Validating Numerical Models - Quantifying Error [1]

Alistair Duffy **Dawn Coleby**
School of Engineering and Technology, De Montfort University, Leicester, UK
apd@dmu.ac.uk
dawn@dmu.ac.uk

Anthony Martin
MIRA Ltd, Watling Street, Nuneaton, UK
anthony.martin@mira.co.uk

Malcolm Woolfson **Trevor Benson**
School of Electrical and Electronic Engineering, Nottingham University, Nottingham, UK
Malcolm.woolfson@nottingham.ac.uk
Trevor.benson@nottingham.ac.uk

Abstract

The validation of numerical models often progresses incrementally from previous models or other numerical solutions or is undertaken by comparison with experimentally obtained reference measurements. Notwithstanding the accuracy of the reference results, quantification of the error between the two is important information in deciding the quality of the model. It is frequently the case that this estimate of error is done by eye. However, for purposes of traceability and objectivity, interest has started to focus on techniques to quantify this error in an algorithmic manner in a way that agrees with the general observations of experienced engineers. This paper reviews two of the most promising techniques, namely Feature Selective Validation (FSV) and Integrated Error against Logarithmic Frequency (IELF), putting them in the context of correlation and reliability functions.

INTRODUCTION

Validation may involve comparing the results from one numerical model with another modeling technique, another implementation using the same technique or with measurements. The issues raised here include:

- Noise, including numerical noise and implementation errors, and experimental error compound differences in the results being compared. Experience may allow those undertaking the validation to accommodate these errors. In many cases, this will not be done overtly, which causes problems in trying to capture the experience being relied upon.
- In many practical situations, for example in EMC, the results are complex. One particularly common example is the coupling to cabling where, typically this will involve cables in cavities, coaxial systems, cross-talk systems and/or multiaperture systems. These resonant structures interact producing complex looking data. The effect of this is that discrepancies in one aspect of the model can mask the accuracy of the rest of the model.
- Comparisons between models and measurements becomes entirely subjective, which increases the communications difficulties within a team, particularly if this team is multidisciplinary (and therefore does not share similar backgrounds and experiences) or geographically separated (and hence will not develop a group tacit knowledge of such matters)

These factors are substantive arguments for the development of techniques to quantify the errors and recommended practice in using this information. This paper is concerned primarily with generating this quantitative data.

Two data sets used in this paper to represent validation data are given in Figure 1. They have been chosen because they have moderate complexity resulting from compound resonances. It is acknowledged that the actual data, to which the techniques discussed later may be applied, may have a much greater or lower feature density.

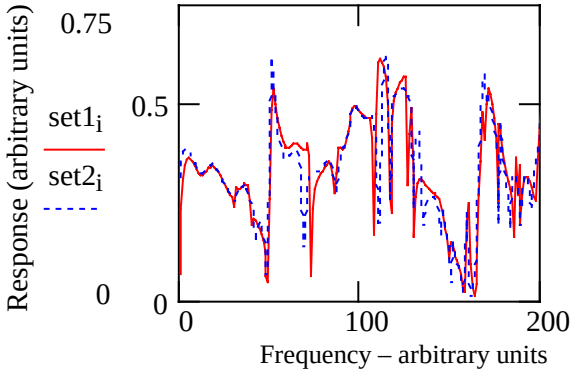

Figure 1 Two typical data sets used for comparison

It can be seen that the set of data shown in Figure 1 exhibits a very similar amplitude trend (mean level). However, there are shifted resonant-like features and some features which appear on one data set but not the other.

As noted previously, engineers typically assess the quality of such comparisons visually, with individual and group experience, both tacit and explicit, being essential in order to comment on the quality of such comparisons [1]. Those used to undertaking these comparisons often 'know' how to

[1] Based on "Progress in quantifying validation data" presented at the IEEE International Symposium on EMC, Boston, MA, August 2003.

interpret a result but can't say why: experience suggests the ambiguities and the assumptions. However, problems with human interpretation include fatigue and experiential differences between those assessing similar results, which produces different interpretations. It is also evident that these interpretations are further guided by the way in which expertise is learned, as much as the expertise at a particular instant. Thus, continuous learning may, itself, result in temporal differences, with different assessments of the same data being forthcoming at different times.

An important consideration when quantifying the differences between data sets, is that any quantitative measure should also provide a basis for investigating the quality of the comparisons. The measure should encourage the user to ask "why is that?", ideally decomposing the original comparison into something that helps in the post-mortem exercise of identifying if and how the comparison can be improved. Moreover, to aid communication, natural language descriptors should be used where possible.

The next section outlines correlation and Reliability factors, neither of which are particularly useful in this context, but worth considering as essential background: there may be circumstances in which these approaches could be beneficial. This is followed by an overview of Feature Selective Validation (FSV) and Integrated Error against Log Frequency (IELF), two techniques which have demonstrated some benefit in this application area.

CORRELATION and RELIABILITY FUNCTIONS

The *Pearson Correlation Coefficient* [2], is usually used to measure whether there is a linear relationship between two variables and the strength of the relationship. Plotting the two signals against each other on a scatter diagram is another easy method to determine the strength of the linear relationship between the compared signals, and the direction of the relationship. One study on correlation is discussed in reference [3]

To ensure that the Pearson Correlation Coefficient is accurate, several assumptions must be met. These assumptions involve the residuals of the data sets, which must be independent, normally distributed and have a constant variance. If the assumptions are not valid, log transformations of the data can be used, or an equivalent non-parametric test called *Spearman Rank Correlation* [2] can be applied. Spearman Rank Correlation measures the correlation of the ranks of the two variables. Both correlation coefficients will lie in the range –1 to +1. In most cases, the Pearson correlation coefficient is the default 'correlation'.

The Pearson Correlation Coefficient is calculated using:

$$R=\frac{\sum xy-(\sum x\sum y)/n}{\sqrt{(\sum x^2-(\sum x)^2/n)(\sum y^2-(\sum y)^2/n)}}$$

x=data set 1

y=data set 2

n= total number of points in both data sets

The Spearman Rank Correlation Coefficient is calculated using:

$$\mathbf{R} = 1-\frac{6\times\sum(D^2)}{n(n^2-1)}$$

D=difference in rank of pairs x and y

The main disadvantage is that despite the general familiarity, a simple correlation does not convey much information about the richness of the data and level of discrimination in high feature density situations. It can, therefore, produce results, which do not reflect the 'by-eye' opinions [4].

A technique developed to overcome some of the limitations of correlation was the introduction of *correlelograms* [5]. These involve cross-correlating the two data sets, incrementally shifting one set against the other to determine the peak correlation and the necessary shift (a similar approach could be used for stretching operations). A symmetry measure of the cross-correlation function is then determined by taking the rms of the differences between the cross-correlelogram at point k and point f_{max} -k where k is a variable (frequency) between f_{min} and f_{max}. The final value is the difference between the auto-correlelogram (where both data sets are the reference data) and the cross-correlelogram. Figure 2 shows the cross- and auto-correlation graphs for the data sets presented in Figure 1.

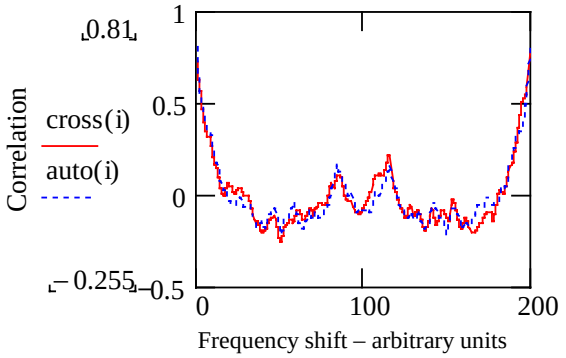

Figure 2 Correlelograms

The results are summarised in table 1

Table 1 Correlation values

Maximum correlation value	0.804
rms difference	0.051

rms symmetry	0.901

The fairly high correlation value indicates a reasonably good comparison, the difference value suggests that the results are close, reasonably uniformly across the range, and the high symmetry measure suggests that there is little skewing of the results.

One advantage of this approach is that it extends a technique which is widely accepted. The additional measures can provide a greater insight into the behavior of the system and the graphical output can assist in the overall assessment of the comparison. The disadvantage is that the additional measures have little, apparent, intuitive relevance, and relating them back to the physical systems may be difficult.

Reliability factors (R-factors) were originally used to compare transmission electron diffraction, TED, [6] and to compare calculated and measured LEED intensity spectra for surface structure determination [7]. Reliability factors were designed to analyse differences between two sets of results e.g. modelled and experimental data. Difference measures are used to compare signals and derivatives are used to compare features. Most R-factors compare the gradients, peaks and troughs of the signals being compared, albeit in different ways.

One of the R-factors considered for this work is that proposed by van Hove [8], which consists of a number of individual elements, generally comparing differences in the original data sets or differences in their derivatives, as detailed here:

$$R1 = \frac{\sum_{f\,\min}^{f\,\max} \left| I_{Set1_f} - C \cdot I_{SET2_f} \right|}{\sum_{f\,\min}^{f\,\max} \left| I_{Set1_f} \right|} \qquad R2 = \frac{\sum_{f\,\min}^{f\,\max} \left(\left| I_{Set1_f} - C \cdot I_{SET2_f} \right| \right)^2}{\sum_{f\,\min}^{f\,\max} (I_{Set1_f})^2}$$

$$R3 = \frac{No.+ve\ \ slopes\ \ SET1}{No.-ve\ \ slopes\ \ SET1} - \frac{No.+ve\ \ slopes\ \ SET2}{No.-ve\ \ slopes\ \ SET2} \qquad C = \frac{\sum I_{set1}}{\sum I_{set2}}$$

$$R4 = \frac{\sum_{f\,\min}^{f\,\max} \left| I'_{SET1}(f) - C \cdot I'_{SET2}(f) \right|}{\sum_{f\,\min}^{f\,\max} \left| I'_{SET1}(f) \right|} \qquad R5 = \frac{\sum_{f\,\min}^{f\,\max} (I'_{SET1}(f) - C \cdot I'_{SET2}(f))^2}{\sum_{f\,\min}^{f\,\max} (I'_{SET1}(f))^2}$$

Ix represent the amplitudes of the data points in the two data sets. R1 and R2 emphasize the agreement in the locations, heights and depths of the peaks and troughs but do not concern themselves with the detail on the peaks nor the nature of curvature. R3, R4 and R5 were proposed by van Hove to overcome this. In order to provide a single figure of merit, it has been proposed [4] that a total value can be obtained from:

$$RT = \sqrt{R1^2 + R2^2 + R3^2 + R4^2 + R5^2}$$

When applied to the data of Figure 1, the van Hove method gives the results of Table 2. The low values of R1 and R2 indicate a generally good agreement. The high value of RT, derived from R4 and R5 suggest large differences in the fine-grain detail.

Table 2 Van Hove assessment of the data sets in Figure 1

R1	R2	R3	R4	R5	RT
0.136	0.052	-0.275	1.079	1.319	1.723

Modifications have been made to the van Hove method to allow the individual R factors to be derived as a function of frequency[6].

FEATURE SELECTIVE VALIDATION (FSV)[4]

The basis of the FSV technique is the decomposition of the results to be compared into only two 'component' measures and then the recombination of the results to provide a global *goodness of fit* measure. The components used are the Amplitude Difference Measure (ADM), which compares the amplitudes and 'trends' of the two data sets and the Feature Difference Measure (FDM), which compares the rapidly changing features (as a function of the independent variable). The ADM and FDM are then combined to form a global difference measure (GDM). All of the ADM, FDM and GDM are usable as point-by-point analysis tools or as a single, overall, measurement.

The ADM and FDM are obtained using the following equations.

$$ADM = \sum_{f_{\min}}^{f_{\max}} \frac{|\, I_{low1} - I_{low2} \,|}{\alpha_{AD1}}$$

$$FDM = 2 \sum_{f_{\min}}^{f_{\max}} \{ FD_1(f) + FD_2(f) + FD_3(f) \}$$

where

$$FD_1(f) = \frac{|\, I'_{low1} - I'_{low2} \,|}{\alpha_{FD1}}$$

$$FD_2(f) = \frac{|\, I'_{high1} - I'_{high2} \,|}{\alpha_{FD2}}$$

Also

$$FD_3(f) = \frac{|\, I''_{high1} - I''_{high2} \,|}{\alpha_{FD3}}$$

$$\alpha_{FD_1} = \frac{2}{f_{\max} - f_{\min}} \sum_{f=f_{\min}}^{f_{\max}} \left[\left| I'_{LOW1}(f) \right| + \left| I'_{LOW2}(f) \right| \right]$$

$$\alpha_{FD_2} = \frac{4}{f_{\max} - f_{\min}} \sum_{f=f_{\min}}^{f_{\max}} \left[\left| I'_{HIGH1}(f) \right| + \left| I'_{HIGH2}(f) \right| \right]$$

$$\alpha_{FD_3} = \frac{6}{f_{\max} - f_{\min}} \sum_{f=f_{\min}}^{f_{\max}} \left[\left| I''_{HIGH1}(f) \right| + \left| I''_{HIGH2}(f) \right| \right]$$

where I_{low1} *and* I_{low2} are the amplitude of data sets 1 and 2 at data point f. The subscript *low* refers to the low frequency components of the data sets. This is obtained by Fourier transforming the data and inverse transforming the lowest 25% of the data, i.e. the range of frequencies $0 \le f \le f_s/8$, where f_s is the sampling frequency. α_{AD1} is an amplitude normalisation factor which is the average absolute energy contained in the signals under investigation.

I_{high} is the high pass component of the data sets, obtained by Fourier Transforming the data sets and inverse transforming the highest 75%, i.e. $f_s/8 \le f \le f_s/2$. The single primes (') indicate the first derivative with respect to the x-axis of a set and the double primes (") indicate the second derivative of the data. The α denominators are normalisation factors obtained for the data set components being compared.

The Global Difference Measure (GDM) is then obtained as:

$$GDM = \sum_{f_{\min}}^{f_{\max}} \sqrt{(ADM(f))^2 + (FDM(f))^2}$$

The value obtained from this equation gives a single figure of merit representing the comparison of the modeled and measured data across the data points. The GDM, like the ADM and FDM, is available on a point-by-point basis. The benefit of the point-by-point results is that these can help to identify regions where attention needs to be focused during validation of the model or in the post-mortem phase.

Natural language descriptors have been assigned to the output from this technique (ideal, excellent, very good, good, fair, poor, very poor) with some success [9].

The comparison of the traces in Figure 1 yields the GDM of Figure 3, ADM of Figure 4 and FDM of Figure 5. The global figures are given in Table 3 along with their natural language descriptors.

Table 3 Overall results of the FSV comparison of the data in Figure 1

Measure	Value	Descriptor
GDM	0.3	Good
ADM	0.1	Excellent
FDM	0.3	Good

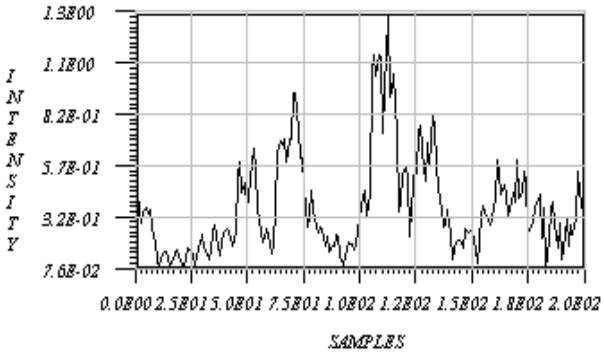

Figure 3 Global Difference Measure of the data of Figure 1

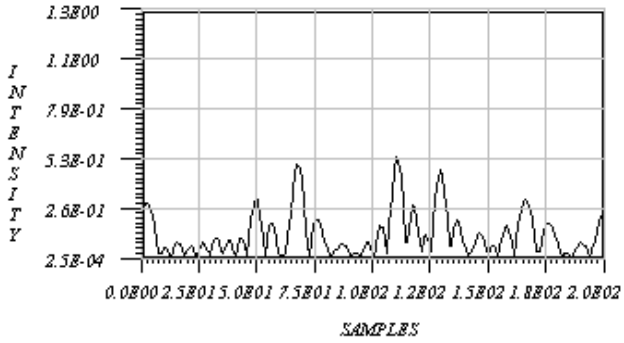

Figure 4 Amplitude Difference Measure of data in Figure 1

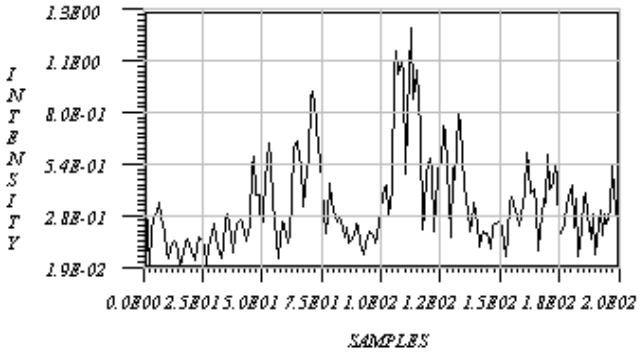

Figure 5 Frequency Difference Measure of data in Figure 1

Further, the probability density function of the individual point-by-point analyses can be plotted to provide a confidence measure. Essentially, this density function provides a visual guide as to how well a comparison conforms to the descriptor discussed above. The probability density function for the Global Difference Measure, shown in Figure 3, is given in Figure 6.

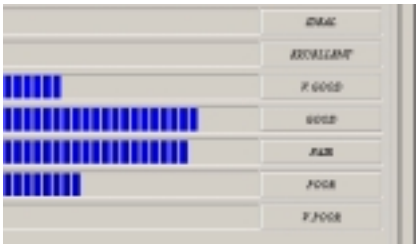

Figure 6 Probability density histogram 'confidence levels' for the GDM

The results confirm the conclusions of a visual inspection of the source data; that is,

1. the amplitude and general trends are in excellent agreement and
2. while there are regions, which are substantially different in terms of the location of features, the overall agreement is good.

This section has demonstrated the basis of the FSV method and described how it can compare data in a tiered manner, ranging from a point-by-point analysis to a single global figure providing an overview of the whole comparison. It has the advantages of breaking down the comparison into the two main aspects generally considered, namely amplitude/trends and features, and of having a natural language description. Previous tests have shown encouraging agreement with the perceptions of practicing engineers.

INTEGRATED ERROR AGAINST LOG FREQUENCY (IELF) [8]

This technique is based on the concept that the difference between the data being compared is the most significant aspect of the comparison. Further, any comparison based on a mean value of the difference would also require the standard deviation in order to give some additional context to this figure. However, the authors' of [8] concluded that a single figure of merit could be obtained by integrating the error over the frequency range of interest. A logarithmic frequency axis was chosen. Essentially, the calculation sums the (error ×data point separation) ÷ overall range. Hence, the proposed equation for this is:

$$IELF = \frac{\sum_0^n |error_n| . \{\ln(f_{n+1}) - \ln(f_{n-1})\}/2}{\ln(f_n) - \ln(f_0)}$$

- n is the number of frequencies for which there is data,
- f_0 is the first frequency, f_n is the last frequency,

and

- $error_n$ is the difference between the two sets of data for the n^{th} frequency.

The source data in Figure 1, with the difference on a logarithmic frequency range is shown in Figure 7, w.

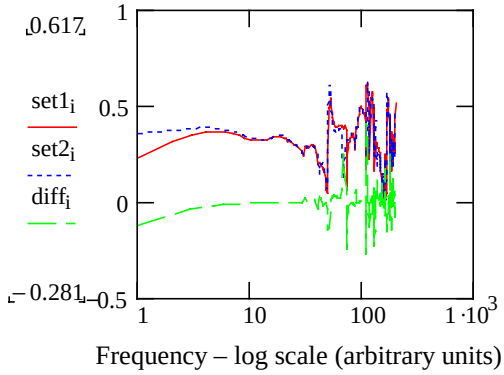

Figure 7 Original and difference data

The scaling of the IELF is such that 0 would be a perfect comparison and subsequently larger values would represent increasingly poor results. Applying the IELF to the data of Figure 1 (keeping the amplitude axis linear) gives a result of 0.9. However, this value is best used as a relative comparison with other results. Both values indicate a high level of agreement between the original data.

An advantage of this approach is that because it produces a single value which has some relevance to the way data is compared by EMC engineers, it is particularly good for ranking data resulting from many comparisons, and because the scaling axis can be readily normalized, it is easy to compare with data ranked by the engineers.

DISCUSSION

The ability to assign a numerical value between results from a numerical model and those a measurement or other models is an important factor in validating numerical models. This paper has reviewed some of the most promising candidate techniques in order to be able to do this. It should be noted that none of the techniques is limited to validation; the applications for them can extend to many other areas of engineering requiring the quantitative comparison of complex data. One of the fundamental challenges is that in order to be of practical use, any technique must provide some mirror of the way in which engineers look at the data and this implies the need to benchmark. This is a topic of current research. A method proposed to do this uses a question and answer approach to quantify comparisons [11].

REFERENCES

[1] AP Duffy, "EMC Comparison: A technical problem or a knowledge management issue?", *Zurich International EMC Symposium*, Feb 2003.

[2] See for example http://mathworld.woolfram.com

[3] Kevin J Mcdonald, *Software Tool To Compare Two Datasets*, M.Sc. Dissertation, York University, UK. April 1999.

[4] Anthony Martin, *Feature Selective Validation*, 1999 PhD Thesis, De Montfort University, Leicester.

[5] AP Duffy, M S Woolfson, T M Benson: "Use of correlation functions to assist the experimental validation of Numerical Modelling Techniques". *Microwave and Optical Technology Letters*, June 5, 1994, 7 (8), pp. 361 - 364.

[6] T. Suzuki, H. Minoda, Y. Tanishiro, K.Yagi, "TED analysis of the Si(113) surface structure."1999 *Surface Science*. Vol 438, part 1/3 pp76-82.

[7] K.Heinz, G.Besold, "Comparison of Zanazzi-Jona and Pendry reliability factors over an extended energy range.", 1983 *Journal of Physics C (Solid State Physics)*. Vol 16, pp1299-1306.

[8] Van Hove, M.A. et al (1977) "Surface structure refinement of 2H-MoS_2, 2H-$NbSe_2$ + w(100)p(2x1)-O via new reliability factors for surface crystallography", Surface Science, Vol 64, pp.85.

[9] AR Ruddle, AJM Martin, DD Ward, "Quantitative data comparisons: Applications and experiences in automotive EMC", *Zurich International EMC Symposium*, Feb 2003.

[10] RJ Simpson and CCR Jones, "IELF (Integrated Error Against Log Frequency) as a solution to quantifying comparisons" *Zurich International EMC Symposium*, Feb 2003.

[11] D Coleby and A Duffy, "The development and Validation of a New Technique to Compare Complex Data Sets", *Zurich International EMC Symposium*, Feb 2003.