SFB 823

Model checks and simultaneous prediction bands for load sharing models in prestressed concrete beams

Kevin Leckey, Jens Heinrich,
Christine H. Müller, Reinhard Maurer

Nr. 32/2021

Discussion Paper

SFB 823

Model checks and simultaneous prediction bands for load sharing models in prestressed concrete beams

Kevin Leckey*† Jens Heinrich‡ Christine H. Müller† Reinhard Maurer‡

January 18, 2022

Abstract

This article presents a new method to test on whether a parametric model is capable of describing data properly. It also introduces a simple procedure to generate simultaneous prediction bands based on independent copies of a process. The performance of these prediction bands, e.g. in a leave-one-out cross-validation, will also be used as another indication of whether data is modeled properly. Both methods are applied to data from fatigue experiments on prestressed concrete beam girders. These experiments highlight a couple of different influences on the fatigue of such girders, namely the so-called cable factor and the deflection force. Both effects are incorporated into different load sharing models for component failures which then are compared and used for predicting these failure times.

Keywords: Accelerated life testing, cable factor, deflection force, load sharing models, K-sign depth.

1 Introduction

Properly modeling and predicting component failures of a system are two key aspects of reliability theory. This article will tackle both problems by first introducing a new approach to check whether a parametric model is capable of describing given data and then presenting a general method to derive prediction bands for stochastic processes.

The methods in this article will be used to compare several new model extensions to a load sharing model for successive failures of tension wires in prestressed concrete beams. There is a vast literature on load sharing models and their applications (Kim and Kvam, 2004; Park, 2010; Phoenix and Newman, 2009; Xu et al., 2017, 2019b; De Oliveira and Colosimo, 2004; Xu et al., 2019a), often in context of so-called k-out-of-n systems

*Corresponding author, Email: kevin.leckey@tu-dortmund.de
†Department of Statistics, TU Dortmund University, Germany
‡Department of Architecture and Civil Engineering, TU Dortmund University, Germany

(Kong and Ye, 2017; Zhang et al., 2020; Cramer and Kamps, 1996; Beutner, 2010) with n components and k failures. The specific models considered in this article are extensions of a load sharing model used in Szugat et al. (2016) and Leckey et al. (2020b). Since the underlying fatigue experiments are expensive and time consuming, only a fairly low number of independent failure series are available. Therefore, distributional assumptions in form of parametric models are necessary. Szugat et al. (2016) and Leckey et al. (2020b) managed to model a first series of experiments published in Heinrich et al. (2016) and Heeke et al. (2019) using a load sharing model with only two parameters. This was only possible since the experiments in this first series only differed in their initial stress range (their "load"), keeping other parameters such as the number of wires and their placement (curvature) within the concrete beams constant.

In order to investigate the influence of certain parameters on the fatigue strength of these prestressed beam girders, additional experiments were conducted which will be presented and analyzed in this paper. Within these experiments, the number and/or size of the tension wires differed from the choice made in the first series of experiments (Heinrich et al., 2016; Heeke et al., 2019). These changes have a major impact on the failure times due to effects like the cable factor and deflection force discussed in Section 3.1. Hence, the first challenge in analyzing this new data set is to construct a suitable model that incorporates these additional effects. Moreover, these experiments are accelerated life time experiments since the experiments would take too long if they ran under realistic stress ranges. Therefore, a second challenge is to extrapolate failure predictions under realistic stress conditions. We are mainly treating the first challenge here, but also present some new results concerning the prediction of a future failure process.

In order to select proper models in our application, a new method based on the so-called 3-sign depth, a special case of the K-sign depth, is introduced. An introduction to the K-sign depth and some basic properties including its asymptotic quantiles can be found in Leckey et al. (2020a) and Malcherczyk et al. (2021). While tests based on the K-sign depth are usually used for parameter testing within a parametric model, the application in this paper shows that they are also capable of detecting whether a model is suited to represent given data. More precisely, they can be used to check whether the median of each coordinate of a process is modeled correctly.

Having found an appropriate model, predicting a new failure process is important. Predicting accelerated life time experiments is treated for example in Patel (1989), Xiong and Milliken (2002), Hong and Meeker (2013), and Leckey et al. (2020b). Moreover, predictions for degradation processes are given for example in De Oliveira and Colosimo (2004), Wang and Xu (2010), and Meeker et al. (1998). Predictions in this article aim to cover the entire trajectory of a process, i.e. all component failures of a system. Such predictions are usually called simultaneous prediction bands and are computationally more challenging than single point predictions. These bands have a variety of different applications such as wind power forecasting (Bessa, 2015), economics (Härdle and Marron, 1991), econometrics (Hymans, 1968), and mortality forecasting (Li and Chan, 2011). A general overview over

simultaneous prediction bands and their most desirable properties can be found in Kolsrud (2007). The predictions introduced in this paper are broadly applicable since they make no specific distributional assumptions other than that i.i.d. copies of the process have to be either available (as data) or can be simulated. In particular, they are very similar to the so-called adjusted intervals in Kolsrud (2007) but are computed more directly with a scoring function based on extreme coordinate-wise ranks.

The paper is organized as follows. Section 2 contains the main statistical methods in their most general setup. In particular, methods for parameter estimation and model checks based on the 3-sign depth are given in Section 2.1 whereas simultaneous prediction bands for arbitrary processes are presented in Section 2.2. The fatigue tests for prestressed concrete beams as well as different parameters/forces that can impact their lifetime will be discussed in Section 3. The resulting statistical models are given in Section 4. These models are compared and used for predictions in Section 5. In particular, they will be compared based on their 3-sign depth as well as the coverage rates that their resulting prediction bands achieve in a leave-one-out cross-validation. Finally, a conclusion and outlook is given in Section 6.

2 Statistical methods

This paper contains a couple of general methods for parameter estimations, model checking, and simultaneous prediction intervals. We start by complementing classical approaches for estimation and model checking with an approach based on 3-sign depth.

2.1 Estimations and model checks in parametric models based on 3-sign depth

Let $\mathbf{Z} = (Z_n)_{1\leq n\leq N}$ be a random vector with some distribution given by a parametric model with a d-dimensional model parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ for some $d \geq 1$. We assume that, for every $n = 1, \ldots, N$,

$$Z_1, \ldots, Z_N \quad \text{are independent,} \tag{1}$$

$$Z_n \text{ has a continuous distribution with some density function } f_{n,\boldsymbol{\theta}}. \tag{2}$$

Note that the density function $f_{n,\boldsymbol{\theta}} : \mathbb{R} \to [0, \infty)$ of Z_n may depend on the index $n \in \{1, \ldots, N\}$. In particular, we do not assume that the coordinates of $\mathbf{Z}$ are identically distributed.

The vector $\mathbf{Z}$ will represent the entire data from all experiments in applications later on. Assumptions (1) and (2) allow us to use the well-known maximum likelihood principle for parameter estimation. This approach as well as the closely related likelihood-ratio (LR) test will be the main classical methods for parameter estimation and confidence sets. For the reader's convenience, these approaches are summarized in the next paragraph.

2.1.1 The classical likelihood approach

Let $\mathbf{z} = (z_n)_{n=1,\dots,N}$ be a realization of the random vector $\mathbf{Z}$. The likelihood function $\mathcal{L}$ maps a parameter $\boldsymbol{\theta}$ to the value of the corresponding joint density function at $\mathbf{z}$, that is

$$\mathcal{L}(\boldsymbol{\theta}) := \mathcal{L}(\boldsymbol{\theta}, \mathbf{z}) := \prod_{n=1}^{N} f_{n,\boldsymbol{\theta}}(z_n). \tag{3}$$

The maximum likelihood estimation (MLE) for $\boldsymbol{\theta}$ is defined as

$$\widehat{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}). \tag{4}$$

With the likelihood function given in (3), an asymptotic $(1-\alpha)$-confidence set for the parameter $\boldsymbol{\theta}$ can be derived using the likelihood-ratio test (Schervish, 1995, pp. 459-461):

$$\mathcal{C}_\alpha^{\mathrm{LR}} := \left\{ \boldsymbol{\theta} \in \mathbb{R}^d;\ -2\ln\left(\frac{\mathcal{L}(\boldsymbol{\theta})}{\mathcal{L}(\widehat{\boldsymbol{\theta}})}\right) \leq \chi^2_{d,1-\alpha} \right\} \tag{5}$$

in which $\chi^2_{d,1-\alpha}$ denotes the $(1-\alpha)$-quantile of the χ^2-distribution with d degrees of freedom.

Model checks with the likelihood approach are only possible by comparing different models via the likelihood. Having only one model, it is not possible to decide whether this model is good or not for a given data set. This is different to the new approach based on sign-depth which is presented in the next paragraphs.

2.1.2 Parameter estimation via 3-sign depth

As a robust counterpart to the likelihood approach, we will use the 3-sign depth as a special case of the K-sign depth that can be found, e.g., in Kustosz et al. (2016) or Malcherczyk et al. (2021). This approach is based on the residuals, which are defined as follows. Let $\mathrm{med}_n(\boldsymbol{\theta})$ denote the median of the coordinate Z_n according to the model with parameter $\boldsymbol{\theta}$, i.e.

$$\int_{-\infty}^{\mathrm{med}_n(\boldsymbol{\theta})} f_{n,\boldsymbol{\theta}}(x)\mathrm{d}x = \frac{1}{2}.$$

If the median is not unique, any choice will be fine for the subsequent method. Then the residuals associated to a parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ are defined as

$$\mathbf{R}(\boldsymbol{\theta}) = (R_1(\boldsymbol{\theta}), \dots, R_N(\boldsymbol{\theta})) \quad \text{with} \quad R_n(\boldsymbol{\theta}) = Z_n - \mathrm{med}_n(\boldsymbol{\theta}).$$

The *3-sign depth* $d_3(\boldsymbol{r})$ of an arbitrary vector $\boldsymbol{r} = (r_1, \dots, r_N) \in \mathbb{R}^N$, $N \in \mathbb{N}$, is defined as the relative number of 3-tuples (r_i, r_j, r_k), $i < j < k$, that have alternating signs. More formally,

$$d_3(\boldsymbol{r}) := \frac{1}{\binom{N}{3}} \sum_{i_1<i_2<i_3} \left(\prod_{j=1}^{3} \mathbb{1}\{(-1)^j r_{i_j} > 0\} + \prod_{j=1}^{3} \mathbb{1}\{(-1)^j r_{i_j} < 0\} \right), \tag{6}$$

in which $\mathbb{1}$ denotes the indicator function, i.e. a $\{0,1\}$-valued function that is 1 if and only if the condition in its brackets holds. An efficient R-implementation of this depth notion can be found in Horn (2020). When applied to the residual vector $(R_1(\boldsymbol{\theta}), \ldots, R_N(\boldsymbol{\theta}))$, a large 3-sign depth indicates a good model fit whereas a small depth is caused by regions with atypically many/few positive residuals and therefore a bad fit. In particular, the 3-sign depth can be used for parameter estimations and confidence sets in the same way $\mathcal{L}$ is used in the likelihood approach. More formally, let $(r_1(\boldsymbol{\theta}), \ldots, r_N(\boldsymbol{\theta}))$ denote a realization of $\mathbf{R}(\boldsymbol{\theta})$. The 3-sign depth of a parameter $\boldsymbol{\theta}$ is defined as

$$d_3(\boldsymbol{\theta}) := d_3(r_1(\boldsymbol{\theta}), \ldots, r_N(\boldsymbol{\theta}))$$

with d_3 given in (6). Using this depth notion, alternatives to the MLE and the LR-based confidence sets can be defined as

$$\begin{aligned} \widehat{\boldsymbol{\theta}}^{\mathrm{SD}} &:= \arg\max_{\boldsymbol{\theta}} d_3(\boldsymbol{\theta}), \\ \mathcal{C}_\alpha^{\mathrm{SD}} &:= \left\{\boldsymbol{\theta} \in \mathbb{R}^3;\, d_3(\boldsymbol{\theta}) \geq q_\alpha^{SD}\right\} \end{aligned} \tag{7}$$

in which the necessary α-quantile q_α^{SD} can be obtained via simulation or the limit theorem in Malcherczyk et al. (2021); see also the R-function `qdepth` from Horn (2020).

Remark 2.1. *Note that by design the 3-sign depth only considers signs of residuals, which makes it outlier robust. Moreover, assuming that the independence* (1) *is given, it only checks whether the model medians* $\mathrm{med}_n(\boldsymbol{\theta})$ *are correct without checking any other properties of the distribution of each coordinate* Z_n*. In particular the latter property makes this approach an interesting alternative to the likelihood method in situations where the distributional assumption based on the density* $f_{n,\boldsymbol{\theta}}$ *is an obvious oversimplification of the true nature behind a given data set. However, both properties automatically cause the 3-sign depth approach to be less powerful than the likelihood method in situations where the distributional assumptions are correct and outlier robustness is not needed.*

Remark 2.2. *The value of the* 3*-sign depth depends on the order of* $r_1(\boldsymbol{\theta}), \ldots, r_N(\boldsymbol{\theta})$*. While the general approach works with any order that does not affect the independence assumption (i.e. orders based on the values* $z_1, \ldots, z_N$ *of the underlying process are not allowed), poorly ordered residuals can result in a test with low power. Suitable orders should ensure that whenever* $\boldsymbol{\theta}$ *deviates from the true underlying parameter, large blocks of consecutive residuals exist that all tend to have the same sign. A discussion on the performance of different orderings in linear models can be found in Horn and Müller (2020); see also Section 4.1 for the order chosen in our application.*

2.1.3 The 3-sign depth for model checks

The 3-sign depth will mainly be used to check which of the models introduced in Section 4 are best suited for the data at hand. In a more general setting, this depth can be used to

test hypotheses of the form

$$H_0 : \text{there is a } \boldsymbol{\theta} \in \mathbb{R}^d \text{ such that } Z_n \text{ has median } \mathrm{med}_n(\boldsymbol{\theta}) \text{ for all } n. \tag{8}$$

If this hypothesis does not hold then, in particular, the parametric model cannot properly describe the distribution of $\mathbf{Z}$. The hypothesis can be tested with a type I error of at most $\alpha \in (0,1)$ via:

$$\text{reject } H_0 \text{ if } d_3(\widehat{\boldsymbol{\theta}}^{\mathrm{SD}}) \leq q_\alpha^{SD}$$

with $d_3(\widehat{\boldsymbol{\theta}}^{\mathrm{SD}})$ and q_α^{SD} as in (7). Moreover, p-values for this test can be generated by computing the value of the distributional function at $d_3(\widehat{\boldsymbol{\theta}}^{\mathrm{SD}})$, e.g., by using `pdepth` from the R-package Horn (2020).

Remark 2.3. *Note that no other model is necessary for this model check. If H_0 in (8) is rejected then the assumed model is not appropriate. However, it also can be used to compare different models as done with the likelihood approach. Then the p-values of the depth tests for the different models can be compared.*

2.2 Prediction bands

With both a model check and parameter estimation at hand, we will now first discuss how to generate prediction bands for a real-valued process $\mathbf{Y} = (Y_i)_{1 \leq i \leq I}$ of length $I \in \mathbb{N}$. In contrast to the previous section, this process will only represent the outcome of a single experiment (with all of its component failures) while the other experiments are usually used to estimate its distribution. We begin with some general theory before adapting this theory to parametric models in Section 2.2.2.

When predicting the trajectory of $\mathbf{Y} = (Y_i)_{1 \leq i \leq I}$, it is reasonable to aim for a so-called simultaneous prediction band defined in, e.g., Kolsrud (2007):

Definition 2.4. *Let $(\ell_i)_{1 \leq i \leq I}$ and $(u_i)_{1 \leq i \leq I}$ be real-valued sequences with $\ell_i < u_i$ for all i. The set $\mathcal{B} = [\ell_1, u_1] \times [\ell_2, u_2] \times \cdots \times [\ell_I, u_I]$ is called simultaneous prediction band for $\mathbf{Y} = (Y_i)_{1 \leq i \leq I}$ with coverage probability $1 - \alpha \in (0,1]$ if*

$$\mathbb{P}(\mathbf{Y} \in \mathcal{B}) = \mathbb{P}\left(\bigcap_{i=1}^{I} \{\ell_i \leq Y_i \leq u_i\} \right) = 1 - \alpha.$$

In contrast to simple one-dimensional prediction intervals, there are several approaches that lead to different simultaneous prediction bands. Roughly speaking, this difference is caused by the different options for the point-wise coverage probabilities $\mathbb{P}(\ell_i \leq Y_i \leq u_i)$, $i = 1, \ldots, I$, and the possibility to add asymmetry by choosing $\mathbb{P}(Y_i < \ell_i) \neq \mathbb{P}(Y_i > u_i)$. A more thorough discussion on different properties and approaches to derive prediction bands can be found in Kolsrud (2007); see also Claeskens and Van Keilegom (2003) and Chew (1968) for other approaches and special cases. The prediction bands in the subsequent analysis will aim for the arguably most natural form of prediction bands with the following properties for all $i, j = 1, \ldots, I$:

(a) $\mathbb{P}(\ell_i \le Y_i \le u_i) = \mathbb{P}(\ell_j \le Y_j \le u_j)$,

(b) $\mathbb{P}(Y_i < \ell_i) = \mathbb{P}(Y_i > u_i)$.

Such a prediction band can be obtained by a procedure called *adjusted intervals* in Kolsrud (2007). This procedure starts by letting $[\ell_i, u_i]$ be $(1-\beta)$-prediction intervals for Y_i for every i, initially with $\beta = \alpha$. Afterwards, β is continuously decreased until the resulting band indeed becomes a simultaneous $(1-\alpha)$-prediction band. A description of a stepwise procedure to generate such intervals by a simulation can be found in Kolsrud (2007).

Since the computation of adjusted intervals given in Kolsrud (2007) is fairly time consuming, we will present a more direct approach that yields similar results. The approach only requires that i.i.d. copies of the process $\mathbf{Y} = (Y_i)_{1 \le i \le I}$ are available. In particular, the coordinates $Y_1, \dots, Y_I$ of the process may be dependent and have an arbitrary distribution (distributional assumptions are only required for simulations and are therefore unnecessary in data applications with i.i.d. samples).

2.2.1 Simulation procedure

First generate[1] a total of M samples from the distribution of $\mathbf{Y} = (Y_i)_{1 \le i \le I}$ where the number M of samples is chosen sufficiently large (we used $M = 1000$ for the prediction bands in Section 5). Let $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(M)}$ denote these realizations of i.i.d. copies of $\mathbf{Y} = (Y_i)_{1 \le i \le I}$. Roughly speaking, we may now generate a prediction band by keeping all samples except for the α-fraction that has either one of the highest or lowest coordinate-wise ranks in one of its coordinate i. More precisely, for each coordinate $i \in \{1, \dots, I\}$, let $\mathbf{r}_i := (r_i^{(1)}, r_i^{(2)}, \dots, r_i^{(M)})$ denote the rank vector of $(y_i^{(1)}, y_i^{(2)}, \dots, y_i^{(M)})$, that is

$$r_i^{(m)} = \#\{n \in \{1, \dots, M\};\ r_i^{(n)} \le r_i^{(m)}\}, \quad m = 1, \dots, M,$$

where $\#\mathcal{S}$ denotes the cardinality of a finite set $\mathcal{S}$. Based on these ranks, a score c_m is assigned to each sample $\mathbf{y}^{(m)}$, $m \in \{1, \dots, M\}$:

$$c_m := \max_{i=1,\dots,I} \left| r_i^{(m)} - \frac{M+1}{2} \right|.$$

Note that a sample has a very high score if and only if at least one of its coordinates has a very low or high rank. To actually build a prediction band which covers all but the α-fraction of samples with the most extreme ranks, we may therefore choose the smallest band that covers the set

$$\mathcal{S} := \left\{ \mathbf{y}^{(m)};\ m \in \{1, \dots, M\} \text{ with } c_m \le q_{1-\alpha}(c_1, \dots, c_M) \right\}$$

[1] or use a given data set consisting of i.i.d. samples

in which $q_{1-\alpha}(c_1, \ldots, c_M)$ denotes the empirical $(1-\alpha)$-quantile of $c_1, \ldots, c_M$. Such a band covering $\mathcal{S}$ is called the *envelope* of $\mathcal{S}$, cf. Kolsrud (2007). Its lower and upper bounds $(\ell_i)_{i=1,\ldots,I}$ and $(u_i)_{i=1,\ldots,I}$ are given by

$$\ell_i := \min\{y_i;\, \mathbf{y} = (y_1, \ldots, y_I) \in \mathcal{S}\},$$
$$u_i := \max\{y_i;\, \mathbf{y} = (y_1, \ldots, y_I) \in \mathcal{S}\}.$$

2.2.2 Prediction bands for parametric models

The simultaneous prediction bands in Section 2.2.1 were designed for a process $\mathbf{Y} = \mathbf{Y}_0$ with a fixed distribution. Since the underlying model parameter $\boldsymbol{\theta}$ in a parametric model is usually unknown, a simulation of these prediction bands is only possible when estimating the parameter in advance based on other processes $\mathbf{Y}_1, \ldots, \mathbf{Y}_J$ resulting in a data set $\mathbf{Z} := (\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_J^\top)^\top$. Two different approaches called *naive* and *confidence-set-based* are presented below.

Let $\alpha \in (0,1)$. For every possible parameter $\boldsymbol{\theta}$, the method in Section 2.2.1 can be used to derive a $(1-\alpha)$-prediction band

$$\mathcal{B}_\alpha(\boldsymbol{\theta}) := [\ell_1(\boldsymbol{\theta}),\, u_1(\boldsymbol{\theta})] \times \cdots \times [\ell_I(\boldsymbol{\theta}),\, u_I(\boldsymbol{\theta})]$$

by generating i.i.d. copies of $\mathbf{Y}_0$ assuming that $\boldsymbol{\theta}$ is the true model parameter. The so-called *naive/plug-in prediction* will simply take the MLE $\widehat{\boldsymbol{\theta}} := \widehat{\boldsymbol{\theta}}(\mathbf{Z})$ (or any other consistent estimation) and derive the prediction band with $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. Hence, the naive/plug-in prediction is given by

$$\mathcal{B}_\alpha(\widehat{\boldsymbol{\theta}}) = [\ell_1(\widehat{\boldsymbol{\theta}}),\, u_1(\widehat{\boldsymbol{\theta}})] \times \cdots \times [\ell_I(\widehat{\boldsymbol{\theta}}),\, u_I(\widehat{\boldsymbol{\theta}})]. \tag{9}$$

Such a prediction usually works well for large sample sizes since in this case, a consistent estimator $\widehat{\boldsymbol{\theta}}$ will be close to the true model parameter. However, if the estimator deviates substantially from the true model, a naive prediction can become liberal in the sense that its prediction error exceeds the α-level it is designed for.

In order to ensure that the α-level is kept, we will use another approach based on confidence sets. To this end, fix $\alpha_1, \alpha_2 \in (0,1)$ with $\alpha_1 + \alpha_2 = \alpha$. First derive a $(1-\alpha_1)$-confidence set $\mathcal{C}_{\alpha_1} := \mathcal{C}_{\alpha_1}(\mathbf{Z})$ for $\boldsymbol{\theta}$, e.g., by using the LR-approach presented in Section 2.1.1. Then derive a $(1-\alpha_2)$-prediction band $\mathcal{B}_{\alpha_2}(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \mathcal{C}_{\alpha_1}$. Finally, use

$$\mathcal{B}_{\alpha_2}(\mathcal{C}_{\alpha_1}) := \bigcup_{\boldsymbol{\theta} \in \mathcal{C}_{\alpha_1}} \mathcal{B}_{\alpha_2}(\boldsymbol{\theta}) \tag{10}$$

as a $(1-\alpha)$-prediction for $\mathbf{Y}$. Note that the prediction error is indeed at most α since if $\boldsymbol{\theta}^*$ denotes the true model parameter then

$$\mathbb{P}_{\boldsymbol{\theta}^*}(\mathbf{Y} \in \mathcal{B}_{\alpha_2}(\mathcal{C}_{\alpha_1})) \geq \mathbb{P}_{\boldsymbol{\theta}^*}\left(\mathbf{Y} \in \mathcal{B}_{\alpha_2}(\boldsymbol{\theta}^*), \boldsymbol{\theta}^* \in \mathcal{C}_{\alpha_1}\right) \geq 1 - \alpha_1 - \alpha_2$$

in which the last inequality follows from the union bound for the complementary event $\{\mathbf{Y} \notin \mathcal{B}_{\alpha_2}(\boldsymbol{\theta}^*)\} \cup \{\boldsymbol{\theta}^* \notin \mathcal{C}_{\alpha_1}\}$. Also note that if the data used for $\mathcal{C}_{\alpha_1}$ is independent from $\mathbf{Y}$ then the last bound can be replaced with $(1-\alpha_1)(1-\alpha_2)$, thus allowing to use slightly larger errors α_1, α_2. Finally note that the confidence set C_{α_1} is usually approximated by a grid. The choice of its mesh width results in a trade-off between the precision and computational effort for $\mathcal{B}_{\alpha_2}(\mathcal{C}_{\alpha_1})$.

Remark 2.5. *Since set-unions are inconvenient to compute, the predictions in Section 5 use the potentially larger prediction band*

$$\widetilde{\mathcal{B}}_{\alpha_2}(\mathcal{C}_{\alpha_1}) := [\ell_1, u_1] \times \ldots \times [\ell_I, u_I]$$

with lower- and upper bounds given by

$$\ell_i := \min\{\ell_i(\boldsymbol{\theta});\, \boldsymbol{\theta} \in \mathcal{C}_{\alpha_1}\}, \quad u_i := \max\{u_i(\boldsymbol{\theta});\, \boldsymbol{\theta} \in \mathcal{C}_{\alpha_1}\}.$$

Remark 2.6. *In the subsequent analysis, the error rates will be chosen as $\alpha = 0.05$ in the naive prediction bands and $\alpha_1 = \alpha_2 = 0.05$ in bands based on confidence sets. Note that if the sample size N (i.e. the dimension of $\mathbf{Z}$) is very large, it is advisable to choose $\alpha_1 := \alpha_1(N)$ in such a way that $\alpha_1(N) \to 0$ with a sufficiently slow rate of convergence that still ensures that $\mathcal{C}_{\alpha_1(N)}$ converges to a single point set consisting of the true model parameter. As mentioned in Leckey et al. (2020b, Remark 2), such a convergence is ensured if $\chi^2_{3,1-\alpha_1(N)}/N$ converges to zero. However, since the amount of data used in Section 5 is fairly small ($N < 300$), we decided to stick to $\alpha_1 = 0.05$ which yields a reasonably sized confidence set. Finally note that the theoretical error for prediction bands based on confidence sets is chosen twice as large ($\alpha_1 + \alpha_2 = 0.1$) as the one in the naive approach since despite this larger theoretical error it always holds that $\widehat{\boldsymbol{\theta}} \in \mathcal{C}^{\mathrm{LR}}_{\alpha_2}$ and therefore*

$$\mathcal{B}_{0.05}(\widehat{\boldsymbol{\theta}}) \subseteq \mathcal{B}_{0.05}(\mathcal{C}^{\mathrm{LR}}_{0.05}).$$

Remark 2.7. *Simultaneous predictions bands in combination with a leave-one-out cross-validation are another good indication for whether a (parametric) model describes the data properly. In particular if the coverage rate of the more conservative predictions $\mathcal{B}_{\alpha_2}(\mathcal{C}^{\mathrm{LR}}_{\alpha_1})$ is much lower than $1-(\alpha_1+\alpha_2)$, then the model is probably not capable of modeling the data correctly.*

3 Component failures in prestressed concrete beams

Heinrich et al. (2016), Szugat et al. (2016), Heeke et al. (2019), and Leckey et al. (2020b) present experiments and their analysis of test series with prestressed beams where only the initial stress range was varied. These tests series are called *reference test series* in the following. They are complemented here by test series called *GS*, *GR* and *GL* where additional test parameters were varied.

All test specimens were 4.50m long, 1m high and 0.30m wide prestressed concrete beams. The tendons of the beams consisted of seven-wire strands installed in a steel duct and was also arranged curved over the length of the girder (cf. Figure 1). The tests were performed as 4-point bending tests with a cyclic loading. Here, the level of cyclic loading had a significant influence on the test duration. As the level of cyclic loading increases, the test duration decreases accordingly. At a test frequency of 3-10Hz, the test duration was already several months with more than 10^8 endured load cycles.

The end of the test was usually initiated by a failure of the test girder due to fatigue. However, the failure did not occur abruptly. Rather, a successive increase in single wire breaks of the tendon was observed. As soon as a critical number of wires broke, a total collapse of the test occurred.

The time of the individual wire breaks (number of load cycles) could already be accurately determined during the test procedure by simultaneous microphone, acceleration and crack width measurements. Further information on the test execution and the results of the reference test series is given in Heinrich et al. (2016) and Heeke et al. (2019).

The tests of the GS, GR and GL test series differ from the reference test series in particular by the variation of the tendon geometry (tendon size and tendon alignment). This was intended to investigate certain influences on the fatigue strength. The differences and the influences are described below in more detail.

3.1 Influences on fatigue strength

3.1.1 General comments

The list of influencing parameters on the fatigue strength of a prestressed concrete beam is long. In Suresh (1998) and Schijve (2009), a selection of the most important influencing parameters is listed.

In the GS, GR and GL test series, the prestressing force was nearly doubled in the tendon. The increase of the preload force therefore also increases the deflection force or the lateral load in the area of the tendon deflection (influence of the deflection force).

The prestressing force was increased to compensate for the enlarged prestressing steel surface, since the prestressing steel stress should be at the same level as in the reference tests. Due to the larger number of prestressing steel strands or wires, it can be expected that the local lateral pressures between the prestressing steel wires adjacent to the duct will increase (influence of the cable factor).

3.1.2 Influence of the deflection force and lateral loads

Post-tensioned tendons are usually installed in a curved arrangement as it was the actual case with the test specimens. The course of the tendon corresponds to the bending moment course. The garland-shaped arrangement of the duct in the concrete beam ensures that the

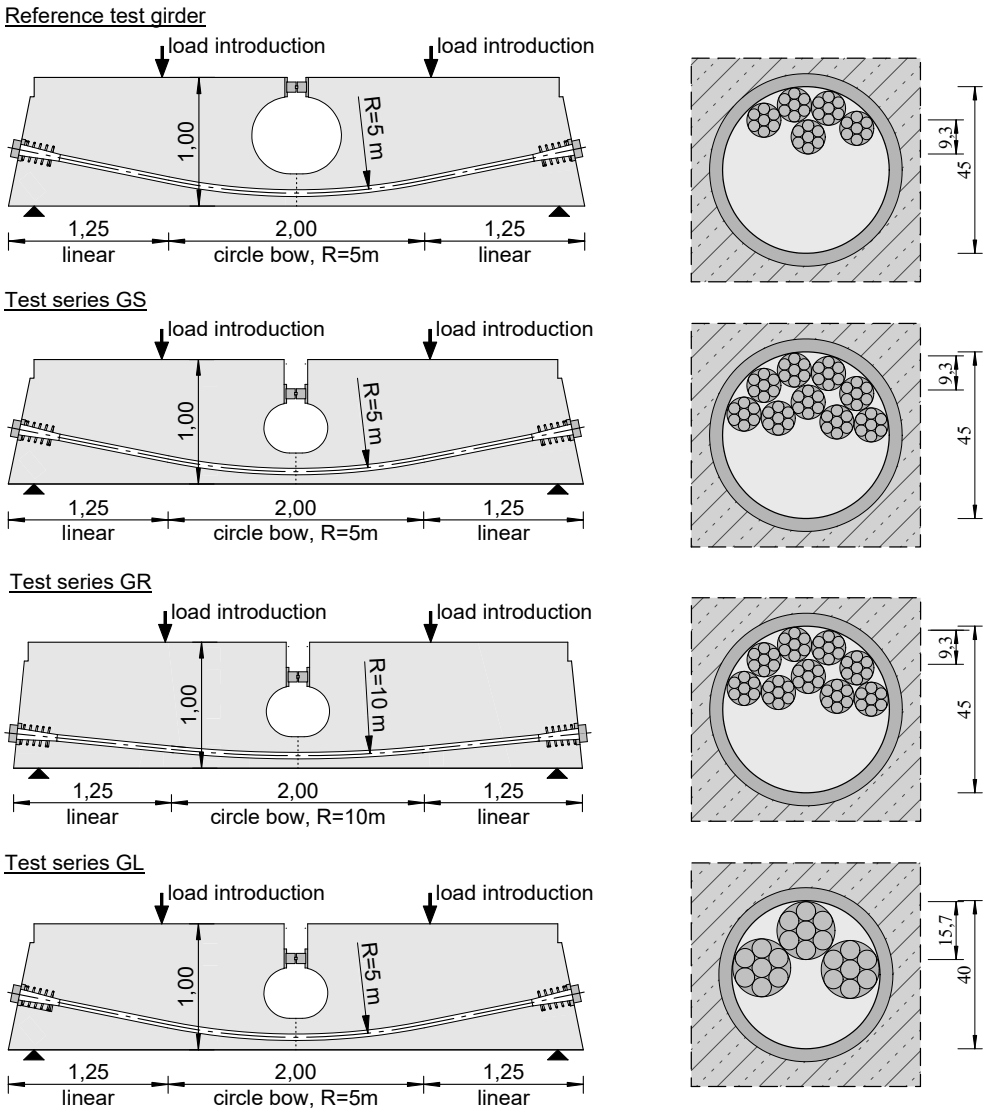

Figure 1: Overview of the conditions and variants of the different test series.

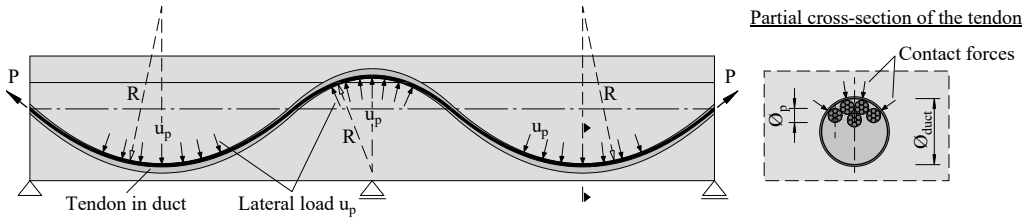

Figure 2: Lateral loads on curved tendons in post-tensioned concrete beams

prestressing steel is in direct contact with the duct when the prestressing load is applied (Figure 2).

The deflection forces at the contact points between the prestressing steel strands and the ducts depend on the prestressing force P and the radius of curvature R and can be determined according to Equation (11):

$$u_p = \frac{P}{R}. \tag{11}$$

As the prestressing force P increases, the deflection force also increases. Various investigations have already shown that lateral pressure can have an impact on the fatigue strength (Remitz and Empelmann, 2018; Hills and Nowell, 1994; Ochi et al., 1970; Nishioka and Hirakawa, 1969). With increasing lateral pressure up to from 0 to 100 MPa, a significant decrease in fatigue strength was observed.

3.1.3 Influence of the cable factor

The deflection force calculated according to Equation (11) is a mean deflection force related to the complete prestressing steel surface. If the radius of curvature is kept constant in the deflection areas (R = const), this also results in a constant deflection force.

The illustration on the right in Figure 2 shows, however, that the deflection force between prestressing steel and duct is not equally distributed over all strands. In the diagram, the middle strand is supported by the two strands located above it. The deflection force from the five strands is therefore only transmitted over four contact points. The same problem also occurs with the single wires of the strands.

This effect is taken into consideration by the cable factor $\kappa_{\max}$, which describes the maximum contact load between prestressing steel and duct in relation to the number of prestressing steel elements.

Two simplified calculation methods for estimating the maximum cable factor are explained in Weiher (2007), Weiher et al. (2008) and Wollmann et al. (1988). In the following, the approach of Weiher (2007), Weiher et al. (2008) will be considered:

$$\kappa_{\max} = 2 \cdot \frac{\varnothing_p}{\varnothing_{duct}} \cdot n_{str}. \tag{12}$$

This equation includes the strand diameter ($\varnothing_p$), the number of strands (n_{str}) and the inner duct diameter ($\varnothing_{duct}$). The maximum deflection force at the contact points between tendon and duct is then determined according to Equation (13):

$$u_{p,\max} = \kappa_{\max} \cdot \frac{P}{R \cdot n_{str}}. \tag{13}$$

3.2 Description of the test specimens

Table 1 and Figure 1 list the relevant test parameters for the GS, GR and GL test series and the reference tests. The following abbreviations are used in Table 1:

- n_{str}: the total number of strands;
- $\mathcal{I}$: the total number of wires, i.e. $\mathcal{I} = 7n_{str}$ since each strand consists of exactly seven wires in all experiments below;
- A_p: the cross-sectional area of the prestressed steel;
- $\varnothing_p$: the strand diameter;
- $\varnothing_{duct}$: the inner duct diameter;
- R: the radius of curvature;
- $P_{\max}$: the maximal prestressing force resulting from the top load of the cyclic load;
- $\sigma_{\min}$: the minimal initial stress (which was 900MPa in all experiments);
- $\sigma_{\max}$: the maximal initial stress ($\sigma_{\max} = P_{\max}/A_p$);
- s: the initial stress range ($s = \sigma_{\max} - \sigma_{\min}$);
- I: the total number of wires breaks detected during the experiment;
- T_{End}: total number of load cycles endured until the end of the experiment;
- $\kappa_{\max}$: the cable factor according to (12);
- u_p: the deflection force according to (11).

Table 1: Relevant test parameters for the GS, GR, GL and the reference test series.

Parameter	Unit	Ref.	GS01	GS02	GR01	GR02	GL01	GL02
n_{str}	[-]	5	9	9	9	9	3	3
$\mathcal{I}$	[-]	35	63	63	63	63	21	21
A_p	$[mm^2]$	260	468	468	468	468	450	450
$\varnothing_p$	$[mm]$	9.3	9.3	9.3	9.3	9.3	15.7	15.7
$\varnothing_{duct}$	$[mm]$	45	45	45	45	45	40	40
R	$[m]$	5	5	5	10	10	5	5
$P_{\max}$	$[kN]$	247 - 352	477	463	477	463	459	445
$\sigma_{\min}$	$[N/mm^2]$	900	900	900	900	900	900	900
$\sigma_{\max}$	$[N/mm^2]$	950 - 1355	1020	990	1020	990	1020	990
s	$[N/mm^2]$	50 - 455	120	90	120	90	120	90
I	[-]	1 - 19	33	29	32	29	6	6
T_{End}	[million]	0.2 - 108	17.6	4.0	6.6	5.6	1.7	19.4
$\kappa_{\max}$	[-]	2.07	3.72	3.72	3.72	3.72	2.36	2.36
u_p	$[kN/m]$	49.4 - 70.5	95.5	92.7	47.7	46.3	91.8	89.1

4 Statistical models

The older data from the reference tests can be modeled fairly well using the so-called Basquin load sharing model from Szugat et al. (2016) or Leckey et al. (2020b). In this model, the interarrival times between consecutive wire breaks are assumed to be independent and exponentially distributed with increasing rates that incorporate the load sharing effect. More precisely, let J denote the total number of experiments and I_j the number of broken wires observed in the j-th experiment, $j = 1, \ldots, J$. Moreover, let $T_{j,i}$ denote the time (in loadcycles) when the i-th wire in the j-th experiment breaks. Then the interarrival times are defined as

$$W_{j,i} := T_{j,i} - T_{j,i-1}, \quad j = 1, \ldots, J,\ i = 1, \ldots, I_j,$$

with the convention that $T_{j,0} := 0$ for all j. The models given in Szugat et al. (2016) or Leckey et al. (2020b) make the following assumptions:

$$W_{j,i},\ j = 1, \ldots, J,\ i = 1, \ldots, I_j, \text{ are independent,} \tag{A1}$$

$$W_{j,i} \text{ is exponentially distributed with rate } \lambda_{\boldsymbol{\theta}}(j, i), \tag{A2}$$

in which $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \mathbb{R}^2$ is a model parameter and the rate is given by

$$\lambda_{\boldsymbol{\theta}}(j, i) := \lambda_{\boldsymbol{\theta}}^{\text{Basq}}(j, i) := \mathrm{e}^{-\theta_1} \cdot \left(\frac{s_j \mathcal{I}_j}{\mathcal{I}_j - (i-1)} \right)^{\theta_2} \tag{14}$$

where $\mathcal{I}_j$ and s_j denote the total number of wires and initial stress range in the j-th experiment. These failure rates are referred to as the *Basquin model* since they are based

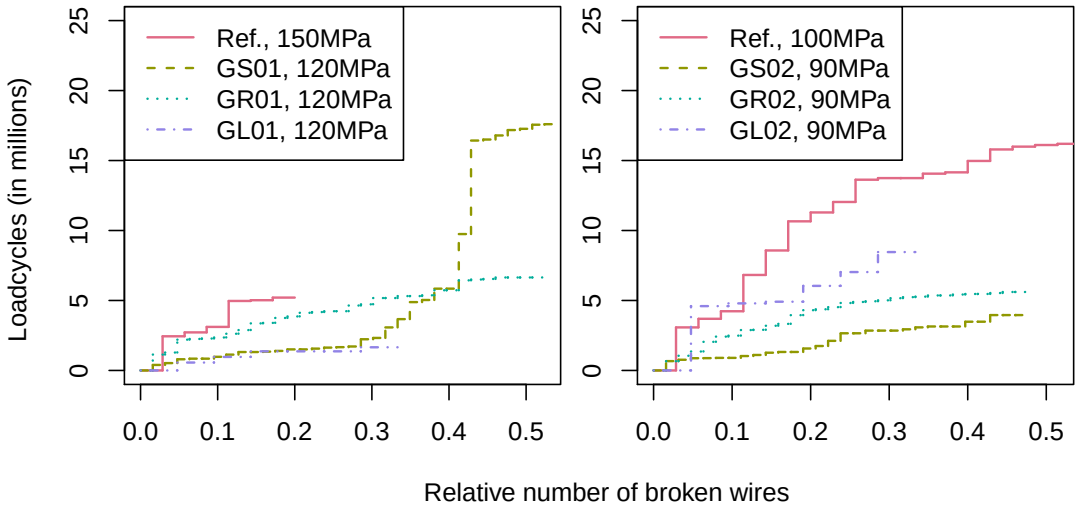

Figure 3: Failure times in different experiments from GS,GR and GL with an initial stress range of 120MPa (left) or 90MPa (right) and an experiment from the reference series with an initial stress range of 150MPa (left) or 100MPa (right).

on a relation between stress and fatigue already proposed by Basquin (1910). Note that the failure rates in this model only depend on the current stress range given by an equal load sharing rule (i.e. the initial stress s_j divided by the relative number of intact wires). However, taking only these features into account seems to be insufficient for the new test series since wires in these experiments tend to break faster than in the reference series; see Figure 3. As already discussed in Section 3.1, this could potentially be caused by the different deflection forces (11) and/or cable factors (12). We therefore introduce several model extensions to include these influence factors and use the methods from Section 2 to determine the most promising models among them.

Remark 4.1. *Experiment GS01 depicted on the left part of Figure 3 shows an extremely atypical behavior: While the time between consecutive wire breaks usually decreases when less wires are intact, the time between breaks seems to increase in this experiment, in particular after around 30% of the wires are broken. This will be very unfortunate for parameter estimations later on since a direct comparison between the GR and GS series would be ideal to consider the effect of deflection forces without taking the cable factor into account.*

For comparison, the following three model expansions will be considered. The first expansion adds the cable factor to the Basquin model by defining new rates with a three-

dimensional parameter $\boldsymbol{\theta} \in \mathbb{R}^3$:

$$\lambda^{\mathrm{cf}}_{\boldsymbol{\theta}}(j,i) = \mathrm{e}^{-\theta_1} \left(\frac{s_j \mathcal{I}_j}{\mathcal{I}_j - (i-1)} \right)^{\theta_2} \kappa^{\theta_3}_{\max,j} \tag{15}$$

in which $\kappa_{\max,j}$ refers to the maximal cable factor $\kappa_{\max}$ in the j-th experiment calculated with (12). Such a model can perform reasonably well when considering only data such as the reference and GR series which have similar deflection forces. When instead considering data such as the GS and GR series, which have the same cable factors but different deflection forces, the following rates are more useful:

$$\lambda^{\mathrm{df}}_{\boldsymbol{\theta}}(j,i) = \mathrm{e}^{-\theta_1} \left(\frac{s_j \mathcal{I}_j}{\mathcal{I}_j - (i-1)} \right)^{\theta_2} u^{\theta_3}_{p,j} \tag{16}$$

in which $u_{p,j}$ refers to the deflection force in the j-th experiment calculated with (11). Finally, the previous two rates can be generalized to the following overall model with a four-dimensional parameter $\boldsymbol{\theta} \in \mathbb{R}^4$:

$$\lambda^{\mathrm{full}}_{\boldsymbol{\theta}}(j,i) = \mathrm{e}^{-\theta_1} \left(\frac{s_j \mathcal{I}_j}{\mathcal{I}_j - (i-1)} \right)^{\theta_2} \kappa^{\theta_3}_{\max,j} u^{\theta_4}_{p,j}. \tag{17}$$

Remark 4.2. *While the rates given by* (17) *are the preferable choice to jointly model all experiment series (Ref., GS, GR, GL), they unfortunately require a four-dimensional model parameter $\boldsymbol{\theta}$. Such a parameter can easily be estimated based on a sufficiently large amount of data. Since all of our experiments combined consist of 17 beams with a total of 272 observed wire breaks, models with higher dimensional parameters automatically bear the risk of overfitting.*

4.1 Ordering the data when applying the 3-sign depth

As already mentioned in Remark 2.2, the power of the 3-sign depth test is affected by the way the data is ordered. While one could use the simple chronological order (i.e. first according to the experiment index j and then according to number i of broken wires), it is advisable to choose an order that tends to lead to large blocks of residuals with the same sign whenever a parameter $\boldsymbol{\theta}$ other than the true model is considered.

The order used in this article will first sort according the cable factor of the experiment and then according to the deflection force, hence:

$$\text{reference data} < GL02 < GL01 < GR02 < GR01 < GS02 < GS01.$$

Finally, the data within each group is ordered according to their current stress range, that is according to

$$(j_1, i_1) < (j_2, i_2) \quad :\Longleftrightarrow \quad \frac{s_{j_1} \mathcal{I}_{j_1}}{\mathcal{I}_{j_1} - (i_1 - 1)} < \frac{s_{j_2} \mathcal{I}_{j_2}}{\mathcal{I}_{j_2} - (i_2 - 1)}.$$

Any remaining ties will be ordered chronologically (hence increasing in the experiment index j). The parameter estimations and model checks in the next section are based on the general approach from Section 2.1 with a process $\mathbf{Z} = (Z_n)_{n=1,\dots,N}$ equal to the sequence $(W_{j,i})_{j,i}$ ordered according to the procedure described above.

5 Data analysis

This section has mainly two objectives: the first one is to identify which of the models in Section 4 are sufficient to describe either all or part of the data from the experiment series mentioned in Section 3.2. The second objective is to use proper models combined with the methods in Section 2.2 to predict the outcome of single experiments. All graphics and simulations in this section are done with the statistics software R (R Core Team, 2020).

5.1 Testing different models with 3-sign depth

In order to study different effects such as the deflection force and cable factor, the following subsets of data will be used:

- **All**: Taking all data (reference, GS, GR, and GL series) into account.
- **RefGR**: Taking data from the reference and GR series. These experiments have similar deflection forces u_p but different cable factors $\kappa_{\max}$. Hence the rates $\lambda^{\mathrm{cf}}_{\boldsymbol{\theta}}(j,i)$ in (15) should yield a sufficient model.
- **GSGR**: Taking data from the GS and GR series. These experiments have the same cable factors $\kappa_{\max}$ but different deflection forces u_p. Hence the rates $\lambda^{\mathrm{df}}_{\boldsymbol{\theta}}(j,i)$ in (16) should yield a sufficient model.
- **Ref**: Taking only data from the reference test series. These experiments have the same cable factors $\kappa_{\max}$ and similar deflection forces u_p. Hence the Basquin model (14) already used in Szugat et al. (2016) and Leckey et al. (2020b) should be sufficient to describe the data.

Moreover, since later breaks, e.g., during the experiment GS01, turned out to be atypically long and thus changed models and predictions significantly, we also consider reduced data where the total number of wire breaks is bounded from above. More precisely, for all data subsets mentioned above, we also consider a reduction of the observations to all breaks up to the point where 25% of the wires are broken. This leads to at most 8 broken wires in each experiment of the reference series, at most 15 broken wires in the GR and GS series, and at most 5 broken wires in the GL series. The corresponding reduced data sets are denoted by $\mathbf{All}^{\mathbf{25\%}}$, $\mathbf{RefGR}^{\mathbf{25\%}}$, $\mathbf{GRGS}^{\mathbf{25\%}}$, and $\mathbf{Ref}^{\mathbf{25\%}}$, respectively.

Table 2: Parameter estimations and p-values according to the 3-sign depth for all models from Section 4. All values are based on all failure times from different subsets of experiments.

		All	RefGR	GRGS	Ref
λ^{full}	$\widehat{\boldsymbol{\theta}}^{\text{SD}}$	$\begin{pmatrix}26\\1.9\\2.9\\0.1\end{pmatrix}$	$\begin{pmatrix}66\\1.7\\3.9\\3.8\end{pmatrix}$	$\begin{pmatrix}12\\0.04\\0.3\\0\end{pmatrix}$	$\begin{pmatrix}184\\0.2\\4.9\\15\end{pmatrix}$
	p-value	0.982	0.994	0.985	0.988
λ^{cf}	$\widehat{\boldsymbol{\theta}}^{\text{SD}}$	$\begin{pmatrix}25\\2\\2.8\end{pmatrix}$	$\begin{pmatrix}26\\2.2\\3\end{pmatrix}$	$\begin{pmatrix}12\\0.04\\0.04\end{pmatrix}$	$\begin{pmatrix}24\\2.6\\-3.1\end{pmatrix}$
	p-value	0.978	0.987	0.985	0.819
λ^{df}	$\widehat{\boldsymbol{\theta}}^{\text{SD}}$	$\begin{pmatrix}74\\2.3\\4.5\end{pmatrix}$	$\begin{pmatrix}3.2\\4.4\\-2.9\end{pmatrix}$	$\begin{pmatrix}9\\0.1\\-0.3\end{pmatrix}$	$\begin{pmatrix}180\\0.2\\15\end{pmatrix}$
	p-value	0.067	0.067	0.985	0.988
λ^{Basq}	$\widehat{\boldsymbol{\theta}}^{\text{SD}}$	$\begin{pmatrix}63\\10\end{pmatrix}$	$\begin{pmatrix}33\\3.9\end{pmatrix}$	$\begin{pmatrix}12\\0.04\end{pmatrix}$	$\begin{pmatrix}26\\2.6\end{pmatrix}$
	p-value	0.008	0.044	0.985	0.819

Table 2 contains a summary of parameter estimations and p-values in the different subsets of data and with respect to the different models introduced in Section 4. All estimations and p-values are based on the 3-sign depth described at the end of Section 2.1. In particular, the test is only focused on whether the average behavior of wire breaks is modeled correctly without assuming that the interarrival times are exponentially distributed. Note that the full model given by λ^{full} in (17) yields very high p-values above 0.98 for all data subsets. Hence, this model yields a very good fit through the data which, however, could also indicate an overfitting since the 3-sign depths became atypically large. Surprisingly, all models yield essentially the same results when considering the data from the GR and GS series only. In particular, the model λ^{df} in (16) that includes the deflection force could not be used to properly describe the difference between the GR and GS series as the matching high p-values of all models indicate. This is probably caused by the atypical behavior of GS01 mentioned in Remark 4.1. As expected, the different cable factors in the reference and GR series can only be modeled properly by λ^{full} and λ^{cf}, whereas all models are capable of describing the reference series only. Finally, it turns out that the rates λ^{cf} in (15) are actually sufficient to model the entire data.

Table 3 contains results similar to Table 2 when only considering data until at most 25% of all wires are broken in each experiment. With this restriction, the most atypical part of GS01 is removed and therefore λ^{df} at least yields an improvement to the simple

Table 3: Parameter estimations and p-values according to the 3-sign depth for all models from Section 4. All values are based on failure times until at most 25% of the wires broke in different subsets of experiments.

		$\text{All}^{25\%}$	$\text{RefGR}^{25\%}$	$\text{GRGS}^{25\%}$	$\text{Ref}^{25\%}$
λ^{full}	$\widehat{\boldsymbol{\theta}}^{\text{SD}}$	$\begin{pmatrix} 42 \\ 1.3 \\ 2.4 \\ 1.8 \end{pmatrix}$	$\begin{pmatrix} 169 \\ 0.4 \\ 5.4 \\ 13.7 \end{pmatrix}$	$\begin{pmatrix} 56 \\ 1 \\ 14.7 \\ 1.8 \end{pmatrix}$	$\begin{pmatrix} 182 \\ -0.05 \\ -3.3 \\ 15.7 \end{pmatrix}$
	p-value	0.997	0.997	0.966	0.975
λ^{cf}	$\widehat{\boldsymbol{\theta}}^{\text{SD}}$	$\begin{pmatrix} 29.5 \\ 2.7 \\ 3.4 \end{pmatrix}$	$\begin{pmatrix} 27.5 \\ 2.3 \\ 3.1 \end{pmatrix}$	$\begin{pmatrix} 38.3 \\ 5.4 \\ 0.4 \end{pmatrix}$	$\begin{pmatrix} 25.8 \\ 2.4 \\ 0.2 \end{pmatrix}$
	p-value	0.983	0.995	0.839	0.946
λ^{df}	$\widehat{\boldsymbol{\theta}}^{\text{SD}}$	$\begin{pmatrix} 195 \\ 0.03 \\ 16.6 \end{pmatrix}$	$\begin{pmatrix} 0.9 \\ 4.8 \\ -3.4 \end{pmatrix}$	$\begin{pmatrix} 36.1 \\ 0.9 \\ 1.8 \end{pmatrix}$	$\begin{pmatrix} 201 \\ -0.3 \\ 17.3 \end{pmatrix}$
	p-value	0.275	0.152	0.966	0.975
λ^{Basq}	$\widehat{\boldsymbol{\theta}}^{\text{SD}}$	$\begin{pmatrix} 114 \\ 20.8 \end{pmatrix}$	$\begin{pmatrix} 34 \\ 4.1 \end{pmatrix}$	$\begin{pmatrix} 35.6 \\ 4.9 \end{pmatrix}$	$\begin{pmatrix} 26.3 \\ 2.5 \end{pmatrix}$
	p-value	0.03	0.091	0.839	0.946

Basquin model λ^{Basq} when considering data from the GR and GS series only. Aside from this improvement, the overall comparison between the different models remains the same: both λ^{full} and λ^{cf} seem to be the best rates to jointly model all data while λ^{full} is likely to actually overfit the data.

5.2 Analysis via prediction bands

With the result from the previous section at hand, we will only include the models λ^{full} from (17) and λ^{cf} from (15) in the upcoming analysis. All estimations are either based on all data or on the reduced data $\text{All}^{25\%}$ which only contains observations until at most 25% of the wires are broken.

Table 4 contains the maximum likelihood estimations obtained when using these different models and data. Note that the estimated value for θ_4 is negative when considering

Table 4: The maximum likelihood estimations $\widehat{\boldsymbol{\theta}}$ based on different models/data.

	All	$\text{All}^{25\%}$
λ^{full}	$(10.79,\ 2.44,\ 2.08,\ -1.51)$	$(42.03,\ 2.63,\ 3.36,\ 1.16)$
λ^{cf}	$(25.15,\ 2.12,\ 1.35)$	$(30.15,\ 2.75,\ 3.62)$

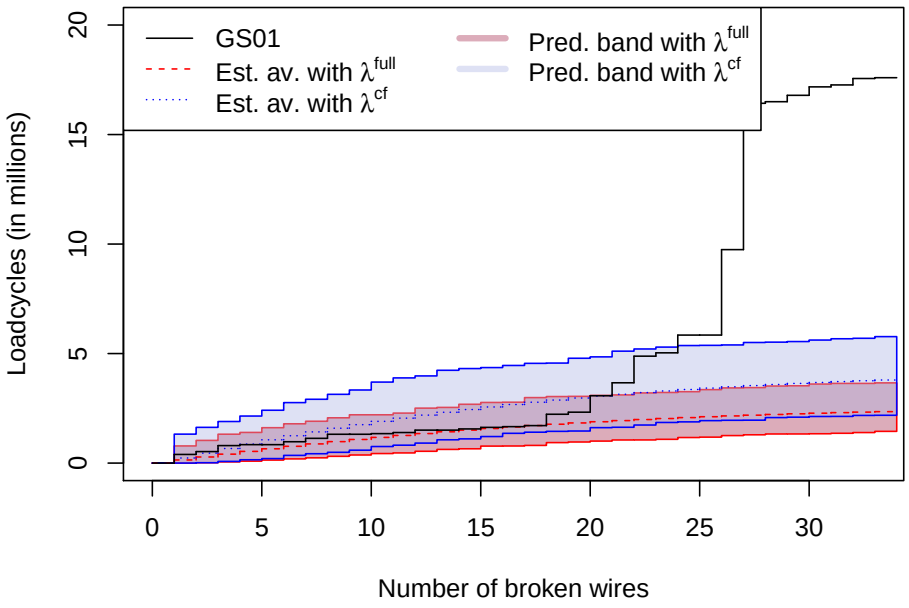

Figure 4: Estimated averages and naive 95%-prediction bands for GS01 based on either model (17) or (15).

the full data set. In fact, when using the LR-approach in (5), the entire 95%-confidence set for $\boldsymbol{\theta}$ contains only parameters with $\theta_4 < 0$. Hence, the results indicate that the deflection force has a negative effect on the rates in the sense that a larger deflection force leads to longer times between consecutive wire breaks. As before, this effect could potentially be caused by the odd behavior of GS01 mentioned in Remark 4.1. In particular, the maximum likelihood estimation is positive in all of its coordinates when only considering the data $\text{All}^{25\%}$ and therefore reflects a more intuitive effect of the deflection force.

In order to compare predictions based on different models and/or methods, we will now conduct a leave-one-out cross-validation. More precisely, for each of the 17 experiments (11 reference and two each from GS, GR, GL), prediction bands are computed based on estimations obtained from the other 16 experiments using either the naive approach (9) or the approach (10) based on the LR-confidence set (5). As mentioned in Remark 2.6, the errors are chosen to be $\alpha = 0.05$ in the naive approach and $\alpha_1 = \alpha_2 = 0.05$ in the LR-confidence set approach. The confidence set is approximated via grid search with mesh widths equal to 0.1 in each dimension.

Figure 4 depicts the naive prediction bands for GS01. Note that the bands do not cover the atypical later breaks in the experiment but at least manage to predict the first 20-25 failures correctly. Surprisingly, the model λ^{cf} with a lower dimensional parameter $\boldsymbol{\theta}$ performed better when predicting GR01 and GL01 as can be seen in Figures 5 and 6. In

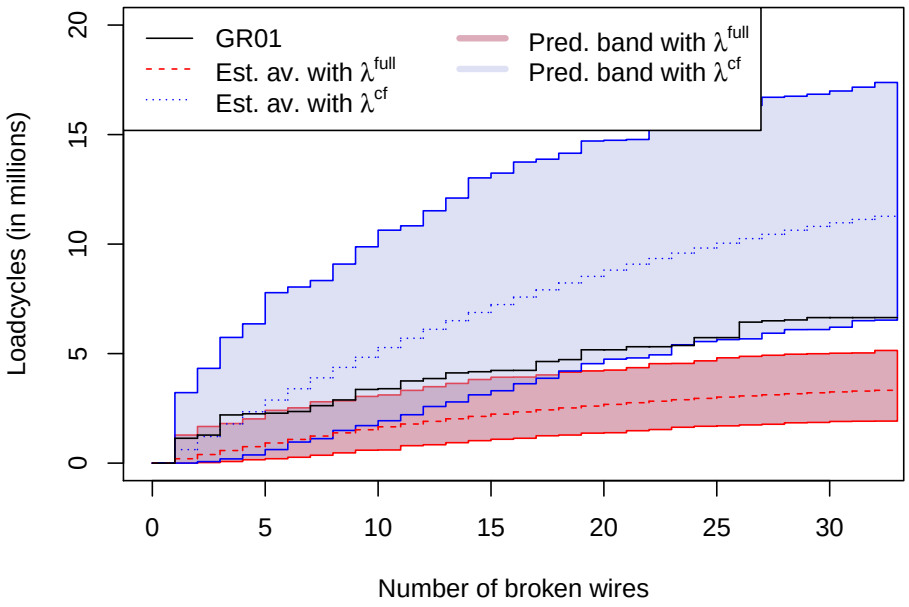

Figure 5: Estimated averages and naive 95%-prediction bands for GR01 based on either model (17) or (15).

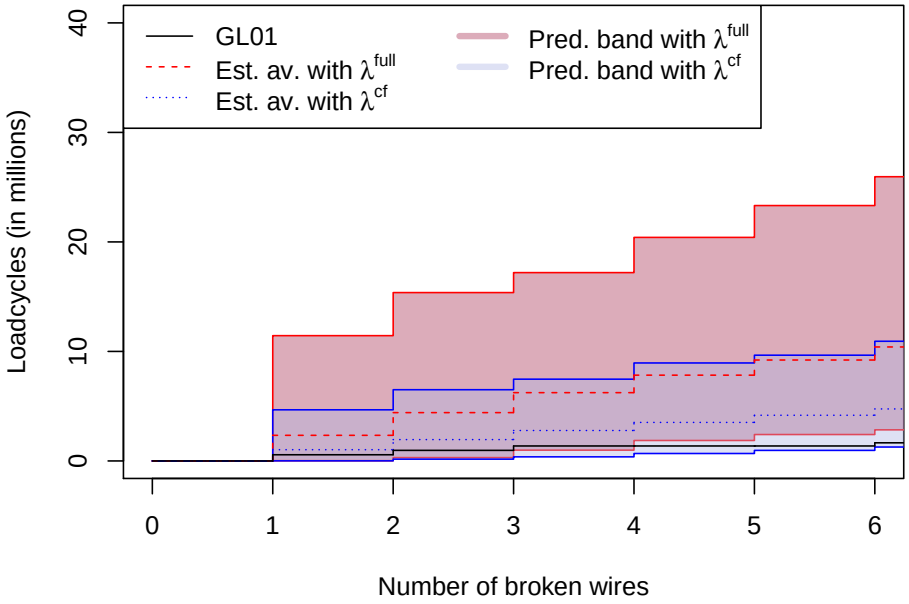

Figure 6: Estimated averages and naive 95%-prediction bands for GL01 based on either model (17) or (15).

Table 5: Coverage rates (in %) of prediction bands in a leave-one-out cross-validation based on the models from (17) and (15).

Method	Model	All	Ref.	GS	GR	GL
Naive	λ^{full}	47.06	63.64	0.00	0.00	50.00
CS-LR	λ^{full}	82.35	90.91	0.00	100.00	100.00
Naive	λ^{cf}	47.06	54.55	0.00	0.00	100.00
CS-LR	λ^{cf}	76.47	90.91	0.00	50.00	100.00
Number of experiments		17	11	2	2	2

Table 6: Coverage rates (in %) of prediction bands in a leave-one-out cross-validation based on the models from (17) and (15) in the reduced data set $\text{All}^{25\%}$.

Method	Model	$\text{All}^{25\%}$	$\text{Ref}^{25\%}$	$\text{GS}^{25\%}$	$\text{GR}^{25\%}$	$\text{GL}^{25\%}$
Naive	λ^{full}	76.47	72.73	100.00	50.00	100.00
CS-LR	λ^{full}	94.12	90.91	100.00	100.00	100.00
Naive	λ^{cf}	70.59	72.73	50.00	50.00	100.00
CS-LR	λ^{cf}	88.24	90.91	100.00	50.00	100.00
Number of experiments		17	11	2	2	2

contrast to the other model, it manages to cover GL01 with its prediction band and also covers all breaks in GR01 except for the 23rd one. On the other hand, model λ^{full} has a slightly higher coverage rate in the reference data; see Table 5 for a summary. This higher coverage rate in the reference series combined with negative θ_4 value in the parameter estimation once again indicates that the extra parameter is used for a better fit (or even overfit) in the reference data rather than a proper modeling of the deflection force itself. In fact, none of the methods was able to produce predictions that cover any of the two GS experiments. Also the overall coverage rate in all 17 experiments is less than 50% for the naive predictions. Since the (asymptotic) coverage rate of the naive methods should be 95%, the results from naive predictions seem a bit low even for the small sample of only 17 experiments[2].

Table 6 contains the coverage rates in the reduce data $\text{All}^{25\%}$. All methods perform significantly better when only considering breaks until at most 25% of all wires are broken. In particular, they now predict either one or both of the GS experiments correctly. Also the confidence set approach based on the four-dimensional model λ^{full} now manages to exceed a 90% coverage rate in all test series. The same approach used with model λ^{cf} still covers almost 90% of the experiments and even the naive approach covers more than 70% of the data. Hence, both models seem to be well suited to predict at least the earlier wire breaks in all test series.

[2]For a proper 95%-prediction band, the coverage rate should have a binomial distribution with $N = 17$ and $p = 0.95$.

6 Conclusion

We have presented a new method for model checking based on sign depth and applied this on component failures in prestressed concrete beams. Since the experiments varied in different parameters such as the number of wires and their diameter, the simple load sharing model from Szugat et al. (2016) was extended by adding extra factors representing the deflection force and cable factor. Then we applied a new simple method to obtain simultaneous prediction bands for the failure times and used them in a leave-one-out cross-validation. This resulted in a second model validation method for the data. The resulting predictions covered the data fairly well, at least up to the point when 25% of all wires are broken. Later breaks were harder to predict, in particular for the GS series with largest deflection force and cable factor that showed unexpectedly long waiting times between these breaks.

Unfortunately, experiments to study the fatigue of prestressed concrete beams are very expensive and time consuming, which is why we only had two experiments for each of the GS, GR and GL series in addition to the 11 reference experiments. In combination with the atypical behavior in the GS01 experiment, this caused our methods to struggle when adding the deflection force to the models, often resulting in negative parameter estimations. The addition of the cable factor, however, significantly improved the simple load sharing model and turned out to be sufficient to properly predict most of the data. It therefore seems like the cable factor has a higher impact on the fatigue than the deflection force but more experiments are required to confirm this hypothesis.

Acknowledgments

The authors gratefully acknowledge support from the Collaborative Research Center "Statistical Modelling of Nonlinear Dynamic Processes" (SFB 823, B5) of the German Research Foundation (DFG).

Data availability statement

The data from all experiments (reference, GR, GS, GL) is publicly available at `https://www.statistik.tu-dortmund.de/1938.html`. The reference series consists of both the TR and SB series.

References

Basquin, O. (1910). The exponential law of endurance tests. *American Society for Testing and Materials Proceedings*, 10:625–630.

Bessa, R. J. (2015). From marginal to simultaneous prediction intervals of wind power. In *18th International Conference on Intelligent System Application to Power Systems (ISAP)*, pages 1–6. IEEE.

Beutner, E. (2010). Nonparametric model checking for k-out-of-n systems. *Journal of statistical planning and inference*, 140(3):626–639.

Chew, V. (1968). Simultaneous prediction intervals. *Technometrics*, 10(2):323–330.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *The Annals of Statistics*, 31(6):1852–1884.

Cramer, E. and Kamps, U. (1996). Sequential order statistics and k-out-of-n systems with sequentially adjusted failure rates. *Annals of the Institute of Statistical Mathematics*, 48(3):535–549.

De Oliveira, V. R. B. and Colosimo, E. A. (2004). Comparison of methods to estimate the time-to-failure distribution in degradation tests. *Quality and Reliability Engineering International*, 20(4):363–373.

Härdle, W. and Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics*, 19(2):778–796.

Heeke, G., Heinrich, J., and Maurer, R. (2019). Neue Erkenntnisse zur Ermüdungsfestigkeit von Spannbeton unter sehr hohen Lastwechselzahlen: Prognose der Lebensdauer. *Beton- und Stahlbetonbau*, 114(4):242–254.

Heinrich, J., Heeke, G., Maurer, R., and Müller, C. H. (2016). Resistance to fatigue and prediction of lifetime of wire tendons cast into concrete up to 10^8 cycles. *19th IABSE Congress Stockholm*, pages 2586–2598.

Hills, D. and Nowell, D. (1994). *Mechanics of Fretting Fatigue*. Springer.

Hong, Y. and Meeker, W. Q. (2013). Field-failure predictions based on failure-time data with dynamic covariate information. *Technometrics*, 55(2):135–149.

Horn, M. (2020). *GSignTest: Robust Tests for Regression-Parameters via Sign Depth*. R package version 1.0.5, `https://github.com/melaniehorn/GSignTest`.

Horn, M. and Müller, C. H. (2020). Tests based on sign depth for multiple regression. *SFB Discussion Paper*, 20(07). `https://www.statistik.tu-dortmund.de/2630.html`.

Hymans, S. H. (1968). Simultaneous confidence intervals in econometric forecasting. *Econometrica: Journal of the Econometric Society*, 36(1):18–30.

Kim, H. and Kvam, P. H. (2004). Reliability estimation based on system data with an unknown load share rule. *Lifetime Data Analysis*, 10(1):83–94.

Kolsrud, D. (2007). Time-simultaneous prediction band for a time series. *Journal of Forecasting*, 26(3):171–188.

Kong, Y. and Ye, Z. (2017). Interval estimation for k-out-of-n load-sharing systems. *IISE Transactions*, 49(3):344–353.

Kustosz, C. P., Leucht, A., and Müller, C. H. (2016). Tests based on simplicial depth for AR(1) models with explosion. *Journal of Time Series Analysis*, 37:763–784.

Leckey, K., Malcherczyk, D., and Müller, C. H. (2020a). Powerful generalized sign tests based on sign depth. *SFB Discussion Paper 12/20*. `https://www.statistik.tu-dortmund.de/2630.html`.

Leckey, K., Müller, C. H., Szugat, S., and Maurer, R. (2020b). Prediction intervals for load-sharing systems in accelerated life testing. *Quality and Reliability Engineering International*, 36(6):1895–1915.

Li, J. S.-H. and Chan, W.-S. (2011). Simultaneous prediction intervals: An application to forecasting US and Canadian mortality. In *Proc. Living 100 Symp.*, pages 1–21.

Malcherczyk, D., Leckey, K., and Müller, C. H. (2021). K-sign depth: From asymptotics to efficient implementation. *Journal of Statistical Planning and Inference*, 215:344–355.

Meeker, W. Q., Escobar, L. A., and Lu, C. J. (1998). Accelerated degradation tests: modeling and analysis. *Technometrics*, 40(2):89–99.

Nishioka, K. and Hirakawa, K. (1969). Fundamental investigations of fretting fatigue:(part 5, the effect of relative slip amplitude). *Bulletin of JSME*, 12(52):692–697.

Ochi, Y., Hayshi, H., Tateno, B., Ishii, A., and Kuroki, T. (1970). Effects of contact pressure and slip amplitude on fretting fatigue properties of rail steel with fish plate. *WIT Transactions on Engineering Sciences*, 13.

Park, C. (2010). Parameter estimation for the reliability of load-sharing systems. *IIE Transactions*, 42(10):753–765.

Patel, J. (1989). Prediction intervals-a review. *Communications in Statistics-Theory and Methods*, 18(7):2393–2465.

Phoenix, S. L. and Newman, W. I. (2009). Time-dependent fiber bundles with local load sharing. II. General Weibull fibers. *Physical Review E*, 80(6):066115.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Remitz, J. and Empelmann, M. (2018). Einfluss von Umlenkbelastungen auf die Ermüdung von Spanngliedern im nachträglichen Verbund. *Beton-und Stahlbetonbau*, 113(8):579–588.

Schervish, M. J. (1995). *Theory of Statistics*. New York, NY: Springer-Verlag.

Schijve, J. (2009). *Fatigue of Structures and Materials*. Springer Netherlands.

Suresh, S. (1998). *Fatigue of Materials*. Cambridge University Press.

Szugat, S., Heinrich, J., Maurer, R., and Müller, C. H. (2016). Prediction intervals for the failure time of prestressed concrete beams. *Advances in Materials Science and Engineering*, 2016. DOI:10.1155/2016/9605450.

Wang, X. and Xu, D. (2010). An inverse gaussian process model for degradation data. *Technometrics*, 52(2):188–197.

Weiher, H. (2007). *Verhalten von PE-HD Schutzhüllen bei der Umlenkung von verbundlosen Spanngliedern*. Ph.D. thesis, TU Munich.

Weiher, H., Specht, E., Pfeiffer, B., Klamroth, K., and Zilch, K. (2008). Determination of the cable factor for deviated tendon bundles. *Structural Engineering International*, 18(1):88–94.

Wollmann, G., Yates, D., Breen, J., and Kreger, M. (1988). Fretting fatigue in post-tensioned concrete. *Research Report 465-2F, Center for Transportation University of Texas at Austin*.

Xiong, C. and Milliken, G. A. (2002). Prediction for exponential lifetimes based on step-stress testing. *Communications in Statistics-Simulation and Computation*, 31(4):539–556.

Xu, J., Hu, Q., Yu, D., and Xie, M. (2017). Reliability demonstration test for load-sharing systems with exponential and Weibull components. *PloS one*, 12(12):e0189863.

Xu, J., Liu, B., and Zhao, X. (2019a). Parameter estimation for load-sharing system subject to Wiener degradation process using the expectation-maximization algorithm. *Quality and Reliability Engineering International*, 35(4):1010–1024.

Xu, J., Xie, M., and Hu, Q. (2019b). Reliability assessment for load-sharing systems with exponential components using statistical expansion as a correction. *Applied Stochastic Models in Business and Industry*, 35(4):998–1010.

Zhang, J., Zhao, Y., and Ma, X. (2020). Reliability modeling methods for load-sharing k-out-of-n system subject to discrete external load. *Reliability Engineering & System Safety*, 193:106603.