

Abstract to be submitted to the
8th - World Congress on Structural and Multidisciplinary Optimization
June 1-5, 2009, Lisbon, Portugal

**ASSESSING THE VALIDITY OF SURROGATE MODELS WITH
RESTRICTED INPUT SPACES**

Pineda, Fregly, Haftka, and Queipo

Surrogate models are often constructed using training data restricted to certain regions of the input space of interest; for example, as a result of correlation between input variables, physical limitations, or lack of information. When using these models in the context of analysis and optimization, the modeler then must decide whether or not: i) it is reasonable to use the surrogate model at a given prediction site, and/or ii) the surrogate model predictions are consistent with the training data outputs. The question of interest, which has received little attention, is how to address issues i) and ii) statistically.

In this work, issue i) is addressed using a statistical process control approach which establishes a likelihood (probability) measure of a given prediction site lying within the training data input space. The likelihood is estimated using the fact that the squared Mahalanobis distances of prediction sites to the center of the training sample (under normality conditions) behave as a Chi-square probability distribution of n degrees of freedom, with n being the number of variables. The Mahalanobis distance is a well-known measure which also accounts for the covariance structure of the available data. It is then possible to establish the probability of interest as the one of having a Mahalanobis distance equal to or greater than that associated with the prediction site. By establishing a threshold for this probability we can discard the locations for which the model is not expected to offer reliable predictions.

Issue ii) is tackled using order statistics with measures such as expected values and bounds for the extremes (maximum and minimum) estimated using the training data outputs. The expected values for the extremes cannot be arbitrarily large even if the range of the model output (with finite variance) is unbounded. Alternative approximations, including some based on sample (training data outputs) quantiles, are discussed. Under some parametric scenarios, that is, assuming a particular distribution for the training data outputs, values for the extremes are calculated using the method of moments (equating expected values and first sample moments). Surrogate model predictions outside these statistical bounds/extremes may not be reliable measures of the original model outputs and could be discarded or subject to further evaluation in the context of surrogate-based analysis and optimization efforts.

Using Kriging as a surrogate model, the effectiveness of the proposed approach was demonstrated with training and test data from, a) several analytical functions with different correlation structure for the training data and degrees of violations of the normality assumption (according to the Kolmogorov-Smirnov test), and b) the modeling of contact forces and torques in a biomechanical problem (i.e., knee implant).