AIAA JOURNAL
Vol. 47, No. 9, September 2009

Technical Notes

Cross Validation Can Estimate How Well Prediction Variance Correlates with Error

Felipe A. C. Viana* and Raphael T. Haftka†
University of Florida, Gainesville, Florida 32611

DOI: 10.2514/1.42162

Nomenclature

$\mathbf{e}$	=	error in prediction
$\mathbf{e}_{XV}$	=	cross-validation error
R	=	correlation coefficient between the absolute value of the errors and the square root of the prediction variance
R_{XV}	=	estimator for R based on cross validation
$\mathbf{s}^2$	=	prediction variance
$\mathbf{s}^2_{XV}$	=	prediction variance (during cross validation)

I. Introduction

THE use of surrogates for facilitating optimization and statistical analysis of computationally expensive simulations has become commonplace [1–4]. They offer easy-to-compute prediction and in some cases (such as kriging [5,6] and polynomial response surfaces [7,8]), they also furnish the prediction variance as a measure of uncertainty [9]. Figure 1a illustrates the concepts of prediction and prediction variance. Adaptive sampling and optimization methods use the prediction variance to select the next sampling point. For example, the Efficient Global Optimization (EGO) [10] and the Enhanced Sequential Optimization [11] algorithms use the kriging prediction variance to seek the point of maximum expected improvement as the next simulation for the optimization process. For such methods, it is important to assess the accuracy of the prediction variance; but presently, this is not available (although there is work on how to improve the uncertainty structure [12]). Cross validation is a standard tool for estimating the mean square errors (see the Appendix), thus the quality of the fit; and it can be used for selecting surrogates in a set [13–15]. Cross validation divides a set of p data points into k subsets. The surrogate is fit to all subsets except one, and the error is checked in the subset that was left out. This process is repeated for all subsets to produce a vector of cross-validation errors, $\mathbf{e}_{XV}$. Figure 1b illustrates cross validation when only one point is omitted.

We propose using cross validation for estimating the correlation between the prediction variance and the errors. Specifically we propose to use the correlation between the absolute values of the cross-validation errors and the square root of the prediction variance at the points that were left out.

Received 12 November 2008; accepted for publication 29 April 2009.

*Research Assistant, Department of Mechanical and Aerospace Engineering; fchegury@ufl.edu. Student Member AIAA.

†Distinguished Professor, Department of Mechanical and Aerospace Engineering; haftka@ufl.edu. Fellow AIAA.

II. Correlation Between Square Root of Prediction Variance and Absolute Errors

There are many possible measures of the quality of the prediction variance (e.g., based on test points we could compute the ratio between integrated square errors and integrated prediction variance, or the mean ratio between absolute errors and square root of prediction variance). However, here we focus on correlation as a possible indicator of usefulness. We calculate the correlation between the square root of the prediction variance, $\mathbf{s}$, and the absolute value of the error in prediction, $|\mathbf{e}|$[‡]:

$$R = r(\mathbf{s}, |\mathbf{e}|) \tag{1}$$

During the process of cross validation, at the points that were taken out, we compute both errors and prediction variance. Once repeated for all subsets, we have vectors of cross-validation errors $\mathbf{e}_{XV}$ and prediction variances $\mathbf{s}^2_{XV}$. Figure 1c illustrates the concepts of $\mathbf{e}_{XV}$ and $\mathbf{s}_{XV}$. In this work, we used the leave-one-out cross-validation for computing $\mathbf{e}_{XV}$ and $\mathbf{s}_{XV}$. This can be expensive for surrogates such as kriging. Congdon and Martin [16] presented a cheap-to-compute alternative to the computation of both $\mathbf{e}_{XV}$ and $\mathbf{s}^2_{XV}$ for kriging models (they discuss how these measures can be used to assess the quality of the fit). The correlation coefficient using cross-validation data is given by

$$R_{XV} = r(\mathbf{s}_{XV}, |\mathbf{e}_{XV}|) \tag{2}$$

We propose that 1) R_{XV} can be used as an estimator of R; and 2) R_{XV} can be used to rank different surrogates according to how well their prediction variance correlates with the errors in prediction.

III. Numerical Experiments

Table 1 gives details about the different basic surrogates used during the investigation. The DACE toolbox of Lophaven et al. [17] and the SURROGATES toolbox of Viana [18] were used to execute the kriging and polynomial response surface algorithms, respectively. The SURROGATES toolbox was also used for easy manipulation of the surrogates. We create different kriging surrogates by varying the regression models.

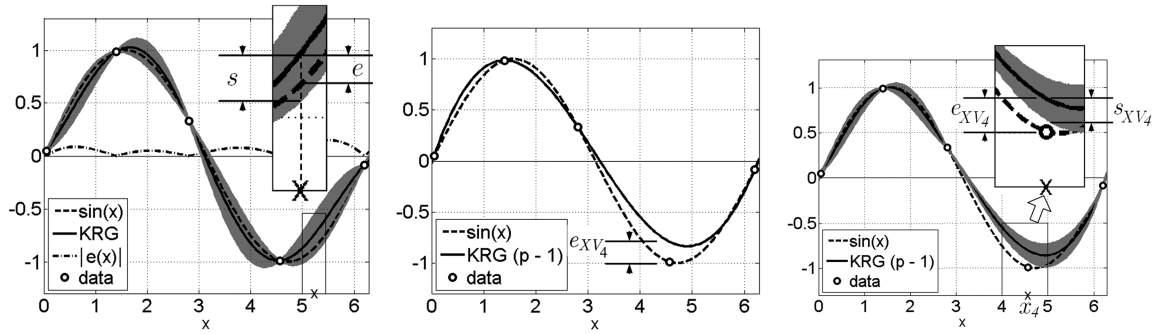
The quality of fit, and thus the performance, depends on the design of experiment (DOE). Hence, for all test problems, a set of 1000 different Latin hypercube designs [19] were used to average out the DOE dependence of the results. We used the MATLAB function *lhsdesign*, set with the *maxmin* option with 1000 iterations to generate the DOEs for fitting.

As test problems, we employed the following two analytical functions, widely used as benchmark problems in optimization [20]:

1) Branin-Hoo (two variables):

$$y(\mathbf{x}) = \left(x_2 - \frac{5.1x_1^2}{4\pi^2} + \frac{5x_1}{\pi} - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10, \qquad -5 \le x_1 \le 10, \qquad 0 \le x_2 \le 15 \tag{3}$$

[‡]We compute R using a set of 10,000 points created by the MATLAB function *lhsdesign* set with the *maxmin* option with 10 iterations. For these points we compute both the prediction variance and the actual errors.

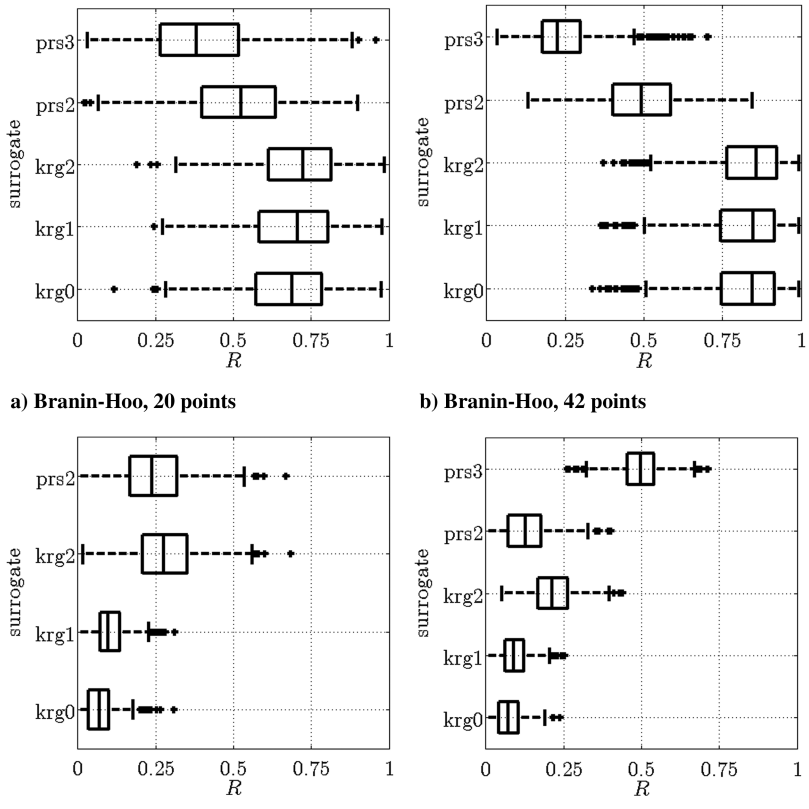


a) Test data: error and uncertainty

b) Cross-validation error

c) Cross-validation data: error and uncertainty

Fig. 1 Prediction and error measurements exemplified by fitting a kriging model (KRG) to $p = 5$ data points of the function $\sin(x)$. Gray areas illustrate the square root of the prediction variance. (a) Prediction, absolute value of the actual errors, $|e|$, and square root of prediction variance, s (details for a given point of the domain); (b) cross-validation error at the fourth point of the DOE, e_{XV_4}; and (c) cross-validation data: error, e_{XV_4}, and square root of the prediction variance, s_{XV_4} (details for the fourth point of the DOE).



a) Branin-Hoo, 20 points

b) Branin-Hoo, 42 points

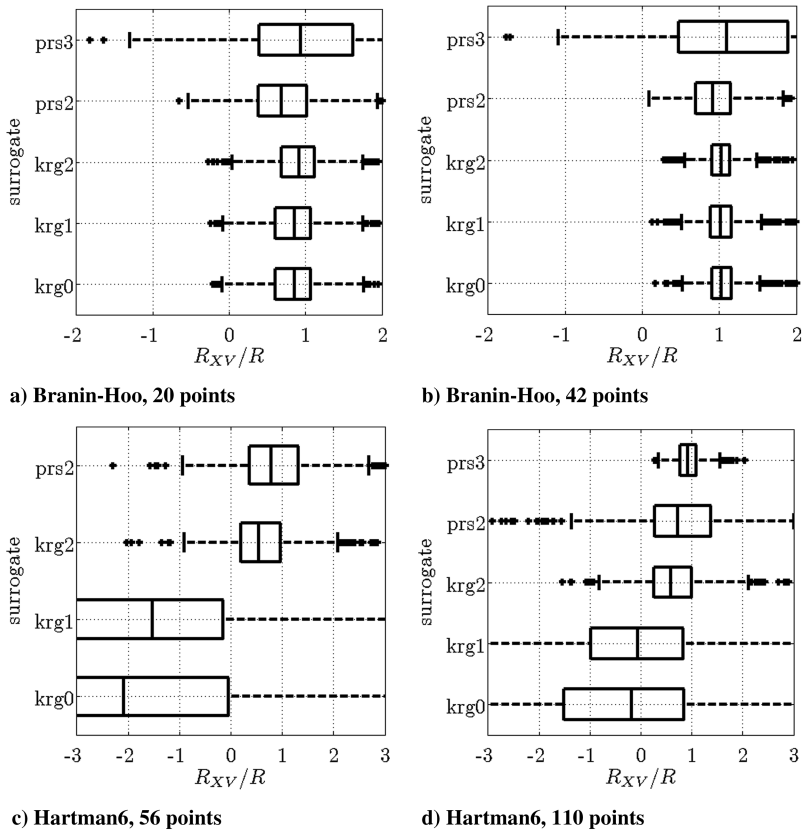
c) Hartman6, 56 points

d) Hartman6, 110 points

Fig. 2 Boxplots of R out of 1000 DOEs.

Table 1 Information about the set of surrogates

Surrogates		Details
1	KRG0	Kriging models.
2	KRG1	KRG0, KRG1, and KRG2 were set with zero, first-, and second-order polynomial trend models, respectively. All use the Gaussian correlation model. In all cases, $\theta_0 = (p^{-1/d}) \times 1_{d\times 1}$, and $10^{-3} \leq \theta_i \leq 2\theta_0$,
3	KRG2	$i = 1, 2, \ldots, d$ (d is the number of design variables), and we used the DACE toolbox for optimizing θ.
4	PRS2	
5	PRS3	Full second- and third-order (PRS2 and PRS3, respectively) polynomial response surface.



a) Branin-Hoo, 20 points **b) Branin-Hoo, 42 points**

c) Hartman6, 56 points **d) Hartman6, 110 points**

Fig. 3 Boxplots of the ratio between R_{XV} and R out of 1000 DOEs.

2) Hartman6 (six variables):

$$y(\mathbf{x}) = -\sum_{i=1}^{4} a_i \exp\left(-\sum_{j=1}^{6} b_{ij}(x_j - d_{ij})^2\right)$$

$$0 \le x_j \le 1, \qquad j = 1, 2, \ldots, 6, \qquad \mathrm{a} = [1.0 \quad 1.2 \quad 3.0 \quad 3.2]$$

$$\mathbf{B} = \begin{bmatrix} 10.0 & 3.0 & 17.0 & 3.5 & 1.7 & 8.0 \\ 0.05 & 10.0 & 17.0 & 0.1 & 8.0 & 14.0 \\ 3.0 & 3.5 & 1.7 & 10.0 & 17.0 & 8.0 \\ 17.0 & 8.0 & 0.05 & 10.0 & 0.1 & 14.0 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} 0.1312 & 0.1696 & 0.5569 & 0.0124 & 0.8283 & 0.5886 \\ 0.2329 & 0.4135 & 0.8307 & 0.3736 & 0.1004 & 0.9991 \\ 0.2348 & 0.1451 & 0.3522 & 0.2883 & 0.3047 & 0.6650 \\ 0.4047 & 0.8828 & 0.8732 & 0.5743 & 0.1091 & 0.0381 \end{bmatrix} \quad (4)$$

To investigate the effect of the point density, we fitted the Branin-Hoo function using both 20 and 42 points and the Hartaman6 function with 56 and 110 points.

IV. Results and Discussion

Figure 2 illustrates through boxplots how changing the DOEs impacts R of different surrogates. We see that the correlation can be substantially smaller than 1, meaning that the prediction variance does not always describe well variations of the actual errors. As expected, the sparseness of the DOE in the six-dimensional space negatively affects both the kriging and polynomial response surface prediction variances. In fact, R is mostly below 0.5. We can also see that the best surrogate in terms of the correlation depends on the problem. Apparently, with sparse data sets, the correlation improves with order of the trend. The Branin-Hoo function fit with 20 points corresponds to a relatively dense set of points (and then there is a marginal difference between the different kriging models). For the Hartman6 function, where data are sparser because of dimensionality, PRS2 is almost as good (and sometimes even better) in terms of the correlation as the best kriging surrogate. This may indicate that, in high dimensions, polynomial response surfaces may be useful with optimization algorithms that use the uncertainty structure.

Figure 3 illustrates how well R_{XV} estimates R. Figures 3c and 3d show that with sparse data sets, R_{XV} is only a rough estimate of R (even though it becomes better with more points).

When multiple surrogates are available, we investigate if R_{XV} can be used to detect which surrogate has the prediction variance that better correlates with the absolute errors. Figure 4 illustrates this idea with the scatter of R_{XV} and R in a single DOE of the Branin-Hoo

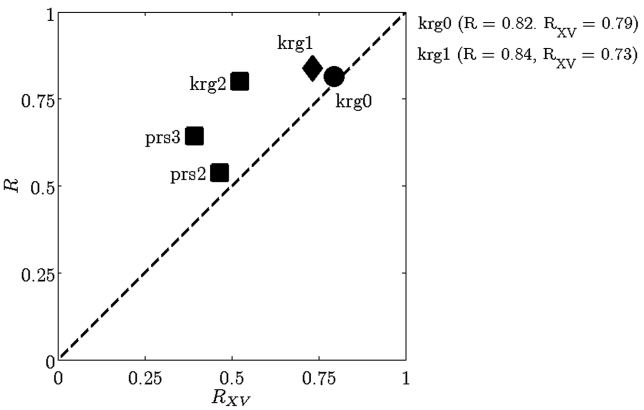


Fig. 4 Scatter plots of R_{XV} and R for an arbitrary DOE with 20 points of the Branin-Hoo function. The best surrogates in terms of R and R_{XV} are shown with diamond and circle markers, respectively.

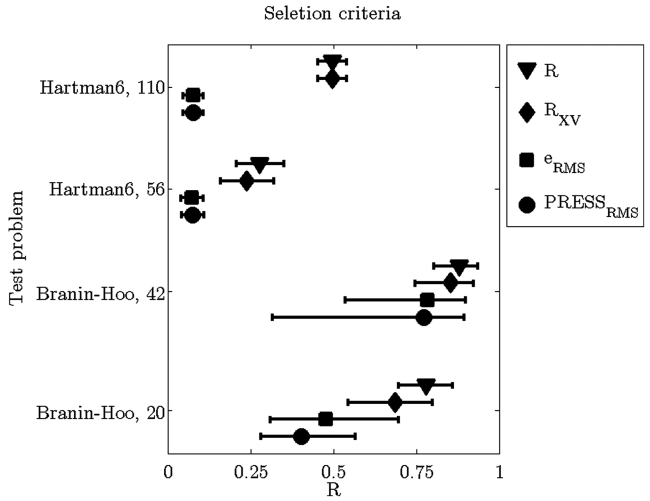


Fig. 5 [25 50 75] percentiles (over 1000 DOEs) of R of surrogates selected according to different criteria.

function fit with 20 points. We can see that the values of R_{XV} and R are not the same (and sometimes not even close, as in the case of krg2) and the surrogates with the best R_{XV} and R are not the same. Nevertheless, selection based on R_{XV} (krg0) and R (krg1) have comparable performance in terms of R (values are close, $R = 0.82$ for krg0 and $R = 0.84$ for krg1). If this happens for the set of studied problems, we consider that our approach succeeds in selection.

Figure 5 shows the [25 50 75] percentiles of the R values (out of 1000 DOEs) for surrogates picked according to different criteria (including both prediction and prediction variance). Unexpectedly, the surrogates chosen based on the prediction (surrogates with smaller values of $\mathbf{e}_{\text{RMS}}$ or $\text{PRESS}_{\text{RMS}}$) offer a rather poor performance in terms of the correlation between absolute errors and square root of the prediction variance. Figure 5 confirms our expectations showing that R_{XV} is more successful for selecting the surrogate with good correlation of error and prediction variance than both $\text{PRESS}_{\text{RMS}}$ and the $\mathbf{e}_{\text{RMS}}$.

V. Conclusions

We proposed using cross validation for estimating the correlation between the absolute value of the errors and the square root of the prediction variance. The approach was tested on two algebraic examples for kriging and polynomial response surface surrogates. For these examples we found that while we may obtain only a rough estimate of the correlation between their prediction variance and actual absolute errors, we succeeded in selecting surrogates with good correlation.

Surprisingly, 1) the statistically based prediction variance may not always correlate well to the errors; 2) the surrogate with the most accurate predictions did not necessarily have the best correlation; 3) with sparse data sets, the trend function influenced the quality of the correlation of kriging surrogates; and 4) the uncertainty structure of polynomial response surfaces was almost as good as (and sometimes better than) the best kriging surrogate.

Appendix: Root Mean Square Error (RMSE) and Prediction Sum of Squares (PRESS)

In this paper, when we check the accuracy of a surrogate, we compute the RMSE by Monte Carlo integration at a large number of p_{test} test points:

$$\text{RMSE} = \sqrt{\frac{1}{p_{\text{test}}}\sum_{i=1}^{p_{\text{test}}} e_i^2} \tag{A1}$$

where $e_i = y_i - \hat{y}_i$ is the error associated with the prediction, $\hat{y}_i$, compared to the actual simulation, y_i, in the ith test point.

For comparing surrogates based on the data only at the p points of the design of experiments, we use cross-validation errors, $\mathbf{e}_{XV}$. The RMSE is estimated from $\mathbf{e}_{XV}$:

$$\text{PRESS}_{\text{RMS}} = \sqrt{\frac{1}{p}\mathbf{e}_{XV}^T\mathbf{e}_{XV}} \tag{A2}$$

Acknowledgments

This work has been supported by the NASA Constellation University Institute Program (CUIP) and the National Science Foundation (Grant No. 0423280).

References

[1] Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P., "Design and Analysis of Computer Experiments," *Statistical Science*, Vol. 4, No. 4, 1989, pp. 409–435. doi:10.1214/ss/1177012413

[2] Toropov, V. V., "Simulation Approach to Structural Optimization," *Structural and Multidisciplinary Optimization*, Vol. 1, No. 1, 1989, pp. 37–46.

[3] Queipo, N. V., Haftka, R. T., Shyy, W., Goel, T., Vaidyanathan, R., and Tucker, P. K., "Surrogate-Based Analysis and Optimization," *Progress in Aerospace Sciences*, Vol. 41, No. 1, 2005, pp. 1–28. doi:10.1016/j.paerosci.2005.02.001

[4] Simpson, T. W., Toropov, V. V., Balabanov, V. O., and Viana, F. A. C., "Design and Analysis of Computer Experiments in Multidisciplinary Design Optimization: A Review of How Far We Have Come—or Not," *Proceedings of the 12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, AIAA, Reston, VA, 10–12 Sept. 2008.

[5] Stein, M. L., *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York, 1999.

[6] Simpson, T. W., Mauery, T. M., Korte, J. J., and Mistree, F., "Kriging Models for Global Approximation in Simulation-Based Multidisciplinary Design Optimization," *AIAA Journal*, Vol. 39, No. 12, 2001, pp. 2233–2241. doi:10.2514/2.1234

[7] Box, G. E. P., Hunter, W. G., and Hunter, J. S., *Statistics for Experimenters*, Wiley, New York, 1978.

[8] Myers, R. H., and Montgomery, D. C., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 2nd ed., Wiley, Hoboken, NJ, 2002.

[9] Apley, D. W., Liu, J., and Chen, W., "Understanding the Effects of Model Uncertainty in Robust Design with Computer Experiments," *Journal of Mechanical Design*, Vol. 128, No. 4, 2006, pp. 745–958.

[10] Jones, D., Schonlau, M., and Welch, W., "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, Vol. 13, No. 4, 1998, pp. 455–492. doi:10.1023/A:1008306431147

[11] Jin, R., Chen, W., and Sudjianto, A., "On Sequential Sampling for Global Metamodeling in Engineering Design," *ASME 2002 Design Engineering Technical Conferences and Computer and Information in Engineering Conference*, DETC-DAC34092, ASME, Fairfield, NJ, 2002.

[12] den Hertog, D., Kleijnen, J. P. C., and Siem, A. Y. D., "The Correct Kriging Variance Estimated by Bootstrapping," *Journal of the Operational Research Society*, Vol. 57, No. 4, 2006, pp. 400–409.

[13] Kohavi, R., "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers, Los Altos, CA, 1995, Vol. 2, No. 12, pp. 1137–1143.

[14] Meckesheimer, M., Booker, A. J., Barton, R. R., and Simpson, T. W., "Computationally Inexpensive Metamodel Assessment Strategies," *AIAA Journal*, Vol. 40, No. 10, 2002, pp. 2053–2060. doi:10.2514/2.1538

[15] Viana, F. A. C., Haftka, R. T., and Steffen, V., "Multiple Surrogates: How Cross-Validation Errors Can Help Us to Obtain the Best Predictor," *Structural and Multidisciplinary Optimization* (to be published).

[16] Congdon, C. D., and Martin, J. D., "On Using Standard Residuals as a Metric of Kriging Model Quality," AIAA 2007-1928, 23–26 April 2007.

[17] Lophaven, S. N., Nielsen, H. B., and Søndergaard, J., *DACE—A MATLAB Kriging Toolbox*, Informatics and Mathematical Modelling, Technical University of Denmark, TR IMM-TR-2002-12, 2002.

[18] Viana, F. A. C., *SURROGATES ToolBox User's Guide*, http://fchegury.googlepages.com, 2008.

[19] Mckay, M. D., Beckman, R. J., and Conover, W. J., "A Comparison of Three Methods for Selecting Values of Input Variables from a Computer Code," *Technometrics*, Vol. 21, No. 1, 1979, pp. 239–245. doi:10.2307/1268522

[20] Dixon, L. C. W., and Szegö, G. P., *Towards Global Optimization 2*, North-Holland, Amsterdam, The Netherlands, 1978.

A. Messac
Associate Editor