JOURNAL OF RESEARCH of the National Bureau of Standards
Vol. 89, No. 2, March–April 1984

An Iterative Calibration Curve Procedure

Clifford H. Spiegelman

National Bureau of Standards, Gaithersburg, MD 20899

Accepted: March 13, 1984

Calibration curves are an important part of many measurement processes. The user of a fitted calibration curve must know its precision and accuracy. These are determined in a timely fashion using the data iteratively. This paper gives a method that divides the data into training and test groups. The test group is iteratively checked to see that a prechosen nominal confidence interval probability of coverage is met. If on the basis of this check the calibration experiment is completed, the nominal probability level is shown to still be valid.

Key words: constants; measurements; observations; probability; statistics; statistical methods.

1. Introduction

Calibration curves are an important part of many measurement processes. The user of a fitted calibration curve must know its precision and accuracy [1][1], and these are determined in a timely fashion by using the data iteratively. This paper gives a method that divides the data into training (calibration curve-producing) and test (check) groups. The test group is iteratively checked to see that a prechosen nominal confidence interval probability of coverage is met. If on the basis of this check the calibration experiment is completed, the nominal probability level is shown to still be valid.

We assume that the measurement process has negligible drift. This is only partially checked by the iterative calibration technique; of course, routine application of control chart procedures is a must [2].

It is also assumed that many measurements are taken between calibrations. Under this circumstance particularly appropriate statistical calibration procedures are found in Scheffé [3], Lieberman, Miller, and Hamilton [4], and Knafl, Sacks, Spiegelman, and Ylvisaker [5]. We concentrate on the Scheffé procedure; it is demonstrated on an engineering example in Lechner, Reeve, and Spiegelman [6].

All of these procedures produce interval estimates such that the true value is contained in $(1-\alpha)\%$ of them in the long run with probability $1-\delta$. The two probability levels α and δ are chosen by the calibrator.

In order to describe the iterative procedure, the necessary notation is given.

About the Author, Paper: C. H. Spiegelman is with the Statistical Engineering Division in NBS' Center for Applied Mathematics. His work was partially supported under Office of Naval Research contract N00014-83-k-0005.

[1]Figures in brackets indicate literature references at the end of this paper.

2. Notation and Method

There are two fundamental variables: Y which is a nonstandard measurement of a property and x which is an exact standard or certified value of a possibly different property. For the example in section 3, x represents the gravimetric value (mass) of liquid in a tank (fig. 1) and Y represents differential pressure.

These two variables are related by the equation $\mathbf{Y}=H\boldsymbol{\beta}+\sigma\mathbf{e}$, where the terms of the equation are defined below. The other observables are $Y_i^*, i=1,2,\ldots$, and they correspond to unknown x_i^*.

Here $\mathbf{Y}$ is an nxl vector of observations, H is an nxp full rank matrix whose i-th row is $\mathbf{h}_i^t=\mathbf{h}^t(x_i)=(h_1(x_i),\ldots,h_p(x_i))$, $\boldsymbol{\beta}^t=(\beta_1,\ldots,\beta_p)$, $\mathbf{e}$ is an nxl vector of independent and identically distributed standard normal random variables having mean zero and covariance matrix $V(\mathbf{e})=\mathbf{I}_n$, and σ is the standard deviation. The Y_i^* are post calibration observations and the goal is to estimate their associated x_i^*; in this paper x_i^* are taken to be unknown constants.

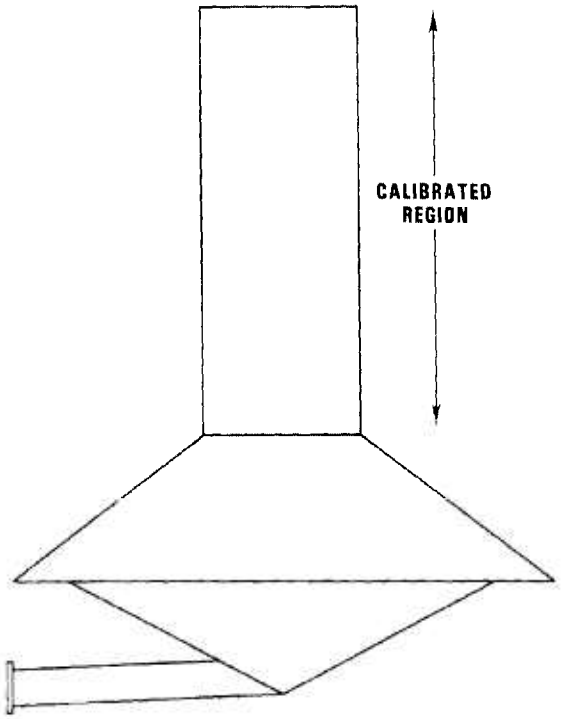

Figure 1—Calibrated tank located at NBS. A cubic model was used to correspond to linear deformation of all the tank walls.

The calibration curve is denoted by $m(x)=\mathbf{h}'(\mathbf{x})\beta$; it is taken to be monotonic. Let the least squares estimate of $m(x)$ be denoted by $\hat{m}(x)$ and its variance by $\sigma^2 s^2(x)$. Initially we assume that σ^2 is known. We discuss estimating σ^2 in section 4.

Data are nearly always collected sequentially; therefore it makes sense to analyze them sequentially. Once the measurement process is out of control additional measurements are of value only in identifying the problem. If a reasonable statistical procedure is available for iteratively analyzing the data, as is the procedure outlined in this section, then it should be used. This will help identify out-of-control situations early.

Of course the ability to detect out-of-control situations depends on the calibration design, i.e., x-values used for the calibration. Such designs have been discussed in detail for linear spline calibration curves [7]. As a byproduct of the present investigation we show the soundness of the advice in the cited work against using the exact optimal design. In fact efficiency under an assumed model and an ability to check when this model holds are competing demands.

A procedure is given for checking in an ad hoc fashion the validity of the previous assumptions. The checks are deliberately for coverage probabilities rather than directly for the assumptions, i.e., the stated $(1-\alpha)$ uncertainty level is checked. This is an indirect check on the underlying assumptions. If an assumption is marginally violated and yet the $1-\alpha$ is met, the author sees little reason to doubt the calibration procedure. If the nominal level is not met, then the calibrator is expected to at least check his measurement procedure and possibly reset his equipment. The novelty of this procedure is that if the experiment is carried to completion, the nominal levels $(1-\alpha)$, and $(1-\delta)$ remain valid.

Our procedure is as follows:

Step 1. After a reasonable amount of data is collected, the data are divided into two groups, SG1 and SG2. New data are placed in either group. Ways in which this may be done are given as comments at the end of this section. Each group should contain approximately half the data, although under some circumstances other divisions are reasonable (see section 4). The partitioning of the available data can be done randomly or according to a well chosen statistical sampling plan (see the comments at the end of this section).

Step 2. Choose the probability levels $1-\alpha$ and $1-\delta$. In order to simplify the notation, anything calculated only from the data in SG1 has subscript 1; anything calculated from all the data has no subscript.

Step 3. Using only data from SG1, determine the least squares estimate of $m(x)$, $\hat{m}_1(x)$ and its variance $\sigma^2 s_1^2(x)$.

Step 4. From the data in SG1, form the Scheffé upper and lower curves $U(x)$ and $L(x)$. The rationale is given in Scheffé [3] and Lechner, Reeve, and Spiegelman [6].

For *all* x

$$U(x)=\hat{m}_1(x)-\sigma(z_\alpha+\chi_\delta(p)s_1(x))$$

$$L(x)=\hat{m}_1(x)+\sigma(z_\alpha+\chi_\delta(p)s_1(x)).$$

Here z_α is the two-tailed α point of a standard normal; $\chi_\delta^2(p)$ is the upper δ point of the chi-squared distribution with p degrees of freedom and

$$\chi_\delta(p)=\sqrt{\chi_\delta^2(p)}.$$

Step 5 (optional). Calculate the minimum of $s_1(x)$ in the calibration region. Denote $\min s_1(x)$ by s. Redefine z_α to be the solution q of the equation

$$\Phi(q+2\chi_\delta s)-\Phi(-q)=1-\alpha. \qquad (1)$$

As explained [5], this step reduces the conservativeness of the Scheffé procedure while maintaining the validity of the probability statements.

Step 6. For each (x_i, Y_i) in SG2 check whether or not $x_i \in [L^{-1}(Y_i), U^{-1}(Y_i)]$. Let

$$T_i = \begin{cases} 1 \text{ if } x_i \in [L^{-1}(Y_i),\ U^{-1}(Y_i)] \\ 0 \text{ otherwise} \end{cases}$$

Recall both $\hat{m}_1(x)$ and $m(x)$ are linear combinations of the $h_j(x)$, $j=1,\ldots,p$. Let $\theta^t=(\theta_1,\ldots,\theta_p)$ be a vector parameter in R^p and $\theta^t \mathbf{h}(x)=\hat{m}_1(x)-m(x)$. Let $p(x,\theta) = \Phi(\theta^t\mathbf{h}(x) + \chi_\delta(p)s_1(x) + z_\alpha) - \Phi(\theta^t\mathbf{h}(x) - \chi_\delta(p)s_1(x) - z_\alpha)$. Finally denote the likelihood conditioned on SG1 and thus also on $\hat{\beta}_1$ by

$$L(\mathbf{x},\theta);\ L(\mathbf{x},\theta) = \prod_{i \in SG2} p(x_i,\theta)^{T_i}(1-p(x_i,\theta))^{1-T_i}.$$

From the likelihood $L(\mathbf{x},\theta)$, get the maximum likelihood estimator $\hat{\theta}$ for θ and compute the maximum likelihood estimator for $p(x,\theta)$, $p(x,\hat{\theta})$. Check whether or not $p(x,\hat{\theta}) \geqslant 1-\alpha$ for all x in the calibration region. If for some x, $p(x,\hat{\theta}) < 1-\alpha$, consider the measurement process possibly defective. If for all x, $p(x,\hat{\theta}) \geqslant 1-\alpha$ and the calibration experiment is not finished, collect the next data point and return to step 1.

Many scientists may not have the computer programs readily available to form the efficient maximum likelihood estimator of $p(x,\theta)$. In these cases we recommend using local averaging or otherwise smoothed estimates (see Stone [8] or Collomb [9]). In particular we recommend a nearest neighbor approach. Choose a number k and then at each point x in the calibration region average the T_i values corresponding to the k closest x_i values to x. For small samples there is little known about choosing k; however, in large samples a value of k approximately equal to $n^{2/5}$ should be satisfactory.

This procedure provides a balanced check on whether the conservativeness of the Scheffé procedure and the lack of the model holding exactly, seriously alter the hoped for uncertainty level $1-\alpha$. The bigger the sample size, the less conservative the Scheffé procedure.

Comments about **design**:

As previously stated, the Scheffé procedure is very conservative when $s_1(x)$ is large. Therefore, some of the best diagnostic information comes from data where $s_1(x)=s$. The optimal (D-optimal) design takes observations where $s(x)$ is at a maximum. Thus for a straight line the optimum design has observations only at the ends of the calibration region. Some of the best diagnostic information occurs at x-values in the middle and will be missed with this design.

Comments about **subgroups**:

Often the calibrator will have a good understanding about the possible malfunction of his measurement system. Then a choice of subgroups will be clear. He should feel free to choose as many combinations as he likes. The validity of uncertainty statements for completed calibrations remains. The check procedures are ad hoc, and if many checks are performed, he should expect some of them to indicate a possible malfunction of his system. The interpretation of these ad hoc checks requires sound scientific and engineering judgment. Some possible choices of subgroups are:

1) If we are mainly interested in detecting drift then SG1 should contain the older measurements and SG2 the newer ones. If we want to check run-to-run variability, SG1 and SG2 should not contain observations from the same run.

2) Suppose we want to check whether or not $m(x)$ has the assumed form over a subinterval $[a,b]$. Then SG1 should *not* contain (if possible) observations with x-values in $[a,b]$.

3. Analysis

We show that if σ is known all the T_i are independent of $\hat{m}_2(x)$; in this case, our iterative check does not affect the coverage probabilities when the model defined in section 2 holds.

THEOREM. *When σ is known the statistics* T_i *are independent of* $\hat{m}(x)$.

PROOF:

$T_i=1$ if and only if

$$\hat{m}_1(x_i) - \sigma(\chi_\delta(p)s_1(x_i)+z_\alpha) \leqslant Y_i$$

$$\leqslant \hat{m}_1(x_i)+\sigma(\chi_\delta(p)s_1(x_i)+z_\alpha) \qquad (2)$$

Clearly eq (2) is equivalent to $-(\chi_\delta(p)s_1(x_i)+z_\alpha)$

$$\leqslant \frac{Y_i-\hat{m}_1(x_i)}{\sigma} \leqslant (\chi_\delta(p)s_1(x_i)+z_\alpha).$$

Given the least squares estimate for β, $\hat{\beta}_2$, $E[Y_i|\hat{\beta}]=\hat{m}(x_i)$; similarly $E[\hat{m}_1(x_i)|\hat{\beta}]=\hat{m}(x_i)$.

Thus, $Y_i-\hat{m}_1(x_i)$ is uncorrelated with $\hat{\beta}$. Since the $Y_i-\hat{m}_1(x_i)$ and $\hat{\beta}$ are jointly normal the T_i are independent of $\hat{\beta}$. Q.E.D.

Suppose σ is not known but estimated independently from the calibration experiment. Then if the upper and lower bounds in Scheffé [3] are modified as he indicated the desired uncertainty statements still apply. This follows from the fact that all of the T_i are independent of $\hat{\sigma}^2$. This is not obvious to the author so the details are included.

Let T_i be modified to incorporate replacement of σ by σ_1; see Scheffé [3] for details. It is to be shown that the T_i are independent of $\hat{\sigma}^2$. Divide all sides of modified eq (2) by $\hat{\sigma}^2$. After some algebra $\hat{\sigma}_1^2/\hat{\sigma}^2$ can be written as a function of the ratio of two independent chi-squares whose sum is proportional to $\hat{\sigma}^2$. By applying standard change-of-variable techniques, it can be seen that $\hat{\sigma}^2$ is independent of this ratio. Finally $[Y_i-\hat{m}(x_i)]/\hat{\sigma}$ are uniformly distributed on a unit sphere and are independent of $\hat{\sigma}^2$. Q.E.D.

Table 1. Mass-pressure calibration data.

Mass	Pressure	Run
567.004	2.06534	1
567.2	2.0655	3
567.22	2.05974	2
585.772	2.32647	4
586.091	2.32747	3
604.913	2.58939	5
604.964	2.5881	3
623.878	2.84772	3
680.441	3.62457	1
680.693	3.61958	2
699.204	3.88191	4
718.321	4.14248	5
737.333	4.39982	3
793.881	5.17109	1
794.134	5.16728	2
812.658	5.4279	4
831.74	5.68723	5
850.749	5.94467	3
907.347	6.71461	1
907.572	6.71065	2
926.108	6.97103	4

4. Example

The pressure mass calibration example is based upon data collected under the direction of J. Whetstone of NBS. The tank is of an experimental nature and is located in the fluid mechanics building at the National Bureau of Standards. The calibration curve relates pressure and mass measurements. In the region where the tank is used the calibration curve is hypothesized to be a straight line. However, due to bowing of the tank walls C. P. Reeve of NBS' Statistical Engineering Division and the author felt a cubic model was more appropriate. This model corresponds to linear deformation of all the tank walls.

The calculations made were done using the updated version of the program fully documented in Lechner, Reeve, and Spiegelman [10]. The updated program allows designation of training and test samples and automatically indicates whether or not a test point is in the calibration interval. Further information about this modification can be obtained from the author or C. P. Reeve.

The data are shown in table 1. In figure 2 residuals from the five runs are shown. Clearly run 2 is quite different from the others. However, as figure 3 indicates, the third run is also quite different from runs 1, 4, and 5.

In all cases σ^2 is estimated from the data. For the data on hand if SG1 contains any data points from run 2 then $p(x,\hat{\theta})$ is identically one. That is, the Scheffé intervals include all the data in SG2. This is true regardless of how many points are in SG1, provided it is five or more. (Note: Five is the minimum number of observations needed). If all of the points from run 2 are in SG2 then the Scheffé intervals cover none of them. In particular if SG2 contains only the data from run 2, $p(x,\hat{\theta})$ is identically zero, see figure 4.

Note that in typical cross validation procedures a fixed number of observations, usually one, is dropped out at a time and the procedure checked [11]. If this is done then the estimate of $p(x,\theta)$ is identically one. It can be shown that even if four or five observations are dropped out at one time the resulting average estimate of $p(x,\theta)$ will be nearly one. Thus, it appears that in this case purposeful choice of SG1 and SG2 is important.

5. Conclusions and Summary

It is important to find out early whether or not a calibration procedure is in control. In particular for the example in section 3, had the new procedure been applied the experiment might have been terminated as a failure after run 3. Alternatively one additional run to compensate for run 2 may have been collected. Surely

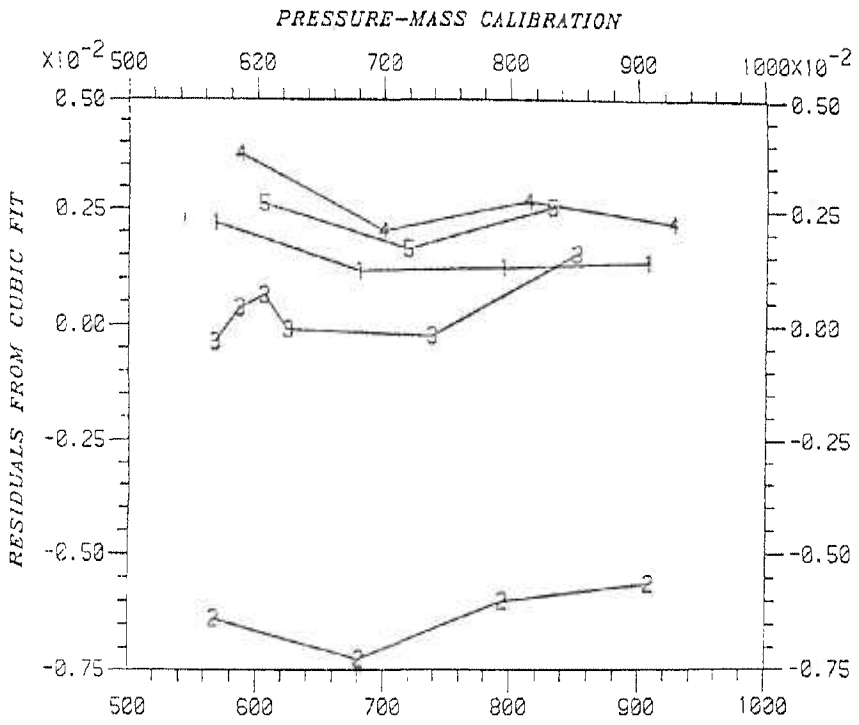

Figure 2—Residuals from runs 1-5.

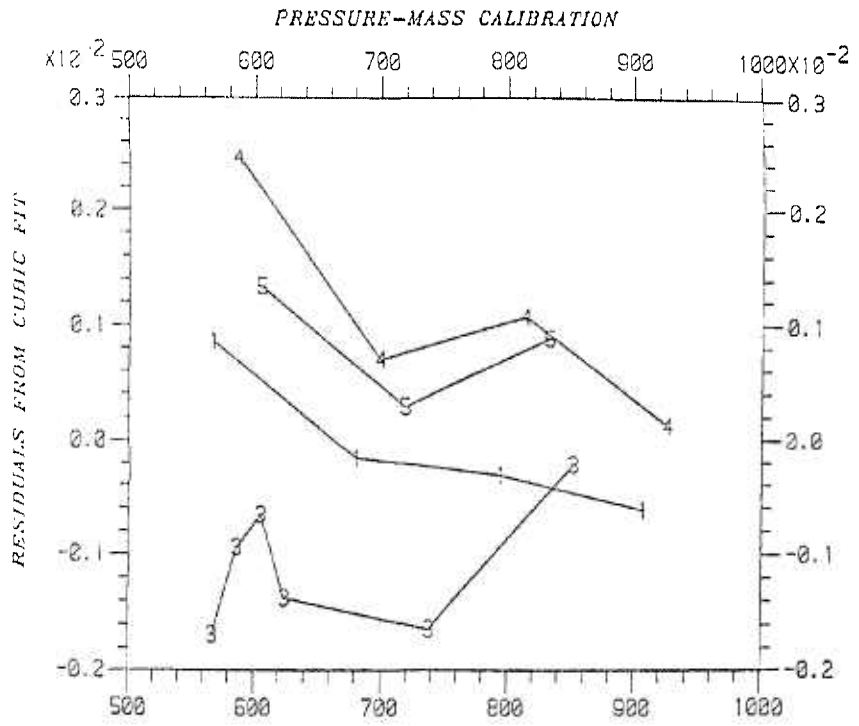

Figure 3—Residuals from runs 1, 3, 4, and 5.

something different would have been done. Clearly, too, the Scheffé procedure is conservative enough to account for some unmodeled run-to-run variation as in run 3.

Thus an iterative calibration can provide insight into the calibration procedure in a timely fashion without doing too much violence to the final uncertainty statements.

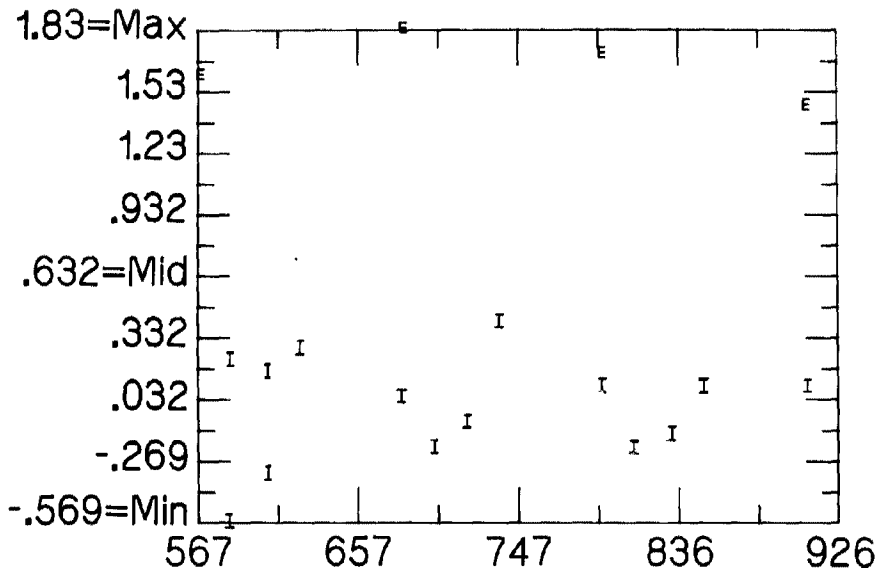

Figure 4—Summary of cross-validation results. Data from runs 1, 3, 4, and 5 are shown in SG1; data from run 2 are shown in SG2. A value bigger than 1 in absolute value indicates an x value outside the calibration interval.

Software Package:CPR*SPLINEUPDATE
Summary of Cross-Validation Results

	SG1	SG2
	X included	X excluded
Inside x C.I.	17	0
Outside x C.I.	0	4
Pct. Inside x C.I.	100%	0%

The author thanks J. Whetstone for providing the data and insight into his calibration system. The data were jointly examined with C. P. Reeve who has written a program to implement many of the procedures shown in this paper.

References

[1] Eisenhart, C. Realistic evaluation of the precision and accuracy of instrument calibration systems. Journal of Research NBS, 67c: 161–187; 1963.

[2] Parobeck, P. Tom B.; H. Ku; J. Cameron. Measurement assurance program for weighings of respirable coal mine dust samples. The Journal of Quality Technology, **13**, No. 3: 157–165; 1981.

[3] Scheffé, H. A statistical theory of calibration. Annals of statistics, **1**: 1–37; 1973.

[4] Lieberman, G. J.; R. G. Miller; M. A. Hamilton. Unlimited simultaneous discrimination intervals in regression. Biometrica **54**: 133–145; 1967.

[5] Knafl, G.; J. Sacks; C. Spiegelman; D. Ylvisaker. Nonparametric calibration. Accepted for publication in Technometrics (1984).

[6] Lechner, J. A.; C. P. Reeve; C. H. Spiegelman. An implementation of the Scheffé approach to calibration using spline functions, illustrated by a pressure-volume calibration. Technometrics **24**, No. 3: 229–234; 1982.

[7] Spiegelman, C. H.; W. J. Studden. Design aspects of Scheffé's calibration theory using linear spines. Journal of Research NBS, **85**: 295–304; 1980.

[8] Stone, C. J. Consistent nonparametric regression. Annals of Statistics, **5**: 595–645; 1977.

[9] Collomb, G. Estimation nonparamétrique de la régression: Revue bibliographique. International Statistical Review, **49**, No. 1: 75–93; 1981.

[10] Lechner, J. A.; C. P. Reeve; C. H. Spiegelman. A new method of assigning uncertainty in volume calibration. NBSIR 80-2151, 101 pages; 1980.

[11] Golub, G.; M. Heath; G. Wahba. Generalized cross-validation as a method for choosing a Good Ridge parameter. Technometrics **21**: 215–223; 1979.