A Review of Statistical Methods for Comparing Two Data Sets

[1]A. Duffy and [2]A. Orlandi

[1] School of Engineering and Technology
De Montfort University, The Gateway, Leicester LE1 9BH
apd@dmu.ac.uk

[2] UAq EMC Laboratory
University of L'Aquila, L'Aquila, Italy I-67040
Orlandi@ing.univaq.it

Abstract – Statistical approaches to compare data for validation of computational electromagnetics have been used for several years. They provide an accepted means of obtaining a numerical value to quantify the data under consideration. However, the use and meaning of these 'numbers' depends, by necessity, on the application. This paper provides an overview of some of the most widely applicable techniques, relating the output of these to visual assessment. It further includes comparison with the FSV (Feature Selective Validation) method allowing a triangulation between statistical approaches, visual approaches and heuristic approaches to validation. It is important that the decision to use or reject a particular technique for validation is based on a rational and objective selection approach. This paper suggests a framework to support this selection approach.

Keywords: Validation, statistical analysis, and feature selective validation.

I. INTRODUCTION

The complexity of electromagnetic systems being analyzed and modeled can produce results which are, themselves, excessively complicated. This is particularly true when models tend 'in the limit' to replicate reality. Statistical electromagnetics is a topic that has become part of the general approach to study the results from these simulation activities. A standard starting point for statistical electromagnetics is [1]. A more recent contribution to the need for a better understanding and application of statistics in electromagnetics is [2] where the *a priori* assumption that there are unknown contributors to the model is acknowledged and these can be treated statistically. In both cases, these publications demonstrate the benefits to be gained by considering a statistical analysis under appropriate circumstances. However, correct application relies on appropriate selection and while there are obvious circumstances where one, or other, technique can be applied; there are many other circumstances where non-statistical approaches are more appropriate. This paper presents a short overview of statistical and non-statistical approaches for validation with the aim of helping those involved in validation make more appropriate selection of techniques to quantify the comparison of numerical data with experimental data or with other numerical models.

Validation of numerical models involves determining whether the agreement of a simulation with experiments, other simulations or analytically resolvable systems is adequate. Identifying what 'adequate' means may, in practice, involve the following:

- Expectations of agreement based on previous experience.
- Accounting for known assumptions embedded in both or either the model or (e.g.) the experiments.
- The end application to which the model is being used.

Clearly, accounting for these does suggest that there will only rarely be an absolute pass/fail decision to be made and more frequently whether there is a high / low probability that the model is good enough. The concept of defining adequacy as part of a model validation framework is likely to become a more relevant and pertinent issue in the near future as concepts such as satisfactions [3-4] and error budgets in models become part of the language of modelers.

The use of statistics in the validation of computational electromagnetics is not clearly defined. Hence, reviewing some of the statistical validation options does appear to be a relevant contribution to the debate on how best to perform quantitative validation. An example of the current state of the debate can be seen in the topic of modeling reverberation chambers, particularly in comparing models against experiments, where, one hand, [5] suggests that the nature of the reverberation chamber is such that the probability distribution of the fields is an appropriate way of comparing the models with measurements. On the other hand, [6] suggests that it is relatively straightforward to get statistical agreement even if there is total disagreement between the actual modeled and measured results.

1054-4887 © 2008 ACES

This paper provides a general overview of some of the more widely used statistical techniques and compares them against a sub-set of visual assessments. Visual assessment, the “eye-balling” of graphs, is probably the most common, widespread and accepted approach to validation. It is important that any approach to quantify results for validation purposes is empathetic to this and not in opposition. Hence, the use of a set of comparisons for which a bank of visual assessments is available is seen as a reasonable starting point to analyze the potential contribution for a bank of possible statistical techniques. An increasingly popular heuristic approach, FSV (Feature Selective Validation) is reviewed and a simple approach to quantifying visual comparisons is also reviewed.

II. TEST DATA

In order to be able to discuss the various techniques later in this paper, some test data is required to illustrate the quality of the comparisons. The first column of Fig. 1 shows three comparison graphs. Approximately 50 engineers were asked their assessment of these comparisons using a six point rating scale which are presented in histogram form in column 2 of Fig. 1 (these results are a subset of those presented in [7], where further methodological details are also presented). The use of the terms for the histograms is based on common natural language descriptors. The visual rating scale used is presented in Fig. 2.

A mean value was determined for these three comparisons by averaging the numerical scores from the survey. According to the visual assessment ‘Graph 4’ (average score 5.95) is the worst, ‘Graph 5’ is the best (average score 4.56) and ‘Graph 8’ is approximately mid-way between the other two (5.36). This provides a benchmark to test candidate statistical approaches. In particular, agreement in rank-ordering the results would be expected from any technique used because, often, the absolute score is not as important as knowing whether one comparison is much better or worse than another comparison.

III. STATISTICAL TECHNIQUES[1] [9 - 12]

III.1. Correlation and visually based approaches

The most common approach to correlate two sets of data is the Pearson r correlation, which provides a numerical measure of how closely related two variables are.

The Pearson Correlation Coefficient is calculated using equation (1),

$$r = \frac{\sum xy - (\sum x \sum y)/n}{\sqrt{(\sum x^2 - (\sum x)^2/n)(\sum y^2 - (\sum y)^2/n)}} \qquad (1)$$

X = data set 1,
Y = data set 2,
n = Total number of points in both data sets.

An alternative correlation technique is the Spearman Rank Correlation. Which measures the association of the ranks of the two variables. The point at which the largest value of the data-set occurs would be given a rank of 1; the next largest point would be given a rank of 2 and so on. The Spearman Rank Correlation is then based on the difference between the ranks for the two data sets. No results are presented for this here, it is simply mentioned to demonstrate that even for something as apparently straightforward as correlation, there are a number of options open to the modeler.

Correlation values have the range [-1,+1] with +1 indicating a perfect positive correlation and -1 a perfect negative correlation (i.e. a change in one variable produces an opposite change in the other variable).

A scatter-plot can be used to provide a visual indication of the correlation of the two sets of data. Here, the numerical values of the data sets are put into two columns; the notional independent axis information presented in the original data sets is ignored. These columns then form the x and y coordinates of the plotted graph. The closeness of the resulting data to a straight line indicates the level of association between the two data sets.

Boxplots provide a summary of the data distribution for the individual data sets by determining the distribution of the data displaying a box representing the upper and lower quartiles of the distribution, with the median value given as a straight line within this box. Fences give the extremes and outliers are specifically highlighted. From here, outlier values can be removed. However, the removal of outliers may not be appropriate in electromagnetics as an outlier could represent a correct, but extreme; result such as a high Q resonance.

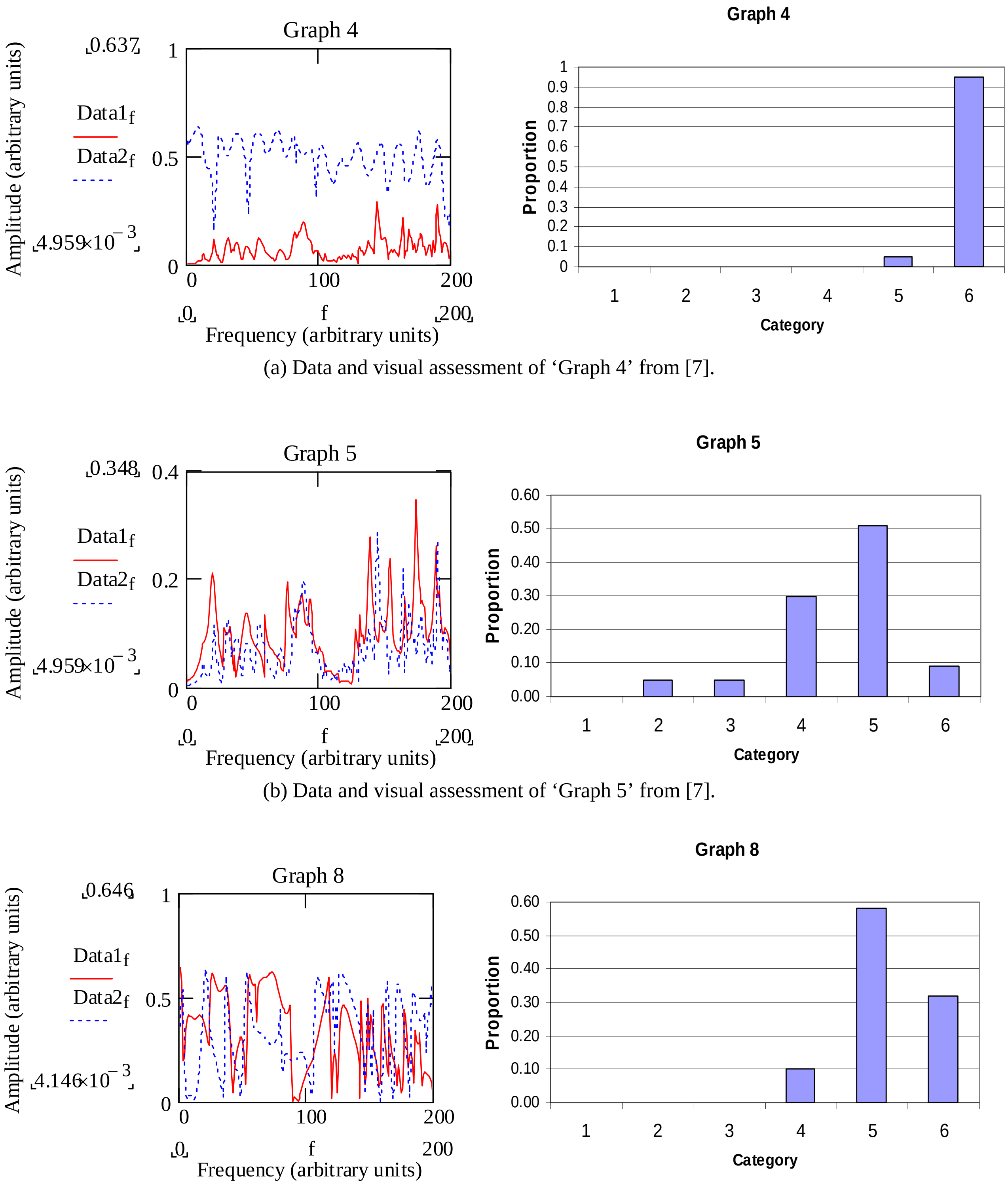
The Pearson correlation coefficients for the three sets of data in Fig. 1 are given in Table 1.

Table 1. Correlation coefficients (Pearson r) for the three comparisons above.

	“Graph 4”	“Graph 5”	“Graph 8”
Correlation Coefficient	0.022	0.383	0.040

The scatterplots and boxplots are given in Fig. 3. “Data A” and “Data B” refer to the two data sets presented on each graph.

[1] All data has been generated using the SPSS statistical Software package



(a) Data and visual assessment of ‘Graph 4’ from [7].

(b) Data and visual assessment of ‘Graph 5’ from [7].

(c) Data and visual assessment of ‘Graph 8’ from [7].

Fig. 1. Original data for comparison and visual assessment based on approximately fifty responses. The Graph number refers to that used in [7]. The categories (x axis) in column 2 are 1- excellent, 2 - very good, 3 - good, 4 - fair, 5 - poor, and 6 - very poor.

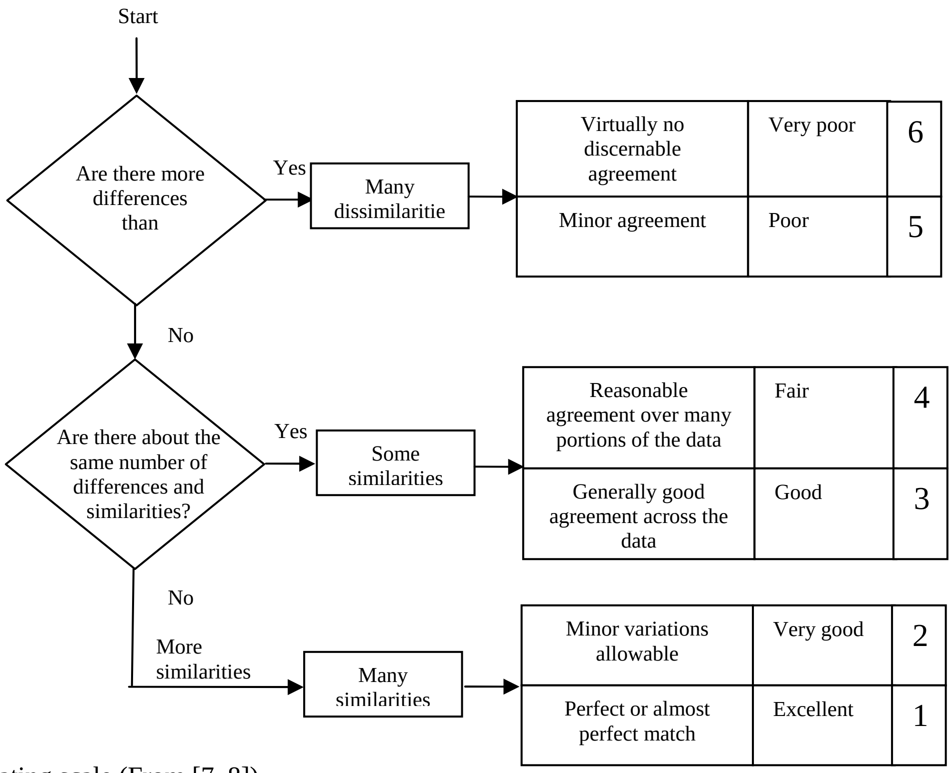

Start
Are there more differences than
Yes
Many dissimilaritie
Virtually no discernable agreement
Very poor
6
Minor agreement
Poor
5
No
Are there about the same number of differences and similarities?
Yes
Some similarities
Reasonable agreement over many portions of the data
Fair
4
Generally good agreement across the data
Good
3
No
More similarities
Many similarities
Minor variations allowable
Very good
2
Perfect or almost perfect match
Excellent
1

Fig. 2. Visual rating scale (From [7, 8]).

These statistics show some interesting properties. The rank ordering of the correlation coefficients is the same as the visual rank-ordering. However, the magnitude of the differences does not reflect the visual assessment. The generally low level of correlation could be seen to reflect the visual assessment. The scatterplots suggest that "Graph 8" is random, "Graph 4" has a slightly better association (the points are less randomly distributed across the graph) and "Graph 5" almost shows a hint of a positive gradient straight line. Interpretations of the scatterplots also support the visual assessment. The boxplots simply relate the data distributions with "Graph 8" clearly showing the greatest agreement.

III.2. Parametric tests

It is inappropriate to make the assumption that the data has a normal probability distribution, an implicit requirement of parametric tests, i.e. tests which consider the comparison of data parameters, such as means. However, for large sample sizes, the Central Limit Theorem does allow parametric tests to be used. The most common of these, the Student's t-test evaluates the difference in means for two groups. The resulting p-level gives a probability of error associated with rejecting the null hypothesis, i.e. the hypothesis that there is no difference in the two groups, when, in fact the hypothesis is correct. The results are summarized in Table 2.

Table 2. t-test results for the three original data sets.

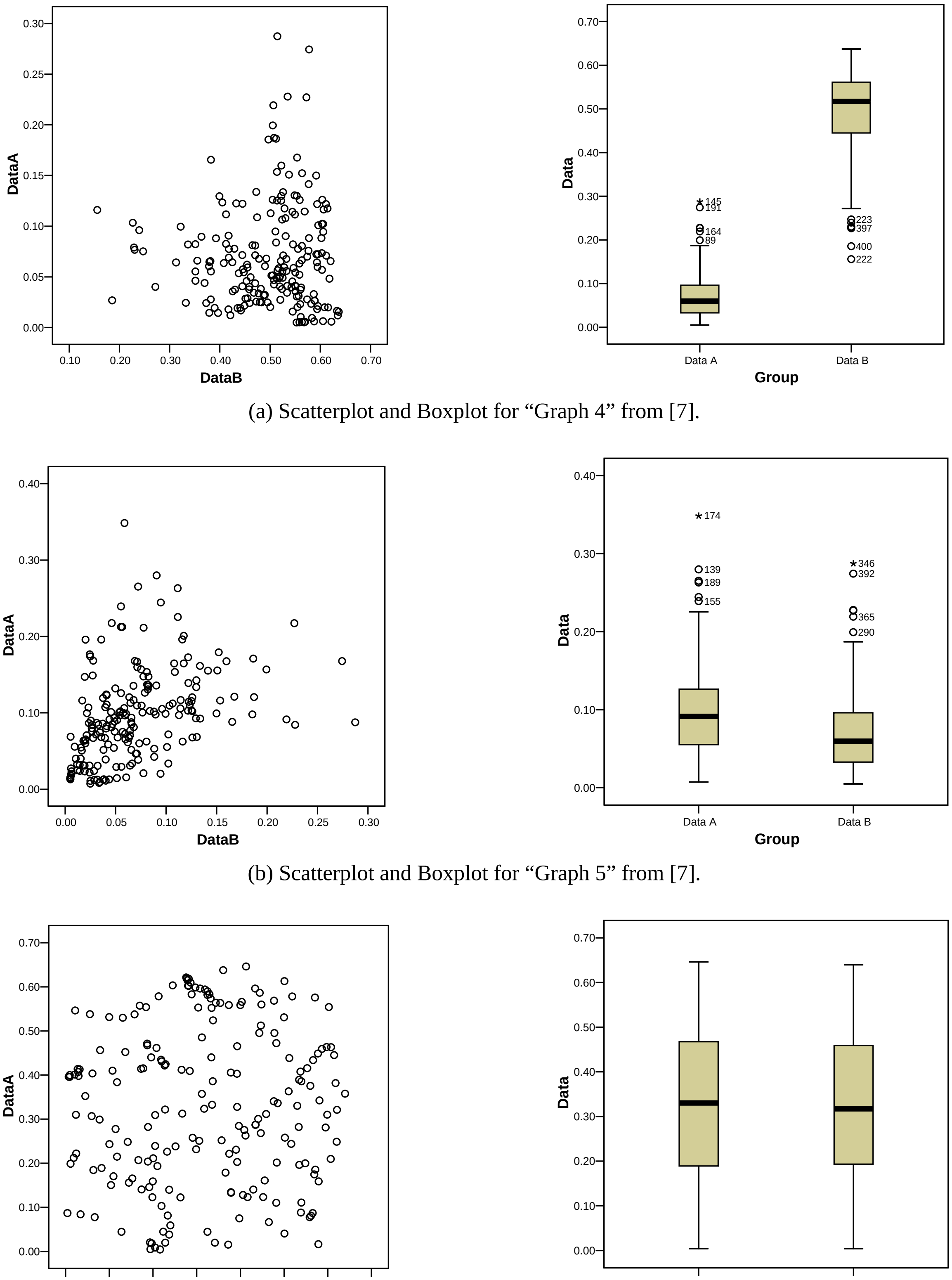
	"Graph 4"	"Graph 5"	"Graph 8"
t-parameter	-56	4.7	0.7
P value	0.000	0.000	0.473

These values show that only the data in "Graph 8" are similar. It shows that "Graph 4" is the worst comparison. Of course, as the purpose of the t-test is to compare means of groups, the results will confirm what has been demonstrated in the Boxplots of Fig. 3

III.3. Non-parametric tests

Non-parametric tests make no assumptions about the normality or otherwise of the data. They take into account the shape of the distributions. Two popular tests in electromagnetics are the χ^2 test and the Kolmogorov-Smirnov (KS) test. The χ^2 test measures the level of association between the two results. The KS test makes an assessment of whether there is sufficient evidence to reject the null hypothesis. It should, however, be noted that outliers can have a serious effect on the results.

The χ^2 test is based on a measure of the difference between two samples. The problem with this test is that it relies on dividing the square of the differences by the value of one of the data points and this results in the potential for different results depending on which data set is taken as a reference. The χ^2 parameter for the three graphs is given in Table 3.



(a) Scatterplot and Boxplot for “Graph 4” from [7].

(b) Scatterplot and Boxplot for “Graph 5” from [7].

(c) Scatterplot and Boxplot for “Graph 8” from [7].

Fig. 3. Scatter plots and Boxplots for “Graph 4”, “Graph 5” and “Graph 8”, respectively.

Table 3. χ^2 parameter.

	"Graph 4"	"Graph 5"	"Graph 8"
χ^2	75	20	149

This surprisingly suggests that "Graph 8" is much worse than "Graph 4".

The KS test converts the data sets into distributions and compares those distributions, looking for the maximum difference. Commonly, it is used to compare a data set with a known distribution but has here been used to compare two independent data sets. The results are given in Table 4. "Graph 8" comes out as a clear best comparison.

Table 4. KS-test results for the three original data sets.

	"Graph 4"	"Graph 5"	"Graph 8"
KS Z parameter	9.8	2.8	0.7
P value	0.000	0.000	0.714

IV. FEATURE SELECTIVE VALIDATION (FSV)

FSV is not a canonical statistical technique. It is, however, an increasingly accepted heuristic technique that finds favor particularly in the EMC community for validation of computation electromagnetics and has been discussed in detail in [7-13]. It is presented here because, like statistical techniques, its aim is to quantify confidence in the comparison of the data sets fed into it. Correlation techniques do this through the value of the correlation coefficient, other techniques do this based on the p-values, FSV does this using a variety of inbuilt metrics, the most general of which being the Global Difference Measure. The FSV tool with which the following results has been computed can be download from the official FSV web page [14] or directly from [15].

In overview, FSV works by taking the two original data sets and low and high pass filtering each of these. The low pass data is differenced, as detailed in [14], to give the Amplitude Difference Measure (ADM) which measures the level of (dis)agreement of the data envelope. First and second derivatives of the low and high pass data are differenced (as in [14]) to give the Feature Difference Measure (FDM), which measures the level of disagreement of the finer detail and features in the original data. The ADM and FDM are then combined to give the GDM as in equation (2)

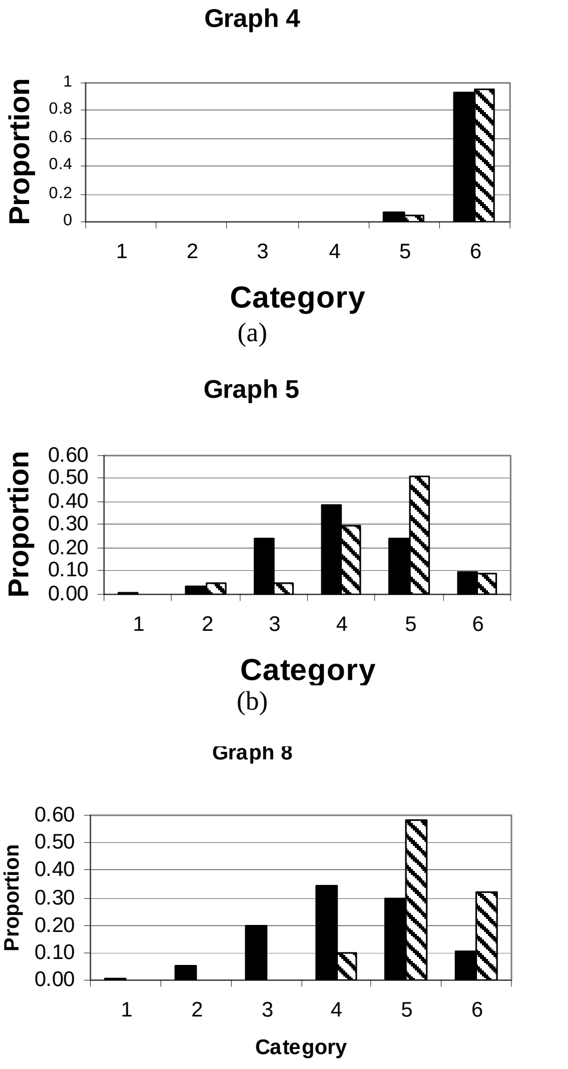
$$GDM = \sqrt{ADM^2 + FDM^2} \qquad (2)$$

As well as the single figure of merit given by the GDM, one particular useful feature of FSV is the confidence histogram, where the proportion of the GDM curve (on a point-by-point basis) is binned into the categories as noted in the visual rating scale of Fig. 2. The resulting probability density function has been shown to provide close analogue of the visual assessment of a group visual assessment.

The GDM values are given in Table 5; the confidence histograms for the data of Fig. 1 is given in Fig. 4.

Table 5. FSV (Global Difference Measure) results for original comparisons.

	"Graph 4"	"Graph 5"	"Graph 8"
GDM	5.26	4.41	4.67



(a)

(b)

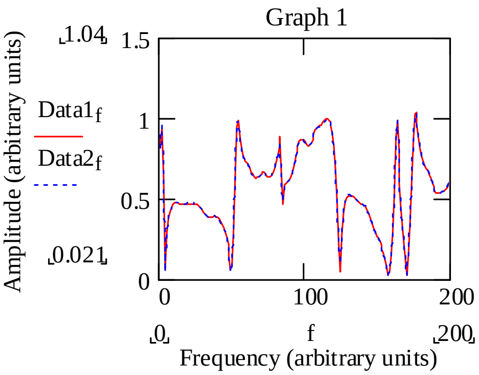
(c)

Fig. 4. FSV (dashed bars) compared to visual evaluation (solid bars) for the data in Fig. 1 from [7]. (a) "Graph 4", (b) "Graph 5", and (c) "Graph 8".

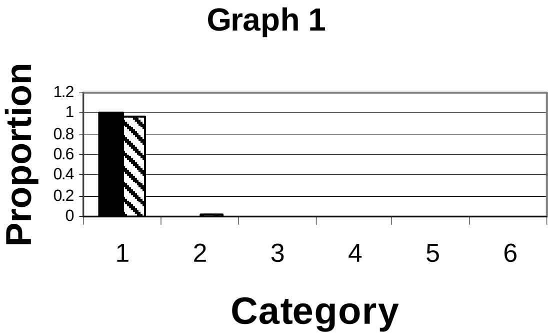
V. DISCUSSION

This paper has presented a non-mathematical summary of some of the most widely used statistical techniques as applied to computational electromagnetic validation. In particular, the emphasis has been to take a set of results already visually assessed by engineers familiar with performing this task and applying the techniques to see whether agreement could be obtained between the statistical techniques and the visual assessment. It should be noted that a paper such as this cannot prove the applicability or otherwise of individual statistical techniques, it can highlight the range of techniques available and suggest which are possibly more suitable than others.

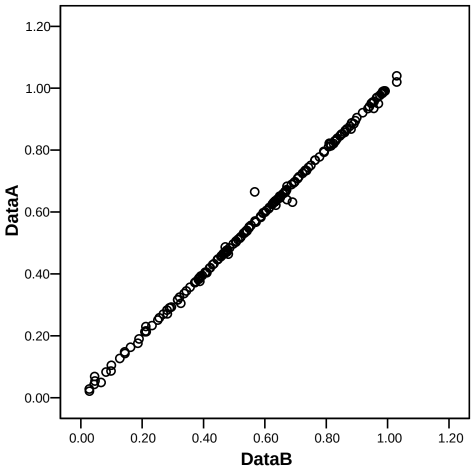
All the comparisons used in this paper have shown a marked difference between the two components. In order to show how the tests compare when there is very little difference, by way of a ‘lower bound’, the data of Fig. 5(a) was compared using the techniques discussed above. Fig. 5(b) shows the visual assessment and FSV assessment, Fig. 5(c) shows the scatter plot and Fig. 5(d) shows the box plots. The Pearson r correlation is 0.999, the χ^2 value is 0.05 (irrespective of the order of variables) and the t-test value is 0.004. Clearly, there is little doubt as to the generally very high level of agreement between the two graphs irrespective of which method is used in the comparison.



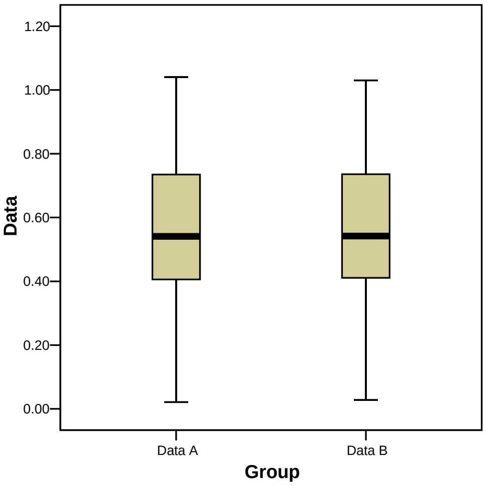
(a) From [7].



(b) From [7] - solid bar is visual assessment, dashed bar is FSV.



(c)



(d)

Fig. 5. Comparison of two data sets with very little difference. (a) Original data, (b) visual comparison compared to FSV comparison, (c) scatterplot, and (d) boxplots.

Scatterplots and correlation demonstrated the same rank-ordering of the data as the visual assessment. The boxplots and t-test results were in agreement but did not agree with visual assessment. The χ^2 test correctly identifies the best comparison whereas the KS test agrees with the t-test. In practice, the non-parametric tests are probably not well suited to large data sets [9] with parametric tests being more reliable due to the Central Limit Theorem. However, where there is a need to compare a data set with a known distribution, such as optimizing a reverberation chamber to produce a Rayleigh channel, then a χ^2 or KS approach would be well suited. A difficulty with applying χ^2 to two sets of data is that it relies on one ‘reference’ set. If using it to compare models against measurements or models against models, the reference set must be unambiguous (for example, by changing the ‘reference’ set in the previous table, the χ^2 parameter for “Graph 4” = 1386!) and the user must be aware of points that are very close to zero in the reference set as this can produce an unnecessarily dominating effect on the final value (for example, by changing the ‘reference’ set in the previous table, the χ^2 parameter for “Graph 4” = 1386!).

In all cases, a real benefit derived from the application of a statistical approach to validation of computational electromagnetics is that it provides an objective starting point to discuss the comparisons and agree a conclusion.

REFERENCES

[1] R. Holland and R. St John, *Statistical Electromagnetics*, Francis, Philadelphia, PA, 1999.

[2] D. Carpenter, “Statistical electromagnetics: an end-game to computational electromagnetics,” *IEEE Int. Symp. on EMC*, pp.736 – 741, 2006.

[3] H. Sasse and A. P. Duffy, “Satisficing in computational electromagnetics,” *Applied Computational Electromagnetics Society Newsletter*, vol. 21, no. 2, 2006.

[4] A. Coates, H. Sasse, D. E. Coleby, A. P. Duffy, and A. Orlandi, “Validation of a three dimensional transmission line matrix (TLM) model implementation of a mode stirred reverberation chamber,” *IEEE Trans. on EMC*, in press.

[5] P. Corona, J. Ladbury and G. Latmiral, “Reverberation chamber research – then and now: a review of early work and comparison with current understanding,” *IEEE Trans. on EMC*, vol. 44, no. 1, pp. 87 – 94, Feb. 2002.

[6] C. Bruns and R. Vahldieck, “A closer look at reverberation chambers – 3D simulations and experimental verification,” *IEEE Trans. on EMC*, vol 47, no. 3, pp. 612 – 626, Aug. 2005.

[7] A. Orlandi, A. P. Duffy, B. Archambeault, G. Antonini, D. E. Coleby, and S. Connor, “Feature selective validation (FSV) for validation of computational electromagnetics (CEM). Part II – assessment of FSV performance,” *IEEE Trans. On EMC*, vol. 48, no. 3, pp. 460 – 467, 2006.

[8] D. E. Coleby and A. P. Duffy, “A visual interpretation rating scale for the validation of numerical models,” *COMPEL: The Int. Journal for Computation and Mathematics in Electrical and Electronic Engineering*, vol. 24, no. 4, pp. 1078 – 92, 2005.

[9] StatSoft Inc., *Electronic Statistics Textbook*, Tulsa, OK, Statsoft, USA, 2006.
WEB: http://statsoft.com/textbook/stathome.html

[10] D. G. Rees, “Essential Statistics, 4/e, 2000, Chapman and Hall / CRC, Boca Raton

[11] J. Devore and R. Peck, *Statistics – the exploration and analysis of data*, 2/e, Duxbury Press, Belmont, California, 1993.

[12] T. T. Soong, *Fundamentals of probability and statistics for engineers*, Wiley, Chichester, UK, 1993.

[13] A. P. Duffy, A. J. M. Martin, A. Orlandi, G. Antonini, T. M. Benson, and M. S. Woolfson, “Feature selective validation (FSV) for validation of computational electromagnetics (CEM). Part I – the FSV method,” *IEEE Trans. on EMC*, vol. 48, no. 3, pp. 449 – 59, 2006.

[14] FSV official webpage:
http://www.eng.dmu.ac.uk/~apd/FSV/FSV%20web/

[15] FSV downloads at:
http://ing.univaq.it/uaqemc/FSV_3_2_2/