R-04-56

Strategy for verification and demonstration of the sealing process for canisters for spent fuel

Christina Müller
Bundesanstalt für Materialforschung und -prüfung (BAM)

Tomas Öberg, Tomas Öberg Konsult AB

August 2004

Svensk Kärnbränslehantering AB
Swedish Nuclear Fuel
and Waste Management Co
Box 5864
SE-102 40 Stockholm Sweden
Tel 08-459 84 00
+46 8 459 84 00
Fax 08-661 57 19
+46 8 661 57 19

ISSN 1402-3091
SKB Rapport R-04-56

Strategy for verification and demonstration of the sealing process for canisters for spent fuel

Christina Müller
Bundesanstalt für Materialforschung und -prüfung (BAM)

Tomas Öberg, Tomas Öberg Konsult AB

August 2004

This report concerns a study which was conducted for SKB. The conclusions and viewpoints presented in the report are those of the authors and do not necessarily coincide with those of the client.

A pdf version of this document can be downloaded from www.skb.se

Summary

Electron beam welding and friction stir welding are the two processes now being considered for sealing copper canisters with Sweden's radioactive waste. This report outlines a strategy for verification and demonstration of the encapsulation process which here is considered to consist of the sealing of the canister by welding followed by quality control of the weld by non-destructive testing.

Statistical methodology provides a firm basis for modern quality technology and design of experiments has been successful part of it. Factorial and fractional factorial designs can be used to evaluate main process factors and their interactions. Response surface methodology with multilevel designs enables further optimisation. Empirical polynomial models can through Taylor series expansions approximate the true underlying relationships sufficiently well. The fitting of response measurements is based on ordinary least squares regression or generalised linear methods.

Unusual events, like failures in the lid welds, are best described with extreme value statistics and the extreme value paradigm give a rationale for extrapolation. Models based on block maxima (the generalised extreme value distribution) and peaks over threshold (the generalised Pareto distribution) are considered. Experiences from other fields of the materials sciences suggest that both of these approaches are useful.

The initial verification experiments of the two welding technologies considered are suggested to proceed by experimental plans that can be accomplished with only four complete lid welds each. Similar experimental arrangements can be used to evaluate process "robustness" and optimisation of the process window.

Two series of twenty demonstration trials each, mimicking assembly-line production, are suggested as a final evaluation before the selection of welding technology. This demonstration is also expected to provide a data base suitable for a baseline estimate of future performance. This estimate can then be used for the safety assessment.

In order to provide data for the analysis of welds it is required to use well defined methods for the investigation of the weld integrity. As the total length of the welds produced during the verification and demonstration exceeds 140 m and these needs to be examined with a reasonable resolution non destructive methods (NDT) with high data acquisition rates are required. Destructive methods like metallographic sectioning and examination or tensile testing can be used only to provide complementary information as the acquisition of these data are labour-intensive and time consuming. As the analysis is relying on the reliability of the NDT it is required to have a strategy for the determination of this. Three different ways to investigate the reliability of NDT signals are described. The performance demonstration, the parameter approach and the Integral Approach: ROC – Receiver Operating Characteristic and its relation to POD (probability of detection). The first step of a performance demonstration is to define the essential technical parameters of the system. The ROC and POD methods are appropriate tools to provide a clear measure of integral performance of the system though it has to be paid by high effort in test series with realistic test samples. With POD the user can learn about the detection capability whereas the ROC gives more information about the system's capability to distinguish between signal and noise. The modular approaches open the door to a promising technique – more efficient and with the capability also to optimize the system.

The data evaluation methods chosen to be applied for this assessment is the quantitative POD (Probability of Detection) method according to MIL-STD 1823 /section B4/.

The (1-POD) curve will provide the probability of missing a defect as function of defect size which can be used as input for probabilistic risk assessment.

Finally, a hypothetical calculation example Appendix 2 is provided to illustrate the fitting of data and extrapolation using the extreme value models.

Contents

Introduction

Svensk Kärnbränslehantering AB (Swedish Nuclear Fuel and Waste Management Co) is responsible for the management and disposal of spent nuclear fuel in Sweden. A RD&D (Research, Development and Demonstration) programme is now being implemented, aiming for final disposal of the radioactive waste in a deep repository.

The aim for this report is to present a strategy for the estimation of the future performance of the encapsulation process consisting of the welding process and quality control process. These estimates will provide input for probabilistic modelling in the overall safety assessment for the deep repository /Hedin, 2003/.

The spent nuclear fuel will be encapsulated in canisters, copper tubes with inserts of cast iron, which are to be sealed by welding. The Canister Laboratory in Oskarshamn is currently evaluating two welding techniques: Electron beam welding (EBW) and friction stir welding (FSW). The goal is to have both the welding processes and the following non destructive testing (NDT) verified and demonstrated in the beginning of 2005. Svensk Kärnbränslehantering AB has assigned Tomas Öberg Konsult AB to develop a general strategy for welding trials and Bundesanstalt für Materialforschung und -prüfung (BAM) for the NDT part.

The report consists of two parts:

Part A. Strategy for verification and demonstration of welding processes
Part B. Strategy for determination of NDT reliability

Part A. Strategy for verification and demonstration of welding processes

The purpose of this part is to outline such a strategy based on a sequential implementation of statistical design of experiments and statistical evaluation of test results.

The investigation and evaluation of the two welding processes shall also provide a best estimate of future performance, with predictions and uncertainty estimates for defects, in an assembly line production.

Statistical methodology is a cornerstone of modern quality technology /Deming, 1986; Snee, 1990; de Mast et al, 2000/. In part this method description is therefore written as a tutorial, since the ambition is to implement the same methodology also in the other parts of the production and manufacturing processes for canisters and inserts.

Part B. Strategy for determination of NDT reliability

All canisters are subjected to non destructive testing (NDT) for the intended aim of quality assurance and production control. NDT reliability is the degree to which the NDT system is capable to reach the intended aim in defect detection and characterization and false calls /1, section B/.

A part of the validation procedure is to demonstrate the acceptable rest risk of the NDT procedures in failing to detect canisters with defects which would violate the requirements for weld integrity. The corresponding parameter matrix for NDT experiments has to be

combined. with the matrix planned for the welding experiments yielding a comprehensive evaluation of the process uncertainties from defect configurations due to welding parameters AND limited NDT sensibility due to NDT process parameters.

The specific tasks to determine for the safety analysis are:

- The risk for failing to detect canisters containing defects in the sealing weld that exceeds the acceptance criteria.
- To define the measurement accuracy of the inspection in terms of defect size and location.

As a complement to the NDT data analysis confirmations by destructive tests will be done. These tests will be done in order to confirm that:

A. The strength and high ductility in copper is valid also in the welds produced in the welding process.

B. The assumption that none-detected defects do not affect the effective corrosion barrier as determined by NDT.

C. The NDT-measurement accuracy is within expected boundaries.

A reasonable number of destructive tests will provide a description of frequency (size distribution), variability and correlation with the NDT methods. A tiered approach is suggested, since the estimates depend both on the population characteristics as well as the sample size. Initially four specimens per weld are suggested to undergo destructive testing, but with low variability this may subsequently be reduced to two per weld. Even with this reduced sampling scheme, results from more than 50 destructive tests will be available for statistical analysis. This will provide a good basis for evaluation and confirmation or falsification of the above statements.

Part A. Strategy for verification and demonstration of welding processes

Tomas Öberg

Abbreviations and acronyms

ANOVA	Analysis of variance
DOE	Design of experiments
EBW	Electron beam welding
FSW	Friction stir welding
GEV	Generalised extreme value distribution
GLM	Generalised linear models
GPD	Generalised Pareto distribution
MLR	Multiple linear regression
NDT	Non-destructive testing
OLS	Ordinary least squares
POD	Probability of detection
SPC	Statistical process control

A1 Statistical methodology

This chapter will outline the suggested statistical methodology for validating, verifying and demonstrating the performance of the two welding processes (EBW and FSW). It is possible, and even likely, that adjustments will be made in the strategy when more experimental results are at hand. This is also in line with the sequential approach employed, i.e. design, experimentation and evaluation will proceed step-by-step to accomplish a continuous improvement /Box, Hunter and Hunter, 1978/ (Figure 1-1).

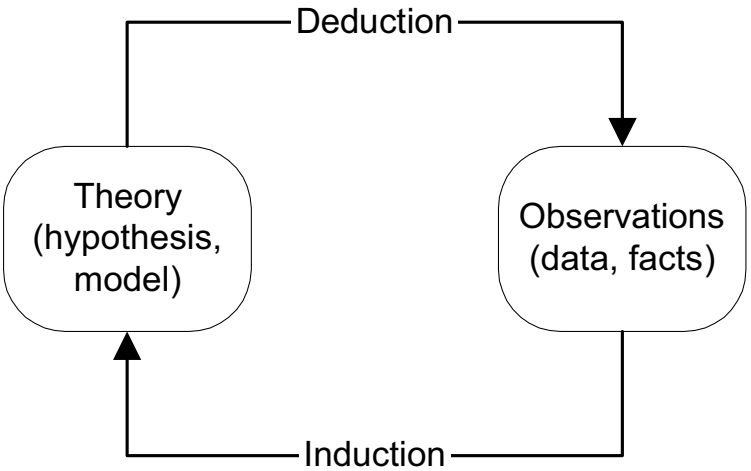

Figure 1-1. *An iterative learning process.*

A1.1 Design of experiments

Design of experiments (DOE) was successfully implemented to improve industrial production processes already in the 1950s /Davies et al, 1956/. Careful planning is essential to allow for further evaluation, statistical analysis and model-building. Statistical design of experiments has developed into an important tool for product and process design, to achieve control and stability, process optimisation and robustness /Montgomery, 1999/. Richards et al and Koleva recently reported results from using multilevel response-surface designs in optimising electron beam welding /Richards et al, 1996; Koleva, 2001/. Experiences gained from other fields of welding are also encouraging /Balasubramanian and Guha, 1999; Gunaraj and Murugan, 1999; Kim et al, 2003; Li, Shiu and Lau, 2003/.

The design of experiments methodology is an essential part in the work to optimise the canister welding technologies and to find robust production conditions, see section A2.2.

A1.1.1 Factorial and fractional factorials

In most systems more than one factor influences the outcome, and often the normal variations show substantial correlations among these factors. A main purpose with experimental design is therefore to break these correlations so that effects of each factor can be estimated independently and without confounding. This has traditionally been solved by employing a "single-factor-at-a time approach", varying only one factor at a time holding the rest constant. This would work in a world without factor interactions, but this is not true for most chemical, physical or technological systems.

Fisher employed a new and ingenious strategy to solve the problem of interactions, introducing factorial experiments /Fisher, 1926/. This approach, first used in agricultural research more than 75 years ago, has subsequently proved its value in most other fields of the experimental sciences.

Factorial experiments are usually carried out at two factor levels and can be used to estimate both linear effects and interactions. This arrangement is among the easiest to learn, but still one of the most versatile approaches in statistical experimental design. A full factorial design is set up so that all possible treatments are investigated: for two factors the number of experimental treatments is four (2^2), for three factors the number of treatments is eight (2^3), and so on. If there are k factors, each at two levels, then a full factorial design has 2^k treatments. A factorial experiment with three factors, A–C, can be graphed as a cube where each treatment is represented as a vertex /Box and Draper, 1969/ (Figure 1-2). The factors have been coded so that the high value is "+1" and the low value is "−1".

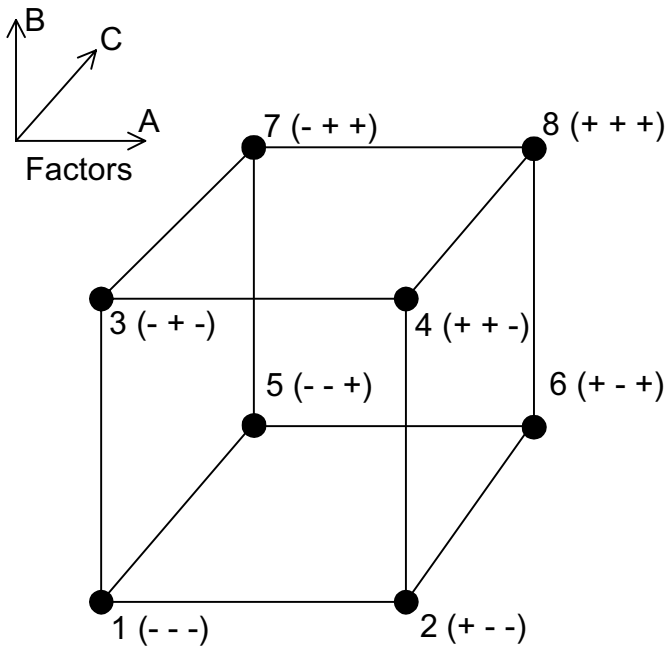

Figure 1-2. *A full factorial experiment, with two levels for three factors.*

It is given in Table 1-1 in tabular form.

Table 1-1. A 2^2 two-level full factorial experiment.

Run	A	B	C
1	−1	−1	−1
2	+1	−1	−1
3	−1	+1	−1
4	+1	+1	−1
5	−1	−1	+1
6	+1	−1	+1
7	−1	+1	+1
8	+1	+1	+1

The main effect of a factor can easily be calculated as the difference of treatments with this factor on the high level (+1) and those at the low level (−1) divided by the number of treatments, i.e. the mean difference.

Full factorials are impractical when the number of factors increases above five (2^5=32 treatments). The solution to the problem is to run a fraction of the suggested treatments, and these designs are consequently called fractional factorials. In the example above it is possible to delete half of the treatments, moving along the diagonals in Figure 1-2, thereby reducing the investigation of the three factors to an experiment consisting of only four runs, i.e. treatments without replication. It is still possible to estimate the main effects, but interactions will be confounded with one another and with the main effects. If one of the main effects is without effect, the fractional factorial will collapse into a full factorial in the remaining two factors /Box, Hunter and Hunter, 1978/ (Figure 1-3).

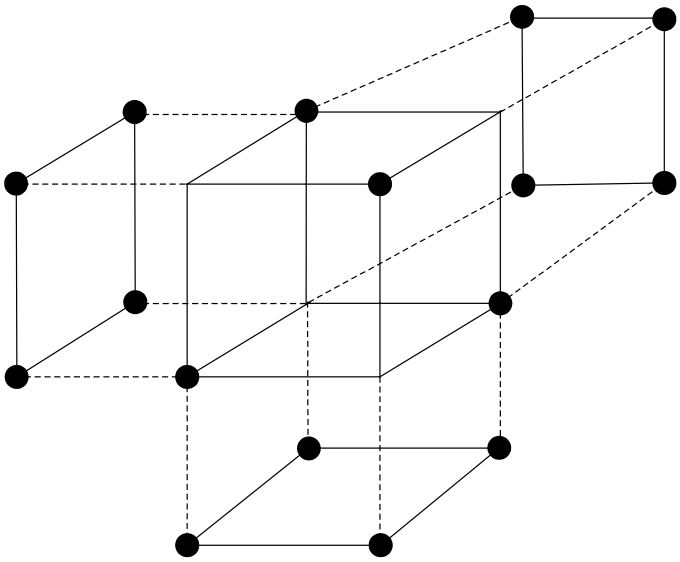

Figure 1-3. *Projections from a fractional factorial in three dimensions to a full factorial in two dimensions.*

The fractional factorial approach to experimentation can easily be extended to any other suitable number of factors and is therefore useful in screening for potentially important effects. The multifactorial experiment plans by Plackett-Burman are similar arrangements, and often used for screening in industrial research and development /Plackett and Burman, 1946/. Factorials and fractional factorials are also suitable arrangements for the evaluation of "robustness"/"ruggedness" of a process and for "robust designing" of products /Hunter, 1985; Phadke, 1989; Bergman et al, 1998/.

A1.1.2 Multilevel response-surface designs

The two-level factorials are limited to describe linear relationships, but it is possible to extend a factorial plan to a multilevel composite design that are capable of describing curvature by assuming quadratic models. The central composite design is such an extension from a factorial or fractional factorial with favourable statistical properties (orthogonality and rotatability) /Box and Wilson, 1951/ (Figure 1-4).

A multilevel design that has proved to be particularly useful in industrial investigations is the Box-Behnken design /Box and Behnken, 1960/. These designs are rotatable or approximately rotatable fractional 3^k designs and, in contrast to the central composite designs, require only three separate levels for each experimental factor (Figure 1-5).

These multilevel designs are also called response surface designs since they allow the fitting and graphing of quadratic polynomials, see Empirical model-building and linear models in the next section.

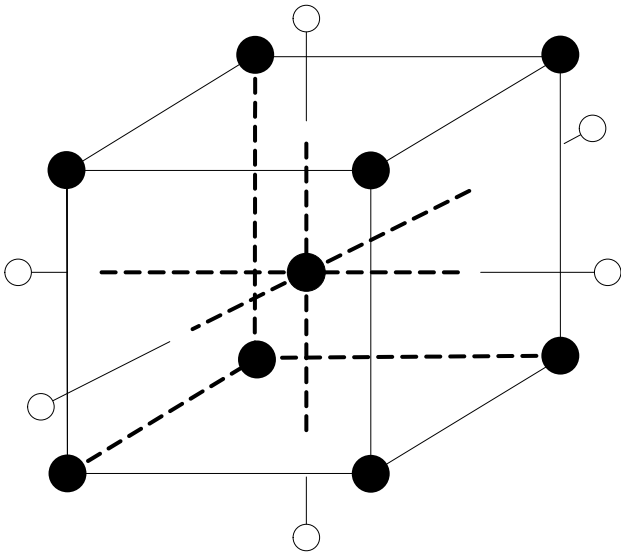

Figure 1-4. *A central composite design for three factors, a full factorial (8 runs) extended with axial (6 runs) and centre points (6 runs).*

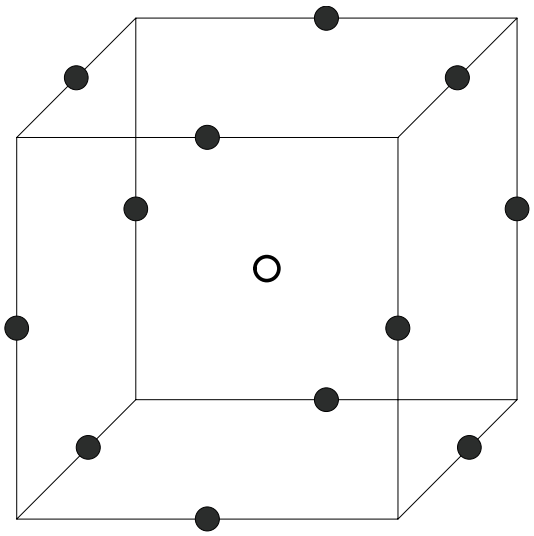

Figure 1-5. *A Box-Behnken design for three factors, with centres of edges (12 runs) and centre points (3 runs).*

A1.1.3 Algorithmic experimental design

The described classical approaches to experimental design all rest on the assumption that the factors can be varied fully independently of one another. In industrial experimentation, the experimental domain is often asymmetrically constrained. One solution to this problem is to shrink the experimental domain. Another approach is to adapt the design both to provide coverage of the irregularly shaped domain and at the same time attain as much mutual independence in the estimated model coefficients as possible. To select the best compromise a criterion is used, the most important being what is now called the D-criterion or D-optimality /Atkinson and Donev, 1992/. A D-optimum design is the choice of candidate points resulting in a model matrix $\mathbf{X}$ that maximises the determinant of $\mathbf{X}^T\mathbf{X}$ or minimises the determinant of its inverse (see section 2.2 below). Computer algorithms are be used to find D-optimal arrangements adapted to the assumed model and domain constraints.

A1.1.4 Blocking and randomisation

A basic assumption in analysing data from designed experiments is that the uncontrolled variation is random. Blocking and randomisation are used to ensure that uncontrolled sources of variation do not influence the interpretation, and both of these ideas are also due to Fisher /Fisher, 1949/.

Changes in environmental factors over time or between different lots of raw material are examples of such uncontrolled systematic variation. Arranging the treatment comparisons in blocks, expected to be more homogenous than the remaining part, may improve the precision of estimates as well as avoid spurious correlations. In a multifactor experiment, blocking is achieved by deliberately confounding the block factor, i.e. time, with higher-order interactions.

However, one must assume that some unknown systematic disturbances can remain, also after blocking. Therefore the run order should be randomised, in order to validate the statistical estimation methods. Randomisation will ensure that undetected systematic variations will not jeopardise the basic assumption of a random error distribution. This leads to the general advice: *"Block what you can and randomise what you cannot"* /Box, Hunter and Hunter, 1978/.

A1.2 Empirical-model building and linear models

In science and technology, theoretical or "first principles models" describe the observed macroscopic variables from possible theoretical mechanisms. In practice, it is often difficult to specify a theoretical model – due to the complexity of the system investigated and a lack of theoretical understanding. This also true for the welding processes considered here.

The current approach is to use empirical models based on polynomials. Polynomials are linear in the parameters, but through Taylor series expansions they can also fit curvature and approximate "reality" sufficiently well. The relationship between empirical models and theoretical models is discussed in various textbooks /Box and Jenkins, 1976; Box, Hunter and Hunter, 1978/.

The following is usually true about polynomial models /Box and Draper, 1987/:

1. The higher the degree of the approximating polynomials, the more closely the Taylor series can approximate the true underlying relationships (within the fitting domain).
2. The smaller the domains over which the approximations are made, the better the approximations will be with polynomials of a fixed degree.

The fitting these polynomials with ordinary least squares (OLS) regression /Draper and Smith, 1981/ and generalised linear models (GLM) /Dobson, 2002/ is described below.

A1.2.1 Linear models

In the most simplistic regression model, there is only one independent variable or factor and one dependent variable or response. Such a model can be described by the equation of a straight line plus error:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \qquad (1\text{-}1)$$

The least squares fitting procedure minimises the vertical squared distances along the y-axis between the regression line and the actual observations, i.e. it minimises the prediction error of y. The coefficients are calculated as:

$$\hat{\beta}_1 = \frac{\mathrm{cov}_{xy}}{s_x^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - x)^2}, \quad \hat{\beta}_0 = y - \hat{\beta}_1 \bar{x} \tag{1-2}$$

The least squares procedure is equal to the maximum likelihood approach for normally distributed data, but the problem is of course to know the underlying distribution.

Linear regression can easily be generalised to fit several independent or explanatory variables. The term for the least squares procedure is multiple linear regression (MLR). The multifactor experiments discussed in the previous section are typical examples of when there is more than one independent variable to consider. A linear model for two independent variables can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{1-3}$$

Such a model can also be extended to describe curvature, the model remaining linear in the parameters:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^{\,2} + \beta_{22} x_2^{\,2} + \beta_{12} x_1 x_2 + \varepsilon \tag{1-4}$$

Geometrically, MLR can be understood as the fitting of a plane or hyperplane to the observation data. In Figure 1-6, this regression surface is shown for two independent variables.

The linear polynomial can be rewritten in more condensed form as a matrix equation:

$$\mathbf{y} = \mathbf{X}\mathbf{b} \tag{1-5}$$

The least squares solution to an overdetermined system is found by first premultiplying by the X-matrix transpose, i.e. the normal equation:

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\mathbf{b} \tag{1-6}$$

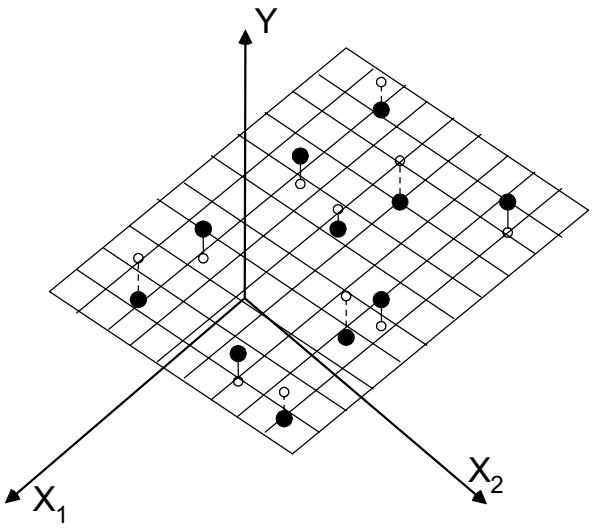

Figure 1-6. *A regression surface.*

Both sides of the expression are then multiplied by the inverse of the variance-covariance matrix:

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\mathbf{b} \qquad (1\text{-}7)$$

The variance-covariance matrix premultiplied by its inverse equals an identity matrix:

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right) = \mathbf{I} \qquad (1\text{-}8)$$

The coefficients estimates are thus:

$$\hat{\mathbf{b}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \qquad (1\text{-}9)$$

This also shows an important characteristic of the MLR method, namely that the variance-covariance matrix $\mathbf{X}^T\mathbf{X}$ must be invertible, i.e. the independent x-variables are truly linearly independent. One purpose of statistical experimental design is to ensure that the data have this characteristic, and MLR is therefore a suitable regression method for designed experiments.

Significance testing of the model parameters is straightforward in MLR, when a normal distribution is assumed. Analysis of variance (ANOVA) is a method of estimating the amount of variation from different sources, e.g. within and among treatments /Sokal and Rohlf, 1973/. The total sum of squares about the mean in a regression model is decomposed into the sum of squares from the model and error:

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 \qquad (1\text{-}10)$$

In Figure 1-7, this decomposition is illustrated graphically (equivalent to the Pythagorean Theorem).

The significance of the model is evaluated by comparing the model sum of squares (divided by its degrees of freedom to give a variance) with the error sum of squares (divided by its degrees of freedom), using the Fisher variance-ratio *F*-test. The model sum of squares can be decomposed further into the contribution from the different factors in the model, and each of these can then be evaluated separately.

Replicates are needed to estimate the experimental or pure error. The error sum of squares can then be further decomposed into the sum of squares due to purely experimental uncertainty and the sum of squares due to lack of fit. The significance of the lack of fit can then again be estimated using the Fisher variance-ratio *F*-test. A significant lack of fit indicates that the model does not adequately describe the influence of the experimental factors, i.e. additional model terms are needed.

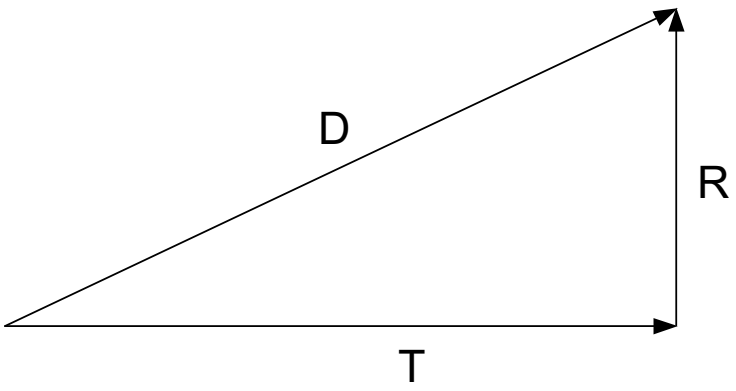

Figure 1-7. *Geometric representation of ANOVA. Deviations from the grand average (D), treatment deviations (T) and residuals (R).*

A1.2.2 Transformations and generalised linear models

Ordinary least squares (OLS) estimation of linear models requires the errors to be normally and independently distributed with constant variance. If an analysis of the residuals reveals systematic deviations, e.g. that the errors are a function of the response variable, then the modelling approach must be modified. If the dependent variable (response) is continuous, then it is often possible to find a variance stabilising transformation, $z = f(y)$. If the standard deviation is proportional to the mean raised to the alpha power, then a suitable power transformation can be found, $z = y^{\lambda}$ where $\lambda = 1 - \alpha$, e.g. square root ($\lambda = 0.5$), log ($\lambda = 0$), inverse square root ($\lambda = -0.5$) and inverse ($\lambda = -1$) /Box and Cox, 1964; Box and Draper, 1987/.

Data for quality characteristics are often expressed as counts, frequencies or proportions. Response variables of this kind do not follow a normal distribution, the variance change with the mean and the effects do not combine additively /Hamada and Nelder, 1997/. Only transforming the observed responses may stabilise variance, but does not necessarily achieve normality and additivity. OLS fitting of linear models will thus often require a more complex model than would otherwise be needed.

The polynomial model can be expressed notationally as:

$$y = \eta + \varepsilon \qquad (1\text{-}11)$$

The distribution of the response variable can be taken into account when fitting the data, by using *generalised linear models* (GLM). With GLM, the error distribution of the response variable is expanded to include a wider class (the exponential family), which also includes the Poisson, binomial and gamma distributions. Instead of trying to find a good transformation for the observation data, one assumes that there exists a transformation called the link function:

$$g(\mu) = \eta = \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k \qquad (1\text{-}12)$$

The choice of link function, to achieve linearity of the systematic effects, is thus separated from the choice of variance function and distribution. The parameters ($\beta_1 \ldots \beta_k$) are estimated iteratively by a maximum likelihood procedure /Hamada and Nelder, 1997; Lewis et al, 2001; Lee and Nelder, 2003/.

Statistical design of experiments has mainly been developed on the basis of normal-theory. For such designs, orthogonal parameter estimates and alias relationships may not hold in a GLM analysis. However, a GLM analysis may still provide useful additional information and should thus be contemplated whenever the normality assumption is in doubt /Myers and Montgomery, 1997; Myers, 1999/.

Additionally, one should observe that the variance is independent of the mean only for normally distributed data. Therefore, replicate runs are not necessary if a non-normal distribution is assumed.

A1.3 Extreme value statistics

The frequently used statistical models and statistics describe usual or ordinary events, e.g. the population mean. However, sometimes the focus is on the unusual events, e.g. the maximum value, and then the ordinary models will be less satisfactory. The unusual events are called extreme values, and extreme value theory has developed to meet the challenges in several fields of engineering, environmental science and finance /Reiss and Thomas, 1997/.

In particular, the estimation of events that are so unusual that they have not yet been observed often require extreme value analyses. This type of extrapolation can of course be put into question, but in many types of applications it will still be necessary to find some rationale for dealing with extreme events. The extreme value paradigm comprises such a rationale for extrapolation, and no other credible alternative has yet been proposed /Coles, 2001/.

The rationale behind the extreme value paradigm is similar to the central limit theorem, but applied to maxima instead of sums. Suppose that $X_1, X_2 \ldots X_n$ is a sequence of independent random variables for an unknown distribution function $F(x)$. Then under some general conditions on the tail of $F(x)$ the normalized maxima, $M_n = max(X_1, X_2 \ldots X_n)$ for large values of n $(n \rightarrow \infty)$, converge to one of the three families of distributions collectively termed "extreme value distributions". These limiting distributions can then be used to make inferences about extreme observations based on finite samples.

The unusual occurrence of large discontinuities in the canister sealing welds is a typical example of a situation where the use of extreme value statistics is appropriate and also the only feasible approach to estimate future process performance.

A1.3.1 The generalised extreme value (GEV) and generalised Pareto distributions (GPD)

Fisher and Tippet showed already in 1928 that there are only three types of distributions, which can arise as limiting distributions of extreme values in random samples /Fisher and Tippett, 1928/. These three types of distributions (Gumbel, Fréchet and Weibull) can be joined together in the "Generalised Extreme Value distribution" (GEV):

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\} \qquad (1\text{-}13)$$

Here $\sigma > 0$, μ and ξ are real parameters, and $z_+ = max(0,z)$. For $\xi = 0$, this should be interpreted as the limit which gives the double exponential or Gumbel distribution.

The GEV family of distributions are suitable for modelling the distribution of maxima in block sequences ("block maxima"). The parameters of GEV (μ, σ and ξ) can be estimated from available observed data with likelihood-based or other technique /Tajvidi, 2004/. Estimates of extreme quantiles are subsequently obtained by inverting Equation (1-14) /Coles and Dixon, 1999/:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1-\{-\log(1-p)\}^{-\xi}\right], & \text{for } \xi \neq 0 \\ \mu - \sigma\log\{-\log(1-p)\}, & \text{for } \xi = 0 \end{cases} \qquad (1\text{-}14)$$

z_p is termed the return level associated with the return period $1/p$, e.g. the level z_p is expected to be exceeded on average once every $1/p$ block of observations /Coles, 2001/.

It could be argued that GEV modelling of maxima is wasteful if complete series of observations are available. The procedure of blocking can be avoided by focusing on excesses of some high threshold u instead of the maxima.

$$y_i = x_i - u \qquad (1\text{-}15)$$

These excesses ("peaks over threshold") can be modelled with the related "Generalised Pareto Distribution" (GPD):

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \qquad (1\text{-}16)$$

Here $\tilde{\sigma} > 0$, $y > 0$ for $\xi \leq 0$ and $0 < y < \tilde{\sigma} / \xi$ for $\xi > 0$. For $\xi = 0$ this should be interpreted as the limit which gives an exponential distribution:

$$H(y) = 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right) \qquad (1\text{-}17)$$

The shape parameter ξ is equal to that of the associated GEV distribution and also the other parameters are connected:

$$\tilde{\sigma} = \sigma + \xi(u - \mu) \qquad (1\text{-}18)$$

If the shape parameter is negative, then there is also an upper limit ($u - \tilde{\sigma} / \xi$) for excesses. The number of excesses can be modelled with a Poisson distribution.

The difference between the GEV and GPD models is thus primarily a matter of how the available data are used. In Figures 1-8 and 1-9 the selection of data from a series of 200 observations, evenly spread across 10 objects, is illustrated for the two modelling approaches. The peaks over threshold GPD model utilise all large observations, but two of the block maxima below the threshold are discarded.

The accuracy of GEV and GPD models are often assessed with probability, quantile and density plots. The estimated return level as a function of the return period can also be graphed together with the confidence intervals (Figure 1-10).

The extreme value models discussed so far all assume an underlying process of independent random variables. Temporal independence is often an unrealistic assumption and the same is of course true for spatial independence. If long-range dependences are weak, this will not influence the GEV family of models for block maxima. However, to avoid invalidating the fitting procedure for the GPD models it may be necessary to adopt declustering (to obtain approximately independent threshold excesses) /Coles, 2001/.

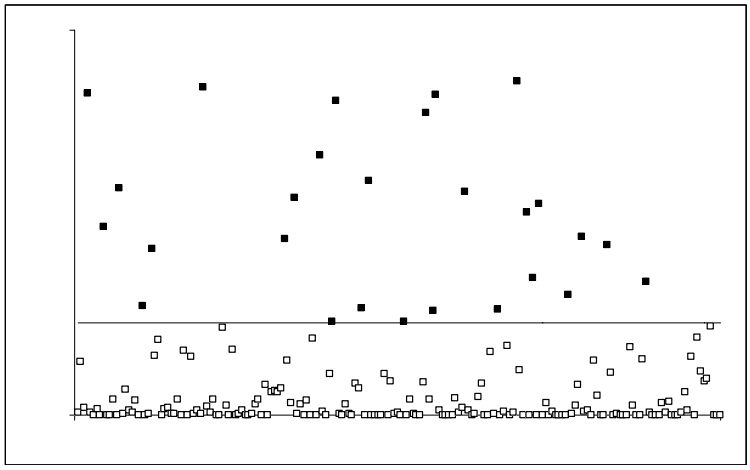

Figure 1-9. *Twenty-seven peaks over threshold marked with filled squares (from 200 observations).*

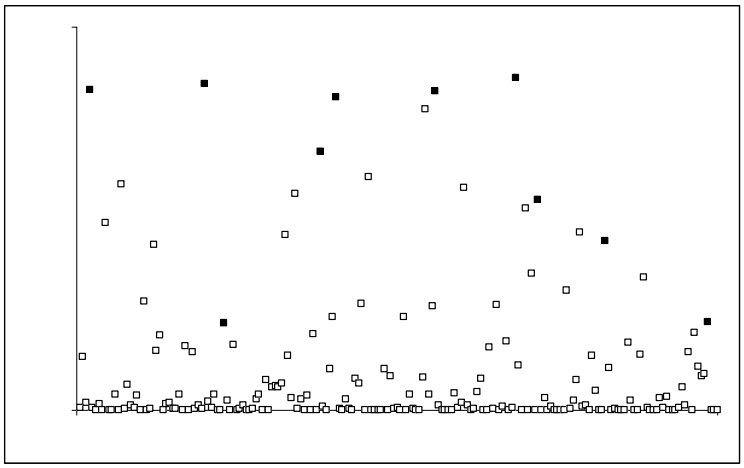

Figure 1-8. *Ten block maxima marked with filled squares (each block with 20 observations).*

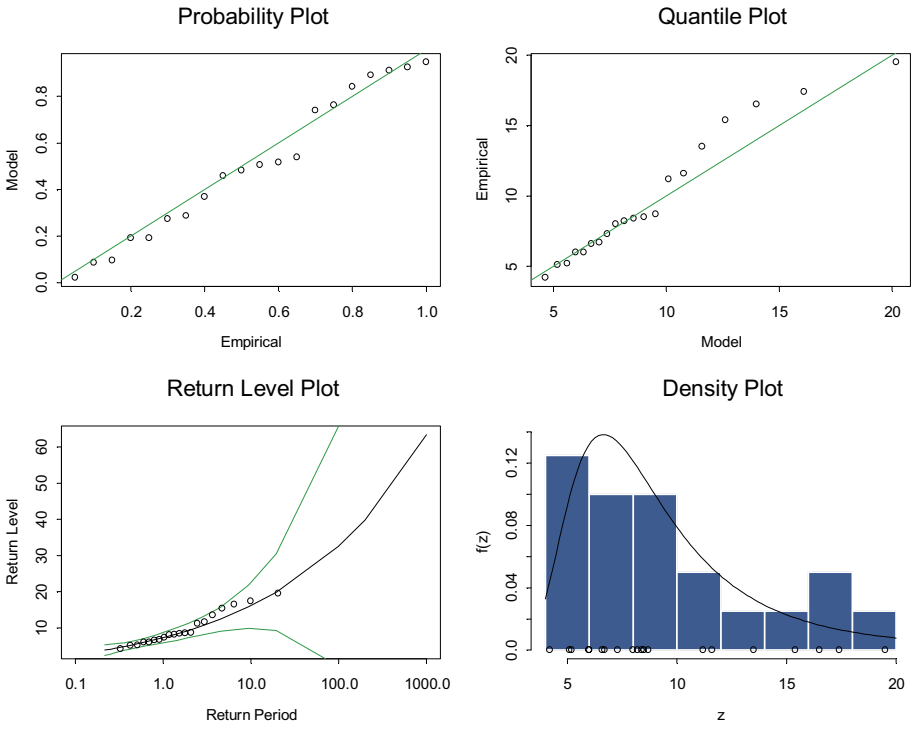

Figure 1-10. *Example of diagnostic plots for a GEV model.*

A1.3.2 Experience with extreme value modelling in material science

The use of extreme value theory is not new to material science. The Weibull distributions, one of the three distributions belonging to the GEV family, has long been used for modelling the breaking strength of materials /Weibull, 1939/, and later also for reliability and lifetime modelling. Similarly, corrosion scientists started to use the Gumbel distributions, also belonging to the GEV family, in the 1950s /Scarf and Laycock, 1996; Zapp, 1996/. GEV type models continue to be used for modelling corrosion extremes /Scarf and Laycock, 1996; Ahari et al, 1997; Vajo et al, 2003/.

Here the focus will be on modelling size and distribution of discontinuities and defects in materials. A substantial number of publications – using extreme value theory to model inclusions and material inhomogeneities – have appeared during the last decade /Atkinson and Shi, 2003/.

Murakami and others have studied the size distribution of defects in steel, inclusions and inhomogeneities, with Gumbel type extreme value distributions /Beretta et al, 1997; Murakami et al, 1998; Atkinson and Shi, 2003/. Murakami's method is based on measuring the maximum size within randomly chosen areas or volumes. This approach permit extrapolation outside of the inspection domain, however, the linear dependence between inclusion size and volume of steel has been put into question /Atkinson and Shi, 2003/. The Gumbel type distributions were also used by Dierickx and co-workers in the modelling of grain size distribution of zirconia /Dierickx et al, 2000/. The model parameters were in both cases estimated with the maximum likelihood method.

Atkinson and co-workers have used the Generalised Pareto Distribution for similar modelling of inclusions in steel /Shi et al, 1999; Anderson et al, 2000; Shi et al, 2001; Yates et al, 2002; Anderson et al, 2003; Atkinson and Shi, 2003/. An interesting property of the GPD is that there is an upper limit when the shape parameter ξ is negative, thus providing means to estimate the maximum inclusion size /Shi et al, 1999; Atkinson and Shi, 2003/. Another argument for using GPD has been the assumption that breaks in specimens are not due to a single inclusion. This dependence on prior assumptions in choosing the reference distribution was observed by Lorén, who also noted that GPD seemed least sensitive to an incorrect distribution assumption /Lorén, 2003/.

Generally, the peaks over thresholds GPD model is robust against nonstationarities in the Poisson process of excesses and it can be used even for weakly dependent time series, where dependence in data can be handled by methods such as declustering. However, it should be noted that the main assumption is that the tail of the distribution can be approximated by a GPD. It is therefore essential for the modelling purpose that any marked deviations from these assumptions are analysed carefully. This is of crucial importance when these models are used for extrapolation, otherwise the results might be misleading.

The distribution assumptions thus have a notable influence on the estimated precision of the predictions, i.e. confidence intervals /Anderson et al, 2000; Anderson et al, 2003; Lorén, 2003/. Assuming a specific distribution from one family will generally give more narrow confidence intervals /Anderson et al, 2000; Anderson et al, 2003/. This may be overoptimistic, if the underlying distribution is unknown. The uncertainty in distribution selection may be substantial /Coles, 2001/. A more conservative approach, applying the full GEV or GPD model, is therefore recommended for future work here.

The choice between a GEV and a GPD model is mainly data driven. Threshold methods, like GPD, utilise more of the available data and will thus give more precise estimates /Anderson et al, 2003/. If only control area maxima is available, then a GEV model would be chosen, but else a GPD model is preferred for making best use of the available data /Anderson et al, 2003/.

In Figure 1-11, adapted from Anderson, Shi et al, both these model choices are illustrated (shape parameter and type of data).

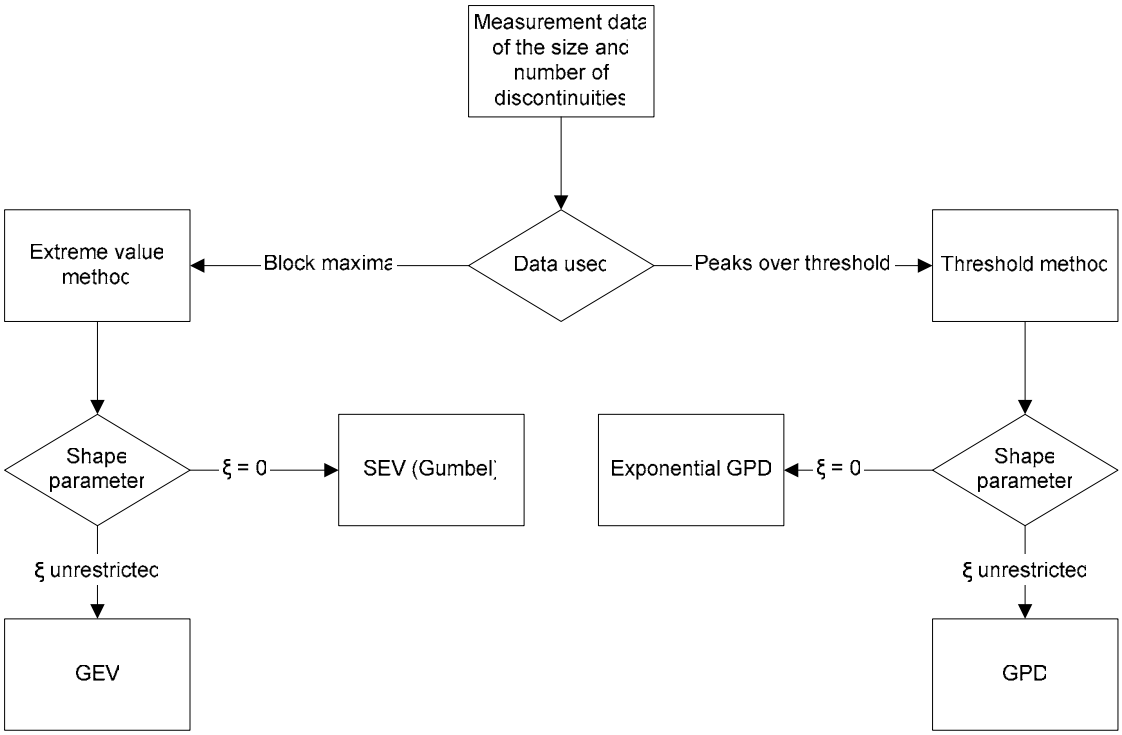

Figure 1-11. *Data and assumptions to choose model.*

A2 Verification of welding processes

Much of the initial work in process and equipment development, for both the EBW and FSW welding techniques, has already been accomplished. It is now needed to better quantify some of the important variables, their interactions and optimal set points for stable operation. It is also required to verify that the processes are stable.

A sequential approach will be used to develop and verify the welding processes (corresponding to the common-sense notion of "not putting all of your eggs in one basket"). The main benefit of this approach is the opportunity to adapt the investigation and the strategy to new findings. However, the experimental effort may to some extent increase compared to a "single-shot" approach. It should also be emphasised that iterative approaches have long played an important role in science and can be employed for continuous quality improvement of production processes /Box and Draper, 1969; Box, 1976; Öberg and Deming, 2000/.

A2.1 Process experiments – design, evaluation and optimization

The outcome for the welding process is measured compared to a number of performance measures, e.g. frequency and size / depth of cavities. The purpose of the first phase is to improve performance and find the optimal parameter settings. Through replication, it is also possible to estimate stability for continuous response variables. However, for response variables expressed as proportions or counts, assuming binomial or Poisson distribution respectively, the variance is a function of the mean making replication less meaningful.

The practical realisation of the first phase has been prepared by:

1. Specifying a list, for each welding technique, of process factors that may have an impact.
2. Specifying a domain to be covered by the process experiments, possible set points, constraints and reasonable number of runs in one experiment.
3. The development of statistical experimental plans to meet the above specification.
4. Review and approval by the project and laboratory management.

The process experiments will be run during spring of 2004, followed by a statistical evaluation. Important response variables are the quality characteristics of the finished weld, analysed with non-destructive testing (NDT), and the performance of the welding machines. If necessary, a follow-up experiment may be added, augmenting the previous experimental runs.

The statistical evaluation will rely on empirical modelling with linear polynomials as described in section 2. Different fitting procedures will be used and compared. The first phase experiment should give an answer to the quantitative importance of each investigated factor on the process outcome for the two welding processes.

A2.1.1 Electron beam welding (EBW)

The number of process factors of interest is more than ten, and this includes qualitative variables such as patterns (with practically unlimited number of settings). The reasonable number of set points (experimental runs) in the first phase investigation is 16–20, with 4–5 lids and four different set points for each lid.

Interactions between different process factors can be assumed to have a major impact and an empirical polynomial model must therefore encompass at least the terms for two factor interactions in addition to the main effects. The number of process factors that can be included is thus limited, if main effects and two-factor interactions should be separated from each other.

A resolution V fractional factorial design with five factors in 16 runs has been selected as a suitable starting-point. This will allow a complete separation of main effects and two-factor interactions, and this design is also easy to extend to include for second-order terms /Box and Draper, 1987/. The prioritised factors selected and a tentative experimental plan is given in Table 2-1, with a randomised run order.

The high and low levels are given as +1 and –1 in the table below. The actual levels for each factor will be specified after initial range finding trials or from previous operating experience. If domain constraints are of importance, then an alternate computer generated optimal design may be considered.

Table 2-1. Experimental plan for EBW process, phase one.

No	Y	Z	Rotation speed	Beam current	Q2 Lens
1	1	1	1	–1	–1
2	–1	–1	–1	–1	1
3	1	1	1	1	1
4	–1	–1	–1	1	–1
5	–1	1	–1	–1	–1
6	1	1	–1	1	–1
7	–1	1	–1	1	1
8	1	–1	1	1	–1
9	1	–1	–1	–1	–1
10	1	–1	–1	1	1
11	–1	–1	1	–1	–1
12	–1	–1	1	1	1
13	–1	1	1	1	–1
14	1	1	–1	–1	1
15	–1	1	1	–1	1
16	1	–1	1	–1	1

A2.1.2 Friction stir welding (FSW)

The number of continuous factors to consider in the friction stir welding process is limited to four. One of these factors, tool angle, can only be changed in between each full lid weld. Eight runs can be performed during each 360° rotation, i.e. 8 × 45°, and two welds can be performed on each lid. The reasonable number of set points (experimental runs) in the first phase investigation is 32, with two lids, two welds on each and eight different set points for each lid.

The constraint regarding change of tool angle will not permit a standard composite design. Instead a Box-Behnken response surface design in the remaining factors will be run twice (for two settings of tool angle). This setup will permit the fitting of a reduced response surface model, separating main effects and interactions in all factors, also describing curvature for the three remaining process factors. In Table 2-2 a tentative experimental plan is given, with changes in tool angle organised in blocks and the other factors in a randomised run order.

The high, intermediary and low levels are given as +1, 0 and –1 in the table below. The actual levels for each factor will be specified after initial range finding trials. These are planned to proceed as a full three level factorial design, investigating three tool angles and three levels of welding speed.

Table 2-2. Experimental plan for FSW process, phase one.

No	Lid	Segment	Tool angle	Welding speed	Tool rotation	Axial force
1	1	0–45°	–1	–1	–1	0
2	1	45–90°	–1	0	1	–1
3	1	90–135°	–1	1	–1	0
4	1	135–180°	–1	–1	1	0
5	1	180–225°	–1	0	1	1
6	1	225–270°	–1	0	0	0
7	1	270–315°	–1	–1	0	–1
8	1	315–360°	–1	1	1	0
9	2	0–45°	1	0	0	0
10	2	45–90°	1	1	0	–1
11	2	90–135°	1	1	0	1
12	2	135–180°	1	–1	–1	0
13	2	180–225°	1	0	0	0
14	2	225–270°	1	–1	0	–1
15	2	270–315°	1	0	–1	1
16	2	315–360°	1	0	0	0
17	3	0–45°	–1	1	0	–1
18	3	45–90°	–1	0	–1	1
19	3	90–135°	–1	0	0	0
20	3	135–180°	–1	1	0	1
21	3	180–225°	–1	0	0	0
22	3	225–270°	–1	–1	0	1
23	3	270–315°	–1	0	0	0
24	3	315–360°	–1	0	–1	–1
25	4	0–45°	1	0	–1	–1
26	4	45–90°	1	–1	1	0
27	4	90–135°	1	1	–1	0
28	4	135–180°	1	1	1	0
29	4	180–225°	1	–1	0	1
30	4	225–270°	1	0	0	0
31	4	270–315°	1	0	1	–1
32	4	315–360°	1	0	1	1

A2.2 Robustness – sensitivity to variations in settings, machinery and environment

It is important to evaluate “robustness”/“ruggedness” in order to establish “process windows” and develop the specifications and tolerances for parameters that are adjustable or controllable. This phase may well be implemented concurrently with the first phase, and the data from these process experiments will together with previous trials form a basis for decision about the specifications. Measurement accuracy and precision of machine controllers should also be accounted for.

The rationale in limiting the investigation is the complexity of the production process. Thirty different factors have already been investigated for the EBW process, e.g. acceleration voltage, beam current, travel speed, focus lens Q1 and Q2, chamber pressure, working distance, gun tilt angle, slope in and out distances and helium flow. Most of these factors interact with each other, and it is not feasible at this stage to include all relevant machine parameters in the statistical process model to verify their influence on robustness.

However, at a later stage it may be of interest to characterise more in detail the influence of potential "nuisance factors" or "noise factors", i.e. external factors that can influence the process outcome. Some examples of nuisance factors to consider are variations in raw material, in environmental factors (e.g. ambient temperature and humidity) and between operators.

An extended ruggedness test could be used to estimate influence of both the environmental factors and the process operating factors. Resolution III fractional factorials or Plackett-Burman designs will limit the number of runs to an absolute minimum, but then it will not be possible to distinguish between main effects and interactions /Plackett and Burman, 1946; ASTM, 2002/.

A3 Demonstration of welding processes

The final development phase, before selection of future welding technology, is a demonstration under conditions similar to the intended assembly-line production. These demonstration tests will provide a basis for estimation of the expected welding quality, and justify the extrapolation to a future welding process.

The proposed scope for this third phase of technology development is to run two series of twenty canisters for each of the two welding processes, EBW and FSW. A primary purpose is to demonstrate that the welding technology development now has reached a level where assembly-line production can meet the design criteria for canister production, as specified in the safety assessment /Hedin, 2003/. The acceptance specification for the seal welds, state that the largest allowed discontinuity is 35 mm in radial extension. The manufacturing and quality control steps should guarantee that no more than 0.1 percent (1/1000) of finished canisters has defects larger than this acceptance criterion.

The two series of welding experiments are also anticipated to supply data for prediction of the expected welding quality from the production plant. Estimation of size and distribution of defects will allow further refinement of the safety assessment. Together with the characterisation of the non-destructive testing (NDT) systems, this could also allow the replacement of the currently assumed probability distribution (binomial with p=0.001) with one derived from empirical data.

A3.1 Process experiments – design and evaluation

The process demonstration experiments will be run in two parallel series for EBW and FSW. Process conditions, set points, will be fixed within the operating specifications. These specifications will be developed from previous experience, results from phase I–II experiments and supplier guidelines for the equipment.

Welding machine operating variables are also variables in evaluating the process stability, but primary response variables are the lid weld quality measured by the NDT system. Defects are characterised by size, geometry and location.

Each of these 2 × 20 experiments consists of a complete canister lid weld. The evaluation will focus on maximum size defects and defects above a selected threshold, modelled either with the generalised extreme value distribution and/or the generalised Pareto distribution. The evaluation report will also illuminate the consequences of these methodological choices. Distribution models will be fitted using a maximum likelihood or other suitable approach. Lack of independence between subsequent experiments, time trends and correlation with environmental factors or/and operating conditions will be given due consideration.

A calculation example is given in Appendix 2.

A3.2 Prediction of future process performance

Experience from industrial manufacturing processes show that performance and quality improve over time. The pace of improvement can be further enhanced by applying statistical process control (SPC).

The future performance of the welding processes is largely unknown, but the demonstration phase experiments will supply data for a baseline estimate. It can be assumed that further development will contribute to an improvement beyond this baseline estimate. However, drawing conclusions from a series of twenty experiments to the outcome of 4500 canister lid welds is a substantial extrapolation. Nevertheless this extrapolation is the only available rationale for estimating future performance and extreme value theory is judged to provide the best available approach.

The extreme value models developed from the results obtained in the 2 × 20 demonstration experiments will subsequently be used to estimate the maximum discontinuity size as a function of number of lid welds, together with an estimate of the confidence in this estimate. Since the details of the underlying distribution are unknown, the generalised extreme value (GEV) and Pareto (GPD) distributions are selected for statistical modelling. The selection of a more specific extreme value distribution could give more narrow confidence bands for the prediction estimates, but it can not be ruled out that a thus simplified model is overoptimistic. The selected conservative approach therefore seems more appropriate at this stage, but will need further verification as soon as measurement data become available.

The calculation example in Appendix 2 is extended with predictions and uncertainty estimates.

A4 Implicatios for SR-Can

The final failure ratio will also depend on the performance of the non-destructive testing methods, described in part B. The extreme value models predicting future welding performance must therefore be combined with the NDT model for estimating the overall process performance.

The figures in Appendix 2 show the cumulative probability of defects below a given size, i.e. $F(x)=P(X\leq x)$. The probability of discontinuities above a given size $P(X>x)$ is then $1-F(x)$ (Figure 4-1).

The probability of not detecting a discontinuity above a given size is in an analogous manner defined from the probability of detection (POD), see part B and Appendix 1 for further details. Figure 4-2 show a hypothetical 1-POD curve.

If we can assume independence between F(x) and POD, then the probability of discontinuities not detected is the product between the two functions specified above:

$$[1-\mathrm{POD}]\cdot[1-F(x)] \qquad (4\text{-}1)$$

The design criteria for the canister production process states that not more than 10^{-3} times the number of deposited canister should have a remaining wall thickness less than 15 mm. The largest allowed discontinuity size is then 35 mm in radial extension (50 mm original wall thickness) and the probability is calculated as outlined above. The probability of a remaining wall thickness $y = 50 - x$ is then:

$$1-[1-\mathrm{POD}]\cdot[1-F(x)] \qquad (4\text{-}2)$$

Both distributions, F(x) and POD, are to be estimated from experimental data and thus have uncertainties in the defining parameters. F(x) is likely to be an extreme value distribution with three parameters μ, σ and ξ (see A1.3.1). POD is a lognormal distribution with parameters μ and σ:

$$\mathrm{POD}(x)=\Phi\left(\frac{\ln x-\mu}{\sigma}\right) \qquad (4\text{-}3)$$

where Φ is the standard normal distribution function.

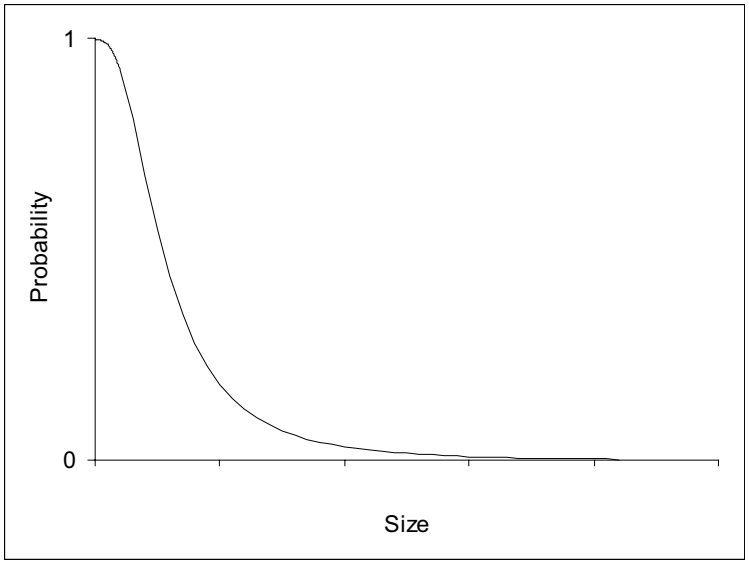

Figure 4-1. *Probability of discontinuities above a given size.*

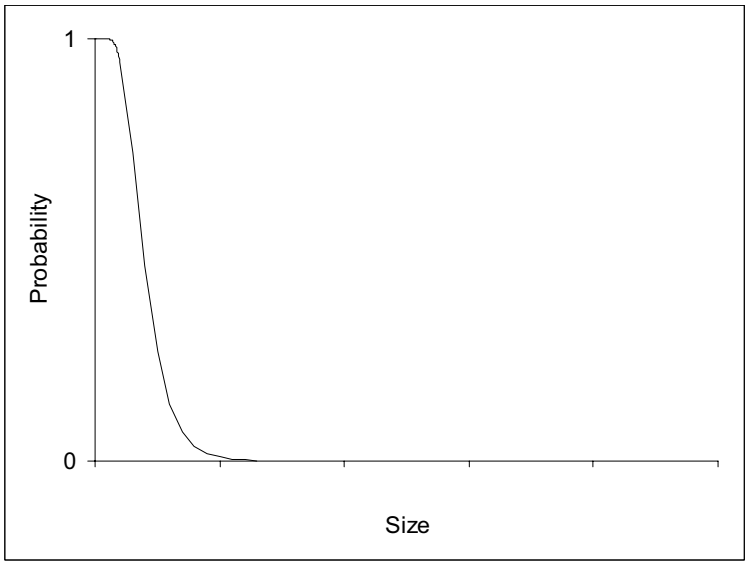

Figure 4-2. *Probability of not detecting discontinuities above a given size.*

To estimate the total variability, these parameter uncertainties must also be accounted for. The uncertainties in parameter estimates for POD can be assumed to be normally distributed while the uncertainty of the extreme value parameters are better estimated by profile likelihood, see Appendix 2 for further details.

It may be possible to develop an analytical solution for estimating the probability for a remaining critical wall thickness equal to or above 15 mm, but a Monte Carlo-simulation will else provide a convenient method to develop an empirical probability distribution based on the assumptions given and the estimated parameter uncertainties.

Several issues cannot be resolved until data from the verification tests are available, notably the questions regarding assumed independence and the choice of statistical models. Several practical issues also remain such as the rejection criteria and potential conflicts between the requirements for production rate versus the overall safety.

A5 Conclusions

A strategy has been proposed for a systematic development of the welding processes. Methods and software are available to accomplish the tasks of verification and demonstration.

The specific choice of approach to build empirical process models will have to wait until data are available from these process experiments. However, it is anticipated that polynomials can be fit to the data using ordinary least squares (OLS) or generalised linear models (GLM).

The uncertainty in extrapolation of results from the demonstration trials to the full-scale assembly line production is substantial. The focus is here on the tail-end statistics and will to rely on extreme value theory. The extreme value approach provides a scientific rationale for this extrapolation.

Models based on the generalised extreme value (GEV) and generalised Pareto distributions (GPD) have been successfully used in other similar applications. The methods for fitting and estimating uncertainties are under continuous development and the current best practice is reviewed. The final choices of models and methods will also here have to wait until measurement data are available.

The final estimation of failure ratios will depend on the performance of the canister production, the welding technology and the non-destructive testing. This combined probability can be evaluated by Monte Carlo simulations from the estimated theoretical distributions.

A6 References

Ahari K G, Coleys, K S Nicholls J R, 1997. Statistical evaluation of corrosion of sialon in burner rig simulated combustion atmospheres. Journal of the European Ceramic Society 17, 681–688.

Anderson C W, Shi G, Atkinson H V, Sellars C M, 2000. The precision of methods using the statistics of extremes for the estimation of the maximum size of inclusions in clean steels. Acta Materialia 48, 4235–4246.

Anderson C W, Shi G, Atkinson H V, Sellars C M, Yates J R, 2003. Interrelationship between statistical methods for estimating the size of the maximum inclusion in clean steels. Acta Materialia 51, 2331–2343.

ASTM, 2002. Standard guide for conducting ruggedness tests. Standard E1169-02.

Atkinson A C, Donev A N, 1992. Optimum experimental designs. Oxford University Press, Oxford, UK.

Atkinson H V, Shi G, 2003. Characterization of inclusions in clean steels: a review including the statistics of extremes methods. Progress in Materials Science 48, 457–520.

Balasubramanian V, Guha B, 1999. Assessment of some factors affecting fatigue endurance of welded cruciform joints using statistical techniques. International Journal of Fatigue 21, 873–877.

Beretta S, Blarasin A, Endo M, Giunti T, Murakami Y, 1997. Defect tolerant design of automotive components. International Journal of Fatigue 19, 319–333.

Bergman R, Johansson M E, Lundstedt T, Seifert E, Aberg J, 1998. Optimization of a granulation and tabletting process by sequential design and multivariate analysis. Chemometr. Intell. Lab. 44, 271–286.

Box G E P, 1976. Science and statistics. J. Am. Stat. Assoc. 71, 791–799.

Box G E P, Behnken D W, 1960. Some new three level designs for the study of quantitative variables. Technometrics 2, 455–475.

Box G E P, Cox D R, 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B-Stat. Methodol. 26, 211–243.

Box G E P, Draper N R, 1969. Evolutionary operation; a statistical method for process improvement. Wiley, New York, USA.

Box G E P, Draper N R,1987. Empirical model-building and response surfaces. Wiley, New York, USA.

Box G E P, Hunter W G, Hunter J S, 1978. Statistics for experimenters: An introduction to design, data analysis, and model building. Wiley, New York, USA.

Box G E P, Jenkins G M, 1976. Time series analysis: Forecasting and control. Holden-Day, San Francisco, USA.

Box G E P, Wilson K B, 1951. On the experimental attainment of optimum conditions. J. R. Stat. Soc. Ser. B-Stat. Methodol. 13, 1–45.

Coles S, 2001. An introduction to statistical modeling of extreme values. Springer, London ; New York.

Coles S G, Dixon M J, 1999. Likelihood-based inference for extreme value models. Extremes 2, 5–23.

Davies O L, Box G E P, Connor L R, Cousins W R, Himsworth F R, G. P. Sillitto G P, 1956. The design and analysis of industrial experiments. Published for Imperial Chemical Industries by Oliver and Boyd, Edinburgh, UK.

de Mast J, Schippers W A J, Does R J M M, van den Heuvel E R, 2000. Steps and strategies in process improvement. Quality and Reliability Engineering International 16, 301–311.

Deming W E, 1986. Out of the crisis. Massachusetts Institute of Technology, Center for Advanced Engineering Study, Cambridge, Mass.

Dierickx D, Basu B, Vleugels J, Van der Biest O, 2000. Statistical extreme value modeling of particle size distributions: experimental grain size distribution type estimation and parameterization of sintered zirconia. Materials Characterization 45, 61–70.

Dobson A J, 2002. An introduction to generalized linear models. Chapman & Hall/CRC, Boca Raton.

Draper N R, Smith H, 1981. Applied regression analysis. Wiley, New York, USA.

Fisher R A, 1926. The arrangement of field experiments. J. Min. Agric. 33, 503–515.

Fisher R A, 1949. The design of experiments. Oliver and Boyd, Edinburgh, UK.

Fisher R A, Tippett L H C, 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proceedings of the Cambridge Philosophical Society 24, 189–190.

Gunaraj V, Murugan N, 1999. Application of response surface methodology for predicting weld bead quality in submerged are welding of pipes. Journal of Materials Processing Technology 88, 266–275.

Hamada M, Nelder J A, 1997. Generalized linear models for quality improvement experiments. J. Qual. Technol. 29, 292–305.

Hedin A, 2003. Planning report for the safety assessment SR-Can. Svensk Kärnbränslehantering AB, Stockholm.

Hunter J S, 1985. Statistical design applied to product design. J. Qual. Technol. 17, 210–221.

Kim I, Son K, Yang Y, Yaragada P, 2003. Sensitivity analysis for process parameters in GMA welding processes using a factorial design method. International Journal of Machine Tools & Manufacture 43, 763–769.

Koleva E, 2001. Statistical modelling and computer programs for optimisation of the electron beam welding of stainless steel. Vacuum 62, 151–157.

Lee Y, Nelder J A, 2003. Robust design via generalized linear models. J. Qual. Technol. 35, 2–12.

Lewis S L, Montgomery D C, Myers R H, 2001. Examples of designed experiments with nonnormal responses. J. Qual. Technol. 33, 265–278.

Li B, Shiu B, Lau K, 2003. Robust fixture configuration design for sheet metal assembly with laser welding. Journal of Manufacturing Science and Engineering 125, 120–127.

Lorén S, 2003. Estimating inclusion distributions of hard metal using fatigue tests. International Journal of Fatigue 25, 129–137.

Montgomery D C, 1999. Experimental design for product and process design and development. Journal of the Royal Statistical Society Series D 48, 159–177.

Murakami Y, Takada M, Toriyama T, 1998. Super-long life tension-compression fatigue properties of quenched and tempered 0.46% carbon steel. International Journal of Fatigue 20, 661–667.

Myers R H, 1999. Response surface methodology – Current status and future directions. J. Qual. Technol. 31, 30–44.

Myers R H, Montgomery D C, 1997. A tutorial on generalized linear models. J. Qual. Technol. 29, 274–291.

Phadke M S, 1989. Quality engineering using robust design. Prentice Hall, Englewood Cliffs, N.J.

Plackett R L, Burman J P, 1946. The design of optimum multifactorial experiments. Biometrika 33, 305–325.

Reiss R D, Thomas M, 1997. Statistical analysis of extreme values : from insurance, finance, hydrology and other fields. Birkhèauser, Basel ; Boston.

Richards N, Chaturvedi M , Liu Y, Mount K, 1996. Optimization of electron beam welding parameters for Incoloy 903. International Journal of Materials & Product Technology 11, 284–300.

Ronneteg U, 2001. OFP – glappanalys 1. Svensk Kärnbränslehantering AB.

Scarf P A, Laycock P J, 1996. Estimation of extremes in corrosion engineering. Journal of Applied Statistics 23, 621–643.

Shi G, Atkinson H V, Sellars C M, Anderson C W, 1999. Application of the Generalized Pareto Distribution to the estimation of the size of the maximum inclusion in clean steels. Acta Materialia 47, 1455–1468.

Shi G, Atkinson H V, Sellars C M, Anderson C W, Yates J R, 2001. Computer simulation of the estimation of the maximum inclusion size in clean steels by the generalized Pareto distribution method. Acta Materialia 49, 1813–1820.

Snee R D, 1990. Statistical thinking and its contribution to total quality. Am. Statistician 44, 116–121.

Sokal R R, Rohlf F J, 1973. Introduction to biostatistics. W. H. Freeman, San Francisco, USA.

Tajvidi N, 2004. Confidence intervals and accuracy estimation for heavytailed generalized Pareto distributions. Extremes in press.

Vajo J J, Wei R, Phelps A C, Reiner L, Herrera G A, Cervantes O, Gidanian D, Bavarian B, Kappes CM, 2003. Application of extreme value analysis to crevice corrosion. Corrosion Science 45, 497–509.

Weibull W, 1939. A statistical theory of the strength of materials. Ingeniörsvetenskapsakade miens Handlingar 151, 1–45.

Yates J, Shi G, Atkinson H, Sellars C, Anderson C, 2002. Fatigue tolerant design of steel components based on the size of large inclusions. Fatigue & Fracture of Engineering Materials & Structures 25, 667–676.

Zapp P E, 1996. Pitting growth rate in carbon steel exposed to simulated radioactive waste. Westinghouse Savannah River Company. 23.

Öberg T, Deming S N, 2000. Find optimum operating conditions fast. Chem. Eng. Prog. 96, 53–59.

Part B. Strategy for determination of NDT reliability

Christina Müller BAM

Abbreviations and acronyms

NDT	Non Destructive Testing
CL	Canister Laboratory Oskarshamn
EBW	Electron Beam Welding
FSW	Friction Steer Welding
NDE	Non Destructive Evaluation
DIN	Deutsches Normungsinstitut
ASTM	American Society for Testing and Materials
PISC	Program for Inspection of Steel Components
EPRI	Electric Power Research Institute, Charlotte, North Carolina
PDI	Performance Demonstration Initiative
ASME	American Society for Mechanical Engineering
ROC	Receiver Operating Characteristic
POD	Probability of Detection
ENSIP	Engine Structural Integrity Program
ENIQ	European Network for Inspection Qualification
TP	True Positive Indicatio
FN	False Negative Indication
FP	False Positive Indication
TN	True Negative Indication
R	Total Reliability
IC	Intrinsic Capability
AP	Application Factors
HF	Human Factor
P(TP)	Probability of a true positive indication
P(FP)	Probability of a false positive indication
adec	signal threshold
SH	Screen Hight
PODAP	
CT	Computerised Tomography
RT	Radiagraphic Technique
UT	Ultrasonic Technique
a	defect dimension
â	signal height
a90/95	a at 90% level of the 95% confidence limit

B1 Non destructive methods

The NDT systems that shall be analyzed are the X-ray system, the ultrasonic system at CL. The result of using the two inspection systems in combination shall also be analyzed.

The inspection tasks to study are inspection of sealing welds produced using EBW and FSW.

The FSW is a less mature process compared to EBW. Inspection procedures for the FSW process is expected to be defined during 2003 and data could be available for analyze at the beginning of 2004.

The X-ray inspection system at CL consists of a 9 MeV Varian accelerator with a16 bits AD and viewing system provided by BIR. The ultrasonic system is a phaseded array 8 bit system however recently upgraded to a 10 bit system. The inspection of the EB-weld is carried out using an 80 element array working at a center frequency of 2,3 MHz. From the physical point of view the two methods are complements to each other. In radiography the signal for defect detection is essentially the X-ray intensity difference in direction of the radiated length of the defect whereas the ultrasonic echo intensity is mainly proportional to the area perpendicular to the ultrasonic beam. Due to these facts the radiographic method is better suited for volumetric defects and the ultrasonic method for area like defects. Though the modern phased array techniques applied in the project are capable to overcome these limitations by the additional degree of freedom in beam angle steering by electronic means the radiography will be kept as accompanying analysis method while welding process development.

B1.1 Introduction to the subject

NDT Reliability is the degree to which the NDT system is capable to reach the intended aim in defect detection and characterization and false calls /1/. For the intended aim of quality assurance and production control of lid welds of copper canisters this intended aim is:

The first is to be able to, in quantitative terms, express the risk for failing to detect canisters containing defects in the sealing weld that exceeds the acceptance criteria.

The second is being able to define the measurement accuracy of the inspection in terms of defect size and location.

The work should result in a general methodology that can be applied for other inspection tasks for the canister. In order to meet the requirements of the overall SKB timetable the methodology should be developed by the end of 2004 and at least preliminary results should be available at that time.

B1.2 Overview of general strategies in measuring reliability of NDT

Three different ways to investigate the reliability of NDT signals will be described. The first way of investigation, the performance demonstration, is preferred e.g. in the US American nuclear power industry. This is an integral consideration of the non destructive test as a system where the whole NDT system is packed in a black box and only the input in terms of

the real existing flaws in the component is considered and compared to the output in terms of the indications of the human inspector or of the automated system. The second – the European tradition – relies on a standardized description of physical/technical parameters of the NDE system which are preconditions for successful system performance. As example for such a standardized set of performance parameters the X-ray system of SKB is characterized completely against CEN … which will be described in detail in the technical report. The third approach – the modular conception – is a marriage of both: The signal chain is cut into main modules. Each module is assessed in a most appropriate individual way e.g. via modeling calculations. The single results are joint together according to the reliability theory of systems where the reliability of the total system is composed of the reliability of the subsystems. Separating criteria for the system were proposed through a reliability formula developed during a series of European-American workshops on NDE reliability. Examples of all three approaches will be given.

The NDE system consists of the procedures, equipment and personnel that are used in performing NDE inspection. According to this definition we consider in the situation in Figure B1-1, where we have on the left hand side the "truth" of the component which is in our case a weld with defects and on the right hand side the corresponding inspection. This means that we have a 100% reliability when we would have a 1:1 correspondence between both sides. Since this is in almost all practical examples not the case we have to set up tools to measure and maintain reliability to raise the acceptance of NDE in between the neighboring engineering sciences and to make global reliability conceptions like risk based inspection or risk based life time management feasible.

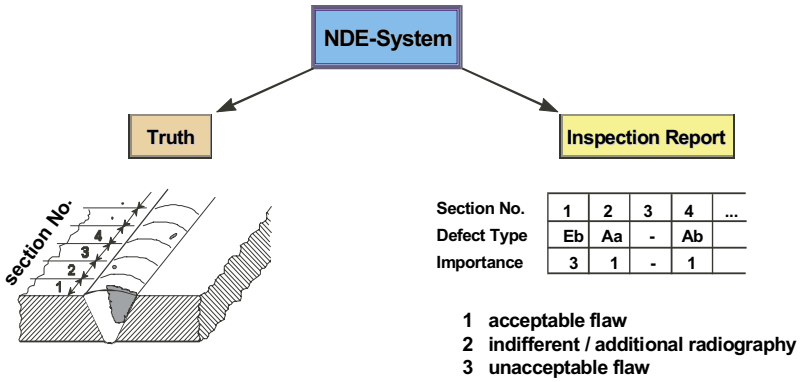

Figure B1-1. *The aim of NDE – describe the real status.*

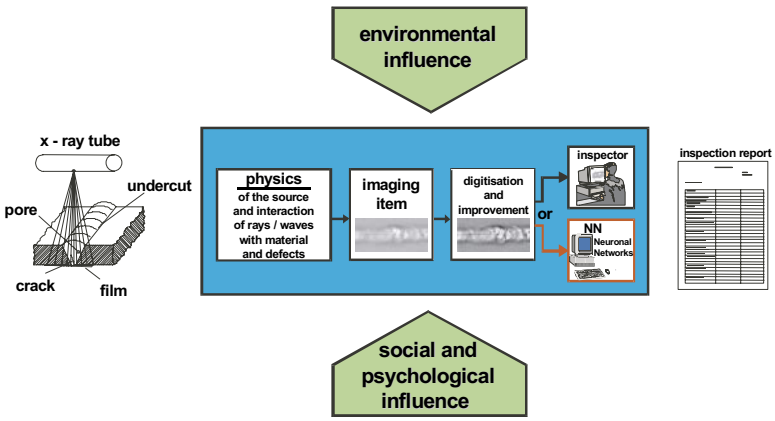

Figure B1-2. *The signal transfer chain of a radiographic system.*

In Figure B1-2 we take a closer look to an NDT signal transfer chain: The signal starts in terms of an energy beam or wave from a source and interacts with matter in terms of a component and its possible defects creating an output signal which is driven by the physics of the method. This output signal is now further influenced by the physical and technical properties of the more or less complex receiver and converter in our case the imaging item for X-rays and the digitization and processing unit. Now we consider how to measure and ensure the reliability. We start with the European approach: This approach is by far the most widely used, and is in use in most, or perhaps all, industries. One or more reference objects (such as DIN image quality indicators for radiography, or ASTME127 reference blocks for ultrasonic inspection) are used, in combination with written procedures controlling details of the inspection method, to reach and demonstrate consistency. The intent is to achieve essentially the same inspection conditions, independent of where, when, or by whom an inspection is conducted. This type of European approach is often used to provide process control information of a qualitative nature: loss of control at some earlier manufacturing stage is indicated by the unexpected occurrence of numerous or large indications, for example. Capability for detection of "real" (naturally-occurring) defects is not given for all applications, but is sometimes inferred from the size of the simulated defects in the reference objects, or from the size of defects that have been detected in past inspections using the same conditions. This inferential process can often lead to false conclusions about the detectability of real defects but it is assumed that the validation occurred during the accomplishment of the standard and its use over many years. Considering the example of a signal transfer chain of a NDE system in Figure B1-2, the application of standards means that the performance of the NDE system is assured in defining the values of parameters in each module. This approach will be demonstrated in section two with the help of examples of recently developed European standards for Radiography and especially X-ray film digitization.

Performance demonstration

The "American Approach", is the Performance Demonstration. We begin with the Performance Demonstration for empirical applications. In its simplest form this approach involves the use of material samples containing known defects as a basis for studying the effects on detectability of factors such as calibration, changes in inspection equipment, or inspector training programs. For the other inspection parameters that are not deliberately changed, consistency is still pursued through the use of reference objects and control procedures. Test programs of this kind are often used in conjunction with "round-robin" or other interlaboratory data acquisition procedures. This type of test can be applied equally well to NDE methods producing qualitative (i.e. pass/fail) or quantitative (i.e. signal amplitude) outputs, but in practice they appear to have been used most widely with qualitative methods in terms of blind trials as indicated in Figure B1-3, where input (the true defect situation in the component) and output (defect indication in the inspection report) are compared and the signal transfer chain from Figure B1-2 is treated as black box. To date, the major efforts in this type of Performance Demonstration have been the PISC program (focusing on characterization of all types of ultrasonic testing of nuclear power plant components), and the PDI program at the Electric Power Research Institute (EPRI) NDE center, in Charlotte, NC, (under whose auspices some hundreds of testing companies have already passed examinations in manual ultrasonic testing, and automated testing of pressure vessel nozzles, according to the ASME code, section XI, appendix VIII) /2/. The reliability measurement tools in terms of ROC (Receiver Operating Characteristic) and POD (Probability of Detection) – see also /3/ – will be described in section 3. Beyond the empirical approach exists the Performance Demonstrations for Quantitative Applications. An example of this type of characterization is provided by the POD evaluations conducted by US aircraft engine manufacturers to satisfy requirements of the Engine Structural

Integrity Program (ENSIP), as described in MIL-STD-1823 /4/. Attention has been focused on the use of NDE techniques for the detection of low-cycle fatigue cracks. Samples containing laboratory-generated cracks in representative surfaces (such as bolt-holes) are inspected, and results are reported in a pass/fail version for qualitative methods such as penetrant inspection, or signal amplitude form for quantitative methods such as eddy-current inspection. Data are presented in the POD versus flaw size format, and inspection thresholds are adjusted to achieve specific POD values for specific defect sizes. Most distributed is the "â versus a" method, which is illustrated schematically in Figure B1-4.

The third approach, the modular approach which is described in detail in /5/, might be considered also as the scientific basis for the well known ENIQ methodology /6/. The objective of the modular approach for measuring the reliability of NDE is to provide a validated testing system that fulfills the requirements of the client in the most efficient and cost effective manner. This capability is especially important where expensive statistical tests are not possible. In developing this concept we divide a system into appropriate sub modules as indicated in Figure B1-5, and evaluate the discrete reliability of each. The knowledge gained within each of the modules allows an optimization of the total system. The reliability of the total system is then determined by combining the single reliabilities of the modules, including their possible correlation according to e.g. fault tree analysis /7/.

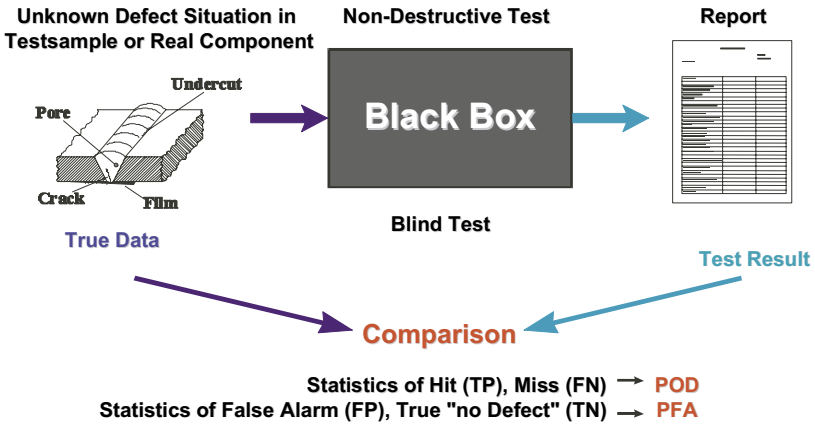

Figure B1-3. *Principle of a "Performance Demonstration".*

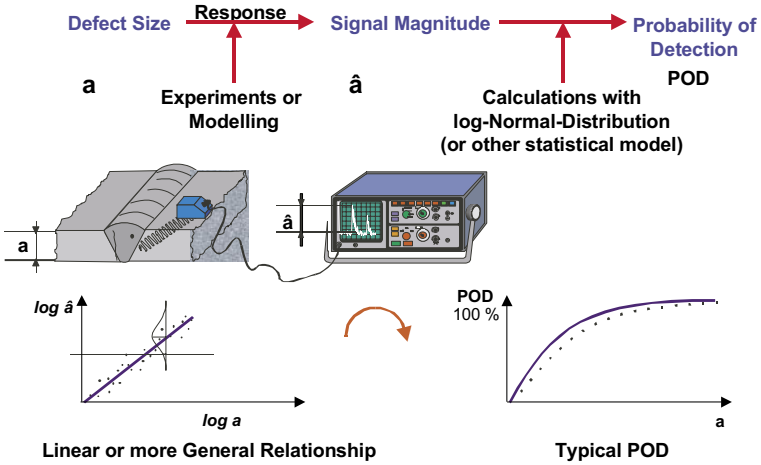

Figure B1-4. *"â versus a" – for automated thresholding systems a forecast of the POD is possible from the statistics of the response signals.*

The modular approach facilitates direct integration of the 1st American-European Workshop Reliability Formula /8/ as illustrated in Figure B1-6. The expression defines a total reliability R, which consists of: an intrinsic capability IC describing the physics and basic capability of the devices, factors of industrial application such as restricted access in the field, AP, and finally the human factors HF.

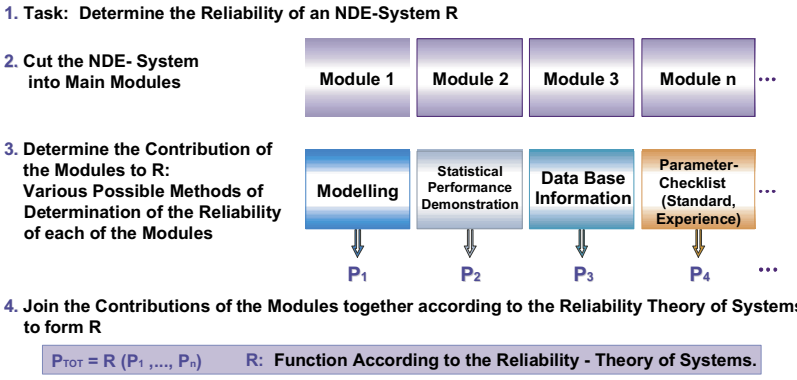

Figure B1-5. *Modular validation – combination of different influencing factors.*

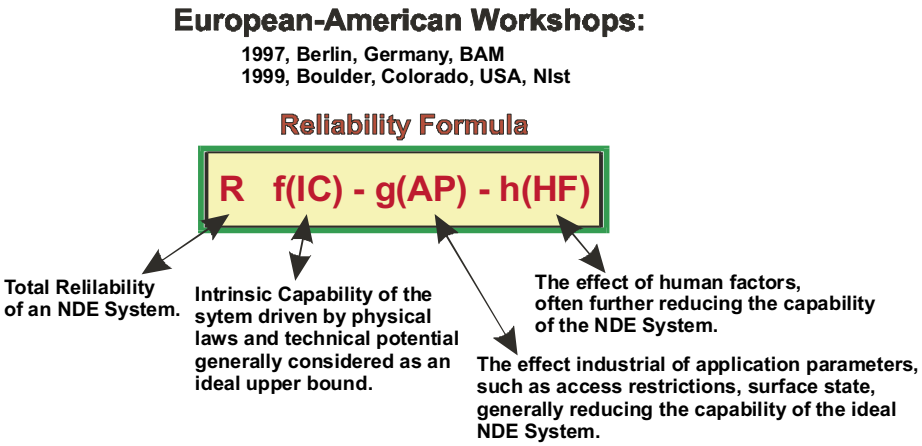

Figure B1-6. *Identifying the main influencing factors by the reliability formula.*

Parameter approach: Definition of essential parameters in European standards

As discussed above the conventional European approach relies on the definition of essential device and process parameters and their allowed variations to provide a basis reliability. We consider here the present status of European standards for industrial radiography with focus on characterization of X-ray systems and digitization /9/.

As already mentioned, these standards ensure the basic functioning of the techniques. The final performance containing all the three factors from IC, AP and HF needs to be demonstrated in an integral performance test as described in the next section or analyzed in a modular validation.

Integral approach: ROC – Receiver Operating Characteristic and its relation to POD

The Receiver Operating Characteristic (ROC) /10, 11/ is deviated from the general theory of signal detection and widely used since over 40 years in fields of evaluation of diagnostic systems like radar techniques, test of human perception and in medical diagnosis and since the eighties also in NDE. The general four possible situations in NDT (Nondestructive Testing) diagnosis are presented in Figure B1-7.

For both "true situations", defect present or no defect present, we have the possibility to recognize the truth (TP, TN) or to miss the truth with a false indication (FN, FP). The idea of the ROC method is to characterise the accuracy of an inspection system by evaluating the true positive detection rate versus the false positive detection rate for a set of possible decision criteria or recording levels in the language of NDT which represents a varying sensitivity. The creation of an ROC curve in this way is shown in Figure B1-8 where – following the curve from the lower left corner to the upper right – the sensitivity of the system raises. So – in the lower part of the curve the highest signals (correct indications) are included and only a small amount of noise (false calls). In the higher part more and more all of the defects are taken into account but also a greater amount of false calls has to be paid as price. The underlying mathematical model in terms of the two Gaussian signal distribution curves for the defect signals and the noise respectively are shown on the right hand side. In practice – especially in the case of manual testing with hit miss results – it is not possible to apply continuously growing signal thresholds and to count correct and false call rates for each because it is too much effort. A practical hit/miss experiment situation for defect evaluation on X-ray images is presented in Figure B1-9.

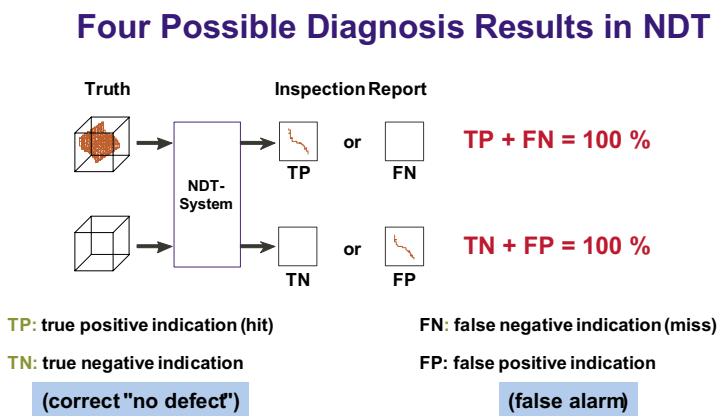

Figure B1-7. *The principles of ROC (Receiver Operating Characteristic). The possible diagnosis results in NDT.*

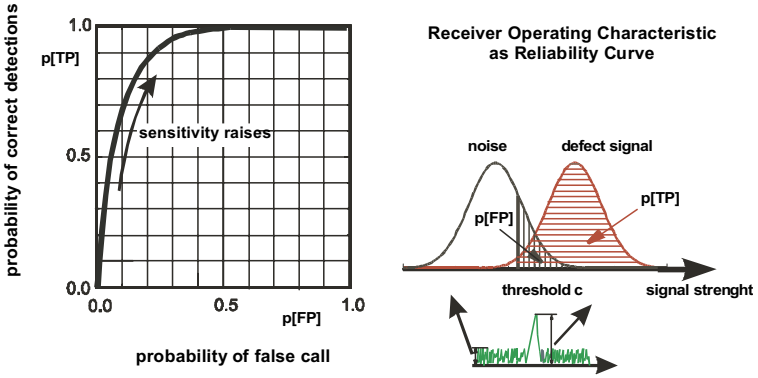

Figure B1-8. *Characteristic of one NDT – system by an ROC curve.*

For the ROC investigation the weld length is partitioned into grading units of 1 cm. The right hand side shows the scheme of the corresponding evaluation form. For the practical evaluation different discrete categories of signal counting are defined to be applied by the inspectors during the non-destructive testing evaluation as indicated in Figure B1-10. These categories might correspond to the visibility of defects on a radiographic film – as demonstrated in Figure B1-9 – or to an echo height in an ultrasonic A-scan like the signal plots in Figure B1-10. With the approach scheduled in Figure B1-10 we yield five different experimental points in the ROC diagram. The maximum point represents the actually reachable maximum operating point of the NDT system. From the whole curve shape – which can be obtained by using a Maximum Likelihood regression method on the basis of the binormal model – the overall fictive capability of the system is indicated. There is e.g. a forecast possible what will happen when the sensitivity of the system will be raised: Is there a valuable gain in defect finding or is only the false alarm rate increasing? Considering the area under the ROC-curve – which give an overall measure of the systems capability to separate a defect signal from noise – (see Figure B1-11) it may vary from 0.5 (pure chance curve 1) up to 1.0 which corresponds to an ideal NDT system belonging to the left corner's step curve.

For the fictive systems shown in Figure B1-11 the performance of the system increases from curve 1 to curve 7. Figure B1-12 illustrates the relation between ROC and POD (Probability of Detection curves): Both types of curves have the same statistical background – only the results are arranged with respect to different variables For illustration we consider one point on the ROC curve with fixed false call probability and take the corresponding POD value

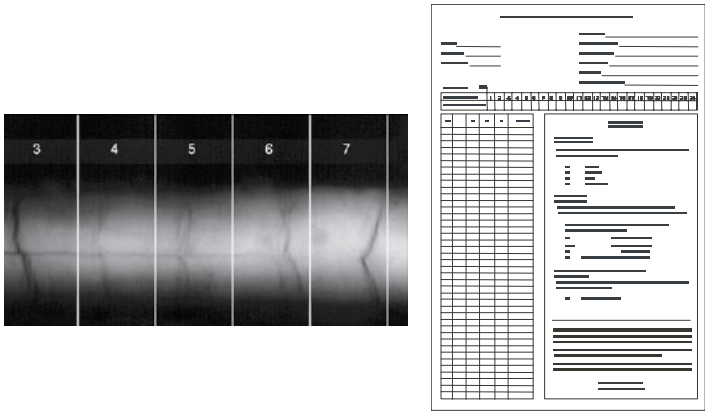

Figure B1-9. *Grading units (length intervals).*

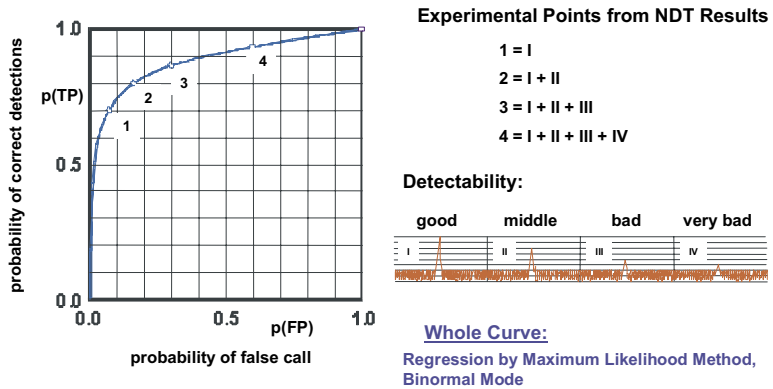

Figure B1-10. *Practical determination of the ROC-Curve.*

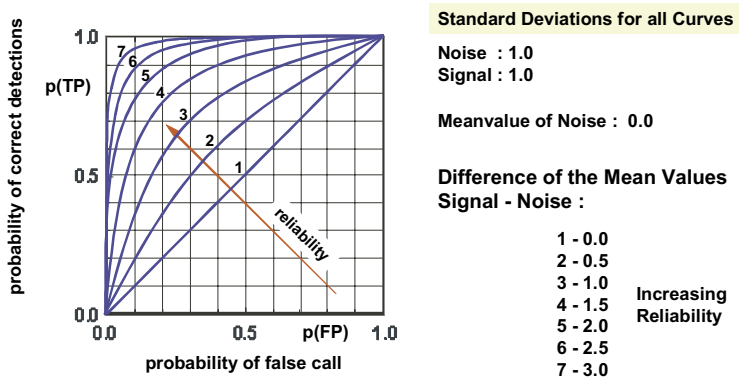

Figure B1-11. *Comparison of different NDT Systems.*

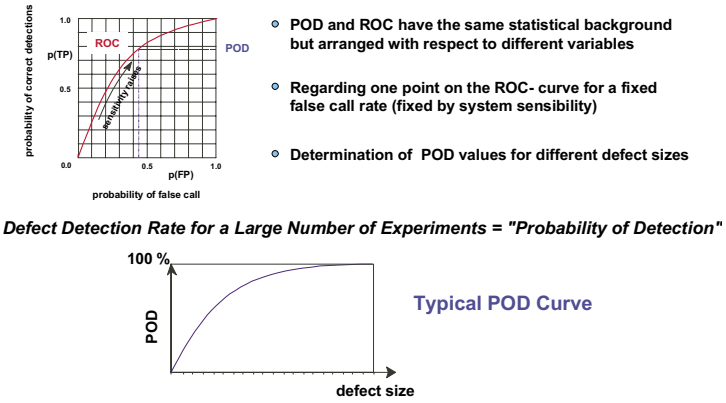

Figure B1-12. *ROC – POD connection (POD – Probability of Detection).*

in terms of the p(TP). This value has then to be spread off over the different defect sizes as indicated in the lower part of Figure B1-12. When there is a very strong dependence on the defect size – the ROC curve has to be recorded for each defect size separately.

Example 1: Reliability investigation of radiographic testing with filmdigitization of austenitic tube welds using ROC and POD

The subject of the example are cracks due to inter granular stress cracking corrosion in austenitic tube welds. During the last decade these inter granular stress cracking corrosion in thin walled austenitic pipe welds of German boiling water reactors has been widely investigated /12/. The cracks had grown during plant operation but are induced by a susceptible material's composition and medium influence /13/. Most of these longitudinal cracks start at the inner tube surface from root undercuts and move along the grains towards the heat influenced zone in the base material. The results are curved shaped cracks more or less open with possible ramifications. According to these properties the radiographic images of those cracks are more blurry and lower in contrast than images of cracks in ferritic steel.

The straight double wall penetration scheme especially designed for these types of cracks is shown in Figure B1-13. The straight double wall penetration scheme especially designed for these types of cracks is shown in Figure B1-13. In Figure B1-14 some realistic cross sections from austenitic welds along with the corresponding radiographs (simulation according to /14/) are shown. From the left to the right the crack changes its shape from a "good" to a "bad" crack and in the same direction the radiographic method is reaching its boundary.

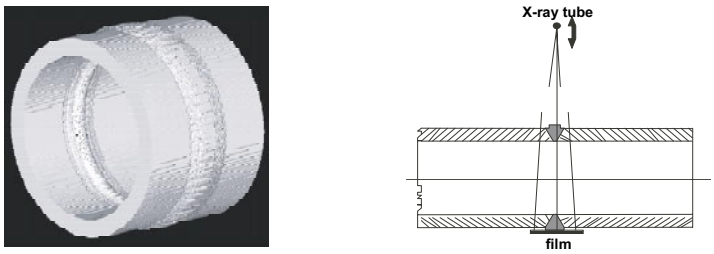

Figure B1-13. *Double wall penetration for in-service-inspection.*

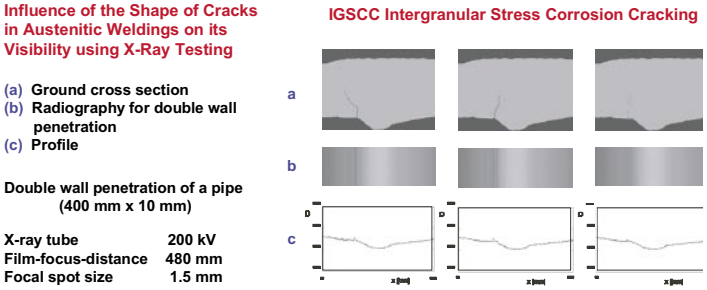

Figure B1-14. *Example: Reliability 1nvestigation of radiographic testing of austenitic tube welds.*

For the ROC investigation – according to Figure B1-15 – forty welds were radiographed and afterwards tested destructively to determine the truth about the defects. This truth has been compared to the results of the following inspection modes carried out by five inspectors: Conventional evaluation on a light box, digitized images on computer screen with 70µm and 35µm space resolution (12 bit dynamic range) and digitized images filtered with a special method dedicated to longitudinal cracks "70 µm*"/15/. In the ROC curves in Figure B1-16 no significant difference can be seen between the conventional film evaluation, the 70 µm and the 35 µm resolution digitized image evaluation for AGFA D4 films and our defect population when no special image filtering is applied. A significant growing in performance can only be observed for the special filtering technique and special trained inspectors.

To see this more clear the operating points only of the individual inspectors are compared for the four evaluation modes.

We learned from this test that the advantage of digitized films occurs only in case the inspectors had enough training with the handling of the device and the digital filters contain "a priori" knowledge about the welds. An other important parameter in addition to the pixel size on the digitized image is the pixel resolution of the monitor as precondition to offer the high resolution data in a proper way to the human eye.

An other interesting way to analyze the recorded data is to look for the POD distributed over the defect sizes in terms of crack depths. The "true valueas" were determined using REM (Raster Electron Microscopy) -images of the ground cross sections illustrated in Figure B1-18. Figure B1-19 shows the POD values for three different set up: Theoretical values for straight notches, the results of film evaluation on the light box and the results of the special filtering technique on 70 µm* digitized images. The deviation from straight notch detection is simply explained by the special zigzag shape of the cracks. Especially of

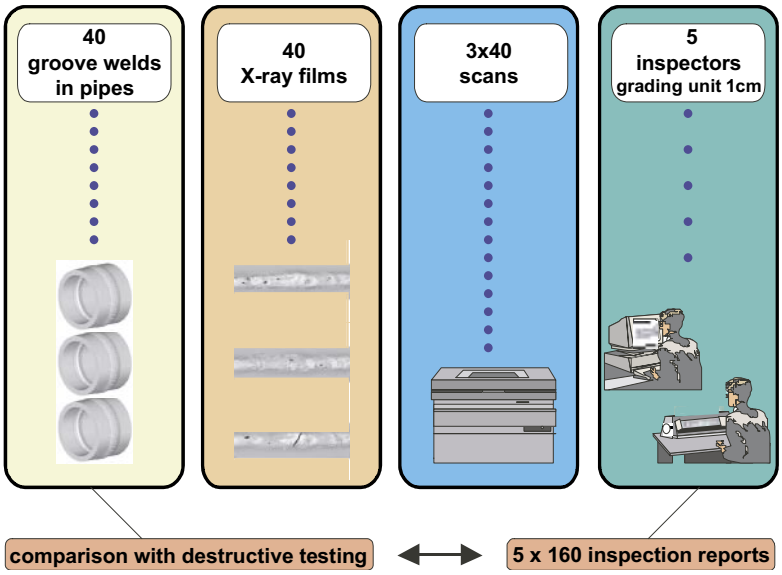

Figure B1-15. *Outline of the experiments.*

70 µm* : digitized image, 70 µm spatial resolution,
spezial filtering, inspector with experience

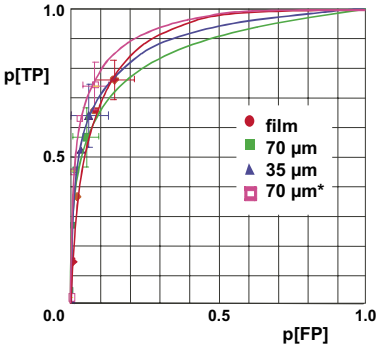

Figure B1-16. *Detection of cracks (inspector average).*

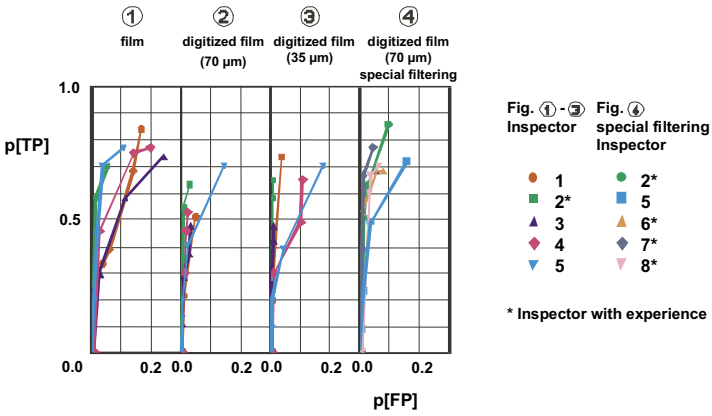

Figure B1-17. *Detection of cracks of each inspector.*

Figure B1-18. *Determination of defect (longitudinal crack) depth using ground cross.*

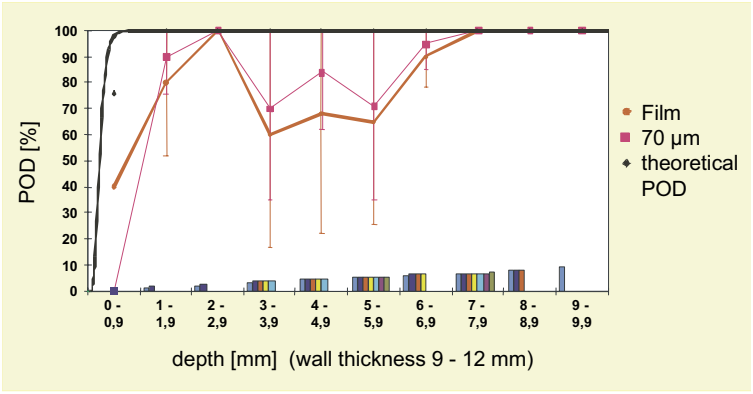

Figure B1-19. *POD versus crack depth.*

interest is the “POD valley” in the middle range. This is the range where the cracks start to deviate remarkable from straight growing which is resulting in a blurred image. The crack images become better detectable again when the cracks are almost wall through and wider.

Example 2: Reliability investigation of ultrasonic testing of gasturbine engine parts using POD and the modular approach

The task here is the detection of porosity in electron beam welds of Ti-alloy aircraft engine parts using automated ultrasonic testing with focussed probes. Of special interest is here the “naturally occurring flaw shape” of the porosities looked for by NDE in comparison to ideal shaped flaws. The detailed approach is described in /16/.

Figure B1-20 shows the scheme of modules applied for the investigation.

Sphere bottom bore holes were used as the ideal flaw. The POD as a function of sphere diameters is presented in Figure B1-21: We see a fast rising POD whereas the POD for the naturally occurring pores (Figure B1-22) rises much more slowly. The shape of the pores seems to strongly influence the POD – which up to now had physically not been clear. We can now formally define the application factor influence by dividing the latter POD by the former – which is shown in Figure B1-23. And we call the corresponding portion of the POD PODAP where AP stands for application factor. This factor could be used to scale other sphere bottom hole values to realistic ones – as a concept.

- **Automated ultrasonic testing using focussed probes**
- **Natural flaw: Pores in the weld**
- **Ideal flaw: Sphere bottom bore hole**

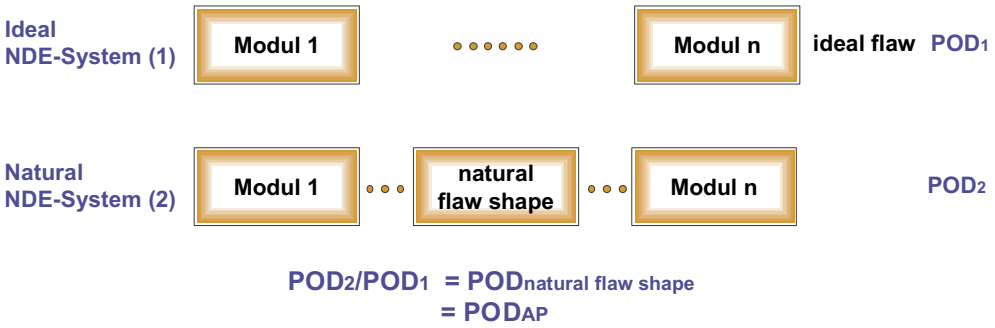

Figure B1-20. *Example 2: Determination of the influence of the "naturally occuring flaw shape" on the total POD.*

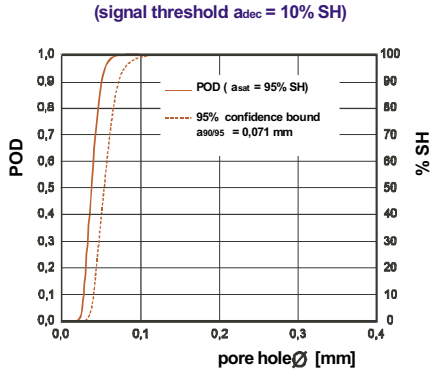

Figure B1-21. *POD1 sphere bottom bore holes.*

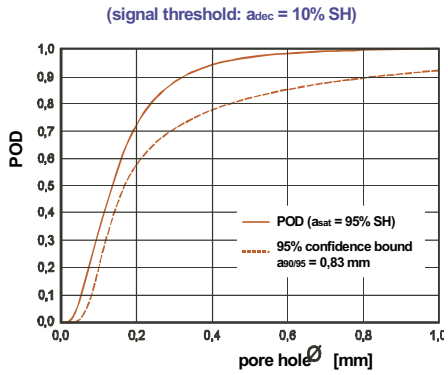

Figure B1-22. *POD2 naturally occuring pore holes in a weld.*

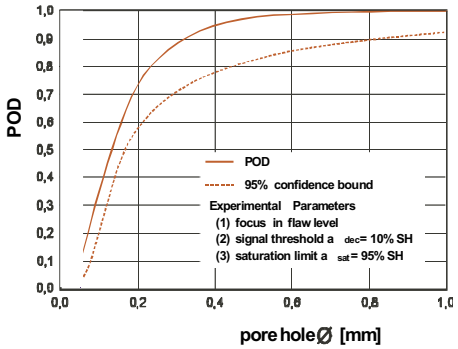

Figure B1-23. *PODAP = POD natural flaw / POD sphere bottom bore hole.*

B1.3 Conclusion

The first step of an performance demonstration is to define the essential technical parameters of the system. The ROC and POD methods are appropriate tools to provide a clear measure of integral performance of the system though it has to be paid by high effort in test series with realistic test samples. With POD the user can learn about the detection capability whereas the ROC gives more information about the system's capability to distinguish between signal and noise. The modular approaches open the door to a promising technique – more efficient and with the capability also to optimize the system.

B2 Selection of reliability data evaluation methods to be applied for SKB – general

Since the main goal is a high precision assessment of the flaw size which might be overseen by NDT the quantitative POD (Probability of Detection) method according to MIL-STD 1823 /4/ was selected for this assessment. The (1-POD) curve will provide the probability of missing a defect as function of defect size which can be used as input for probabilistic risk assessment of failing with the assumption of a certain flaw growth behavior under a certain load /17/.

The conventional procedure provides a mean POD curve as function of defect size with a corresponding 95% confidence limit curve due to the scatter in the experimental values. The defect size "a" which is counted as practically sure detectable is when the 95% confidence limit curve reaches the 90% POD level. It will be necessary to optimize the NDT methods that way the lower 99,9 confidence limit of the probability of detection of the critical a = 35 mm is close to 1.0.

B2.1 Signal POD, hit miss and ROC and plan of experiments

Both SKB-NDT-Methods are 100% digitized and fully mechanized methods so that a full signal analysis is possible according to the "â versus a" method. The method analyzes the full potential power of the NDT method to detect a signal â from a noisy background caused originally by a physical defect dimension of size "a" in the object under test. The shape of the final POD curve depends on the selection of the threshold from which on a signal is counted as a real signal from a defect in contrast to a noise signal.

Considering the reliability formula in Figure B1-6 this covers mainly the assessment of the intrinsic capability and controlled application factors. As long as human beings carry out the signal detection a so called “hit miss” POD has to be determined where the Human Factor is also assessed via a statistically reliable experiment using a number of different human interpreters. These results will then further be analyzed via the ROC method as described above to allow to select the optimum operation point also with respect to the economy of false calls. While the mathematical apparatus of “â versus a” and “hit miss” is described in annex A we explain here the strategy in the different phases of evaluation:

In the start up phase of the project the “State of the Art” PODs of the existing NDT systems at SKB will be evaluated on the basis of examples of full weld rings with minimum 60 defects (according to MIL1823) typical for the expected production situation. Which is accomplished for EB in the beginning of 2004 and for FSW in the second half of 2004. On the basis of this “state of the art” results, the NDT methods will be optimized empirically and on the basis of modeling calculations. Using the optimized NDT methods the plan of experiments will be specified on the basis of requirements from possible parameters of the welding process yielding typical defect distributions with possible significant differences and combined with parameter set ups of the experiments considered for further optimizing the NDT of the welds. Depending on the number of parameter sets to be combined and its occupation a full or reduced test matrix will be applied.

B2.2 Determination of the “true” defect locations, dimensions and shape

The common approach is to determine the real defect sizes via destructive tests. In the original approaches in the gas turbine industry the defects were even surface cracks to be easily measurable via optical surface methods. In our approach a high value was assigned to the possibility to have still access to the complete defects in the welds even in sections. When they are once destroyed no more reconstruction of the full 3D situation is possible since during destructive tests parts of information get lost. For this reason a full 3D high energy CT is used as reference methods to characterize all volumetric defects down to 1 mm diameter. Since the flat – area like – defects are too flat for reliable indication by the existing HECT an additional high precision UT analysis by the BAM UT facilities were accomplished for the reference. For a selected number of locations a partly destroying micro-CT analysis and destructive confirmations are planned to achieve full evidence about the defect reality.

B2.3 Measurement accuracy

Joint ROC

The so-called joint ROC can be used to get insight into the measurement accuracy. The joint ROC is an extension of the ROC introduced earlier in this document. Within this approach, an inspection is considered successful if the defect was not only detected (i.e. its location was determined) **but also** measured correctly (i.e. its size was determined with a given accuracy). This means that the joint ROC curve is always lower than or equal to the “normal” ROC curve, and both curves coincide in the ideal case. By comparing “normal” and joint ROCs one can get an impression about the sizing accuracy.

Uncertainty in measurement of the defect sizes

Every measurement process is inevitably influenced by uncontrolled factors. This leads to random errors in measurements. Therefore it is necessary not only to indicate the measured value, but also to state the associated uncertainty of measurement. The uncertainty can be characterized by the standard error of the measured quantity which is defined as the standard deviation of the measured magnitude's sampling distribution.

The main value of interest in the present project is the linear size of defects in the radial direction. In the present project the quantities obtained from the BAM NDT systems are considered to be true values. The uncertainty of the sizing can be expressed as follows:

$$\Delta l = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(S_i - B_i\right)^2}$$

where S_i is the radial size of the defect number i obtained using the SKB equipment, B_i is the radial size of the defect number i obtained using the BAM equipment and N is the total number of defects in the experiment.

B3 Selection of reliability data evaluation methods to be applied for SKB – specific

The application of the classical "â versus a" method is possible within an "effective" 1D approach applied separately to the X-ray technique for volumetric defects and UT-technique for area like defects.

For this approach the real problem of 3D defects characterized by 2D or even 3D signal arrays has to be transformed to an "effective 1D"–problem.

For X-ray "a" is assigned to maximum penetrated length difference between the bulk background and the defect in ray direction and â is assigned to the maximum contrast on the detector screen.

The example in Figure B3-1 shows the first POD date analysis for the state of the art SKB X-ray system.

In this example a is not exactly the penetrated length in 35° but the radial projection. The curves growing from left to right are the POD and its 95% confidence bound according to Mil1823. According to Mil1823 it is usual to work with a90/95 – this is the size at which the 95% confidence bound curve crosses 90% – as defect size which is detected for sure. For our example of volumetric flaws detected by radiography it is 2.65 mm. The curve growing from right to left is the 1-POD curve describing the probability of missing a defect of a certain size. In case of spherical pores this would already meet the requirements but since the real pores have non-symmetric shapes further studies have to be undertaken to investigate the additional influences presumably in a 2D POD approach.

For the UT-technique "a" is assigned to the area of the defect perpendicular to the UT "beam" and â is assigned to be the maximumecho hight. The corresponding POD curve with 95% confidence bound and 1-POD can be seen in Figure B3-2. The a90/95 value is 67 square mm which would correspond to a diameter of 9.2 mm in case of a sphere shape.

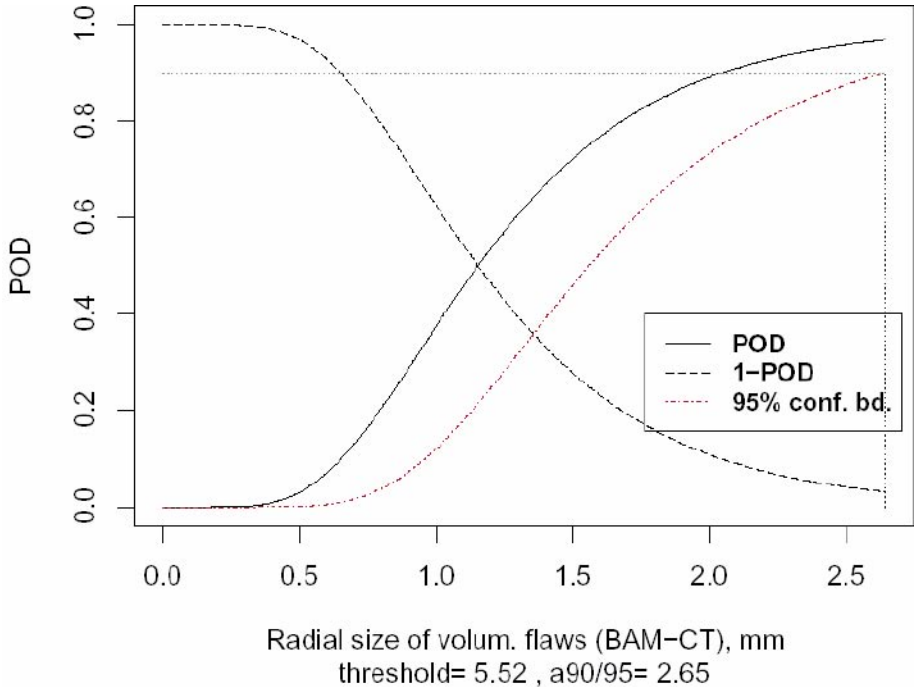

Figure B3-1. *POD curve for the radial size (radiographic testing, EB-weld).*

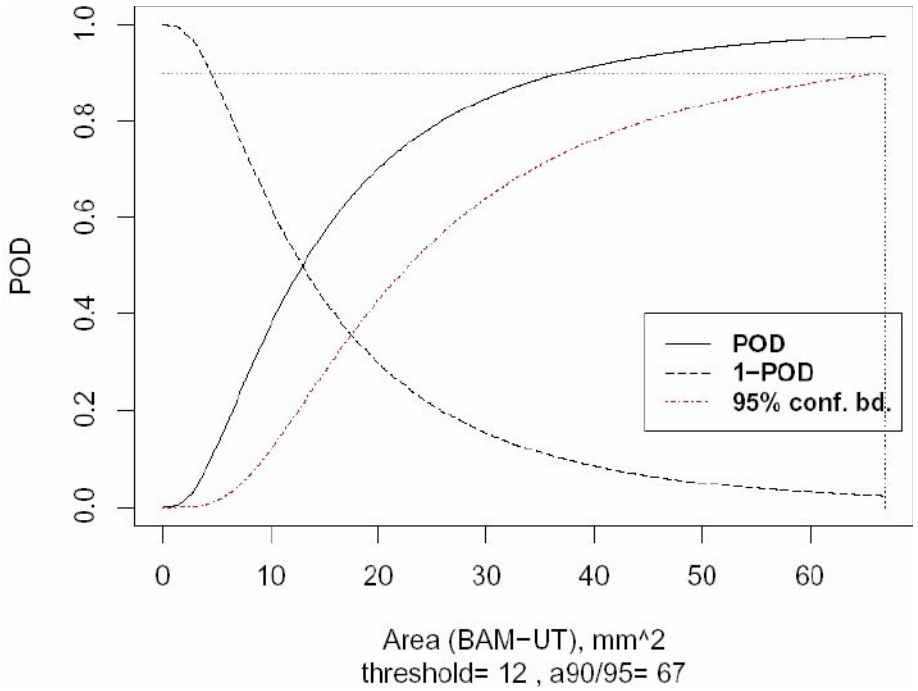

Figure B3-2. *POD curve for the area (ultrasonic testing of EB,-weld) .*

Conclusion and outlook

For the reliability assessment of the intrinsic capability and part of application factors of NDT- reliability the signal-POD (or â versus a) method according to Mil1823 is well suited to determine the remaining probability of the applied NDT method to oversee a defect of a certain size. The originally 1- dimensional conception can be used in defining “effective” 1D parameters. A comprehensive treatment of the problem, especially in separating all the physically actually influencing factors from the factor of interest, the radial dimension of the defects, the full 2D approach (see annex A) needs to be applied. This will be worked out during 2004.

B4 References

/1/ **ASNT, 1999.** Topical Conference Paper Summaries Book of the American-European Workshop on Nondestructive Inspection Reliability, September 21–24, 1999, NIST, Boulder, CO, USA, ISBN: 1-57117-041-3.

/2/ **ASME, 1992.** Boiler and Pressure Vessel Code, Section XI, Rules for Inservice Inspection of Nuclear Power Plant Components. American Society of Mechanical Engineering. 1992 Edition.

/3/ **Nockemann C et al, 1994.** Performance Demonstration in NDT by Statistical Methods: ROC and POD for Ultrasonic and Radiographic Testing, Proceedings, 6th European Conference on Non Destructive Testing, pp 37–44.

/4/ Non-Destructive Evaluation System Reliability Assessment, MIL-STD-1823 (U.S. Department of Defense).

/5/ **Mueller C, Fritz T, Tillack G-R, Bellon C, Scharmach M, 2001.** Theory and Application of the Modular Approach to NDT Reliability, Materials Evaluation, vol 59, no 7, pp 871–874.

/6/ European methodology for qualification of non-destructive tests, second issue document: ENIQ.SC(96)2 (E970055/pl).

/7/ **Barlow R E, Fussel J B, Singpurvalla N D, eds, 1975.** “Reliability and Fault Tree Analysis, SIAM, Philadelphia”.

/8/ Proceedings of the European-American Workshop Determination of Reliability and Validation Methods on NDE, Berlin, Germany, June 18–20, 1997, ISBN 3931381-18-8.

/9/ **Zscherpel U.** Filmdigitization systems for DIR: Standards, Requirements, Archiving and Printing on http://www.ndt.net/article/v05n05/zscherp/zscherp.htm

/10/ **Nockemann C, Heidt H, Thomsen N, 1991.** Reliability in NDT: ROC study of radiographic weld inspections, NDT&E International 24 5, pp 235–245.

/11/ **Metz C E, 1978.** Basic Principles of ROC analysis, Seminars in Nuclear Medicine 8 4.

/12/ **Csapo G, Just T, Eggers H, Hein E, Nimtz R, 1994.** "Flaw Detectability of NDT on Thin Walled Austenitic Pipe Welds", Proceedings of the 6th ECNDT, Nice, October 24–28th, pp 995.

/13/ **Erve M, Wesseling U, Kilian R, Hardt R, Brümmer G, Maier V, Ilg U, 1994.** Cracking in Stabilized Austenitic Stainless Steel Piping of German Boiling Water Reactors – Characteristic Features and Root Cause, Proceedings of the 20th MPA-Seminar, Safety and Reliability of Plant Technology, October 6–7, pp 29.1.

/14/ **Tillack G-R, Bellon C, Nockemann C, 1994.** Modeling within a global conception of Reliability of NDE, Non-Destructive Examination Practice and Results, State of the art and PISCIII Results, Proceedings of the Joint CEC OECD IAEA Specialists Meeting held at Petten on 8–10 March, Edited by E. Borloo and P. Lemaitre, EUR 15906 EN NEA/CSNI/R (94) 23, pp 389.

/15/ **Zscherpel U, Nockemann C, Heinrich W (Siemens AG, KWU Berlin), Mattis A (Siemens AG, KWU Erlangen), 1995.** Neue Entwicklungen bei der Filmdigitalisierung, DGZfP- Berichtsband der Jahrestagung, Aachen 1995, 47 1, S. 369–376.

/16/ **Feist W D, Tillack G R, 1997.** Ultrasonic Inspection of Pores in Electron Beam Welds – Evaluation of Detectability, Proceedings of the European-American Workshop Determination of Reliability and Validation Methods on NDE, June 18–20, BAM, Berlin, Germany, pp 291–298.

/17/ 3rd European-American Workshop on Reliability of NDE, September,11–13, 2002, Berlin, Germany, Organized by DGZfP, BAM and ASNT, supported by EPERC and several European NDE societies Christina.Mueller@bam.de

Appendix 1

Signal response analysis

Introduction

The annex describes the mathematical model and calculation method for the signal response analysis or "â versus a" method. According to Figure B1-4 from the main paper it is asked what is the relation between a defect dimension a in the weld and the physical signal magnitude â caused by the defect of size a. Further the distribution of the the signals â can be transformed to a POD Probaility of Detection under the assumption of a threshold which is used to distinguish between noise and signals from the dfects. The scheme is desribed for 1D and 2D signal distributions.

General description

Consider a quantitative NDT system. Upon being presented a stimulus a, it generates a response â. If the response exceeds a certain decision threshold $\hat{a}_{dec}$, the system registers a flaw detection. As the NDT system is influenced by uncontrolled factors, stimuli of the same magnitude, i.e. discontinuities of the same size, can cause responses of different strength. For this reason the strength of the response â to the stimulus of size a is considered as a random value and associated with a probability density $g_a(\hat{a})$. The relation between a and â can be expressed as follows:

$$\hat{a} = \mu(a) + \delta$$

Here $\mu(a)$ equals the mean value of $g_a(\hat{a})$ and δ is the random error whose distribution determines the probability density $g_a(\hat{a})$.

In practice, it is often assumed that δ is distributed normally with zero mean and constant (independent of a) variance. $g_a(\hat{a})$ is then the normal density function with mean $\mu(a)$ and variance equal to that of δ.

The probability of detection (POD) as function of the size of the discontinuity is:

$$\mathrm{POD}(a) = P\{\hat{a}(a) > \hat{a}_{\mathrm{dec}}\} = \int_{\hat{a}_{\mathrm{dec}}}^{+\infty} g_a(\hat{a}) d\hat{a}$$

Figure A1-1 illustrates this formula. The probability of detection is represented as hatched part of the area under the bell curve.

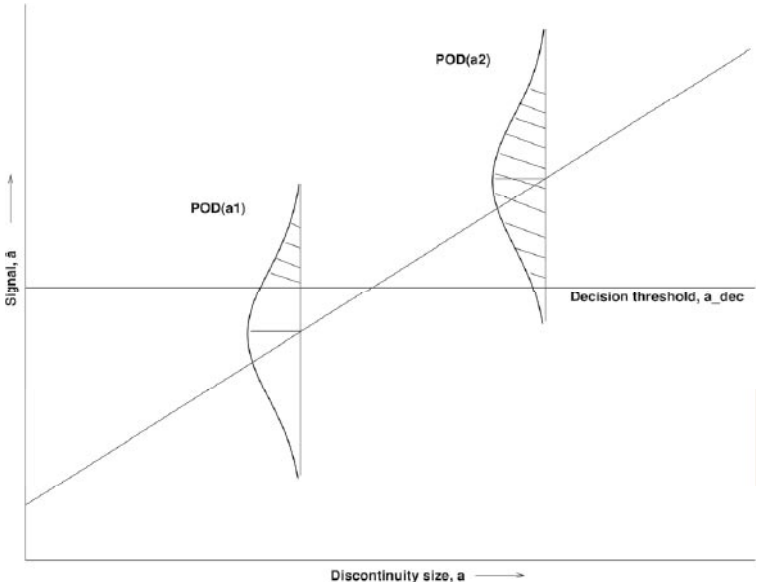

Figure A1-1. *Probability of detection.*

Calculation of the POD

Source data are *a* and â – arrays of length *n* that contain sizes of the defects and response magnitudes, respectively, and the decision threshold $\hat{a}_{dec}$. Note that the theory for dealing with censored data has been developed (see /1/) but is not used here, because the data sets available to us do no contain censored data.

Calculation of the POD function parameters

The following formula is commonly used to model the relation between *a* and â:

$$\ln \hat{a} = \beta_0 + \beta_1 \ln a + \delta \qquad \text{(A1-1)}$$

Here δ is normally distributed with zero mean and constant variance σ_δ^2.

Under the assumptions of the model, the POD function has the following form:

$$\mathrm{POD}(a) = P\{\hat{a} > \hat{a}_{\mathrm{dec}}\} = P\{\ln(\hat{a}) > \ln(\hat{a}_{\mathrm{dec}})\} = \Phi\left(\frac{\ln a - \mu}{\sigma}\right),$$

where Φ is the standard normal distribution function, and

$$\mu = \frac{\ln \hat{a}_{\mathrm{dec}} - \beta_0}{\beta_1}$$
$$\sigma = \frac{\sigma_\delta}{\beta_1}$$

The parameters β_0, β_1 and σ_δ describe the linear dependency of â on *a* and have the following meaning:

β_0 Intercept
β_1 Slope
σ_δ Standard deviation of the residuals

Their values are estimated from the arrays *a* and â using the method of maximum likelihood.

The 95% lower confidence POD

The 95% lower confidence bound is given by the following formula:

$$\mathrm{POD}_{95}(a) = \Phi(\hat{z} - h)$$

where

$$h = \sqrt{\frac{\gamma}{nk_0}\left(1 + \frac{(k_0\hat{z} + k_1)^2}{k_0k_2 - k_1^2}\right)}$$

and

$$\hat{z} = \frac{\ln(a) - \mu}{\sigma}$$

Parameter γ reflects the sample size and is taken from Table A1-1.

Table A1-1. Values of γ for 95% lower confidence bound on the POD(a) function.

Sample size	γ for 95% conf.
20	5.243
25	5.222
30	5.208
40	5.191
50	5.180
60	5.173
80	5.165
100	5.159
∞	5.138

Variables k_0, k_1, k_2 are the components of the matrix:

$$I(\mu,\sigma) = \frac{n}{\bar{\sigma}^2}\begin{pmatrix} k_0 & -k_1 \\ -k_1 & k_2 \end{pmatrix} = \left(\frac{1}{\beta_1^2} TV(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma})T^{\mathrm{T}}\right)^{-1}$$

Here *T* is the transformation matrix:

$$T = \begin{pmatrix} 1 & \mu & 0 \\ 0 & \sigma & -1 \end{pmatrix}$$

Thus all variables needed to compute the POD curve and the confidence bound are described.

POD curves are often summarized by stating the single value $a_{90/95}$, i.e. the value of *a* at which the 95% lower confidence bound reaches the value of 0.9. Figure A1-2 shows a typical POD curve (solid line) with the 95% lower confidence bound (red dashed line) and the $a_{90/95}$.

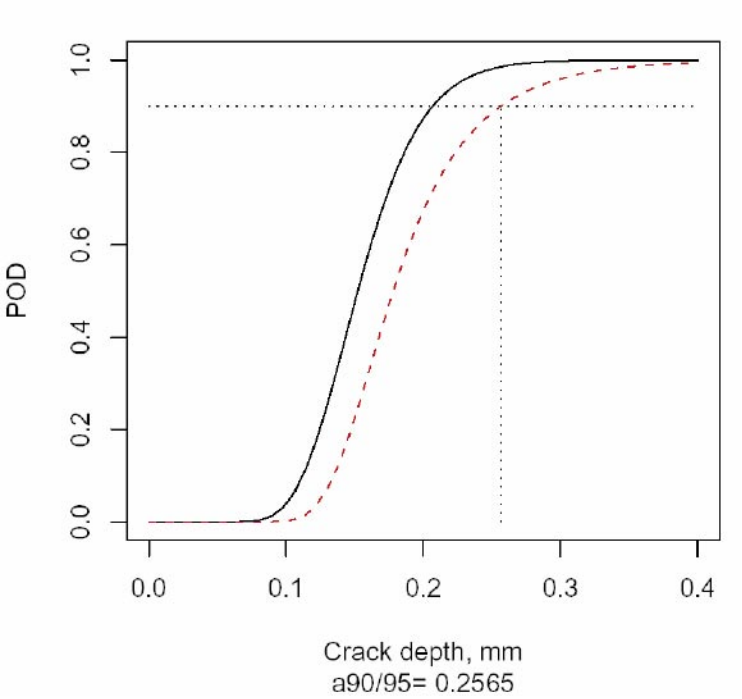

Figure A1-2. *A typical POD curve.*

Two-dimensional approach

If we denote the independent variable as a_1 and introduce an additional independent variable a_2 into the Equation (A1-1) that models the signal, then the equation takes the following form:

$$\ln(\hat{a}) = \beta_0 + \beta_1 \ln(a_1) + \beta_2 \ln(a_2) + \beta_{12} \ln(a_1) \ln(a_2) + \delta$$

Here $\beta_{12}\ln(a_1)\ln(a_2)$ is the interaction term and δ is, as in Equation (A1-1), the normally distributed random error with zero mean and constant variance σ^2_δ.

The POD curve then transforms to a surface. Figure A1-3 shows a contour plot of such POD.

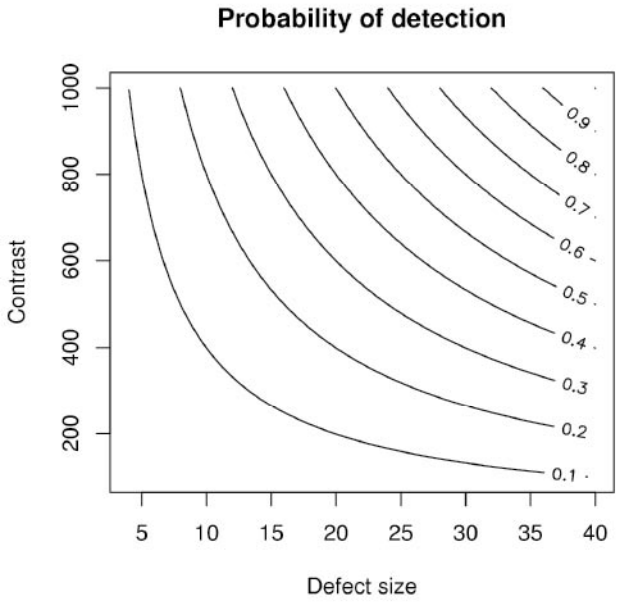

Figure A1-3. *Contour plot of the probability of detection.*

References

/1/ **Berens A P, 1988.** NDE Reliability Data Analysis. ASM International.

Appendix 2

Calculation example

Two hypothetical examples have been constructed to illustrate the methodology of fitting GEV and GPD models. The data was taken from a previous study of NDT-methodology /Ronneteg, 2001, section A6/, and does not reflect the assumed performance during the demonstration phase. Results from these test welds (L003, L005, L009 and L010) are listed in Table A2-1. The calculations were done with *S-plus*® using functions developed by Stuart Coles. Profile likelihood estimation of confidence limits was done with a function modified by Nader Tajvidi.

Table A2-1. Radial size (mm) of 43 defects in four test welds.

5.8	4.2	13.1	9.7
7.0	6.9	7.4	3.5
27.5	4.2	6.9	5.8
2.1	3.3	3.7	2.7
2.7	0.7	12.3	5.6
4.1	3.4	15.1	10.9
2.6	3.9	3.1	4.1
2.4	10.2	7.3	12.1
2.0	8.9	9.5	19.0
1.9	4.5	4.3	7.2
3.9	7.4	13.6	

These data were recorded from only four test welds, but let us assume for the purpose of this calculation example that the data represent the maxima from 43 independent objects (canisters). A GEV-model can be fitted by maximum likelihood estimation (Figure A2-1). The maximum likelihood estimates of the parameters with standard errors (within parentheses) are:

$\mu = 4.33$ (0.485)
$\sigma = 2.78$ (0.407)
$\xi = 0.285$ (0.139)

The quality of the fitted model can be evaluated from the probability and quantile plots (upper left and right), showing the agreement between the model predictions and empirical data expressed with either scale (Figure A2-2). The uncertainty in the predictions can be viewed in the return level plot with the 95% confidence bands added (lower left), based on an assumed asymptotic normal distribution of the maximum likelihood estimates. A more accurate estimate is provided by the profile likelihood method, but either way the uncertainty in model predictions is substantial when extrapolating from observed to unobserved levels.

The GEV model predicts an increased return level with an increased return period, also beyond the physical limit of discontinuity size (the lid weld is 50 mm). This illustrates the fact that the performance must substantially exceed the design criteria in order to

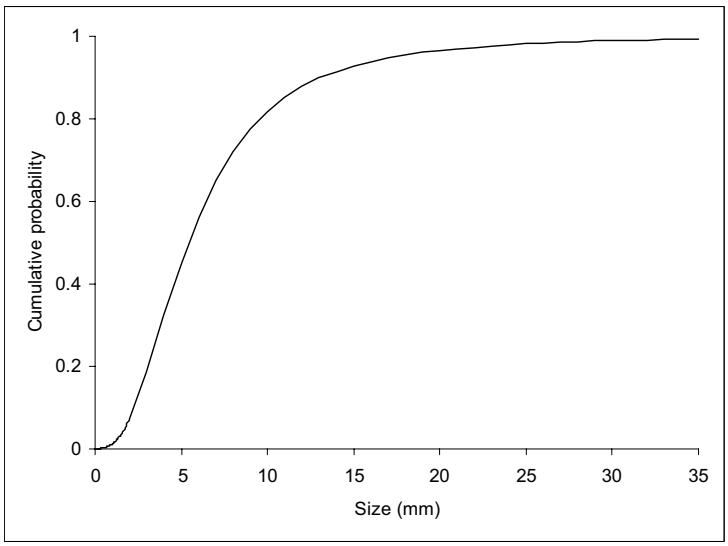

Figure A2-1. *Cumulative probability distribution for the fitted GEV model.*

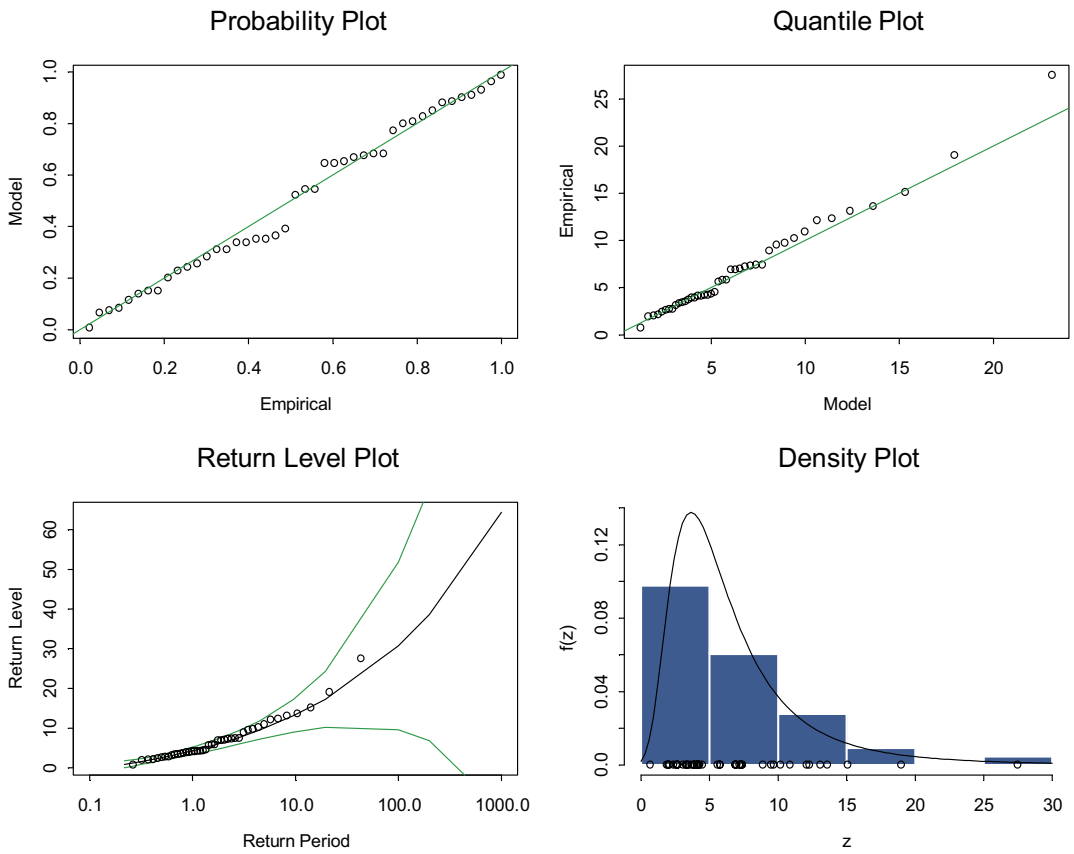

Figure A2-2. *Diagnostic plots for the fitted GEV model.*

demonstrated future compliance, if the statistical modelling is based on single maximum values and the GEV distribution.

Let us now assume, for the purpose of illustrating the other modelling approach, that the data in Table A2-1 represent measurements on 10 independent objects. Let us furthermore assume that we have established that the extremes do not have a tendency to cluster.
A GPD-model can then be fitted by maximum likelihood estimation (Figure A2-3).
A threshold of 5 was selected by graphical evaluation of model stability over a range of thresholds. The maximum likelihood estimates of the parameters with standard errors (within parentheses) are:

$\tilde{\sigma}$ = 5.81 (1.69)
ξ = –0.0728 (0.198)

The quality of the fitted model can also here be evaluated from the probability and quantile plots (Figure A2-4). The uncertainty in the predictions is shown in a return level plot with 95% confidence bands added, but also here the confidence bands are better estimated with the profile likelihood method and then become asymmetric (Figure A2-5). The deviance function can be approximated with the χ^2-distribution:

$$D(\theta) = 2\left\{\ell(\hat{\theta}_0) - \ell(\theta)\right\} \sim \chi^2_d \qquad \text{(A2-1)}$$

The 95% confidence limits are estimated from Figure A2-5 by applying the above formula (A2-1), i.e. *log-likelihood* $= -59.115 - \frac{1}{2}\chi^2_{0.05[1]} = -59.115 - \frac{1}{2} * 3.84146 = -61.036$ corresponding to a confidence interval of (16.3, 39.9).

The profile likelihood uncertainty estimates for small samples can be further refined by incorporating a scaling factor /Tajvidi, 2004, section A6/. The uncertainty in model predictions is still substantial, but less than for the GEV model. It can be noticed that the model predictions themselves do not stretch beyond the physical barriers.

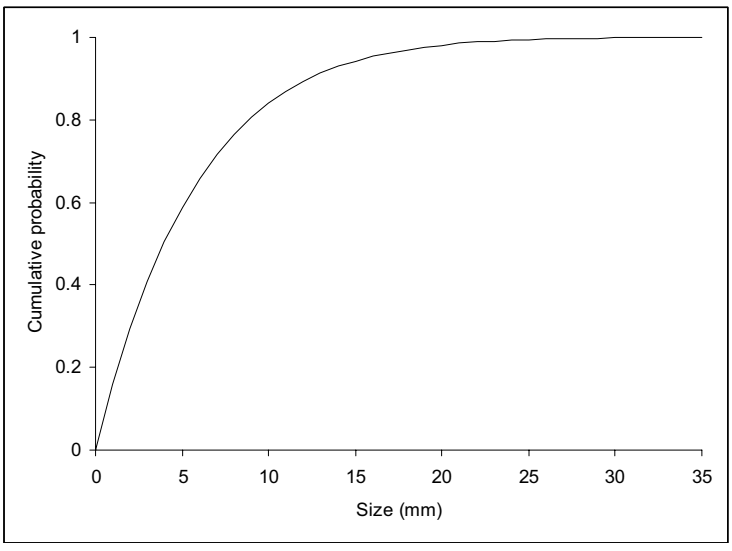

Figure A2-3. *Cumulative probability distribution for the fitted GPD model.*

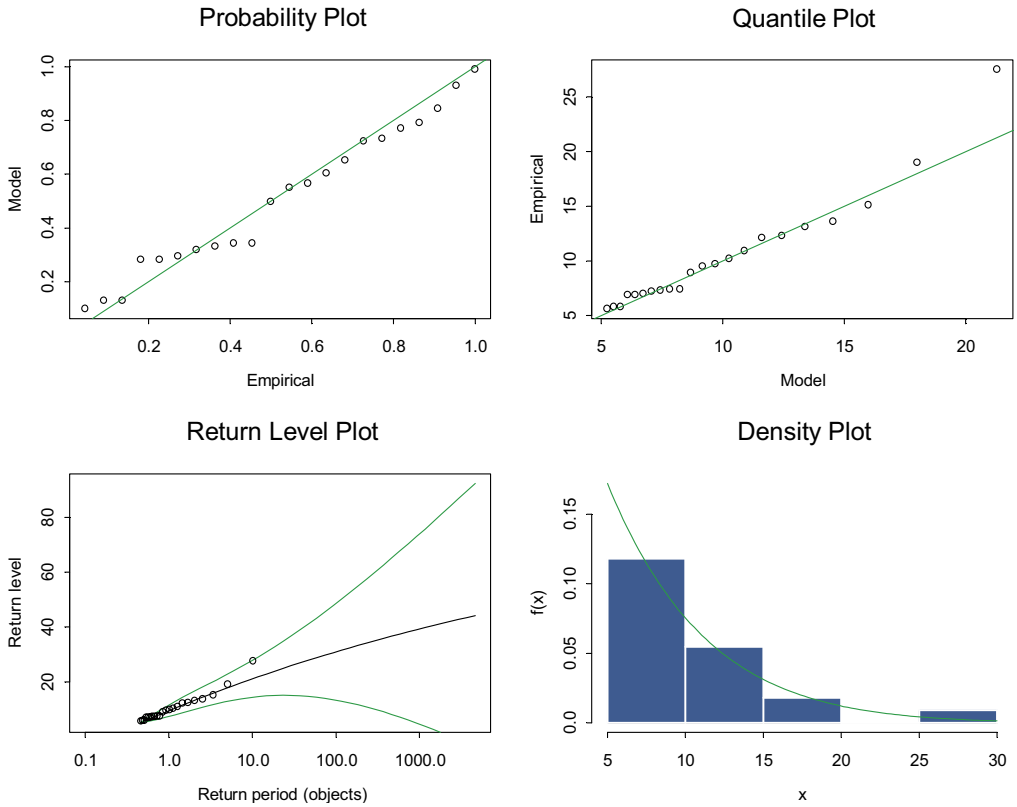

Figure A2-4. *Diagnostic plots for the fitted GPD model (assuming 4.3 observations per object).*

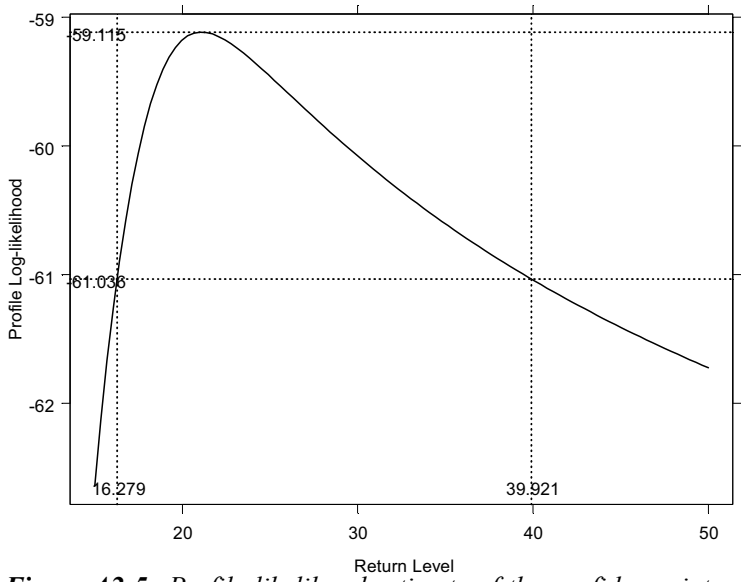

Figure A2-5. *Profile likelihood estimate of the confidence interval for the fitted GPD model at the return period 10 (assuming 4.3 observations per object).*