Regular Paper

An Advanced Cost Estimation Methodology for Engineering Systems

C.G. Hart,[1,] * Z. He,[2] R. Sbragio,[2] and N. Vlahopoulos[3]

[1]*Naval Architecture and Marine Engineering Department, College of Engineering, University of Michigan, Ann Arbor,* MI 48105
[2]*Michigan Engineering Services,* LLC, *Ann Arbor,* MI 48105
[3]*Naval Architecture and Marine Engineering Department, Mechanical Engineering Department, College of Engineering, University of Michigan, Ann Arbor,* MI 48105

Received 5 January 2010; Revised 2 November 2010; Accepted 18 February 2011, after one or more revisions
Published online 15 November 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI 10.1002/sys.20192

ABSTRACT

A mathematically advanced method for improving the fidelity of cost estimation for an engineering system is presented. In this method historical cost records can be expanded either through the use of local metamodels or by using an engineering build-up model. In either case, the expanded data set is analyzed using principal component analysis (PCA) in order to identify the physical parameters, and the principal components (PCs) which demonstrate the highest correlation to the cost. A set of predictor variables, composed of the physical parameters and of the multipliers of the principal components which demonstrate the highest correlation to the cost, is developed. This new set of predictor variables is regressed, using the Kriging method, thus creating a cost estimation model with a high level of predictive capability and fidelity. The new methodology is used for analyzing a set of cost data available in the literature, and the new cost model is compared to results from a neural network based analysis and to a cost regression model. Further, a case study addressing the fabrication of a submarine pressure hull is developed in order to illustrate the new method. The results from the final regression model are presented and compared to results from other cost regression methods. The technical characteristics of the new novel general method are presented and discussed. © 2011 Wiley Periodicals, Inc. Syst Eng 15: 28–40, 2012

Key words: multidisciplinary optimization; ship design; naval architecture; complex systems

1. INTRODUCTION

The main objective of the work presented in this paper is the development of a mathematically advanced regression type of cost model which can associate physical parameters of a system with the system's cost. Principal component analysis is employed for identifying parameters and physical characteristics with strong correlation to the cost. The Kriging method is employed either in a conventional or in an adaptive mode for creating the cost regression model. The parameters and the physical characteristics with a strong correlation to the cost comprise the input parameters for the cost regression model. The new cost model can be used within a multidisci-

Contract grant sponsors: US Office of Naval Research Phase II SBIR Contract #N00014-08-C-0647 "Risk and Uncertainty Management for Multidisciplinary System Design and Optimization" (TPOC: Dr. Scott Hassan) and the National Defense Science and Engineering Graduate Fellowship

* Author to whom all correspondence should be addressed at: 3293 Taney Lane, Falls Church VA 22042 (e-mail: hartcg@umich.edu; chris.hart@ee.doe.gov).

© 2011 Wiley Periodicals, Inc.

plinary design environment [Hart and Vlahopoulos, 2010] for associating the values of the physical parameters with the cost. Thus, as the decision-making process progresses, the cost performance can be considered along with the other attributes of the design. In such applications the constraints imposed on the engineering performance characteristics will retain the parameters used as input to the cost regression model within the overall feasible region. Curran, Raghunathan, and Price [2004] summarize alternative engineering cost modeling methods applied to design and manufacture in aerospace engineering. The influence of cost in a concurrent engineering environment is addressed. The cost techniques presented in Curran, Raghunathan, and Price [2004] are divided in the classic estimating techniques (analogous costing, parametric technique, and bottom up technique), the advanced estimating techniques (feature-based technique, fuzzy logic, neural networks, uncertainty, and data mining), and the genetic causal cost modeling. Among the classic estimating techniques, the analogous costing methodology estimates the cost of a product, based on its differences from the previous similar products. Parametric techniques use regression equations between cost and independent variables (cost drivers) that are developed based on the historical data. Finally, the bottom-up approach is a very detailed estimate that requires substantial and detailed data based on a cost breakdown structure. Among the advanced estimating techniques, feature-based technique estimates costs are based on the design features of the product, such as number of holes, flanges, bends, and corner fillets, which are related to the production time. Fuzzy logic can be used to model systems that deal quantitatively with imprecision and uncertainty. Neural network techniques are based on the concept of a system that learns to estimate cost. After being trained with previous cost data, the neural network can compute the input data to predict the cost. The uncertainty method is based on statistic models of cost, such as Monte Carlo simulations, in which the cost estimation is linked to its probability of occurrence. The new method presented in this paper fits in the category of mathematically advanced estimating techniques that require some historical cost data to be available. A genetic causal cost modeling method is presented in Curran, Raghunathan, and Price [2004] as a manufacturing cost model linked to the structural analysis. It uses a breakdown of the overall cost into cost elements such as material cost, fabrication cost, and assembly cost. For each of these cost elements, the product is submitted to a breakdown in product families (for example, the cost of a panel is divided in the cost of riveting, stringers, skin, etc.). Each cost element is linked to the same design variables of the structural analysis through semiempirical equations. Other comprehensive reviews of cost modeling applied to the aerospace industry are presented in Meisl [1988a, 1988b]. Methods used to estimate cost in early program stages in space vehicle projects are presented, when insufficient knowledge of the parameters makes it necessary to employ previous cost experiences with new requirements. These methods use the cost breakdown structure of the product and compare its complexity with the complexity of previous projects in order to translate it into man-hours. A principal component based regression model is presented in Chan and Park [2005] and Williams [2003] for estimating the cost of construction projects. Fifty-seven variables related to the project, the construction team and the contractor were identified to have influence on the project cost. Data from 87 projects were used as sample data. The use of principal component analysis reduced the analysis to seven significant variables used to construct a regression model for the cost estimation. A cost modeling approach based on regression analysis is presented in Shtub and Versano [1999]. The model is applied for estimating the cost associated with a steel pipe bending process. It is concluded that the neural network approach leads to a smaller average square error and variance than other regression analysis models. The complete set of data presented in Shtub and Versano [1999] is used in this paper for demonstrating the validity and capabilities of the new cost estimation methodology presented here. In the work presented in this paper, the historical cost information may or may not be enhanced through the use of a local meta model or an engineering build-up model, depending if the conventional or if the adaptive Kriging method is used, respectively. The PCA is employed first for analyzing the data and identifying the important PC and the important physical characteristics of the system which present the highest correlation to the cost. A Kriging method is used either in the conventional or in the adaptive mode for developing the regression model. When the conventional mode is used, only the original historical cost data provide the information for constructing the cost model. When the adaptive mode is selected, then either local metamodels or an engineering build-up approach are used for enriching the original set of historical cost data. The important physical parameters and the participation factors of the important PC comprise the input variables to the regression model, while the cost comprises the attribute which is being evaluated. Figure 1 outlines the cost estimation approach presented in this paper. The steps in Figure 1 are discussed in depth in the next section of this paper. In the two examples presented in this paper the performance of the new cost estimation approach is compared to results from a neural network approach, a regression analysis, and PLS and CART cost predictions. Improvements are observed in the perform-

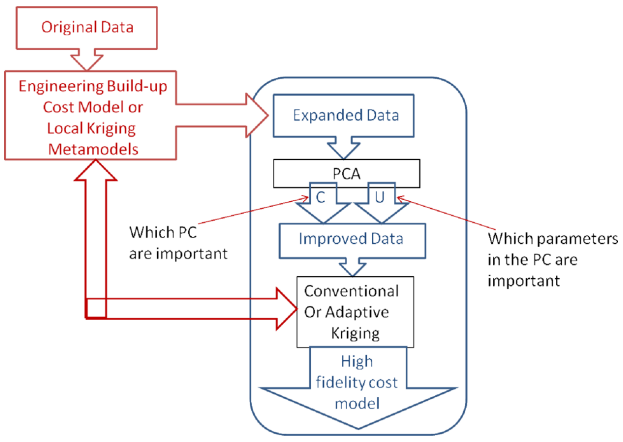

Figure 1. A graphical representation of the steps of the cost estimation method presented in this paper. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

ance of the new method for the two applications presented in this work.

Since the development of a cost model requires mathematical processing of the system's physical characteristics and cost data, the corresponding literature was reviewed, and the results are summarized next. A key element in cost estimation involves improving the accuracy of parametric models by analyzing historical cost data in a novel way. In this vein, the authors concentrate first on examining the literature on data mining, and the related knowledge discovery from databases (KDD) [Fayyad, Piatetsky-Shapiro, and Smyth, 1996]. Recently-published general overviews were identified as an introduction to the world of data mining and KDD [Wu et al., 2008; Han and Kamber, 2006; Srikant and Agrawal, 1996; Tan, Steinbach, and Kumar, 2005; Breiman, 1998]. The information contained in these papers provides an excellent understanding of the current state of the art and a broad foundation upon which to build. Of particular interest for potential adaptation and application in this problem were data mining methods broadly identified as classification [Wu et al., 2008] and association analysis [Tan, Steinbach, and Kumar, 2005] methods. Specifically, the classification and regression trees (CART) [Breiman, 1998] approach was very interesting in its broad application in the literature [Wu et al., 2008] as well as its fairly logical and straightforward theoretical base. Data mining methods have been employed in general business applications of data mining [Bingham, 2003; Apt, 2002], and in customer relationship management [Berson, Smith, and Thearling, 1999].

Principles of linear and matrix algebra are also used for analyzing data held in matrix form [Gifi, 1990]. In the past these methods have been employed by the authors in algorithms for developing time-dependent regression models [Sun et al., 2007; Sun, Vlahopoulos, and van Ast, 2007]. Methods such as singular value decomposition (SVD), data envelope analysis (DEA), and especially principal component analysis (PCA) have a wide spectrum of applications such as analyzing gene expression data in bioinformatics [Wall, Rechtsteiner, and Rocha, 2003; Ringnér, 2008], the effects of deregulation on the airline industry [Adler and Golany, 2001], and several other types of relationships, including those in finance [Jackson, 2005; Jolliffe, 2002].

Alternative methods for creating regression models are based on polynomials, interpolating and smoothing splines [Craven and Wahba, 1978], neural networks [Hajela and Burke, 1993; Rumelhart, Widrow, and Lehr, 1994; Cheng and Titterington, 1994; Ellacott, Mason, and Anderson, 1997], radial basis functions [Dyn, Levin, and Rippa, 1986], and wavelets [Jansen, Maifait, and Bultheel, 1996]. Cressie [1988] and Sacks et al. [1989] introduced the concept of the Kriging method. The Kriging method combines a regression model with Gaussian kernels for expressing the spatial correlation correction functions. Thus, it increases the accuracy compared to regression models due to presence of the spatial correlation correction term. A small sample of applications of the Kriging meta models are in variable fidelity optimization strategies [He and Vlahopoulos, 2009]; for managing system level uncertainty during conceptual design [He, Zhang, and Vlahopoulos, 2008]; and for approximating deterministic computer models [Martin and Simpson, 2005]. An adaptive feature has been added to the Kriging capability employed in this work. This adaptive feature will be described in the "Adaptive Kriging Method" section below. The mathematical method developed in this paper for cost modeling utilizes a PCA approach for analyzing the cost data and identifying the physical parameters and the principal components that demonstrate the highest correlation to the cost. The identified physical parameters and the participation factors of the important principal components comprise the predictor variables. An adaptive Kriging method is developed and employed in this paper for creating a high-fidelity metamodel for evaluating the response variables (cost parameters). The next section of this paper will cover the steps of the new mathematical methodology for cost modeling. This section will be followed by three sections which provide background and a theoretical foundation for the algorithms and methods used by the process. Cost data available from the literature that have been analyzed in the past by a neural network based cost model and by a regression analysis are analyzed with the new methodology, and the results are compared. Finally, a case study is presented associated with the fabrication of a submarine pressure hull, and cost prediction results are computed by the new method are compared with PLS and CART cost predictions. For the two examples analyzed in this paper, the new cost estimation method demonstrates better performance with respect to all other methods considered in the comparison. All methods which are compared require available historical cost data.

2. MISERLY: STEP-BY-STEP

The acronym MISERLY is short for Method of Improved Cost Estimation from Historical data of Engineering Systems. Figure 1 above provides a graphical representation of the steps taken by MISERLY in the creation of a mathematically advanced cost regression model. These steps are described in further detail in this section.

1. *Original Data:* The original data set is made up of physical parameters and historical cost records that are gathered for a given engineering system. Examples of an engineering system include, but are not limited to, ships, aircraft, spacecraft, advanced ground vehicles, submarines, offshore oil and gas platforms and other energy production facilities, and large land-based structures subjected to a diverse set of dynamic loads. Ideally, several copies and/or versions of this engineering system have already been constructed. The original data set consists of the X (physical parameter) and Y (cost parameter) matrices that should be arranged as follows:

$$X=\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}; \qquad Y=\begin{pmatrix} y_{11} & \cdots & y_{1q} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nq} \end{pmatrix}. \tag{1}$$

In these relationships, the X matrix consists of the p physical parameters of the n different designs upon

which data was gathered. Similarly, the Y matrix consists of the q historical cost parameters which were gathered for each of the n designs. Availability of historical cost data is necessary for the first step of the MISERLY process.

2. *Enhancement of Historical Cost Data:* If the Kriging method is used in the adaptive mode for creating the cost regression model, additional cost data points need to be created. Either a local metamodel (i.e., a Kriging model developed by a small number of historical data residing in the vicinity of the region where extra data points are needed), or an engineering build-up model is used for creating the additional data points requested by the adaptive Kriging method. The engineering build-up model must be formulated from a combination of sound engineering judgment and experience in the particular domain of the system in question. It also must be capable of recreating the values contained in the original data set. It is imperative that the utmost care and skill are employed in the creation of this cost model.
3. *Expanded Data:* An expanded data set is generated by combining the historical cost data with those created by the local metamodel or the engineering build up model. This step is only required if the Kriging method is used in the adaptive mode for creating the regression cost model.

$$X_e = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mp} \end{pmatrix}; \quad Y_e = \begin{pmatrix} y_{11} & \cdots & y_{1q} \\ \vdots & \ddots & \vdots \\ y_{m1} & \cdots & y_{mq} \end{pmatrix} \text{where } m >> n \tag{2}$$

Here X_e is the expanded matrix of physical parameters and Y_e is the matrix of expanded cost parameters. The physical parameters chosen to populate the expanded data set should represent a uniform distribution throughout the feasible design space. The expanded data set should include the original data set.

4. *PCA and Component-Influenced Parameters:* The expanded data set of physical parameters X_e is then analyzed using principal component analysis (PCA). If the initial set of data was not expanded, then $X_e = X$ and $Y_e = Y$ during the remaining of the process. Two outputs from the PCA, the C array and U matrix, are used to identify which of the original p physical parameters, and which of the principal components (PCs) that are output from PCA, account for the greatest amount of variation in the design. Each high-variation PC is composed of values that are treated as weights and are denoted by the variable w. These weights are used to create one component-influenced parameter, or a weighted sum of the original design variables, for each PC. Component-influenced parameters are denoted by the variable t. When the PCA step is finished, there will be r physical parameters and s PCs identified as high-variation. A more thorough discussion of the concept of component-influenced parameters is undertaken in the next section.
5. *Improved Data*: A set of predictor variables is developed using the high-variation physical parameters from the original set and $_s$ component-influenced parameters which were calculated using a weighted sum of the physical parameters in the original data set. Equation (3) shows the structure of this improved data set, X^*.

$$X^* = \begin{pmatrix} x_{11} & \cdots & x_{1r} & p_{11} & \cdots & p_{1s} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} & p_{n1} & \cdots & p_{ns} \end{pmatrix}. \tag{3}$$

6. *Regression using the Kriging method*: This new set of predictor variables is regressed, using the Kriging method (either in the conventional or in the adaptive mode) for developing a mathematical relationship between the predictor variables and the cost.

The next three sections discuss the main theoretical foundation upon which the MISERLY process was built.

3. PRINCIPAL COMPONENT ANALYSIS

PCA is a method for reducing the dimensionality of a set of data while still retaining most of the variation in the data set. This reduction in dimensionality allows a large, complex data set which relates thousands, or even tens of thousands, of variables, to be expressed by a new set of variables, usually much smaller in number. In most cases, these new variables are linear combinations of the original variables, and identify directions in the data along which variation is the maximum [Ringnér, 2008]. A specific relationship between these individual components is not defined in PCA. The mechanics of PCA are summarized in Jackson [2005], and are presented here for completeness.

There are several methods for performing principal component analysis, the most popular being dubbed R-, Q-, and N-analysis. The most general method of these is N-analysis, also known as singular value decomposition (SVD). N-analysis was chosen for this work due to this generality, as well as the method's simplicity. Although N-analysis requires familiarity with a new algorithm (SVD), it is a single-stage procedure that contrasts with the two-stage R- and Q-analyses.

The fundamental identity of SVD,

$$X = ZL^{1/2}U', \tag{4}$$

decomposes the centered and scaled $n \times p$ data matrix X into the $p \times p$ U and $L^{1/2}$ matrices and the $n \times p$ Z matrix. The U matrix represents the characteristic vectors of XX', Z represents the characteristic vectors of XX', and $L^{1/2}$ is a diagonal matrix that is function of their characteristic roots.

For this work, the matrices of interest from SVD are the U and $L^{1/2}$ matrices. The diagonal elements of the $L^{1/2}$ matrix were converted into an array and then modified in order to facilitate interpretation using the following relationship:

$$C_i = \left(\frac{L_i^{1/2}}{\sqrt{n-1}} \right)^2, \text{ where } i = 1, 2, \ldots, p. \tag{5}$$

By performing this modification, the C array, like its source $L^{1/2}$ matrix, contains a ranking of the amount of variance accounted for by each of the p PCs. The benefit of the modification is that the sum of the values in the C array now equals p. This is powerful because it relates each of the PCs to the physical parameters. If one of the PCs has a value in the C array that is greater than one, it accounts for greater variation than any of the design variables.

The U matrix is not modified in this work. The columns of the U matrix correspond to each PC while its rows correspond to each of the original physical parameters. Therefore, each PC contains weights for each of the original predictor variables. These weights can also be interpreted as an indication of the importance of each of these predictor variables in each of the PCs.

When the high-variability PCs are engaged in creating the component-influenced parameters, each of their values is treated as a weight for the corresponding physical parameter. The component-influenced parameters are then a sum of these weighted physical parameters:

$$t_{jk} = \begin{Bmatrix} w_{1k} \\ \vdots \\ w_{pk} \end{Bmatrix} \times \{ x_{j1} \cdots x_{jp} \}, \tag{6}$$

where $j = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, s$.

In this relationship, j refers to the number of designs in the original data set, and k to the number of principal components determined to be high variation PCs.

4. KRIGING REGRESSION MODELS

The Kriging model is based on treating $Z(\tilde{x})$, the difference between the actual performance variable $y(\tilde{x})$ and a regression model prediction $\hat{F}(\tilde{x})$, as a stochastic process:

$$y(\tilde{x}) = \hat{F}(\tilde{x}) + Z(\tilde{x}), \tag{7}$$

where $\tilde{x}$ is the d-dimensional vector of the variables that defines the point where the performance variable is evaluated, and d is the number of variables. A regression model which is a linear combination of m selected functions $f(\tilde{x})$ is used here:

$$\hat{F}(\tilde{x}) = \beta_1 f_1(\tilde{x}) + \cdots + \beta_m f_m(\tilde{x}) = f^T(\tilde{x}), \tag{8}$$

where $\beta^T = \{\beta_1, \beta_1, \ldots, \beta_m\}$ are regression parameters. $Z(\tilde{x})$ is considered as a normal process with zero mean and a covariance that can be expressed as

$$\text{cov}(Z(\tilde{x}_i), Z(\tilde{x}_j),) = \sigma^2 R(\tilde{x}_i, \tilde{x}_j), \tag{9}$$

where σ^2 is the process variance and $R(\tilde{x}_i, \tilde{x}_j)$ is the spatial correlation function. The equation used for the spatial correlation function is a Gaussian spatial correlation function:

$$R(\tilde{x}_i, \tilde{x}_j) = \prod_{k=1}^{d} \exp(-\theta_k (\tilde{x}_{i,k} - \tilde{x}_{j,k})^2), \tag{10}$$

and it indicates a process with infinitely differentiable paths in the mean square sense. θ_k is the correlation parameter that corresponds to the kth component of the d-dimensional vector of the random variables $\tilde{x}$, i.e., $k = 1, 2, \ldots, d$; and θ represents the vector of the θ_k parameters. For a set $\tilde{x}_s$ comprised of n number of sample points,

$$\tilde{x}_{si}$$

$$\tilde{x}_s^T = \{\tilde{x}_{s1}, \tilde{x}_{s1}, \ldots, \tilde{x}_{sn}\} \tag{11}$$

where $i = 1, 2, \ldots, n$.

The corresponding performance variable $\tilde{y}_s$ is considered known and its values are defined as

$$\tilde{y}_s^T = \{y(\tilde{x}_{s1}), y(\tilde{x}_{s2}), \ldots, y(\tilde{x}_{sn})\}. \tag{12}$$

The vector of correlations between the sample points $\tilde{x}_s$ and the evaluation point $\tilde{x}$ can be expressed as

$$\tilde{r}^T(\tilde{x}) = \{R(\tilde{x}, \tilde{x}_{s1}), R(\tilde{x}, \tilde{x}_{s2}), \ldots, R(\tilde{x}, \tilde{x}_{sn})\}. \tag{13}$$

The correlation matrix $[R]$ is also defined among all the sample points:

$$[R] = [R(\tilde{x}_{si}, \tilde{x}_{sj})]_{n \times n}. \tag{14}$$

The spatial correlation function in Eqs. (13) and (14) has been defined by Eq. (10). In the Kriging method the value of the performance function evaluated by the metamodel at the evaluation point $\tilde{x}$ is treated as a random variable. The computation of β and $Z(\tilde{x})$ in Eq. (7) is based on minimizing the mean square error (MSE) in the response

$$MSE[\hat{y}(\tilde{x})] = E[\hat{y}(\tilde{x}) - y(\tilde{x})]^2, \tag{15}$$

subject to the unbiased constraint

$$E[\hat{y}(\tilde{x})] = E[y(\tilde{x})]. \tag{16}$$

The matrix R and the parameters β and σ^2 depend on θ. Once θ is determined, the regression parameter β and the variance σ^2 can be computed as

$$\hat{\beta} = (\tilde{F}^T R^{-1} \tilde{F})^{-1} \tilde{F}^T R^{-1} \tilde{y}_s \tag{17}$$

$$\hat{\sigma}^2 = \frac{1}{n} (\tilde{y}_s - \tilde{F}\hat{\beta})^T R^{-1} (\tilde{y}_s - \tilde{F}\hat{\beta}), \tag{18}$$

where the matrix $\tilde{F}$ is defined as $\tilde{F} = [f_j(\tilde{x}_{si})]_{n \times m}$. The value for the response of interest is computed as

$$\hat{y}(\tilde{x}) = f(\tilde{x})^T \hat{\beta} + \tilde{r}^T(\tilde{x}) R^{-1} (\tilde{y}_s - \tilde{F}\hat{\beta}). \tag{19}$$

In the traditional Kriging method, the optimal value of θ is computed as the maximum likelihood estimator of the likelihood function:

$$L(\theta, \beta, \sigma^2|\tilde{y}_s) = p(\tilde{y}_s|\theta, \beta, \sigma^2),$$

$$p(\tilde{y}_s|\theta, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2(\det(R))}}\exp\left(-\frac{(\tilde{y}_s - \tilde{F}\hat{\beta})^T R^{-1}(\tilde{y}_s - \tilde{F}\hat{\beta})}{2\sigma^2}\right). \quad (20)$$

where $p(\tilde{y}_s|\theta, \beta, \sigma^2)$ is the multivariate normal distribution for the n observations $\tilde{y}_s$ given the model parameters θ, β, and σ^2. In the ordinary Kriging method this is accomplished by minimizing the product $[(\det(R))^{1/n}(\sigma^2)]$ while neglecting the variations in the model parameters θ, β, and σ^2.

5. ADAPTIVE KRIGING

This section presents an adaptive Kriging approach for metamodel development. The adaptive algorithm determines a specified number of adaptive sample points that are generated from an initial metamodel in regions of the variable domain defined according to the specified target result of the modeled function.

As mentioned above, a metamodel is a technique used to predict the response of a process and intends to reduce the number of expensive numerical simulations, and hence reduce the computational cost. In this work, it has been applied as a sophisticated interpolation function with the purpose of increasing the predictive capability of a parametric cost estimation model. The metamodel predictor of a process defined by a relation $Y(X)$, where $X = \{X_1, X_2, \ldots, X_m\}$ and $Y = \{Y_1(X), Y_2(X), \ldots, Y_n(X)\}$ is constructed based on experiments, which consist of a series of sample points X and their correspondent response Y.

The adaptive metamodel intends to generate more accurate responses Y of the modeled function $Y(X)$ when the responses Y are inside or in the vicinity of a desired target response T. The target T can be defined either as a single value or as a response region with minimum and maximum values, T_{min}, T_{max}, as shown in Figure 2. Hence, the adaptive metamodel generated in these regions with response inside or in the vicinity of T will be more accurate due to the presence of a larger number of SPs conveniently distributed than the metamodel created by sampling all SPs with random generators.

The adaptive metamodel is generated from a previous conventional metamodel. The criteria for determining the new adaptive SPs is to reduce the mean square error (MSE) of the metamodel for responses Y inside the target response region T. It is also desirable to consider regions in the vicinity of T, such that responses Y that are outside T due to errors in the initial metamodel evaluation can also be improved.

From the initial conventional metamodel, the mean square error (MSE) is estimated inside the domain of the input variable X. By defining the target region T, a weight function W is calculated, such that W is higher when the modeled function result approaches or is inside T. The new adaptive SPs are attracted to the target region based on the values of W

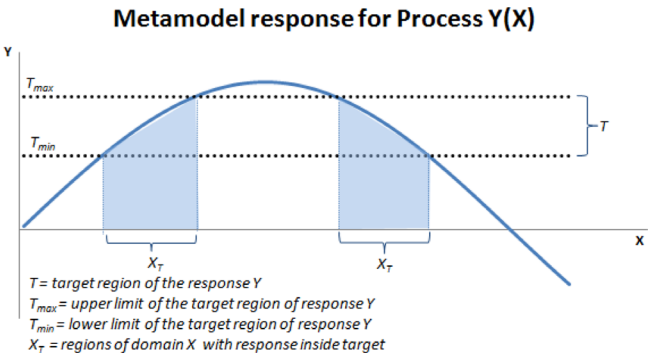

Figure 2. Metamodel response $Y(X)$ and target regions of response T. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$\times MSE$. Hence, if in the variable domain a position has a high value of MSE and is inside T (high value of W), it is a potential candidate to receive a sample point there. Regions with very small MSE inside T or regions with high MSE with results out of T have a smaller probability of attracting adaptive SPs. Regions outside and far from T (small value of W) and with small MSE will have virtually zero probability of receiving a SP.

The weight function W must be flexible in order to attract adaptive SPs to positions where the MSE is large enough to provoke an error in the estimated function value such that it can take this value out of the target region. Several types of weight functions can be used. These functions are similar to probability density functions, with a higher probability of sampling an SP inside the target region and a smaller probability outside it. A suitable weight function is based on an exponential decay and considers both the MSE and a user-defined decay parameter σ_ε^2 for selecting the size of the domain of interest. This function is shown in

$$W(X) = e^{\left(-\frac{(m_k(X) - T)^2}{2\sigma_\varepsilon^2 + s_K^2(X)}\right)}, \quad (21)$$

where

$W(X)$ = weight function at X coordinates,
X = coordinates of the variables (example: for a two dimensional metamodel: $X = x_1, x_2$),
T = target value,
$m_k(X)$ = metamodel prediction of $Y(X)$ at X coordinates,
σ_ε^2 = parameter that defines the size of the domain of interest (works like a variance, which increases or decreases the length of decay of the exponential),
$s_k^2(X)$ = MSE of the Kriging metamodel at coordinates.

If the response target value T is defined as a region delimitated by values $\{T_{min}, T_{max}\}$ as shown in Figure 2, Eq. (18) is used to determine the weight function in the following way:

$$W(X) = \begin{cases} 1 & \text{if } m_k(X) \text{ is inside } T, \\ e^{\left(-\frac{(m_k(X)-T)^2}{2\sigma_\varepsilon^2 + s_k^2(X)}\right)} & \text{if } m_k(X) \text{ is outside } T. \end{cases} \quad (22)$$

If the target value T is defined as a single value (not a region), Eq. (21) is used directly, presenting a maximum value of $W(X)$ equal to 1 if the metamodel prediction of function $Y(X)$, $m_k(X)$ is equal to the target T. The value of the weight function $W(X)$ decays exponentially as the value of $m_k(X)$ goes away from the target T.

The relative maxima of the product $W \times MSE$ define the coordinates X where the adaptive SPs must be positioned.

A summary of steps to be performed in the adaptive Kriging model follow.

a. Create an initial metamodel with a specified number of SPs.
b. Create the grid coordinates for determination of $W \times MSE$, according to the domain.
c. Determine the MSE (Mean Square Error) of the original metamodel at the grid coordinates.
d. Determine W at the grid coordinates, according to the target value, for the Adaptive Metamodel.
e. Calculate $W \times MSE$ at the grid coordinates.
f. The peak values of the product of W and MSE indicate the positions that are close or inside the target region and that have a large MSE. These positions are the candidates to receive a new adaptive SP.
g. Generate the new adaptive SPs .The regions where the peaks of the function $W \times MSE$ occurred are identified and the adaptive SPs are generated in these regions. In order to eliminate very small peaks, a filter is used, which considers only the positions where the values of $W \times MSE$ are larger than the average value of $W \times MSE$ calculated over the grid. If more adaptive sample points than peaks are needed, the remaining SPs are generated at positions defined according to a probability determined as

$$\text{Probability} = \frac{W\sqrt{MSE}}{\sum_{i=1}^{G} (W\sqrt{MSE})_i}, \text{ where } G = \text{size of the grid,} \quad (23)$$

which considers only the filtered regions. Positions with higher value of $W\sqrt{MSE}$ have more chance to receive an adaptive SP.
h. The results (new adaptive SPs) are passed to the solver.

6. COMPARATIVE STUDY

In order to demonstrate the application of the MISERLY process for creating a cost estimation model using an existing set of cost data, and for comparing MISERLY's performance with another advanced cost estimation approach based on neural networks, information from Shtub and Versano [1999] is used. This work presents the cost estimation of a steel pipe bending study case. The costs of 36 pipe bending configurations were estimated using neural networks and regression analysis and compared with their actual cost measurements (Table I; the information originates from Shtub and Versano [1999] and is presented here for completeness).

The input parameters of each pipe bending configuration are: the external diameter of the pipe (d_1), the internal diameter (d_2), the number of bends (K_f), the number of axes in space in which the pipe is bent (Z_r), and the distance between the bending location and the end of the pipe (L_g). The cost of bending is influenced by these parameters as follows: for pipes with external diameter smaller than 5 mm and difference between external and internal diameters less than 3 mm, a special treatment is required, as the pipes tend to deform during bending operations. The cost increases with the number of bends as additional time is required for each bend. The number of axes, which is always smaller or equal to the number of bends, affects the cost as it complicates the required operations. Finally, the distance between the bending location and the end of the pipe increases the number of special operations to bend the pipe when the pipe is chopped less than 100 mm from a bend. This last variable is expressed as the number of required special operations (can take value 0, 1, or 2).

The cost data shown in Table I are for the bending of hollow and chopped pipes. They were obtained from the measurements of at least twenty repetitions of the bending of each pipe configuration.

The estimated results are compared to the actual costs measured in experiments. The neural network process was employed in Shtub and Versano [1999] for creating a regression model for the cost. The following process was followed for training the neural network. From the n pipe configurations in the database, one is selected to be left out and the others, $n - 1$, are used to train the neural network. Then, the neural network is used to estimate the cost of the configuration that is left out. The square error between the estimated cost and the actual cost is determined. This process is repeated n times and the average square error and its standard deviation are determined and used as comparison parameters.

The same procedure is repeated using the MISERLY process in this paper. First the original data are expanded by developing local metamodels in regions identified by the adaptive method to contain limited data points. The local metamodels are created from the limited actual data residing in these regions. In this particular application, the PCA part of the MISERLY process identified the multipliers of three principal components as suitable predictor variables. In order to compare the performance of MISERLY to the one presented in Shtub and Versano [1999], for each original pipe bending configuration (or sample point) left out, a metamodel is created and the square error determined. After generating the 36 metamodels, the average square error and its standard deviation are computed. The results are then compared with the neural network results and also with the regression analysis results included in Shtub and Versano [1999].

The results obtained by the MISERLY process along with results from neural network analysis, and results from a regression cost model are summarized in Table II. The last two sets of results originate from Shtub and Versano [1999].

Table I. Cost Estimation of the Neural Network and Regression Models Compared to the Actual Cost (from Shtub and Versano [1999])

	Parameters					Actual cost	Forecasting		Squared errors	
	d_1	d_2	K_f	Z_r	L_g		Reg	Neu	ESTIM i(reg)	ESTIM i(neu)
1	16	13	5	3	1	239.5	149.3	157.0	8136	6802
2	16	13	3	2	1	106.2	99.3	124.0	47	318
3	18	16	5	4	2	408.0	312.4	419.4	9139	129
4	16	14	1	1	2	156.6	163.8	150.8	52	34
5	10	8	5	5	2	433.5	301.0	365.4	17571	4639
6	20	17	3	3	1	109.0	151.2	109.3	1777	0
7	18	15	1	1	2	157.2	151.0	140.9	38	265
8	11	9	3	3	1	101.6	137.1	117.8	1261	262
9	20	18	4	2	1	170.4	139.3	154.0	964	268
10	12.7	11.3	6	3	2	210.0	311.1	187.1	10207	525
11	11	9	2	1	2	148.6	163.5	163.3	222	217
12	30	28	4	1	2	289.4	231.7	304.5	3328	228
13	10	9	3	2	1	166.0	101.4	97.5	4169	4694
14	25	23	5	2	2	306.0	273.3	275.0	1070	962
15	11	8	4	3	1	127.2	132.3	156.5	26	858
16	14	12	2	1	2	153.8	172.6	159.7	353	34
17	18	16	6	4	2	425.0	326.2	358.7	9753	4400
18	28	26	2	2	1	118.5	143.7	96.0	637	507
19	11	8	5	2	2	162.5	232.6	170.3	4912	61
20	14	12	5	3	1	112.7	175.6	195.9	3954	6911
21	25.4	22.9	6	3	2	412.0	300.8	377.1	12377	1217
22	16	13	4	2	1	170.8	107.2	153.9	4051	288
23	11	9	4	2	2	200.1	223.8	175.4	563	609
24	16	14	2	1	1	88.0	67.7	120.9	410	1085
25	18	16	3	2	1	113.9	122.6	131.9	77	324
26	11	9	3	1	2	151.0	178.5	172.2	756	450
27	19.1	16.6	5	3	2	207.8	290.8	276.2	6888	4670
28	15	13	5	4	1	145.1	212.0	163.1	4474	326
29	22	20	2	1	2	204.0	193.0	144.4	120	3547
30	15	12	4	1	2	182.4	186.3	163.5	16	357
31	15	13	5	3	2	221.3	283.9	241.8	3922	421
32	12.7	11.3	4	2	2	199.3	240.3	175.6	1686	559
33	15	13	6	2	1	237.8	131.2	183.1	11366	2990
34	14	12	5	5	1	170.6	248.2	159.2	6021	129
35	12.7	10.9	2	1	1	84.0	60.7	93.9	544	97
36	10	9	2	1	1	76.1	65.1	71.0	122	26

ESSTIM i(reg) – The squared error between the actual cost and the regression estimation.
ESSTIM i(neu) – The squared error between the actual cost and the neural network estimation.

The results are expressed as the average square error of the 36 metamodels and the associated standard deviation. The cost results obtained by the MISERLY process offer an improvement compared to the neural network based cost model, and they are far more accurate from the conventional regression analysis results.

7. CASE STUDY ASSOCIATED WITH CONCEPTUAL SUBMARINE APPLICATION

A conceptual submarine manufacturing cost was chosen as a generic and representative case study since it highlights the nature of many manufacturing processes. This case study is

Table II. Comparison between the MISERLY Results and the Results Obtained by Neural Network and Regression Analysis [Shtub and Versano, 1999]

	MISERLY	Neural networks (from [7])	Regression analysis (from [7])
Average square error	1113.16	1366.97	3639.10
Standard Deviation	1944.63	1998.57	4403.64

adapted from previous work by the authors [Hart and Vlahopoulos, 2010]. Since no real cost data are available for this application, a commercially available engineering build-up cost software package was utilized to create the "historical" cost data. The same package was also used for creating extra data points in the regions requiring more data points when the Kriging method was used in the adaptive mode. The MISERLY process does require an initial set of historical data to be available. However, it also provides the flexibility to augment the historical data with information created by an engineering build-up model. In this case study the engineering build-up model is employed for creating even the "historical" data. The purpose of the case study is to demonstrate how the overall MISERLY process operates when a set of historical cost data is available, and how its cost assessment capability compares to the PLS and CART methods. Table III summarizes the data used in the development of the bottom-up cost model. The column on the left contains the parameters which are considered constant, while the column on the right contains the parameters which vary. In all discussions throughout this paper, the varying parameters are normalized between values of 0.5 and 1.5.

The work breakdown structure (WBS) is summarized in Figure 3. The WBS is used by SEER (a commercially available engineering build-up cost software package) for assessing the cost of fabricating a particular product. The version used in this work is SEER-DFM, taking its name from the popular industry concept of "design for manufacturing." SEER-DFM allows for the modeling of the manufacturing costs of a product, in this case, a submarine pressure hull. In order to use SEER, the following steps are necessary:

a. Develop the Work Breakdown Structure (WBS) of the product to be developed.
b. Define all the types of production operations that are needed.
c. Define the geometry of each component.
d. Gather data about the production operations.
e. Input the data in the code (can be accomplished remotely).

The costs for the components of the product are determined in the lower levels of the WBS and are basically divided into

a. Labor costs/unit—calculated according to the time needed to do the work and the hourly labor cost. Includes the setup costs for the machines needed to do the work.
b. Material costs/unit—calculated according to the material selected for the components.
c. Tooling costs/unit—calculated according to the machines and tools needed for the components.

Based on these costs, the SEER-DFM code determines the total cost/unit, using a bottom-up strategy, adding all the costs until the top level of the WBS is reached. In the absence of any historical cost data, an engineering build-up cost computational method is the only viable approach for creating a cost estimate.

As is illustrated in the WBS for the submarine pressure hull in Figure 3 of this paper, the first step in the process is to fabricate the hull itself. Industrial knowledge indicates that submarines are constructed in a series of modules or hoops, which are joined together. These hoops must then be formed and populated with decks and a limited outfit before they are joined together to form the pressure hull itself.

The process of breaking the hull down further into hoops is accomplished in two steps. First, the pressure hull is segmented using bulkheads. The remaining length of the hull is considered to be a uniform cylinder interrupted only periodically by kingposts for stiffness. These three sections are then split up further into actual hoops in the following manner.

In the case study manufacturing cost code, the dimensions of each hoop are determined automatically. The standard hoop width is set to four times the frame spacing. This hoop width is then used to divide the three lengths of the pressure hull into the number of hoops that will make up each of these sections. The result of this breakdown is a certain number of standard hoops, and one "leftover" hoop of some nonstandard width, for each section. The diameter of the pressure hull is then used to determine the length of the piece of material that will be formed into these hoops. The thickness of the plate is a varying parameter. Finally, since the hoop widths are not uniform along the length of the hull, an array of all possible hoop widths is generated.

Creating values that define the dimensions of the endcaps is not as involved as creating the dimensions of the hoops.

Table III. Definitions of Constant and Varying Parameters for Submarine Case Study

Definitions of Constant Parameters			Definitions of Varying Parameters			
Param.	Definition	Value	Variable	Definition	Max	Min
H_{td}	'tween deck height	2.286 m	L_{pmb}	Length of the parallel midbody	40 m	1 m
H_b	Bilge height	2.438 m	D	Maximum hull diam.	13 m	8.4 m
H_s	standard separation between hydrodynamic pressure hulls	0.607 m	n_a	Aft form factor	5	2
$H_{s,\min}$	minimum separation between hydrodynamic pressure hulls	0.25 m	n_f	Forward form factor	5	2
$n_{d,\max}$	Maximum number of decks	4	L_f	Frame spacing	0.75 m	1.5 m
n	Discretizations of hull	1000	t_p	Plate thickness	0.0127 m	0.0191 m

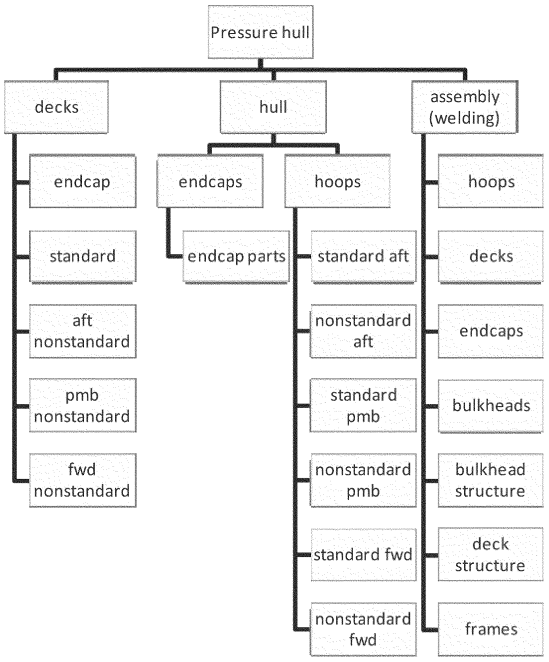

Figure 3. The WBS associated with the submarine case study.

Since each endcap is modeled as a hemisphere, the radii of these hemispheres serve as the main dimension necessary to define their construction. The fabrication process for the endcaps is assumed to include the dual-axis bending of four plates into the shape of ¼ of each hemispherical endcap. The dimensions of these plates are easily obtained from the geometry of a hemisphere, and therefore the input dimensions are also readily available. The thickness of the endcap was calculated in the structures discipline and is passed to the affordability discipline accordingly. Once the dimensions for the hoops and endcaps have been created and stored, the next step is to populate the matrix which contains the dimensions for the decks of the vessel. During this stage the bulkheads are also manufactured.

The dimensions of the material that makes up each section of the pressure hull are not the only inputs into the SEER-DFM program which must be entered. There are also several additional inputs, most of which are hardcoded into the case study manufacturing cost code, that provide specifics to the program concerning the manufacturing process being modeled. These additional inputs include such things as material used (low carbon steel), the material yield for each part (varies), the type of procedure being performed (plate roll bending on several), and specific information about the details of the procedure (the form diameter and number of passes for the plate roll bending procedure).

After every major piece of the pressure hull has been fabricated, it must then be assembled into the complete hull. In order to accomplish this task, the case study manufacturing cost code counts the total number of parts that must be assembled and measures the total length of the weld that must be run in order to complete the pressure hull. A general weight category is assigned to each part type, i.e., hoops, decks, endcaps, bulkheads, structure for each of them, etc., along with a general distance traveled.

All of this manufacturing data comprises the input to SEER-DFM package to take the data and process it as a command file input. Once SEER-DFM has completed its simulation, it provides per unit cost information for each level of the WBS, as well as a summary for each type of cost, i.e., labor, material, and tooling, and total cost for the entire manufacturing process.

For this particular case study, the manufacturing code is used to generate the "historical" total cost information and the total labor hours for a given set of varying parameters. These varying parameters and the "historical" cost information are combined to form the original data set. During the MISERLY process, SEER is also engaged in order to provide granularity in the regions identified by the adaptive Kriging method. The results of the case study are presented in a manner which matches closely the step-by-step discussion of how the MISERLY process operates.

7.1. Original Data

The MISERLY process does require an initial set of actual cost data to be available. The original data set for the case study consists of physical parameters and historical cost information for a hypothetical set of 50 designs. As has been discussed, it is assumed that this data set would originate from actual cost data in a practical application of this method. For this case study, the SEER-DFM model described in the previous section was used to generate the "historical" submarine cost data. The case study is intended to demonstrate how the MISERLY process operates when at least an initial set of historical cost data is available, and how the resulting cost predictions compare to results from a PLS and a CART approach.

7.2. Engineering Build-Up Model

The same SEER-DFM model that created the "historical" data set for this case study can be used as a proxy for the engineering build-up model. In a "real-world" application of this process, the engineering build-up model would have been created using the actual design data from the manufacturing of a class, or several classes, of submarines.

7.3. Expanded Data

The engineering build-up model creates the expanded data set. The varying physical parameters in the right column of Table III and the total cost and the total labor hours are used. The columns of the case study expanded data set are summarized in Table IV.

7.4. PCA and Component-Influenced Parameters

In this paper PCA is used for determining which physical parameters, and which PCs, account for the majority of the variation in the data. The C array and U matrix from these calculations are included in Tables V and VI.

Table IV. Data Variable Definition

Model variable	Description
x_1	Length of the parallel midbody
x_2	Maximum hull diam.
x_3	Aft form factor
x_4	Forward form factor
x_5	Frame spacing
x_6	Plate thickness
y_1	Total labor hours
y_2	Total cost

The *C* array and *U* matrix contain a significant amount of information regarding the data set in question. The *C* array indicates how much of the variation in the model is accounted for by each of the PCs. As was indicated earlier, the numbers in the matrix shown in Table IV have added value—their importance can be compared directly to the original predictor variables due to a modification procedure followed in this work. For this example, the 1st, 2nd, 3rd, and 4th PCs account for more variation than any single of the original predictor variables. This fact is noted, and the four PCs will be used in the creation of the component-influenced parameters that will make up part of the improved data set.

In addition to identifying which of the PCs account for more variability than any of the physical parameters, the raw numbers contained in the *C* array show not only that the first PC accounts for the largest amount of the variation, but also by looking at the change in the amount of variation accounted for by each PC, that there is a much larger drop in importance between the first and second PCs. In fact, the relative importance of the 2nd, 3rd, 4th, and 5th PCs are very close indeed, and they are much lower than the importance of the 1st PC. It isn't until the last PC that there is another significant drop in importance. This relationship between the PCs aids in the next step of the method—looking at the original predictor variables in an effort to select the most influential.

Like the *C* array, the *U* matrix also carries a tremendous amount of information. Most importantly for this application, it relates each of the principal components to the original predictor variables. For this particular example it was determined that the first PC accounts for a large amount of the variation of the data set and that the second, and subsequent, PCs account for a significantly lesser amount of the variation. With this in mind, the first PC will be used to determine the most influential of the original values. It is very obvious, when examining the absolute values in the first column of the *U* matrix, that the first two variables, the length of the parallel midbody and the diameter, have by far the largest impact on the data. In fact, their influence is so strong that the first PC can be assigned a general physical interpretation as the "Large Dimension Component."

In addition to their absolute values, there is significance in their opposite sign. After careful evaluation of the application of PCA to several instances, the authors determined that the absolute signs of the weights do not matter, as has been observed in other applications of PCA as well, but the relative signs of the weights do. The opposite sign of the two weights in a PC indicates an inverse relationship. In this particular instance, this inverse relationship has an important physical interpretation: There is a strong correlation between large values of the parallel midbody length and small values of the diameter and vice versa.

Table V. C Matrix for Conceptual Submarine Design Case Study

C array					
PC1	PC2	PC3	PC4	PC5	PC6
1.36	1.05	1.03	1.00	0.99	0.57

Table VI. U Matrix for Conceptual Submarine Design Case Study

	U matrix					
	PC1	PC2	PC3	PC4	PC5	PC6
L_{pmb}	0.69	-0.30	0.14	0.00	0.02	-0.64
D	-0.72	-0.19	0.09	0.05	-0.02	-0.66
n_a	0.05	0.88	0.19	-0.08	0.28	-0.31
n_f	0.06	0.23	-0.84	-0.03	-0.42	-0.24
L_f	-0.04	-0.05	0.11	-0.98	-0.18	0.00
t_p	-0.02	-0.19	-0.47	-0.19	0.84	-0.02

7.5. Improved Data

In this case study, the improved data set was composed of the first two original physical parameters, and multiplier for the first four PC. The matrix for the improved data is identical to that shown in Eq. (3) with $n = 50$, $r = 2$, and $s = 4$.

7.6. Creating the Cost Regression Model Using Kriging

At the final step of the MISERLY process a cost model was created from the 50 "historical" points using the Kriging method. Ten different configurations of the submarine which were not part of the "historical" data were used as evaluation points. In order to assess the MISERLY performance in this application, two other cost regression approaches, PLS and CART, were also used to create cost models model from the 50 "historical" points and they were also tested against the 10 evaluation points. Table VII summarizes these results. For this particular application the MISERLY performance provides a significant improvement compared to PLS and CART models. In the absence of any historical data, only an engineering build-up approach can be employed for cost estimation. However, this example demonstrates that when historical data are available, the MISERLY process offers increased accuracy compared to PLS and CART cost predictions.

In order to demonstrate the operation of adaptive mode of operation of the Kriging method for creating the cost model,

Table VII. Comparison of Regression Results (Avg Absolute % Error)

	PLS	CART	MISERLY
Labor hours	38.6	33.5	2.28
Total cost	303.4	520.7	4.21

25 additional sample points (i.e., different combinations of the varying parameters) were identified for creating additional cost data. The original set of the 50 "historical" points was expanded to include the additional 25 sample points, and a new cost metamodel was created. The inclusion of the additional 25 sample points reduced the mean error further from 4.1% to 3.6%.

8. Closure

A general and mathematically advanced method for cost estimation is proposed, and existing cost data from the literature are used for demonstrating its performance. Further, a case study addressing the fabrication of a submarine pressure hull is developed in order to illustrate the new method. Technical elements from PCA and the Kriging method for creating metamodels comprise the foundation of the MISERLY process. In the two applications presented in this paper, the MISERLY approach demonstrated better cost predictive capabilities compared to a neural network method, a regression analysis, and also with respect to PLS and CART models. The MISERLY approach does have similar limitations with all other cost modeling methods that require historical cost data to be available. Thus, MISERLY will not be useful in situations where completely new technologies or designs are considered. It does provide a mathematically advanced framework for conducting the cost estimation, and it does provide the flexibility to combine historical cost data with information created by engineering build-up cost models. In the future the new cost modeling capability can be used within a design decision making environment for assessing the implications of design changes to the cost. This information can be used in parallel with other performance assessments of an engineering system for guiding the overall decision making process.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the US Office of Naval Research Phase II SBIR Contract #N00014-08-C-0647 "Risk and Uncertainty Management for Multidisciplinary System Design and Optimization" (TPOC: Dr. Scott Hassan) and the National Defense Science and Engineering Graduate Fellowship for the support which made this research possible.

REFERENCES

N. Adler. and B. Golany, Evaluation of deregulated airline networks using data envelopment analysis combined with principal component analysis with an application to Western Europe, Eur J Oper Res 132(2) (2001), 260–273.

C. Apte et al., Business applications of data mining, Commun ACM 45(8) (2002), 49–53.

A. Berson, S. Smith, and K. Thearling, Building data mining applications for CRM, McGraw-Hill Professional, New York, 1999.

E. Bingham, Advances in independent component analysis with applications to data mining, Helsinki University of Technology, Helsinki, 2003.

L. Breiman, Classification and regression trees, Chapman & Hall/CRC, New York, 1998.

S.L. Chan and M. Park, Project cost estimation using principal component regression, Construction Management Econom 23(3) (March 2005), 295–304.

B. Cheng and D.M. Titterington, Neural networks: A review from a statistical perspective, Statist Sci 9(1) (1994), 2–54.

P. Craven and G. Wahba, Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the methods of generating cross-validation, Numer Math 31 (1978), 377–403.

N. Cressie, Spatial prediction and ordinary kriging, Math Geol 20(4) (1988), 405–421.

R. Curran, S. Raghunathan, and M. Price, Review of aerospace engineering cost modelling: The genetic causal approach, Prog Aerospace Sci 40 (2004), 487–534.

N. Dyn, D. Levin, and S. Rippa, Numerical procedures for surface fitting of scattered data by radial functions, SIAM J Sci Statist Comput 7(2) (1986), 639–659.

S.W. Ellacott, J.C. Mason, and I.J. Anderson, Mathematics of neural networks: Models, algorithms, and applications, Kluver Academic, Boston, MA, 1997.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, From data mining to knowledge discovery in databases, Commun ACM 39(11) (1996), 24–26.

A. Gifi, Nonlinear multivariate analysis, Wiley, New York, 1990.

P. Hajela and L. Berke, Neural networks instructural analysis and design: An overview, 4th AIAA/USAF/NASA/OAI Symp Multidisciplinary Anal Optim, Cleveland, OH, AIAA 2, 1993, AIAA-92-4805-CP, pp. 901–914.

J. Han, J. and M. Kamber, Data mining: concepts and techniques, Morgan Kaufmann, Elsevier, New York, 2006.

C.G. Hart and N. Vlahopoulos, A Multidisciplinary design optimization approach to relating affordability and performance in a conceptual submarine design, J Ship Prod (2010), to appear.

Z. He and N. Vlahopoulos, Utilization of response surface methodologies in the multi-discipline design optimization of an aircraft wing, 2009 SAE Cong, SAE Paper 2009-01-0344.

Z. He, G. Zhang, and N. Vlahopoulos, Uncertainty propagation in multi-disciplinary design optimization of undersea vehicles, 2008 SAE Congress, SAE Int J Mater Manuf 1(1) (2008), 70–79.

J. Jackson, A user's guide to principal components, Wiley-Interscience, Hoboken, NJ, 2005.

M. Jansen, M. Maifait, and A. Bultheel, "Generalized cross validation for wavelet thresholding," Signal processing, Elsevier, New York, 1996, No. 56, pp. 33–44.

I. Jolliffe, Principal component analysis, Springer, New York, 2002.

J.D. Martin and T.W. Simpson, Use of kriging models to approximate deterministic computer models, AIAA J 43(4) (April 2005), 853–863.

C.J. Meisl, Techniques for cost estimating in early program phases, Eng Costs Prod Econom 14 (1988a), 95–106.

C.J. Meisl, Missile propulsion cost modeling, Eng Cost Prod Econom 14 (1988b), 117–129.

M. Ringnér, What is principal component analysis? Nature Biotechnol 26(3) (2008), 303–304.

D.E. Rumelhart, B. Widrow, and M.A. Lehr, The basic ideas in neural networks, Commun ACM 37(3) (1994), 87–92.

J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn, Design and analysis of computer experiments, Statist Sci 4(4) (1989), 409–435.

A. Shtub and R. Versano, Estimating the cost of steel pipe bending, a comparison between neural networks and regression analysis, Int J Prod Econom 62 (1999), 201–207.

R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements, 5th Int Conf Extend Database Technol Adv Database Technol, Springer, New York, 1996.

J. Sun, N. Vlahopoulos, and K. Hu, Model update under uncertainty and error estimation in shock applications, 2005 SAE Noise and Vibration Conference, Traverse City, MI, 2005, SAE Paper No. 2005-01-2373.

J. Sun, J. He, N. Vlahopoulos, and P. van Ast, Model update and statistical correlations metrics for automotive crash simulations, 2007 SAE Cong, 2007, SAE Paper No. 2007-01-1744.

P. Tan, M. Steinbach, and V. Kumar, Introduction to data mining, Addison-Wesley Longman, Boston, MA, 2005.

M. Wall, A. Rechtsteiner, and L.M. Rocha, "Singular value decomposition and principal component analysis," A practical approach to microarray data analysis, Kluwer, Norwell, MA, 2003, pp. 91–109.

T.P. Williams, Predicting final cost for competitively bid construction projects using regression models, Int J Project Management 21 (2003), 593–599.

X. Wu et al., Top 10 algorithms in data mining, Knowledge Inform Syst 14(1) (2008), 1–37.

Chris Hart is currently the Offshore Wind Manager at the US Department of Energy. He graduated from the US Naval Academy with a degree in Naval Architecture and Marine Engineering and immediately accepted a commission as a Special Operations Officer in the US Navy. After 10 years of active duty, including combat tours in support of Operations Iraqi and Enduring Freedom, Chris began his graduate studies at the University of Michigan. In the ensuing 44 months, Chris earned a PhD and MSE in Naval Architecture and Marine Engineering and an MBA.

Zhijiang He received a BS and an MS in Engineering Mechanics from Tsinghua University, Beijing, China, in 2000 and 2002, respectively, and a PhD in Mechanical Engineering from the University of Michigan, Ann Arbor, in 2007. In October 2006, he joined Michigan Engineering Services, LLC, Ann Arbor, MI, as a Research and Development Engineer. He has published over 20 papers in journals and conferences. His research focuses on multidisciplinary optimization, energy finite element analysis, energy boundary element analysis, Kriging modeling, statistical reliability engineering analysis, reduced order modeling, aeroelasticity, and vibration and forced response analysis for turbo-machinery.

Ricardo Sbragio works currently as a Naval Engineer at CTMSP, a technological center of the Brazilian Navy, in the project and implantation of hydrodynamic test facilities. Before that he was a Research Scholar at the University of Michigan and a Research and Development Engineer at Michigan Engineering Services, LLC, working on projects related to acoustics and structural simulations. He also worked as Head of the Engineering Department at CTMSP in the supervision and execution of projects related to naval propellers, submarine conception, risk analyses, and nuclear propulsion. He holds a PhD in Naval Architecture and Marine Engineering and a Professional Degree in Nuclear Engineering, both from the University of Michigan.

Nick Vlahopoulos is a Professor at the University of Michigan. He joined the University of Michigan in 1996 after working in the Industry for seven years. He has graduated 15 PhD students, with another five PhD students and one research associate work currently under his supervision. He has published over 60 journal papers and 80 conference papers. The areas of his research are: numerical methods in structural acoustics; modeling of blast events; and performing multidiscipline design optimization for complex systems.