NPL Report CISE 19/99

of the National Measurement

System Programme on

Software Support for Metrology

# MODEL VALIDATION IN THE CONTEXT OF METROLOGY: A SURVEY

BY

B.P. BUTLER, M.G. COX, A.B. FORBES,

P.M. HARRIS AND G.J. LORD

February 1999

# MODEL VALIDATION IN THE CONTEXT OF METROLOGY: A SURVEY

## NPL Report CISE 19/99

B.P. Butler, M.G. Cox, A.B. Forbes, P.M. Harris and G.J. Lord
Centre for Information Systems Engineering, NPL

February 1999

## ABSTRACT

Models are used in metrology to characterise measurement data and, where appropriate, the process which produced that data. Indeed, modelling is a procedure designed to extract *information* from *data*. Therefore, for any proposed model, the question then arises: is the information provided by the model comprehensive and reliable? This question is resolved by model validation. This report describes some theoretical concepts which prove useful in analysing models, and presents some practical examples from various metrological areas where they have been applied.

This report is the first deliverable of the Model Validation project within the UK Department of Trade and Industry's National Measurement System *Software Support for Metrology* Programme 1998-2001.

© Crown Copyright 1999
Reproduced by permission of the controller of HMSO

ISSN 1361-407X

Extracts from this report may be reproduced provided the source is acknowledged and the extract is not taken out of context.

Authorised by Dr Dave Rayner,
Head of the Centre for Information Systems Engineering

National Physical Laboratory, Queens Road, Teddington, Middlesex, TW11 0LW

# MODEL VALIDATION IN THE CONTEXT OF METROLOGY: A SURVEY

B Butler, M G Cox, A B Forbes, P M Harris, G J Lord
Scientific Software Section, NPL, Teddington, TW11 0LW, UK

25 February 1999

## Abstract

Models are used in metrology to characterise measurement data and, where appropriate, the process which produced that data. Indeed, modelling is a procedure designed to extract *information* from *data*. Therefore, for any proposed model, the question then arises: is the information provided by the model comprehensive and reliable? This question is resolved by model validation. This report describes some theoretical concepts which prove useful in analysing models, and presents some practical examples from various metrological areas where they have been applied.

# Contents

# List of Figures

# Chapter 1

# Introduction

This report is the first output of the Model Validation project of the National Measurement System (NMS) Software Support for Metrology (SSfM) programme, funded by the UK Department of Trade and Industry. The current SSfM programme runs from April 1998 until March 2000 [78].

The aim of this report is to support further work in the Model Validation project, collating knowledge and experience of model validation in metrology and elsewhere, and drawing on the research literature as appropriate.

A framework is presented for modelling in which validation is shown to play an important role. Validation methods are identified, and discussed in the context of some examples. It is intended that these methods would be applied in the model validation case studies and constitutes part of the technical input to the best-practice guidelines at the end of the project.

## 1.1 Outline of contents

The motivation and key concepts are introduced in this section. In particular, a framework to support the analysis of models is described. Model validation requirements are described, suggesting some of the reasons why model validation is important in metrology. Some of the particular difficulties are described; it is clear that these are far from trivial to solve. One of the main concepts introduced is that of *modelling risk* and of how this may be controlled by model validation.

In Chapter 2, the issues met when validating conceptual models are discussed, with particular reference to examples drawn from length metrology and thermal measurement within ionising radiation metrology. The section also summarises the main discrete modelling issues discussed in more depth in Appendix A.

This treatment and the latter example continue in Chapter 3. The reason for decomposing the problem into the two phases (conceptual and numerical) becomes clear on the grounds that

- the issues are different; and
- the phased approach provides a basis for minimising modelling risk.

In particular, the important modelling concepts of consistency, stability and convergence of numerical schemes for continuous models are introduced and discussed.

Chapter 4 summarises the main lessons from Chapter B for the important case of linear regression, or equivalently at every iteration of a nonlinear regression procedure.

Chapter 5 returns to applications, this time emphasising the breadth of models and the concomitant need to apply appropriate model validation techniques. The example applications include those drawn from metrology, where it is clear that traditional statistical procedures are common, partly because metrologists seek to control the nature of their experiments such that this is the case, but that this is not always possible, e.g., when assessing trends in air quality. In such circumstances, more advanced techniques are needed. By contrast, other branches of applied science, especially those studying natural phenomena have less control over the nature of their data or the underlying processes which generate that data. Hence, they have generally been more ready to draw upon more novel approaches. Hence, it can be argued that there is scope for technology transfer from these non-metrology applications to metrology in respect of model validation.

Conclusions are drawn in Chapter 6.

Appendix A presents a mathematical treatment of regression, which may be unfamiliar to many metrologists, but which has the benefit of identifying many of the main issues (particularly model spaces and error functions) in discrete modelling. Some of the concepts can be carried over to continuous modelling.

Appendix B seeks to classify the techniques which are available to assist in modelling and analysing models which have an inherently statistical nature, particularly those associated with regression problems. No survey of this kind can be truly comprehensive, but this section sets out what are believed to be the leading methods. Some of these will be unfamiliar to most metrologists, and the section assumes a moderate level of familiarity with statistical reasoning. Hence, it is probably best viewed by metrologists as a reference section, to be called upon if they encounter difficulties with modelling discrete metrological data. Perhaps the most practical material relates to how plotting various quantities can provide at-a-glance confirmation that, qualitatively, the model is not seriously wrong.

Appendix C identifies the three main sources of error when Finite Element Methods are used to solve continuous models, and indicates why *a posteriori* error analysis is gaining popularity among theorists and can also be useful for the validation of metrological models with respect to the residuals of the model.

With regard to the scope of this report, it is clear from the above that the software implementation, as such, receives relatively little attention in this report, since it is addressed more fully elsewhere in the SSfM programme, particularly in the

- *Testing spreadsheets and other packages used in metrology* project in respect of third-party software packages and the
- *Software development methods* theme in relation to software developed by and for the metrologist.

## 1.2 A metamodel of the modelling process

In this section we define the modelling process in the most general setting and introduce the concepts which underpin model validation.

The main steps of the modelling process are as follows [16]

1. requirements capture leading to a statement of the problem
2. development of the mathematical model
   (a) conceptual
   (b) discretised (algebraic)
3. software implementation and testing.

The process progresses in the order shown but, being a design process, there are feedback loops which are informed by model validation (largely in step 2) and software testing (largely in step 3).

Note the distinction between the conceptual model and the discretised model. The *conceptual model* can be written down on paper, in words and mathematical symbols.

We define the *model function* $F$ as follows:

$$F(\mathbf{y}, \mathbf{x}, \mathbf{a}) = \mathbf{0}, \tag{1.1}$$

where $\mathbf{y}$, $\mathbf{x}$, $\mathbf{a}$ are vectors of response variables, independent variables and parameters, respectively. Note that some of these variables and parameters might appear in associated functions $G$ (see below) and hence have values which are set elsewhere, i.e., they relate to the application domain and might not appear explicitly in the model. Furthermore, $F$ is to be seen as a general function in that it can include differential and integral operators.

Appendix A provides a geometrical interpretation of the theory of data approximation (discrete modelling, in metrology terms). The main points are that, in practice,

1. discrete modelling is characterised by a search over a space of models;
2. the "best" model is that for which the associated error (loss) function is minimised;
3. model validation needs to test that
   (a) the underlying metrological assumptions,
   (b) the model space and
   (c) the error function (and possibly its approximation)

   are consistent with the metrological requirements.

The problem statement above is adequate for *discrete models*, where $F$ is a algebraic model of its inputs. For *continuous models*, which include explicit differential terms (e.g., the formulation includes a "rate of change"), it is also necessary to specify *problem data* such as

- the problem domain;

- the boundary conditions;
- the initial conditions, if $F$ contains a term of the form $\frac{\partial y}{\partial t}$

We use the term *modelling* to be the totality of all operations associated with

1. choosing a measurement strategy to determine the form and parameters of $F$;
2. determining a suitable functional form for $F$;
3. estimating parameters $\mathbf{a}$ given data $\mathbf{x}$ and $\mathbf{y}$;
4. estimating associated functions $G = G(\mathbf{y}, \mathbf{x}, \mathbf{a})$.
5. estimating the uncertainties associated with the steps above.

> *Example* Estimable functions in regression include
>
> - $G = F$, where $F$ is evaluated at a set of points,
> - $G = F^{-1}$, where $F^{-1}$ is the inverse of $F$ required when evaluating the inverse of a calibration curve. This might occur, e.g., in temperature measurement, where the temperature scale is determined by fitting voltage ($y$) PRT (Platinum Resistance Thermometer) data to given temperature ($x$) fixed points but, in use, we require temperature ($y$) for a given recorded PRT voltage $x$.

We defer discussion of the form of $F$ and methods for estimating the parameters $\mathbf{a}$ until later. At this stage, it is sufficient to note that a reliable model of the behaviour of the errors in the input quantities is required as input to the modelling process, and that all features of this process must be validated if the overall process is to be successful.

## 1.3 Model validation requirements: high-level

The case for validated metrological models is founded upon three wider requirements, as described below.

### 1.3.1 Traceability to standards

Valid metrological results must be traceable to standards. Software typically forms part of the traceability chain. Tests exist for the analysis of methods and software, but there is often little guidance as to *which* model validation tests to use and *how* to use them effectively. This is true even though the software tools are (arguably) becoming easier to use, and particularly now that many appear in non-specialist packages such as spreadsheets.

### 1.3.2 Quality certification

The growth in certification schemes and Quality Management Systems like ISO 9001 [65] has accelerated moves towards *formal*, *objective* quality requirements, rather than expert, though *subjective*, opinion. Measurement traceability (the audit requirement of metrology) is increasingly dependent on validated modelling and (measurement, analysis and reporting) procedures, in addition to traditional physical standards.

### 1.3.3 Improved uncertainty estimation

One of the key drivers for progress in metrology is the need to reduce the uncertainty of the measurement information. By definition, this requires good modelling, to push the boundary of what is "known", i.e., *characterised*, ever further, hence reducing the remaining uncertainty [64]. Conversely, characterisation needs to be well-founded so that spuriously low uncertainty estimates are avoided. Therefore good uncertainty estimation is associated with good (i.e., validated) models.

## 1.4 Difficulties in modelling

Deficiencies can occur in respect of the data, the problem formulation, the algorithm design and the software implementation. These are discussed in turn.

### 1.4.1 The data

Note that this includes both its *quantity* and *quality*. Quantity is governed both by the *number* of data elements and by its *distribution*. Circle fitting is a well-known example. At least three distinct points are required. If the points lie on a small arc, the traditional parametrisation (centre co-ordinates and radius) is ill-conditioned and the individual parameters are highly correlated with large uncertainties [7]. Even if the points are (apparently) well-distributed, a finite sample will generally underestimate the form (systematic) error; in pathological circumstances, it could estimate it as zero! [35]. With regard to *quality*, the errors should be tightly-controlled and/or well characterised. The problem is reduced by better metrological awareness and obtaining more *representative* data, which adds to costs.

### 1.4.2 The problem formulation

Often, users have an intuitive understanding of the data and its errors, and this needs to be incorporated in the estimation problem. It often helps to prepare a formal *user requirement document*, emphasising objectives, assumptions and constraints. This is supported and informed by documentation, model validation and, possibly, an intelligent user interface.

### 1.4.3 The algorithm design

Previous deficiencies related to the conceptual model and the data. When considering the algorithm, some of the focus of modelling control shifts towards the tools. Only low-level fine-tuning remains, which is heavily constrained by implementation details. This stage is the "last chance" for validation. Whether formal or not, a *functional specification* is the outcome.

### 1.4.4 The software implementation

This can fail because of the presence of

- unstable methods;
- logic errors;
- inadequate exception handling;
- language errors due to the developer and/or the compiler.

Many of these are subtle and can persist after gross errors have been removed. Numerical stability and correctness in some statistical procedures, including regression, have been considered [20].

## 1.5 Model validation requirements: low-level

As viewed by the metrologist, models are becoming more comprehensive and hence more extensive validation is required to

1. test whether a given model is *fit for purpose* in the sense of providing reliable metrological information;
2. test whether increased complexity in $F$ and/or the model of the errors adds to or detracts from the metrological information determined by the model;
3. select a suitable model from a number of candidates;
4. indicate improvements to a model where it is found to be inadequate;
5. indicate the domain of validity of the model;
6. assist in choosing a measurement strategy.

In this report, the term *weak* model validation is used to address only requirement 1 above. It corresponds to what, in the statistics literature, is termed a *goodness of fit* test, where a hypothesis is tested and no alternative is specified.

By constrast, *strong* model validation also addresses (at least one of) requirements 2-6 above. Requirement 2 is defined as *estimator validation*; see chapter 2. Requirement 3 is defined as *model* *selection*; see chapters 2 and 3. Requirement 4 relates to the iterative nature of model building, which is best seen as a design process with feedback loops informed by strong model validation.

> *Example* Feedback is required when determining a model (and associated measurement strategy) for the UK's new primary CMM, which is based on multilateration principles [33]. In this example, use of naïve models without sufficient analysis leads to ill-posed problems, resulting in rank-deficiency and the inability to identify all parameters in the model..

In other cases, poor models suffer from unnecessary ill-conditioning and hence loss of precision of some of the estimated parameters, typically because the model fails to take account of all the metrological information available and/or to specify what is missing.

Requirement 5 recognises that, for practical, nontrivial problems, it is infeasible to design a model which will work for all possible input data.

> *Example* The classical estimator of location of a univariate data set is the sample arithmetic mean. However this estimator is not always appropriate, e.g., if the data distribution is skewed and/or the data contains outlying *contaminants*. Either a trimmed mean or the sample median is a more robust estimator [70]. Being more robust, their practical *domain of validity* is greater.

Requirement 6 addresses an inverse problem, namely, given the limitations of the model, determine the data that is needed to maximise its usefulness. Examples include choice of data for form assessment of workpieces [79].

Note that metrologists rarely need to *validate* the continuous models in respect of the *problem data*, such as boundary conditions. However, metrologists frequently need to test the sensitivity of the model to errors in the initial or boundary conditions, material constants and/or the geometry of the boundary. This forms the basis of a technique for model selection which is based on the assumption that, given a set of models which are otherwise equivalent, choose that model which is least sensitive to errors in the problem data.

> *Example* Deformation of a pressure vessel due to applied pressure depends on problem data such as the vessel's material properties (Young's modulus, etc.). Model embedding can be used to test whether the result is sensitive to assumptions such as omission of temperature-dependent terms. Furthermore, the model can be solved by several variants of common techniques such as finite element analysis, boundary element analysis, etc. Error analysis can be used to predict/bound the errors associated with each method.

## 1.6 Fitness for purpose

With any form of problem approximation, such as that described in Section 1.2, it is necessary to compare the uncertainty introduced by the approximation process to the uncertainty contributed by other sources. The latter provides a natural scale for assessing the importance of errors introduced by an approximation.

*Example* If the uncertainty contribution due to other sources is $u$ and the uncertainty introduced by the approximation process is $u/10$, the combined uncertainty (assuming the errors are uncorrelated) is $u\sqrt{1.01}$, so the difference is probably negligible and the simplified estimator is *fit for purpose*.

Other measures of *fitness for purpose* exist, providing various criteria against which to judge candidate models.

The fitness for purpose requirement implies, by definition, that model validation must be linked to the application, i.e., generic model validation is a contradiction in terms. However, generic model validation *techniques* exist and form part of the subject matter of this report. Model validation requires measures (metrics) of model success which are meaningful in the application context.

*Example* For a regression problem, the root-mean-square residual is an estimate of the noise in the data, which is a useful measure of metrological data quality. Furthermore, the metrologist might have a good idea of the size of the noise, perhaps through experience or knowledge of the error behaviour. Hence it makes sense to use this metric as (one of) the quantitative criteria of success..

In the context of fitness for purpose, consideration is needed of (i) the reasons for failure, (ii) the nature of the modelling objectives and (iii) the modelling risk.

# Chapter 2

# Methods for validating the conceptual model

Since the conceptual model is intimately linked with the physics/chemistry etc., of the metrological application area, it tends to be heavily application-specific in practice. However, there are some key generic principles which can be applied[16]:

1. Test that a known property is satisfied, e.g.,
   - coordinate system invariance,
   - a constitutive relation such as conservation of energy.
2. Look for *model embedding* e.g.,
   - simple models (asymptotic case/linearised/fewer terms/assume some errors are negligible);
   - this model;
   - complicated models (full nonlinear/extra terms/comprehensive error model).
3. Perform some sensitivity analysis [64, 13] to determine the influence of each assumption. Model validation activity can be prioritised according to the sensitivity of the model to its assumptions/"known" parameter values.

## 2.1 Testing of properties

Validating models against benchmark problems is one way of checking that the model has the correct type of behaviour. This is the approach taken by NAFEMS [74] whose objective is "To promote the safe and reliable use of finite element and related technology", which is realised by disseminating benchmark problems.

Typically in a physical process certain properties are known—it might be that energy is dissipated or mass is conserved. Different benchmark problems provide a variety of properties to test against. An alternative is to ask that the numerical scheme used preserves key properties of interest (in some cases this can be proved).

> *Example* For the model of the calorimeter (see section 2.3) the exponential decay of the solution is a well known property and hence it is sensible to choose a numerical method that also has this property.

As progress is made, the assumptions made at each stage, i.e., those relating to the

- data
- conceptual model
- discretised model
- software implementation

are tested, refined and justified. Furthermore, a "profile" of the model is built up during the process, e.g., with regard to the domain of validity (see section 1.3).

Since assumptions later in the process build on (refine) their predecessors, there is a natural interpretation of model validation as a risk management strategy. In risk management terms, hazard probabilities and, by extension, project lifecycle costs are reduced by "getting it right first time".

> *Example* Consider the problem of determining the free-field correction for a microphone. The following assumptions are made
>
> - the acoustic source is at a single point in space;
> - the microphone is mounted at the end of a semi-infinite cylinder;
> - the source emits spherical wavefronts with constant frequency and amplitude;
> - reflecting surfaces are known and are planar;
> - reflections are attenuated by a constant factor;
> - the governing equation is a Helmholtz second-order partial differential equation.
>
> Clearly, all of these assumptions can be tested. In particular, it is advisable to use some sensitivity analysis to determine the effects of relaxing any subset of the assumptions.

## 2.2 Examples of conceptual discrete models

### 2.2.1 Fitting geometric features to co-ordinate data

Co-ordinate metrology is concerned with comparing the shape and size of manufactured workpieces with their ideal or *nominal* shape and size. Often the nominal shapes are common geometric features such as lines, planes, circles, sphere, cylinders, cones, etc. A standard calculation is to find the best fit feature to a set of measurement data.

> *Example* Let this nominal feature be a circle in a plane. To perform this calculation it is necessary to define
>
> 1. a parametrization of the space of circles, e.g. by the two centre co-ordinates and radius $\mathbf{a} = (\mathbf{x_0}, \mathbf{y_0}, \mathbf{r_0})^{\mathrm{T}}$,
> 2. an error function $E(X, \mathbf{a})$ that provides an aggregate measure of the distance from the circle specified by $\mathbf{a}$ to the data $X = \{\mathbf{x_i}\}$, usually involving the $d_i = d(\mathbf{x_i}, \mathbf{a})$ from the data points $\mathbf{x_i}$ to the circle and
> 3. a method of minimising $E$ with respect to $\mathbf{a}$.

Modelling issues in feature fitting are concerned with:

> The choice of parametrization.
>
> The definition of the error function.
>
> The choice of algorithm to determine the optimal parameters.

The parametrization of geometric features involves the topology of the space of geometric elements. For elements involving the direction of an axis (cylinders and cones, for example) the model space is topologically nontrivial which means that more than one parametrization is necessary to represent all elements within the space.

Issues in model validation relate to the definition of the parametrisation(s) (determining the form of $F$; this requires mathematical analysis and peer review) and the choice of error function $E(X, \mathbf{a})$ (which is validated by analysis and review as before). Often approximate estimators are implemented and their behaviour has to be characterised (estimator validation).

The choice of estimation algorithm is a discretised model requirement. In practice, the application of the model to data (both representative and "extreme") provides overall validation of the model.

For more information see [7, 6, 18, 44, 45, 47].

### 2.2.2 Calculation of form error

Related to geometric element fitting is the calculation of form error of workpieces, a measure of how far the workpiece departs from its nominal form. In order to determine sensible measurement strategies it is required to develop models for the likely departure from form.

Modelling and model validation issues concern:

1. Choice of empirical modelling basis functions: polynomials, splines, Fourier series, wavelet resolution.
2. The number and nature of the basis functions required to simulate likely form error.

3. The distribution of likely form error within the space of error models. For example, should the coefficients of the basis functions be generated from normal or rectangular distributions.

The model validation techniques are based on sensitivity analysis and peer review, supplemented by *between family* validation for (1), *within family* model validation for (2) and testing using validation data for (3).

## 2.3 Example of conceptual continuous models

As an illustrative example we consider the flow of heat through a graphite block with an embedded thermistor. This is an example of a calorimeter used to measure heat (key to the Thermal NMSPU Programmme). The description here originates from the Ionising Radiation NMSPU Programme.

### 2.3.1 Physical description and assumptions

Thermistors are placed in a graphite core and surrounded by graphite blocks. These are in turn surrounded by polystyrene and this by a further graphite horseshoe with the whole assembly encased in polystyrene (see figure 2.1). The assembly is then irradiated and the temperature rise monitored from the thermistors in the core. From this temperature rise the radiation dose may be calculated. The aim is to model the heat flow through the system to reduce the error in the estimated heat rise from the radiation. A more complete description of the physical set up may be found in [71].

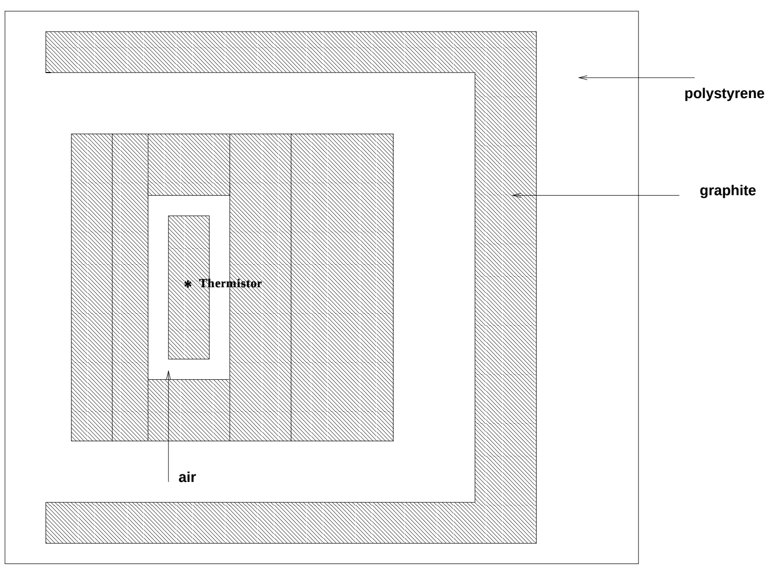


Figure 2.1: Diagram of set up for radiated calorimeter

In the calorimeter example governing equations for the flow of heat are required. The processes and related equations are well known: we need to consider heat transfer by conduction, convection and radiation (see for example [23]). There are several assumptions that may be made: for example convection and radiation is considered to be negligible and the heat flow is dominated by conduction.

Furthermore material properties have to be considered, for example each slab of graphite is treated as isotropic. An initial temperature profile needs to be assumed across the system and appropriate boundary conditions imposed—for example it might be assumed that the polystyrene inside the horseshoe surrounding the central slabs is perfectly insulating.

### 2.3.2 Developing the model

For the case of the calorimeter we consider heat flow by conduction in each slab in the main block. Around the block we impose the condition that there is no heat flow, i.e., the polystyrene layer is assumed to be perfectly insulating. Between slabs and for the air gaps we assume a heat transfer proportional to the temperature difference. With these boundary conditions we solve for each slab the linear *Heat Equation*

$$F(u, [t, x, y], \kappa) = \frac{\partial u}{\partial t} - \kappa \Delta u = 0,$$

where we have further assumed that the materials properties $\kappa$ are neither time nor temperature dependent. Material properties (such as conductivity and specific heat) may either be taken as 'book' values or fitted as parameters from separate experiments.

It is often important to account for the uncertainty in the material constants and similar parameters, and to consider their influence on the solution and their contribution to the uncertainty of the measurand.

It is well known that solutions to the *Heat Equation* decay exponentially fast to a constant value (that depends on the initial condition) [23]. Plotting this decay, we can verify the appropriateness of the model comparing experimental results from the thermistor.

## 2.4 Practical validation of conceptual continuous models

In the discrete modelling scenario, the conversion from the abstract conceptual model to a discretised model (suitable for implementation in algorithm form on a computer) is relatively trivial. However, in the case of a continuous model, the mapping from the abstract (mathematical) domain to a numerical algorithm is not always obvious and requires some design.

For formulating or choosing a conceptual model of a physical process there are a number of useful principles to follow:

1. Consider the requirements: Be clear what phenomenon is to be modelled and what outputs are required. This will aid in selecting the level of detail required in the model.

2. Decide on criteria (qualitative and/or quantitative) to be used in determining a model's fitness for purpose. A qualitative criterion has either a pass or fail result. A quantitative criterion may be used for ranking candidate models.

3. Consider the assumptions that were used for deriving the model. If the model is derived from first principles, keep a record of assumptions made.

4. Look for features in the model that might either be redundant or added to improve the model.
5. Check that qualitatively the model behaves correctly, e.g., it has the correct behaviour for limiting cases.
6. Check that quantatively the model behaves correctly, e.g., by using simple test data.
7. Compare the model to others (between-family and within-family) for the same or similar situations.
8. At all times compare to experiments where possible.

# Chapter 3

# Methods for validating the discretised model

As discussed in [16], the decisions can be classified into macro or micro decisions. Using an approach motivated by object-oriented programming, the macro decision is associated with selecting a technique *class* for a given *instance* of a problem.

> *Example* In acoustic metrology, displacement of a plane microphone diaphragm could be modelled using isoparametric finite elements in 2-D. The corresponding micro decisions (relating to class implementation) might relate to the discretisation type (uniform versus adaptive), choice of basis, e.g., quadratic shape functions, etc.

The same approaches above can be followed, namely,

- testing that properties are satisfied
- looking for model embedding. Note that in the case of discretised models, this can include the basis function degree, choice of tolerances, coarseness of discretisation grid (in space and/or time), etc.

Error analysis (of model and data) and visualisation are both very helpful when validating the discretised models.

The treatment above focuses on the problem of discretising continuous models. However, "discrete" conceptual models also require some conversion to a form which is suitable for implementation as an algorithm, particularly when the model functions are derived from physical considerations [38] or are used in measurement strategy validation [33].

> *Example* If we continue to develop the microphone correction conceptual model from chapter 2, we might decide to employ the following numerical scheme for the discretised model
>
> - Use a Boundary Element Method (because part of the boundary is at infinity) with $M$ nodes

- Represent the solution as a series of Hankel functions with $N$ terms

Validation techniques would include comparison with an equivalent Finite Element Model, changing $M$ and $N$, the number of nodes on the boundary and terms in the series, respectively.

Of course, these techniques need to be supported with a reliable software implementation, by means of module and interface testing, and the use of high-quality library software for evaluating special functions, performing numerical quadrature and solving dense linear systems.

Methods based on multiresolution analysis, e.g., multigrid methods for solving continuous models[15], or wavelet analysis of discrete data [39] have some built-in validation, particularly of the level of discretisation.

## 3.1 Stability of the discretised model

From a continuous model we discretize in order to get a numerical scheme. Traditionally numerical schemes rely on results such as

$$\mathrm{C}onsistent + \mathrm{S}table \Longrightarrow \mathrm{C}onvergent.$$

Essentially,

- a *consistent* scheme will approximate the set of equations as mesh sizes tend to zero.
- a *convergent* scheme is one which tends towards the solution of the full continuous problem as more resources (finer discretisation, more model terms, etc.) are provided. See also Appendix C.

There are several forms of stability that can be considered (see for example [56, 57]). The stability conditions ensure technical properties hold which control the behaviour of the algorithm as the step sizes are reduced. This leads to the most basic methods of validation of a numerical scheme which are

*halve the discretization and re-solve*
and,
*halve the tolerances and re-solve*

This form of validation (in effect, testing that model embedding occurs) must be considered as one of the most important as it can be applied to whatever problem is being solved and can be used to estimate the error in the numerical solution.

Many codes have error estimation techniques built in which might be used in addition for error control. Such techniques essentially rely on an approximation of the residual error. A popular method for estimating this is to use two methods of different order (i.e., with different accuracies) for example the MATLAB routines ode23 and ode45 are examples of this. As part of the validation process such information should be examined.

Another traditional approach to the validation of numerical schemes is to consider test problems. There have been a number of attempts to produce test data sets. For initial value problems in particular there are a number of test datasets available [67, 58, 24].

These types of validation are useful for verifying that the methods used for solving the continuous method can produce the right type of results. It is dangerous however to rely exclusively on these benchmark tests as there is no guarantee that the codes will produce the same accuracy for other problems. Such benchmarking of codes can lead to codes being written to perform well on those benchmark problems.

For PDEs, numerical methods are validated against the classical problems in the class of equations being solved. Thus, for example, for parabolic PDEs the heat equation used for modelling the calorimeter is a typical test problem. In more complicated systems the ability to reproduce the solution over a range of parameters (for example picking out bifurcation points accurately) is used for validation purposes.

This has led in recent years to validation against “known” solutions or test cases from other respected code. For the calorimeter it would be possible to test results from a Finite Element Package against, for example, those from a Finite Difference approximation.

In a time-dependent (or time-like) form of problem for the asymptotic behavior it can be instructive to validate the time dependent case against a time independent case.

## 3.2 Examples of Discretised Continuous Models

### 3.2.1 Heat flow in a calorimeter

Given a valid conceptual model of the calorimeter (see Section 2.3), a discretised Finite Element Method (FEM) model can be developed. The macro decision was to choose FEM rather than, say, a Finite Difference Model.

Several micro decisions need to be made concerning order of basis functions, mesh discretisation, etc. These decisions are guided by stability considerations from Section 3.1.

In Figure 3.1, we see the effect of a poor choice of stepsize when a Finite Element Method is used, in that a linear stability criterion has not been met in the second plot, which exhibits spurious local variations, reflecting the fact that the time steps and space grids are unbalanced.

The heat flow equation can also be compared with the steady-state heat equation. This is equivalent to comparing $U_s$, the solution of Laplace's equation $\Delta u = 0$, and $U_d$, where

$$U_d = \lim_{t \to \infty} u(t)$$

and $u(t)$ is a solution to the heat equation.

## 3.3 Practical validation of discretised models and numerics

Validation of discrete models can be achieved by analytical methods and by direct numerical computation. This assumes that the numerical implementation of the method is correct and that robust stable algorithms have been used. If numerical "glitches" occur in solving a practical problem it should be remembered that part of the difficulty might be caused by software errors.

1. Be clear on the data required from the numerics.

2. Check that the numerical results confirm physical intuition.

3. Validate the numerics against the conceptual model. In particular take advantage of any known solutions.

4. Validate the numerics against experimental data, if necessary at different parameter values or geometries.

5. Check the convergence of the numerical method.

   - Refine the spatial/temporal mesh and resolve the problem. Be aware that linear stability results often place restrictions on mesh densities.
   - Test convergence by increasing the order of the method used for the problem.

   These techniques allow estimates of the

   - rates of convergence (which can be compared to theory)
   - error

   to be found. Convergence rates can be viewed graphically.

   In particular look for the appearance or disappearance of small scale structures in the solutions as these are often indicative of numerical instabilities.

   In general spurious solutions will diverge as the mesh is refined.

6. Test that specific properties in the numerical solutions are maintained. For example there might be symmetries in the system that can be verified or there might be a conservation property that is important.

7. Verify the chosen method against other numerical simulations either from different packages or from published results.

8. Be aware of the scales that you are trying to resolve and choose appropriate methods and techniques.

   > *Example* Electromagnetic modelling of high-frequency microwave radiation require short time steps to resolve temporal behaviour. Similarly, fine mesh discretisation might be required for high frequency vibration/acoustic phenomena.

9. Examine the stability of the solution to changes in parameters, boundary conditions and initial conditions. Note that a change in stability may be a truly physical phenomena or it may be an artifact introduced by the discretisation.

10. Where possible, plot the solutions fully as often numerical “glitches” are best recognised by graphical means.

## 3.4 Graphical methods

The following plots are recommended, since they provide graphical evidence (both quantitative and qualitative) in support of model validation:

- Plot of root-mean-square residual $\sigma$ against number of terms in a model family *Use*: Looking for saturation to a “minimum”. This can be repeated across model families, e.g., polynomial, spline, Fourier series, etc., which are “equivalent” empirical functions for fitting univariate data.
- Plot of each covariate against the residuals. *Use*: Looking for systematic behaviour not already modelled.
- Plot of coefficient of successive terms of a model, e.g., Chebyshev coefficients against polynomial degree. *Use*: Looking for saturation to a minimum.
- Plot of Cook’s distance, which is a measure of the influence of each observation $x_i$ in $X$ [70]. *Use*: Assessing the quality of the data, effectiveness of weighting strategy. Ideally, each $x_i$ will have the same Cook’s distance.
- Plot of residuals against the independent variable $\mathbf{x}$. *Use*: Looking for local changes in variance, assessing the distribution of the residuals.
- Plot of sorted residuals against distribution scores. *Use*: Assessing the distribution of the residuals; the better this plot of residuals can be fitted by a line, the better the residuals agree with the assumed distribution.
- Plot of the predicted model. *Use*: For discrete models, “wobbles” are symptomatic of numerical difficulties resulting from overfitting (use of too many terms in the model) or use of unstable fitting algorithms. For continuous models, the usual reason is failure of the numerical scheme to meet the stability criteria. Note that sometimes wobbles or discontinuities are inherent in the problem, e.g., if a continuous model includes ‘point’ loads or impulse terms.

Many other plots are available in statistical packages [70] and recommended in the literature for exploratory data analysis, etc. [31].

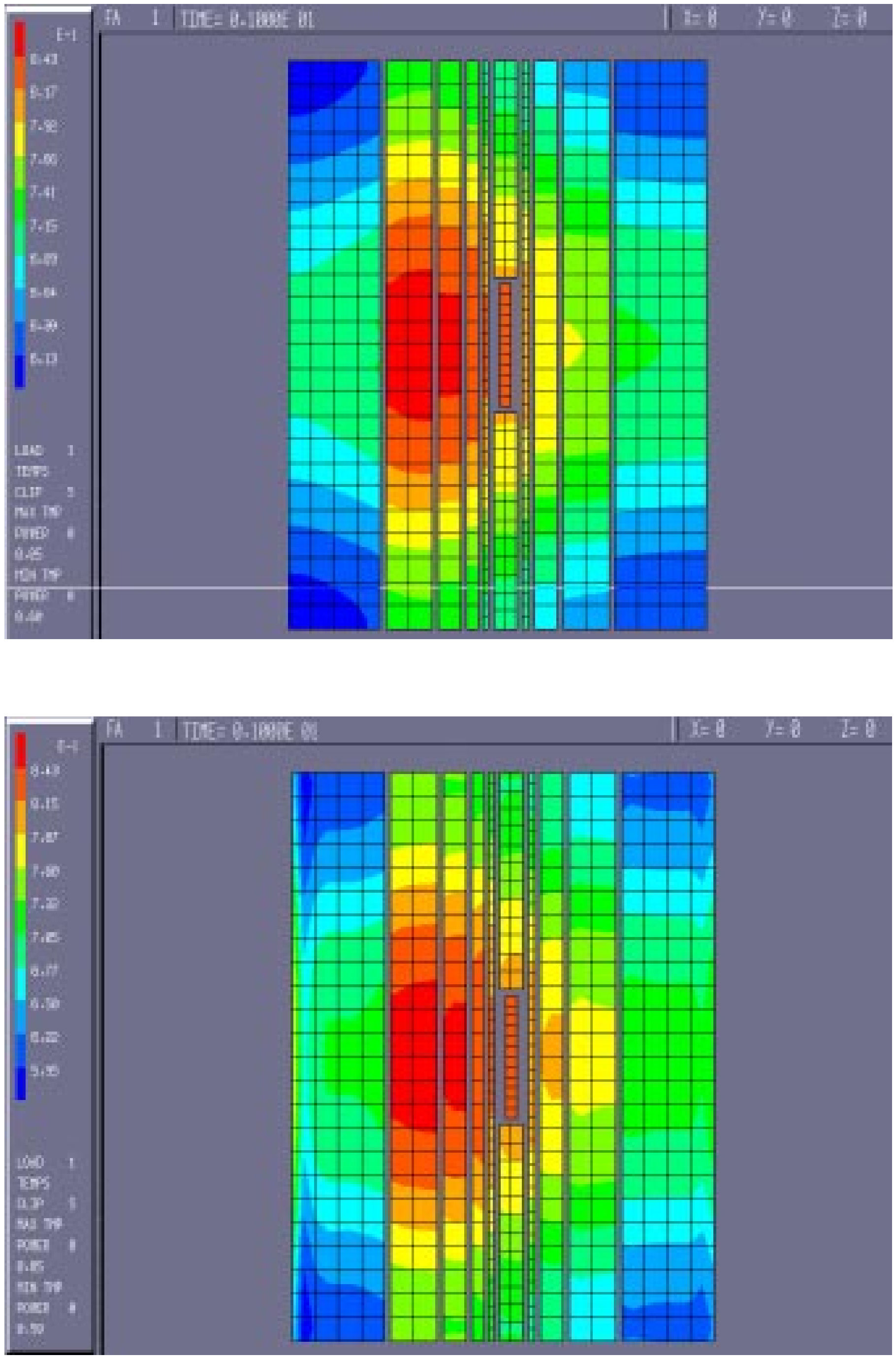

Figure 3.1: Heat distribution through the radiated calorimeter at time $T = 1$. By contrast with the first plot, a linear stability criterion is not satisfied in the second and spurious “wobbles” appear.

# Chapter 4

# Methods for validation of regression models

Regression models are common in metrology for relating a measured output in response to changes in a (set of) control variables, where this relationship is assumed to have an algebraic form. Hence this section examines the modelling and validation issues in more detail.

The regression *decision tree* has many options (e.g., model function form, nature and use of error models). (Regression) model validation is usually concerned with providing *statistical* answers in support of model selection.

Initially, high-level *between-family* decisions are required, which in turn motivate low-level *within-family* decisions. For instance, sequential decisions might be to use

1. an empirical model (as opposed to one based on the physics such as the Edlen equation [11, 12]);
2. a polynomial;
3. least-square norm;
4. degree 5.

Note that decisions 1) and 2) may be classed as high-level decisions, and 3) and 4) as low-level decisions.

Model selection is supported by software tools, which provide evidence in favour of various hypotheses, but it is generally agreed that decisions are best left to the user.

*Testing of (regression) software* differs in that it comprises *mathematical* tests on the *implementation* of the selected model. The key aim is *generic correctness* of the regression tools, with regard to *reliability* (the software tools perform to specification), *robustness* (they perform adequately for extreme cases) and *efficiency* (they use resources well, particularly data precision).

There is a special form of estimator validation which lies somewhere between traditional model validation and software testing, in that it seeks to address whether an *approximate estimator*, e.g., one obtained by linearising the optimal estimator and hence solving a simplified computational problem, remains fit for purpose.

## 4.1 Issues governing the quality of a regression model

With particular reference to regression, the following issues govern the quality of a regression model.

**The data** Poor data limits what can be achieved (see Section 1.4). Also, it must be assumed that most of the variability in the data can be ascribed to (a linear combination of) known independent (control) variables. We note in passing that sometimes a linear combination has improved properties (e.g., orthogonality) for modelling and this feature is exploited in regression techniques such as Principal Components Regression (PCR).

If the data is poor, even the best model will provide large uncertainty estimates for each regression parameter.

**The error model** When developing the error model, the following need to be considered (at the very least):

1. which variables are subject to error;
2. whether these errors are correlated;
3. whether the errors are heteroscedastic (have constant variance);
4. analysis of systematic and stochastic components, and
5. size of the errors, possibly characterised by a Signal to Noise ratio or, better, by a covariance matrix.

If more than one variable is subject to error, generalised distance regression (GDR) [46] is appropriate. Firstly, the error covariance matrix $V$ of the input parameters should be estimated; it is often necessary to use information from other sources. Let $X^{\mathrm{T}}V^{-1}X$ be the Gauss-Markov estimator of the parameters. If, as is often the case, the errors are uncorrelated, $V$ is diagonal, so errors weighted by the inverse of the square roots of the diagonal values are homoscedastic. If the errors are *unbiased*, the Gauss-Markov estimate will be unbiased [82]. If the errors have a non-Normal distribution, least-squares ($l_2$-norm) estimation is no longer the best linear unbiased estimator (in the context of maximum likelihood estimation) and another fitting criterion such as $l_1$ or $l_\infty$ might be more appropriate.

In practice, it is difficult to distinguish between a heavy-tailed distribution and a normal distribution with a small number of outlying contaminants. The two favoured approaches are

- Reject the outliers [9] and proceed with the normal assumption;
- Accommodate them by employing a robust estimator [75, 86, 84].

**Additional/prior information** Such information (e.g., from experience and handbooks) is not only useful for defining the error model [64]. Users often require the regression model to satisfy simple constraints, e.g., that strain never decreases as load is increased. This amounts to a *shape constraint* on the fit. More generally, users prefer simple fitting functions, which are much easier to interpret. This shape can be lost if the data is poor and the model functions are not *designed* to satisfy the constraints.

**The functional requirements** These help to define the overall regression aims. Requirements include:

- **data reduction** The sole purpose of the model is to interpolate the data, using fewer parameters than there are data points. Hence the user has full flexibility to choose a functional form for $F$ which provides reliable estimates of the underlying "signal".
- **parameter estimation** The model is semi-physical at least, reflecting the presence of physical parameters in the model. Further, there are often constraints on these parameters, e.g., mass cannot be negative. Consequently, the space of models is reduced and some compromises are introduced, i.e., physical reality (e.g., assuming there is no phase change, temperature is a nondecreasing function of pressure if the volume of gas is maintained constant, or the resistance of an electrical component cannot be negative) versus closeness to the data.
- **system identification/parameter reduction** In such problems, there is considerable (near-) redundancy in the parameters, and part of the task is to find a linear combination such that a lower-dimensional regression model is adequate (e.g., see [68]). The singular value decomposition (SVD) [55] is a very useful tool for such problems.
- **evaluating estimable functions of the parameters** Estimable functions include the regression model function itself, its analytical derivative, etc.
- **estimating uncertainties of parameters and derived quantities** This is the subject of much metrological activity and guidelines have been published [64].

## 4.2 Regression diagnostic plots

Viewing the results of a regression plot is helpful for validating models, and particularly for model selection. In the following example, the objective was to model the nonlinear response of a piezoelectric transducer (PZT) used in measurements of displacement $d$. The response $V$ of an ideal PZT is linear over its operating range, i.e., $V \propto d$.

Figure 4.1 contains a plot of root-mean-square residual against polynomial degree. Since the linear response has been removed from the data, the main improvement is to be found when adding a quadratic term. However, it is also clear that more improvement is possible and that the RMS residual does not settle until about degree 20.

The evidence that a cubic fit, say, is insufficient can be seen in Figure 4.2, where it is clear that the residuals display systematic behaviour. Such behaviour persists as the degree is increased until about degree 22, when the residuals begin to "appear random", see Figure 4.3.

Figure 4.4 displays the symptoms of *overfitting*, when there are too many terms (i.e., fifty-one terms) in the model. Notice that the fit is beginning to wobble towards the ends of the interval, and that this is made even more noticeable in the residual plot when the $95\%$ confidence limit is seen to grow dramatically by comparison with Figure 4.3.

Figures 4.5, 4.6 and 4.7 provide further evidence (using the AIC, AICc and SBC criteria, respectively; see section B.1.2) in support of the recommendation that, for this data set, the "best" model from the family is a degree 22 polynomial. Note that each criterion has been scaled by a constant factor for convenience, and that the minimum is generally well-defined.

Figure 4.8 indicates that the optimum degree, using the $C_p$ criterion (see section B.1.2), is approximately 35 (the criterion has been scaled so that unity is best). This estimate of degree is anomalously high, although we note that $C_p$ is almost constant over an interval which includes degree 22. We conclude that it is best to weigh up the evidence from several criteria.

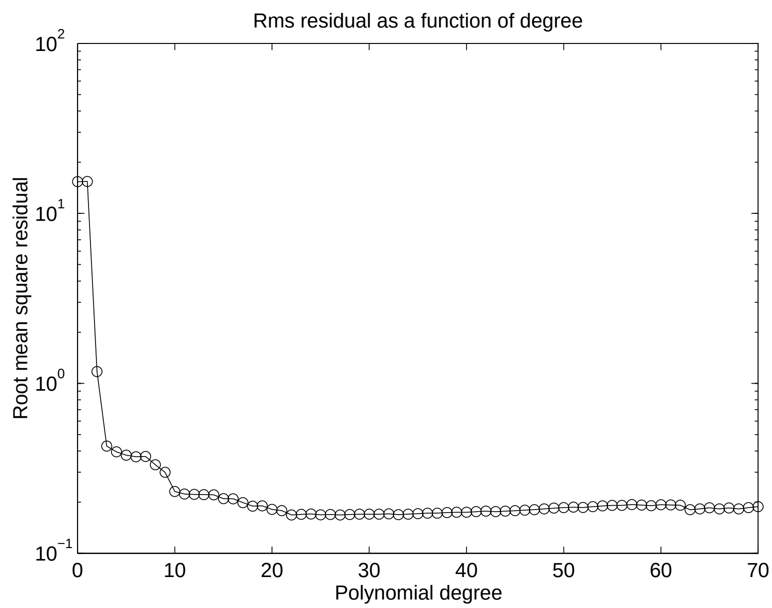


Figure 4.1: PZT data: Root mean square (RMS) residual as a function of polynomial degree

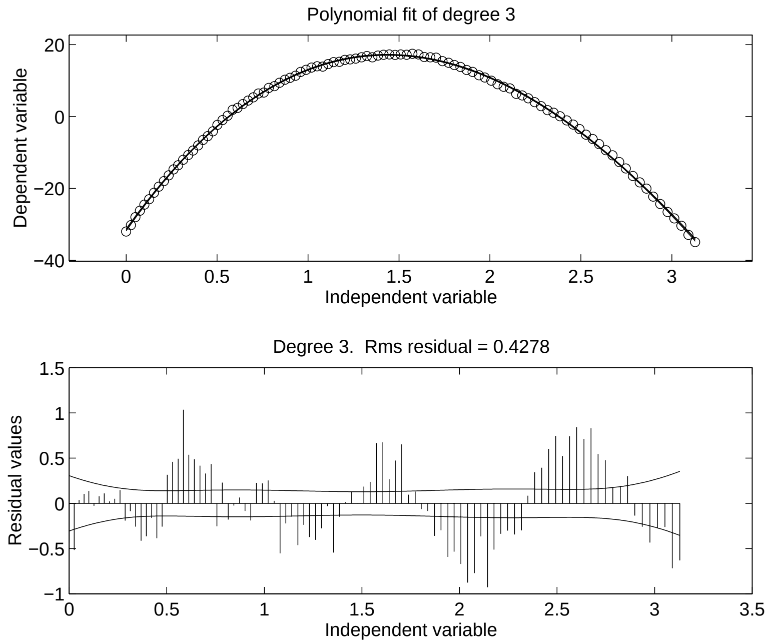


Figure 4.2: PZT data: Degree 3 polynomial fit, also showing residuals and confidence limits on the fit and residuals

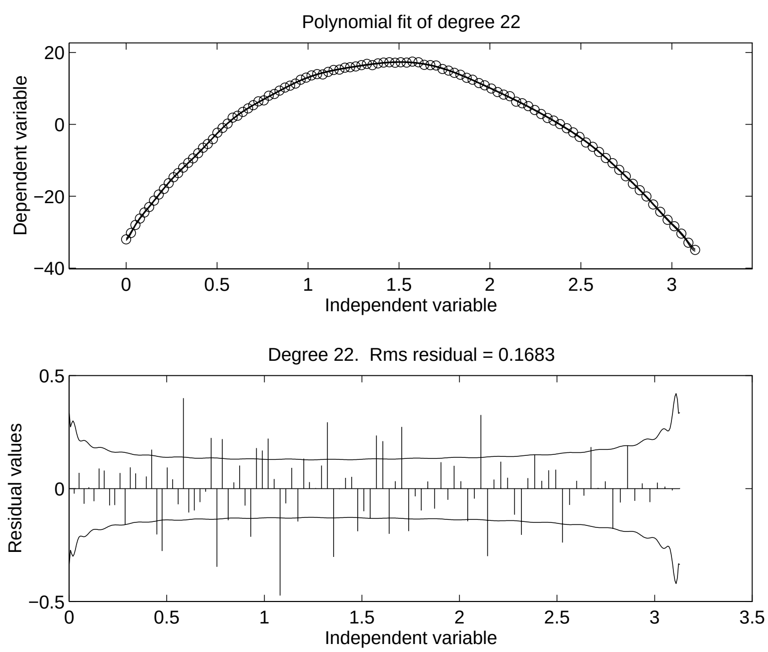


Figure 4.3: PZT data: Degree 22 polynomial fit, also showing residuals and confidence limits on the fit and residuals

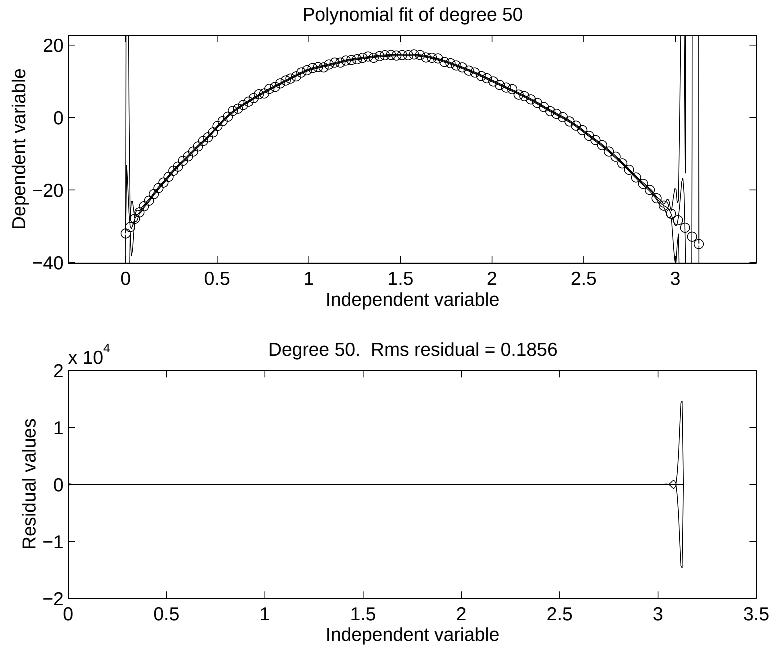


Figure 4.4: PZT data: Degree 50 polynomial fit, also showing residuals and confidence limits on the fit and residuals

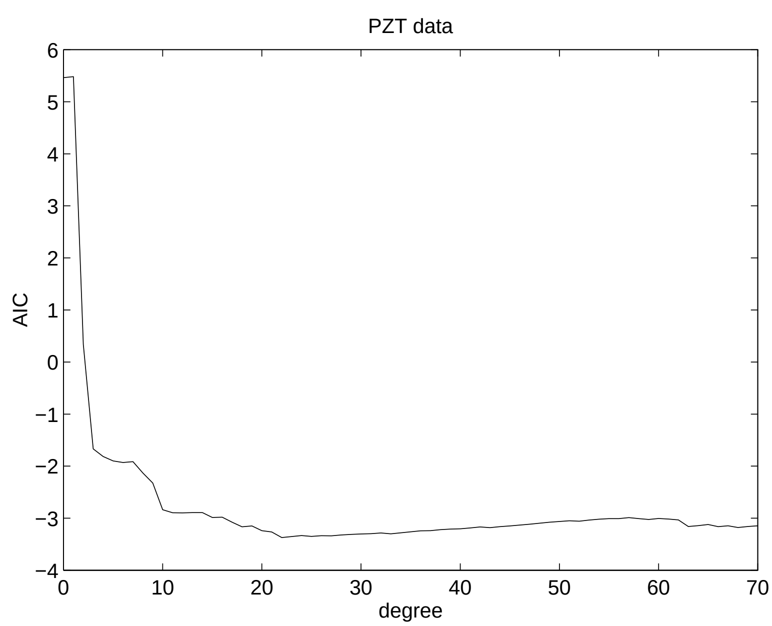


Figure 4.5: PZT data: Akaike Information Criterion (AIC) computed for polynomial degrees 0 to 70 inclusive

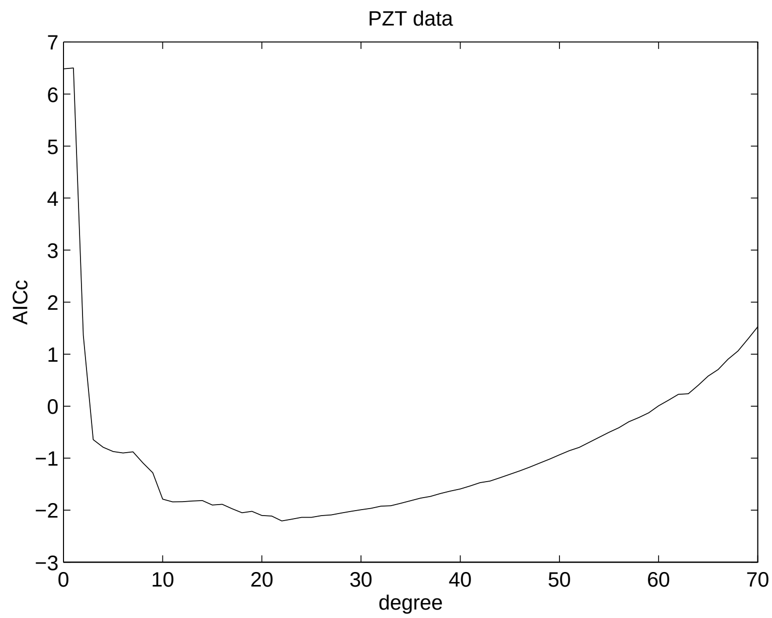


Figure 4.6: PZT data: Corrected Akaike Information Criterion (AICc) computed for polynomial degrees 0 to 70 inclusive

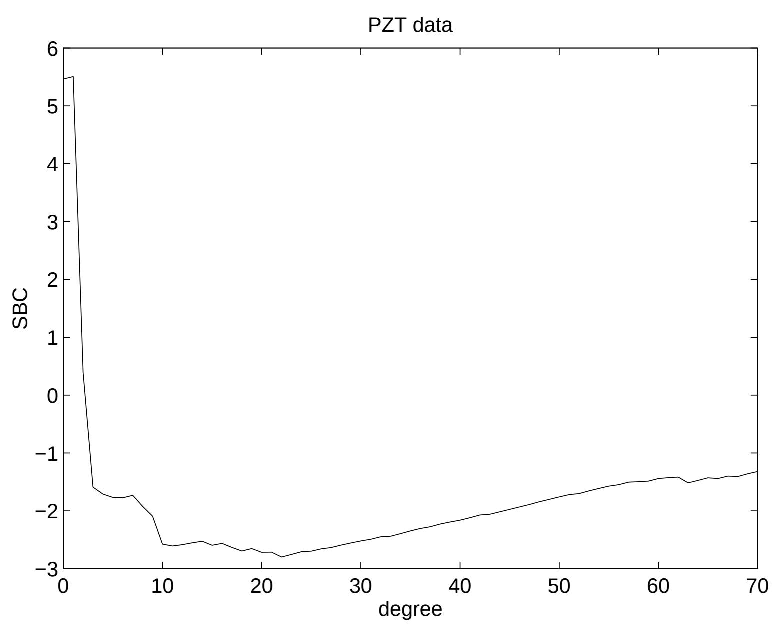


Figure 4.7: PZT data: Schwartz Bayesian Criterion (SBC) computed for polynomial degrees 0 to 70 inclusive

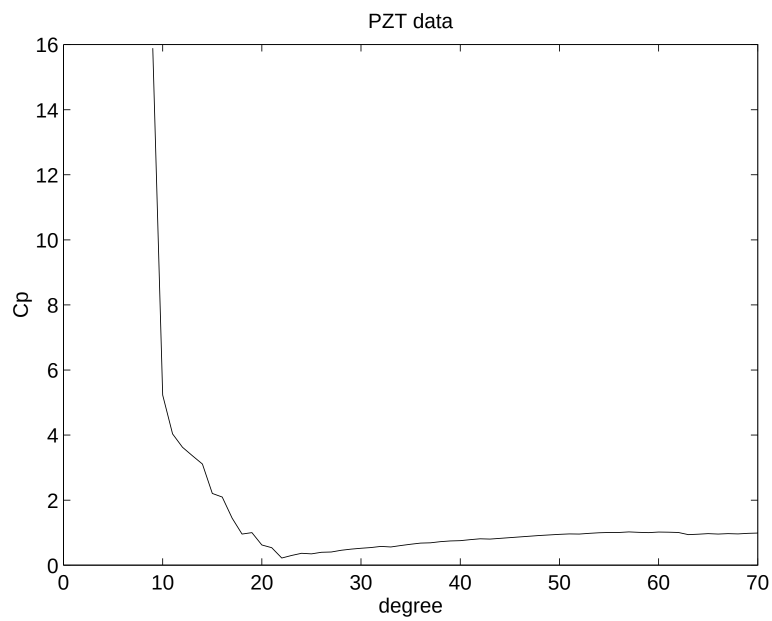


Figure 4.8: PZT data: Mallows $C_p$ criterion computed for polynomial degrees 9 to 70 inclusive

# Chapter 5

# Selected applications of model validation

General frameworks for model validation are worthless unless they find practical application. In this section, a number of nontrivial examples are presented in which model validation has played a crucial part.

In the metrology-based examples, model validation has been applied to add credibility to information derived from models. The statistics used have generally been classical, though supplemented as needed by robust and non-parametric approaches.

The non-metrology examples have been chosen

- to indicate the “state of the art” of model validation more generally in applied science, and
- to point to key projects, some of which are designed to provide application-specific guidelines

## 5.1 Examples in metrology

### 5.1.1 Retroreflector calibration

**The Model** Retroreflectors are key optical components of the new primary CMM being designed at NPL. The form error of the nominally spherical surfaces (and of the nominally spherical reference wavefront at the calibration stage) contributes to the uncertainty of coordinates measured by the new CMM. The objective is to model these form errors so that corrections can be made for their effects, such that the resulting coordinate estimates have lower uncertainty. The (discrete) form error models are complicated, with additional nonlinearities being introduced by misalignment effects, among other factors.

**Validation** By means of careful perturbation analysis, the effects of linearisations such as $\cos\theta = 1$ for $|\theta| \ll 1$ and truncation of Taylor series can be estimated and compared with other error sources. In particular, the domain of validity of the linearised models is indicated [33].

### 5.1.2 Coordinate measurement using multilateration principles

**The Model** The retroreflectors form part of a larger system, in which laser displacement measurements (from target retroreflector to tracking station) are combined with suitable calibration information by means of a model [33]. The model outputs include estimates of the coordinates (and the associated standard uncertainty) of the probe centre when a point is measured. The uncertainty requirements of such a system, viewed as the primary UK CMM, are very challenging. Also, it is not possible to identify all the parameters of the system by displacement measurements alone, hence additional calibration data is required.

**Validation** The *system identifiability* issue was resolved by analysis of the singular vectors which form a basis of the null-space of the system matrix. As a consequence, it was possible to devise measurement strategies such that the additional data would add materially to the data available to the model. Extensive simulation and testing was performed to validate the resulting measurement strategy, to the extent that a *simulated instrument* was produced as a by-product. This needs further validation with measured test data.

### 5.1.3 Use of virtual features to design CMM sampling strategies

**The Model** The main use of a CMM is in part inspection, particularly for geometric dimensioning and tolerancing. Thus the measured coordinate data (e.g., that obtained from a system like that above) is input to software which

- fits a mathematical function, which could be a geometric feature such as a circle, or a general free-form spline surface, and
- performs further processing based on the parameters of the fitted mathematical function and on the deviation of the measured data from this functional representation.

The uncertainty in the measured coordinates is but one contribution to the overall measurement uncertainty of the task. Other contributions include the use of a finite sample of points, such that estimates of form error size are generally *underestimates*, since extreme points will generally be missed when a fixed sampling interval is used. This underestimation of the maximum form error is *unsafe* in the context of tolerancing applications.

**Validation** Numerical simulation can be used to estimate the degree of negative bias in the form error size estimate, together with estimates of the uncertainty of this bias. The quality of such estimates depends critically upon the quality, particularly in terms of how well they represent typical form error on actual testpieces, of the simulated *virtual features* (= nominal feature + form error function + stochastic measurement error). The simulation was based on fine sampling of relatively few testpieces. Hence model validation is an important issue.

Further model validation techniques used include

- Use of orthogonal basis functions (such as the Chebyshev basis for a polynomial-, and sinusoids for a Fourier series-representation);

- Consequently model embedding (between-family validation) is possible;
- In some instances, both Fourier series and polynomial models could be used, which enabled across-family validation to be carried out;
- Plots of residuals were used to provide visual evidence that the residuals of the chosen representation were random;
- Plots of root-mean-square residual (equivalently, mean-square-error) as a function of number of terms was used to indicate when this criterion saturates.

### 5.1.4 Parametric Error Models of CMMs

**The Model** In practice, a conventional CMM will exhibit position-dependent systematic errors in the reported coordinate measurements. That is, the scale readings will be biased, and that bias will depend on the probe's position in the working volume. Full calibration (Class 1 checking) by the manufacturer will provide input to an error correction model which seeks to remove that bias. However, that model is static, and the machine will drift in response to influences such as changes in its environment, resettings and normal "wear and tear" in use. Hence periodic reverification is required, using measurements of *Class 2* artefacts such as ball plates.

A mixed physical/empirical model has been developed [38], wherein the errors are decomposed into 21 components due to kinematic errors such as axis roll, lack of axis squareness, etc. Each component kinematic error is then represented as a low-degree polynomial. NPL has suggested an alternative, fully empirical model [36].

**Validation** Clearly, the latter formally includes (to first order) the kinematic error model as a special case (hence embedding is possible in principle, although the parametrisation is very different, and so the models can almost be viewed as independent and belonging to separate families). Note that

1. The same ballplate, holeplate and step gauge data were used for fitting both types of model;
2. Within-family and between-family validation was performed
3. the same (largely graphical) model validation tests were performed for each model, and these tests were consistent with those employed when validating the form error models (see Section 5.1.3).

### 5.1.5 Determination of volume ratios in pressure metrology

**The Model** To generate low pressures, gas in a small *inlet* volume and at a moderate pressure is expanded into much larger evacuated volume. If the ratio of the volumes is known and the pressure in the inlet volume, the pressure in the expanded volume can be estimated. In a multiple expansion experiment, an estimate of the volume ratio $\rho$ is determined by repeatedly expanding gas from the inlet volume into the large volume and the pressures $p_i$, $P_i$, and temperatures $t_i$, $T_i$, in the inlet and large volumes measured at each expansion. The pressures, temperatures and volume ratio are related through the virial equation of state (a prescribed physics-based model $F$) so that at each expansion it is possible to derive an estimate of the ratio.

Modelling issues are concerned with:

The form of the virial equation of state. This equation defines an implicit relationship between pressure, volume and temperature in the form of a power series expansion about pressure $p$. The ideal gas law includes only the constant term. For more accuracy, the second virial coefficient is included and conceivably higher order terms could also be represented.

The temperature distribution in the small and more importantly the large volume. A small number of temperature sensors are placed around the large volume, a metal cylinder. Due to temperature gradients, heat sources, etc., these temperature readings will not be equal and an estimate of the mean temperature of the gas within has to be determined taking into account this behaviour.

The error behaviour of the pressure and temperature sensors.

Prior calibration information associated with the sensors.

Correlation of uncertainties amongst observation equations.

The method of analysis of the measurement and calibration data, i.e, estimator design.

Calculation of uncertainties for the computed model parameters taking into account the uncertainties associated with the measurements and the calibration information.

Correlations arising from the use of the same sensors in different expansions.

**Validation** Issues in model validation relate to i) the models for the subcomponents, e.g., the empirical models for the sensor behaviour, ii) the selection of components in the combined model, e.g. the uncertainty in the estimate of the second virial coefficient for the gas has a neglible effect on the overall uncertainty, iii) the proper formulation of the combined model and iv) estimator validation.

At its simplest, the measurement data analysis consists of finding the mean of a set of numbers. This is embedded within the most comprehensive formulation, where the parameter estimates involve the solution of (structured) systems of nonlinear equations.

For more information see [46, 50, 51].

### 5.1.6 Pressure balance calibration

**The Model** In a liquid pressure balance, pressures are applying a force (via a set of masses) to a piston moving within a cylinder containing a liquid. The pressure generated is a function of the force applied and the effective cross sectional area of the piston. The effective area $A(p)$ itself is a function of pressure and a component of the calibration of a balance is to estimate this function. In a cross-floating experiment, a test balance is compared with a reference balance. The balances are linked and a sequence of masses $\{M_i\}$ are applied to the the reference balance. For each such mass $M_i$, the mass $m_i$ on the test balance is adjusted until it is judged that each balance is generating the same pressure.

From the two sets of mass measurements $\{M_i\}$ and $\{m_i\}$ and the calibrated estimate of the reference balance effective area $A(p)$, it is possible to estimate the effective area $a(p)$ of the test balance as a function of pressure.

Modelling issues concern

> The empirical function $A(p)$ chosen to represent the effective area. Usually $A$ is chosen to be a linear function but nonlinear behaviour is sometimes observed.
>
> Correlation amongst the errors associated with the masses. In practices the masses $\{M_i\}$ and $\{m_i\}$ are assembled from a finite number of reference masses. This means that there is an explicit correlation in the uncertainties associate with the applied masses.
>
> Uncertainty associated with the assessment of when the pressures are balanced.
>
> Incorporation of the calibration information associated with the reference balance.
>
> Incorporation of calibration information associated the mass of the piston assembly, air and fluid buoyancy, etc.
>
> The method of analysis of the measurement and calibration data, i.e, estimator design.
>
> Calculation of uncertainties for the computed model parameters taking into account the uncertainties associated with the measurements and the calibration information.

**Validation** Model validation issues concern the design of an estimator that properly takes into account all the main sources of measurement uncertainty. Like the volume ratio problem, the basic data analysis problem is very simple: fitting a straightline to data. The comprehensive approach again requires the solution of a large system of nonlinear equations.

For more information see [48, 49].

### 5.1.7 Gas effusion modelling

**The Model** Orifice flow devices are often used in low pressure measurement. Gas molecules diffuse across the orifice but, since the orifice is not infinitesimally thin, some molecules may be reflected and hence the transmission probability is not unity, and so needs to appear as a term in a molecular conductance model [21]. A physical model for a cylindrical geometry was developed by Clausing [30] in the form of a linear Fredholm integral equation of the second kind in an auxiliary probability density function $h(z)$. Since the NPL orifice geometry was more complicated, it was decided to estimate the transmission probability by Monte Carlo simulation of millions of molecular events.

**Validation** The results were compared with tabulated results, and it proved impossible to find sufficient agreement across all published references. A new model was used which takes advantage of features of the geometry of the NPL orifice to reduce the integral equation to an analytical formula with acceptable uncertainty estimates. The results were consistent with the published sources, where these existed and showed the correct asymptotic behaviour, etc.

### 5.1.8 Evaluating trends in ozone calibration data

**The Model** The UK has a network of rural ground-level ozone monitoring stations, which has been in operation over seven years. The vapour concentration of ground-level ozone is believed to be correlated with human activity and ambient temperature. There is a concern that this concentration might be rising (i.e., the trend is positive). The concentration is highly variable over many timescales (e.g., hourly, daily, weekly, monthly, quarterly), and across monitoring stations. There are also gaps in the data, e.g., due to breakdown or planned maintenance of a monitoring station.

The model described in [37] comprises trend terms and varying numbers of seasonal terms, depending on the timescale over which the data is averaged. Since it unsafe to assume a Gaussian distribution of the (weighted) errors, the approach chosen was to use bootstrap resampling to provide numerous pseudo data sets on which to estimate the trend using weighted least-squares regression. The nonparametric bootstrap $95\%$ confidence interval (based on percentiles) was found to be strictly positive for the network of stations.

**Validation** Model validation received a high priority in the project. Key parameters for model selection include:

- The number of seasonal terms: the root-mean-square residual is used as the criterion, but needs to be interpreted carefully, because the error in the data (as seen by the regression procedure) reduces as averaging increases. However, it was found that, suitably scaled, the prediction variance is reduced as the number of seasonal terms is increased.
- The number of terms in the yearly trend model: it was found that the bootstrap confidence limits for quadratic and cubic models of yearly trend were contained within those for linear trend, and hence that linear trend was sufficient.
- The nonparametric coefficient of variation of the estimate of RMS error is smallest for the linear trend model, hence providing more evidence that the linear trend model is adequate.
- The *bootstrap bias* [41] can be estimated and corrected for, and compared with 0.5 (which corresponds to “no-bias”); estimates of bias ranged from 0.48 to 0.53, implying the bias in not serious.
- Berthouex and Brown [10] show that the precision of bootstrap estimates decreases rapidly as the estimates move toward the extreme tailes of the distribution. They provide formulae for quantifying this precision.
- The model needs to take account of missing data. Weighting by the number of samples helps. However, there are concerns that extreme data points could be missed and that, if the pattern of missing data varies with time, e.g., more data is missing from the end of a series than the start, that another form of bias could be introduced. The latter can be minimised by choosing more seasonal terms (hence less averaging); the weighting strategy would then downgrade the contribution of periods in which there was substantial missing data. Therefore, this issue is addressed in the model and provides more evidence in favour of models with large numbers of seasonal terms.

- Bootstrap resampling can also be applied to individual site data. Here, each individual site is processed separately and the confidence intervals determined, from which the hypothesis that ground-level ozone concentration is not increasing at a given site can be tested. From these binary results (i.e., the hypothesis is either rejected, or not, with 95% confidence), a binomial test can be applied. Although this test assumes a binomial distribution, the assumption is much weaker than, say, that of a Gaussian: it applies to the *signs* of the values rather than the values themselves. The result of the binomial test supports that from other tests.

In summary, model validation played a very large part in this project. The number of modelling decisions was large, and the risk of inferring misleading information would be costly.

### 5.1.9 Calibration of underwater electroacoustic transducers

**The Model** The response to a stepped sinusoid is contaminated by the transient response of the device and the response reflections of tge driving signal from the tank walls and the water surface [60]. The aim of the measurement is to estimate the free-field calibration response of the transducer, and hence modelling is used to estimate the calibration parameters and remove the resonance effects.

The physical model of such a signal is a linear combination of exponentials with complex conjugate exponents, obtained by solving a linear, constant-coefficient, ordinary differential equation that describes a linear damped harmonic oscillator. The exponents must contain contain nonpositive real parts (consistent with damping and decay of transient terms). Models with such behaviour are known to be difficult to solve due to *ill-conditioning*. That is, the values of the decay parameters are not well predicted by the data, particularly if two or more decay parameters coalesce.

Classical algorithms for such problems include the Prony and modified Prony methods (both are linear, processing the data forwards and backwards in time (the independent variable), with difficulties of bias and (statistical) inefficiency) and linear/nonlinear weighted least squares, where the problems of ill-conditioning are much more obvious.

**Validation** The model parameter estimates are validated by

- applying different estimators to the same data set;
- applying the same estimator to different data windows;
- changing the frequency of pulse input (hence changing the number of periods in the “clean” part of the signal).

The quantities examined included

- the model parameter estimates;
- the RMS residual of each fit;
- the resonance Q-factor of the transducer derived from the modedl parameter estimates.

Both simulated and actual data were considered during validation.

The results of the validation were to indicate the domain of applicability of the least-squares estimators.

### 5.1.10 Analysis of surface texture of diffraction gratings

**The Model** For diffraction gratings that are nominally sinusoidal surfaces, the key functional parameters are the pitch (period) and the amplitude of the grating. In practice, profile measurements are taken on a replica of the grating and an offset curve recorded. The surface is not perfectly sinusoidal, and the resulting form error (including offset effects) can be modelled as a Fourier series. Furthermore, there is an underlying trend in diffraction grating profile measurements due to instrument drift, modelled by a polynomial.

**Validation** The modelling procedure [32] consists of an outer iteration (to find the best value of the period) within which there is an iterative model building procedure to determine $q$, the number of Fourier harmonics required to fit the grating profile, and $n$, the number of polynomial terms required to fit the drift effects. The criterion used is the RMS residual which is inspected for "convergence". Generally, the updates to the period estimate are relatively small, which indicates that estimates are precise and there is no evidence to prove that they are inaccurate.

### 5.1.11 Interpolation of (electrical) load impedance data

**The Model** Reliable representation (calibration) of electrical load impedance data, as a (complex-valued) function of frequency, is required in electrical standards work. Difficulties are introduced by the presence of large gaps in the data, e.g., between DC (frequency = 0) and 45 MHz. Data in the (45–100MHz range) is measured in 1 MHz frequency intervals. Physical models (based on Taylor series expansion across the interval) were rejected, because of concerns about the radius of convergence and lack of flexibility. Unconstrained empirical functions were also rejected because the presence of the gap introduces ill-conditioning such that the uncertainty associated with each estimate would be unacceptable. The underlying physical model for the voltage reflection coefficient takes the form of a ratio of transmission line admittances, which depend on frequency. From this analysis, two frequency-derivative-based physical parameters are derived, together with boundary conditions to be satisfied by the curve which "crosses the gap".

**Validation** The proposed model [34] is to use empirical functions (polynomials) subject to boundary conditions on the value and derivatives around the gap, possibly supplemented by shape constraints and/or regularisation (ridge regression, see B.1.2).

Within-family validation based on saturation of the RMS residual is being considered, together with graphical plots of the residuals and their variation with frequency. Between-family validation will focus on the use of variable transformations, such that the gap in the transformed variable would be less than that in the original (frequency) variable.

### 5.1.12 Roundness assessment

**The Model** NPL and Taylor-Hobson are developing the new national primary roundness-measurement instrument which operates in a similar form to the established Talyrond instrument (recording suppressed radius values as a function of angle of rotation). However there are numerous refinements, particularly the design of a multi-indexing measurement strategy which enables the metrologist to separate component form error (the object of the measurement) from instrument spindle error (a major source of systematic error in such measurements) [35].

**Validation** The model is supported by in-depth error analysis which evaluates the uncertainties of the component form error and the instrument spindle error in terms of the number of measurement traces, the number of points per trace and the indexing angle associated with each trace. The individual fits to component form error, etc., are displayed in polar plots and are consistent with expectation.

### 5.1.13 Design of a non-contact probe

In making measurements, for example using a co-ordinate measuring machine it is desirable to minimize contact between the probe tip and the artifact being investigated, in particular to avoid impact damage that would alter the measurements of the artefact. A probe tip that detects proximity to the artifact would help to minimize the effects of impact and hence improve the overall accuracy. The feasibility of a proximity probe based on changes in an electrical field was investigated. The basic set up is illustrated in Figure 5.1

The two plates and the shield then form ’the probe’ and we are then interested in the change in the electric field between plate 1 and plate 2 as the position of the whole ’probe’ assembly is varied. We measure the distance of the assembly by considering the distance of the shield tip to the artifact.

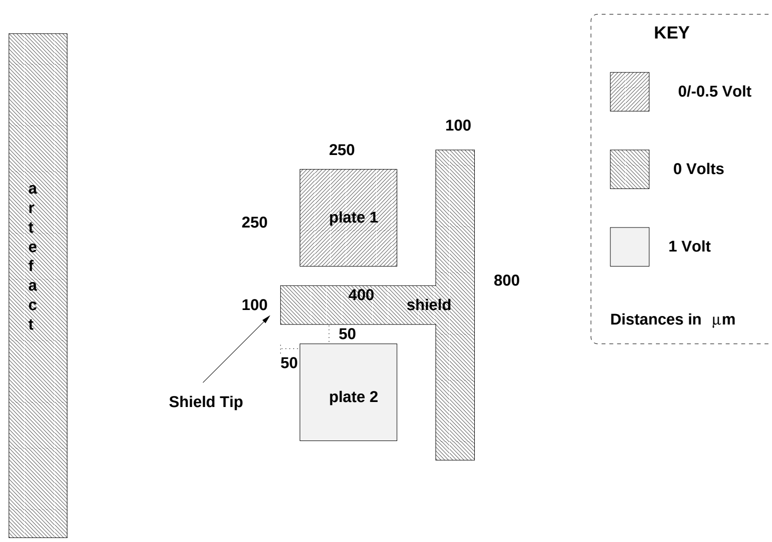


Figure 5.1: Geometry of example probe. In the simulations the distance between the shield tip and artifact is varied.

**The Conceptual Model** The following assumptions were made in writing down a 2-Dimensional model

1. Artifact is made of a conducting material
2. No free charges in air
3. The potential difference between the plates and shield is known
4. The potential difference between the shield and conducting surface is known.
5. Edges of the boundary, other than the artifact, are non-conducting.

In this case the electrical potential $U$ may be found by solving Laplace's equation

$$\Delta U = 0$$

with boundary conditions imposed from the known potentials on the conducting surface, shield and probes. In order to track the change in potential as the distance between the shield tip and artifact varies - we monitor the potential around plate 1 at a distance of 25 $\mu$m.

**Numerics and validation** A standard finite element code was tested on two examples (for which analytic solutions of the potential are known) as a verification of the code. In these examples results were shown to be correct to four decimal places (with a fixed and relatively course mesh).

In figure 5.2 we have plotted details from an example contour plot with a distance of $8\mu$m between the shield tip and artifact with a voltage of $-0.5$V on plate 1.

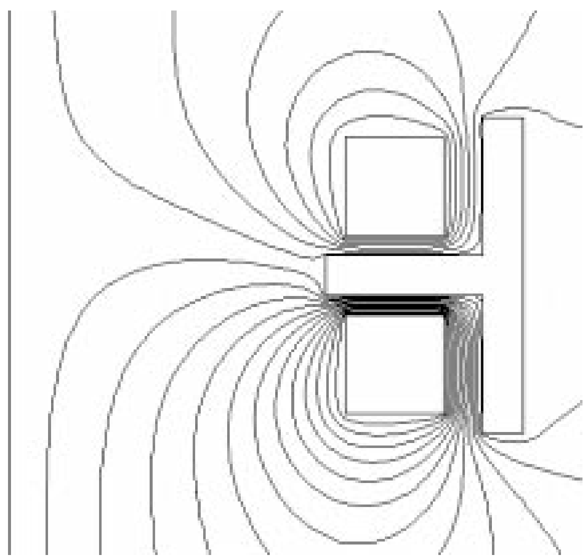

Figure 5.2: Example run with a distance of $8\mu$m between the shield tip and artifact with a voltage of $-0.5$V on plate 1

The effect of decreasing the mesh size (and hence accuracy) was investigated for the problem under consideration. In figure 5.3 we have plotted the difference between the average potential around plate 1 of 2000 and 4000 nodes as a function of distance from the artifact. This gives an estimate of the error. We see from the figure that the error is of the order $10^{-4}$ and that the error decreases with distance. Note that although it appears that at the smallest distance the error is zero this is infact due to the scale of the plot.

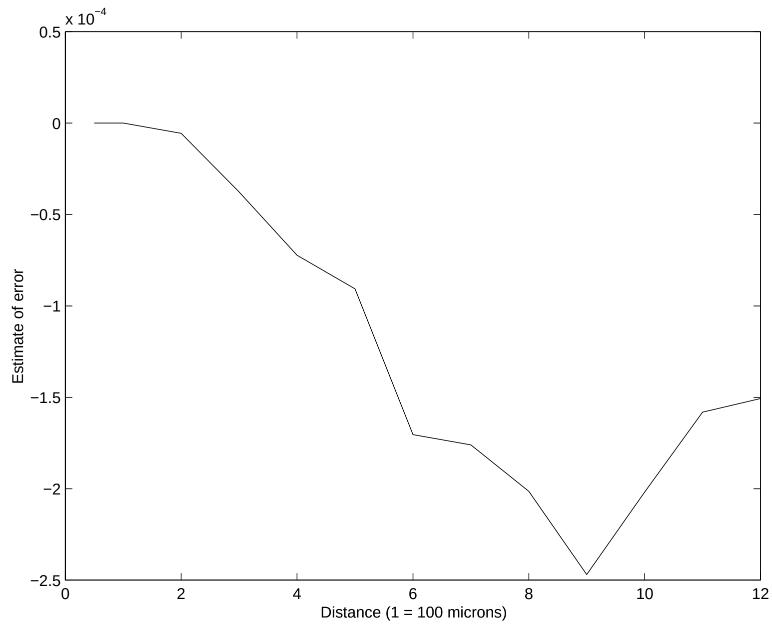


Figure 5.3: Error as the mesh is doubled as a function of distance from the artifact. Error is measured as the difference between the average potential around plate 1 solved with 2000 and 4000 nodal points. Voltage on plate 1 = 0 Volts.

## 5.2 Examples in areas other than metrology

### 5.2.1 Biomedical signal analysis

Nonlinear autoregressive modelling of EEG signals [87] is clearly an area where modelling risk and human safety overlap. Introduction of an additional term into a linear or nonlinear autoregressive model always improves the fit (in the mean-squared sense). Akaike showed that, for a linear autoregressive model, a significant improvement in the fit is associated with a reduction in the residual variance of at least $2V/N$, where $V$ is the variance without the candidate additional term, and $N$ is the number of data points.

The authors show that the same criterion, a reduction in residual variance by at least $2V/N$, is a criterion for the significance of a single nonlinear term as well.

### 5.2.2 Quantitative Population Ecology

Sharov [83] has used $t$-tests and cross-validation on regressive and autoregressive models. He used $t$-tests on parameters in an autoregressive model of yearly-sampled fox populations in Labrador and found that "odd-years" data, though satisfying internal model validation tests, is not a good predictor of "even-years" data and hence the model, when used to predict all the data, is invalid.

The jack-knife was applied to show that, for the same data, an autoregressive model with a lag of two years, e.g., where the population at year $t$ is predicted by the population at year $t - 2$, is reliable. The correlogram product test is used to show that the hypothesis that butterfly activity is correlated with sunspot activity can be rejected at the $95\%$ confidence level.

### 5.2.3 Structural System Identification

Sandia National Naboratory[81] in the US has several projects relating to the validation of continuous (structural) models. The aim is to develop a systematic methodology and tools for developing predictive analytical models based on a physical structure's dynamic response characteristics. Sandia's goal is to develop structural system identification techniques and software suitable for both research and production applications to solve problems of national interest.

One of these projects, MP-SSID, is to compute, for a particular configuration of design variables, the modal parameters of a finite element model using the linearized sensitivities of the model's mass and stiffness. MP-SSID then uses these parameters to compute measures of the model's predictive accuracy with respect to available experimental data. MP-SSID will also compute the nonlinear modal parameter sensitivities of the model about the given design variable configuration. A large number of these single-node modules will be controlled simultaneously by the DAKOTA optimization package to perform global and local parameter estimation in parallel using both gradient and non-gradient optimization techniques, such as genetic algorithms.

SSID is actually a toolbox of software programs which not only perform the parameter estimation but also control execution of NASTRAN (a commercial aerospace-standard finite element package) to compute model sensitivities. SSID employs a Bayesian parameter estimation theory which treats both the model parameters and the data used for model adjustment as unknown or subject to some random errors and uses the relative confidence in the parameters and the data to guide the estimation process.

SSID uses stiffness and mass sensitivities together with state-of-the-art numerical analysis tools for sparse matrix manipulation. This enables the efficient handling of very large order problems and minimizes the need to recompute model sensitivities.

The Error Localization and Parameter Selection/Estimation (ELAPSE) algorithm was developed to address the problem of localizing model errors using dynamic response data. The algorithm is based on computation of a degree-of-freedom error vector using the mass and stiffness of a finite element model together with the modal parameters identified from dynamic testing. The analytical and experimental parameters are used to compute a force imbalance which is indicative of the errors in the model. The location of modeling errors can then be used to guide the analyst in selection of parameters for model building.

### 5.2.4 Agro-ecosystems modelling

One of the aims of the EC-supported CAMASE concerted action RTD project [22] is to improve the reliability of models used in agro-ecosystems modelling. Among other outputs, the project made available a first draft of guidelines relating to validation, sensitivity and uncertainty analysis and calibration supplemented with definitions. The long-term aim is to produce a manual with more explicit guidelines, procedures, tools and examples.

### 5.2.5 Circuit simulator models

Both NIST [61] and FSA (the Fabless Semiconductor Association) [53] have set up projects relating to the validation of circuit simulations. Such models are highly complicated and hence proving their validity is non-trivial. NIST have chosen to look at their validation in a more general sense than FSA, whose activity will emphasise the widely-used SPICE simulation package. In both cases, the emphasis is on practical issues such as choice of validation data sets, etc.

### 5.2.6 Atmospheric models

The National Environmental Research Institute, Denmark (NERI) [76] has an initiative on "Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes". This initiative produced a "model validation kit" comprising three data sets and software for a particular model.

### 5.2.7 Climate models

The Climatic Research Unit at the University of East Anglia recently (1997) had a contract [63] with the DoE to facilitate and provide an independent evaluation of Hadley Centre climate model performance in relation to international efforts to model the climate system, to detect human-induced climate change and to predict future climate change. This grew out of a Climate Impacts LINK project.

**Validation** The main issues were that the models are being refined constantly as they are updated with more data, but that certain geographical anomalies persist, indicating areas for improvement. Some validation tests indicate that the presence of different scales (both in space and time) seem to have adverse effects on the solution. A key test is that correlation anomalies between model and new data should decay rapidly, otherwise the model is ill-conditioned with respected to data updates.

### 5.2.8 Learning from human control strategies

A research group [77] is using neural networks in a control application, namely, attempting to model human skills such as driving a vehicle. The neural network model requires validation, not just with validation data held back from the training set, but also from more theoretical models of behaviour. The aim is to verify the human control system (HCS) model fidelity through a novel stochastic model validation procedure. A *dynamic* similarity measure, based on Hidden Markov Model (HMM) observation probabilities, compares dynamic human and model trajectories. By connecting this validation procedure with simultaneously perturbed stochastic approximation (SPSA) in an iterative loop, the system can automatically select that input representation which maximizes model fidelity.

### 5.2.9 Neural network tools

In the present release of the Stuttgart Neural Network Simulator (SNNS), three model selection criteria are implemented: Schwartz Bayesian Criterion (SBC), Akaike Information Criterion (AIC) and the conservative mean square error of prediction (CMSEP). The SBC, the default criterion, is more conservative compared to the AIC. Thus, pruning via the SBC will produce smaller networks than pruning via the AIC. The authors note that both SBC and AIC are selection criteria for linear models, whereas the CMSEP does not rely on any statistical theory, but happens to work pretty well in an application. These selection criteria for linear models can sometimes be applied directly to nonlinear models, if the sample size is large.

### 5.2.10 Population pharmacokinetics

The US FDA Center for Drug Evaluation and Research has issued draft guidelines for industry on population pharmacokinetic modelling [29]. Population pharmacokinetic data analysis can be carried out in three interwoven steps: (a) exploratory data analysis, (b) population pharmacokinetic model development, and (c) model validation. The data analysis plan should be clearly defined in the study protocol.

Exploratory data analysis isolates and reveals patterns and features in the population data set using graphical and statistical techniques. It also serves to uncover unexpected departures from familiar models. An important element of the exploratory approach is its flexibility, both in tailoring the analysis to the structure of the data and in responding to patterns that successive analysis steps uncover.

Most population analysis procedures are based on explicit assumptions about the data, and the validity of the analyses depends upon the validity of assumptions. Exploratory data analysis techniques provide powerful diagnostic tools for confirming assumptions or, when the assumptions are not met, for suggesting corrective actions.

**Validation** The reliability of the analysis results should be checked using diagnostic plots; confidence intervals (standard errors) for key parameters should be checked using nonparametric techniques (such as the jackknife) and the profile likelihood plot (mapping the objective function). This is appropriate because the possibility of biased results is increased by the complexity of the population models and by the sparse individual data available. The nonlinearity of the statistical model and ill-conditioning of a given problem can produce numerical difficulties and force the estimation algorithm into a false minimum. Because the maximum likelihood procedure is sensitive to bizarre observations, it may be necessary to check the *stability* of the model. It is important to evaluate the quality of the results of a population study/analysis for robustness. Evaluation for robustness may be approached by sensitivity analysis; the use of case deletion diagnostics is also encouraged. Evidence of robustness renders the results reasonable and independent of the analyst.

For testing predictive performance, data-splitting is an effective method of model validation when it is not practical to collect a new set of data to test the model. Cross-validation is a form of repeated data-splitting and has the advantage of less variability in the estimates. Bootstrapping is also encouraged.

Key model validation criteria include

- Standardized Prediction Errors (residuals)
- estimates of bias and variance of the model parameters
- Plot of Residuals Against Covariates (explanatory variables)

#### 5.2.11 Computer-aided mechanical control

Glasgow University Department of Mechanical Engineering has a project to produce robust model validation tools for nonlinear dynamic parametric models, in particular for the rotor controls of a helicopter. Preliminary findings are that such tools as exist are fragmentary, poorly integrated and difficult to use.

#### 5.2.12 Stratospheric ozone and human health

The US Socioeconomic Data and Applications Center's (SEDAC) Stratospheric Ozone and Human Health Project [43] uses two forms of model validation

- comparison of measured and modeled data: clear-sky UV irradiance; here some difficulty is caused by the variability in the data, since the comparisons are performed at time steps of 10 minutes
- inter-model comparison with peer-reviewed model; the latter uses a different approach (based on radiative transport) and difficulties are caused by having different natural parameters for comparison.

#### 5.2.13 Defence war-gaming and simulation

The US Department of Defense (DoD) has a strong commitment to Validation, Verification and Accreditation (VV&A) of the models and simulation (M&S) used in training and evaluation of new military systems. Schemes include

- US Defence Modelling and Simulation Office VV&A Technical Working Group [4]
- US Navy Fleet Modelling and Simulation Fleet Project Team [3]
- The US Joint Simulation System, which has published a draft VV&A strategy [2]

Many wargames include humans "in the loop" and validation of such simulations is analogous to measurement system validation. Other simulations are performed wholly in software and hence require extensive VV&A if the users (senior military figures) are to have confidence in the results.

A highly-structured process is favoured [2], where the following may be found:

> *Conceptual Model V&V*
> Conceptual model *verification* can be loosely defined as "Did I model the thing right?" and conceptual model validation as "Did I model the right thing?" Verification addresses

> the M&S function. Validation addresses the M&S fidelity. The verified and validated conceptual model can be used as the structural foundation for the rest of the development effort.
>
> Conceptual model verification and validation should occur before M&S development to avoid the potential pitfall of inaccurate representation and failure to meet the proposed requirements. Errors caught at this early stage of development are easier and less expensive to fix.
>
> *Design V&V*
>
> The Design V&V activity is associated with the design phase of development. At the design phase the developer moves into the implementation specific detail, defining how the M&S will be organized and built. The system architecture, software components, interfaces, and data are all elements of the resulting design. The purpose of design V&V is to assess and document the degree to which all detailed requirements are satisfied by the design, as well as the degree to which the design satisfies the conceptual model. The quality and sufficiency of the design and design documents are also assessed and documented to determine their sufficiency in supporting subsequent development and/or V&V activities.
>
> *Implementation V&V*
>
> The Implementation V&V activity is associated with the implementation phase of software development, where the design is translated into the implemented code. Implementation verification seeks to assess and document the degree to which the implemented code conforms to the design specifications and any defined coding standards. This V&V activity should be automated to the greatest extent possible, through application of CASE tools, CARE tools, and/or data comparison, analysis, and test tools Requirements traceability is a critical feature of implementation V&V. An automated tool that supports the trace from requirements to the implemented code will be used. Once the implementation of the design is completed in code, the results of the M&S are formally reviewed. Responses of the M&S are compared against known or expected behavior from the subject represented to determine if the M&S responses are sufficiently accurate for the intended use.

Other useful references include [8] and [80].

It can be seen that there are many similarities with the approach which the present document's authors have proposed, independently, for metrological models.

Simulation in this sense largely relates to discrete events governed by probabilities, such that the simulation can be modelled as a Markov chain. Key modelling techniques include Monte Carlo simulation and queueing. The former, in particular, has many possible applications in metrology, e.g., when modelling quantum phenomena and estimating uncertainties.

The assumptions of the conceptual model need to be checked, by techniques such as expert review. The discretised model needs to be checked by techniques such as model embedding, e.g., by comparing long-run simulations against shorter runs.

### 5.2.14 Defence and aerospace procurement

Defence war-gaming is heavily influenced by the properties of the resources available to the war-gamer. This has led to VV&A being extended to include procurement, by helping to set requirements. The US Army has a *Virtual Proving Ground* [5] which is intended to assist with both specification and testing. Large defence suppliers such as Boeing have incorporated VV&A in their processes and it now requires this of its business partners [14].

Clearly the bulk of this activity applies to whole systems, and hence the nearest SSfM project would be Measurement System Validation. However, there are individual models, e.g., those which predict the stress distribution in a wing flap, which play an important role. Models of this kind also appear in metrology, e.g., heat flow in a calorimeter (see Sections 2.3 and 3.2).

# Chapter 6

# Conclusions

The joint main objectives of this report were

- to define the concept of model validation and to indicate its importance in the modelling process and, by implication, in metrology.
- to survey the state of the art in model validation, where this might apply, albeit indirectly, to practical metrological problems.

Metrology is built upon sound practice; it is clear that modelling is part of that practice, hence its state should be kept under review. We believe this report is timely, in that the requirement for formal validation has never been greater.

Within the SSfM Model Validation project, this report constitutes the first deliverable. As such, it provides useful background material for the Case Studies which will run in concert with those in the Modelling project. Furthermore, it is also intended to provide technical input to the Model Validation Best Practice Guide which will be the ultimate deliverable of the Model Validation project.

# Appendix A

# Geometrical interpretation of discrete modelling

A geometrical interpretation assists with understanding some of the issues in regression (data approximation). A given data set $X$ can be visualised as a point in $m-$dimensional space. The space of models can be viewed as a surface $S(\mathbf{a})$ in that space, where the surface is of dimension $n$, where $n$ is the length of the parameter vector $\mathbf{a}$.

Generally, each family of models $S(\mathbf{a})$ can be associated with a particular surface, e.g., the model space of all polynomials is $S_{\text{poly}}(\mathbf{a})$. Furthermore, each surface (model space) can be *parametrised* in several ways [7]. Parametrisation is based on the notion of choosing suitable *basis functions* such that a linear combination of these basis functions defines the $n$-dimensional model space $S(\mathbf{a})$. Apart from trivial cases, most parametrisations are limited to certain parts of the model space.

> *Example* The parametrisation of a cone in standard position with a narrow apex angle (with constraints depending on vertex position) is complementary to that of a cone with a large apex angle (with constraints depending on cone radius). Indeed, the optimum parametrisation of the latter tends to that of a plane as the apex angle size approaches $\pi$ radians [52]. That is, both cone parametrisations together define the model space of all cones $S_{\text{cone}}(\mathbf{a})$.

The goal of modelling used in data approximation problems is to find that point $X^*$ in $m-$dimensional space which represents the *expected* value of $X$, i.e., $X^* = E(X)$. To achieve this, the standard statistical approach is to employ *maximum likelihood estimation*, MLE (see, for example, [88, 69]). The aim with MLE is to determine the model $F$ which maximises the likelihood of the data that was observed, where the likelihood $L(X;\mathbf{a})$ is

$$L(X;\mathbf{a}) = \prod_{\mathbf{i=1}}^{\mathbf{m}} \mathbf{f}(\mathbf{x_i};\mathbf{a}),$$

where $f(x_i,\mathbf{a})$ is theprobability of observing the $i^{\text{th}}$ data point. In practice, algorithms for MLE use the log-likelihood function $l(X;\mathbf{a}) = \log(\mathbf{L}(\mathbf{X};\mathbf{a})$. It can be shown that MLE reduces to ordinary least-squares estimation if the errors in $X$ are normally-distributed and uncorrelated.

In practice, modelling does not attempt merely to find $X^*$, though that is often valuable in itself. The reasons are as follows:

1. Searching $m-$dimensional space for $X^*$ is not a simple task. In particular, it is necessary to constrain that search to a suitable subspace $S(\mathbf{a})$ which has the properties (smoothness, asymptotic behaviour, etc,) desired/assumed by the metrologist.

2. Generally, knowledge of $S(\mathbf{a})$ in the neighbourhood of $\mathbf{a}^*$ is helpful in validating the model (since it is then possible to test more than mere closeness to the data) and in estimating uncertainties.

Consequently, for a given subspace $S(\mathbf{a})$, the approximation algorithm searches for $X^*$ returning the model $X_S^* = S(\mathbf{a}^*)$, where $\mathbf{a}^*$ is a particular choice of parameter values $\mathbf{a}$. If the algorithm which searches for $\mathbf{a}^*$ is successful, we can claim that $X_S^* = S(\mathbf{a}^*)$ is the best estimate of $X^*$ from the model space $S(\mathbf{a})$.

In terms which are more familiar to the metrologist, $X$ is the metrologist's data set, say length of a gauge block at different values of the temperature), $S(\mathbf{a})$ might be the set of all linear relationships between temperature and length of a gauge block, $X^*_{\mathrm{line}}$ is the best-fit line to the data set $X$ of length and temperature, and $\mathbf{a}^*$ is the vector of linear parameters, e.g., line slope and intercept.

The *error function* (also known in the literature as a *loss function*) $E(X, \mathbf{a})$ [17] is a scalar function of the "distance" between the data set $X$ and a point $\hat{X}_S = X(\hat{\mathbf{a}})$ in $S(\mathbf{a})$:

$$E(X, \mathbf{a}) = \|\mathbf{X} - \mathbf{X}(\hat{\mathbf{a}})\|, \tag{A.1}$$

where the form of $E$, particularly the choice of norm, depends on the error model assumed by the metrologist.

> *Example* : Fixed precision data, e.g., tabulated data which is supplied with fair rounding to a given number of significant digits, has errors which may be assumed to belong to a uniform distribution, hence Chebyshev (minimax) fitting may be appropriate.

> *Example* : A *smoothing spline* [70] may be appropriate if the data errors are large and normally distributed. In such a case, conventional splines are fitted by least-squares, say, and the standard least-squares error function is penalised by a term which is a function of the second derivative of the spline fit.

Some error functions and/or their derivatives with respect to the parameters $\mathbf{a}$ are difficult to calculate. Furthermore, if the error function is nonlinear in the parameters, a solution algorithm such as the Gauss-Newton method for nonlinear least-squares problems [54] uses a linearised error function at each iteration to compute update steps towards the solution of the nonlinear least-squares problem. Sometimes, application-specific linearisations are suggested. *Estimator validation* is the term given to the process of validating estimators which employ approximate error functions and/or simplified error models, such as neglecting correlations between the errors.

The choice of conceptual model is equivalent to choice of $S(\mathbf{a})$ and its parametrisation. The choice of discretised model is associated with choice of error function, estimator validation, etc.

The gemetrical interpretation of regression problems can be extended to continuous models, even though quite different solution techniques are used. Again, the metrologist has some data $X$, possibly in the form of measurements on a grid in the case of hybrid models of, say, strain due to loading a component on a CMM, or prescribed boundary data (with its uncertainty) in the case of a continuous predictive model used for simulation.

The model equation defines relationships between the function of interest and its derivatives. The required function can be represented by a point in infinite-dimensional parameter space. In practice, however, the space is finite-dimensional, since it is sufficient to know the function and/or its derivatives at a finite set of points, and/or the function is approximated by a simpler function defined on a set of basis functions, with parameters $\mathbf{a}$.

# Appendix B

# Model validation lessons from statistical theory

Model validation is an important research topic in its own right. It is beginning to appear as part of university statistics courses and is built-in to the regression and modelling modules in statistics packages. A gap is beginning to open up between statistical theorists, who are developing proofs and techniques using Bayesian and information-theoretic analysis, and the majority of metrology users, who continue to use traditional (i.e., pre-1980) criteria such as $C_p$, etc.

The aim of this section is to review some of the better-known techniques, all of which continue to be very valuable to the data analyst, and to present brief descriptions of some of the newer concepts. Where possible, it is shown how the newer concepts build on the old.

## B.1 Statistical model validation

Discrete models in metrology have parameters which can be estimated by regression techniques. The statistics literature contains many references to regression, and validation of regression models is a major topic (see, e.g., [70, 13, 72]).

Statistical tests of model validity can be classified according to whether the form of the distribution of the residuals is assumed or not.

### B.1.1 Tests based on parametric statistics

**Tests on metrics defined on the residuals**

1. The residual norm is consistent with the expected error norm. Clearly, this is norm-dependent. For example, if the minimax norm is used as a criterion, we would expect the residual and predicted error norms to be qualitatively similar. Alternatively, if the Euclidean norm/least-squares criterion is used, and
   - the residuals $d_i, i = 1, \ldots, m$ are unbiased (asymptotically, the mean $\overline{d} \to 0$);

- the weighted residuals $w_i d_i$ each have variance 1;

then the expected value of the residual norm is $\sqrt{m-n}$ where $m-n$ is the number of degrees of freedom of the regression.

2. The residual norm is a minimum over all fits satisfying relevant constraints.
3. An $F$ (variance-ratio) test can be applied to determine whether additional terms in a model family make a significant reduction in the norm of the residuals [82]. This test is potentially quite powerful, but requires the assumption that the sequence of residual norms converges uniformly. Such an assumption does not hold when, for instance,
   - a full Fourier series representation of an even function is required, although it might still hold for the coefficients of the cosine terms, or, more generally,
   - a consecutive group of terms (of unknown size) gives a significant improvement, but a single term does not.

**Tests on the distribution of the residuals**

1. The (weighted) residuals have the same distribution as the expected distribution of the errors.
2. The residuals are unbiased. Moreover, for most practical fitting criteria (which are symmetric about zero, i.e., the key quantity is the magnitude of a residual, not its sign), the distribution of the residuals should also be symmetric, at least qualitatively. The most noteworthy exception is one-sided regression, which occurs in some geometric tolerancing applications, e.g., maximum inscribed and minimum circumscribing circles [18].
3. The cumulative distribution functions of the residuals and the hypothesised residual model are qualitatively similar. Many statistical packages, such as S-Plus with its `cdf.compare` function, offer regression diagnostic plots which provide a graphical means of indicating whether there is any similarity between two cumulative distribution functions.
4. There are several generic tests which are based upon assumed distributions of the residuals, being based on differences between the postulated distribution and the empirical distribution of the residuals. These inclused the well-known Chi-squares ($\chi^2$), Kolmogorov-Smirnov and Cramer-Smirnov-von Mises hypothesis tests which compare a summary statistic of the difference between the two distributions against tabulated values of the difference statistic.
5. If we make the relatively weak assumption that the residuals are symmetrically distributed about zero, we can derive hypothesis tests on the lengths of *runs* of residuals having the same sign. The simplest test is to assume that each residual has an equal probability of being positive or negative, and hence apply a binomial test to the signs of the residuals [19]. However, for moderate to large samples, this can lead to underfitting. A more useful test is to produce a histogram of run lengths and apply a $\chi^2$ test to compare against the expected run lengths for a postulated sample distribution. This concept is taken further in a test proposed in [13] as follows. If $m$ positive numbers and $n$ negative numbers appear in a random sequence with all orderings equally probable, then the expected number of runs $r$ and its variance are:

$$E(r) = 1 + \frac{2mn}{m+n}$$

$$V(r) = \frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)} \quad \text{(B.1)}$$

One first forms the histogram of the difference between the two histograms to be compared, or of the difference between the histogram and the function to be compared, and then one counts the number of runs in the difference. This number is then compared with that expected under the null hypothesis, which is such that all orderings of sign are equally probable.

The runs test is usually not as powerful as the Kolmogorov test or the $\chi^2$ test, but it can be combined with the latter test since it is (asymptotically) independent of it.

**Tests on parameters and auxiliary requirements**

1. Each additional parameter is significant in the sense of being different from zero. Given such a parameter value and an estimate of its uncertainty, a $t-$test can be performed and its significance estimated. However, this is based on the assumption that the parameter is drawn from a Normal distribution, which cannot be guaranteed. If the conditions of the Central Limit Theorem are met, i.e., the data errors are independent and of comparable magnitude, the errors in the parameters will generally be nearer to a normal distribution than data errors from which they are derived. Otherwise, non-normality is a concern, as stated above.
2. The values/uncertainties of critical parameters are consistent with what was expected. As an example, an unconstrained fit to data might be sufficient and hence, for example, convexity constraints are not essential.
3. Constraints on the fit and the parameters are satisfied. If constraints were applied, it should become clear from a plot whether the constrained fit is acceptably close to the data. If not, it suggests that data violation of the constraints is a serious problem and more data might be required.
4. When evaluated and plotted, the fit is "taut" and, if additional data is obtained, the deviations are consistent. This is a test of the predictive capability of the model. It is a qualitative counterpart of cross-validation which is a non-parametric test described below.
5. The *correlogram product method* described in [83] is a method for validating correlations between autocorrelated factors. If the correlation coefficient of two such sets is computed, it may appear to be significant, but this could be spurious.

   The standard deviation of the correlation between 2 independent autoregressive processes, $X$ and $Y$, equals the square root of the weighted product of correlograms for variables $x$ and $y$:

   $$\mathrm{SD}_r = \sqrt{\sum_h w_h \rho_{x(h)} \rho_{y(h)}} \quad \text{(B.2)}$$

   where $h$ is the temporal or spatial lag, $\rho_{x(h)}$ and $\rho_{y(h)}$ are correlograms for variables $x$ and $y$ and weights $w_h$, are equal to the proportion of data pairs separated by lag, $h$.

   Thus, to test for significance of the correlation between processes $X$ and $Y$, we need to estimate the standard error using the equation above. If the absolute value of empirical correlation is larger than $2\times$ SD approximately, then the correlation is significant at the $95\%$ confidence level.

### B.1.2 Tests based on nonparametric statistics

**Tests on the metrics defined on the residuals** The following tests are based on the residual sum of squares (RSS) and hence are most suited to traditional least-squares regression. However, no explicit assumption of distributional form is made.

1. The value of $R^2 \equiv (\text{TSS} - \text{RSS})/\text{TSS}$, where TSS is the total sum of squares, that proportion of the variance of the dependent variable which is explained by the independent variables in a linear regression model, is a maximum. Note that $R^2$ depends on the residual sum of squares and hence is related to the norm of the residuals: generally, $R^2 \to 1$ as $\|\mathbf{d}\| \to \mathbf{0}$. As well as the implicit dependence on the least-squares criterion, the difficulty with this criterion is that $R^2$ will generally increase with $n$, the number of terms in the regression model. Hence there is a danger of overfitting unless some account is also taken of the number of terms.

2. The value of Mallow's $C_p$ statistic [75, 70, 88] is as close as possible to $n$, where $n$ is the number of terms in the regression model. The $C_p$ statistic is defined to be

$$C_p = \frac{\mathrm{R}SS}{\hat{\sigma}^2} - (m - 2n), \tag{B.3}$$

where RSS is the residual sum of squares for a model with $n$ terms, $\hat{\sigma}^2$ is an estimate of the "true" residual sum of squares, $m$ is the number of data points and $n$ is the number of regression terms. A good model will have $C_p \approx n$.

3. Akaike's Information Criterion (AIC) [70, 1, 25, 27] is defined as follows:

$$\text{AIC} = m \log \frac{\text{RSS}}{m} + 2n. \tag{B.4}$$

Following [25, 25, 26, 28], this is an estimate of the Kullback-Leibler *discrepancy* (also known as the relative *entropy*) between the model generating the data and a fitted candidate model. For small samples, this has a negative bias, hence the corrected AIC, denoted AICc is defined to be

$$\text{AICc} = m \log \frac{\text{RSS}}{m} + \frac{m(m+n)}{m-n-2}. \tag{B.5}$$

Regardless of which version is used, the best model is that with the minimum discrepancy measure. Both are asymptotically unbiased estimators of the expected Kullback-Leibler discrepancy.

4. The Schwarz Bayesian Criterion (SBC) [59] (also known as the Bayes Information Criterion) is defined as follows:

$$\text{SBC} = m \log \frac{\text{RSS}}{m} + n \log m. \tag{B.6}$$

The SBC is derived by taking a Bayesian view of discrepancy. Again, the best model in a set is that which minimises the SBC. Generally, use of SBC rather than AIC leads to more parsimonious models (fewer parameters per model) [85]. Indeed, the BIC criterion is *consistent* in the sense that, if there is a "true model" of dimension $n$, as $m \to \infty$, the use of SBC will identify that model with probability 1. Use of AIC will tend to lead to over-fitting.

**Resampling methods for internal validation** Some authors distinguish between internal and external validation. The former is conducted by reusing the original data, often by data-splitting [29]. The latter requires new data to test the model. We can classify the following methods as providing some degree of internal validation:

1. Cross-validation
   The CAMASE Guidelines [22] define *cross-validation* as a procedure for calibrating and validating a model with a limited number of representative data sets. It consists of repeated subdivision of all the data into calibration and validation data, followed by corresponding calibration and validation. The average of the observed prediction errors over the subdivisions provides an estimate of the prediction error in an entirely new situation. There are several variants of cross-validation. In the most popular one, called leave-one-out validation, each independent data set gets the role of validation data exactly once, at which occasion the complementary set gets the role of calibration set. Some statisticians do not consider cross-validation to be a model validation procedure [83] because of a lack of methods for testing the statistical significance of the results of cross-validation. Standard ANCOVA cannot be applied because it does not account for the variation in regression parameters when some data are not used in the analysis. Cross-validation determines the variability of regression results when some data from the given set are omitted but is unable to predict the variability of regression results in other possible data sets with similar characteristics.

2. Jack-knife
   The *jack-knife* [40], and originally due to Tukey, is defined in [13] to be a method in statistics allowing one to judge the uncertainties of estimators derived from comparatively small (say 20) samples, without assumptions about the underlying probability distributions. The method consists of forming new samples by omitting, in turn, one of the observations of the original sample. For each of the samples thus generated, the estimator under study can be calculated, and the probability distribution thus obtained will allow one to draw conclusions about the estimator's sensitivity to individual observations. Sharov [83] warns that the jack-knife still does not qualify, strictly, as a validation procedure, however it is better than cross-validation in the sense that can be used for testing the significance of the regression. He provides a worked example for an autoregressive ecology model. The key advance made in the jack-knife is to derive pseudo-values $\mathbf{b^{(i)}}, \mathbf{i} = \mathbf{1}, \ldots, \mathbf{m}$ from the $m$ estimates of linear regression parameters $\mathbf{a^{(i)}}$ obtained by omitting data value $\mathbf{x_i}, \mathbf{i} = \mathbf{1}, \ldots, \mathbf{m}$, as follows

   $$\mathbf{b^{(i)}} = \mathbf{m}\mathbf{a}^* - (\mathbf{m} - \mathbf{1})\mathbf{a^{(i)}} \tag{B.7}$$

   where $\mathbf{a}^*$ is the estimate of the parameters derived using all the data, and to perform Student's $t$-tests on these pseudo-values as a means of determining the significance of the regression model. Sharov [83] notes that the jack-knife is less sensitive to the shape of the distribution than standard regression and hences is more reliable.

3. Boot-strap resampling
   Following [13], the *bootstrap* is a method allowing one to judge the uncertainty of estimators obtained from comparatively small (say 20) samples, without prior assumptions about the underlying probability distributions. The method consists of forming many new samples of the same size as the observed sample, by drawing *with replacement* a random

selection of the original observations. The estimator under study is then formed for every one of the samples thus generated, and will show a probability distribution of its own. From this distribution, confidence limits for the bootstrap estimates can be given. If this procedure is repeated across a number of models, the width of confidence intervals can be used to quantify the quality of the model, i.e., better models have smaller confidence intervals.

**Stepwise algorithms** Some *algorithms* for modelling include model validation at each step. These include

1. Traditional *stepwise selection*, e.g., those which involve adding terms to a smaller model, or dropping terms from a larger model, takes advantage of embedded model development to generate sequences of models. Such approaches are supported in packages such as S-Plus [70] and the NAG library [75]. Generally, criteria such as the residual sum of squares (RSS), AIC, SBC, $C_p$, etc., are used to judge the quality of a regression.
2. *Shrinkage (Ridge) regression* seeks to minimise the mean-squared error (MSE) risk, which comprises variance plus squared bias. Shrinkage regression minimises the former at the expense of the latter. The softRx package offers maximum likelihood statistical inferences and graphics to assist the user in deciding when to stop shrinking, where the shrinkage parameter can take any one of a continuous range of values. It is best seen as an alternative estimator to least-squares rather than a model selection procedure.
3. *Principal Components Regression* (PCR) is based on the use of Principal Components Analysis (PCA) to apply dimensionality reduction of data by rotating the (mean-centred data) so that the principal axes of the data hyperellipsoid coincide with the coordinate axes, and a thresholding procedure is applied to determine the effective dimension of the data. This can be coupled with a stepwise approach to regression [66].
4. *Partial Least Squares* (PLS) is an iterative empirical model building procedure which is common in the chemometrics community. Like PCR, explanatory components are derived from the data, but unlike PCR, the method takes account of the response variable and does not have a separate regression step.
5. The *H-principle*[62], originally due to Wold, is related to PLS and is based on an analogy between data fitting and the Heisenberg Uncertainty Principle of quantum mechanics. The key idea is that it is not possible to minimise both the prediction variance (e.g., RSS) and the model uncertainty (which is related to the *condition* (difficulty) of the approximation problem) simultaneously. In general, adding terms to the model causes the fit variance to decrease and the model uncertainty to increase. The H-principle algorithm uses the product of these variances as its criterion and attempts to minimise the resulting quantity (called the *prediction variance*) by means of a stepwise procedure in which rank–1 updates to the model are made at each step. The resulting model is H-optimal.

## B.2 Information theory

Following [85], model selection using the popular $C_p$, AIC, SBC, etc., is formally equivalent to choosing the best compression for the data. Information/coding interpretation of selection as data coding offers

- a consistent, alternative insight into the various criteria
- a means of comparing criteria;
- a means of generating new criteria.

Following [13], we note that entropy is a concept used in information theory; if $n$ states are possible, each characterized by a probability $p_i$, with $\sum_i p_i = 1$, then

$$S = -\sum_{k=1}^{n} p_i \log(p_i) \tag{B.8}$$

is the entropy, the lowest bound on the number of bits needed to describe all parts of the system; it corresponds to the information content of the system. This is used in data compression: entropy encoding makes use of the non-uniform occurrence of bit patterns in some quantized scheme. The relative entropy of two codes is then formally equivalent to the Kullback-Liebler discrepancy between two models. A further criterion, this time from coding theory is the Minimum Description Length (MDL) criterion, which can be shown to approximate SBC in special cases.

It should be noted that most proofs of consistency, relative bias, etc., of the various criteria do not assume a particular form for the distribution of the errors. However there is often an implicit assumption that model terms are orthogonal.

# Appendix C

# Model validation lessons from the numerical analysis of differential equations

The reliability of numerical schemes for continuous models traditionally depends on convergence as step or mesh sizes tend to zero. The errors involved in solving partial differential equations using finite elements result from the following:

Galerkin discretization error : from the polynomial approximation

quadrature error : from approximating the integrals in the Galerkin formulation

discrete solution error : from solving the discrete system of equations, being either the truncation error from an iterative method or (possibly aggravated) roundoff error from a direct method.

Standard error estimates are of the form

$$\|u - u_h\| \leq S\|h^{\alpha} Du\|$$

where $\|\|$ is a given norm, $S$ is a stability factor for the discretized problem, $D$ denotes derivatives and $\alpha$ is an exponent.

*A-posteriori* error estimates give an estimate of the local error in the problem and are typically used for grid refinement algorithms. They are of the form

$$\|\|\|u - u_h, p\| \leq S\|h^{\beta} R(u_h)\|$$

where $R(u_h)$ represents the residual if $u_h$, $S$ is a stability constant and $\beta$ is a non-negative exponent.

Conceptually an *a priori* estimate yields:

small interpolation error + stability of discrete problem $\rightarrow$ small error whereas an a posteriori estimate yields:

$$\text{small residual} + \text{stability of continuous problem} \rightarrow \text{small error}$$

The important feature to note is that the a posteriori estimates give a bound on the error in terms of a residual - which can be estimated. Thus, in a computation, it is possible to get an estimate of the local error and to control that error. This is why a posteriori error estimates have become so popular for adaptive finite element methods. For a general introduction see, for example, [42], in the context of far-field predictions see [73]

# Bibliography


[1] H. Akaike. A new look at statistical model identification. *IEEE Trans. Automat. Contr*, 19:776–723, 1974.

[2] Anon. Joint Simulation System Verification, Validation and Accreditation Strategy. Web page: http://www.jsims.mil/Product_Assurance/Product%20Assurance%20Web%20Page/vvaplan.htm.

[3] Anon. Navy Modelling and Simulation Project Team Charter. Web page: http://www.bmh.com/FPT/gen_info/Charter_Draft.html. Draft.

[4] Anon. Verification, Validation and Accreditation Technology Working Group (VV&A TWG). Web page: http://www.dmso.mil/dmso/wrkgrps/twg/vva/.

[5] Anon. VPG Verification, Validation and Accreditation. Web page: http://vpg.apg.army.mil/vva.html.

[6] G. T. Anthony, H. M. Anthony, B. Bittner, B. P. Butler, M. G. Cox, R. Drieschner, R. Elligsen, A. B. Forbes, H. Groß, S. A. Hannaby, P. M. Harris, and J. Kok. Reference software for finding Chebyshev best-fit geometric elements. *Precision Engineering*, 19:28 – 36, 1996.

[7] G. T. Anthony, H. M. Anthony, M. G. Cox, and A. B. Forbes. The parametrization of fundamental geometric form. Technical Report EUR 13517 EN, Commission of the European Communities (BCR Information), Luxembourg, 1991.

[8] O. Balci. Special issue of ACM Transactions on Modeling and Computer Simulation for Training: Foundations and Techniques. Vol. 7, No.3, 1997. Guest editor.

[9] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley and Sons, Chichester, 1984. 2nd Edition.

[10] P.M. Berthouex and L.C. Brown. *Statistics for Environmental Engineers*. CRC Press, USA, 1994.

[11] K.P. Birch and M.J. Downs. An updated Edlén Equation for the Refractive Index of Air. *Metrologia*, 30:155–162, 1993.

[12] K.P. Birch and M.J. Downs. Correction to the updated Edlén Equation for the Refractive Index of Air. *Metrologia*, 31:315–316, 1994.

[13] R.K. Bock and W. Krischer. *The Data Analysis Briefbook*. Springer, 1998. See also http://www.cern.ch/Physics/DataAnalysis/BriefBook/.

[14] Boeing. Web page: http://www.boeing.com/companyoffices/doingbiz/, 1999.

[15] W.L. Briggs. *A Multigrid Tutorial*. SIAM, Philadelphia, 1987.

[16] B. Butler. Validation of continuous models in metrology, November 23. Presentation at SSfM Continuous Modelling workshop, NPL.

[17] B. P. Butler, M. G. Cox, S. Ellison, and W. A. Hardcastle, editors. *Statistical Software Qualification*. Royal Society of Chemistry, Cambridge, 1996.

[18] B. P. Butler, A. B. Forbes, and P. M. Harris. Geometric tolerance assessment problems. In P. Ciarlini, M. G. Cox, R. Monaco, and F. Pavese, editors, *Advanced Mathematical Tools in Metrology*, pages 95–104, Singapore, 1994. World Scientific.

[19] B.P. Butler. A Framework for model Validaton and software Testing in Regression. In P. Ciarlini, M.G. Cox, F. Pavese, and D. Richter, editors, *Advanced Mathematical Tools in Metrology*, volume III, pages 158–164, Singapore, 1997. World Scientific.

[20] B.P. Butler, M.G. Cox, A.B. Forbes, S.A. Hannaby, P.M. Harris, and S.M. Hodson. *Statistics software qualification—Reference data sets*. Royal Society of Chemistry, Cambridge, 1996. edited by B P Butler, M G Cox, S L R Ellison, W A Hardcastle.

[21] B.P. Butler, V. Music, and F.J. Redgrave. Influence of orifice geometry on the simplification of transmission probability calculations. In *Proceedings of the Fourth International Vacuum Congress, Birmingham, 31 August - 4 September, 1998*, Bristol, 1999. Institute of Physics. *To appear*.

[22] CAMASE Committee. Guidelines for modelling. CAMASE News (Newsletter of Agro-ecosystems Modelling) extra edition, Web page: http://www.bib.wau.nl/camase/modguide.html, November 1995.

[23] H. Carslaw and Jaeger. *Conduction of Heat in Solids*. Clarendon, 1959.

[24] J. Cash. Software for initial value problems. Technical report, Imperial College, 1997. www.ma.ic.ac.uk/ jcash/IVP_software/finaldae/readme.html.

[25] J. Cavanaugh. Unifying the derivations of the akaike and corrected akaike information criteria. *Statistics and Probability Letters*, 31:201–208, 1997.

[26] J. Cavanaugh and A.A. Neath. Regression and time series model selection using variants of the Schwarz information criterion. *Communications in Statistics—Theory and Methods*, 26:559–580, 1997.

[27] J. Cavanaugh and R.H. Shumway. A bootstrap corrected variant of AIC for state-space model selection. *Statistica Sinica*, 7:473–496, 1997.

[28] J. Cavanaugh and R.H. Shumway. An Akaike information criterion for model selection in the presence of incomplete data. *Journal of Statistical Planning and Inference*, 67:45–65, 1998.

[29] Center for Drug Evaluation and Research, FDA. (Draft) Guidance for Industry: Population Pharmacokinetics. Section VIII Data Analysis. Web page: http://www.fda.gov/cder/guidance/1852pk.htm, September 1997.

[30] P. Clausing. The flow of highly rarefied gases through tubes of arbitrary length. *J. of Vacuum Science and Tech.*, 8(5):636–646, 1971.

[31] R.D. Cook and S. Weisberg. *Introduction to Regression Graphics*. JohnWiley, New York, 1994.

[32] M. Cox and P. Harris. Length 6: Surface texture data analysis. Technical report, Scientific Software Section, CMOT, NPL, January 1997. Unpublished.

[33] M. G. Cox, M. P. Dainton, A. B. Forbes, P. M. Harris, and E. B. Hughes. Multi-lateration co-ordinate measuring systems. Technical report, National Physical Laboratory, Teddington, 1999. *In preparation.*

[34] M.G. Cox. Development of RF coaxial impedance standards: interpolation routines for load impedance values: Study of existing interpolation routines. Issue 2.0, January 1999.

[35] M.G. Cox. The separation of component form error and instrument spindle error in roundness measurement to ultra-high accuracy using an instrument with a rotary indexing table. Technical report, NPL, 1999. *In preparation*.

[36] M.G. Cox, A.B. Forbes, P.M. Harris, and G.N. Peggs. Determining CMM behaviour from measurements of standard artefacts. Technical report, NPL, CISE, 1998.

[37] M.G. Cox, P.M. Harris, M.J.T. Milton, and P.T. Woods. Method for Evaluating Trends in Ozone Concentration Data and its Application to Data from the UK Rural Ozone Monitoring Network—Part II: Validation and Aggregation of Individual Site Results. *NPL Report (in preparation)*, 1997. Summary published as *A method for evaluating trends in ozone-concentration data and its application to data from the UK Rural Ozone Monitoring Network* in Advanced Mathematical Tools in Metrology, vol III, World Scientific, pp 171–178. 1997.

[38] P.C. Cresto. Self-Calibration with Application to CMMs Geometry Error Correction. In P. Ciarlini, M.G. Cox, R. Monaco, and F. Pavese, editors, *Advanced Mathematical Tools in Metrology*, volume 1, pages 167–174, Singapore, 1994. World Scientific.

[39] I. Daubechies. *Ten Lectures on Wavelets*. CBMS-NSF Lecture Notes No. 61. SIAM, Philadelphia, 1992.

[40] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1982.

[41] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

[42] K. Eriksson, D. Estep, P.Hansbo, and C. Johnson. Introduction to adaptive methods for differential equations. *Acta Numerica*, pages 105–158, 1995. CUP.

[43] Center for International Earth Science Information Network (CIESIN). Stratospheric ozone and human health project—model validation. Columbia University and web page: http://sedac.ciesin.org/ozone/rtm/mval.html.

[44] A. B. Forbes. Least-squares best-fit geometric elements. Technical Report DITC 140/89, National Physical Laboratory, Teddington, 1989.

[45] A. B. Forbes. Robust circle and sphere fitting by least squares. Technical Report DITC 153/89, National Physical Laboratory, Teddington, 1989.

[46] A. B. Forbes. Generalised regression problems in metrology. *Numerical Algorithms*, 5:523–533, 1993.

[47] A. B Forbes. Model parametrization. In P. Ciarlini, M. G. Cox, F. Pavese, and D. Richter, editors, *Advanced Mathematical Tools for Metrology*, pages 29–47, Singapore, 1996. World Scientific.

[48] A. B Forbes and P. M. Harris. A comparison of methods used for the calculation of effective area in the calibration of pressure balances. Technical Report CISE 2/95, National Physical Laboratory, December 1995.

[49] A. B Forbes and P. M. Harris. Estimation methods in pressure balance calibration. In G. Molinar, D. Simpson, and C. Tilford, editors, *Pressure Metrology from Ultra-High Vacuum to Very High Pressures*, 1999. To appear.

[50] A. B Forbes, P. M. Harris, and F. J. Redgrave. A discussion of methods for the estimation of expansion ratios determined by multiple expansions. In *Proceedings of the Fourth International Vacuum Congress, Birmingham, 31 August - 4 September, 1998*, Bristol, 1998. Institute of Physics. *To appear.*

[51] A. B Forbes, P. M. Harris, and F. J. Redgrove. Estimation of volume ratios determined by multiple expansions. Technical report, National Physical Laboratory, 1999. In preparation.

[52] A.B. Forbes. Least squares best fit geometric elements. In J.C. Mason and M.G. Cox, editors, *Algorithms for Approximation*, volume II. Chapman and Hall, 1990.

[53] FSA Modeling Committee. Model Validation Guidelines, with Modeling and Characterization data package. Web page: http://www.fsa.org/modeling/mod_valid.html, October 1997.

[54] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, London, 1981.

[55] G.H. Golub and C.F. van Loan. *Matrix Computations*. John Hopkins Press, Baltimore, USA, 1996.

[56] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems.*, volume 8 of *Computational Mathematics*. Springer, 1993.

[57] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential Algebraic Problems*, volume 14 of *Computational Mathematics*. Springer, 1996.

[58] E. Hairer and G. Wanner. Testset of stiff odes. Technical report, Université de Genève, 1998. www.unige.ch/math/folks/hairer/testset/testset.html.

[59] S.R. Hare. *Low frequency climate variability and salmon production*. PhD thesis, School of Fisheries,Univerity of Washington, 1996.

[60] P.M. Harris and S.P. Robinson. Acoustic Signal Modelling and its Application to the Calibration of Underwater Electroacoustic Transducers in Reverberant Laboratory Tanks. In *Underwater Acoustics Calibration and Measurement*, volume 30 of *Proceedings of the Institute of Acoustics*, 1998.

[61] A.R. Hefner. NIST Working Group on Model Validation (of circuit simulators). Web page: http://ray.eeel.nist.gov/modval.html.

[62] A. Höskuldsson. Prediction Methods in Science and Technology (H-principle). Technical University of Denmark and Web page: http://www.predict.dk/.

[63] M. Hulme. Climate Model Validation. Climate Research Unit, University of East Anglia and web page: http://www.cru.uea.ac.uk/ mikeh/research/cmodval.htm.

[64] ISO. Guide to the expression of uncertainty in measurement. Technical report, ISO, Geneva, 1993.

[65] ISO. ISO 9001: Quality Systems—Model for quality assurance in design, development, production, installation and servicing. Geneva, 1994.

[66] A. Jones. An Introduction to Multivariate Calibration. See web page: http://144.124.112.51/auj/depttalk96/index.html. From the Institute of Biological Sciences, University of Wales, Aberystwyth.

[67] W. Lioen and J. Swart. Test set for initial value problem solvers. Technical report, CWI, PO Box 94079, 1090 GB,Amsterdam,NL, 1998. www.cwi.nl/cwi/projects/IPVtestset/.

[68] H.K. Liu. High-Dimensional Empirical Linear Prediction. In P. Ciarlini, M.G. Cox, F. Pavese, and D. Richter, editors, *Advanced Mathematical Tools in Metrology*, volume III, pages 79–90, Singapore, 1997. World Scientific.

[69] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979.

[70] Mathsoft Inc., Seattle. *S-Plus 4 Guide to Statistics*, July 1997.

[71] M.R. McEwen and S.Duane. Development of a pertable graphite calorimeter for photons and electrons. Technical Report NPL CIRM 18, CIRM, National Physical Laboratory, 1998.

[72] Allan D.R. McQuarrie and Chih-Ling Tsai. *Regression and Time Series Model Selection*. World Scientific, Singapore, June 1998.

[73] P. Monk and E Süli. The adaptive computation of far-field patterns by a posteriori error estimation of linear functionals. *SIAM J. of Num Anal.*, 36(1):251–274, 1998.

[74] NAFEMS. Nafems home page. http://www.nafems.org/.

[75] NAG. G02 chapter introduction. In *NAG Fortran Library Document, Mark 17*. The Numerical Algorithms Group Ltd, Oxford, UK, 1996.

[76] National Environment Research Institute (NERI), Denmark. Initiative on ”Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes. Web page: http://www.dmu.dk/AtmosphericEnvironment/harmoni.htm.

[77] M.C. Nechyba and Y. Xu. Learning and Transfer of Human Control Strategies. Web page: http://www.cs.cmu.edu/afs/cs/project/space/www/hcs/hcs.html.

[78] NPL. National Measurement System Software Support for Metrology Programme, 1998. Brochure available from NPL.

[79] NPL. Traceability of CMM measurements: Influence of the workpiece errors on the measurement uncertainty (Virtual Workpiece). Technical report, NPL, Teddington, Middx, UK, January 1999. Deliverable under EC Contract MAT1-CT94-0076.

[80] E.H. Page, B.S. Canova, and J.A. Tufarolo. A Case Study of Verification, Validation and Accreditation for Advanced Distributed Simulation. http://ms.ie.org/page/papers/tomacs/alspPaper/baseline.html, 1997.

[81] Sandia. Computational Methods for Model Validation and System Identification for Structural Dynamics. Web page: http://sass2537.sandia.gov/BB/comp_str_dyn/model_val.html.

[82] G. A. F. Seber. *Linear regression analysis*. John Wiley and Sons, New York, 1977.

[83] A Sharov. Lecture 4: Statistical Analysis of Population Dynamics. Web page: http://www.gypsymoth.ento.vt.edu/ sharov/PopEcol/lec4/statist.html. Course: Quantitative Population Ecology at the Dept. of Entolomogy, Virginia Tech, Blacksburg, VA.

[84] W. Stahel. Robust alternatives to least squares. In *Advanced Mathematical Tools in Metrology III*, Singapore, 1997. World Scientific. To be published in this volume.

[85] R Stine. Lectures on Model selection, Bayesian model selection, Information Theory and model Selection. Department of Statistics, Wharton School, Univ of Penn, Philadelphia, May 1998. Web page: http://www-stat.wharton.upenn.edu/ bob/select.intro.pdf.

[86] D.F. Vecchia and J.D. Splett. Outlier-resistant methods for estimation and model-fitting. In *Proceedings of the International Workshop: Advanced Mathematical Tools in Metrology*, pages 143–154, Singapore, 1994. World Scientific.

[87] J.D. Victor and A. Canel. A relation between the Akaike criterion and reliability of parameter estimates, with application to nonlinear autoregressive modelling of ictal EEG. *Annals of Biomedical Eng.*, 20:167–180, 1992.

[88] S. Weisberg. *Applied Linear Regression*. JohnWiley, New York, 1985. 2nd edition.