Int. J. Simul. Multisci. Des. Optim. 2014, **5**, A05
© H. Hamad et al., Published by EDP Sciences, 2014
DOI: 10.1051/smdo/2013003

IJSMDO

Available online at:
www.ijsmdo.org

ARTICLE OPEN ACCESS

Interpretability and variability of metamodel validation statistics in engineering system design optimization: a practical study

Husam Hamad[1,a,*], Awad Al-Zaben[2] and Rami Owies[3]

[1] Electronics and Communications Engineering Department, College of Engineering at Al-Lith, Umm Al-Qura University, Makkah Al-Mukarramah, KSA
[2] Biomedical Engineering Department, Hijjawi College of Engineering Technology, Yarmouk University, Irbid, Jordan
[3] Biomedical Engineering Department, College of Engineering, Jordan University of Science and Technology, Irbid, Jordan

Received 14 December 2012 / Accepted 5 November 2013 / Published online 4 February 2014

Abstract – Prediction accuracy of a metamodel of an engineering system in comparison to the simulation model it approximates is one fundamental criterion that is used in metamodel validation. Many statistics are used to quantify prediction accuracy of metamodels in deterministic simulations. The most frequently used ones include the root-mean-square error (RMSE) and the *R*-square metric derived from it, and to a lesser degree the average absolute error (AAE) and its derivates such as the relative average absolute error (RAAE). In this paper, we compare two aspects of these statistics: interpretability of results returned by these statistics and their sample-to-sample variations, putting more emphasis on the latter. We use the difference-mode to common-mode ratio (DMCMR) as a measure of sample-to-sample variations for these statistics. Preliminary results are obtained and discussed via a number of analytic and electronic engineering examples.

Key words: Simulation, Modeling, Metamodel validation.

1. Introduction

Simulators are used in the design of many engineering systems. For example, the design of electronic integrated circuits usually involves the use of transistor-level simulators such as PSPICE. Because such simulators are often expensive in terms of simulation times, different approximation models-that are often termed metamodels- may be used to replace simulation models. Metamodels are built and validated using simulation results for samples of data points in the input space. Two fundamental criteria are used as the basis for accepting or rejecting a metamodel: efficiency and accuracy. Efficiency is indicative of how expeditiously predictions can be obtained; accuracy is indicative of how good these predictions are.

Efficiency of a metamodel can be determined prior to metamodel construction, and without any computational cost in terms of the simulation runs needed, e.g., the time taken to evaluate a second-order polynomial metamodel in a given number of dimensions is the same regardless of the underlying simulation model. On the other hand, determining the accuracy of a metamodel is closely linked to the number of data points used in error calculations.

The accuracy of a metamodel is determined using quantitative methods which are mostly based on average statistics, or subjective methods using data displays such as box plots as in Sargent [1] and Kleijnen and Deflandre [2]. Hamad and Al-Hamdan use circle, ordinal, and marksman plots in [3] and [4].

Two of the most popular quantitative measures used in deterministic simulations to validate metamodels in terms of their accuracy of prediction are the root-mean-square error (RMSE) and the average absolute error (AAE), or some of their derivates. Calculation of these statistics is obtained usually using all of the available points in validation test samples, but in some techniques cross-validation methods are employed using subsets of the available test data; see Martin and Simpson [5] and Meckesheimer et al. [6].

Derivates of the RMSE and AAE statistics which are in essence relative error averages are sometimes used in the literature to give more interpretability to the results returned by RMSE and AAE, or otherwise to enable comparisons of metamodels especially when responses from different disciplines are approximated. Two of the mostly used ones include *R*-square and relative average absolute error (RAAE) or the like; see for example Jin et al. [7]. The *R*-square metric is in essence the square of the RMSE relative to the variance of the response data in the test sample, while RAAE may be obtained by

[a] Husam Hamad is on a sabbatical leave from Yarmouk University, Jordan
*e-mail: husam@yu.edu.jo

relating AAE to the standard deviation (defining equations for these statistics are given later). Note that some applications use AAE and RMSE relative to the average response instead, e.g., Qu et al. [8]. Other work relates RMSE to the range of response values in the test data samples, e.g., Eeckelaert et al. [9].

An important assumption for the validity of the results obtained by average-based statistics is related to the number of data points used in test samples. Of course, results of these statistics are meaningful only if the data used is sufficient in number. It is often the case that obtaining a sufficient number of observations is impractically expensive for complex simulation models. For such cases, average-based metrics such as RMSE and AAE may be "sensitive" to the number of observations used.

This paper provides a comparative study of four of the average-based statistics used for metamodel assessment in terms of prediction accuracy. The statistics used in this study are the RMSE and *R*-square, and AAE and RAAE. Two aspects of these metrics are considered: interpretability of results and sample-to-sample variation in the results. We put more emphasis on the issue of sample-to-sample variation, introducing a measure to quantify this variation. The term given to this measure is the common-mode to difference-mode ratio (DMCMR), as defined in the next section.

The remainder of this paper is organized as follows. In Section 2, the four statistics mentioned above are defined and contrasted in terms of their results interpretability and variability, after defining what we mean by these terms. Preliminary results are presented via examples in Section 3 with a discussion in Section 4. The paper is then concluded by Section 5.

2 Statistics for metamodel prediction accuracy

In this section we define and compare the four statistics of RMSE, *R*-square, AAE, and RAAE used for expressing prediction accuracy of metamodels in relation to their respective simulation models. We compare two aspects of these statistics: interpretability of the results they return and the sample-to-sample variation of these results. We start by defining these four statistics. We then clarify the term interpretability used in this context, followed by the definition of DMCMR – the measure we use to quantify sample-to-sample variations.

2.1 RMSE and *R*-square

Two of the more important measures used for model accuracy assessment including deterministic simulation models are RMSE and *R*-square. They are defined by

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i-1}^{n} (\hat{y}_i - y_i)^2}{n}}, \tag{1}$$

$$R^2 = 1 - \frac{\mathrm{MSE}}{\sigma^2}, \tag{2}$$

where MSE is the mean square error and σ^2 is the variance. In these equations and the discussion to follow, y_i is used to denote the response modeled by $\hat{y}_i$ for the ith data point in a validation test sample having n observations. Note that *R*-square is essentially derived from RMSE by squaring it then relating the result to the variance of y_i data. Theoretical thresholds for best accuracies are zero and unity for RMSE and *R*-square, respectively.

2.2 AAE and RAAE

The other two statistics that are compared in this paper are the AAE and its derivate RAAE. AAE is defined by

$$\mathrm{AAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|. \tag{3}$$

Rather than relating AAE to the standard deviation of the response data y_i as in Jin et al. [7], we obtain RAAE by relating AAE to the average of absolute values of the response data y_i in the validation sample. Defining RAAE this way gives more interpretability to the results as explained shortly and provides a simpler means for comparing results for different responses. RAAE is then defined as

$$\mathrm{RAAE} = \frac{\mathrm{AAE}}{\frac{1}{n}\sum_{i=1}^{n} |y_i|} \times 100\,\%. \tag{4}$$

As can be seen from these defining equations, RAAE can be thought of as representing a measure for the percentage error in the metamodel. The lower thresholds indicative of best accuracies for both AAE and RAAE are zero.

2.3 Interpretability of results

Assume that in a test problem RMSE is found to be 0.05 and 50 in another, then which of the two metamodels has better prediction accuracy for its respective simulation model? Assume on the other hand that the statistics for *R*-square are given instead as 0.78 and 0.98 for the two metamodels, respectively, then which one is better?

Note that results returned by RMSE cannot be interpreted without referring to the context of the problem. For the example given, if the response values are within the range 0.04–0.06 for the first metamodel with RMSE = 0.05, and 10,000–20,000 for the second metamodel having RMSE = 50, then obviously the second metamodel is superior. On the other hand, the closer *R*-square result for a metamodel is to unity the better its fit quality is regardless of response values. Looking at interpretability of results of these statistics from another perspective, what thresholds can we set on the values of the four statistics defined above for adequate metamodels? Note that thresholds can be readily set for *R*-square and RAAE, but not for RMSE and AAE without reference to the context of the problem. For example, the lower and upper thresholds in a given situation may be set at 0.95 and 5% for *R*-square and RAAE,

respectively. However, thresholds for RMSE and AAE cannot be easily set without reference to the context of the problem.

Based on this discussion, we use two classifiers in this paper for interpretability of statistical results in terms of prediction accuracy. The results are either: (1) interpretable, or (2) not interpretable without context. Hence, *R*-square and RAAE as defined in equations (3) and (4) above are classified as having "interpretable results", while results returned by RMSE and AAE are classified as "not interpretable without context".

2.4 Sample-to-sample variation using DMCMR

Another probably more important issue related to the results returned by these statistics concerns the sample-to-sample variations in the results. Ideally, variations should approach zero for deterministic simulation models. However, achieving such ideal results comes at the cost of increased test samples sizes. In this work, we quantify sample-to-sample changes in results using DMCMR, the difference-mode to common-mode ratio, where DMCMR for the two quantities ζ_1 and ζ_2 is defined by

$$\text{DMCMR} = \frac{\zeta_1 - \zeta_2}{\frac{1}{2}(\zeta_1 + \zeta_2)} \times 100\,\%, \tag{5}$$

where absolute values are taken to calculate the common-mode component in the dominator of equation (5).

Variations with sample size n in the statistics defined by equations (1)–(4) above are expected to be noticeable for test sample sizes which are not adequate for taking averages, like all other average-based quantities. However, since the sample size n appears in different forms in these equations (RMSE $\propto n^{-1/2}$, AAE $\propto n^{-1}$, while n cancels out in the numerator and dominator for *R*-square and for RAAE), then variations with n are expected to be different for these four statistics, as will be demonstrated by the examples of the next section.

3 Examples

We compare in this section variation with sample size for RMSE and *R*-square on one hand, and AAE and RAAE and on the other hand, via three examples. For these examples:

- Polynomials metamodels are used. The number of coefficients q for a polynomial in k dimensions with a degree d is $q = (k+d)!/k!d!$. These q coefficients are determined by the method of least squares in the examples.
- Latin hypercube validation test samples are used to determine the four statistics above. Sample sizes of ωq are used, where the number of coefficients multiplier ω is varied in steps of 1 starting at $\omega = 1$. Latin hypercube sampling is used to provide flexibility with sample sizes and good uniformity over the input space.

The examples are taken from Hamad [10]. Example 1 uses a one-dimensional analytic function for the response. For this example, two metamodels having different number of coefficients q are studied. In Example 2, a two-dimensional function that is frequently used in the literature is modeled, also via two metamodels with different complexities. The third example involves simulation results for an electronic circuit with three design variables (inputs).

3.1 Example 1

The following response is defined for the space $x \in [-1,1]$:

$$y = 1 \times 10^{-5}\left(e^{20x} - 1\right) + 1000. \tag{6}$$

Two metamodels are derived for this response: the first one is a second-order polynomial built using a minimum bias design having four points, and the second metamodel is a fifth-order polynomial derived using another minimum bias design with 10 points.

Accuracy tests are carried out using 50 samples for each metamodel. The number of observations for these samples are ωq; $\omega = 1, 2, \ldots, 50$. The number of coefficients q for the second-order and fifth-order polynomials is three and six, respectively. Calculations of RMSE, *R*-square, AAE, and RAAE are carried out for the second-order polynomial metamodel using the 50 test samples in turn. Results are shown in Figure 1 for one Latin hypercube sampling trial for each of the 50 samples.

Note from the figure that each metric settles at a "final" value shown by solid lines crossing the plots in the middle. These final values for RMSE in part (a) of the figure and AAE in part (c) are at approximately 750 and 250, respectively. Based on these two results, is the second-order polynomial an acceptable metamodel? Moving on to parts (b) and (d) of the figure for *R*-square and RAAE final values of approximately 0 and 19%, respectively, now is the metamodel acceptable?

Note that results from RMSE and AAE cannot be interpreted without reference to the context of the problem to examine the range of response values. On the other hand, merely taking into consideration that the value of *R*-square is much lower than the upper limit of one, then it is concluded that there is a problem with the metamodel prediction capability. Similarly, RAAE results can be interpreted by the mere consideration of the values returned. A value of 19% for RAAE also indicates that there is a problem. Note that some kind of quantification to the size of the problem may also be inferred from RAAE results; it can be further concluded that the size of the problem is such that the prediction performance of the metamodel is erroneous by 19% on average. In light of the discussion given so far, we may therefore say that results returned by *R*-square and RAAE are classified as "interpretable" by comparison to RMSE and AAE results which are "not interpretable without context".

Sample-to-sample variations measured by DMCMR defined in equation (5) above are depicted in Figures 2 and 3. Figure 2a shows variation results for RMSE superimposed on those for *R*-square. Similarly, Figure 2b shows superimposition of results for AAE and RAAE. Figure 3 shows sample-to-sample variations for all four metrics for sample sizes of $10q$ or less.

Note from Figure 2a that the minimum sample size is at approximately $25q$ for RMSE to settle to within ±10% of the final value (the two dotted lines running across the figure),

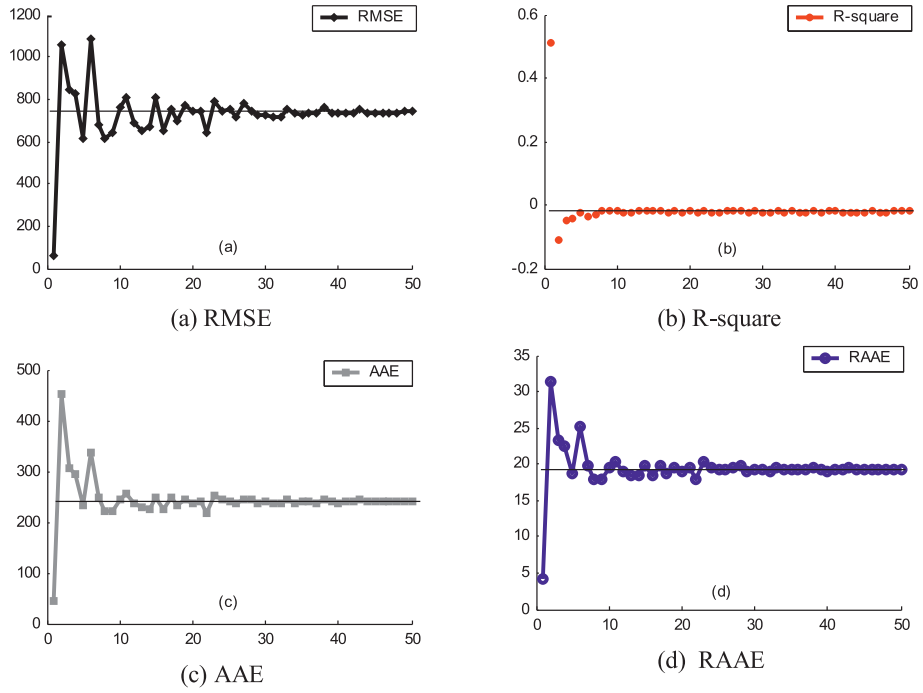


Figure 1. Validation results for the second-order polynomial metamodel vs. the number of coefficients multiplier ω.

and approximately 30q for R-square. For smaller sample sizes of 10q or less, Figure 3 reveals that as a whole R-square performs worse than RMSE vis-à-vis sample-to-sample variations. This unexpected result may be explained by noting that the response y in equation (6) above is almost constant for most of the input space and increases sharply for $x > 0.8$, a condition which is not favorable for R-square calculation. To explain, refer to equation (2) above. It can be seen that the dominator of the second term in equation (2) is small for nearly constant response values, and even if the metamodel returns nearly zero MSE, the value of the second term in equation (2) resulting from dividing two small numbers is not without numerical problems. A similar situation leading to questioning the validity of R-square results that is dealt with in the literature rises for the case when n is close to the number of coefficients q in the metamodel. For such cases, R-square is "adjusted" to accommodate the relative size of n to q; see Kleijnen and Deflandre [2].

Sample-to-sample variations for AAE and RAAE can be compared by reference to Figures 2b and 3 for the smaller sample sizes. It can be seen from both figures that RAAE performs slightly better that AAE. Note from Figure 2b that both AAE and RAAE settle down to ±10% variations at a minimum sample size of around 25q, which is the same result for RMSE as discussed above. Referring to Figure 3, it can be seen that RMSE behaves only marginally worse that AAE or RAAE.

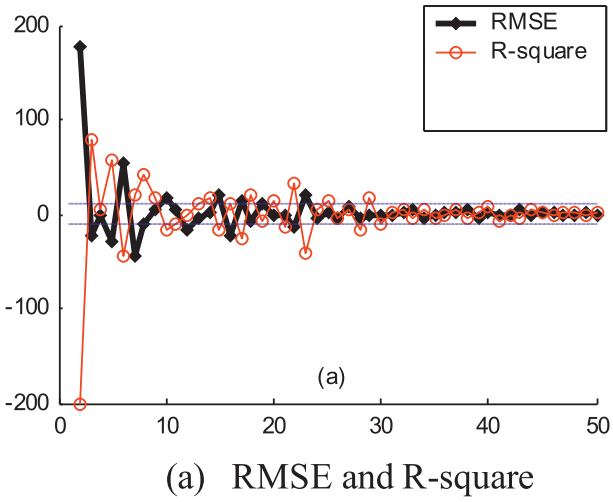
In order to investigate the sample-to-sample variations of the four metrics in more detail, 10 trials are carried out for each of the 50 Latin hypercube samples and DMCMRs are calculated for each metric. Then, minimum sample sizes after which the corresponding metric is confined within DMCMR levels of ±10% are noted and plotted in Figure 4. The average of such minimum sample sizes for the 10 trials for each metric are also calculated and given in Table 1. Figure 5 explains how the minimum size is determined for trial 1 for DMCMR calculations for RMSE.

As can be seen from Figure 4, R-square performs consistently worse for all 10 trials, while RAAE performs consistently the best. The results in Table 1 also show that RAAE has the best performance on average in terms of sample-to-sample variations in its results, with minimum average size of 11.7q. This means that a test sample of size 11.7q on average is considered "adequate" in the sense that the results returned will not be more than ±10% of the "true" value that would be obtained if the test sample size were infinite. By the same token, adequate sample sizes for RMSE, R-square, and AAE are 21q, 29.2q, and 16.2q respectively, as given in Table 1. Note that for this example R-square unexpectedly performs worse than RMSE. The reason for the unexpected performance of R-square was mentioned above, where it was explained that R-square results cannot be used in two situations: (1) if the number of observations n is close to the number of coefficients q, and (2) if the response is nearly constant for a sizable portion of the input space. Note in addition that R-square is close to zero for most test samples; see Figure 1b above.

In summary, RAAE outperforms the other three statistics for Metamodel 1 of Example 1 because: (1) it has the best interpretable results, and (2) it has the smallest adequate sample size of 11.7q, i.e., the minimum sample size for ±10% deviation from the "true" result is 11.7q or nearly 35 observations for this case.

3.1.1 Metamodel 2

The order of the metamodel polynomial is changed to five. The final values are approximately 450, 0.65, 170, and 13% for RMSE, R-square, AAE, and RAAE, respectively.



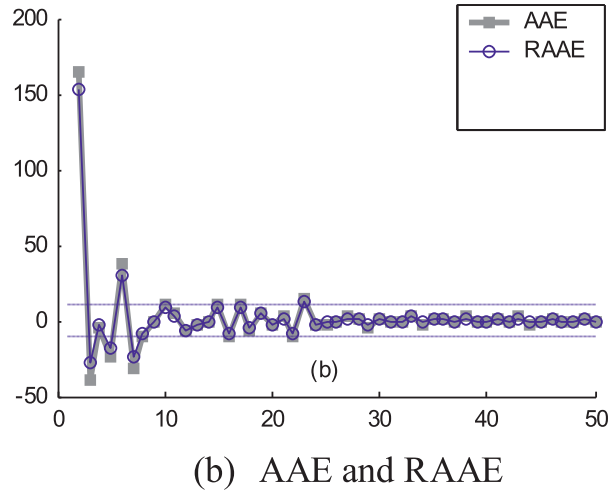


Figure 2. Sample-to-sample variations in DMCMR vs. ω.

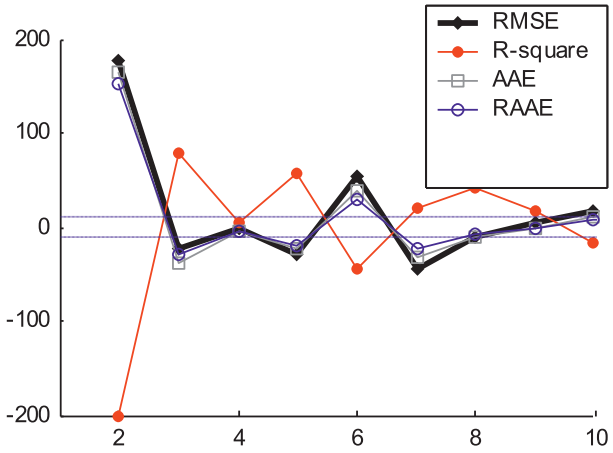


Figure 3. Sample-to-sample variations in DMSMR vs. ω for "smaller" samples.

Sample-to-sample variations for the four metrics are calculated, again using 50 Latin hypercube samples. Results for DMCMRs are shown for one sampling trial in Figure 6 superimposed for

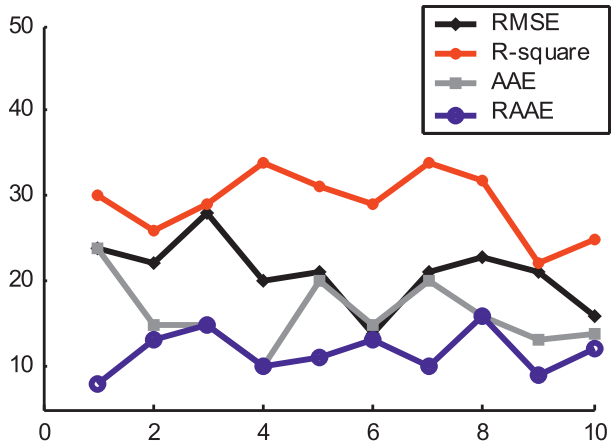


Figure 4. Minimum sample sizes for confinement within ±10% DMCMR using 10 trials for each of the 50 test samples.

Table 1. Average of minimum sample sizes for ±10% DMCMR levels confinement using 10 trials for each test sample.

Statistic	Metamodel 1	Metamodel 2
RMSE	21q	24.9q
R-square	29.2q	14.6q
AAE	16.2q	17.5q
RAAE	11.7q	16.7q

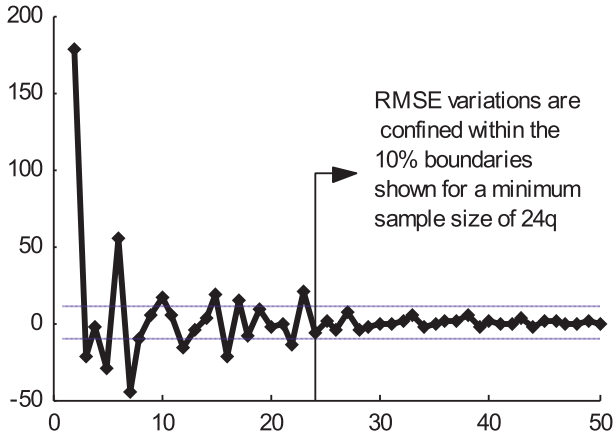
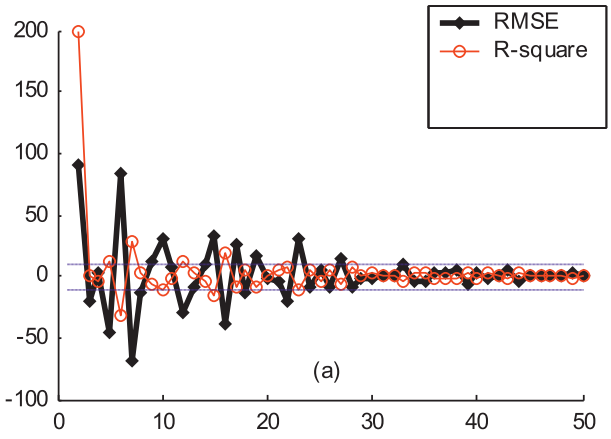


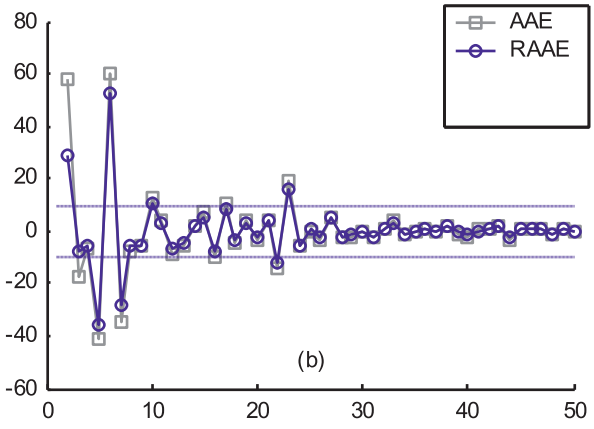
Figure 5. RMSE sample-to-sample variation for Latin hypercube sampling trial 1 showing the minimum sample size of 24q for confinement within ±10% DMCMR.

RMSE and R-square in part (a), and AAE and RAAE in part (b) of the figure.

As seen in Figure 6a, performance of R-square in terms of sample-to-sample variation is overall better than RMSE performance, as would be expected. Note for example that the minimum sample size for ±10% confinement of DMCMR level variations for the sampling trial shown is 28q for RMSE and 17q for R-square. The situation for AAE and RAAE is similar to that for Metamodel 1 above, with RAAE performing marginally better than AAE as seen in Figure 6b. Sample-to-sample variations for the four metrics are shown in Figure 7 for sample



(a) RMSE and R-square



(b) AAE and RAAE

Figure 6. Sample-to-sample variations in DMCMR vs. ω.

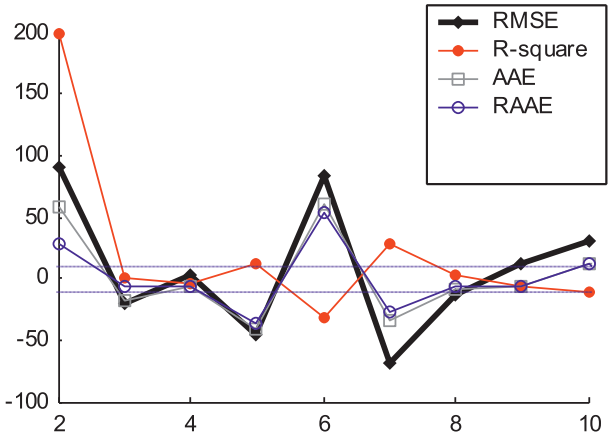


Figure 7. Sample-to-sample variations vs. ω for "smaller" samples.

sizes of $10q$ or smaller. It can be seen from the figure that for the smaller samples with sizes between $2q$ and $3q$ the variability is largest for R-square, while variability becomes largest for RMSE for sample sizes of $4q$ up to $10q$.

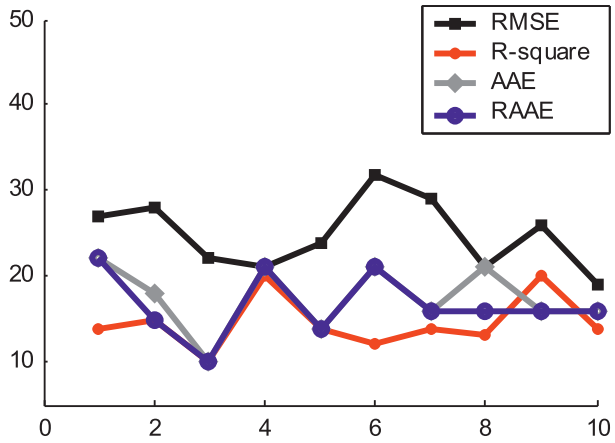


Figure 8. Minimum sample sizes for confinement within ±10% DMCMR using 10 trials for each of the 50 test samples.

In order to investigate the sample-to-sample variations of the four metrics in more detail, the scenario used for Metamodel 1 above is followed here where 10 sampling trials are carried out for each of the 50 Latin hypercube samples and DMCMRs are calculated for each metric. Then, minimum sample sizes after which the corresponding metric is confined within DMCMR levels of ±10% are noted and plotted in Figure 8. The average of such minimum sample sizes for the 10 trials for each metric are also calculated and given in Table 1 above.

Note from Figure 8 that RMSE performs consistently worse for all 10 trials. It can also be seen from the figure that R-square performs marginally better than AAE and RAAE for all 10 trials except trial 9. RAAE performs slightly better than AAE for all 10 trials; see Figure 8. These are the same conclusions in relation to the average minimum sample sizes in Table 1, where RMSE has the worst average of $24.9q$, and the other three metrics having nearly equal averages, with R-square slightly the best.

Therefore, we can conclude for Metamodel 2 of Example 1 that RMSE has the worst performance in relation to sample-to-sample variations. Also, interpretability of its results does not put it up in front of the other three metrics either. While the sample-to-sample variation for RAAE is slightly higher than R-square ($16.7q$ for RAAE and $14.6q$ for R-square as given in Table 1), however, its interpretability of results may make it the best choice in this situation as well. Thus, while the grounds for rejecting a metamodel because of an R-square of 0.65 are shaky, but on the other hand a metamodel with a RAAE of 13% is understood to have an error of 13% on average, may be giving clearer indications of whether to accept or to reject the metamodel.

3.2 Example 2

The second example involves the two dimensional response studied in other work including Martin and Simpson [5], Jin et al. [7], and Hamad et al. [11]; given by:

$$y = \cos\left(6(x_1 - 0.5)\right) + 3.1 \times |x_1 - 0.7| + 2(x_1 - 0.5) + \sin\left(\frac{1}{|x_1 - 0.7| + 0.31}\right) + 0.5x_2 \; x_1, x_2 \in [0, 1]. \tag{7}$$

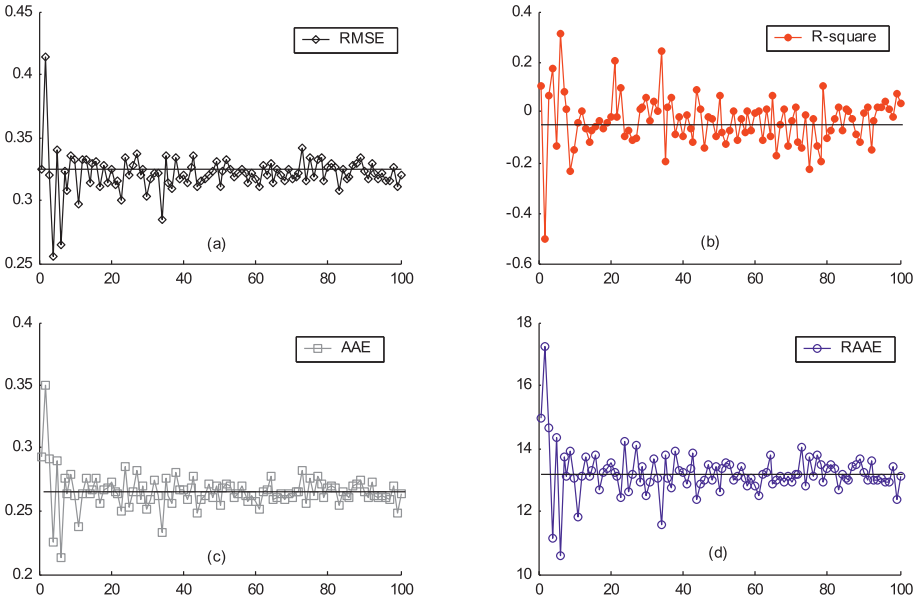


Figure 9. Validation results for Metamodel 1 vs. the number of coefficients multiplier ω.

Two metamodels are derived and tested for this response. Metamodel 1 is a second-order polynomial having $q = 6$, while Metamodel 2 is a piecewise metamodel consisting of two second order polynomials: one for each part of the input space partitioned along x_1 in two halves, and with $q = 6$ for each polynomial; see Hamad et al. [11].

3.2.1 Metamodel 1

Figure 9 shows validation results for Metamodel 1 using one-hundred Latin hypercube samples with sizes $\omega q = 1q$, $2q$, …, $100q$ observations. Part (a) of the figure shows that most of the one-hundred test samples have RMSEs around 0.32, while AAEs for most samples are around 0.27 as depicted in part (c); what can be inferred about the prediction validity of Metamodel 1 from these results? Again, prediction accuracy cannot be judged by merely referring to the results for RMSE or AAE.

Referring to *R*-square results of Figure 9b, however, it can be immediately concluded that the metamodel is not accepted in terms of its prediction merits because *R*-square is far from the upper threshold of unity for all test samples; what is wrong with Metamodel 1? This question can be answered by interpreting RAAE results of Figure 8d. RAAE results show that for the one-hundred samples tested, their RAAE levels are between 10% and 17%, with most points having 12–14% RAAE as depicted in the figure, i.e., the error for observations in these samples is between 12% and 14% on average.

Sample-to-sample variations for the four metrics are calculated, again using one-hundred Latin hypercube samples. Results for DMCMRs are shown in Figure 10 for one sampling trial for the first 50 samples superimposed for RMSE and *R*-square in part (a), and AAE and RAAE in part (b) of the figure. Superimposition of the four metrics is shown in Figure 11 for the first 10 test samples. As can be seen from these figures, *R*-square performance is poor; however, the results for the other three metrics are comparable.

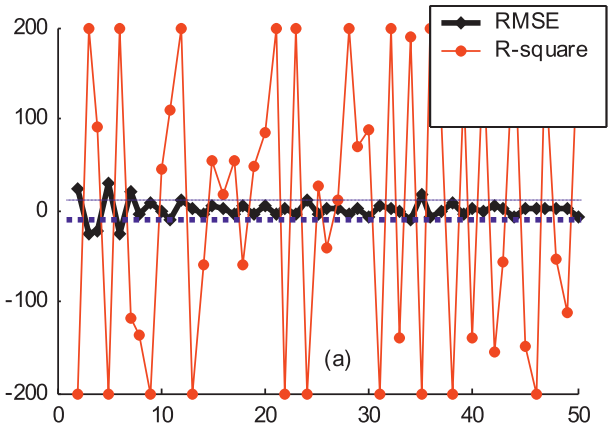
In order to investigate the sample-to-sample variations of the four metrics in more detail, we use 10 sampling trials as before for each of the first 50 Latin hypercube samples and DMCMRs are calculated for each metric. Then, minimum sample sizes after which the corresponding metric is confined within DMCMR levels of ±10% are noted and plotted in Figure 12. The average of such minimum sample sizes for the 10 trials for each metric are also calculated and given in Table 2.

It can be seen from Figure 12 and Table 2 that sample-to-sample variation is worst for *R*-square, with RMSE slightly better that AAE and RAAE. Note from Figure 9b that *R*-square is close to zero; this is a similar situation to that obtained for Metamodel 1 of Example 1 and shown in Figure 1b.

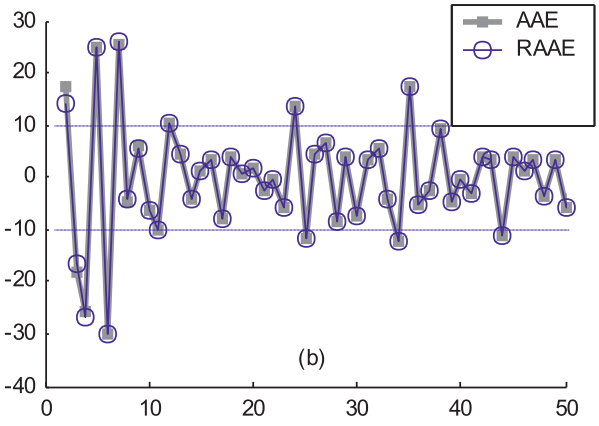
The response considered in this example is waving in the direction of x_1, and the global second-order Metamodel 1 cannot follow this waving response consistently leading to increased sample sizes as shown in Table 2. Results are improved by using the piecewise Metamodel 2 as demonstrated below.

3.2.2 Metamodel 2

The piecewise Metamodel 2 is validated to test its prediction accuracy using the same scheme outlined above for Metamodel 1. Results are shown in Figure 13 for *R*-square and RAAE only in order to save space. Part (a) of the figure shows that *R*-square is around 0.82 for most of the test samples indicating that the piecewise second-order metamodel is a good improvement relative to the global Metamodel 1. With this improved *R*-square value of 0.82, is Metamodel 2 acceptable? RAAE provides more information that helps in answering this question. Refer to Figure 12b, it is seen that RAAE is 4–5% for most samples, meaning that Metamodel 2 has 4–5% error on average. This allows for a more informed basis for judging the acceptability of the metamodel.



(a) RMSE and R-square



(b) AAE and RAAE

Figure 10. Sample-to-sample variations in DMCMR vs. ω.

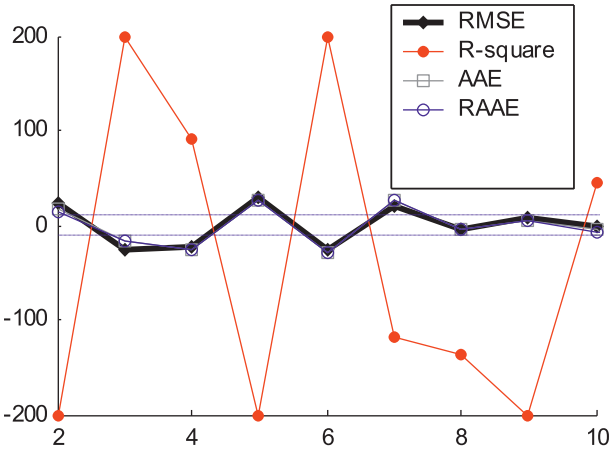


Figure 11. Sample-to-sample variations vs. ω for "smaller" samples.

Sample-to-sample variations are shown in Figure 14a for RMSE and *R*-square, and for AAE and RAAE in Figure 14b, and Figure 15 shows these variations for the smaller samples with sizes of 10q or less. To investigate these variations in more

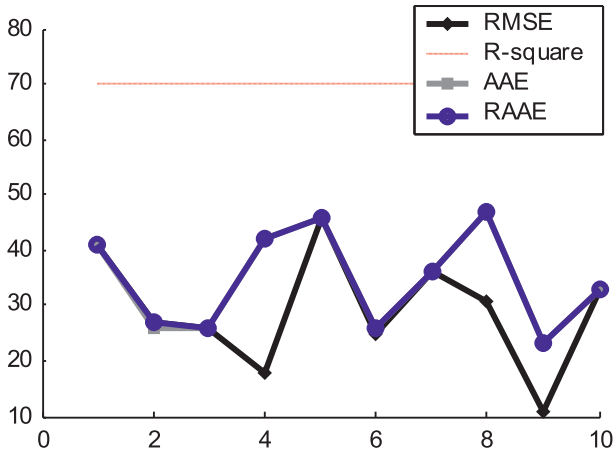


Figure 12. Min. sample sizes for confinement within ±10% DMCMR using 10 trials for each of the 50 test samples.

Table 2. Average of minimum sample sizes for ±10% DMCMR levels confinement using 10 trials for each test sample.

Statistic	Metamodel 1	Metamodel 2
RMSE	29.4q	6.4q
R-square	>100q	4.6q
AAE	34.6q	6.4q
RAAE	34.7q	6.4q

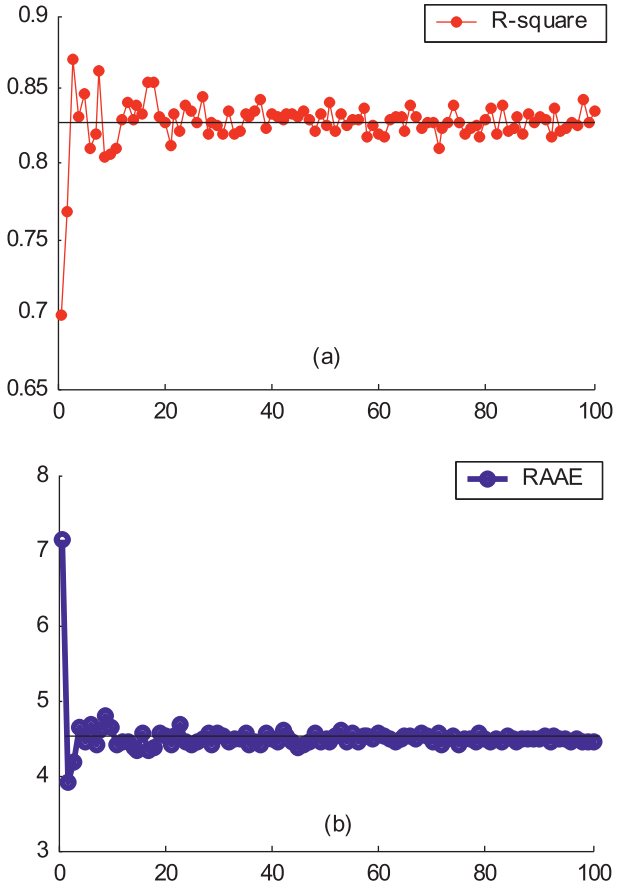


Figure 13. Validation statistics for Metamodel 2 vs. the number of coefficients multiplier ω.

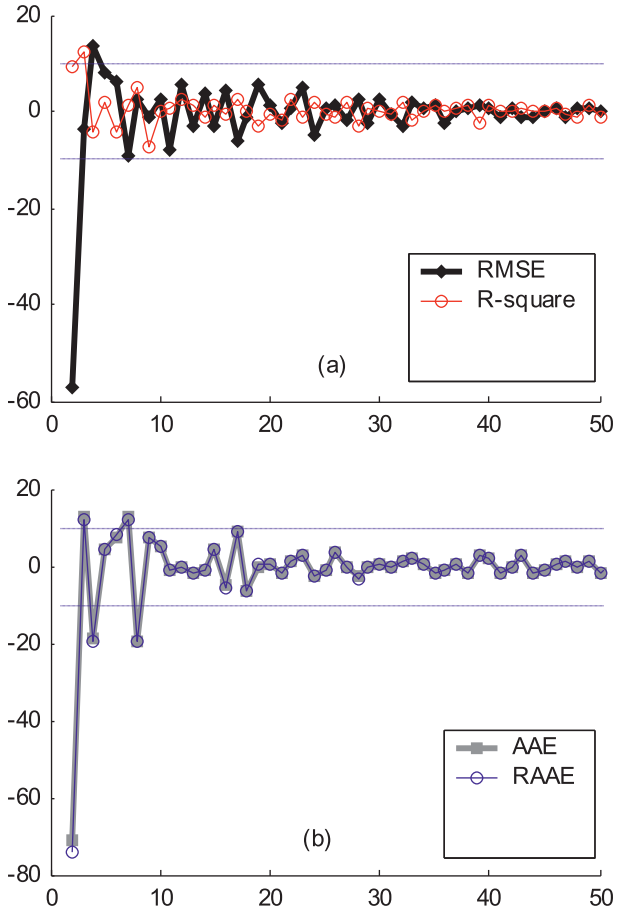


Figure 14. Superimposition of DMCMR results.

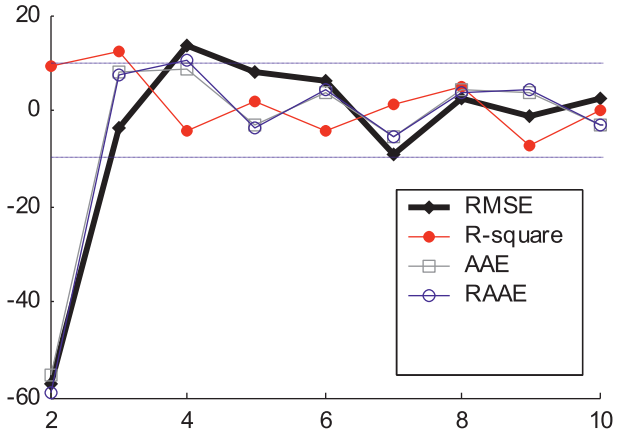


Figure 15. Variations vs. ω for "smaller" samples.

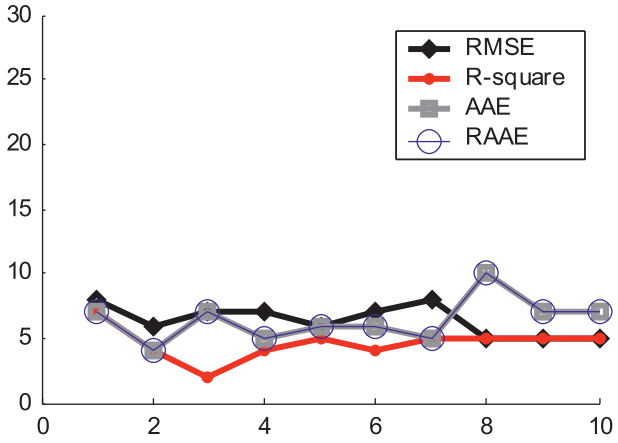


Figure 16. Minimum sample sizes for confinement within ±10% DMCMR.

details, the method used above for Metamodel 1 is repeated here for Metamodel 2, where 10 sampling trials for each of the 50 Latin hypercube test samples are examined for minimum sizes after which the respective metrics are confined to ±10% variations. The results are summarized by Figure 16, and Table 2.

It can be seen from Figure 16 and Table 2 that *R*-square performs marginally better than the other three metrics. RAAE performs the same as AAE for all 10 trials, and RMSE is slightly worse for the first seven trials; see Figure 16.

Similar conclusions can be made for Metamodel 2 in this example as those given above for Metamodel 2 in Example 1, where it was mentioned that while the sample-to-sample variation for RAAE is slightly worse than *R*-square, however, its interpretability of results may make it the best choice. Thus, while the grounds for rejecting a metamodel because of an *R*-square of 0.82 are shaky (see Figure 13a), but on the other hand a metamodel with a RAAE of 4–5% is understood to have an error of 4–5% on average, may be giving clearer indications of whether to accept or to reject the metamodel.

3.3 Example 3

The three-dimensional problem in this subsection is an electronic engineering problem that relates the portion H of the input signal that appears as an output in the circuit of Figure 17.

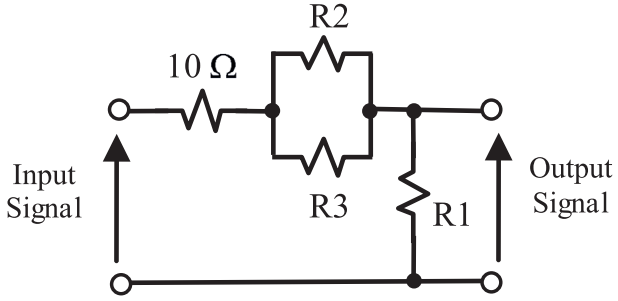


Figure 17. Electric circuit for Example 3.

The portion H is dependent upon the three design variables R_1, R_2, and R_3 connected as shown in the figure. Using a circuit simulator gives results which are identical to those given by the following equation obtained from elementary circuit analysis techniques

$$H = \frac{R_1R_2 + R_1R_3}{R_1R_2 + R_1R_3 + R_2R_3 + 10R_2 + 10R_3}. \quad (8)$$

A second-order polynomial is constructed from a minimum bias experimental design having seventeen points in the space

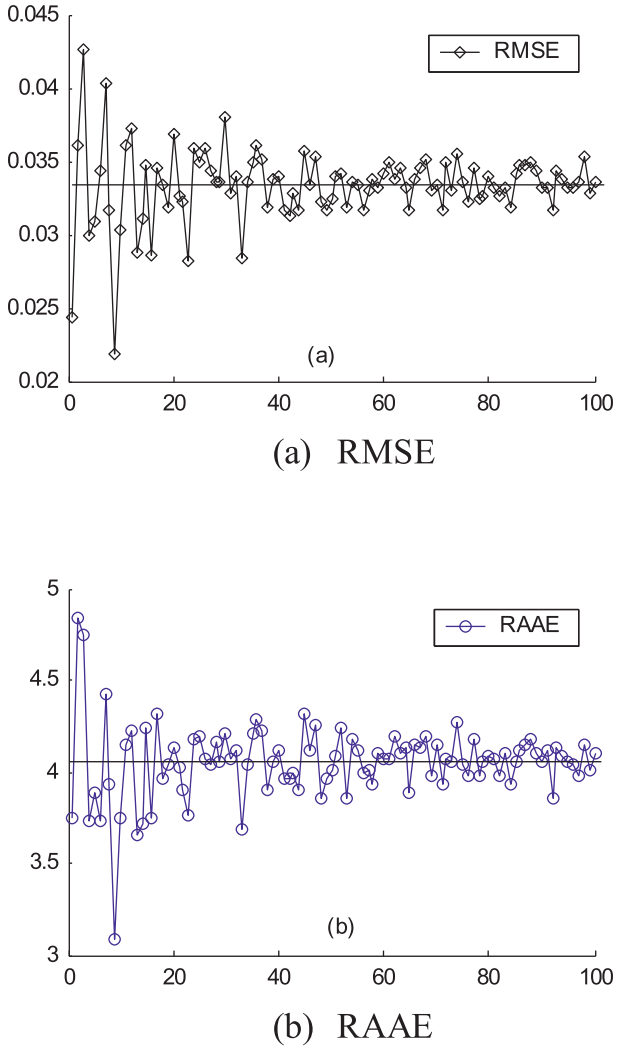


Figure 18. Validation statistics for Example vs. the number of coefficients multiplier ω.

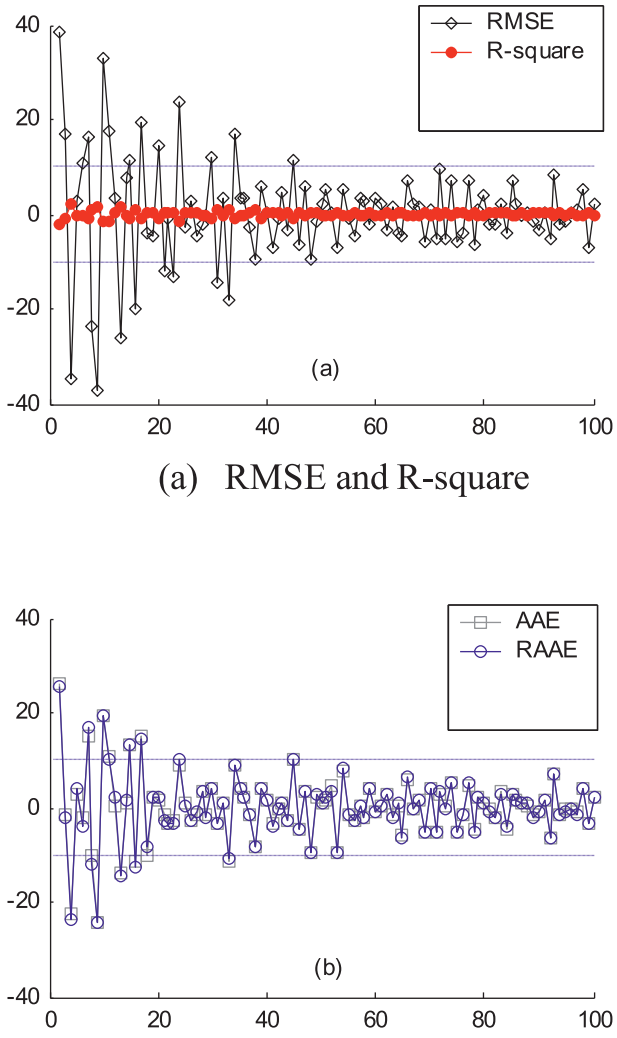


Figure 19. Superimposition of DMCMR results.

[1,100]^3. Accuracy tests are then carried out using one-hundred Latin hypercube samples. The number of observations for these samples are ωq; $\omega = 1, 2, \ldots, 100$. The number of coefficients q for this case is 10. Calculations of RMSE, R-square, AAE, and RAAE are carried out for the metamodel using the one-hundred test samples in turn. Results are shown in Figure 18 for RMSE and RAAE using one sampling trial for each of the one-hundred Latin hypercube test samples used. Results for AAE and R-square are omitted to save space.

Figure 18a shows that RMSE varies from a little above 0.02 to nearly 0.043, being approximately 0.033 ± 0.002 for most samples; is this metamodel accurate based on these RMSE results? Note that in equation (8) above $0 < H < 1$ for the entire input space; now the quality of the metamodel can be determined after this information about the response is known. Interpreting the results from R-square can be carried out without reference to the context of the problem; R-square values (not shown) are between 0.959 and 0.991 for all of the one-hundred test samples, a performance which is close to the upper theoretical threshold of unity.

Note that interpreting the results for RAAE shown in Figure 18b gives more information about the prediction accuracy of the metamodel. The figure reveals that for all of the one-hundred test samples RAAE is 3.1–4.8%, indicating that the error in metamodel predictions is on average between 3.1% and 4.8%.

Sample-to-sample variations are depicted by Figure 19a for RMSE and R-square, and by Figure 19b for AAE and RAAE, while Figure 20 depicts these variations superimposed for small sizes having $10q$ observations or less. The same scales are used in both parts of Figure 19 for easy comparison. Note from the figures that sample-to-sample variation is worst for RMSE, with the R-square performance being up in front of the other three statistics.

These sample-to-sample variations are investigated further by carrying out 10 sampling trials for each the first 50 of the 100 Latin hypercube samples and DMCMRs are calculated for each of the four metrics considered. Minimum sample sizes

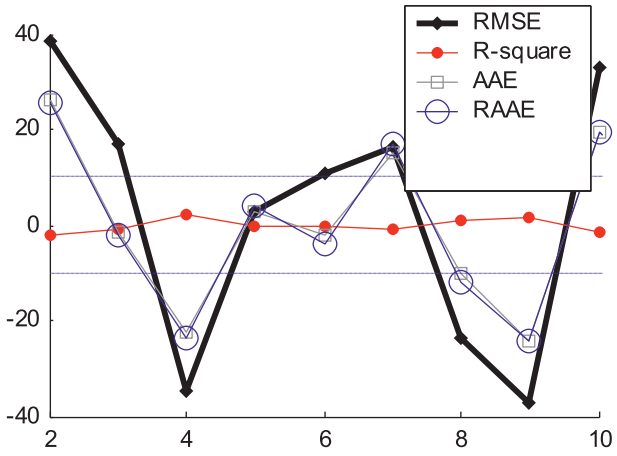


Figure 20. Sample-to-sample variations vs. ω for "smaller" samples.

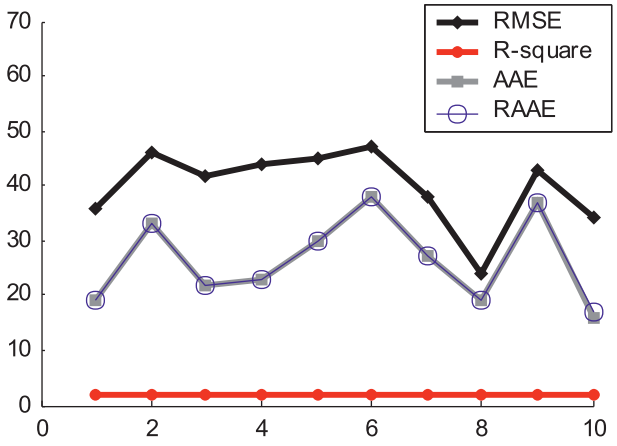


Figure 21. Minimum sample sizes for confinement within ±10% DMCMR.

Table 3. Average of minimum sample sizes for ±10% DMCMR levels confinement.

Statistic	Average size
RMSE	39.9q
R-square	2.0q
AAE	26.4q
RAAE	26.5q

after which the corresponding metric is confined within DMCMR levels of ±10% are plotted in Figure 21. The average of such minimum sample sizes for the 10 sampling trials for each metric are also calculated and given in Table 3.

It can be seen from Figure 21 and Table 3 that R-square far outperforms the other three metrics with an average minimum size of 2q, and RMSE showing worst results in terms of sample-to-sample variations with an average minimum size of nearly 40q. RAAE performs the same as AAE for nearly all 10 trial at averages of about 26q. Thus, R-square is up in the front being robust against sample-to-sample variations. Also, the results returned by it can be interpreted without context, as compared to those results returned by RMSE or AAE.

4. Discussion

Five metamodels are considered in this comparative study: two metamodels for each problem in Examples 1–2, and one metamodel in Example 3. These metamodels are validated in terms of their prediction accuracy using four statistics: RMSE and its derivate R-square, and AAE and its derivate RAAE. Results obtained using these four statistics are compared based on two criteria: interpretability and sample-to-sample variations. With regard to interpretability of results:

- RMSE and AAE results are not interpretable without referring to the problems context; a metamodel with RMSE of 0.05 and AAE of 0.02, say, may be an accurate model or otherwise depending on the response values used.
- On the other hand, there is no need to refer to the problem's context for R-square and RAAE results; an R-square of 0.999 is always welcome and a RAAE of 0.5% is also always taken positively. However, there is more "substance" in the information returned by RAAE since it points to the size of average error in prediction and not merely giving a yes/no answer to the metamodel acceptability as in the case for R-square.

Therefore, RAAE has some preference over R-square in terms of interpretability of results, with both RMSE and AAE way behind in this respect. This leading role for RAAE vis-à-vis interpretability does not necessarily guarantee the first position for it among the other three statistics; the other more important aspect related to sample-to-sample variability is still to be considered.

The performance of the four statistics in terms of sample-to-sample variation is thoroughly studied in this work: at least 50 test samples are used for any of the five metamodels and the number of observations used in these samples being 1q, 2q, up to 50q, with each of the 50 test samples undergoing 10 Latin hypercube sampling trials. Variability is quantified by the newly introduced measure referred to as difference-mode to common-mode ratio DMCMR. The minimum sample size for a given sampling trial for which DMCMR is confined within ±10% (see Figure 5) is noted for the 10 sampling trials and plotted for the five metamodels in the five Figures 4, 8, 12, 16, and 21. Averages of results in these figures are summarized in three tables: Tables 1–3. These five figures and three tables, summarized in Figure 22 for convenience, are the basis upon which preliminary results for sample-to-sample variability are given. The following can be concluded by reference to the summary in Figure 22:

- Sample-to-sample variability for RAAE shows marginal improvement over AAE variability in worst cases. Another more important merit for RAAE over AAE is related to interpretability of results. Therefore, if the choice has to be made between AAE and RAAE, then RAAE would be a better choice since it conveys more information about the metamodel prediction validity with reduced sample-to-sample variability.

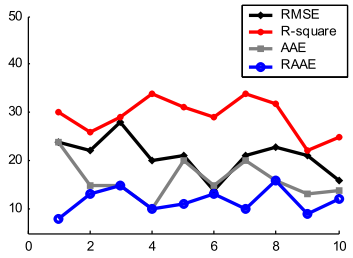
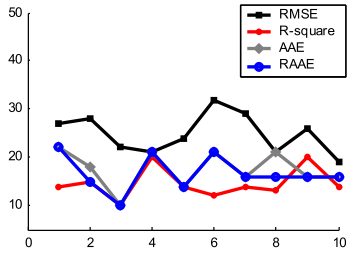
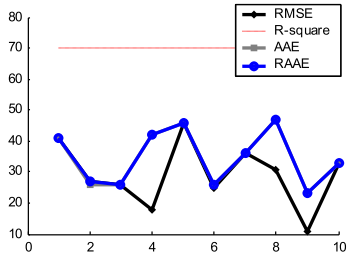
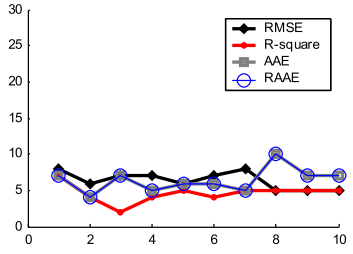
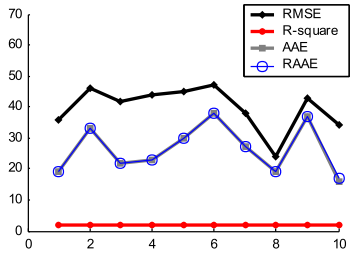
	Trial-by-trial results	Averages
Example 1: Metamodel 1 (see Fig. 4)		21q 29.2q 16.2q 11.7q
Example 1: Metamodel 2 (see Fig. 8)		24.9q 14.6q 17.5q 16.7q
Example 2: Metamodel 1 (see Fig. 12)		29.4q >100q 34.6q 34.7q
Example 2: Metamodel 2 (see Fig. 16)		6.4q 4.6q 6.4q 6.4q
Example 3 (see Fig. 21)		39.9q 2.0q 26.4q 26.5q

Figure 22. Summary of sample-to-sample variations results (averages in the last column are for, from top: RMSE, *R*-square, AAE, and RAAE).

- *R*-square has the best performance in relation to sample-to-sample variability provided that the following conditions are satisfied by the response data used in its calculation: (1) The number of such data points is larger than the number of coefficients in the metamodel, and more importantly (2) The data variance is not close to zero.

Based on this preliminary study we are inclined to conclude that using both *R*-square and RAAE in metamodel predictability validation serves two purposes at the same time: (1) it minimizes the number of observations in test samples for robustness against sample-to-sample variations, and (2) it allows for a more solid ground for judging the acceptability of the metamodel. To clarify this latter point, recall the results obtained for Metamodel 2 in Example 2 above, with *R*-square values in Figure 13a close to 0.82 and RAAE results are 4–5% as shown in Figure 13b. Using *R*-square results alone only permits a choice between either yes or no for acceptability of the metamodel. However, if in addition it is known that the prediction error is 4–5% on average, then this may constitute a better grounds for accepting (or rejecting) the metamodel.

Finally, it can be seen by referring to the summary in Figure 22 that minimum sample sizes for confinement within ±10% variability are impractically high for most cases. For instance, the average of minimum sample sizes for the 10 sampling trials is 4.6*q* for *R*-square results related to Metamodel 2 in Example 2, and as can be seen from the figure, this is the second best case among the 20 averages given. This result means that the required sample size is nearly twenty-eight observations for the two-dimensional problem considered in the example with six coefficients used for the corresponding metamodel. This then raises the question about the need to develop a "substitute" statistic which can be exclusively used in validating metamodels for deterministic simulations. This statistic should be chosen to reduce the minimum size of adequate validation samples. The Metamodel Acceptance Score (MAS) discussed in [10] is one such statistic. MAS sample-to-sample variability is reduced since MAS value depends on error "count" in a sample rather than error average.

5. Conclusions

This paper presented a comparative study of four statistics used in validating metamodels in deterministic simulations: RMSE along with its derivate *R*-square, and AAE and its derivate RAAE. RMSE and AAE are "purely" average-based, while their derivates are "less" dependant on averages through the cancellation of the sample size terms in the numerators and dominators of their defining equations. It is shown that the derivates are less prone to variations in sample sizes by comparison to their respective statistic, at least marginally as in the case of AAE and RAAE, provided that the derivate is used correctly in the first place, e.g., the variance of response data used for *R*-square calculations not approaching zero. In addition to being more robust against sample size variations, the derivate statistics provide better interpretability of results without reference to the context of the problem. Based on the results of this preliminary study, we are not hesitant to advocate the use of *R*-square and RAAE together in reporting metamodel validation results for deterministic simulations instead of RMSE or AAE. However, due to the excessive sample sizes required for ±10% sample-to-sample variations, we are even less hesitant to report the need to develop a substitute statistic which can be used exclusively for deterministic simulations.

References

1. Sargent R. 2004. Validation and verification of simulation models. In Proceedings of IEEE Winter Simulation Conference. IEEE, p. 13–24.
2. Kleijnen J, Deflandre D. 2006. Validation of regression metamodels in simulation: bootstrap approach. European Journal of Operational Research, 170, 120–131.
3. Hamad H, Al-Hamdan S. 2005. Two new subjective validation methods using data displays. In Proceedings of IEEE Winter Simulation Conference. IEEE, p. 2542–2545.
4. Hamad H, Al-Hamdan S. 2007. Discovering metamodels' quality-of-fit via graphical techniques. European Journal of Operational Research, 178, 543–559.
5. Martin J, Simpson T. 2005. Use of Kriging models to approximate deterministic computer models. American Institute of Aeronautics and Astronautics Journal, 43, 853–863.
6. Meckesheimer M, Booker A, Barton R, Simpson T. 2002. Computationally inexpensive metamodel assessment strategies. American Institute of Aeronautics and Astronautics Journal, 40, 2053–2056.
7. Jin R, Chen W, Simpson T. 2002. Comparative studies of metamodeling techniques under multiple modeling criteria. Journal of Structural Optimization, 23, 1–5.
8. Qu X, Venter G, Haftka R. 2004. New formulation of a minimum-bias central composite experimental design and Gauss quadrature. Structural and Multidisciplinary Optimization, 28, 231–242.
9. Eeckelaert T, Daems W, Gielen G, Sansen W. 2004. Generalized simulation-based posynomial model generation for analog integrated circuits. Analog Integrated Circuits and Signal Processing, 40, 193–203.
10. Hamad H. 2011. Validation of metamodels in simulation: a new metric. Engineering with Computer, 27, 309–317.
11. Hamad H, Al-Hamdan S, Al-Zaben A. 2010. Space partitioning in engineering design: a graphical approach. Structural and Multidisciplinary Optimization, 41, 441–452.

Cite this article as: Hamad H, Al-Zaben A & Owies R: Interpretability and variability of metamodel validation statistics in engineering system design optimization: a practical study. Int. J. Simul. Multisci. Des. Optim., 2014, **5**, A05.