

# Model-Based Identification of Systematic Errors in Measurements

H. de Jong, V.J. de Wit, N.J.I. Mars, P.E. van der Vet

Department of Computer Science

University of Twente

P.O. Box 217, 7500 AE Enschede, the Netherlands

Phone: +31-53-4894622 Fax: +31-53-4892927

Email: hdejong@cs.utwente.nl, wit@cs.utwente.nl, mars@cs.utwente.nl, vet@cs.utwente.nl

## Abstract

Large-scale databases and knowledge bases with property measurements obtained in scientific experiments or at observation sites have become available in the past decade. With the growing use of these so-called measurement bases, the task of ascertaining the accuracy of measurements is becoming increasingly important. We introduce an implemented method for the model-based identification of systematic errors which has been applied in a case-study in materials science. The method is formalized in terms of QR and MBD techniques, which provides correctness guarantees and facilitates extensions.

## Introduction

In the past decade large-scale databases and knowledge bases have become available to researchers working in a range of scientific disciplines. In many cases these databases and knowledge bases contain measurements of properties of (physical) objects obtained in experiments or at observation sites. As examples, one can think of databases with molecular structures in crystallography and mechanical property databases in materials science.

These large collections of measurements, henceforth called *measurement bases*, form interesting resources for scientific research. By analyzing the contents of a measurement base, one may be able to find patterns that are of practical and theoretical importance (e.g., Fayyad, Haussler & Stolorz [1996]). The discovery of a sudden change in the temperature dependency of a certain material above 1000 K, for instance, may call for a scientific explanation or affect the use of the material at high temperatures.

An important task accompanying the use of measurement bases is the identification of *systematic errors* in measurements of a certain property. The occurrence of systematic errors detracts from the accuracy of measurements and compromises any patterns found in the measurement base.

Statistical treatments of measurement focus on random errors and usually brush aside the problem of systematic errors with general recommendations like:

"The detection of ... systematic errors ... depends on the observer's alertness and knowledge of the natural, instrumental, and personal factors that can influence his procedures." (Barry [1978], p. 13). In order to identify systematic errors we have to take recourse to knowledge about the experiments in which the measurements were conducted and knowledge about the processes occurring in the physical systems investigated. Although in some situations it will suffice to quickly locate a systematic error by means of heuristic rules acquired through experience, in others it is necessary to gain a deeper understanding of the source of a systematic error by using appropriate models of the experiments. The latter situations will interest us here.

Given the large and increasing scale of measurement bases, in which a corresponding high numbers of systematic errors are expected, it would be highly desirable to develop computer support for the *model-based identification of systematic errors*. We will introduce a method which has been implemented in a system called KIMA. Given a property measurement, models of the physical systems created and controlled in the actual and in the ideal experiment, and supplementary measurements of quantities, our method generates all possible systematic errors in the reported value of the property.

The method employs concepts and techniques from the fields of qualitative reasoning (QR) and model-based diagnosis (MBD). The algorithm generating the possible systematic errors is formalized in terms of QR techniques for the qualitative simulation and comparative analysis of dynamical systems (Kuipers [1994]; Weld [1990]; de Jong & van Raalte [1997]). In addition, it incorporates the basic MBD principle of predicting behaviors from the models of a physical system and matching the predictions with available measurements (Davis & Hamscher [1988]). The work described in this paper thus establishes an interesting connection between measurement analysis problems on the one hand and QR and MBD techniques on the other.

After an introduction of some basic concepts, we will briefly review the QSIM and CEC\* techniques for

qualitative simulation and comparative analysis. The next section defines the algorithm for error identification and shows that it produces all possible genuine systematic errors but occasionally spurious errors as well. We will then present the results of the application of the method in a case-study in a realistic domain, the fracture strength of brittle materials. This is followed by a discussion of the method and its results in the context of related work.

## Experimental systems and systematic errors

The measurements that are stored in a measurement base are assumed to be measurements of properties of physical objects obtained in scientific experiments. An *experiment* will here be viewed as the activity of creating and sustaining a controlled physical system, also referred to as an *experimental system* (de Jong [1998]). Control over a physical system is achieved by experimenters who attempt to actively create and maintain the structure of the system and regulate its behavior by imposing certain *experimental conditions*.

*Measurements* are determinations of the quantities of an experimental system at a certain time-point in the interval on which the system is investigated. If several quantities are measured at a time-point, we obtain a (partially) *measured state* of the system. The measurements are not made in an arbitrary way, but they are directed at the determination of *properties* of a physical object, such as the fracture strength of a material. A property measurement can be seen as the determination of the value of a certain quantity when an experimental system has been brought into a certain state.

A *systematic error* is the (significant) deviation of the measured value from the true value of a property. The true value is the value that would have been obtained if the measurement had been carried out on an *ideal experimental system*, that is, an experimental system with an ideal structure and an ideal behavior exhibited under ideal conditions. What counts as 'ideal' is dependent upon the aim of the measurement, e.g., the approximation of an idealized theoretical situation or specific service conditions. It will often be impossible to realize an ideal experimental system in an experiment. In other words, the ideal experimental system is a *hypothetical* system and the true value of the property a *hypothetical* value.

The causes of a systematic error are assumed to lie in differences between the structure of the actual and ideal experimental system, and differences in the experimental conditions imposed upon the systems. The task of *error identification* is concerned with analyzing these differences in order to predict possible systematic errors.

## Simulation and comparative analysis of experimental systems

The conceptualization of experimental systems as controlled physical systems has the advantage of suggesting a fruitful way of modeling the structure and behavior of experimental systems. We will use differential equations, more specifically *qualitative differential equations* (QDEs), for this purpose (Kuipers [1994]). QDEs are appropriate devices for describing experimental systems, since in many cases much of the knowledge will be qualitative in nature, especially when certain idealized circumstances cannot be realized.

Once an experimental system has been modeled by a QDE, a *qualitative simulation* algorithm can be used to infer the possible qualitative behaviors  $QB_1, \dots, QB_m$  of the system from an initial qualitative state  $QS(init)$  which represents the experimental conditions. We have employed the QSIM algorithm for this purpose (Kuipers [1994]). A qualitative behavior is a sequence of qualitative states of the system.

The QSIM algorithm can be seen as a theorem prover deriving theorems of the following form:

$$QDE \wedge QS(init) \rightarrow QB_1 \vee \dots \vee QB_m.$$

QSIM is sound and incomplete in that the disjunction of possible qualitative behaviors contains all genuine behaviors of the system, but occasionally spurious behaviors as well (see theorem 6 and the discussion in sec. 5.6.2 in Kuipers [1994]).

In order to compare the behavior of two experimental systems described by QDEs one can perform a *comparative analysis*. Given a qualitative model and behavior of the first system ( $QDE$  and  $QB$ ) and a qualitative model and behavior of the second system ( $\hat{QDE}$  and  $\hat{QB}$ ), the comparative analysis algorithm CEC\* derives the possible comparative behaviors  $CB_1, \dots, CB_n$  of the two systems from an initial comparative state  $CS(init)$ . The comparative behaviors are contained in a comparative environment (de Jong, Mars & van der Vet [1996]; de Jong & van Raalte [1997]; de Jong [1998]).

Whereas a qualitative behavior consists of a sequence of qualitative states of a *single* system, a comparative behavior is a graph of comparative states of *two* systems. Each comparative state  $CS(pc_i)$  represents the difference in value of shared quantities of the experimental systems at a pair of comparison  $pc_i$ . A pair of comparison is a pair of time-points from the qualitative behavior of the first and second system at which the systems can be (meaningfully) compared. The difference between a shared system variable  $q$  at pair of comparison  $pc_i$  is expressed as a relative value  $RV(q, pc_i)$  which can be  $\uparrow$  (greater),  $\parallel$  (equal), or  $\downarrow$  (smaller).  $RV(q, pc_i) = \uparrow$  abbreviates to  $q \uparrow_{pc_i}$ .

The pairs of comparison arising from the comparison of two qualitative behaviors are partially ordered. This ordering contains a pair of comparison  $pc_0$  referring to the initial time-points of the behaviors and

a pair of comparison  $pc_n$  referring to the final time-points. The comparative behaviors produced by CEC\* show either how given differences in response at  $pc_n$  can be explained by differences in initial conditions at  $pc_0$  (*explanatory comparative analysis*) or which differences in response at  $pc_n$  can be predicted from given differences in initial conditions at  $pc_0$  (*predictive comparative analysis*).

CEC\* derives theorems of the following form:

$$[QDE \wedge QB] \wedge [Q\hat{D}E \wedge Q\hat{B}] \wedge CS(init) \rightarrow CB_1 \vee \dots \vee CB_n.$$

Just like QSIM, CEC\* is sound and incomplete. It generates all genuine comparative behaviors, but is not guaranteed to exclude spurious comparative behaviors (theorems 15 and 16 in de Jong [1998]).

An important feature of CEC\* is that the technique is able to compare behaviors of systems with a different structure, that is, of systems described by different QDEs. This ability is important for error identification, since an experiment may be carried out in such a way as to give rise to an experimental system structurally different from the ideal experimental system.

### Error identification method

Error identification starts with a property measurement from a measurement base. The property measurement is accompanied by candidate models of the experimental system investigated, candidate descriptions of the experimental conditions imposed upon the system, and a sequence of measured states. It is confronted with a specification of the ideal property measurement in the form of a model of the ideal experimental system and a description of the ideal experimental conditions. The result of error identification is a set of possible systematic errors in the property measurement. The systematic errors are explained in terms of differences in the structure and differences in the experimental conditions of the systems. Fig. 1 provides a schematic overview of the error identification method.

### Candidate models and candidate behaviors of experimental system

Due to a lack of information about the execution of an experiment, it is often not possible to unambiguously determine the structure of the experimental system and the conditions under which it has been investigated. For instance, a bend test might be performed on a material specimen with or without surface damage, at or well above room temperature. Therefore, the property measurement specifies a set  $\hat{C}\hat{M}$  of tuples of a *candidate model* of the experimental system and *candidate experimental conditions*. Each tuple represents a different possible structure of the system and a different possible set of experimental conditions.

For every pair  $\langle Q\hat{D}E, Q\hat{S}(init) \rangle \in \hat{C}\hat{M}$ , one can predict a set of *candidate behaviors* of the experimental system by performing a qualitative simulation and

then eliminating the qualitative behaviors which do not agree with the measured states of the system. The candidate behaviors are alternative qualitative descriptions of the behavior of the physical system investigated in the experiment.

In order to count as a candidate behavior of the experimental system, a behavior  $QB$  resulting from simulation must be consistent with the sequence of measured states of the system. This consistency check lies at the heart of work on measurement interpretation (Forbus [1987]; DeCoste [1991]; Dvorak & Kuipers [1991]) and is here performed by a simple algorithm given in de Jong [1998]. The information about the experimental system provided by the measured states may occasionally rule out *all* qualitative behaviors for a particular candidate model  $Q\hat{D}E$ . Assuming that the measurements are correct, this implies that the experimental system cannot adequately be described by  $Q\hat{D}E$  and  $Q\hat{S}(init)$ , and we must remove the tuple  $\langle Q\hat{D}E, Q\hat{S}(init) \rangle$  from the set  $\hat{C}\hat{M}$  of candidate models.

### Model and behavior of ideal experimental system

A property measurement is evaluated by comparing the measured value with the value that would have been obtained if the measurement had been conducted on a reference system, the ideal experimental system. The behavior of the ideal experimental system is derived from a user-specified pair  $\langle QDE, QS(init) \rangle$ , representing the structure and experimental conditions of the ideal experimental system. If several qualitative behaviors result from the simulation of  $QDE$  with  $QS(init)$ , the user is invited to select one of these.

### Systematic errors in measurements

From the information provided with the property measurements we thus derive a triple  $\langle QDE, QS(init), QB \rangle$  of the model, conditions, and behavior of the ideal experimental system and a set  $\hat{C}\hat{M}\hat{B}$  of triples  $\langle Q\hat{D}E, Q\hat{S}(init), Q\hat{B} \rangle$  of a candidate model, candidate conditions, and candidate behavior of the actual experimental system.

A property value will be seen as the value of a certain *definition quantity*  $q_{def}$  when the experimental system has reached its *final state*. A systematic error can hence be viewed as a differential final response of the actual and ideal experimental system. Error identification can then be reformulated as performing a predictive comparative analysis. The goal of this analysis is to predict whether the response of the experimental system actually investigated would differ from the response of the ideal experimental system if the latter were realized in an experiment. The model  $QDE$  and behavior  $QB$  of the ideal experimental system have to be compared with every candidate model-behavior triple in  $\hat{C}\hat{M}\hat{B}$ , since different assumptions about the

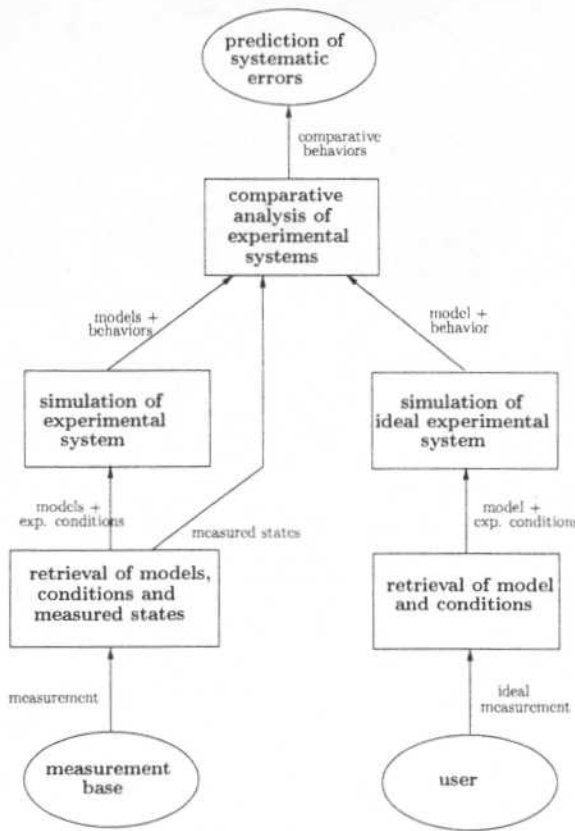


Figure 1: Schematic overview of model-based error identification.

model and behavior of the actual physical system may yield different predictions of the systematic error.

A predictive comparative analysis by means of CEC\* starts with a set  $CS(init)$  of RVs at the first pair of comparison  $pc_0$  which represents differences in the experimental conditions of the systems. In the case of error identification, initial RVs can be derived by comparing measurements of the initial state of the actual system with a quantitative specification of the ideal experimental conditions. If the values of a quantity  $q$  are significantly different according to a statistical or heuristic criterion (de Jong [1998]), we obtain  $q \downarrow_{pc_0}$  or  $q \uparrow_{pc_0}$ , depending on the sign of the difference. Otherwise, we have  $q \parallel_{pc_0}$ . Additional RVs in  $CS(init)$  might be directly supplied by the user.

For each combination of the model  $QDE$  and the behavior  $QB$  of the ideal system and a candidate model  $\hat{QDE}$  and a candidate behavior  $\hat{QB}$  of the actual system, CEC\* will return a comparative environment consisting of comparative behaviors  $CB_1, \dots, CB_m$ . A possible systematic error in a measured property value appears as a relative value  $\uparrow$  or  $\downarrow$  for  $q_{def}$  at  $pc_n$  in a comparative behavior.<sup>1</sup> More precisely, the RV of  $q_{def}$  is a qualitative abstraction of the systematic error.

<sup>1</sup>If the RV is  $\parallel$ , the systematic error in the measured property value equals 0. Adhering to common usage of the

error. The comparative behavior explains how deviations from the ideal experimental conditions at  $pc_0$ , added to deviations from the structure and behavior of the ideal experimental system, result in a systematic error at  $pc_n$ .

The notion of *systematic error* is formally captured as follows.

**Definition 1 (Systematic error)** Given an ideal property measurement and an actual property measurement. The experimental systems of the property measurements are described by a model-behavior triple  $\langle QDE, QS(init), QB \rangle$  and a set of candidate model-behavior triples  $\hat{CMB}$ , respectively. Suppose a predictive comparative analysis is carried out involving  $\langle QDE, QS(init), QB \rangle$  and  $\langle \hat{QDE}, \hat{QS}(init), \hat{QB} \rangle \in \hat{CMB}$ . In addition, we have a set  $CS(init)$  of initial RVs at the first pair of comparison  $pc_0$ . The comparative environment constructed has comparative behaviors  $CB_1, \dots, CB_m$ .

Let  $q_{def}$  be the definition quantity of the property. Each comparative behavior  $CB_k$ ,  $1 \leq k \leq m$ , in which  $q_{def} \uparrow_{pc_n}$  or  $q_{def} \downarrow_{pc_n}$  indicates a possible systematic error in the measured value of the property.  $q_{def} \uparrow_{pc_n}$

term, we will say that there is no systematic error in this case.

or  $q_{def} \Downarrow_{pc_n}$  is the qualitative abstraction of this systematic error.  $\square$

It is important to emphasize that *different* comparative behaviors in a comparative envisionment may point at *different* systematic errors, even when they have the same RV.  $RV(q_{def}, pc_n)$  is a qualitative abstraction of a numerical systematic error which is not guaranteed to be the same in different comparative behaviors. If in one melting-temperature experiment the specimen reacts with its container (a source of systematic error) and in another the specimen reacts with its container and has a low purity in addition (another source of systematic error), there will be a systematic error in both observed melting temperatures. However, these systematic errors do not have to be the same.

### Algorithm for error identification

The algorithm for error identification given below summarizes how a measurement can be evaluated by detecting possible systematic errors in the reported value of the property. Notice that since an ideal experiment is assumed to be a hypothetical experiment, no measured states of the experimental system are available.

**Algorithm 1 (Error identification)** Given an actual property measurement and an ideal property measurement. The ideal property measurement refers to a hypothetical experimental system described by the qualitative model  $QDE$  and considered under experimental conditions  $QS(init)$ . The actual property measurement specifies the set  $\hat{CM}$  of tuples of candidate models and candidate conditions, and the sequence  $\hat{MSS}$  of measured states. The following algorithm detects possible systematic errors in the measured value of the property:

**Step 1** Perform a qualitative simulation for every tuple  $\langle \hat{QDE}, \hat{QS}(init) \rangle \in \hat{CM}$ . Call the resulting sets of candidate model-behavior triples  $\hat{CMB}$ . Similarly, determine a behavior  $QB$  for the ideal experimental system from  $QDE$  and  $QS(init)$ .

**Step 2** Use the measured states  $\hat{MSS}$  to reduce the sets of candidate model-behavior triples  $\hat{CMB}$ . Candidate behaviors and models that are not consistent with the measurements are eliminated.

**Step 3** Perform a predictive comparative analysis for the model-behavior triple  $\langle \hat{QDE}, \hat{QS}(init), QB \rangle$  and every candidate model-behavior pair  $\langle \hat{QDE}, \hat{QS}(init), \hat{QB} \rangle \in \hat{CMB}$ . The initial comparative state information  $CS(init)$  is obtained from the ideal experimental conditions and the measured states  $\hat{MSS}$  supplemented by user-specified RVs.

**Step 4** Return the comparative envisionments generated in the previous step. Each comparative behavior in which  $q_{def} \Uparrow_{pc_n}$  or  $q_{def} \Downarrow_{pc_n}$  indicates a possible systematic error.

The algorithm may occasionally produce empty comparative envisionments. Under the assumptions that the measurements are correct and the specification of the ideal experimental system is consistent, we must conclude that the combination of a candidate model  $\hat{QDE}$  and candidate behavior  $\hat{QB}$  does not adequately describe the structure and behavior of the experimental system realized in the experiment.

### Properties of the algorithm

The reliability of the algorithm for error identification can be assessed by inquiring into its ability to find all and only possible systematic errors in a measured property value. Obviously, a systematic error originating from a spurious comparative behavior is not possible, since a spurious comparative behavior does not adequately describe differences in the behaviors of the actual and ideal experimental system. This insight underlies the definition of *spurious* and *genuine* systematic errors.

**Definition 2 (Spurious systematic error)** Given a comparative behavior  $CB$  indicating a possible systematic error by predicting that  $q_{def} \Uparrow_{pc_n}$  or  $q_{def} \Downarrow_{pc_n}$ . The systematic error is spurious, iff  $CB$  is a spurious comparative behavior. A genuine systematic error is a systematic error that is not spurious.  $\square$

This definition allows one to prove the following guarantees on the outcome of an error identification.

**Proposition 1** The algorithm for error identification generates all genuine systematic errors in a measured property value.  $\square$

**Proposition 2** The algorithm for error identification may occasionally generate spurious systematic errors in a measured property value.  $\square$

The propositions are obvious consequences of the soundness and incompleteness properties of QSIM and CEC\* that we as mentioned above.

The computational complexity of the error identification algorithm is determined by the computational complexity of CEC\*. The number of comparative states generated by the algorithm, a good metric of its complexity, is the product of the number of comparative envisionments generated and the number of comparative states generated to construct each of them. The first number is a product of the sizes of the sets  $\hat{CMB}$  and  $\hat{CMB}$ , whereas the second number is of the order  $\mathcal{O}(n3^{rq})$ , with  $n$  representing the number of pairs of comparison,  $q$  the (maximum) number of shared quantities of the experimental systems, and  $r$  the maximum number of predecessors in the graph of ordered pairs of comparison (de Jong [1998]).

Although the worst-case behavior of the error identification algorithm is exponential, the average-case behavior will be much more agreeable. The estimation of the number of comparative states by  $\mathcal{O}(n3^{rq})$  is quite conservative, since it assumes that the variables are not

constrained in any way, a rare event. This is confirmed by our practical experience with the algorithm, which will be discussed below.

## Results

The error identification algorithm has been fully implemented in Common Lisp as a part of the KIMA system for Knowledge-Intensive Measurement Analysis (de Jong [1998]; de Wit & de Jong [1998]). The KIMA system is built on top of the implementations of QSIM and CEC\* whose main functions are repeatedly called. Apart from error identification, the system also carries out the tasks of conflict detection and conflict resolution. Roughly, a conflict between two property measurements is resolved by systematically comparing the candidate models and candidate behaviors of the two experimental systems in question.

KIMA has been successfully applied on realistic though simplified problems in a case-study in the domain of materials science: the analysis of measurements of the fracture strength of brittle materials. We have used KIMA to identify possible systematic errors in a dozen of strength measurements of a brittle material assumed to have been obtained in tension tests and four-point bend tests (de Jong [1998]). KIMA has been able to reproduce a number of interesting empirically established relations, pointed out by domain scientists, between measured property values and features of the experiments. We will show how it deals with an example of a systematic error in a strength measurement from a tension test (fig. 2).

The fracture strength of a brittle material is the nominal stress  $\sigma_a$  applied at the initiation of fracture. A systematic error in a measured strength value can be traced back to deviations from desired micro- and macro-structural characteristics of the material samples and deviations from the preferred loading conditions in the tension test. For example, a sample might be damaged at the surface as a consequence of certain machining operations. Further, it might be eccentrically instead of concentrically loaded, which causes a non-uniform stress distribution (fig. 2(b)-(c)). Surface damage and eccentric loading tend to give rise to measured fracture strengths that are too low.

By studying the literature on brittle fracture (e.g., Davidge [1979]) and tension tests (e.g., Marschall & Rudnick [1974]), and by additional consultation of a domain expert, we have constructed a space of qualitative models and associated initial qualitative states. Each QDE in the model space is a possible description of the structure of a physical system created in a tension test and each initial qualitative state  $QS(init)$  is a possible description of the experimental conditions under which the system evolves.

Consider a measurement of the strength of alumina ( $400 \pm 5$  MPa) which has been obtained in an experiment in which the specimens are damaged, and hence are likely to have fractured through a large machining

crack at the surface. In addition, eccentric loading of the specimens may have occurred. This gives rise to the following candidate models and candidate conditions for the experimental system (fig.3):

$$\hat{CM} = \{ \langle Q\hat{DE}_{ttsurf}, \hat{QS}(init)_{ttsurf} \rangle, \\ \langle Q\hat{DE}_{tteccsurf}, \hat{QS}(init)_{tteccsurf} \rangle \},$$

In the ideal case, specimens without surface damage are tested under concentric loading. Since even without machining damage fracture is likely to initiate at the surface of the specimen,  $QDE_{ttsurf}$  describes the structure of the ideal experimental system. Except for surface damage and possibly eccentric loading, the actual and ideal experiment are believed to have been carried out in the same way.

After simulation with the appropriate experimental conditions, each of the candidate models of the actual experimental system gives rise to a single candidate behavior which is assumed to be consistent with the measured states. The behaviors predicted from the models agree with each other as to the main features of the dynamics of the system (see fig. 4 for a few key quantities). As can be seen, the stress and strain increase until the the maximum stress at the crack tip reaches the theoretical strength ( $\sigma_m = \sigma_{th}$ , the Orowan criterion), after which the material fails almost instantaneously. The fracture stress is the value of  $\sigma_a$  at  $t_1$ . The behavior of the ideal experimental system is also given by fig. 4.

The possible systematic errors in the strength measurement are identified by comparing the model and behavior of the ideal system with the candidate models and candidate behaviors of the actual system. Two comparisons need to be made: (1) a concentrically loaded sample with no surface damage vs. a concentrically loaded sample with surface damage and (2) a concentrically loaded sample with no surface damage vs. an eccentrically loaded sample with surface damage. In the former case we compare two behaviors that are both derived from  $QDE_{ttsurf}$ , whereas in the latter case we compare a behavior derived from  $QDE_{ttsurf}$  with a behavior derived from  $QDE_{tteccsurf}$ . In both cases CEC\* finds two pairs of comparison:  $pc_0$  involving the initial time-points and  $pc_1$  involving the final time-points of the behaviors. The RVs in  $CS(init)$  are derived from the measured states and from information about the experiment:  $s_c \uparrow_{pc_0}, E \parallel_{pc_0}, l_0 \parallel_{pc_0}, d \parallel_{pc_0}, \gamma \parallel_{pc_0}, lr \parallel_{pc_0}, a_0 \parallel_{pc_0}$ .<sup>2</sup>  $s_c \uparrow_{pc_0}$  accounts for the difference in surface condition and is justified by the fact that cracks produced by machining damage tend to be larger than inherent cracks (Davidge [1979]).

Comparative analyses performed with this input lead to the comparative behaviors shown in fig. 5(a)

<sup>2</sup>In order not to clutter the envisionment with comparative behaviors expressing distinctions that do not add much to the overall picture, the crack shape parameter  $s_c = c/\rho$  has been used here.

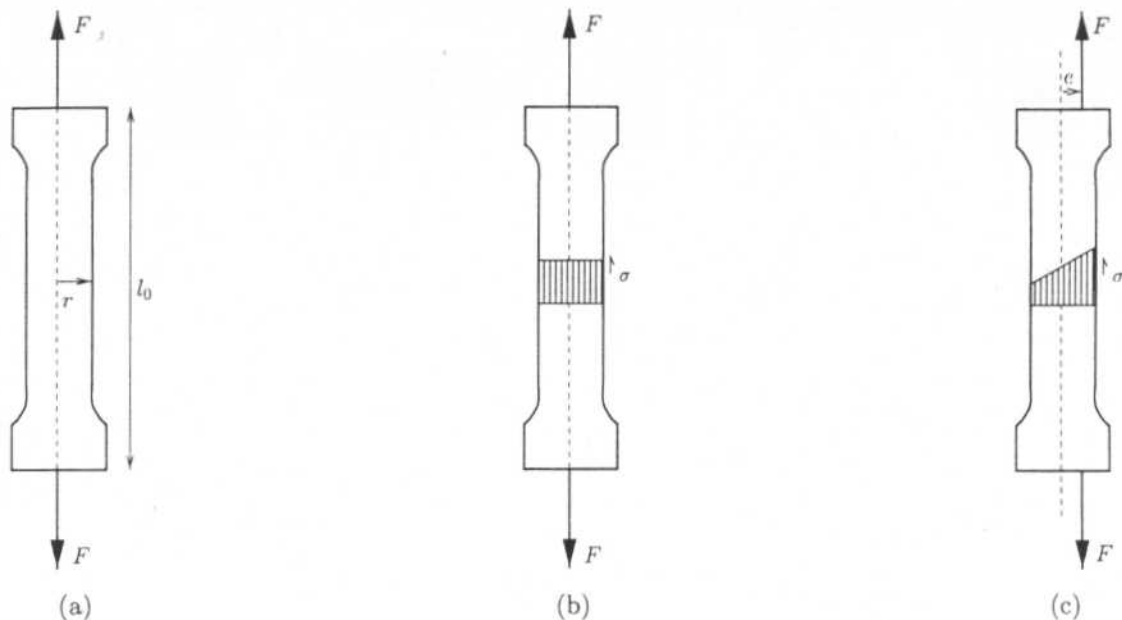


Figure 2: (a) Schematic representation of the tension test. (b)-(c) Stress distribution achieved in a concentrically and eccentrically loaded specimen, respectively (adapted from Marschall & Rudnick [1974]).

and (b). In both instances, the measured property value is predicted to be lower than the true value. In (b) the effect of eccentric loading is seen to amplify the effect of surface damage. Both comparative behaviors hypothesize the relative duration  $RV(T)$  of the intervals between  $pc_0$  and  $pc_1$  to be  $\downarrow$ . In other words, it would take longer in the ideal experiment to break the sample.

Now suppose instead that the specimens used in the actual experiment are stiffer than desired, i.e.  $E \uparrow_{pc_0}$ . This might be caused by a lower volume fraction porosity of the specimens (Dörre & Hübner [1984]). On repeating the comparative analyses with the new  $CS(init)$ , the envisionments in fig. 5(c) and (d) are obtained. Whereas a higher elasticity modulus increases  $\sigma_{th}$ , and thus tends to make the measured fracture strength too high, surface damage and eccentric loading tend to make it too low. The effect of the higher elasticity modulus counteracts the effects of surface damage and eccentric loading, which causes ambiguities. As shown in the figure, there may be a positive systematic error in the measured value ( $\sigma_a \uparrow_{pc_1}$ ), a negative systematic error ( $\sigma_a \downarrow_{pc_1}$ ), or no systematic error at all due to the masking of one error by another ( $\sigma_a \parallel_{pc_1}$ ). In a similar way, ambiguities about the relative durations of the ideal and actual experiments arise.

Whereas in the first example only two possible systematic errors are identified (both in the same direction), the second example leads to a sum total of 12 possible systematic errors (in opposite directions). On the one hand, the potentially large number of ambigu-

ities points at a weakness of using purely qualitative knowledge about the experimental systems. On the other hand, the complexity of the example at hand should be emphasized. The actual experiment may deviate from the ideal experiment in three respects: surface damage, eccentric loading, and too high stiffness. Given the available information, all systematic errors generated by the algorithm are *real* systematic errors.

## Discussion and related work

An interpretation of the model-based identification of systematic errors within an MBD framework helps to clarify both the problem and the method. Generally speaking, model-based diagnosis is concerned with finding explanations for deviations of the behavior of an observed system from that of a known reference system. The MBD approach proceeds by hypothesizing fault models and fault conditions for the observed system, making predictions from these models and conditions, and matching the predictions with the measured states of the system. There are many ways to realize this general approach when diagnosing dynamical systems, using a variety of techniques, such as semi-quantitative simulation (Dvorak & Kuipers [1991]), qualitative simulation and comparative analysis (Neitzke [1997]), or state consistency checking without simulation (Malik & Struss [1996]).

Though closely related, the problem addressed here differs in two essential aspects from the problems usually addressed by MBD. In the first place, the reference system used for error analysis is a hypothetical

$$\begin{aligned}
\epsilon_a &= \frac{1}{E} \sigma_a \\
\epsilon_a &= \frac{l - l_0}{l}, \quad \sigma_a = \frac{F}{A} \\
\sigma_c &= \sigma_a \\
\frac{dl}{dt} &= rl \\
\sigma_{th} &= h\left(\frac{\gamma E}{a_0}\right), \quad h \in M_0^+ \\
\sigma_m &= 2\sqrt{2}\sigma_c f\left(\frac{c}{\rho}\right), \quad f \in M_0^+ \\
&(QDE_{ttsurf})
\end{aligned}$$

$$\begin{aligned}
\epsilon_a &= \frac{1}{E} \sigma_a \\
\epsilon_a &= \frac{l - l_0}{l}, \quad \sigma_a = \frac{F}{A} \\
\sigma_c &= \mathbf{f_{ec}} \sigma_a \\
\mathbf{f_{ec}} &= 1 + \frac{4e}{r} x_{rel} \\
x_{rel} &= \frac{x}{r} \\
\frac{dl}{dt} &= rl \\
\sigma_{th} &= h\left(\frac{\gamma E}{a_0}\right), \quad h \in M_0^+ \\
\sigma_m &= 2\sqrt{2}\sigma_c f\left(\frac{c}{\rho}\right), \quad f \in M_0^+ \\
&(\mathbf{QDE_{tteccsurf}})
\end{aligned}$$

Figure 3:  $QDE_{ttsurf}$  describes a specimen which is loaded along the neutral axis in a tension test, whereas  $QDE_{tteccsurf}$  describes a specimen which is eccentrically loaded (differences with the former in **bold**). The variables have the following interpretation:  $\sigma_a$  nominal applied stress,  $\sigma_c$  applied stress near crack,  $F$  applied force,  $A$  cross-sectional area of specimen,  $l$  and  $l_0$  instantaneous and initial length,  $r$  radius,  $\epsilon_a$  nominal strain,  $c$  crack half-length,  $\gamma$  surface energy per unit area,  $E$  Young's modulus,  $\sigma_{th}$  theoretical strength,  $\sigma_m$  maximum stress at crack tip,  $a_0$  interatomic distance,  $\rho$  crack tip radius,  $e$  eccentricity,  $x$  distance between crack and neutral axis,  $f_{ec}$  stress concentration factor,  $lr$  elongation rate. The crack shape parameter  $s_c = \rho/c$  is also used.

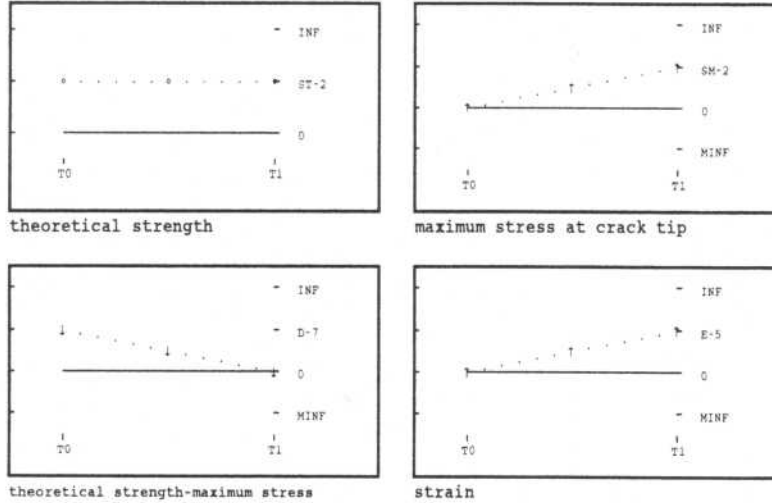


Figure 4: Qualitative behavior obtained by simulation of the models in fig. 3.

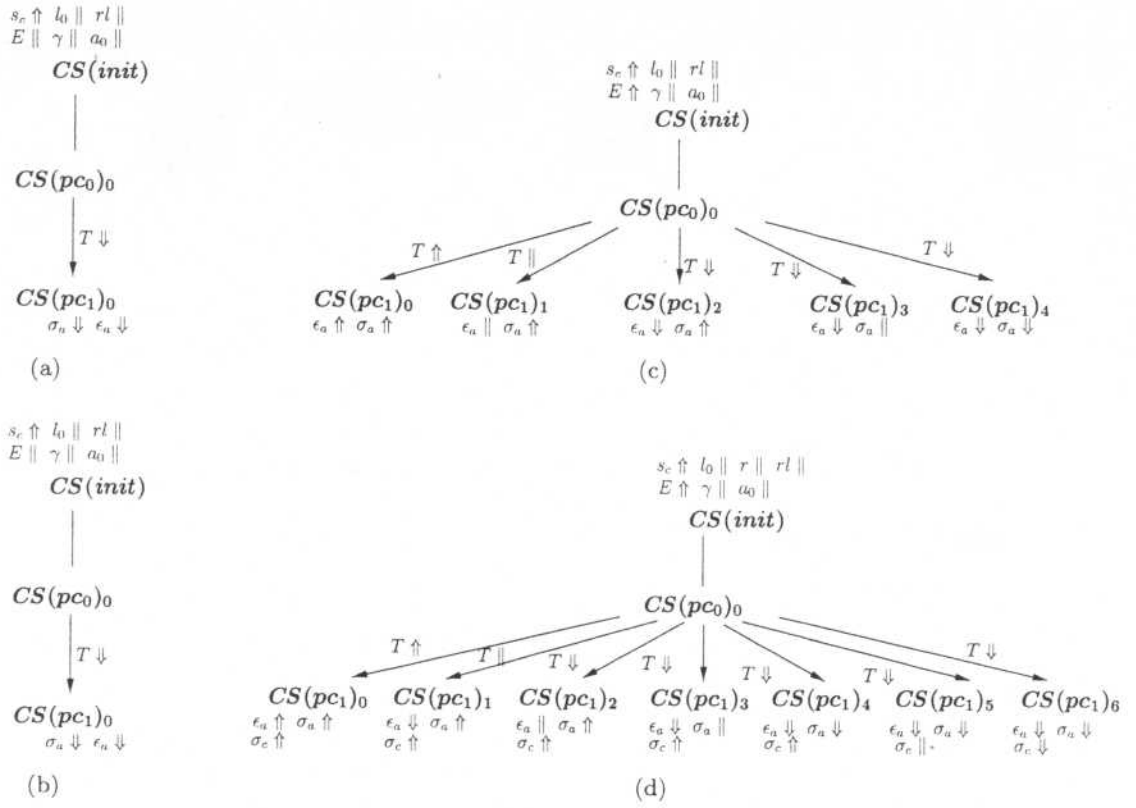


Figure 5: The comparative envisionments arising from the comparison of (a) a concentrically loaded sample with no surface damage ( $QDE_{ttsurf}$ ) vs. a concentrically loaded sample with surface damage ( $Q\hat{D}E_{ttsurf}$ ) and (b) a concentrically loaded sample with no surface damage ( $QDE_{ttsurf}$ ) vs. an eccentrically loaded sample with surface damage ( $Q\hat{D}E_{ttsurf}$ ). In (c) and (d) the same comparisons are carried out with the additional deviation  $E \uparrow$ . A few distinctive RVs are indicated at the comparative states.

system. Since the ideal experiment has not been carried out, the true value of the property is not known and a discrepancy between the true value and measured value cannot be directly detected. The unavailability of the true value leads to a second difference, a difference in the kind of solution that is expected. Whereas model-based diagnosis attempts to *explain* observed deviations from the behavior of the reference system, model-based error identification is directed at the *prediction* of deviations that would be observed if the ideal experiment were carried out.

The formalization of model-based error identification within a QR and MBD framework has the advantage of suggesting improvements and extensions of the method. One improvement would be the generalization of the method to a semi-quantitative approach, in which quantitative knowledge about the actual and ideal physical system, when available, is taken advantage of. This would allow one to reduce the number of ambiguities, and thus the number of possible systematic errors, and derive more precise estimations of the errors. The formalization of the method in terms of mathematically well-founded QR techniques

facilitates this extension. Techniques for the semi-quantitative simulation of dynamical systems already exist (e.g., Berleant & Kuipers [1997]), and de Jong & van Raalte [1997] suggest how to generalize CEC\* to a semi-quantitative approach.

The remark in Davis & Hamscher [1988] that “all model-based reasoning is only as good as the model” points at an aspect of model-based error identification that has been largely neglected in this paper. The systematic errors found by the algorithm depend on the specific models that are used to describe the experimental systems. This raises a host of modeling issues, as we need to configure adequate models from a description of the experiment and from background knowledge about the domain and the test. Together, the models in the model space should cover all sources of error that are currently deemed to be relevant (bearing in mind that the list is open-ended in principle). Work on the automated modeling of physical systems (e.g., Iwasaki & Levy [1994]; Nayak [1995]; Farquhar [1994]; Falkenhainer & Forbus [1991]) seems to provide an interesting starting-point for the automatic generation of models of experimental systems.

## Conclusions

An implemented method for the model-based identification of systematic errors in measurement bases has been introduced and successfully applied to a case-study in materials science. The formalization of the algorithm in terms of QR and MBD techniques has the advantage of providing guarantees on the results of the method and facilitating improvements and extensions. In time, a tool for the identification of systematic errors could become part of developing computer-supported discovery environments in science (de Jong & Rip [1997]), alongside tools for model building, model revision, and data mining.

**Acknowledgements** The authors would like to thank Louis Winnubst and Jeroen Nijhuis for their contributions to this paper.

## References

- B.A. Barry [1978], *Errors in Practical Measurement in Science, Engineering, and Technology*, John Wiley & Sons, New York.
- D. Berleant & B. Kuipers [1997], "Qualitative and quantitative simulation: Bridging the gap," *Artificial Intelligence* 95, 215–256.
- R.W. Davidge [1979], *Mechanical Behaviour of Ceramics*, Cambridge University Press, Cambridge.
- R. Davis & W.C. Hamscher [1988], "Model-based reasoning: troubleshooting," in *Exploring Artificial Intelligence: Survey Talks from the National Conferences on Artificial Intelligence*, H.E. Shrobe, ed., Morgan Kaufmann, San Mateo, CA, 297–346.
- D. DeCoste [1991], "Dynamic across-time measurement interpretation," *Artificial Intelligence* 51, 273–341.
- E. Dörre & H. Hübner [1984], *Alumina: Processing, Properties, and Applications*, Springer-Verlag, Berlin.
- D. Dvorak & B. Kuipers [1991], "Process monitoring and diagnosis, a model-based approach," *IEEE Expert* 5, 67–74.
- B. Falkenhainer & K.D. Forbus [1991], "Compositional modeling: Finding the right model for the job," *Artificial Intelligence* 51, 95–143.
- A. Farquhar [1994], "A Qualitative Physics Compiler," in *Proceedings of the 12th National Conference on Artificial Intelligence, AAAI-94*, AAAI Press/MIT Press, Menlo Park, CA/Cambridge, MA, 1168–1174.
- U. Fayyad, D. Haussler & P. Stolorz [1996], "Mining scientific data," *Communications of the ACM* 39, 51–57.
- K.D. Forbus [1987], "Interpreting observations of physical systems," *IEEE Transactions on System, Man, and Cybernetics* 13, 350–359.
- Y. Iwasaki & A.Y. Levy [1994], "Automated model selection for simulation," in *Proceedings of the 12th National Conference on Artificial Intelligence, AAAI-94*, AAAI Press/MIT Press, Menlo Park, CA/Cambridge, MA, 1183–1190.
- H. de Jong [1998], "Computer-supported analysis of scientific measurements," University of Twente, Memorandum UT-KBS-98-05, Enschede, the Netherlands.
- H. de Jong, N.J.I. Mars & P.E. van der Vet [1996], "CEC: Comparative analysis by envisionment construction," in *Proceedings of the 12th European Conference on Artificial Intelligence, ECAI-96*, W. Wahlster, ed., John Wiley & Sons, Chichester, 476–480.
- H. de Jong & F. van Raalte [1997], "Comparative analysis of structurally different dynamical systems," in *Proceedings of the 15th International Joint Conference on Artificial Intelligence, IJCAI-97*, M.E. Pollack, ed., Morgan Kaufmann, San Francisco, CA, 486–491.
- H. de Jong & A. Rip [1997], "The computer revolution in science: Steps towards the realization of computer-supported discovery environments," *Artificial Intelligence* 91, 225–256.
- B. Kuipers [1994], *Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge*, MIT Press, Cambridge, MA.
- A. Malik & P. Struss [1996], "Diagnosis of dynamic systems does not necessarily require simulation," in *Working Notes of the Tenth International Workshop on Qualitative Reasoning, QR'96*, .
- C.W. Marschall & A. Rudnick [1974], "Conventional strength testing," in *Fracture Mechanics of Ceramics. Volume 1: Concepts, Flaws, and Fractography*, R.C. Bradt, D.P.H. Hasselman & F.F. Lange, eds., Plenum Press, New York, 69–92.
- P.P. Nayak [1995], *Automated Modeling of Physical Systems*, Springer Verlag, Berlin.
- M. Neitzke [1997], *Relativsimulation von Störungen in technischen Systemen*, Dissertationen zur Künstlichen Intelligenz, Bd. 154, Infix, Sankt Augustin.
- D.S. Weld [1990], *Theories of Comparative Analysis*, MIT Press, Cambridge, MA.
- V.J. de Wit & H. de Jong [1998], "Implementatie van KIMA," University of Twente, Memorandum UT-KBS-98-06, Enschede, the Netherlands, in preparation.