The Influence of Data Density on the Consistency of Performance of the Feature Selective Validation (FSV) Technique

Alistair Duffy*, **Antonio Orlandi**+
* De Montfort University, Leicester, UK
+ *UAq EMC Laboratory*, University of L'Aquila, L'Aquila, Italy
contacts: apd@dmu.ac.uk, orlandi@ing.univaq.it

ABSTRACT — The human visual system has an immense capacity for compensating for poor or incomplete data. Psycho-visual coding schemes make use of the brain's ability to extrapolate and interpolate independently of conscious awareness to reduce data (bit) rates but maintain the same level of 'information' within a video signal. However, when attempting to produce a simple method for comparing data-sets, primarily for validation of computational electromagnetics, could give rise to a problem. Namely that someone undertaking the visual inspection of (e.g.) modeled data against experimental data will see the same picture whether sampled at N, 100N or 0.01N data points whereas the software undertaking the comparison would process three very different data sets. The Feature Selective Validation (FSV) method was developed to attempt to mimic the group response of a number of experts undertaking the visual comparison. Hence, the quality of performance of the FSV method should not be severely affected by the number of data points if this assertion is to hold, despite the obvious potential for variation. This paper investigates the FSV performance as a function of data density and shows that the accuracy of its performance remains largely unimpeded by variations in the precision of the data supplied.

I. INTRODUCTION

In order for this paper to investigate the effect of data density on the FSV method, it needs to consider two issues. The first is to review the FSV method, clarifying what data is being used (including what this paper considers as data density), how it is being used, why and how the data density impacts on the underlying equations. The second factor is how does a normal graphical representation of data differ from how the data is presented to the FSV method. This section will overview the FSV method and the issues surrounding graphical representation and then lead on to a more detailed review of the FSV heuristics and then to tests to ascertain the quality of performance of the FSV method.

I.1. FSV OVERVIEW

A typical scenario for the validation of computational electromagnetics involves the modeling of a system that can be directly measured. The resulting pair of data sets are then presented graphically to those involved in this exercise who will ascribe a quality level to the comparison: such as 'good'. Closer inspection may add a qualifier to this descriptor: such as 'good, but…'. The Feature Selective Validation (FSV) method was conceived as a technique to support this exercise, providing a numerical value to the quality of a comparison constructed from components analogous to the general approach used by humans. Namely, comparing the trends of the two data sets and comparing the individual features, perhaps resonant features, and combining these to give an overall confidence in the goodness of fit. The reason for the development of FSV over using existing methods, such as correlation, was that the existing methods do not offer sufficient discrimination or potential for feedback to the users.

The basis for the FSV method is to low pass and high pass filter the original two data sets, take differences of the low pass data to give the Amplitude Difference Measure (ADM) and take a mix of differences of derivatives of the low and high pass data to give the Feature Difference Measure (FDM). These represent the trend / envelope difference and the resonant-type feature difference discussed earlier. These are then treated as independent components and a Global Difference Measure (GDM) obtained from the ADM and FDM. More detail is given in section 2 and in references [1 – 3].

I.2. REPRESENTATION ISSUES

When a set of measurements is taken and represented graphically, it is usual to present the data with lines joining the points, as in Figure 1.

Figure 2 shows the same data but represented more in the manner that a computer would 'see' it.

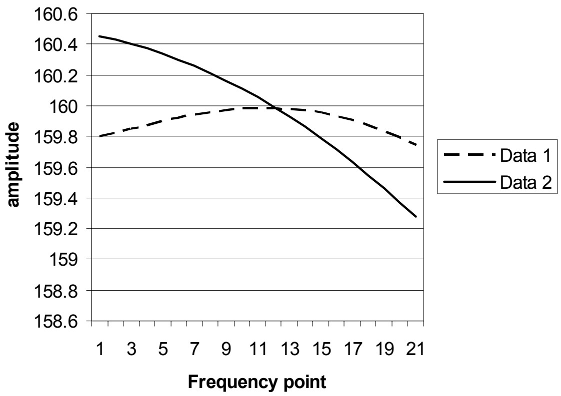

Figure 1. Normal presentation of data using lines.

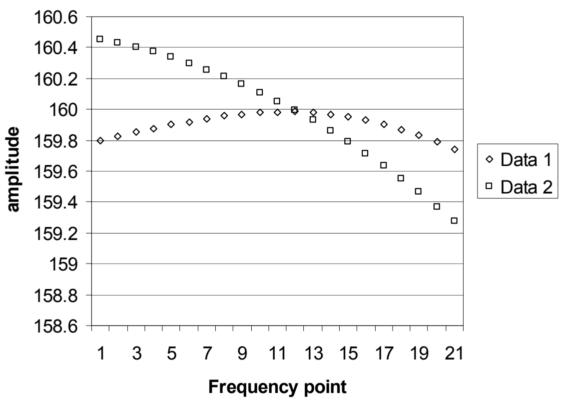

Figure 2. "Non-interpolated" presentation of data.

While we look at continuous lines on a graph, the computer 'sees' points (more correctly a table of numbers). The presence of lines gives a sense of certainty to the locations in both ordinate and abscissa, however, representation as points shows the possibility of uncertainty in the location. While this is a trivial example, it can be seen that having fewer points on a graph increases the uncertainty, but presents less noise in the form of high frequency components. However, more data points increases the precision of the data but increases the noise present. Effectively, manipulating the number of data points may have virtually no effect on the visual representation of the data but will have a substantial effect on the data presented for analysis; for example halving the number of data points reduces the amount of data available for analysis by 3 dB but results in virtually no difference in the visual effect.

I.3. PURPOSE OF STUDY

Bringing together the two themes discussed so far we can see that the use of a tool like the FSV method for the validation of CEM should, as far as possible, predict the response of a large group of users. By doing this, a level of confidence can be attributed to the quality of the comparison. The response of the large group of users will be done 'by-eye' based on lines on a graph, suggesting continuous functions, further implicitly suggesting a level of precision that may not be realistically expected from the data. However, the response of the computer program will be obtained by following a clear and predetermined algorithm based on discrete points, i.e. non-continuous functions with the implication of uncertainty between these points.

Hence, the purpose of, and the research question for, this study is to see if varying the number of points presented to the FSV method will leave the output relatively unchanged, and certainly in line with user opinion. From this, it may be possible to issue some guidelines recommending good practice in the use of the FSV method so as to ensure a high level of consistency between applications.

There is the further issue that should be considered regarding the fact that as the number of data points used to represent the systems being compared is increased or decreased, the information content is either increased or decreased, the precision in the data (i.e. the tolerance of each point along the x axis) varies and the noise content (or aliasing effect due to sampling effects) varies. Hence, while comparing data with, for example, 400 points in one instance and 100 points in another instance, may look identical, they are clearly separate sets of results. Issues surrounding data density and the veracity of the conclusions to be drawn from this are common in other walks of engineering (e.g. [4, 5]).

The next section will review the mathematics behind the FSV method and this will be followed by the tests to address the research question set out above.

II. THE FSV METHOD

The FSV method was outlined above. This section reviews the governing equations and the methods used to represent the FSV output to users.

II.1. FSV EQUATIONS

The governing equations are as follows. Note x is the independent variable, Lo_i and Hi_i are the low pass and high pass filtered versions of i^{th} data set i = 1,2 (the subscript indicating the data set) and the single and double primes show the first and second derivatives with respect to x obtained using a central difference scheme,

$$ADM(x) = \left| \frac{\left(\left|Lo_1(x)\right| - \left|Lo_2(x)\right|\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(\left|Lo_1(i)\right| + \left|Lo_2(i)\right|\right)} \right| \tag{1}$$

$$FDM(x) = 2\left(\left|FDM_1(x) + FDM_2(x) + FDM_3(x)\right|\right) \tag{2}$$

$$FDM_1(x) = \frac{\left|Lo_1'(x)\right| - \left|Lo_2'(x)\right|}{\frac{2}{N}\sum_{i=1}^{N}\left(\left|Lo_1'(i)\right| + \left|Lo_2'(i)\right|\right)} \tag{3}$$

$$FDM_2(x) = \frac{\left|Hi_1'(x)\right| - \left|Hi_2'(x)\right|}{\frac{6}{N}\sum_{i=1}^{N}\left(\left|Hi_1'(i)\right| + \left|Hi_2'(i)\right|\right)} \tag{4}$$

$$FDM_3(x) = \frac{\left|Hi_1''(x)\right| - \left|Hi_2''(x)\right|}{\frac{7.2}{N}\sum_{i=1}^{N}\left(\left|Hi_1''(i)\right| + \left|Hi_2''(i)\right|\right)} \tag{5}$$

$$GDM(x) = \sqrt{ADM(x)^2 + FDM(x)^2}\ . \tag{6}$$

In the summary, single value representations of the ADM, FDM, and GDM are obtained by taking the mean value over the range of x of interest.

Given that a central difference scheme has been used for the derivatives, it is assumed that the data points are evenly spaced. So, from Table I it can be seen that

$$y'(x4) = (y(x5) - y(x3)) / (x5 - x3). \tag{7}$$

Table I. Representation of trial data.

Point number	Data value
1	$y(x_1)$
2	$y(x_2)$
3	$y(x_3)$
4	$y(x_4)$
5	$y(x_5)$
6	$y(x_6)$

However, in an undersampled version of this, as indicated in Table II

$$y'(x4) = (y(x6) - y(x5)) / (x6 - x5). \tag{8}$$

Table II. Under sampled data of Table I.

Point number	Data value
3	$Y(x_2)$
4	$y(x_4)$
5	$Y(x_6)$

There are two practical implications for using this data in FSV. The first is that where there is a low rate of change in the data, then $y(x_5) - y(x_3) \approx y(x_6) - y(x_5)$ and a high rate of change may render this approximation incorrect. The second point is that in Table II the separation of data points is $2\Delta x$, where Δx is the separation in Table I. However, in the FSV equations employing derivatives, the derivatives appear in both the numerator and denominator of the equations, so the Δx effect will cancel. Of course, this is only true if the Δx used for derivatives of dataset 1 is equal to the Δx used for derivatives of dataset 2. This assumption has been made in this analysis.

This leaves the issue of whether Δy_a is sufficiently close to Δy_b where a and b represent two different sampling rates. This will be implicitly investigated in the next section.

In order to help interpret the results in the next section, methods used to represent the FSV data to users will be reviewed.

II.2. FSV REPRESENTATION

The basic representation of the FSV output can be either simply numeric (i.e. single figure values for the ADM, FDM and GDM) or point-by-point values ({A, F, G}DM(x) as in the previous equations). However, one of the design requirements for FSV was to provide a range of diagnostic information [2]. Bearing in mind the aim for the FSV method to mirror the opinions of a group of engineers, it has been found that the proportions of each of the measures that falls into the definitions of the natural language descriptors, given in Table III, provides a useful histogram which is suggestive of the proportions of a large group who, when assessing the original comparisons would categorize them according to the categories [3]. These are called the confidence histograms for each of the measures.

Table III. Natural language descriptors for FSV {A, F, G}DM.

FSV value (quantitative)	Natural language descriptor
Value ≤ 0.1	Excellent
0.1 ≤ Value < 0.2	Very good
0.2 ≤ Value < 0.4	Good
0.4 ≤ Value < 0.8	Fair
0.8 ≤ Value < 1.6	Poor
1.6 ≤ Value	Very poor

Given that the confidence histograms aim to provide a synthetic group response, a wide spread of similar height categories would lead to the interpretation that there can be a low confidence in attributing a single epithet to a comparison. For example, if the confidence histogram showed an approximately even distribution between Good, Fair, and Poor, it would be inappropriate to describe the comparison as Fair; it would be better to describe it as Good – Poor. In order to capture this in a more algorithmic manner, the ‘spread’ of the confidence histograms has been introduced [6]. The Spread is the number of categories that contains 85% of the data points (taking the difference between the most and least favourable category). So, for example, if in the previous example, 30% of the points fell into each of the Good, Fair, and Poor categories, then the Spread would be 3. If, on the other hand, approximately half fell into the excellent category and nearly half into the Very Poor category, then the spread would be 6. Effectively, the Spread is a measure of the variance of the histogram data.

In order to balance the Spread, where a Spread of 2 could equally result from a combination of Excellent-Very Good as it could from Poor – Very Poor, a Grade measure has also been introduced alongside the Spread. Whereas the Spread could be thought of as a variance measure, the Grade is similar to an upper action line in process control. The Grade is the number of categories, starting with Excellent that need to be included for 85% of the data points in the particular measure to be counted. Thus, in the Good-Fair-Poor illustration the Grade would be 5; in the Excellent-Very Good illustration, the Grade would be 2 but the Poor-Very Poor illustration would have a Grade of 6. So together, the Grade-Spread gives a simple indication of the quality and reliability of the comparison.

The following analysis uses Grade-Spread in addition to the summary values to quantify the differences that varying the data density has on the comparison results. The results were obtained using a stand-alone FSV application [7, 8].

III. TESTS

In order to assess the performance of FSV when faced with varying data density, tests with three different data types have been performed. These are (1) EMC modelling of a via performance (2) very high feature density performance and (3) sinusoid representations. In doing this (1) is representative of real data that will commonly be presented to FSV, (2) is representative of data at one extreme of complexity and (3) is representative of data at the other extreme of complexity.

III.1. VIA PERFORMANCE

The system was modeled with 5768 points per data set, this is shown in Figure 3, and the data was then down sampled to a minimum of 177 points: Figure 4.

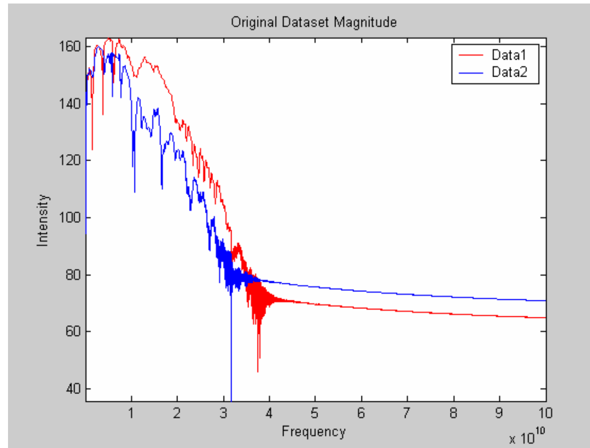

Figure 3. Via models: 5768 samples.

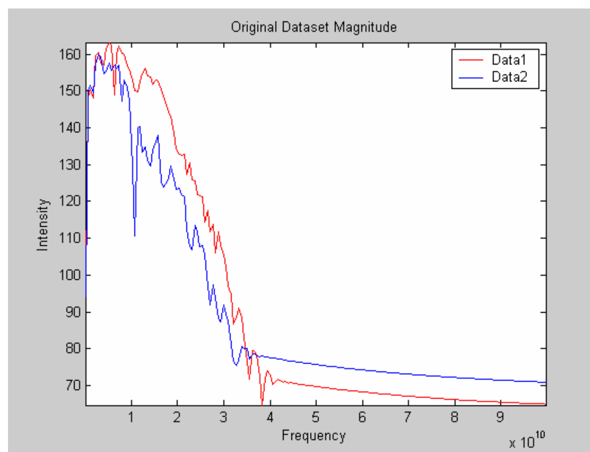

Figure 4. Downsampled data of Figure 3: 177 samples.

Table IV. FSV summary values for various data densities between the representations of Figures 3 and 4.

	ADM	FDM	GDM
No of Data points			
5768	0.165	0.604	0.665
2839	0.152	0.536	0.587
1419	0.139	0.395	0.448
709	0.132	0.321	0.374
384	0.132	0.277	0.332
177	0.140	0.274	0.334

Initial visual observations of this data suggest that the downsampling, as would be expected, has reduced the level of 'high-Q' features resulting in an overall visual improvement in the Feature (i.e. the 'high-Q' aspect) component of the original data, leaving the Amplitude (i.e. the envelope / trend) information relatively unaffected. The effect of the changes can be seen in Table IV which lists the ADM, FDM and GDM components for various data densities.

From a natural language descriptor equivalent, the ADM is unchanged at Very Good, whereas the GDM has gone from Fair to Good, reflecting a likely visual analysis.

However, in order to address possible researcher bias, the Grade-Spread information, as discussed in Section 2 was noted. For all the data densities, the ADM Grade remained at 3 and the Spread at 3. On the other hand, the Grade for the FDM and GDM was 5 for 5768 points – 709 points inclusive and was 4 for the remaining two comparisons. The Spread was mostly 4 for the FDM and GDM (it did nudge into 5 for the FDM and GDM of 2839 samples and for the FDM of 1419 samples). However, together this data suggests that the FSV routine is only particularly sensitive to the number of data points used as far as the eye is sensitive to the filtering and smoothing effects of the reduction in the number data points used to present the graphs.

III.2. HIGH FEATURE DENSITY

In this test, very highly structured and noisy data was downsampled. Figure 5 gives the original data and Figures 6 and 7 give undersampled versions of this data.

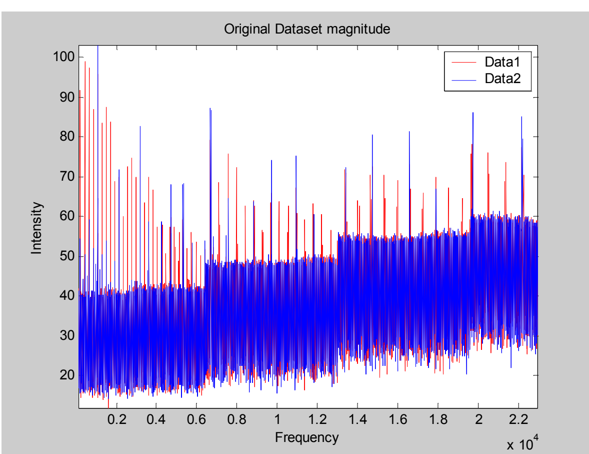

Figure 5. Original data: 3000 samples.

From these three figures, it is clear that the general structure of the data becomes visibly clear once the original data has been downsampled to 50%. Once the structure has become visibly clearer, it is not immediately obvious what effect the reduction in data density will have on the quality of comparison. Table V gives the Difference Measures for various data densities.

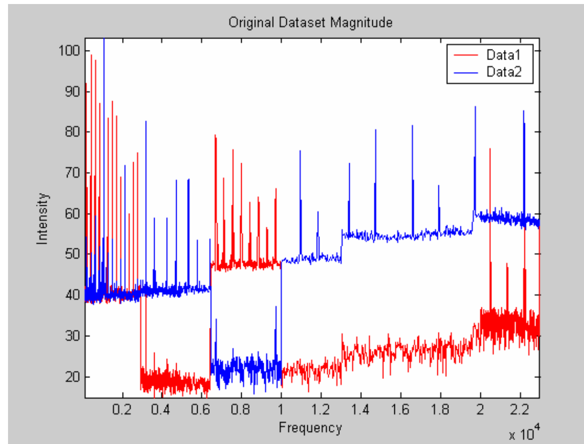

Figure 6. Downsampled data of Figure 5: 1500 samples.

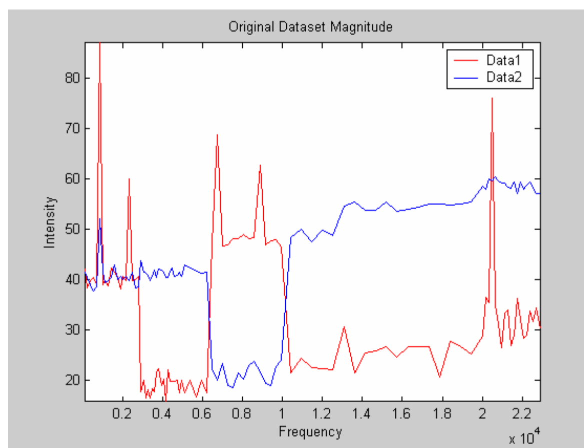

Figure 7. Downsampled data of Figure 5: 97 samples.

Table V. FSV summary values for various data densities between the representations of Figures 5 and 7.

	ADM	FDM	GDM
No of Data points			
3000	0.354	0.609	0.772
1500	0.547	0.681	0.977
750	0.550	0.709	0.995
375	0.521	0.658	0.928
187	0.558	0.649	0.937
93	0.584	0.727	1.027

It is interesting to note that as the graphs become visibly more structured, i.e. moving from Figure 5 to Figure 6, the FSV analysis becomes reasonably constant (below one decimal place). This is interesting because the similarities between Figures 6 and 7 are well defined and there is a high probability that they would be noted as such by most users. However, the similarity between Figures 5 and 6 are not so easy to see.

This observation is also reflected in the Grade-Spread. Excluding the 3000 point comparison, the grade is constant at 5 for the ADM, GDM and FDM for all comparisons except the GDM for 750 and 93 where it tips over to 6. Never-the-less a consistent performance.

For the Spread, the FDM is constant at 5 and the ADM and GDM both go from 4 (1500 sample) to 3 (93 sample). These results reflect the effect of reducing the data density on the visual interpretation, increasing the probability for consistency between users.

So far, the tests have only shown that the FSV method reflects the likely interpretation of the results by expert users because this data is highly structured and, therefore, relatively easy to identify visual changes to the structure of the data. The more demanding task, in these circumstances, is to take a simple structure, which will barely change visually and analyze this with FSV.

III.3. SINUSOIDS

In this final test, two sinusoids were compared after differences were introduced between them. These differences were changing the amplitude so one was 10% of the amplitude of the other and secondly, one was 90% of the amplitude of the other; the final test was to introduce a phase shift between the two curves.

III.3.1. 10% Relative Amplitude

The two curves, one at peak amplitude of 10 V and the other with peak amplitude of 1 V were compared over one period with a range of data points from 50 to 800. The comparison of these curves is illustrated in Figure 8. The summary values for the ADM, FDM and GDM are listed in Table VI.

Clearly, there is a consistency in all the figures. The slight variations in the GDM arise through the point-by-point nature of the summation and are illustrated in the confidence histogram in Figure 9.

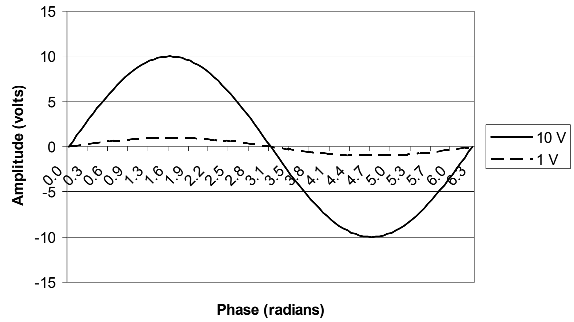

Figure 8. Two sinusoids of very different amplitude.

Table VI. FSV summary values for various data densities for a 10 V and 1 V sinusoid.

	ADM	FDM	GDM
No of Data points			
800	1.091	1.318	1.803
400	1.091	1.318	1.802
200	1.091	1.318	1.811
100	1.091	1.318	1.823
50	1.091	1.318	1.831

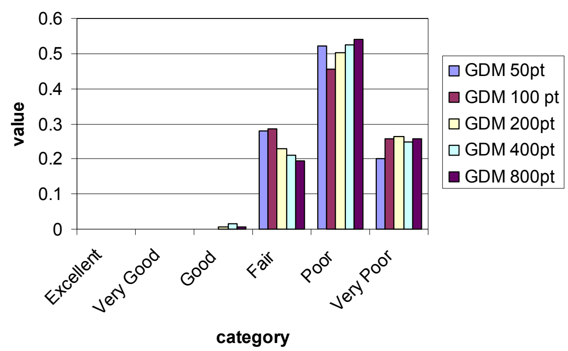

Figure 9. GDM confidence histogram for two sine waves of 10 V and 1 V amplitude.

III.3.2. 90% Relative Amplitude

The same test as in section 3.3.1 was performed, but this time 10 V and 9 V amplitude sine waves were compared, as illustrated in Figure 10. Table VII lists the summary values.

Again, there is practically very little difference in the results, as illustrated with the GDM confidence histogram in Figure 11

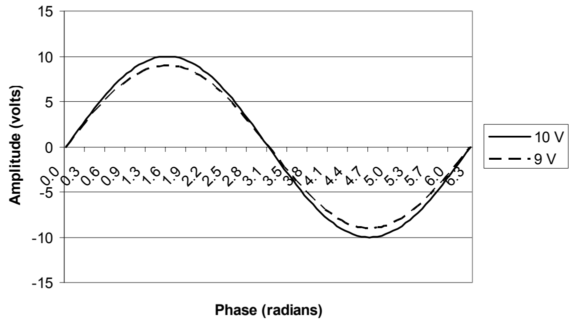

Figure 10. Two sinusoids of similar amplitude.

Table VII. FSV summary values for various data densities for a 10 V and 9 V sinusoid.

	ADM	FDM	GDM
No of Data points			
800	0.071	0.127	0.156
400	0.07	0.102	0.132
200	0.07	0.086	0.118
100	0.07	0.085	0.117
50	0.07	0.085	0.118

The greatest difference that can be seen from this figure is that there has been some shifting between Excellent and Very Good for the 800 point data. This is probably because of the proportion of the GDM at approximately 0.1, i.e. the boundary between the two categories.

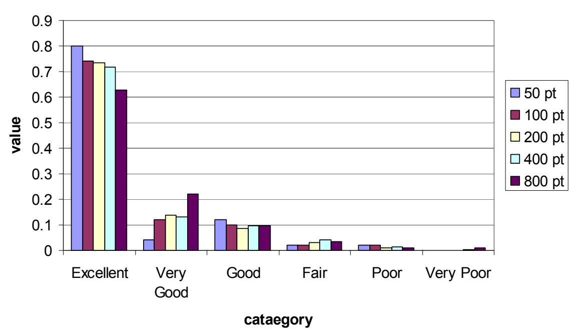

Figure 11. GDM Confidence histogram for two sine waves of 10 V and 9 V amplitude.

III.3.3. Phase Difference

The last of the sinusoidal tests involved taking two similar sinusoids and adding in a one radian phase shift. To add additional difficulty in the tests, the comparison was made over two cycles of the original data with data densities ranging from 10000 points to 20 points. In the case of 20 points, i.e. 10 data points per cycle, the curves clearly depart from smooth sinusoids. The 20 data point curve is shown in Figure 12 and the 100 data point curve, for comparison, is shown in Figure 13.

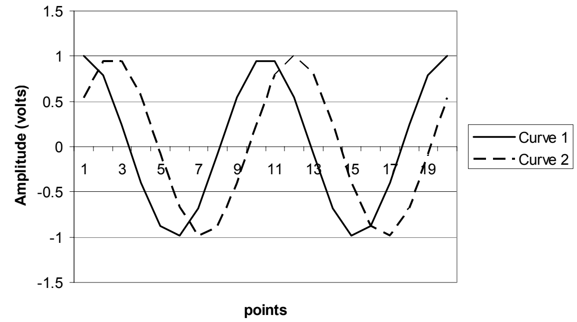

Figure 12. Two poorly defined sinusoids, each with 10 samples per cycle.

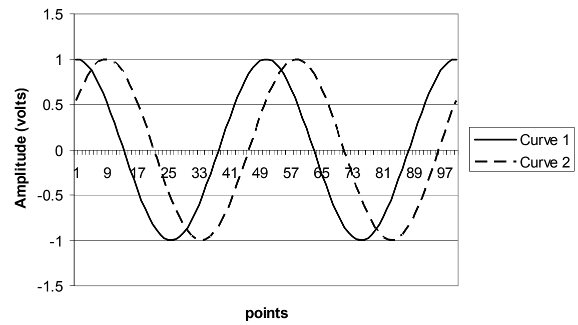

Figure 13. Two well defined sinusoids.

The comparison of the Difference Measure results is given in Table VIII.

Table VIII. FSV summary values for various data densities for two offset sinusoids.

	ADM	FDM	GDM
No of Data points			
10000	0.673	0.894	1.074
1000	0.674	0.689	1.063
100	0.674	0.722	1.094
20	0.712	0.844	1.206

Figure 14 shows the effect of changing the data densities on the GDM.

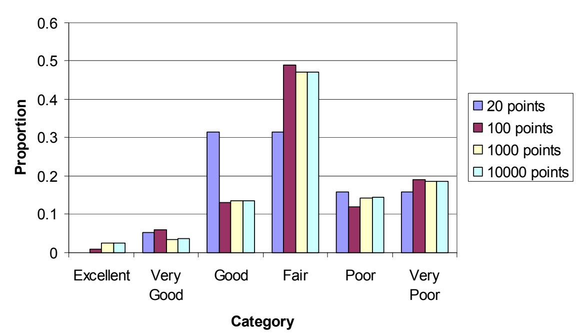

Figure 14. GDM for two offset sinusoids as a function of data density.

Figure 14 shows how consistently the FSV method assesses the comparison of these two data sets.

IV. CONCULSIONS

This paper has considered the effects of varying data density on the results of FSV analysis of varying data sets.

The FSV method is a promising approach to the formal and quantifiable comparison of data for tasks such as computational electromagnetic (CEM) validation, experimental repeatability and assessment of the amount of change when models are manipulated. However, as part of the validation of FSV itself, its robustness to changes in input data needs to be assessed. This paper has done this with three different, typical data sets, ranging from the structurally trivial (sinusoids) to the structurally highly complex.

The main conclusion that that can be drawn from the results is that FSV is robust in that it does not change the output values appreciably when the input data density changes, providing the data remains visually unchanged. This was clearly seen with the sinusoidal data where the overall levels of agreement did not change substantially once the data density was past the point where it appeared smooth and continuous.

The important general recommendation from this work is that there should be consistency in the number of points used to represent the data and that number should be chosen because of the way in which the data is represented. That is, it is important to choose the correct number of points to represent the data to be compared so that the visual representation is precisely that which needs to be compared. It was clear with the first two examples in this paper that the effective data to be compared could be changed considerably depending on whether more or fewer points were displayed. Clearly, this factor has more of an impact on the assessment of the quality of a comparison than anything else. Perhaps a standard question during validation should be "why has the data been displayed with this level of data density?"

ACKNOWLEDGMENT

The authors are indebted to Professor Charles Bunting (Oklahoma State Univ.) for suggesting this line of investigation and to V. Rajamani (Oklahoma State Univ.), Dr. B. Archambeault (IBM) for the numerous discussions during the research for this paper and Mr. F.Campitelli (*UAq EMC Lab.*) for performing part of the calculations.

REFERENCES

[1] AJM Martin, *Quantitative data validation (automated visual evaluations)*, PhD Thesis, De Montfort University, UK, 1999.

[2] AP. Duffy, AJM. Martin, A. Orlandi, G. Antonini, TM. Benson, and MS. Woolfson, "Feature Selective Validation (FSV) for validation of computational electromagnetics (CEM). Part I – The FSV method," *Submitted to IEEE Transactions on EMC*.

[3] A. Orlandi, AP. Duffy, B. Archambeault, G. Antonini, D. Coleby, and S. Connor, "Feature Selective Validation (FSV) for validation of computational electromagnetics (CEM). Part II – Assessment of FSV performance," *Submitted to IEEE Transactions on EMC*.

[4] G.C. Bishop "Gravitational field maps and navigational errors," *Proceedings of the 2000 International Symposium on Underwater Technology*, 2000, p 149-54.

[5] B. Morse, W. Liu, and L Otis, "Empirical analysis of computational and accuracy tradeoffs using compactly supported radial basis functions for surface reconstruction," *Proceedings - Shape Modeling International SMI* 2004, p. 358-361.

[6] A. Orlandi, G. Antonini, C. Ritota, and A. Duffy, "Enhancing Feature Selective Validation (FSV) interpretation of EMC/SI results with Grade-Spread".

[7] A free, stand-alone, FSV application is available from http://ing.univaq.it/uaqemc/public_html/.

[8] FSV web page at http://www.eng.cse.dmu.ac.uk/FSVweb.

Alistair P Duffy (M'93–SM'04) was born in Ripon, UK, in 1966. He obtained a First Class BEng(Hons) degree from University College, Cardiff, in 1988 in Electrical and Electronic Engineering, and the MEng degree the following year. He joined Nottingham University in 1990 receiving a PhD in 1993 for his work on experimental validation of numerical modeling. He also holds an MBA.

He is currently Reader in Electromagnetics and Head of the Division of Engineering at De Montfort University, Leicester, UK and has particular research interests in CEM Validation, communications cabling and technology management. He has published over 100 papers in journals and international symposia.

Dr Duffy is a Fellow of the Institution of Engineering and Technology (IET) and a member of the Chartered Management Institute (CMI). He is active in the IEEE standards activity on the validation of CEM. He is a member of the International Compumag society and the Applied Computational Electromagnetics Society.

Antonio Orlandi (M'90-SM'97) was born in Milan, Italy in 1963. He received the Laurea degree in Electrical Engineering from the University of Rome "La Sapienza", Italy, in 1988.

He was with the Department of Electrical Engineering, University of Rome "La Sapienza" from 1988 to 1990. Since 1990 he has been with the Department of Electrical Engineering of the University of L'Aquila where he is currently Full Professor at the *UAq EMC Laboratory*. Author of more than 100 technical papers he has published in the field of electromagnetic compatibility in lightning protection systems and power drive systems. Current research interests are in the field of numerical methods and modeling techniques to approach signal/power integrity, EMC/EMI issues in high speed digital systems. Dr. Orlandi received the IEEE TRANSACTIONS ON ELECTROMAGNETIC COMPATIBILITY Best Paper Award in 1997, the IEEE EMC SOCIETY TECHNICAL ACHIEVEMENT AWARD in 2003, the IBM SHARED UNIVERSITY RESEARCH AWARD (2004, 2005) and the CST UNIVERSITY AWARD in 2004. He is member of the Education, TC-9 Computational Electromagnetics and TC-10 Signal Integrity Committees of the IEEE EMC Society, Chairman of the "EMC INNOVATION" Technical Committee of the International Zurich Symposium and Technical Exhibition on EMC. From 1996 to 2000 has been Associate Editor of the IEEE TRANSACTIONS ON ELECTROMAGNETIC COMPATIBILITY and now serves as Associate Editor of the IEEE TRANSACTIONS ON MOBILE COMPUTING.