/z)^{2/s} \tag{5}$$

for use in the calculation of N 'fidelities'

$$f_i(z,s) = 2/(1/q_i(z,s) + q_i(z,s)) \tag{6}$$

and 'irrelevancies'

$$h_i(z,s) = (1/q_i(z,s) - q_i(z,s))/(1/q_i(z,s) + q_i(z,s)). \tag{7}$$

Within the frame-work of the gnostical theory, irrelevance plays the role of the distance between z and Z_i, the fidelity being the weight of the datum Z_i.
Introduce the arithmetical mean

$$\bar{f}(z,s) = \sum_{i=1}^{N} f_i(z,s)/N \tag{8}$$

of the fidelities and define the symbol $\bar{h}(z,s)$ for the irrelevances analogously. Let $w(z,s)$ be the weighting function defined by the relation

$$w(z,s) = \sqrt{(\bar{f}(z,s))^2 + (\bar{h}(z,s))^2}. \tag{9}$$

The distribution function generated by the individual datum Z_i is then

$$L_i(z,s) = (1 + h_i(z,s))/2 \tag{10}$$

having the density

$$l_i(z,s) = \frac{d(L_i(z,s))}{dz} = f_i^2(z,s)/(zs). \tag{11}$$

There are two kinds of distribution functions (d.f.) supported by the gnostical theory: 'local' and 'global'. The local d.f. $L(z,s)$ is simply the arithmetical mean of the d.f.'s (10) of individual data:

$$L(z,s) = \sum_{i=1}^{N} L_i(z,s)/N. \tag{12}$$

The global d.f. $G(z,s)$ is obtained using the weighting function w (9),

$$G(z,s) = \sum_{i=1}^{N} L_i(z,s)/w(z,s). \tag{13}$$

Under the condition of a weak influence of uncertainty (small errors of data, a small value of the scale parameter s), the two functions differ only slightly. However, their behaviour is quite different under gross errors.

In the sense of being a monotonous function for an arbitrary data sample, the local d.f. always exists. In a special case of admissibility of the statistical view on data, the d.f. $L(z,s)$ (12) can be interpreted as a nonparametric estimate of the probability d.f. and the function $l_i(z,s)$ (11) as a proper kernel of Parzen's type (Parzen 1962). In this case, the gnostical theory is used only as a background motivating the choice of the special kernel (11). For this purpose, this approach generates remarkably nice and smooth density curves even in the case of small data samples. If it is of interest, the asymptotic features of the estimate can then be studied by established statistical methods. But in the more general case of data not having a statistical model, formula (11) still has significance as a continuous model of the data sample's distribution and as an estimate of the expectation of an another datum of the same origin. The steep descent of the gnostical kernel (11) has an important consequence: the individual subclusters of data influence each other only weakly. This enables the characterization of local details of the data sample and opens an efficient method of cluster analysis.

Unlike the local d.f., the global function $G(z,s)$ has a theoretical support only for homogeneous data samples, i.e. for data with a unimodal density function. When applied to the multimodal case, this function may lose the fundamental feature of a d.f., its monotony. (For an example, see [1]). This allows the hypothesis on the homogeneity of the data sample to be tested. The global d.f. has no known statistical analogy. Its practical importance is connected with its remarkable robustness with respect both to outlying data and outlying subclusters of data: in estimation of probability for extremal quantiles, the 'central', 'main' part of the data sample plays the dominating role. The global d.f. thus characterizes the overall distribution law of the data. Practical consequences of this include the d.f.'s good performance in application to small samples and its applicability to samples generated by different distribution laws (documented in [1]).

The local and global d.f.s differ substantially in their dependence on the scale parameter (s). Let F_p be the point estimate PEDF of the d.f. of the data sample. The local d.f. $L(z,s)$ of the same sample can be made to approach F_p as close as required choosing a sufficiently small positive value for the scale parameter. By contrast, the maximum distance of the global d.f. $G(z,s)$ has a minimum for a 'best' $\tilde{s}$, which can be taken as a robust estimate of the scale parameter. In this case, the d.f. $G(z,\tilde{s})$ is as close as possible to the empirical distribution function F_p, it represents its best fit.

Both gnostical estimate of d.f.'s are applied to survival modelling: Using the global d.f. G, the best estimate of the parameter Θ is determined. The resulting function $G(z,s)$ is then tested for monotonicity. Positive result – end of procedure. Negative result – further processing using the local d.f. $L(z,s)$ to analyse inhomogeneities (subclusters) of the data causing multimodality of distribution of residuals.

Finiteness of the data support

Many 'popular' theoretical distribution functions are defined over an infinite interval. Within the frame–work of the gnostical theory, the variable z can also acquire an arbitrary value from the open interval $(0, \infty)$. In contrast to this, real data are always finite. Moreover, the bounds of the data support may sometimes represent the most valuable information on observed process. For example, this is the case of reliability where the extremely interesting point is the largest value of the life–time (or of a covariate) beneath that the probability of defect is still zero. Naturally, such a point is not known a priori and its estimate is a desirable result of data treatment.

To include the finiteness of the interval $\mathcal{D} = (\mathcal{Z}_{\mathcal{L}}, \mathcal{Z}_{\mathcal{U}})$ into the estimating process and to enable the estimation of the bounds, one introduces the transformation $\tau := \mathcal{D} \rightarrow \mathcal{R}_+$ (where the R_+ is the interval of strictly positive reals) between data and gnostical formulae. For data $\zeta \in \mathcal{D}$ and $z \in R_+$ the applied transformation had the form

$$z = \frac{\zeta - Z_L}{1 - \zeta/Z_U}. \tag{14}$$

Optimization procedure

To evaluate the quality of fitting the PEDF of model residuals by the gnostical d.f. of global type, the quadratic penalty function Q was introduced:

$$Q = \sum_{i=1}^{N} (F_p(R_{[i]}) - G(R_{[i]}, s(R_{[i]})))^2. \tag{15}$$

Each residual is a function of the unknown parameters; the penalty Q is therefore dependent on the vector Θ. Values of parameters as well as values of their functions may be in a general case subjected to some given constraints having the form of both equations and inequalities. The task can be thus formulated as a nonlinear constrained minimization problem

$$\check{\Theta} = \arg\min(Q) \tag{16}$$

solution of which ensures finding such a surviving model t, such a data transformation τ and such a global distribution function G that the mean–square fitting error reaches its minimum whereby the constraints are respected. Several optimization methods have been tested to solve solve this problem using the user–friendly optimization program OPTIA [4]. As a most suitable, the fast and reliable Schittkowski's algorithm [6] was applied yielding a local minimum. Repeating runs with different initial conditions were used to verify the global optimality.

Example

We give an example using real data (shown in Tab.1) taken from life–time tests of truck components.

Item	Amplitude	Survival	Censoring
1	30	64300	1
2	30	29130	1
3	30	105220	1
4	30	108672	1
5	25	209675	1
6	20	293255	1
7	20	324887	1
8	20	309062	1
9	20	353372	1
10	15	625956	1
11	15	1041291	1
12	15	2000000	0
13	15	5000000	0

Table 1. Data

Each component was subjected to periodical load with a chosen amplitude z. The model of the life–time was $t(z,\theta) = B\cdot(z+C)^{-A}+D$ where t is the survival (number of cycles till the breakdown of the component or till the end of the experiment – in the case of censored datum). Parameters A, B as well as the values of $z+C$ are supposed to be positive. The first stage estimation of parameters A, B, C (supposing $D=0$) has been done in $\log_{10}$ form of the model, which is near to linear, namely $y(z,\theta) = B_1 - A\cdot\log_{10}(z+C)$, where $B_1 = \log_{10} B$. After some EM iterations the values $\hat{A} = 2.146$, $\hat{B}_1 = 7.622$, $\hat{C} = -10.575$ have been obtained. Then the value of parameter D has been found maximizing product of "fidelities" of the data fit. It yielded $\hat{D} = 1881$. The estimate of regression curve (after this preliminar estimation) is displayed in Figure 1.

The multiplicative residuals from this approximation of this regression curve have been then computed to iterate the parameters S_0, S_1, R_L, R_U by means of the optimization procedure minimizing the distance between the gnostical and empirical distribution functions of the residuals. Resulting estimates of parameters were

$$\hat{S}_0 = 1.415, \quad \hat{S}_1 = -0.0155, \quad \hat{R}_L = 0.194, \quad \hat{R}_U = 0.708,$$
$$\hat{A} = 1.815, \quad \hat{B}_1 = 7.291, \quad \hat{C} = -10.493, \quad \hat{D} = 877.$$

The resulting model function $t(z,\hat{A},\hat{B},\hat{C},\hat{D})$ now describes the dependence of median of multiplicative residuals' distribution on z. The final fit of the PEDF by the gnostical distribution function of multiplicative residuals is characterized by Table 2.

Ordered residuals	GDF	PEDF	Abs. difference	Item
1	0.01034	0.03846	0.02812	2
2	0.07132	0.11538	0.04407	10
3	0.17218	0.19231	0.02013	1
4	0.31645	0.26923	0.04722	11
5	0.38294	0.34615	0.03679	6
6	0.43582	0.42308	0.01274	8
7	0.48812	0.50000	0.01188	7
8	0.57743	0.57692	0.00051	9
9	0.68513	0.65385	0.03128	3
10	0.71916	0.73077	0.01161	4
11	0.81513	0.80769	0.00744	5
12	0.87233	0.88462	0.01229	12
13	0.99693	0.95857	0.03836	13

Table 2. Comparison of global gnostical distribution function (GDF) with the point emperical distribution function (PEDF).

Figure 2 shows the curve of median and selected quantiles of (global) gnostical distribution of residuals. The curves denoted by R_L and R_U correspond to bounds of the data support, i.e. to the bounds of region where $R_L < R < R_U$ holds. Figures 3 and 4 display graphically the form of gnostical density and distribution function (the global and the local versions) of the multiplicative residuals, under the load amplitude $z = 25$. It is seen how the local density function is able to reveal the outlying results and how the global one suppresses their influence.

Conclusions

The method presented based on the new nonstatistical (gnostical) theoretical background has been proved to be useful in the identification of reliability models under hard practical conditions such as small sets of strongly dispersed data. Only the structure of the formula characterizing the dependence of the life–time on the covariates has to be supposed. The parameters of this model together with a nonparametric estimate of the distribution function of residuals and even the bounds of data support result from the procedure. The method exhibited its remarkable robustness and efficiency in practical applications to reliability of vehicles.

References

[1] R. H. Baran, "Comments on 'A New Theoretical and Algorithmical Basis for Estimation, Identification and Control' by P. Kovanic", *Automatica*, **24**, 283-287 (1988).

[2] D. M. Bates, D. G. Watts, *Nonlinear Regression Analysis and Its Applications*, Willey, N. Y., 1988.

[3] A. P. Dempster, N. M. Laird, D. B. Rubin, (1977), "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm", *J. Roy Statist. Soc.*, ser. **B 39**, 1–22 (1977).

[4] J. Fidler, J. Doležal, J. Pacovský, "Dialogue System OPTIA for Minimization of Functions of Several Variables – User's Guide", Inst. of Inf. Theory and Automation, Prague, 1991.

[5] P. Kovanic, "A New Theoretical and Algorithmical Basis for Estimation, Identification and Control", *Automatica*, **22**, 657-674 (1986).

[6] K. Schittkowski, "NLPQL: A Fortran Subroutine Solving Constrained Nonlinear Programming Problems", *Annals of Operation Research*, **5**, 485-500 (1985/6).

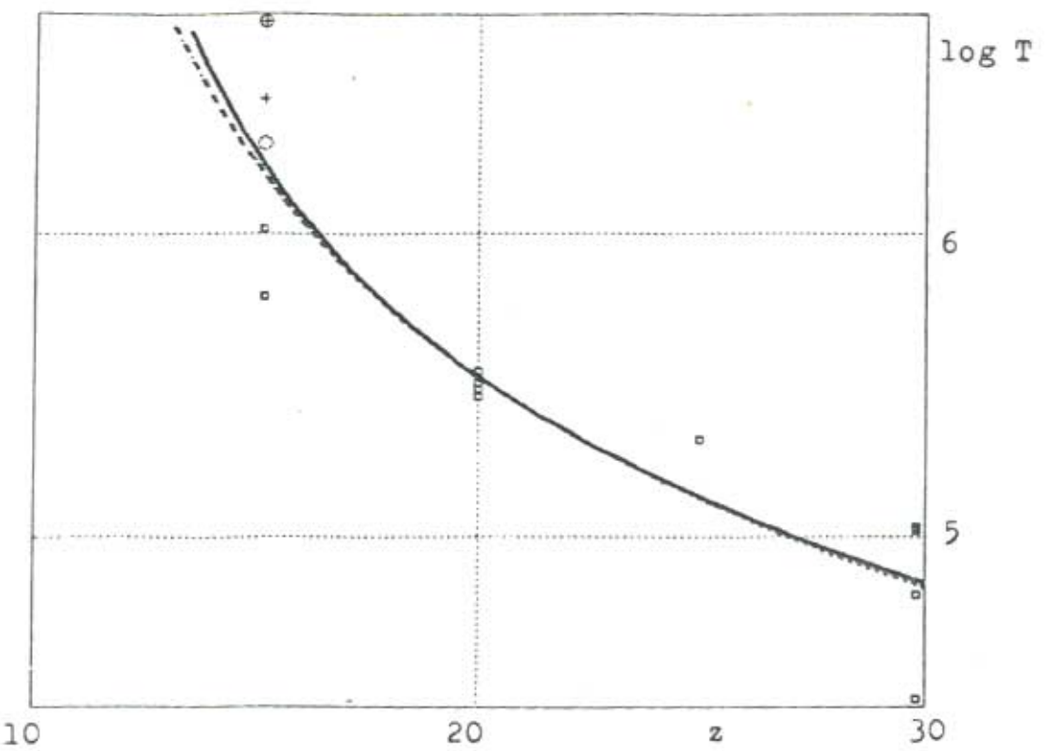
[7] W. Stahel, S. Weisberg, "Directions in Robust Statistics and Diagnostics", *The IMA Volumes in Mathematics and its Applications*, **33** and **34**, Springer Verlag, 1991.

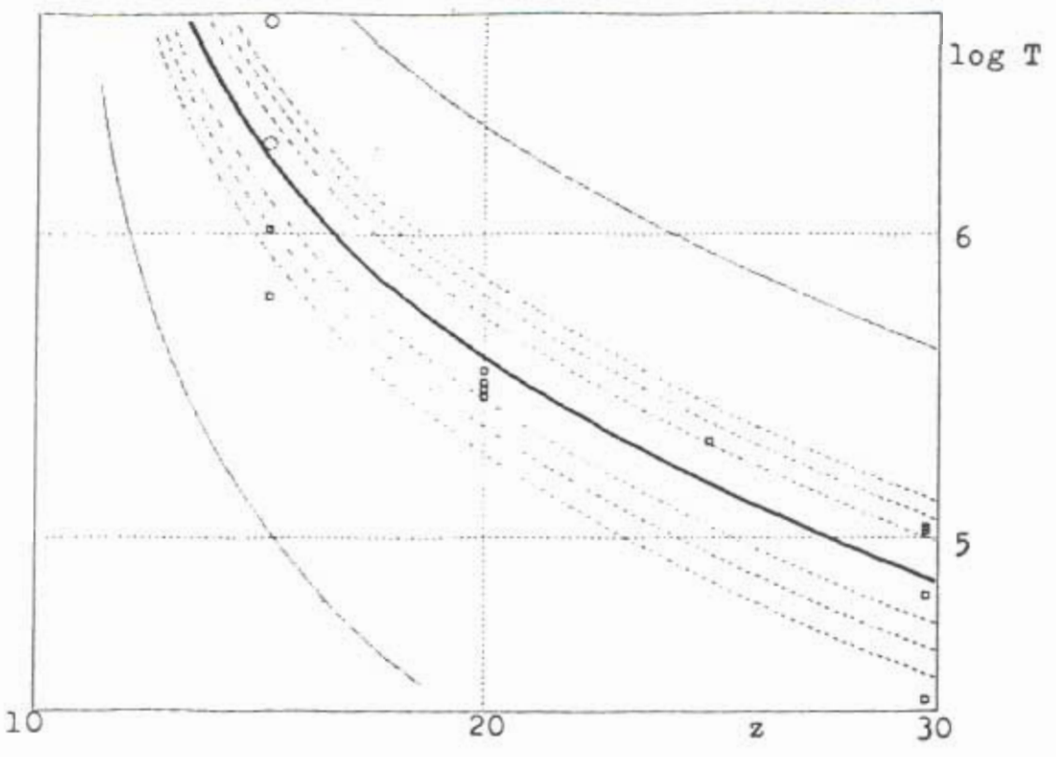
Fig. 1. First estimates of regression curve: dashed line – initial loglinear approximation, full line – least squares approximation, ∘ – censored data, + – reconstructed censored data.

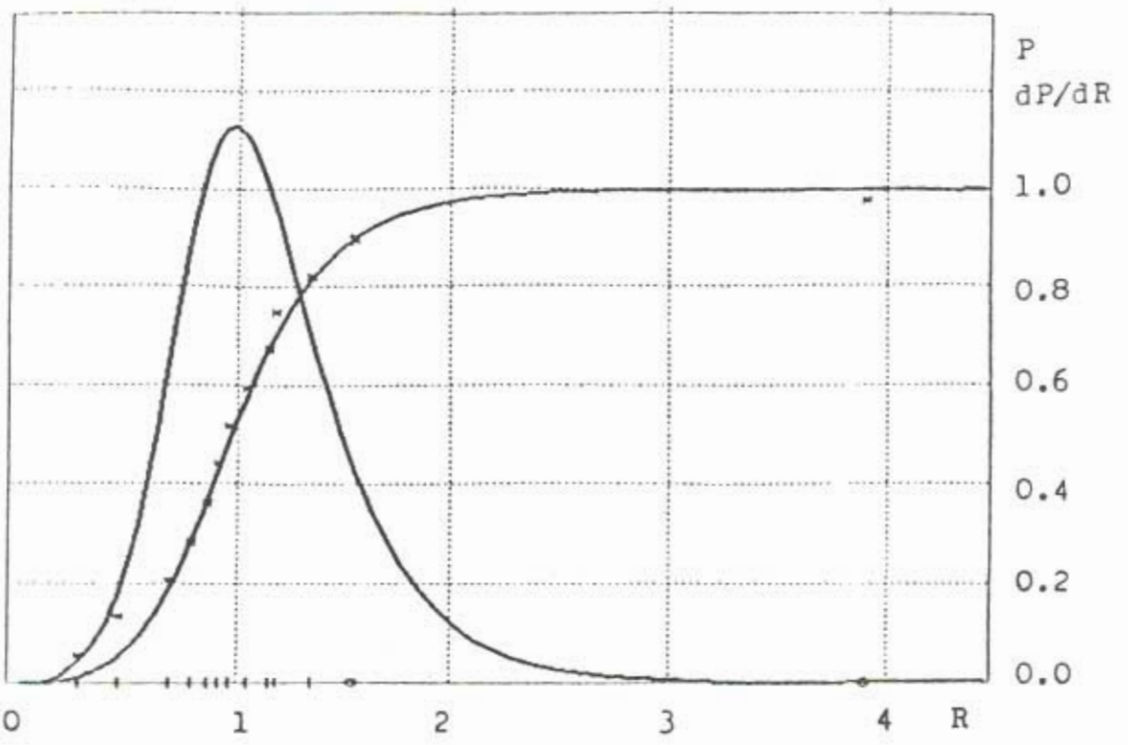
Fig. 2. Resulting curve of medians, of lower and upper 5 %, 10 %, 20 % quantiles and boundary R_L and R_U curves.

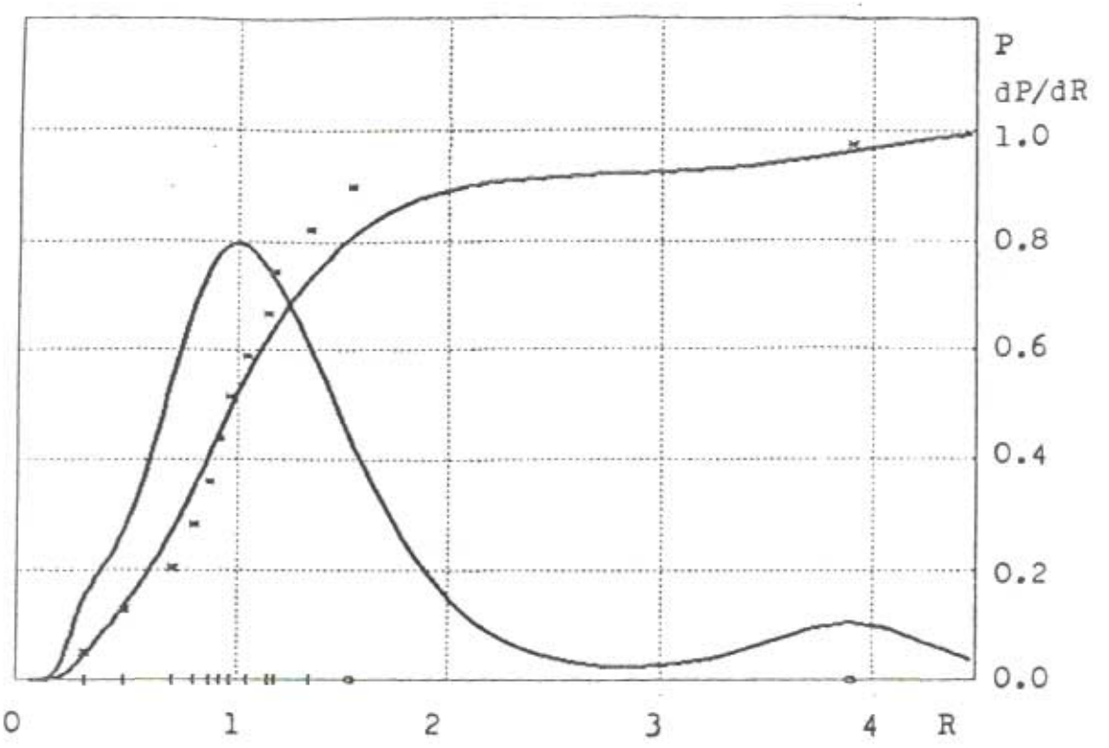
Fig. 3. Global distribution function and density, + – noncensored residuals, ∘ – censored residuals, × – values of PEDF.

Fig. 4. Local distribution function and density, + – noncensored residuals, ∘ – censored residuals, × – values of PEDF.


log T
6
5
10
20
z
30


log T
6
5
10
20
z
30


P
dP/dR
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
R


P
dP/dR
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
R

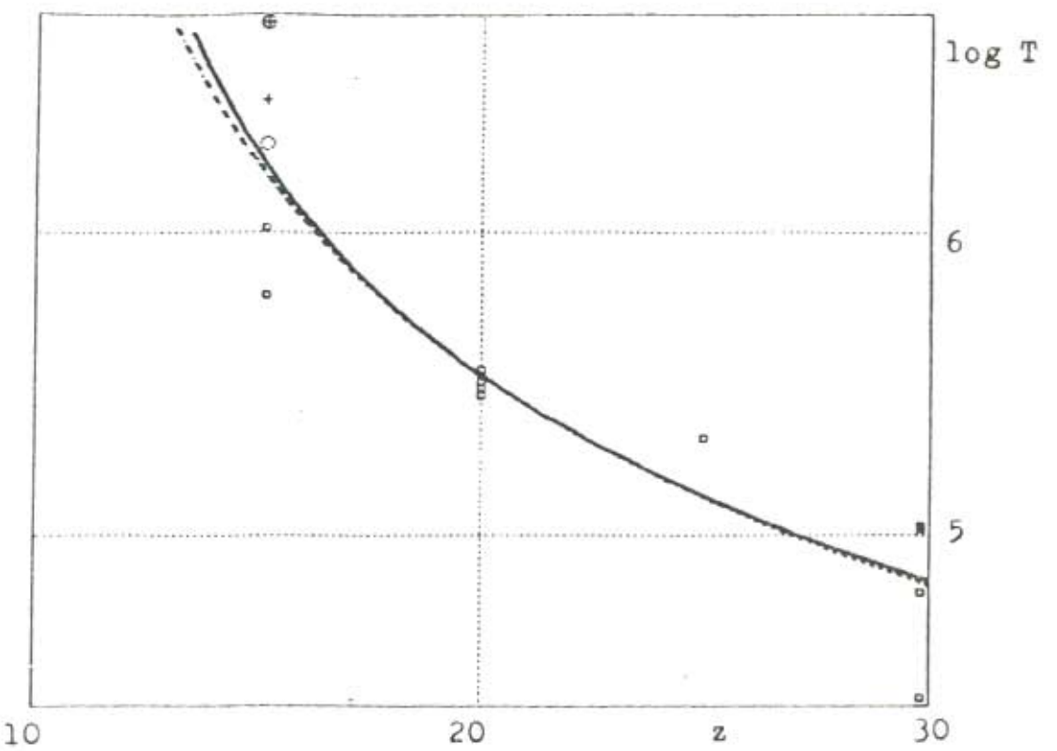


Fig. 1

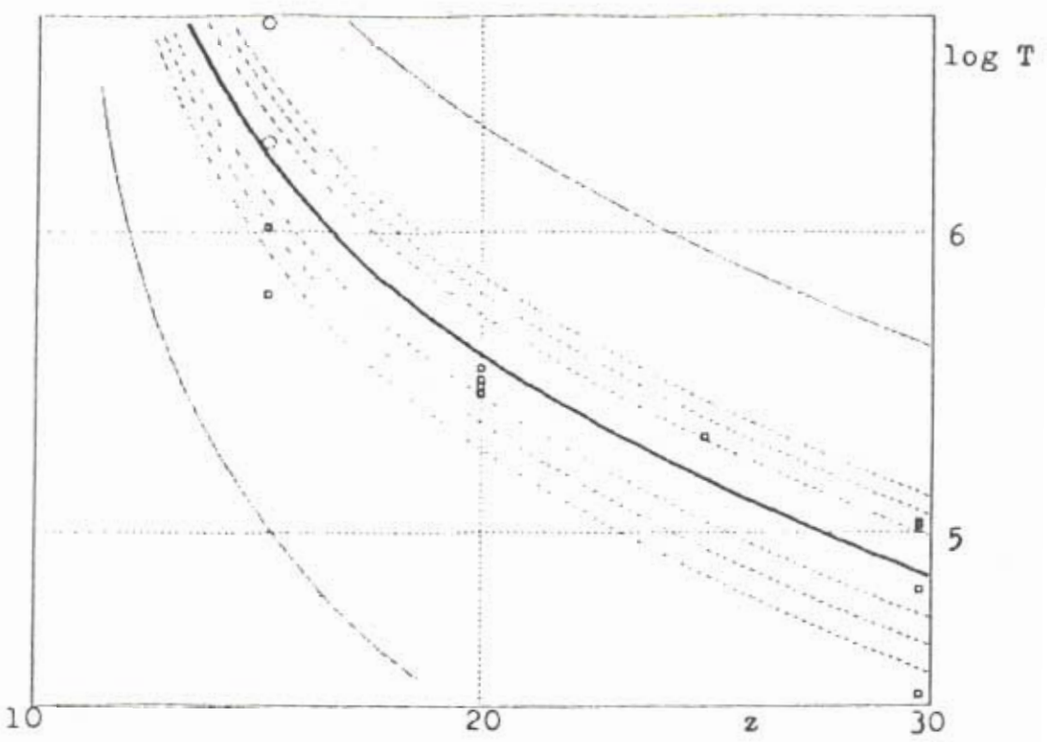


Fig. 2

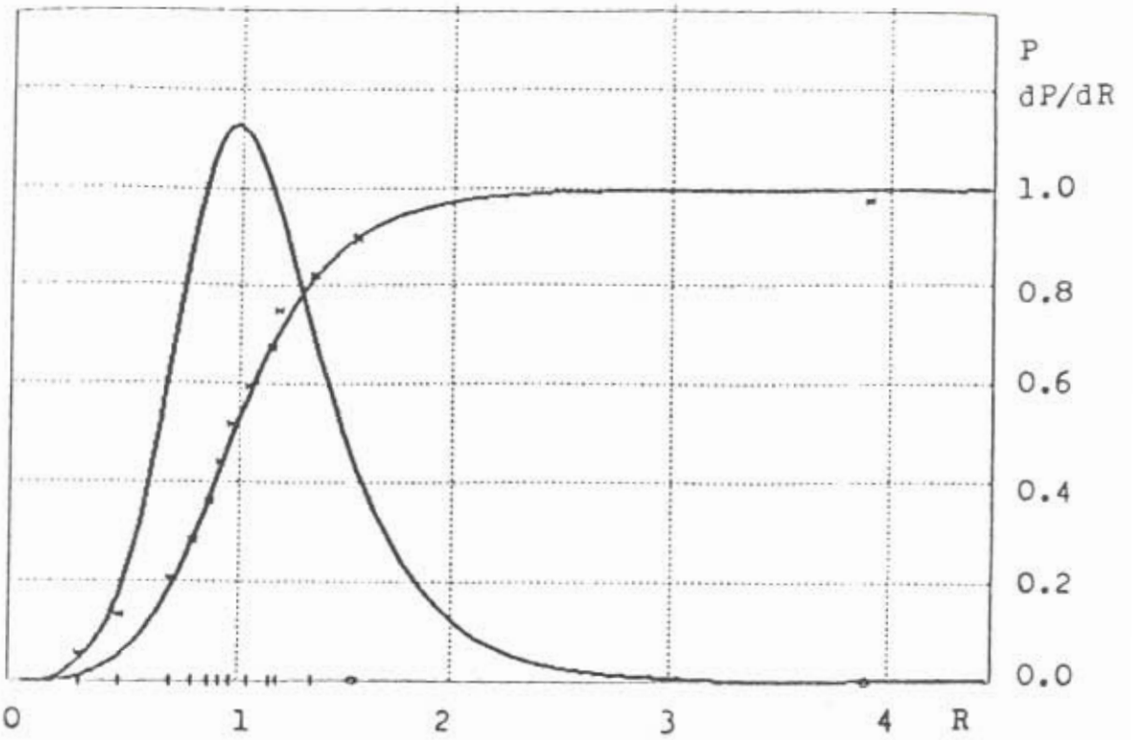


Fig. 3

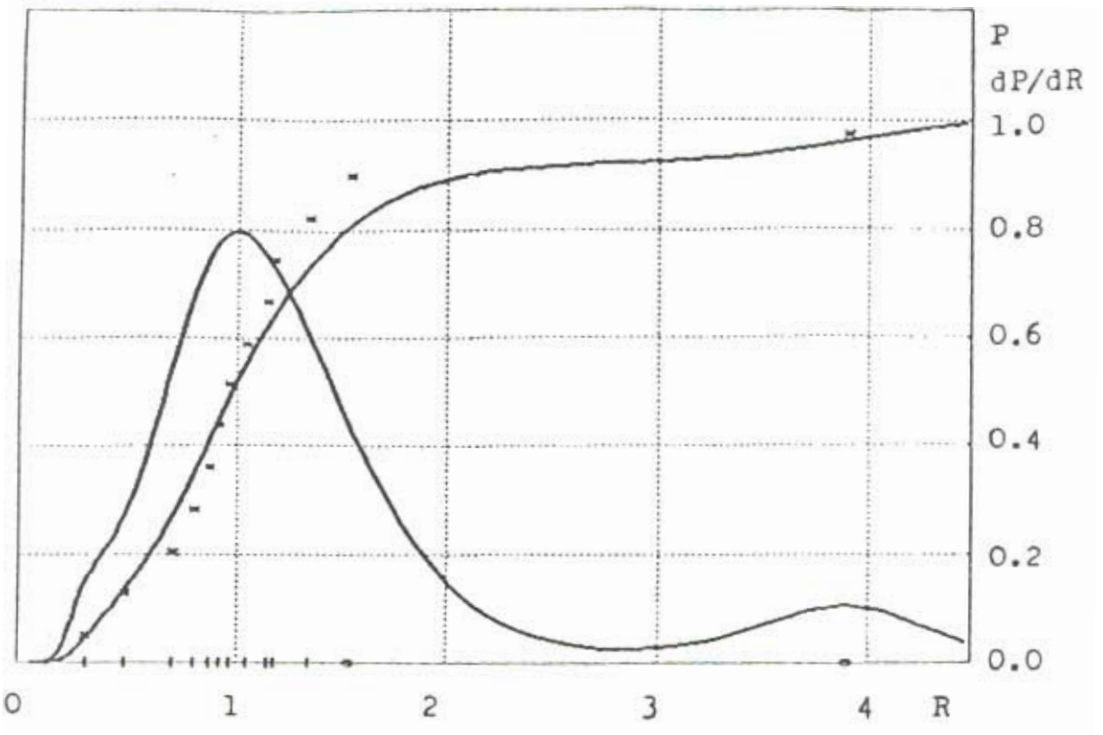


Fig. 4