Applied Mathematical Modelling 149 (2026) 116284

Contents lists available at ScienceDirect

Applied Mathematical Modelling

journal homepage: www.elsevier.com/locate/apm

A flexible model of ordered random variables for non-metallic inclusions in steels and related statistical inference

Stefan Bedbur [a], Udo Kamps [a], Anja Bettina Schmiedt [b,*]

[a] RWTH Aachen University, Institute of Statistics, Aachen, 52056, Germany
[b] OTH Regensburg University of Applied Sciences, Department of Computer Science and Mathematics, Regensburg, 93053, Germany

ARTICLE INFO

Keywords:
Generalized order statistics
Engineering steel
Weibull distribution
Link function

ABSTRACT

In a data set of non-metallic inclusion sizes in samples from engineering steel, common order statistics fail to serve as a suitable model for ascendingly ordered measurements within single samples. Therefore, a flexible model of ordered random variables is proposed, which allows for changes of distributions described by model parameters. Joint maximum likelihood estimation of these parameters and the shape parameter of an underlying left-truncated Weibull distribution is considered, and a model test is developed for the null-hypothesis of common order statistics being an adequate model. To overcome small data situations, a link-function approach is examined in order to reduce the number of involved model parameters as well as to propose to use a link-function parameter as a material indicator. An asymptotic test is provided to check for the presence of a linear link function, and tests for hypotheses about two link-function parameters are studied. Moreover, the construction of simultaneous confidence regions for the link-function parameters as well as of confidence bands for the entire graph of the link function are presented. Throughout, the findings are applied to the real metallurgical data set. Similar problems and data structures arise in other fields of material science and applications such as geology.

1. Introduction

In the manufacturing process of steel, non-metallic inclusions, primarily sulfides and oxides, arise unavoidably. While very small inclusions, on the order of a few nanometers, are generally considered irrelevant or of minor interest concerning the material's properties, larger inclusions, on the micrometer scale, can lead to significant issues such as fatigue, corrosion, and particularly crack initiation. Larger inclusions therefore serve as quality indicator of a steel under test. Inclusion size is hence assumed to be the most relevant geometrical parameter; see Murakami [1] and the references therein.

Throughout, we consider a data set of non-metallic oxide inclusion sizes, which was provided by the Department of Ferrous Metallurgy at RWTH Aachen University. The respective engineering steel and the data are described in detail by Schmiedt et al. [2,3]. On the surfaces of polished planes, in total 60 equal-sized control areas were considered to determine the sizes of non-metallic inclusions above some threshold following ASTM-standard E2283-08 [4]; see also Murakami [1] and Murakami and Beretta [5]. That is, every control area was scanned by optical microscopy to detect those inclusions that intersect the surface. The sizes of their two-dimensional cross-sections were measured and stored, here, in terms of the square root of the projected area (the so-called $\sqrt{\text{area}}$-

* Corresponding author.
E-mail addresses: stefan.bedbur@rwth-aachen.de (S. Bedbur), udo.kamps@rwth-aachen.de (U. Kamps), anja.schmiedt@oth-regensburg.de (A.B. Schmiedt).

https://doi.org/10.1016/j.apm.2025.116284
Received 5 May 2025; Received in revised form 3 June 2025; Accepted 30 June 2025

Available online 16 July 2025

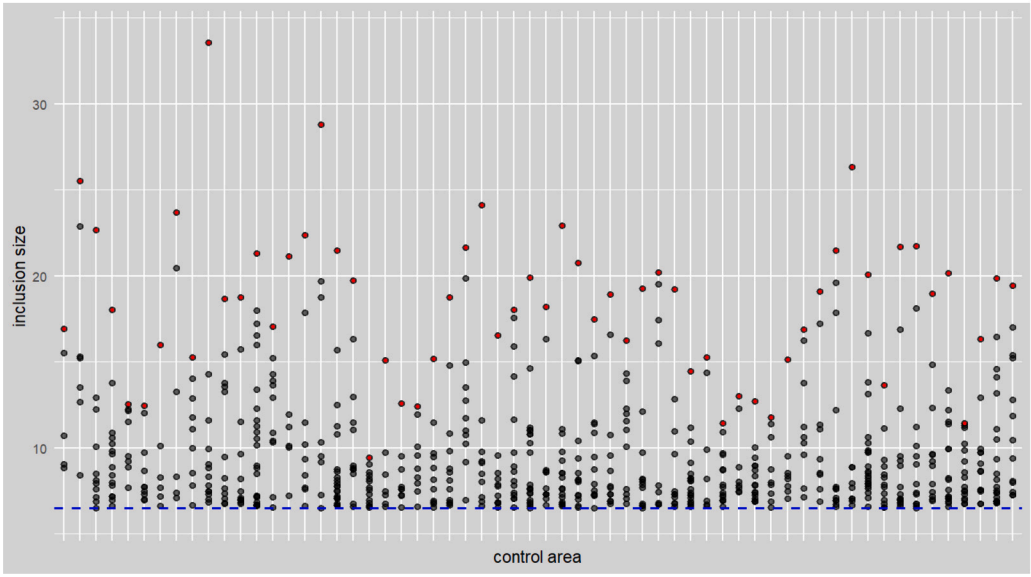

Fig. 1. Illustration of the inclusion sizes (in μm, measured in terms of the $\sqrt{}$area-parameter) exceeding the threshold $\tau = 6.5$ (dashed blue line) in each of the 60 control areas, with the control area maxima marked as red dots.

parameter); see the technical recommendation ESIS P11-02 by Anderson et al. [6]. Further details on the experimental materials and measurement methods are stated in Schmiedt et al. [3].

According to the ASTM-standard E2283-08 [4], only the maximum observed inclusions in the control areas are used in statistical studies by means of extreme value analysis. Given the importance of large observed inclusions, we consider all inclusion sizes from every control area that exceed a certain threshold in our analysis. Our objective is to describe and study their arrangement in ascending order, as well as their progression, to derive a potential quality measure for steels. Thus, we are concerned with fitting an appropriate statistical model of ordered random variables to ascendingly ordered inclusion sizes with a special interest on the large ones.

Similar problems and data structures arise in other fields of material science and applications such as geology.

For illustration of the data considered here, the recorded oxide inclusion sizes in micrometer (μm) above a threshold of $\tau = 6.5$ μm are shown in Fig. 1, sorted in ascending order within each control area. The respective largest inclusion sizes in each control area, i.e., the control area maxima, are marked as red dots. In total, there are measurements of 719 inclusions above threshold τ assigned to 60 control areas.

In statistical models in metallography, so far, the assumption is commonly made that increasing inclusion sizes in each control area can be suitably modelled as realizations of common order statistics (OSs) based on independent and identically distributed (i.i.d.) random variables. However, in the data under consideration, large inclusions, which are of primary interest, do not seem to be well described by OSs. Hence, a more flexible model of ordered random variables should be fitted to appropriately describe ordered non-metallic inclusion sizes. It will turn out that, within our setup, a statistical test will reject the hypothesis of the presence of common OSs based on the aforementioned data.

We propose to apply a model of ordered random variables out of the class of generalized order statistics. In this model termed 'ordering via truncation of distributions', which has not been applied to real data sets so far, the ordered quantities are based on different distributions in contrast to the model of OSs; see Kamps [7,8,9]. Details are given in the next section. In a particular parametrical setup, successive underlying distributions may be characterized by parametrically adjusted hazard rates when starting with a baseline distribution. We develop methods of statistical inference for these parameters. For the baseline distribution, we consider a left-truncated Weibull distribution. Weibull distributions are widely used in the literature for modelling particle size and inclusion size distributions; see, e.g., Anderson et al. [6], Cao et al. [10], Fang et al. [11], Ghosh [12], Hilliard and Lawson [13], Murakami and Beretta [5], Sczerzenie et al. [14]. Here, the suitability of a left-truncated Weibull distribution as a baseline for the data under consideration throughout this manuscript is discussed in Example 2.1. For references on left-truncated Weibull distribution with a known threshold, see Wingo [15], Kreer et al. [16], and Kizilersu et al. [17].

The outline of this manuscript is as follows. In Section 2, a flexible model of ordered random variables within the class of generalized order statistics is proposed for describing ordered non-metallic inclusion sizes in steels. Moreover, the assumption of a left-truncated Weibull baseline distribution is justified. Joint maximum likelihood (ML) estimation of the model parameters and a Weibull shape parameter is then considered in Section 3, which enables to fit the model to data. A model test is provided in Section 4, which is applied to give evidence (subject to the level) against common OSs as an adequate model for our data set of ordered non-metallic inclusion sizes. The flexibility of the proposed model is related to a large number of parameters, which, regarding precision of parameter estimation, may not be desirable in applications with fairly small samples. Therefore, we examine a link-function approach to reduce the number of involved parameters as well as to propose to use the link-function parameters as material indicator. An asymptotic test is provided in Section 5 to check for the presence of a linear link function, which turns out to be suitable

for our data set. Statistical tests for hypotheses about two link-function parameters are then studied in Section 6, where optimality properties are also established. Finally, the construction of simultaneous confidence regions for the link-function parameters as well as of confidence bands for the entire graph of the link function are presented in Section 7. Throughout, the findings are applied to our data set of oxide inclusion sizes and the results are discussed.

2. A general model for ordered inclusion sizes

Let us first consider a single sample of ascendingly ordered random variables $X_1 \le \cdots \le X_r$ representing increasing inclusion sizes within some control area. Moreover, let F denote an absolutely continuous distribution function, referred to as baseline distribution function, with density function f, and let $\gamma_1, \ldots, \gamma_r$ be positive parameters.

We apply a model for ordered random variables termed 'ordering via truncation of distributions' in a particular form within the class of generalized order statistics, which have various applications; see Kamps [7,8,9] and the references therein.

The joint density function of $X_1, \ldots, X_r$ is given by

$$f^{X_1,\ldots,X_r}(x_1,\ldots,x_r) = \left(\prod_{j=1}^{r}\gamma_j\right)\left(\prod_{j=1}^{r}(1-F(x_j))^{\gamma_{r-j+1}-\gamma_{r-j}-1}f(x_j)\right) \tag{1}$$

for $x_1 \le \cdots \le x_r$ and $\gamma_0 = 0$.

It should be noted that, compared to the literature on generalized order statistics, the numbering of parameters is reversed here, which is more convenient in the present modelling.

It turns out that the joint density function in formula (1) coincides with the one of the first r Pfeifer record values; cf., e.g., Kamps [7,9], Arnold and Villasenor [18], Schmidt and Kamps [19]. For the choice $\gamma_j = j$, $1 \le j \le r$, we obtain the joint density function of OSs based on F.

The ordered quantities $X_1, \ldots, X_r$ form a Markov chain with transition probabilities

$$P\big(X_j > t \mid X_{j-1} = s\big) = \left(\frac{1-F(t)}{1-F(s)}\right)^{\gamma_{r-j+1}}, \qquad s < t, \quad 2 \le j \le r.$$

Hence, the conditional distribution of X_j, which represents the j^{th} smallest inclusion size in the control area, given $X_{j-1} = x_{j-1}$ has hazard rate

$$\gamma_{r-j+1}\frac{f}{1-F} \quad \text{on} \quad (x_{j-1},\infty), \quad 1 \le j \le r, \quad x_0 = \tau.$$

Since the hazard rates of all (conditional) inclusion size distributions are proportional to that of the baseline distribution, the parameters $\gamma_1, \ldots, \gamma_r$ allow for adjustments. As larger inclusions occur with a lower incidence that smaller ones in our situation, it is reasonable to assume that the model parameters $\gamma_1, \ldots, \gamma_r$ are increasingly ordered.

One-dimensional marginal density functions of $X_2, \ldots, X_r$ can be stated in terms of Meijer G-functions. However, the minimum X_1 has the simple density function $\gamma_r(1-F(x))^{\gamma_r-1}f(x_r)$ (see formula (1)) and thus distribution function $1-(1-F(x))^{\gamma_r}$ on (τ,∞).

In the Weibull setting, we obtain

$$F^{X_1}(x) = 1 - e^{-\gamma_r(x^\beta-\tau^\beta)}, \quad x \in (\tau,\infty),$$

which is again a truncated Weibull distribution. This is no longer valid for $X_2, \ldots, X_r$.

As mentioned in the introduction, Weibull distributions are widely used in models of particle sizes. Nevertheless, the use of a Weibull distribution for our data set has to be justified by using all 60 data of minimum inclusions.

Example 2.1. Regarding a Weibull baseline assumption for the real data illustrated in Fig. 1, we fit a left-truncated Weibull distribution to the control area minima. Specifically, we consider a theoretical distribution function of the form $1-e^{-\gamma(x^\beta-\tau^\beta)}$, $x > \tau$. Given the threshold $\tau = 6.5$ and the observed minima, we estimate the unknown distribution parameters using the method of maximum likelihood. In Fig. 2, histogram plots of the control area minima are shown (using two different bin widths) along with the estimated Weibull density (blue line). Because of a reasonable fit we continue working with an underlying Weibull distribution.

The results on statistical inference will be based on s independent control areas with r_i inclusion sizes in control area i, $1 \le i \le s$. The corresponding data are denoted by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s$, where the vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ir_i})$ with $x_{i1} \le \cdots \le x_{ir_i}$ contains the ascendingly ordered recorded inclusion sizes within the i^{th} control area, i.e., the observation x_{ij} denotes the j^{th} smallest recorded inclusion size in control area number i for $1 \le j \le r_i$ and $1 \le i \le s$. The observation vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s$ are considered realizations of jointly independent random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_s$ defined by $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ir_i})$, $1 \le i \le s$, where the density function of $\boldsymbol{X}_i$ is supposed to be given by formula (1) with r_i taking the place of r. Hence, we assume that the baseline distribution function is the same for all control areas and that the model parameters do not depend on the control area number. As baseline distribution, we choose a truncated Weibull distribution with distribution function

$$F(x) = 1 - e^{-(x^\beta-\tau^\beta)}, \qquad x > \tau,$$

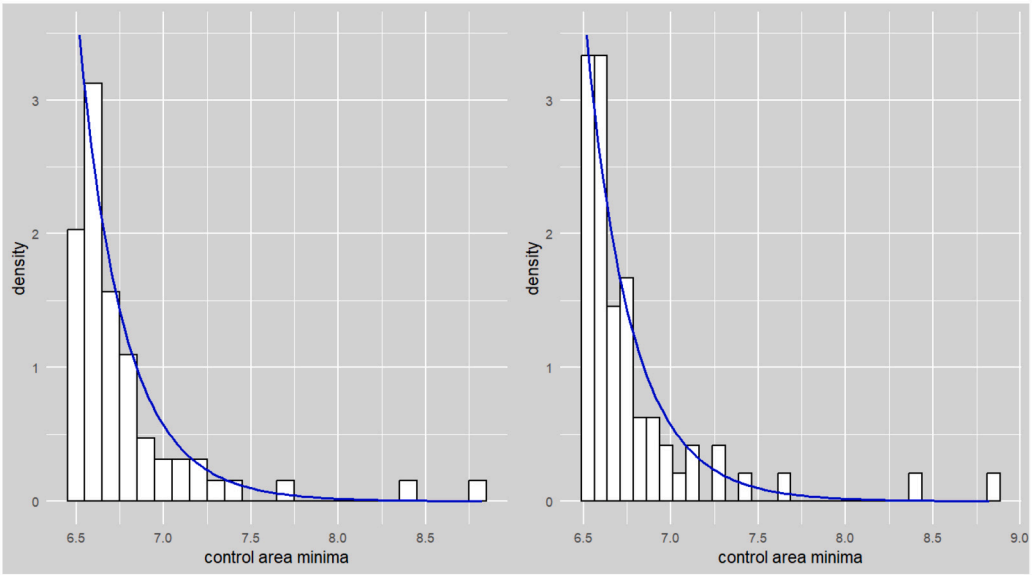

Fig. 2. Histogram plots of the minima of the 60 control areas (using different bin widths on the left and right) along with the density function of a fitted left-truncated Weibull distribution (blue line).

Table 1
Number c_υ of control areas with at least υ observations.

υ	1	2	3	4	5	6	7	8	9	10	11	12
c_υ	60	60	60	60	60	59	56	53	49	40	38	31
υ	13	14	15	16	17	18	19	20	21	22	23	
c_υ	22	16	14	10	7	6	5	5	4	3	1	

for some shape parameter $\beta > 0$ and a known threshold $\tau > 0$. W.l.o.g., let the numbers of recorded inclusions in the control areas be arranged in descending order, i.e., $r_1 \geq r_2 \geq \cdots \geq r_s \geq 1$, leading to r_1 model parameters $\gamma_1, \ldots, \gamma_{r_1}$ in total. Further, we denote by

$$c_\upsilon = |\{i \in \{1, \ldots, s\} : r_i \geq \upsilon\}| = \max\{i \in \{1, \ldots, s\} : r_i \geq \upsilon\}, \qquad 1 \leq \upsilon \leq r_1,$$

the number of control areas with at least υ observations.

Using these notations and upon some algebra, the density function of $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_s)$ can be represented as

$$f^{\mathbf{X}}(\mathbf{x}) = \exp\left\{\sum_{\upsilon=1}^{r_1} \gamma_\upsilon T_\upsilon(\mathbf{x}) + \sum_{\upsilon=1}^{r_1} c_\upsilon \ln(\gamma_\upsilon)\right\} \beta^c \prod_{\upsilon=1}^{r_1} \prod_{i=1}^{c_\upsilon} x_{i\upsilon}^{\beta-1} \tag{2}$$

with

$$T_\upsilon(\mathbf{x}) = \sum_{i=1}^{c_\upsilon} (x_{i,r_i-\upsilon}^{\beta} - x_{i,r_i-\upsilon+1}^{\beta}), \qquad 1 \leq \upsilon \leq r_1, \tag{3}$$

for $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s)$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ir_i})$ and $\tau = x_{i0} < x_{i1} \leq \cdots \leq x_{ir_i}$ for $1 \leq i \leq s$, and $c = \sum_{\upsilon=1}^{r_1} c_\upsilon$.

Example 2.2. Referring to the metallurgical data set displayed in Fig. 1, observations from a total of $s = 60$ control areas are available. Above the threshold $\tau = 6.5$ (µm), in total $c = 719$ oxide inclusion sizes were recorded. The respective number r_i, $1 \leq i \leq 60$, of inclusions per control area varies, ranging from $r_{60} = 5$ to $r_1 = 23$, leading to 23 model parameters $\gamma_1, \ldots, \gamma_{23}$. The estimation of γ_υ, $1 \leq \upsilon \leq 23$, is based on c_υ observations, where the latter denotes the number of control areas with at least υ recorded inclusion sizes. The numbers c_υ are listed in Table 1. It is observed that the estimation of $\gamma_1, \ldots, \gamma_5$ is based on 60 observations each, whereas fewer than 10 control areas provide information for estimating the model parameters $\gamma_{17}, \ldots, \gamma_{23}$. The estimation of γ_{23} is based on a single inclusion size, only.

In Fig. 3, the inclusion sizes $x_{i,r_i-\upsilon+1}$ for six observation vectors $\boldsymbol{x}_i$ are graphically represented. These are each sorted in descending order and plotted against their respective index $\upsilon - 1$. To illustrate that the structures can differ, the plot on the left shows the descending sizes of the three observation vectors with the largest, second-largest, and third-largest inclusion size in the entire data set, while the plot on the right shows those of the three observation vectors with the most, second-most, and third-most inclusions recorded above the threshold $\tau = 6.5$.

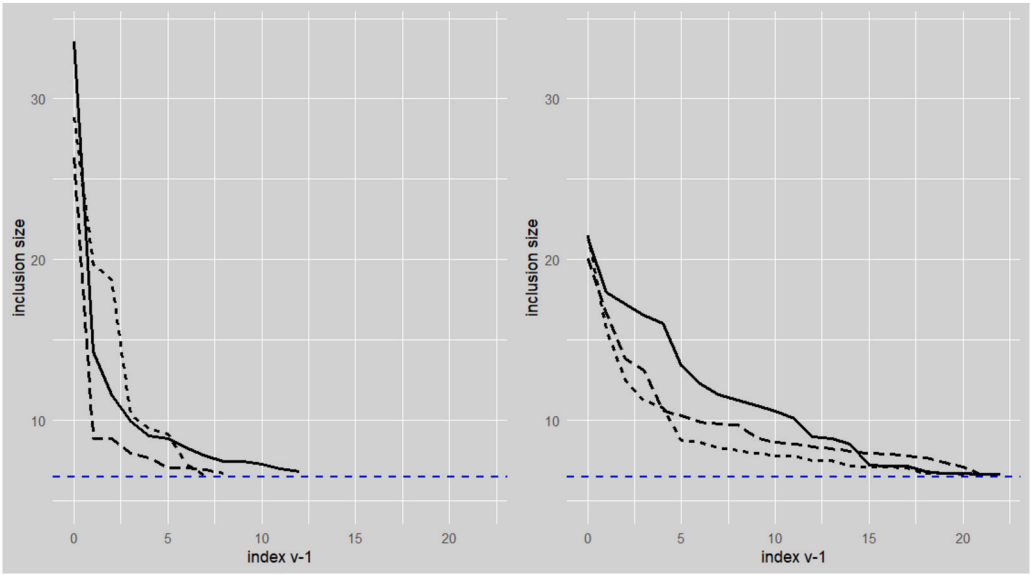

Fig. 3. Illustration of the inclusion sizes above $\tau = 6.5$, where x_{i,r_i-v+1} is plotted against the index $v-1$, for the control areas with the largest, second-largest, and third-largest inclusion size (left) and for the control areas with the largest, second-largest, and third-largest number of inclusions (right), respectively.

3. Joint ML estimation of model parameters and a shape parameter

As a starting point of our analysis, we consider joint ML estimation of $\beta, \gamma_1, \ldots, \gamma_{r_1}$ in model (2). To this end, we first state the corresponding log-likelihood function

$$\ell(\boldsymbol{\vartheta}) = \sum_{v=1}^{r_1} \gamma_v \sum_{i=1}^{c_v} (x_{i,r_i-v}^{\beta} - x_{i,r_i-v+1}^{\beta}) + \sum_{v=1}^{r_1} c_v \ln(\gamma_v) + c\ln(\beta) + (\beta-1)\sum_{v=1}^{r_1}\sum_{i=1}^{c_v} \ln(x_{iv}) \tag{4}$$

along with its first partial derivatives

$$\frac{\partial}{\partial\beta}\ell(\boldsymbol{\vartheta}) = \sum_{v=1}^{r_1} \gamma_v \sum_{i=1}^{c_v} (\ln(x_{i,r_i-v})x_{i,r_i-v}^{\beta} - \ln(x_{i,r_i-v+1})x_{i,r_i-v+1}^{\beta}) + \frac{c}{\beta} + \sum_{v=1}^{r_1}\sum_{i=1}^{c_v} \ln(x_{iv})$$

and

$$\frac{\partial}{\partial\gamma_v}\ell(\boldsymbol{\vartheta}) = \sum_{i=1}^{c_v} (x_{i,r_i-v}^{\beta} - x_{i,r_i-v+1}^{\beta}) + \frac{c_v}{\gamma_v}, \qquad v = 1, \ldots, r_1$$

for $\boldsymbol{\vartheta} = (\beta, \gamma_1, \ldots, \gamma_{r_1}) \in (0,\infty)^{r_1+1}$. It turns out that existence of the ML estimate for $\boldsymbol{\vartheta}$ is indeed an issue, while uniqueness can be guaranteed in case of existence. The results are summarized in the following theorem, the proof of which is technical and therefore moved to the appendix.

Theorem 3.1.

(i) *For $s = 1$, an ML estimate of $\boldsymbol{\vartheta}$ based on $\mathbf{x}$ does not exist.*
(ii) *For $s \geq 2$, an ML estimate of $\boldsymbol{\vartheta}$ based on $\mathbf{x}$ exists if and only if the equation*

$$\frac{c}{\beta} + \sum_{v=1}^{r_1}\sum_{i=1}^{c_v} \ln(x_{iv}) - \sum_{v=1}^{r_1} c_v \frac{\sum_{i=1}^{c_v}(x_{i,r_i-v+1}^{\beta}\ln(x_{i,r_i-v+1}) - x_{i,r_i-v}^{\beta}\ln(x_{i,r_i-v}))}{\sum_{i=1}^{c_v}(x_{i,r_i-v+1}^{\beta} - x_{i,r_i-v}^{\beta})} = 0 \tag{5}$$

has a solution $\hat{\beta}$ with respect to $\beta > 0$, where $x_{i0} = \tau$, $1 \leq i \leq s$. In case of existence, the ML estimate is uniquely determined and given by $\hat{\boldsymbol{\vartheta}} = (\hat{\beta}, \hat{\gamma}_1, \ldots, \hat{\gamma}_{r_1})$ with

$$\hat{\gamma}_v = c_v \left[\sum_{i=1}^{c_v}\left(x_{i,r_i-v+1}^{\hat{\beta}} - x_{i,r_i-v}^{\hat{\beta}}\right)\right]^{-1}, \qquad 1 \leq v \leq r_1. \tag{6}$$

In Lemma 3.2, a simple sufficient condition is provided for the existence of the ML estimate based on a given data set $\mathbf{x}$; its proof can be found in the appendix.

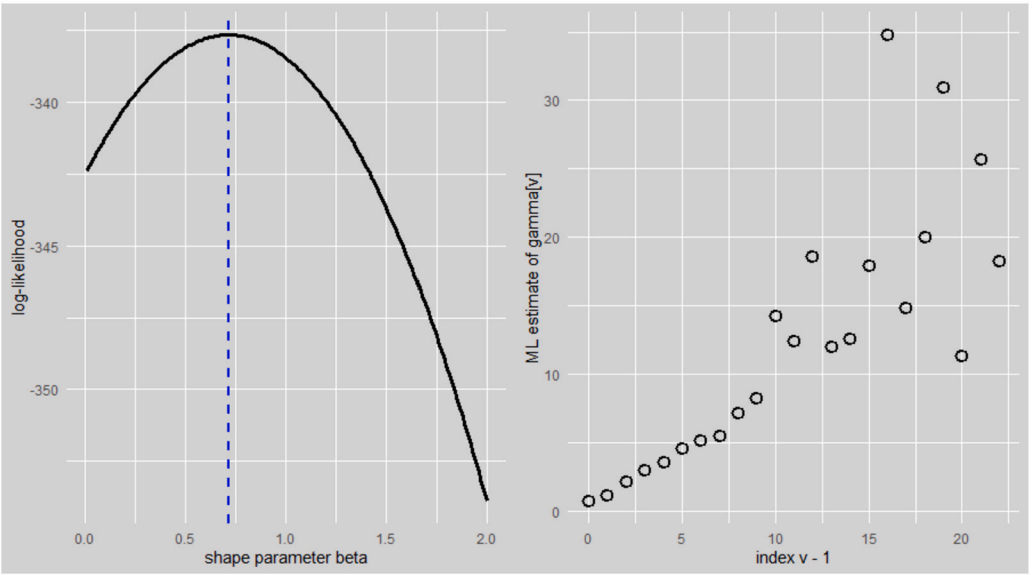

Fig. 4. Profile log-likelihood function for β along with the ML estimate $\hat{\beta}(\mathbf{x}) = 0.72$ in blue (left), and ML estimates $\hat{\gamma}_v(\mathbf{x})$ of γ_v plotted against $v - 1$ (right).

Lemma 3.2. *For $s \geq 2$, the ML estimate of $\boldsymbol{\vartheta}$ based on $\mathbf{x}$ exists if*

$$\min_{1\leq i\leq c_v} x_{i,r_i-v+1} < \max_{1\leq i\leq c_v} x_{i,r_i-v+1}$$

for some $v \in \{1, \dots, r_1\}$ and

$$\sum_{v=1}^{r_1}\sum_{i=1}^{c_v} \ln(x_{iv}) - \sum_{v=1}^{r_1} \frac{c_v \sum_{i=1}^{c_v}(\ln^2(x_{i,r_i-v+1}) - \ln^2(x_{i,r_i-v}))}{2\sum_{i=1}^{c_v}(\ln(x_{i,r_i-v+1}) - \ln(x_{i,r_i-v}))} > 0\,. \tag{7}$$

For $\tau > 0$, it cannot be established that the ML estimate of $\boldsymbol{\vartheta}$ exists with probability 1. This is different from the particular case $\tau = 0$, in which almost-sure existence (and uniqueness) of the ML estimator of $\boldsymbol{\vartheta}$ is guaranteed for $s \geq 2$; see Cramer and Kamps [20] and Bedbur et al. [21] for respective results in an i.i.d. setting ($r_1 = \cdots = r_s$) of sequential OSs.

Example 3.3. Continuing the real data analyses in Example 2.2, the sufficiency condition (7) turns out to be fulfilled for the data set $\mathbf{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_{60})$ of oxide inclusion sizes above $\tau = 6.5$. Using equation (5), the ML estimate of β is obtained as $\hat{\beta}(\mathbf{x}) = 0.72$. The ML estimates of the model parameters $\gamma_1, \dots, \gamma_{23}$ are determined according to formula (6). In Fig. 4, the inferential results are illustrated. On the left, the profile log-likelihood function (4) is plotted for selected values of the shape parameter β, where the model parameters γ_v, $1 \leq v \leq 23$, are replaced by their respective estimates according to formula (6); the maximum is attained at $\hat{\beta}(\mathbf{x}) = 0.72$. On the right, the ML estimates $\hat{\gamma}_v(\mathbf{x})$, $1 \leq v \leq 23$, are shown and plotted against the index $v - 1$, indicating a linear increase in the model parameters.

The idea of a linear increase in the model parameters is further developed in Sections 5-7, including a hypotheses test for a respective functional relationship as well as related statistical inference. In particular, the intercept parameter may serve as a material indicator. However, we shall first demonstrate that there is indeed a need for a more flexible model than OSs to describe our data set.

4. Model check: OSs versus general model (2)

From now on, we consider the model introduced in Section 2 in case that the shape parameter $\beta > 0$ is fixed. By setting $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_{r_1})$, the density in formula (2) is then seen to be of the form

$$\tilde{f}_{\boldsymbol{\gamma}}(\mathbf{x}) = \exp\left\{\sum_{v=1}^{r_1} \gamma_v T_v(\mathbf{x}) - \kappa(\boldsymbol{\gamma})\right\} h(\mathbf{x}) \tag{8}$$

with statistics T_v, $1 \leq v \leq r_1$, as in formula (3) being free of $\boldsymbol{\gamma}$, constant $\kappa(\boldsymbol{\gamma}) = -\sum_{v=1}^{r_1} c_v \ln(\gamma_v)$, and a mapping h being free of $\boldsymbol{\gamma}$, too. The corresponding distributions $\tilde{P}_{\boldsymbol{\gamma}}$, $\boldsymbol{\gamma} \in (0,\infty)^{r_1}$, therefore form a regular exponential family with natural parameter $\boldsymbol{\gamma}$, sufficient and complete statistic $\boldsymbol{T} = (T_1, \dots, T_{r_1})$, and cumulant function κ; cf. Bedbur et al. [22,23], Bedbur and Kamps [24, Section 6.3]. It

is well-known that in this model, the statistics $T_1, \dots, T_{r_1}$ are jointly independent, where $-T_\upsilon \sim \Gamma(c_\upsilon, 1/\gamma_\upsilon)$ has a gamma distribution with shape parameter c_υ and scale parameter $1/\gamma_\upsilon$ for $1 \le \upsilon \le r_1$. Moreover, the ML estimator of $\boldsymbol{\gamma}$ uniquely exists and is given by $\hat{\boldsymbol{\gamma}} = (\hat\gamma_1, \dots, \hat\gamma_{r_1})$ with $\hat\gamma_\upsilon = -c_\upsilon/T_\upsilon$, $1 \le \upsilon \le r_1$; see, e.g., Cramer and Kamps [25]. These findings can be utilized to find a likelihood ratio (LR) statistic for a model check in the sense whether OSs are appropriate to describe the data or have to be rejected in favour of the more general model; see Cramer and Kamps [20] and Bedbur et al. [26] for respective tests in i.i.d. settings ($r_1 = \cdots = r_s$) of sequential and generalized order statistics. In terms of our parametrization, the test problem reads

$$\begin{aligned} H_0 &: \exists \gamma > 0 : \gamma_j = j\gamma, \quad 1 \le j \le r_1 \quad \text{against} \\ H_1 &: \nexists \gamma > 0 : \gamma_j = j\gamma, \quad 1 \le j \le r_1, \end{aligned} \tag{9}$$

where a rejection of H_0 yields that common OSs from any distribution function $F(x) = 1 - \exp\{-\gamma(x^\beta - \tau^\beta)\}$, $x > \tau$, for some $\gamma > 0$ are inappropriate to describe the data.

For $\Theta_0^M = \{(\gamma, 2\gamma, \dots, r_1\gamma) : \gamma > 0\}$, the LR statistic is obtained as

$$\Lambda_M = -2\ln\left(\sup_{\boldsymbol{\gamma}\in\Theta_0^M} \frac{\tilde f_{\boldsymbol{\gamma}}}{\tilde f_{\hat{\boldsymbol{\gamma}}}}\right) = \inf_{\gamma>0}\left(2\sum_{\upsilon=1}^{r_1} c_\upsilon \left\{\frac{\upsilon\gamma}{\hat\gamma_\upsilon} - \ln\left(\frac{\upsilon\gamma}{\hat\gamma_\upsilon}\right) - 1\right\}\right).$$

As the global minimum value is attained at $\hat\gamma = c[\sum_{\upsilon=1}^{r_1}(\upsilon c_\upsilon/\hat\gamma_\upsilon)]^{-1}$, it can be rewritten as

$$\Lambda_M = -2\sum_{\upsilon=1}^{r_1} c_\upsilon \ln\left(\frac{\upsilon\hat\gamma}{\hat\gamma_\upsilon}\right) \quad \text{with} \quad \frac{\upsilon\hat\gamma}{\hat\gamma_\upsilon} = \frac{c}{c_\upsilon}\frac{\upsilon T_\upsilon}{\sum_{k=1}^{r_1} kT_k}, \quad 1 \le \upsilon \le r_1.$$

Note that the statistics $-\upsilon T_\upsilon$, $1 \le \upsilon \le r_1$, are jointly independent and that, under the null hypothesis, $-\upsilon T_\upsilon$ is gamma-distributed with shape parameter c_υ and scale parameter $1/\gamma$ for $1 \le \upsilon \le r_1$ and some $\gamma > 0$. As a consequence, Λ_M has a single null distribution, i.e., a distribution not depending on the specific value of γ if H_0 is true. The LR test with level $p \in (0,1)$ therefore rejects the null hypothesis if and only if $\Lambda_M > d_M$, where d_M is the $(1-p)$-quantile of the random variable

$$-2\sum_{\upsilon=1}^{r_1} c_\upsilon \ln\left(\frac{cD_\upsilon}{c_\upsilon}\right) \tag{10}$$

with the random vector $(D_1, \dots, D_{r_1})$ following a Dirichlet distribution with parameters $c_1, \dots, c_{r_1}$. Simulation can now be used to generate a large sample from random variable (10), the empirical $(1-p)$-quantile of which may then serve as an estimate of d_M.

Example 4.1. Referring to the real data set of oxide inclusion sizes, we continue with Example 3.3 and perform the model test for test problem (9) with $r_1 = 23$. Based on the recorded inclusion sizes $\mathbf{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_{60})$, the LR statistic attains the value $\Lambda_M(\mathbf{x}) = 34.73$. Assuming a significance level of $p = 0.05$, the critical value is obtained as $d_M = 34.62$ based on 10^6 simulations. Hence, the null hypothesis of test problem (9) can be rejected at level 5% (and at every larger significance level). This finding indicates that common OSs based on i.i.d. random variables (with truncated Weibull distributions) are not appropriate for modelling the observed inclusion sizes in our data set.

5. Testing for a linear link function

If the number s of samples in the model proposed in Section 4 is small, every component of $\boldsymbol{\gamma}$ is estimated on the basis of only a few observations, which may result in inaccurate values. We therefore aim at a reduction of the number of involved parameters by following a link-function approach. Such a link function connects the parameters among themselves and yields an additional model assumption, which, of course, has to be justified in some sense, for instance, by a statistical test; see Balakrishnan et al. [27], Volovskiy et al. [28], and Bedbur and Seiche [29] for a respective account in an accelerated lifetime model based on generalized order statistics.

Let $y_1, \dots, y_{r_1}$ be fixed real numbers taking at least two different values. We are interested in testing the null hypothesis

$$H_0 : \quad \exists (a,b) \in \Xi : \gamma_\upsilon = a + by_\upsilon, \quad 1 \le \upsilon \le r_1, \tag{11}$$

corresponding to the presence of a linear link function connecting $\gamma_1, \dots, \gamma_{r_1}$. Here,

$$\Xi = \{(a,b) \in \mathbb{R}^2 : a + by_\upsilon > 0, 1 \le \upsilon \le r_1\} \tag{12}$$

denotes the set of admissible link-function parameters.

In what follows, we develop an asymptotic likelihood ratio test to check for null hypothesis (11). To this end, ML estimation of $\boldsymbol{\gamma}$ under the null hypothesis is considered, first.

Lemma 5.1. *The ML estimator of $\boldsymbol{\gamma}$ in*

$$\Theta_0^L = \{(a + by_1, \dots, a + by_{r_1}) : a, b \in \Xi\}$$

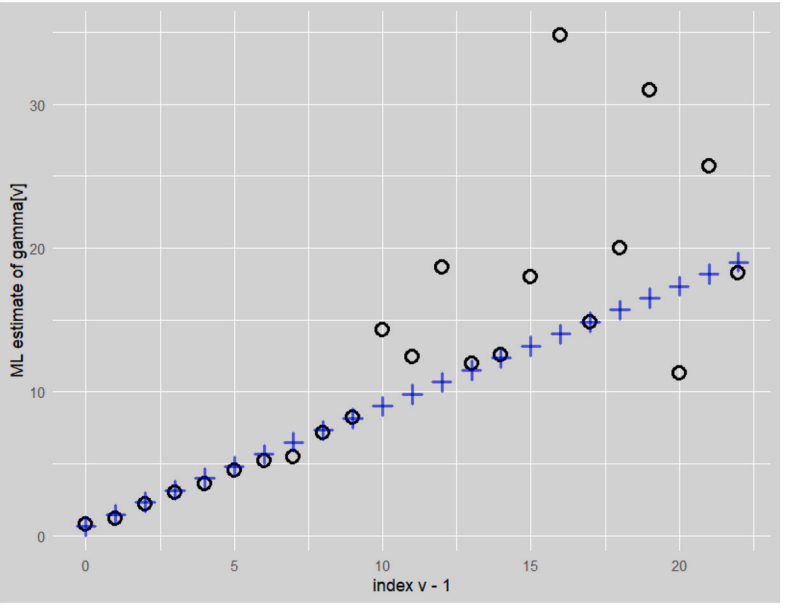

Fig. 5. ML estimates $\hat{\gamma}_v(\mathbf{x})$ in black and plug-in estimates $\hat{\gamma}_v^L(\mathbf{x})$ in blue of γ_v plotted against $v-1$ (weights $y_v = v-1$).

exists, is unique, and given by

$$\hat{\boldsymbol{\gamma}}^L = \left(\hat{a}+\hat{b}y_1, \ldots, \hat{a}+\hat{b}y_{r_1}\right),$$

where $(\hat{a},\hat{b})$ *is the only solution of the equations*

$$a = -\frac{bW+c}{V} \qquad \text{and} \qquad \sum_{v=1}^{r_1} \frac{c_v y_v V}{c+b(W-y_v V)} = W \tag{13}$$

with respect to $(a,b)\in\Xi$. *Here, the statistics* V *and* W *are defined by* $V=\sum_{v=1}^{r_1} T_v$ *and* $W=\sum_{v=1}^{r_1} y_v T_v$.

Proof. Setting $\gamma_v = a+by_v$, $1\le v\le r_1$, in formula (8), the restricted distribution class $\{P_{\boldsymbol{\gamma}} : \boldsymbol{\gamma}\in\Theta_0^L\}$ is seen to form a regular exponential family with natural parameter (a,b) and sufficient statistic (V,W). By using the density transformation theorem, (V,W) can be shown to have a Lebesgue density, and applying Corollary 2.3.1 and Theorem 2.3.2 in Bickel and Doksum [30] then yields that the ML estimator $(\hat{a},\hat{b})$, say, of (a,b) in Ξ exists, is unique, and given by the only solution of the likelihood equations

$$V+\sum_{v=1}^{r_1}\frac{c_v}{a+by_v}=0 \qquad \text{and} \qquad W+\sum_{v=1}^{r_1}\frac{c_v y_v}{a+by_v}=0$$

with respect to $(a,b)\in\Xi$. These formulas can be rewritten as stated in the theorem by using that $aV+bW=-c$. □

Example 5.2. We continue with Example 3.3 and first recall that fewer than ten control areas provide information on the parameters $\gamma_{17},\ldots,\gamma_{23}$; see Table 1. The corresponding ML estimates will therefore tend to be less accurate than those with smaller indexes. Motivated by Fig. 4, a linear relationship between the model parameter is assumed, i.e., $\gamma_v^L = a+b(v-1)$, $1\le v\le 23$, for some unknown vector $(a,b)\in\Xi$ of link-function parameters. Using formula (13), the ML estimates of a and b are obtained as $\hat{a}(\mathbf{x})=0.64$ and $\hat{b}(\mathbf{x})=0.84$. For $1\le v\le 23$, the plug-in estimate $\hat{\gamma}_v^L(\mathbf{x})=\hat{a}(\mathbf{x})+\hat{b}(\mathbf{x})(v-1)$ is shown in Fig. 5 along with the ML estimate $\hat{\gamma}_v(\mathbf{x})$ of γ_v. A comparison yields a good fit for model parameters with small indexes, which are associated with the occurrence of large inclusions and therefore of more relevance in our study. However, to statistically substantiate our link-function approach, a hypotheses test would be useful.

Having derived the ML estimator of $\boldsymbol{\gamma}$ in Θ_0^L, the LR statistic for testing $H_0 : \boldsymbol{\gamma}\in\Theta_0^L$ against $H_1 : \boldsymbol{\gamma}\notin\Theta_0^L$ is given by $\Lambda_L = -2\ln(\tilde{f}_{\hat{\boldsymbol{\gamma}}^L}/\tilde{f}_{\hat{\boldsymbol{\gamma}}})$. It can be simplified as follows.

Lemma 5.3. *The LR statistic for testing* $H_0 : \boldsymbol{\gamma}\in\Theta_0^L$ *against* $H_1 : \boldsymbol{\gamma}\notin\Theta_0^L$ *is given by*

$$\Lambda_L = 2\sum_{v=1}^{r_1} c_v \ln\left(\frac{\hat{\gamma}_v}{\hat{\gamma}_v^L}\right). \tag{14}$$

Proof. Upon inserting $\hat{\gamma}$ and $\hat{\boldsymbol{\gamma}}^L$ in density (8), we obtain

$$\begin{aligned}\Lambda_L &= 2\sum_{v=1}^{r_1}\left[\left(\hat{\gamma}_v - \hat{\gamma}_v^L\right) T_v + c_v(\ln(\hat{\gamma}_v) - \ln(\hat{\gamma}_v^L))\right] \\ &= 2\sum_{v=1}^{r_1} c_v \left[\frac{\hat{\gamma}_v^L}{\hat{\gamma}_v} - \ln\left(\frac{\hat{\gamma}_v^L}{\hat{\gamma}_v}\right) - 1\right]\end{aligned}$$

by using that $T_v = -c_v/\hat{\gamma}_v$ for $1 \le v \le r_1$. Since

$$\sum_{v=1}^{r_1} c_v \frac{\hat{\gamma}_v^L}{\hat{\gamma}_v} = \sum_{v=1}^{r_1} c_v \left(\hat{a} + \hat{b} y_v\right) \frac{-T_v}{c_v} = -\hat{a}V - \hat{b}W = c$$

by the first equation in formula (13), the expression for Λ_L can be simplified as stated. □

Note that Λ_L does not have a single null distribution, i.e., its distribution depends on the specific parameter $\boldsymbol{\gamma} \in \Theta_0^L$. This gives rise to an asymptotic approach.

Theorem 5.4. *For $1 \le v \le r_1$, let $\lambda_v = c_v/c > 0$ be constant as $c \to \infty$, and let $r_1 \ge 3$. Then, for every $\boldsymbol{\gamma} \in \Theta_0^L$, Λ_L has a $\chi^2(r_1 - 2)$-distribution, i.e., a chi-square distribution with $r_1 - 2$ degrees of freedom, as $c \to \infty$.*

Proof. The statistics $T_{i,v}(\mathbf{x}) = x^{\beta}_{i,r_i-v} - x^{\beta}_{i,r_i-v+1}$, $1 \le i \le c_v$, $1 \le v \le r_1$, are jointly independent, where $-T_{1,v}, \ldots, -T_{c_v,v}$ have an exponential distribution with scale parameter $1/\gamma_v$ for $1 \le v \le r_1$. That is, we have r_1 independent samples with c_v i.i.d. random variables in sample v, $1 \le v \le r_1$, having density $g_{\gamma_v}(x) = \gamma_v e^{-\gamma_v x}$, $x > 0$. Now, introducing the functions

$$\zeta_v(\boldsymbol{\gamma}) = \gamma_v - \frac{y_{r_1} - y_v}{y_{r_1} - y_{r_1-1}}\gamma_{r_1-1} - \frac{y_v - y_{r_1-1}}{y_{r_1} - y_{r_1-1}}\gamma_{r_1}, \qquad \boldsymbol{\gamma} \in (0,\infty)^{r_1},$$

for $1 \le v \le r_1 - 2$, null hypothesis (11) can be rewritten as

$$H_0 : \zeta_1(\boldsymbol{\gamma}) = \ldots = \zeta_{r_1-2}(\boldsymbol{\gamma}) = 0,$$

and the asymptotic distribution of Λ_L for $\boldsymbol{\gamma} \in \Theta_0^L$ is obtained from Bradley and Gart [31, Section 2.4]. □

Applying Theorem 5.4, a test for test problem $H_0 : \boldsymbol{\gamma} \in \Theta_0^L$ against $H_1 : \boldsymbol{\gamma} \notin \Theta_0^L$ with asymptotic level p is now given by

$$\varphi_L = \mathbb{1}_{(\chi^2_{1-p}(r_1-2),\infty)}(\Lambda_L),$$

where $\chi^2_{1-p}(r_1 - 2)$ denotes the $(1-p)$-quantile of $\chi^2(r_1 - 2)$ and the notation $\mathbb{1}_A$ is used for the indicator function of set A.

Example 5.5. Continuing with Example 5.2, we compute the value of the LR statistic in formula (14) and obtain $\Lambda_L(\mathbf{x}) = 31.22$. Assuming an (asymptotic) significance level of $p = 0.05$, the critical value is the 0.95-quantile of the chi-square distribution with $r_1 - 2 = 21$ degrees of freedom, i.e., $\chi^2_{0.95}(21) = 32.76$. Consequently, the null hypothesis $H_0 : \boldsymbol{\gamma} \in \Theta_0^L$ of a linear link function cannot be rejected at level 5%.

6. Tests for the link-function parameters

Once having validated the linear link-function model proposed in Section 5, we arrive at the submodel with densities

$$f^L_{a,b}(\mathbf{x}) = \exp\left\{aV(\mathbf{x}) + bW(\mathbf{x}) + \sum_{v=1}^{r_1} c_v \ln(a + b y_v)\right\} h(\mathbf{x})$$

for $(a,b) \in \Xi$ with Ξ as in formula (12) and V, W as in Lemma 5.1. The uncertainty of the model is then fully captured within the slope b and intercept a of the link function, the latter of which could serve as a material indicator. By making use of the exponential family structure of the densities, uniformly most powerful unbiased (UMPU) tests for each of the link-function parameters can be established. Here, we focus on one-sided test problems; UMPU tests for two-sided test problems and test problems with interval hypotheses can similarly be obtained.

Theorem 6.1. *Let $(a_0, b_0) \in \Xi$ and $p \in (0,1)$.*

(i) *For the one-sided test problem $H_0 : a \le a_0$ against $H_1 : a > a_0$, a UMPU level-p test is given by*

$$\varphi = \mathbb{1}_{(-v_p(W),\infty)}(V),$$

where, for every $w \in \mathbb{R}$, $v_p(w)$ denotes the p-quantile of the conditional distribution of $\sum_{j=1}^{r_1} Z_j$ given $\sum_{j=1}^{r_1} y_j Z_j = -w$ for independent random variables $Z_j \sim \Gamma(c_j, 1/(a_0 + by_j))$, $1 \le j \le r_1$, for some arbitrary $(a_0, b) \in \Xi$.

(ii) *For the one-sided test problem $H_0 : b \le b_0$ against $H_1 : b > b_0$, a UMPU level-p test is given by*

$$\varphi = \mathbb{1}_{(-w_p(V),\infty)}(W),$$

where, for every $v < 0$, $w_p(v)$ denotes the p-quantile of the conditional distribution of $\sum_{j=1}^{r_1} y_j Z_j$ given $\sum_{j=1}^{r_1} Z_j = -v$ for independent random variables $Z_j \sim \Gamma(c_j, 1/(a + b_0 y_j))$, $1 \le j \le r_1$, for some arbitrary $(a, b_0) \in \Xi$.

Proof. The assertions directly follow from Theorem 4.4.1 and Lemma 2.7.2 (ii) in Lehmann and Romano [32]. □

The critical values in Theorem 6.1 have to be obtained numerically via simulation. In case (i), we may set $b = 0$ if $a_0 > 0$, and in case (ii), we may choose $a = 0$ if $b_0 y_j > 0$ for $1 \le j \le r_1$.

Example 6.2. We apply Theorem 6.1 to our data set $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{60})$ of oxide inclusion sizes in order to decide between the hypotheses $H_0 : a \le a_0$ and $H_1 : a > a_0$. Since $y_1 = 0$, this test problem can also be interpreted in terms of the parameter $\gamma_1 = a$ associated with the occurrence of maximum inclusions. The threshold a_0 may be specified, for instance, as

$$a_0 = -\frac{\ln(1-q)}{y_q^{0.72} - \tau^{0.72}},$$

where y_q denotes the empirical q-quantile of all data in $\mathbf{x}$. Here, we choose $q = 0.8$, which gives $y_q = 12.08$ and $a_0 = 0.76$. The test statistic is realized as $V(\mathbf{x}) = -251.15$ and, for a significance level of $p = 0.05$, we obtain the critical value $-v_p(W(\mathbf{x})) = -215.75$ based on 10^6 simulated samples using rejection sampling. Hence, the null hypothesis $H_0 : a \le 0.76$ cannot be rejected at level 5%, which is consistent with the ML estimate $\hat{a}(\mathbf{x}) = 0.64$ of a computed in Example 5.2.

To accept or reject a particular link function in a given situation, a test should be available to check for a simple null hypothesis about (a, b).

Lemma 6.3. *For $(a_0, b_0) \in \Xi$, the LR statistic for the test problem $H_0 : (a, b) = (a_0, b_0)$ against $H_1 : (a, b) \ne (a_0, b_0)$ is given by*

$$\Lambda_0 = 2\left[-c - a_0 V - b_0 W + \sum_{v=1}^{r_1} c_v \ln\left(\frac{\hat{a} + y_v \hat{b}}{a_0 + y_v b_0}\right)\right].$$

Proof. Lemma 5.1 yields that

$$\Lambda_0 = -2\ln\left(\frac{f^L_{a_0,b_0}}{f^L_{\hat{a},\hat{b}}}\right) = 2\left[(\hat{a} - a_0)V + (\hat{b} - b_0)W + \sum_{v=1}^{r_1} c_v \ln\left(\frac{\hat{a} + y_v \hat{b}}{a_0 + y_v b_0}\right)\right],$$

from which the representation is obtained by noting that $\hat{a}V + \hat{b}W = -c$. □

Lemma 6.3 can be used to construct a test

$$\varphi_0 = \mathbb{1}_{(-\infty, c_{1-p})}(\Lambda_0)$$

for $H_0 : (a, b) = (a_0, b_0)$ against $H_1 : (a, b) \ne (a_0, b_0)$ with exact level p. Here, the critical value c_{1-p} is the $(1-p)$-quantile of Λ_0 for $(a, b) = (a_0, b_0)$, which can be obtained empirically via simulation (sample independently from gamma distributions $\Gamma(c_v, 1/(a_0 + b_0 y_v))$, $1 \le v \le r_1$).

7. Confidence bands for the link function

Finally, we point out how to obtain simple Wald-type confidence regions for (a, b), which can be used to construct confidence bands for the graph of the link function in the two-dimensional plane. For this, we first address some asymptotic properties of the ML estimator $(\hat{a}, \hat{b})$.

Lemma 7.1. *For $1 \le v \le r_1$, let $\lambda_v = c_v/c > 0$ be constant as $c \to \infty$. Then, as c tends to infinity, $\sqrt{c}[(\hat{a}, \hat{b}) - (a, b)]$ converges in distribution to a bivariate normal distribution with mean zero and covariance matrix $[\mathbf{J}(a, b)]^{-1}$, where*

$$\mathbf{J}(a,b) = \begin{pmatrix} \sum_{v=1}^{r_1} \frac{\lambda_v}{(a+by_v)^2} & \sum_{v=1}^{r_1} \frac{\lambda_v y_v}{(a+by_v)^2} \\ \sum_{v=1}^{r_1} \frac{\lambda_v y_v}{(a+by_v)^2} & \sum_{v=1}^{r_1} \frac{\lambda_v y_v^2}{(a+by_v)^2} \end{pmatrix} \quad \text{for } (a,b) \in \Xi.$$

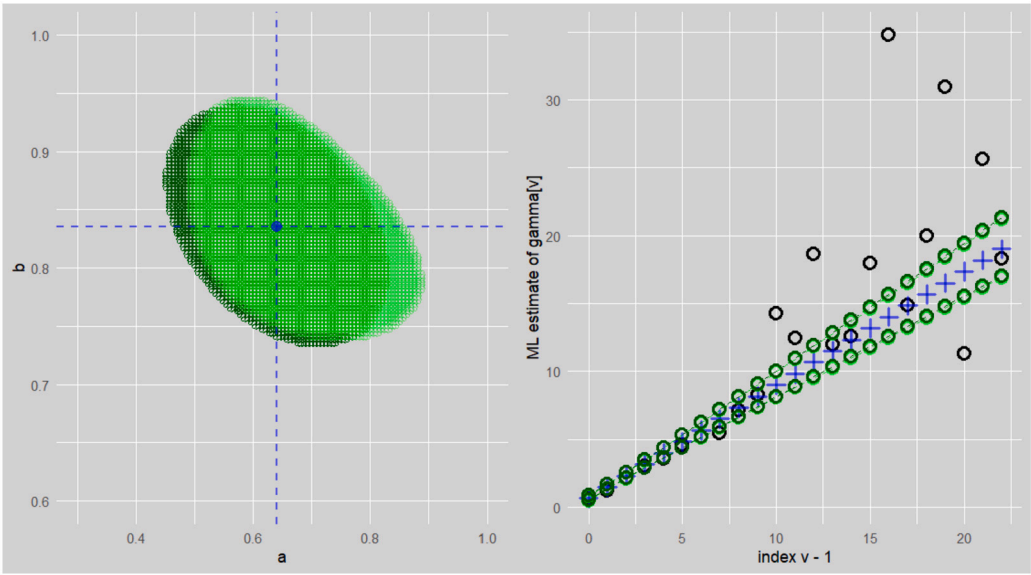

Fig. 6. Confidence regions $C_1(\mathbf{x})$ in light green and $C_2(\mathbf{x})$ in dark green for (a,b) with asymptotic level 95% (left) along with the corresponding confidence bands $B_{C_1(\mathbf{x})}$ and $B_{C_2(\mathbf{x})}$ for the link function (right). The ML estimate of (a,b) and the associated estimated link function are indicated in blue.

In particular, $(\hat{a},\hat{b})$ *is consistent for* (a,b).

Proof. Using the notation in the proof of Theorem 5.4, the data situation is distribution-theoretically equivalent to having r_1 independent samples, where sample $v \in \{1,\dots,r_1\}$ comprises c_v i.i.d. random variables with density $g^{(v)}_{a,b}(x) = (a+by_v)e^{-(a+by_v)x}$, $x>0$; cf. Bedbur and Seiche [29, Section 6]. Here, the Fisher information matrix in sample v reads

$$\mathbf{I}_v(a,b) = \begin{pmatrix} \frac{1}{(a+by_v)^2} & \frac{y_v}{(a+by_v)^2} \\ \frac{y_v}{(a+by_v)^2} & \frac{y_v^2}{(a+by_v)^2} \end{pmatrix}, \qquad (a,b)\in\Xi, \qquad 1\le v\le r_1\,.$$

Theorem 7.1 in Lehmann and Casella [33, p. 475] yields the assertion. □

From Lemma 7.1, the asymptotic distribution of

$$(\hat{a}-a,\hat{b}-b)\mathbf{J}(a,b)(\hat{a}-a,\hat{b}-b)^t$$

is seen to be an exponential distribution with scale parameter 2, which yields that the Wald-type confidence regions

$$C_1 = \left\{(a,b)\in\Xi:\ c\,(\hat{a}-a,\hat{b}-b)\,\mathbf{J}(a,b)\,(\hat{a}-a,\hat{b}-b)^t \le -2\ln(p)\right\}$$

and $$C_2 = \left\{(a,b)\in\Xi:\ c\,(\hat{a}-a,\hat{b}-b)\,\mathbf{J}(\hat{a},\hat{b})\,(\hat{a}-a,\hat{b}-b)^t \le -2\ln(p)\right\}$$

for (a,b) have asymptotic level $1-p$. Among others, such a confidence region C, say, can be used to define a confidence band

$$B_C = \bigcup_{(a,b)\in C} \left\{(v, a+by_v):\ v\in\{1,\dots,r_1\}\right\} \tag{15}$$

for the graph of the linear link function (considered a function of index v). By construction, B_C then has a level not smaller than that of C.

Example 7.2. Continuing with Example 5.2, the ML estimates $\hat{a}(\mathbf{x}) = 0.64$ and $\hat{b}(\mathbf{x}) = 0.84$ for the intercept and slope of the linear link function are used to determine the Wald-type confidence regions $C_1(\mathbf{x})$ and $C_2(\mathbf{x})$ for an (asymptotic) confidence level of $1-p = 0.95$. In Fig. 6, the resulting regions are shown, where $C_2(\mathbf{x})$ is an ellipse by construction and C_1 has a similar shape. Moreover, the corresponding confidence bands $B_{C_1(\mathbf{x})}$ and $B_{C_2(\mathbf{x})}$ are depicted along with the ML estimates $\hat{\gamma}_v(\mathbf{x})$ and plug-in estimates $\hat{\gamma}^L_v(\mathbf{x})$ of γ_v, $1\le v\le 23$. Although both confidence bands may be conservative in the sense that their actual confidence levels are higher than 95%, the realized bands turn out to be narrow and thus informative.

8. Discussion

For a data set with several independent samples of ascendingly ordered measurements above a given threshold, where common order statistics fail to serve as an adequate model within the samples, a model of ordered random variables is fitted, which is flexible concerning growth behaviour of the measurements. Specifically, we consider a data set on sizes of non-metallic (oxide) inclusions in an engineering steel consisting of a total number of 719 observations above the threshold 6.5 (μm), assigned to 60 control areas. Here, inclusion size is regarded as the most relevant geometrical parameter, as large inclusions are identified as primary contributors to damage development during testing. Similar problems and data structures appear, e.g., in other fields of material science.

In the proposed consistent parametric model of ordered random variables, a baseline distribution has to be determined; based on the metallurgy data set there is evidence to choose a left-truncated Weibull distribution. In the model under consideration, joint maximum likelihood estimation of the Weibull shape parameter and model parameters is studied. A model test is offered with the null-hypothesis constituting common order statistics. In case of rejection, we proceed using the general model.

In order to reduce the number of model parameters in case of smaller data sets or to increase precision of inferential procedures, a link-function approach is examined. The parameters of a linear link function, say, may serve as material indicators, in addition. Here, the intercept parameter is associated with the occurrence of maximum inclusions. We provide an asymptotic test to check for the presence of a linear link function as well as tests concerning the parameters of a linear link function. Moreover, in the latter situation, simultaneous confidence regions for the link-function parameters and confidence bands for the graph of the link function are developed. Throughout, the methods are exemplified using the metallurgy data set.

Future work may refer to comparisons of materials and respective statistical inference by means of independent samples of non-metallic inclusion sizes above some threshold in order to detect material differences.

CRediT authorship contribution statement

Stefan Bedbur: Writing – original draft, Software, Formal analysis, Conceptualization, Writing – review & editing, Visualization, Methodology, Data curation. **Udo Kamps:** Writing – original draft, Software, Formal analysis, Conceptualization, Writing – review & editing, Visualization, Methodology, Data curation. **Anja Bettina Schmiedt:** Writing – original draft, Software, Formal analysis, Conceptualization, Writing – review & editing, Visualization, Methodology, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Proofs

A.1. Proof of Theorem 3.1

For $\beta > 0$, let

$$\gamma_v(\beta) = c_v \left[\sum_{i=1}^{c_v} (x_{i,r_i-v+1}^{\beta} - x_{i,r_i-v}^{\beta}) \right]^{-1}, \qquad 1 \le v \le r_1,$$

and

$$d = \sum_{v=1}^{r_1} \sum_{i=1}^{c_v} \ln(x_{iv})$$

for a short notation. Then, by using that $\ln(y) \le \ln(z) - 1 + y/z$ for $y, z > 0$ with strict inequality for $y \ne z$,

$$\begin{aligned} \ell(\boldsymbol{\vartheta}) &= \sum_{v=1}^{r_1} c_v \ln(\gamma_v) + c \ln(\beta) - \sum_{v=1}^{r_1} \frac{c_v \gamma_v}{\gamma_v(\beta)} + (\beta - 1)d \\ &\le \sum_{v=1}^{r_1} c_v [\ln(\gamma_v(\beta)) - 1] + c \ln(\beta) + (\beta - 1)d = \tilde{\ell}(\beta), \quad \text{say}, \end{aligned}$$

with equality if and only if $\gamma_v = \gamma_v(\beta)$ for $1 \le v \le r_1$. Setting

$$\rho_v(\beta) = \frac{1}{\gamma_v(\beta)} = \frac{1}{c_v} \sum_{i=1}^{c_v} (x_{i,r_i-v+1}^{\beta} - x_{i,r_i-v}^{\beta}), \qquad \beta > 0, \qquad 1 \le v \le r_1,$$

the first two derivatives of $\tilde{\ell}$ are given by

$$\tilde{\ell}'(\beta) = -\sum_{v=1}^{r_1} c_v \frac{\rho_v'(\beta)}{\rho_v(\beta)} + \frac{c}{\beta} + d \tag{A.1}$$

$$\text{and} \qquad \tilde{\ell}''(\beta) = \sum_{v=1}^{r_1} c_v \frac{[\rho_v'(\beta)]^2 - \rho_v''(\beta)\,\rho_v(\beta)}{\rho_v^2(\beta)} - \frac{c}{\beta^2}, \qquad \beta > 0.$$

In particular, for $s = 1$, we have $c_1 = \cdots = c_{r_1} = 1$ and thus

$$\begin{aligned}
\tilde{\ell}'(\beta) &= -\sum_{v=1}^{r_1} \frac{\ln(x_{1,r_1-v+1})x_{1,r_1-v+1}^{\beta} - \ln(x_{1,r_1-v})x_{1,r_1-v}^{\beta}}{x_{1,r_1-v+1}^{\beta} - x_{1,r_1-v}^{\beta}} + \frac{r_1}{\beta} + \sum_{v=1}^{r_1} \ln(x_{1v}) \\
&= -\sum_{v=1}^{r_1} \frac{\ln(x_{1v})x_{1v}^{\beta} - \ln(x_{1,v-1})x_{1,v-1}^{\beta}}{x_{1v}^{\beta} - x_{1,v-1}^{\beta}} + \frac{r_1}{\beta} + \sum_{v=1}^{r_1} \ln(x_{1v}) \\
&= -\sum_{v=1}^{r_1} \frac{[\ln(x_{1v}) - \ln(x_{1,v-1})]\, x_{1,v-1}^{\beta}}{x_{1v}^{\beta} - x_{1,v-1}^{\beta}} + \frac{r_1}{\beta} \\
&= -\frac{1}{\beta}\sum_{v=1}^{r_1} \ln\left(\left(\frac{x_{1v}}{x_{1,v-1}}\right)^{\beta}\right) \frac{x_{1,v-1}^{\beta}}{x_{1v}^{\beta} - x_{1,v-1}^{\beta}} + \frac{r_1}{\beta} \\
&> -\frac{1}{\beta}\sum_{v=1}^{r_1} \left(\left(\frac{x_{1v}}{x_{1,v-1}}\right)^{\beta} - 1\right) \frac{x_{1,v-1}^{\beta}}{x_{1v}^{\beta} - x_{1,v-1}^{\beta}} + \frac{r_1}{\beta} = 0,
\end{aligned}$$

such that an ML estimate of β and thus of $\boldsymbol{\vartheta}$ does not exist in that case.

Now, let $s \geq 2$. To establish that $\tilde{\ell}$ is strictly concave, it is sufficient to show that

$$\beta^2 \left([\rho_v'(\beta)]^2 - \rho_v''(\beta)\,\rho_v(\beta)\right) < \rho_v^2(\beta) \tag{A.2}$$

for $\beta > 0$ and $1 \leq v \leq r_1$. For this, let $\beta > 0$ and $v \in \{1, \ldots, r_1\}$ be arbitrary but fixed. Introducing the quantities

$$u_k = \sum_{i=1}^{c_v} t_{ik}$$

with

$$t_{ik} = x_{i,r_i-v+1}^{\beta} \ln^k(x_{i,r_i-v+1}^{\beta}) - x_{i,r_i-v}^{\beta} \ln^k(x_{i,r_i-v}^{\beta}), \qquad 1 \leq i \leq c_v, \qquad k \in \{0,1,2\},$$

inequality (A.2) can be written as $u_1^2 - u_2 u_0 < u_0^2$ or, equivalently, as

$$u_1^2 < u_0\,(u_0 + u_2).$$

Let $1 \leq i \leq c_v$ and $a = x_{i,r_i-v+1}^{\beta}$ and $b = x_{i,r_i-v}^{\beta}$ for brevity. Then we have that $a > b$ and that

$$\begin{aligned}
& (t_{i0} + t_{i2})\, t_{i0} > t_{i1}^2 \\
\Leftrightarrow \quad & [a\,(1 + \ln^2(a)) - b\,(1 + \ln^2(b))]\,(a - b) > [a\ln(a) - b\ln(b)]^2 \\
\Leftrightarrow \quad & a^2 + b^2 - ab\,(\ln^2(a) + \ln^2(b) + 2) > -2ab\ln(a)\ln(b) \\
\Leftrightarrow \quad & (a - b)^2 > ab\,(\ln(a) - \ln(b))^2 \\
\Leftrightarrow \quad & \sqrt{\frac{a}{b}} - \sqrt{\frac{b}{a}} > \ln\left(\frac{a}{b}\right),
\end{aligned}$$

which is true, since the mapping $z \mapsto \sqrt{z} - \sqrt{1/z} - \ln(z)$ with domain $[1, \infty)$ is strictly increasing and thus has a global minimum at $z = 1$. Applying the Cauchy-Schwarz inequality now yields that

$$\begin{aligned}
u_1^2 &\leq \left(\sum_{i=1}^{c_v} \sqrt{t_{i1}^2}\right)^2 < \left(\sum_{i=1}^{c_v} \sqrt{t_{i0} + t_{i2}}\,\sqrt{t_{i0}}\right)^2 \\
&\leq \left(\sum_{i=1}^{c_v} (t_{i0} + t_{i2})\right)\left(\sum_{i=1}^{c_v} t_{i0}\right) = (u_0 + u_2)\,u_0.
\end{aligned}$$

Hence, $\tilde{\ell}$ is strictly concave and the equation $\tilde{\ell}'(\beta) = 0$ with $\tilde{\ell}'$ as in formula (A.1) has at most one solution with respect to $\beta > 0$, which gives the ML estimate of $\boldsymbol{\vartheta}$ based on $\mathbf{x}$ in case of existence. □

A.2. Proof of Lemma 3.2

Let $s \geq 2$, and $\min_{1\leq i\leq c_v} x_{i,r_i-v+1} < \max_{1\leq i\leq c_v} x_{i,r_i-v+1}$ for some $v \in \{1,\dots,r_1\}$, and inequality (7) be true. To show existence of a solution of equation (5) or, equivalently, of the equation $\tilde{\ell}'(\beta) = 0$, $\beta > 0$, with $\tilde{\ell}'$ as in formula (A.1), let $m_v = \max\{x_{1,r_1-v+1},\dots,x_{c_v,r_{c_v}-v+1}\}$ for $1 \leq v \leq r_1$. Then, for every $v \in \{1,\dots,r_1\}$,

$$\begin{aligned}\frac{\rho_v'(\beta)}{\rho_v(\beta)} &= \frac{\sum_{i=1}^{c_v}(x_{i,r_i-v+1}^{\beta}\ln(x_{i,r_i-v+1}) - x_{i,r_i-v}^{\beta}\ln(x_{i,r_i-v}))}{\sum_{i=1}^{c_v}(x_{i,r_i-v+1}^{\beta} - x_{i,r_i-v}^{\beta})}\\ &= \frac{\sum_{i=1}^{c_v}[(x_{i,r_i-v+1}/m_v)^{\beta}\ln(x_{i,r_i-v+1}) - (x_{i,r_i-v}/m_v)^{\beta}\ln(x_{i,r_i-v})]}{\sum_{i=1}^{c_v}[(x_{i,r_i-v+1}/m_v)^{\beta} - (x_{i,r_i-v}/m_v)^{\beta}]}\\ &\to \ln(m_v)\end{aligned}$$

as $\beta \to \infty$, by using that $x_{i,r_i-v} < x_{i,r_i-v+1} \leq m_v$ for $1 \leq i \leq c_v$. Hence, we find

$$\lim_{\beta\to\infty}\tilde{\ell}'(\beta) = -\sum_{v=1}^{r_1} c_v\ln(m_v) + d = \sum_{v=1}^{r_1}\sum_{i=1}^{c_v}(\ln(x_{i,r_i-v+1}) - \ln(m_v)) < 0\,,$$

since all summands are non-positive and at least one summand is negative by assumption.

Moreover, for $1 \leq v \leq r_1$, we have that $\rho_v(0+) = 0$ and

$$\rho_v'(0+), \rho_v''(0+), \rho_v'''(0+) \in (0,\infty)\,,$$

which, by applying l'Hospital's rule twice, yields that

$$\lim_{\beta\to 0+}\frac{\rho_v(\beta) - \beta\rho_v'(\beta)}{\beta\rho_v(\beta)} = \lim_{\beta\to 0+}\frac{-\beta\rho_v''(\beta)}{\rho_v(\beta) + \beta\rho_v'(\beta)} = \lim_{\beta\to 0+}\frac{-\rho_v''(\beta) - \beta\rho_v'''(\beta)}{2\rho_v'(\beta) + \beta\rho_v''(\beta)} = -\frac{\rho_v''(0+)}{2\rho_v'(0+)}\,.$$

Hence,

$$\begin{aligned}\lim_{\beta\to 0+}\tilde{\ell}'(\beta) &= \lim_{\beta\to 0+}\left(d + \frac{c}{\beta} - \sum_{v=1}^{r_1} c_v\frac{\rho_v'(\beta)}{\rho_v(\beta)}\right)\\ &= \lim_{\beta\to 0+}\left(d + \sum_{v=1}^{r_1}\frac{c_v(\rho_v(\beta) - \beta\rho_v'(\beta))}{\beta\rho_v(\beta)}\right)\\ &= d - \sum_{v=1}^{r_1}\frac{c_v\rho_v''(0+)}{2\rho_v'(0+)} > 0\,,\end{aligned}$$

by inequality (7). Since $\tilde{\ell}'$ is continuous, existence of a solution of $\tilde{\ell}'(\beta) = 0$ with respect to $\beta > 0$ is shown, which ensures existence of the ML estimate by Theorem 3.1(ii). □

Data availability

The data set is described by Schmiedt et al. [2,3] and has been provided by the Department of Ferrous Metallurgy at RWTH Aachen University. We have the right to use the data but not to share it.

References

[1] Y. Murakami, Inclusion rating by statistics of extreme values and its application to fatigue strength prediction and quality control of materials, J. Res. Natl. Inst. Stand. Technol. 99 (1994) 345.
[2] A. Schmiedt, H. Dickert, W. Bleck, U. Kamps, Multivariate extreme value analysis and its relevance in a metallographical application, J. Appl. Stat. 41 (2014) 582–595.
[3] A. Schmiedt, H. Dickert, W. Bleck, U. Kamps, Evaluation of maximum non-metallic inclusion sizes in engineering steels by fitting a generalized extreme value distribution based on vectors of largest observations, Acta Mater. 95 (2015) 1–9.
[4] Standard practice for extreme value analysis of nonmetallic inclusions in steel and other microstructural features, Standard E2283-08, ASTM International, West Conshohocken, Pennsylvania, 2003.
[5] Y. Murakami, S. Beretta, Small defects and inhomogeneities in fatigue strength: experiments, models and statistical implications, Extremes 2 (1999) 123–147.
[6] C. Anderson, S. Beretta, J. De Maré, T. Svensson, Technical Recommendations for the Extreme Value Analysis of Data on Large Nonmetallic Inclusions in Steels, P11-02, European Structural Integrity Society, Geesthacht, Germany, 2002.
[7] U. Kamps, A concept of generalized order statistics, J. Stat. Plan. Inference 48 (1995) 1–23.
[8] U. Kamps, A Concept of Generalized Order Statistics, Teubner, Stuttgart, 1995.
[9] U. Kamps, Generalized order statistics, in: N. Balakrishnan, P. Brandimarte, B. Everitt, G. Molenberghs, W. Piegorsch, F. Ruggeri (Eds.), Wiley StatsRef: Statistics Reference Online, Wiley, Chichester, 2016, pp. 1–12.
[10] Z. Cao, Z. Shi, F. Yu, G. Wu, W. Cao, Y. Weng, A new proposed Weibull distribution of inclusion size and its correlation with rolling contact fatigue life of an extra clean bearing steel, Int. J. Fatigue 126 (2019) 1–5.

[11] Z. Fang, B.R. Patterson, M.E. Turner Jr, Modeling particle size distributions by the Weibull distribution function, Mater. Charact. 31 (1993) 177–182.
[12] S. Ghosh, Micromechanical Analysis and Multi-Scale Modeling Using the Voronoi Cell Finite Element Method, CRC Press, 2011.
[13] J.E. Hilliard, L. Lawson, Stereology and Stochastic Geometry, vol. 28, Springer Science & Business Media, 2003.
[14] F. Sczerzenie, G. Vergani, C. Belden, The measurement of total inclusion content in nickel-titanium alloys, J. Mater. Eng. Perform. 21 (2012) 2578–2586.
[15] D.R. Wingo, The left-truncated Weibull distribution: theory and computation, Stat. Pap. 30 (1989) 39–48.
[16] M. Kreer, A. Kızılersü, A.W. Thomas, A.D. Egídio dos Reis, Goodness-of-fit tests and applications for left-truncated Weibull distributions to non-life insurance, Eur. Actuar. J. 5 (2015) 139–163.
[17] A. Kizilersu, M. Kreer, A.W. Thomas, Goodness-of-fit testing for left-truncated two-parameter Weibull distributions with known truncation point, Austrian J. Stat. 45 (2016) 15–42.
[18] B.C. Arnold, J.A. Villasenor, Generalized order statistic processes and Pfeifer records, Statistics 46 (2012) 373–385.
[19] J.P. Schmidt, U. Kamps, Almost sure limit behaviour of Pfeifer record values, Statistics 54 (2020) 830–840.
[20] E. Cramer, U. Kamps, Sequential order statistics and k-out-of-n systems with sequentially adjusted failure rates, Ann. Inst. Stat. Math. 48 (1996) 535–549.
[21] S. Bedbur, M. Johnen, U. Kamps, Inference from multiple samples of Weibull sequential order statistics, J. Multivar. Anal. 169 (2019) 381–399.
[22] S. Bedbur, E. Beutner, U. Kamps, Generalized order statistics: an exponential family in model parameters, Statistics 46 (2012) 159–166.
[23] S. Bedbur, U. Kamps, M. Kateri, Meta-analysis of general step-stress experiments under repeated Type-II censoring, Appl. Math. Model. 39 (2015) 2261–2275.
[24] S. Bedbur, U. Kamps, Multivariate Exponential Families: A Concise Guide to Statistical Inference, Springer, Cham, 2021.
[25] E. Cramer, U. Kamps, Sequential k-out-of-n systems, in: N. Balakrishnan, C.R. Rao (Eds.), Advances in Reliability, in: Handbook of Statistics, vol. 20, Elsevier, Amsterdam, 2001, pp. 301–372.
[26] S. Bedbur, E. Beutner, U. Kamps, Multivariate testing and model-checking for generalized order statistics with applications, Statistics 48 (2014) 1297–1310.
[27] N. Balakrishnan, E. Beutner, U. Kamps, Modeling parameters of a load-sharing system through link functions in sequential order statistics models and associated inference, IEEE Trans. Reliab. 60 (2011) 605–611.
[28] G. Volovskiy, S. Bedbur, U. Kamps, Link functions for parameters of sequential order statistics and curved exponential families, Probab. Math. Stat. 41 (2021) 115–127.
[29] S. Bedbur, T. Seiche, Testing the validity of a link function assumption in repeated type-II censored general step-stress experiments, Sankhya, Ser. B 84 (2021) 106–129.
[30] P.J. Bickel, K.A. Doksum, Mathematical Statistics: Basic Ideas and Selected Topics, Volume I, 2nd ed., Taylor & Francis, Boca Raton, 2015.
[31] R.A. Bradley, J.J. Gart, The asymptotic properties of ML estimators when sampling from associated populations, Biometrika 49 (1962) 205–214.
[32] E.L. Lehmann, J.P. Romano, Testing Statistical Hypotheses, 3rd ed., Springer, New York, 2005.
[33] E.L. Lehmann, G. Casella, Theory of Point Estimation, 2nd ed., Springer, New York, 1998.