

An Investigation of Nonparametric **DATA MINING TECHNIQUES** for Acquisition Cost Estimating

 *Capt Gregory E. Brown, USAF, and Edward D. White*

The Department of Defense (DoD) cost estimating methodology traditionally focuses on parametric estimating using ordinary least squares (OLS) regression. Given the recent advances in acquisition data collection, however, senior leaders have expressed an interest in incorporating “data mining” and “more innovative analyses” within cost estimating. Thus, the goal of this research is to investigate nonparametric data mining techniques and their application to DoD cost estimating. Using a meta-analysis of 14 cost estimating studies containing 32 datasets that predominantly relate to commercial software development, the predictive accuracy of OLS regression is measured against three nonparametric data mining techniques. The meta-analysis results indicate that, on average, the nonparametric techniques outperform OLS regression for cost estimating. Follow-on data mining research that incorporates DoD-specific acquisition cost data is recommended to extend this article’s findings.


DOI: https://doi.org/10.22594/dau.16_756.24.02

Keywords: cost estimation, data mining, nonparametric, Cost Assessment Data Enterprise (CADE)



We find companies in industries as diverse as pharmaceutical research, retail, and insurance have embraced data mining to improve their decision support. As motivation, companies who self-identify into the top third of their industry for data-driven decision making—using ‘big data’ techniques such as data mining and analytics—are 6 percent more profitable and 5 percent more efficient than their industry peers, on average (McAfee & Brynjolfsson, 2012). It is therefore not surprising that 80 percent of surveyed chief executive officers identify data mining as strategically important to their business operations (PricewaterhouseCoopers, 2015).

We find that the Department of Defense (DoD) already recognizes the potential of data mining for improving decision support—43 percent of senior DoD leaders in cost estimating identify data mining as a most useful tool for analysis, ahead of other skillsets (Lamb, 2016). Given senior leadership’s interest in data mining, the DoD cost estimator might endeavor to gain a foothold on the subject. In particular, the cost estimator may desire to learn about *nonparametric* data mining, a class of more flexible regression techniques applicable to larger data sets.



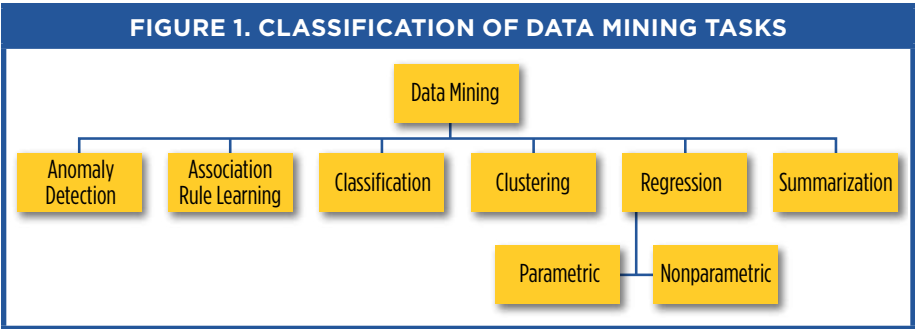
Initially, the estimator may first turn to DoD-provided resources before discovering that cost estimating coursework from the Defense Acquisition University (DAU) does not currently address nonparametric data mining techniques. Coursework instead focuses on parametric estimating using ordinary least squares (OLS) regression, while omitting nonparametric techniques (DAU, 2009). Subsequently, the cost estimator may turn to past research studies; however, this may prove burdensome if the studies occurred outside the DoD and are not easily found or grouped together. For this reason, we strive to provide a consolidation of cost-estimating research that implements nonparametric data mining. Using a technique known as meta-analysis, we investigate whether nonparametric techniques can outperform OLS regression for cost-estimating applications.

Our investigation is segmented into five sections. We begin with a general definition of data mining and explain how nonparametric data mining differs from the parametric method currently utilized by DoD cost estimators. Next, we provide an overview of the nonparametric data mining techniques of nearest neighbor, regression trees, and artificial neural networks. These techniques are chosen as they are represented most frequently in cost-estimating research. Following the nonparametric data mining overview, we provide a meta-analysis of cost estimating studies, which directly compares the performance of parametric and nonparametric data mining techniques. After the meta-analysis, we address the potential pitfalls to consider when utilizing nonparametric data mining techniques in acquisition cost estimates. Finally, we summarize and conclude our research.

Definition of Data Mining

So exactly what is *data mining*? At its core, data mining is a multidisciplinary field at the intersection of statistics, pattern recognition, machine learning, and database technology (Hand, 1998). When used to solve problems, data mining is a decision support methodology that identifies unknown and unexpected patterns of information (Friedman, 1997). Alternatively, the Government Accountability Office (GAO) defines data mining as the “application of database technologies and techniques—such as statistical analysis and modeling—to uncover hidden patterns and subtle relationships in data, and to infer rules that allow for the prediction of future results” (GAO, 2005, p. 4). We offer an even simpler explanation—data mining is a collection of techniques and tools for data analysis.

Data mining techniques are classified into six primary categories, as seen in Figure 1 (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). For cost estimating, we focus on regression, which uses existing values to estimate unknown values. Regression may be further divided into parametric and nonparametric techniques. The parametric technique most familiar to cost estimators is OLS regression, which makes many assumptions about the distribution function and normality of error terms. In comparison, the nearest neighbor, regression tree, and artificial neural network techniques are nonparametric. Nonparametric techniques make as few assumptions as possible, as the function shape is unknown. Simply put, nonparametric techniques do not require us to know (or assume) the shape of the relationship between a cost driver and cost. As a result, nonparametric techniques are regarded as more flexible.



Nonparametric data mining techniques do have a major drawback—to be effective, these more flexible techniques require larger data sets. Nonparametric techniques utilize more parameters than OLS regression, and as a result, more observations are necessary to accurately estimate the function (James, Witten, Hastie, & Tibshirani, 2013). Regrettably, the gathering of ‘more observations’ has historically been a challenge in DoD cost estimating—in the past, the GAO reported that the DoD lacked the data, both in volume and quality, needed to conduct effective cost estimates (GAO, 2006; GAO, 2010). However, this data shortfall is set to change. The office of Cost Assessment and Program Evaluation recently introduced the Cost Assessment Data Enterprise (CADE), an online repository intended to improve the sharing of cost, schedule, software, and technical data (Dopkeen, 2013). CADE will allow the cost estimator to avoid the “lengthy process of collecting, formatting, and normalizing data each time they estimate a program and move forward to more innovative analyses” (Watern, 2016, p. 25). As CADE matures and its available data sets grow larger, we assert that nonparametric data mining techniques will become increasingly applicable to DoD cost estimating.

Overview of Nonparametric Data Mining Techniques

New variations of data mining techniques are introduced frequently through free, open-source software, and it would be infeasible to explain them all within the confines of this article. For example, the software R—currently the fastest growing statistics software suite—provides over 8,000 unique packages for data analysis (Smith, 2015). For this reason, we focus solely on describing the three nonparametric regression techniques that comprise our meta-analysis: nearest neighbor, regression trees, and artificial neural networks. The overview for each data mining technique follows

a similar pattern. We begin by first introducing the most generic form of the technique and applicable equations. Next, we provide an example of the technique applied to a notional aircraft with unknown total program cost. The cost of the notional aircraft is to be estimated using aircraft data garnered from a 1987 RAND study, consolidated in Appendix A (Hess & Romanoff, 1987, pp. 11, 80). We deliberately select an outdated database to emphasize that our examples are notional and not necessarily optimal. Lastly, we introduce more advanced variants of the technique and document their usage within cost-estimating literature.

Analogous estimating via nearest neighbor, also known as case-based reasoning, emulates the way in which a human subject matter expert would identify an analogy

Nearest Neighbor

Analogous estimating via nearest neighbor, also known as case-based reasoning, emulates the way in which a human subject matter expert would identify an analogy (Dejaeger, Verbeke, Martens, & Baesens, 2012). Using known performance or system attributes, the nearest neighbor technique calculates the most similar historical observation to the one being estimated. Similarity is determined using a distance metric, with Euclidian distance being most common (James et al., 2013). Given two observations, p and q , and system attributes $1, \dots, n$, the Euclidean distance formula is:

$$Distance_{p,q} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} \quad (1)$$

To provide an example of the distance calculation, we present a subset of the RAND data in Table 1. We seek to estimate the acquisition cost for a notional fighter aircraft, labeled F-notional, by identifying one of three historical observations as the nearest analogy. We select the observation minimizing the distance metric for our two chosen system attributes: *Weight* and *Speed*. To ensure that both system attributes initially have the same weighting within the distance formula, attribute values are standardized to have a mean of 0 and a standard deviation of 1, as shown in italics.

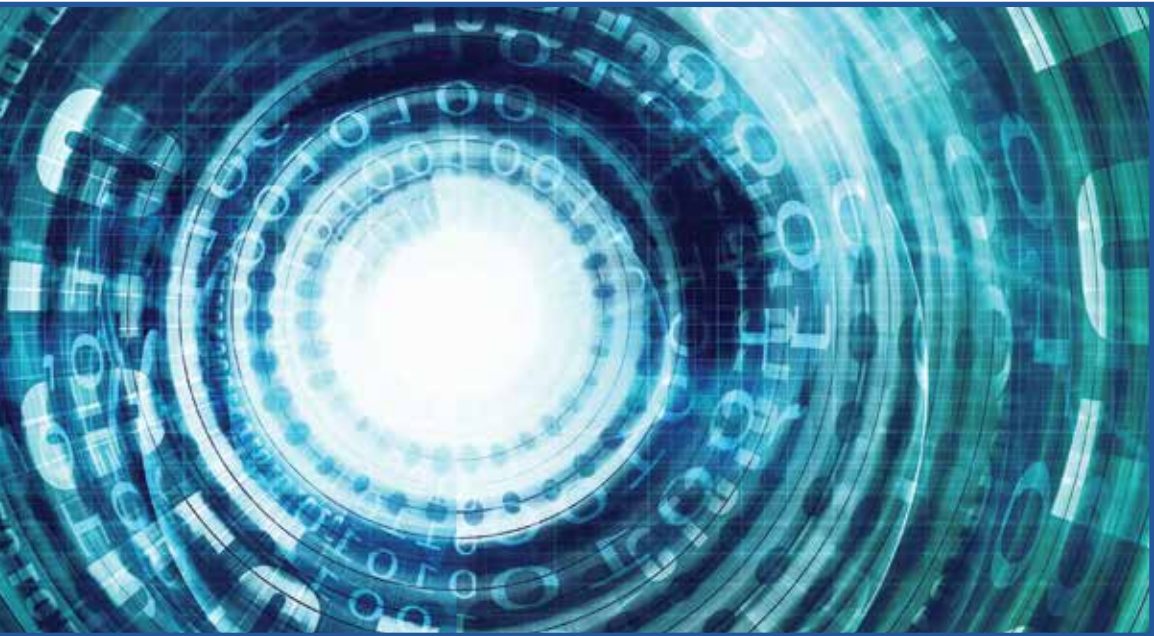
TABLE 1. SUBSET OF RAND AIRCRAFT DATA FOR EUCLIDIAN DISTANCE CALCULATION

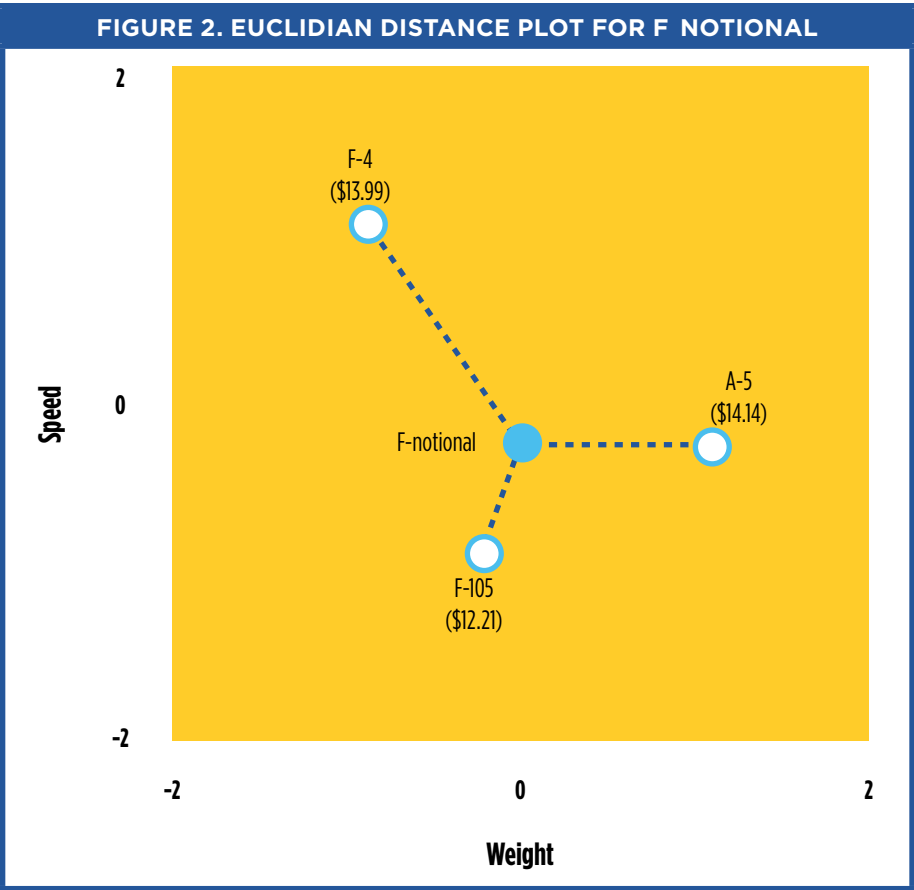
	Weight (Thousands of Pounds)		Speed (Knots)		Cost (Billions)
F-notional	20.00	0.00	1,150	-0.18	unknown
F-4	17.22	-0.87	1,222	1.10	13.99
F-105	19.30	-0.22	1,112	-0.86	12.21
A-5	23.50	1.09	1,147	-0.24	14.14

Using formula (1), the resulting distance metric between the F-notional and F-4 is:

$Distance_{F-notional,F-4} = \sqrt{([0.00 - (-0.87)]^2 + [-0.18 - (1.10)]^2)} = 1.54.$ (2)

The calculations are repeated for the F-105 and A-5, resulting in distance calculations of 0.71 and 1.10, respectively. As shown in Figure 2, the F-105 has the shortest distance to F-notional, and is identified as the nearest neighbor. Thus, the cost estimator would identify the unknown acquisition cost for the notional aircraft to be \$12.21 billion, analogous to the F-105.





Moving beyond our notional example, we find that more advanced analogy techniques are commonly applied in cost-estimating literature. When using nearest neighbor, the cost of multiple observations may be averaged when $k > 1$, with k signifying the number of analogous observations referenced. However, no k value is optimal for all data sets and situations. Finnie, Wittig, and Desharnais (1997) and Shepperd and Schofield (1997) apply $k = 3$, while Dejaeger et al. (2012) find $k = 2$ to be more predictive than $k = 1, 3$, or 5 in predicting software development cost.

Another advanced nearest neighbor technique involves the weighting of the system attributes, so that individual attributes have more or less influence on the distance metric. Shepperd and Schofield (1997) explore the attribute weighting technique to improve the accuracy of software cost estimates. Finally, we highlight clustering, a separate but related technique for estimating by analogy. Using Euclidian distance, clustering seeks to partition

a data set into analogous subgroups, whereby observations within a subgroup, or ‘cluster’, are most similar to each other (James et al., 2013). The partition is accomplished by selecting the clusters, minimizing the within cluster variation. In cost-estimating research, the clustering technique is successfully utilized by Kaluzny et al. (2011) to estimate shipbuilding cost.

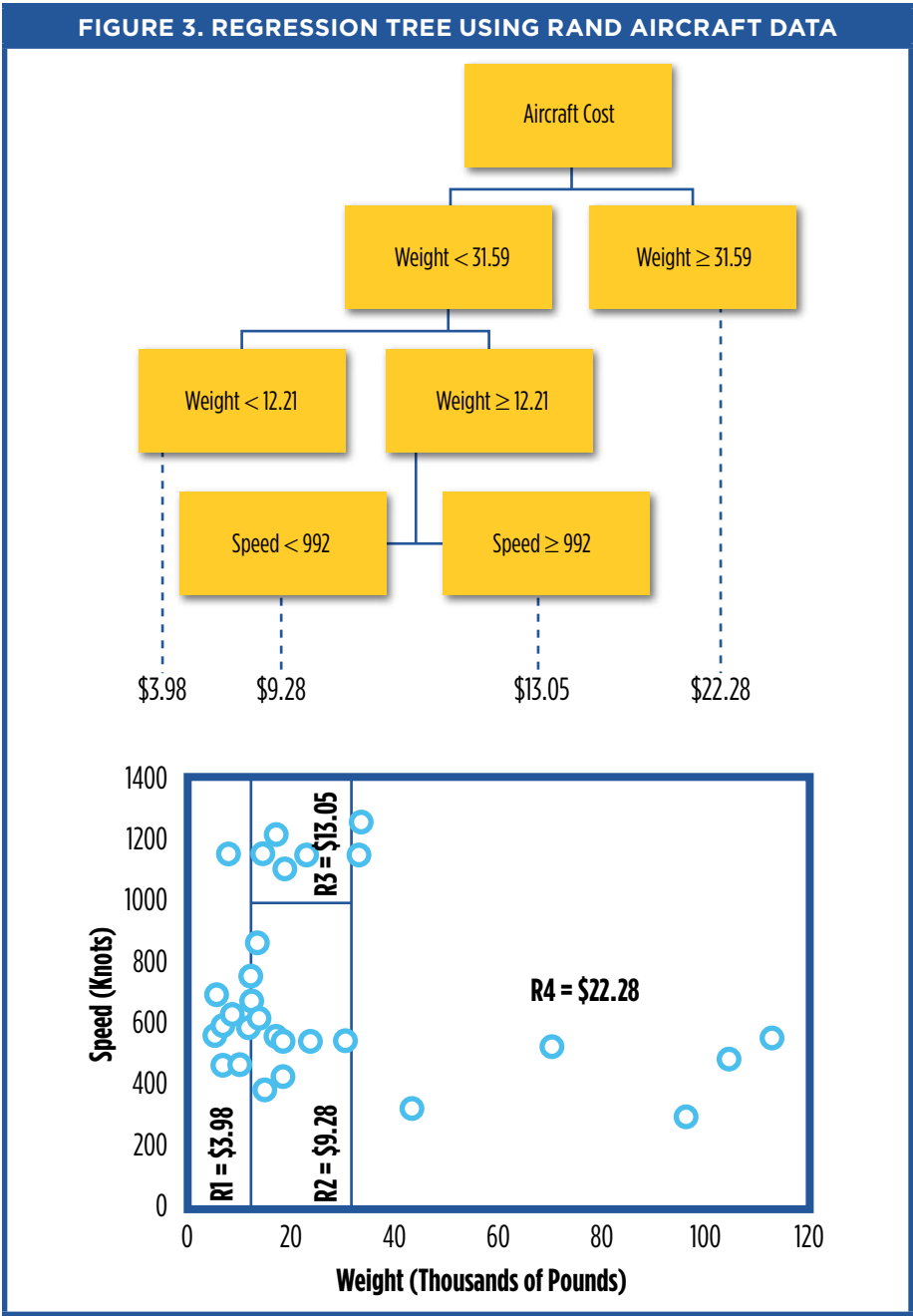
Regression Tree

The regression tree technique is an adaptation of the decision tree for continuous predictions, such as cost. Using a method known as recursive binary splitting, the regression tree splits observations into rectangular regions, with the predicted cost for each region equal to the mean cost for the contained observations. The splitting decision considers all possible values for each of the system attributes, and then chooses the system attribute and attribute ‘cutpoint’, which minimizes prediction error. The splitting process continues iteratively until a stopping criterion—such as maximum number of observations with a region—is reached (James et al., 2013). Mathematically, the recursive binary splitting decision is defined using a left node (L) and right node (R), and given as

$$\min \sum_{i \in L} (e_i - \bar{e}_L)^2 + \sum_{i \in R} (e_i - \bar{e}_R)^2 \quad (3)$$

where e_i = the i th observation’s *Cost*.

To provide an example of the regression tree, we reference the RAND dataset provided in Appendix A. Using the *rpart* package contained within the R software, we produce the tree structure shown in Figure 3. For simplicity, we limit the tree’s growth—the tree is limited to three decision nodes, splitting the historical observations into four regions. Adopting the example of the notional aircraft with a weight of 20 pounds (thousands) and a top speed of 1,150 knots, we interpret the regression tree by beginning at the top and following the decision nodes downward. We discover that the notional aircraft is classified into Region 3. As a result, the cost estimator would identify the unknown acquisition cost for the notional aircraft to be \$13.05 billion, equivalent to the mean cost of the observations within Region 3.



As an advantage, regression trees are simple for the decision maker to interpret, and many argue that they are more intuitive than OLS regression (Provost & Fawcett, 2013). However, regression trees are generally

outperformed by OLS regression, except for data that are highly nonlinear or defined by complex relationships (James et al., 2013). In an effort to improve the performance of regression trees, we find that cost-estimating researchers apply one of three advanced regression tree techniques: bagging, boosting, or piecewise linear regression.

‘Bagging’ involves application of the bootstrap method, whereby many regression trees are built on the data set, but each time using a different subset of the total data set. The predicted responses across all trees are then averaged to obtain the final response. Within cost-estimating research, the bagging technique is used by Braga, Oliveria, Ribeiro, and Meira (2007) to improve software cost-estimating accuracy. A related concept is ‘boosting’, for which multiple trees are also developed on the data. Rather than resampling the original data set, boosting works by developing each subsequent tree using only residuals from the prior tree model. For this reason, boosting is less likely to overfit the data when compared to bagging (James et al., 2013). Boosting is adopted by Shin (2015) to estimate building construction costs.

‘Bagging’ involves application of the bootstrap method, whereby many regression trees are built on the data set, but each time using a different subset of the total data set.

In contrast to bagging and boosting, the ‘M5’ technique—a type of piecewise linear regression—does not utilize bootstrapping or repeated iterations to improve model performance. Instead, the M5 fits a unique linear regression model to each terminal node within the regression tree, resulting in a hybrid tree/linear regression approach. A smoothing process is applied to adjust for discontinuities between the linear models at each node. Within cost research, the M5 technique is implemented by Kaluzny et al. (2011) to estimate shipbuilding cost, and by Dejaeger et al. (2012) to estimate software development cost.

Artificial Neural Network

The artificial neural network technique is a nonlinear model inspired by the mechanisms of the human brain (Hastie, Tibshirani, & Friedman, 2008). The most common artificial neural network model is the feed-forward multilayered perceptron, based upon an input layer, a hidden layer, and an output layer. The hidden layer typically utilizes a nonlinear logistic

sigmoid, transformed using the hyperbolic tangent function ('tanh' function), while the output layer is a linear function. Thus, an artificial neural network is simply a layering of nonlinear and linear functions (Bishop, 2006). Mathematically, the artificial neural network output is given as

$$o^u = f\left(\sum_j W_j V_j^u\right) = f\left[\sum_j W_j g_j\left(\sum_k w_{jk} I_k^u\right)\right], \quad (4)$$

where

I_k^u = inputs, normalized between -1 and 1

w_{jk} = connection weights between input and output layers

W_j = connection weights between hidden and output layer

V_j^u = output of the hidden neuron N_j , N_j = input element at the output neuron N

$g_j(h_j^u) = \tanh(\beta/2)$

h_j^u is a weighted sum implicitly defined in Equation (4).

For the neural network example, we again consider the RAND data set in Appendix A. Using the JMP® Pro software, we specify the neural network model seen in Figure 4, consisting of two inputs (*Weight* and *Speed*), three hidden nodes, and one output (*Cost*). To protect against overfitting, one-third of the observations are held back for validation testing, and the squared penalty applied. The resulting hidden nodes functions are defined as:

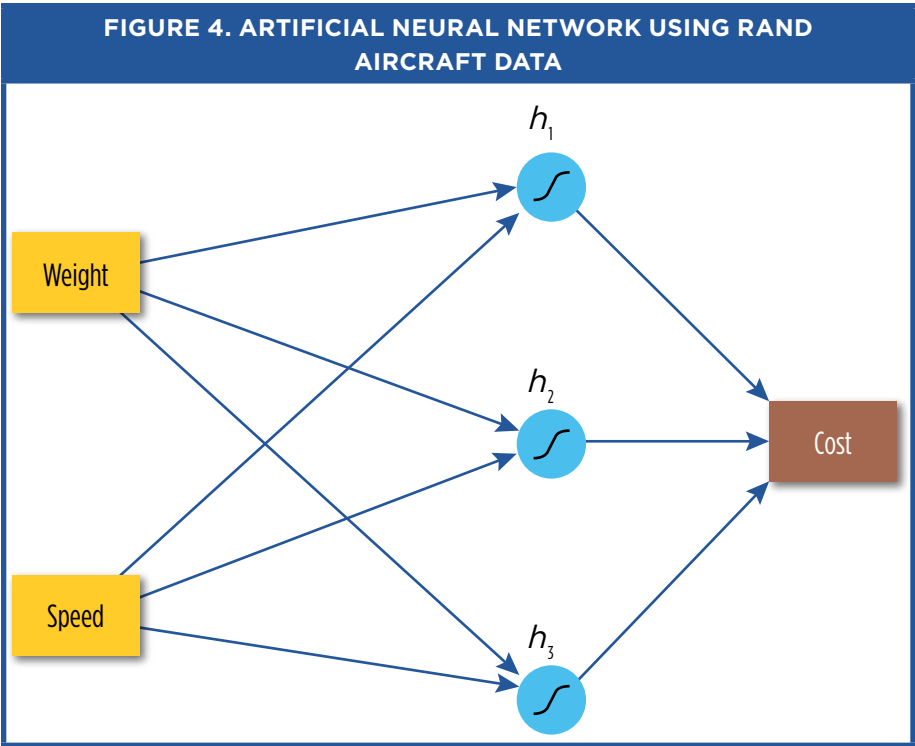
$$h_1 = \text{TanH}[(4.1281 - 0.0677 \times \text{Weight} + 0.0005 \times \text{Speed})/2] \quad (5)$$

$$h_2 = \text{TanH}[-(2.8327 + 0.0363 \times \text{Weight} + 0.0015 \times \text{Speed})/2] \quad (6)$$

$$h_3 = \text{TanH}[-(6.7572 + 0.0984 \times \text{Weight} + 0.0055 \times \text{Speed})/2] \quad (7)$$

The output function is given as

$$O = 14.8727 + 24.1235 \times h_1 + 71.2283 \times h_2 - 16.6950 \times h_3 \quad (8)$$



To calculate the cost of the notional aircraft with a weight of 20 pounds (thousands) and a top speed of 1,150 knots, the cost estimator would first compute the values for hidden nodes h_1 , h_2 , and h_3 , determined to be 0.9322, -0.1886, and 0.6457, respectively. Next, the hidden node values are applied to the output function, Equation (8), resulting in a value of 13.147. Thus, the cost estimator would identify the unknown acquisition cost for the notional aircraft to be \$13.15 billion.

In reviewing cost-estimating literature, we note that it appears the multilayer perceptron with a logistic sigmoid function is the most commonly applied neural network technique. Chiu and Huang (2007), Cirilovic, Vajdic, Mladenovic, and Queiroz (2014), Dejaneger et al. (2012), Finnie et al. (1997), Huang, Chiu, and Chen (2008), Kim, An, and Kang (2004), Park and Baek (2008), Shehab, Farooq, Sandhu, Nguyen, and Nasr (2010), and Zhang, Fuh, and Chan (1996) all utilize the logistic sigmoid function. However, we discover that other neural network techniques are used. To estimate software development cost, Heiat (2002) utilizes a Gaussian function, rather than a logistic sigmoid, within the hidden layer. Kumar, Ravi, Carr, and Kiran

(2008) and Dejaeger et al. (2012) test both the logistic sigmoid and Gaussian functions, finding that the logistic sigmoid is more accurate in predicting software development costs.

Meta-analysis of Nonparametric Data Mining Performance

Having defined three nonparametric data mining techniques common to cost estimating, we investigate which technique appears to be the most predictive for cost estimates. We adopt a method known as meta-analysis, which is common to research in the social science and medical fields. In contrast to the traditional literature review, meta-analysis adopts a quantitative approach to objectively review past study results. Meta-analysis avoids author biases such as selective inclusion of studies, subjective weighting of study importance, or misleading interpretation of study results (Wolf, 1986).

Data

To the best of our ability, we search for all cost-estimating research studies comparing the predictive accuracy of two or more data mining techniques. We do not discover any comparative data mining studies utilizing only DoD cost data, and thus we expand our search to include studies involving industry cost data. As shown in Appendix B, 14 unique research studies are identified, of which the majority focus on software cost estimating.

We observe that some research studies provide accuracy results for multiple data sets; in this case, each data set is treated as a separate research result for a total of 32 observations. When multiple variations of a given nonparametric technique are reported within a research study, we record the accuracy results from the best performing variation. After aggregating our data, we annotate that Canadian Financial, IBM DP Services, and other software data sets are reused across research studies, but with significantly different accuracy results. We therefore elect to treat each reuse of a data set as a unique research observation.

As a summary, 25 of 32 (78 percent) data sets relate to software development. We consider this a research limitation, and address it later. Of the remaining data sets, five focus on construction, one focuses on manufacturing, and one focuses on shipbuilding. The largest data set contains 1,160 observations and the smallest contains 19 observations. The mean data set contains 144.5 observations, while the median data set contains 65.5 observations.

Methodology

It is commonly the goal of meta-analysis to compute a ‘pooled’ average of a common statistical measure across studies or data sets (Rosenthal, 1984; Wolf, 1986). We discover this is not achievable in our analysis for two reasons. First, the studies we review are inconsistent in their usage of an accuracy measure. As an example, it would be inappropriate to pool a Mean Absolute Percent Error (MAPE) value with an R^2 (coefficient of determination) value. Second, not all studies compare OLS regression against all three nonparametric data mining techniques. Pooling the results of a research study reporting the accuracy metric for only two of the data mining techniques would potentially bias the pooled results. Thus, an alternative approach is needed.

We adopt a simple win-lose methodology, where the data mining techniques are competed ‘1-on-1’ for each data set. For data sets reporting error—such as MAPE or Mean Absolute Error Rate (MAER)—as an accuracy measure, we assume that the data mining technique with the smallest error value is optimal, and thus the winner. For data sets reporting R^2 , we assume that the data mining technique with the greatest R^2 value is optimal, and thus the winner. In all instances, we rely upon the reported accuracy of the validation data set, not the training data set. In a later section, we emphasize the necessity of using a validation data set to assess model accuracy.

Results

As summarized in Table 2 and shown in detail in Appendix C, nonparametric techniques provide more accurate cost estimates than OLS regression, on average, for the studies included in our meta-analysis. Given a ‘1-on-1’ comparison, nearest neighbor wins against OLS regression for 20 of 21 comparisons (95 percent), regression trees win against OLS regression for nine of 11 comparisons (82 percent), and artificial neural networks win against OLS regression for 19 of 20 comparisons (95 percent).

TABLE 2. SUMMARY OF META ANALYSIS WIN LOSS RESULTS												
	OLS	Nearest N.	OLS	Tree	OLS	ANN	Nearest N.	Tree	Nearest N.	ANN	Tree	ANN
Wins-Losses	1-20	20-1	2-9	9-2	1-19	19-1	8-6	6-8	10-5	5-10	9-5	5-9
Win %	5%	95%	18%	82%	5%	95%	57%	43%	67%	33%	64%	36%

We also report the performance of the nonparametric techniques in relation to each other. It appears that the nearest neighbor technique is the most dominant nonparametric technique. However, for reasons explained in our limitations, we assert that these results are not conclusive. For the practitioner applying these techniques, multiple data mining techniques should be considered, as no individual technique is guaranteed to be the best tool for a given cost estimate. The decision of which technique is most appropriate should be based on each technique's predictive performance as well as consideration of potential pitfalls, to be discussed later.

Limitations and Follow-on Research

We find two major limitations to the meta-analysis result. As the first major limitation, 78 percent of our observed data sets originate from software development. If the software development data sets are withheld, we do not have enough data remaining to ascertain the best performing nonparametric technique for nonsoftware applications.

As a second major limitation, we observe several factors that may contribute to OLS regression's poor meta-analysis performance. First, the authors cited in our meta-analysis employ an automated process known as stepwise regression to build their OLS regression models. Stepwise regression has been shown to underperform in the presence of correlated variables and allows for the entry of noise variables (Derksen & Keselman, 1992). Second, the authors did not consider interactions between predictor variables, which indicates that moderator effects could not be modeled. Third, with the exception of Dejaeger et al. (2012), Finnie et al. (1997), and Heiat (2002), the authors did not allow for mathematical transformations of OLS regression variables, meaning the regression models were incapable of modeling nonlinear relationships. This is a notable oversight, as Dejaenger et al. (2012) find that OLS regression with a logarithmic transformation of both the input and output variables can outperform nonparametric techniques.

Given the limitations of our meta-analysis, we suggest that follow-on research would be beneficial to the acquisition community. Foremost, research is needed that explores the accuracy of nonparametric techniques for estimating the cost of nonsoftware, DoD-specific applications such as aircraft, ground vehicles, and space systems. To be most effective, the research should compare nonparametric data mining performance against the accuracy of a previously established OLS regression cost model, which considers both interactions and transformations.

Potential Data Mining Pitfalls

Given the comparative success of nonparametric data mining techniques within our meta-analysis, is it feasible that these techniques be adopted by the program office-level cost estimator? We assert that nonparametric data mining is within the grasp of the experienced cost estimator, but several potential pitfalls must be considered. These pitfalls may also serve as a discriminator in selecting the optimal data mining technique for a given cost estimate.



Interpretability to Decision Makers

When selecting the optimal data mining technique for analysis, there is generally a trade-off between interpretability and flexibility (James et al., 2013, p. 394). As an example, the simple linear regression model has low flexibility in that it can only model a linear relationship between a single program attribute and cost. On the other hand, the simple linear regression

offers high interpretability, as decision makers are able to easily interpret the significance of a single linear relationship (e.g., as aircraft weight increases, cost increases as a linear function of weight).

As more flexible data mining techniques are applied, such as bagging, boosting, or artificial neural networks, it becomes increasingly difficult to explain the results to the decision maker. Cost estimators applying such data mining techniques risk having their model become a ‘black box’, where the calculations are neither seen nor understood by the decision maker. Although the model outputs may be accurate, the decision maker may have less confidence in a technique that cannot be understood.

Risk of Overfitting

More flexible, nonlinear techniques have another undesirable effect—they can more easily lead to overfitting. Overfitting means that a model is overly influenced by the error, or noise, within a data set. The model may be capturing the patterns caused by random chance, rather than the fundamental relationship between the performance attribute and cost (James et al., 2013). When this occurs, the model may perform well for the training data set, but perform poorly when used to estimate a new program. Thus, when employing a data mining technique to build a cost-estimating model, it is advisable to separate the historical data set into training and validation sets, otherwise known as holdout sets. The training set is used to ‘train’ the model, while the validation data set is withheld to assess the predictive accuracy of the model developed. Alternatively, when the data set size is limited, it is recommended that the estimator utilize the cross-validation method to validate model performance (Provost & Fawcett, 2013).

Extrapolation

Two of the nonparametric techniques considered, nearest neighbor and regression trees, are incapable of estimating beyond the historical observation range. For these techniques, estimated cost is limited to the minimum or maximum cost of the historical observations. Therefore, the application of these techniques may be inappropriate for estimating new programs whose performance or program characteristics exceed the range for which we have historical data. In contrast, it is possible to extrapolate beyond the bounds of historical data using OLS regression. As a cautionary note, while it is *possible* to extrapolate using OLS regression, the cost estimator should be aware that statisticians consider extrapolation a dangerous practice (Newbold, Carlson, & Thorne, 2007). The estimator should generally avoid extrapolating, as it is unknown whether the cost estimating relationship retains the same slope outside of the known range (DAU, 2009).

Spurious Correlation

Lastly, we introduce a pitfall that is common across all data mining techniques. As our ability to quickly gather data improves, the cost estimator will naturally desire to test a greater number of predictor variables within a cost estimating model. As a result, the incidence of ‘spurious’, or coincidental, correlations will increase. Given a 95 percent confidence level, if the cost estimator considers 100 predictor variables for a cost model, it is expected that approximately five variables will appear statistically significant purely by chance. Thus, we are reminded that correlation does not imply causation. In accordance with training material from the Air Force Cost Analysis Agency (AFCAA), the most credible cost models remain those that are verified and validated by engineering theory (AFCAA, 2008).

Summary

As motivation for this article, Lamb (2016) reports that 43 percent of senior leaders in cost estimating believe that data mining is a most useful tool for analysis. Despite senior leadership endorsement, we find minimal acquisition research utilizing nonparametric data mining for cost estimates. A consolidation of relevant, non-DoD research is needed to encourage the implementation of data mining techniques in acquisition cost estimating.

A consolidation of relevant, non-DoD research is needed to encourage the implementation of data mining techniques in acquisition cost estimating.

We turn to academic research utilizing industry data, finding relevant cost estimating studies that use software, manufacturing, and construction data sets to compare data mining performance. Through a meta-analysis, it is revealed that nonparametric data mining techniques consistently outperform OLS regression for industry cost-estimating applications. The meta-analysis results indicate that nonparametric techniques should, at a minimum, be at least considered for the DoD acquisition cost estimates.

However, we recognize that our meta-analysis suffers from limitations. Follow-on data mining research, utilizing DoD-specific cost data, is strongly recommended. The follow-on research should compare nonparametric data mining techniques against an OLS regression model, which considers both

interactions and transformations. Furthermore, we are honest in recognizing that the application of nonparametric data mining is not without serious pitfalls, including decreased interpretability to decision makers and the risk of overfitting data.

Despite these limitations and pitfalls, we predict that nonparametric data mining will become increasingly relevant to cost estimating over time. The DoD acquisition community has recently introduced CADE, a new data collection initiative. Whereas the cost estimator historically faced the problem of having too little data—which was time-intensive to collect and inconsistently formatted—it is entirely possible that in the future we may have more data than we can effectively analyze. Thus, as future data sets grow larger and more complex, we assert that the flexibility offered by nonparametric data mining techniques will be critical to reaching senior leadership's vision for more innovative analyses.

References

- AFCAA. (2008). *Air Force cost analysis handbook*. Washington, DC: Author.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.
- Braga, P. L., Oliveira, A. L., Ribeiro, G. H., & Meira, S. R. (2007). Bagging predictors for estimation of software project effort. *Proceedings of the 2007 International Joint Conference on Neural Networks*, August 12-17, Orlando, FL. doi:10.1109/ijcnn.2007.4371196
- Chiu, N., & Huang, S. (2007). The adjusted analogy-based software effort estimation based on similarity distances. *Journal of Systems and Software*, 80(4), 628-640. doi:10.1016/j.jss.2006.06.006
- Cirilovic, J., Vajdic, N., Mladenovic, G., & Queiroz, C. (2014). Developing cost estimation models for road rehabilitation and reconstruction: Case study of projects in Europe and Central Asia. *Journal of Construction Engineering and Management*, 140(3), 1-8. doi:10.1061/(asce)co.1943-7862.0000817
- Defense Acquisition University. (2009). *BCF106: Fundamentals of cost analysis* [DAU Training Course]. Retrieved from <http://www.dau.mil/mobile/CourseDetails.aspx?id=482>
- Dejaeger, K., Verbeke, W., Martens, D., & Baesens, B. (2012). Data mining techniques for software effort estimation: A comparative study. *IEEE Transactions on Software Engineering*, 38(2), 375-397. doi:10.1109/tse.2011.55
- Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45(2), 265-282. doi:10.1111/j.2044-8317.1992.tb00992.x
- Dopkeen, B. R. (2013). *CADE vision for NDIA's program management systems committee*. Presentation to National Defense Industrial Association, Arlington, VA. Retrieved from http://dcarc.cape.osd.mil/Files/CSDRSR/CSDR_Focus_Group_Briefing20131204.pdf
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth P. (1996, Fall). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37-54.
- Finnie, G., Wittig, G., & Desharnais, J. (1997). A comparison of software effort estimation techniques: Using function points with neural networks, case-based reasoning and regression models. *Journal of Systems and Software*, 39(3), 281-289. doi:10.1016/s0164-1212(97)00055-1
- Friedman, J. (1997). Data mining and statistics: What's the connection? *Proceedings of the 29th Symposium on the Interface: Computing Science and Statistics*, May 14-17, Houston, TX.
- GAO. (2005). *Data mining: Federal efforts cover a wide range of uses* (Report No. GAO-05-866). Washington, DC: U.S. Government Printing Office.
- GAO. (2006). *DoD needs more reliable data to better estimate the cost and schedule of the Shchuch'ye facility* (Report No. GAO-06-692). Washington, DC: U.S. Government Printing Office.
- GAO. (2010). *DoD needs better information and guidance to more effectively manage and reduce operating and support costs of major weapon systems* (Report No. GAO-10-717). Washington, DC: U.S. Government Printing Office.
- Hand, D. (1998). Data mining: Statistics and more? *The American Statistician*, 52(2), 112-118. doi:10.2307/2685468

- Hastie, T., Tibshirani, R., & Friedman, J. H. (2008). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
- Heiat, A. (2002). Comparison of artificial neural network and regression models for estimating software development effort. *Information and Software Technology*, 44(15), 911-922. doi:10.1016/s0950-5849(02)00128-3
- Hess, R., & Romanoff, H. (1987) *Aircraft airframe cost estimating relationships: All mission types*. Retrieved from <http://www.rand.org/pubs/notes/N2283z1.html>
- Huang, S., Chiu, N., & Chen, L. (2008). Integration of the grey relational analysis with genetic algorithm for software effort estimation. *European Journal of Operational Research*, 188(3), 898-909. doi:10.1016/j.ejor.2007.07.002
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York, NY: Springer.
- Kaluzny, B. L., Barbici, S., Berg, G., Chiomento, R., Derpanis, D., Jonsson, U., Shaw, A., Smit, M., & Ramaroson, F. (2011). An application of data mining algorithms for shipbuilding cost estimation. *Journal of Cost Analysis and Parametrics*, 4(1), 2-30. doi:10.1080/1941658x.2011.585336
- Kim, G., An, S., & Kang, K. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Journal of Building and Environment*, 39(10), 1235-1242. doi:10.1016/j.buildenv.2004.02.013
- Kumar, K. V., Ravi, V., Carr, M., & Kiran, N. R. (2008). Software development cost estimation using wavelet neural networks. *Journal of Systems and Software*, 81(11), 1853-1867. doi:10.1016/j.jss.2007.12.793
- Lamb, T. W. (2016). *Cost analysis reform: Where do we go from here? A Delphi study of views of leading experts* (Master's thesis). Air Force Institute of Technology, Wright-Patterson Air Force Base, OH.
- McAfee, A., & Brynjolfsson, E. (2012). Big data—the management revolution. *Harvard Business Review*, 90(10), 61-67.
- Newbold, P., Carlson, W. L., & Thorne, B. (2007). *Statistics for business and economics*. Upper Saddle River, NJ: Pearson Prentice Hall.
- Park, H., & Baek, S. (2008). An empirical validation of a neural network model for software effort estimation. *Expert Systems with Applications*, 35(3), 929-937. doi:10.1016/j.eswa.2007.08.001
- PricewaterhouseCoopers LLC. (2015). *18th annual global CEO survey*. Retrieved from <http://download.pwc.com/gx/ceo-survey/assets/pdf/pwc-18th-annual-global-ceo-survey-jan-2015.pdf>
- Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. Sebastopol, CA: O'Reilly Media.
- Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage Publications.
- Shehab, T., Farooq, M., Sandhu, S., Nguyen, T., & Nasr, E. (2010). Cost estimating models for utility rehabilitation projects: Neural networks versus regression. *Journal of Pipeline Systems Engineering and Practice*, 1(3), 104-110. doi:10.1061/(asce)ps.1949-1204.0000058
- Shepperd, M., & Schofield, C. (1997). Estimating software project effort using analogies. *IEEE Transactions on Software Engineering*, 23(11), 736-743. doi:10.1109/32.637387
- Shin, Y. (2015). Application of boosting regression trees to preliminary cost estimation in building construction projects. *Computational Intelligence and Neuroscience*, 2015(1), 1-9. doi:10.1155/2015/149702

- Smith, D. (2015). *R is the fastest-growing language on StackOverflow*. Retrieved from <http://blog.revolutionanalytics.com/2015/12/r-is-the-fastest-growing-language-on-stackoverflow.html>
- Watern, K. (2016). Cost Assessment Data Enterprise (CADE). *Air Force Comptroller Magazine*, 49(1), 25.
- Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills, CA: Sage Publications.
- Zhang, Y., Fuh, J., & Chan, W. (1996). Feature-based cost estimation for packaging products using neural networks. *Computers in Industry*, 32(1), 95–113. doi:10.1016/s0166-3615(96)00059-0

Appendix A

RAND Aircraft Data Set

Model	Program Cost, Billions (Base Year 1977)	Airframe Weight, Thousands (Pounds)	Maximum Speed (Knots)
A-3	10.15	23.93	546
A-4	3.73	5.07	565
A-5	14.14	23.50	1,147
A-6	8.88	17.15	562
A-7	3.3	11.62	595
A-10	6.29	14.84	389
B-52	32.03	112.67	551
B-58	32.43	32.69	1,147
B/RB-66	12.93	30.50	548
C-130	11.75	43.45	326
C-133	18.35	96.31	304
KC-135	15.55	70.25	527
C-141	18.91	104.32	491
F3D	3.03	10.14	470
F3H	7.57	13.90	622
F4D	7.1	8.74	628
F-4	13.99	17.22	1,222
F-86	2.48	6.79	590
F-89	5.42	18.12	546
F-100	4.21	12.12	752
F-101	8.93	13.42	872
F-102	11.05	12.30	680
F-104	5.04	7.96	1,150
F-105	12.21	19.30	1,112
F-106	11.88	14.62	1,153
F-111	26.93	33.15	1,262
S-3	12.33	18.54	429
T-38	4.37	5.38	699
T-39	2.57	7.03	468

Note. Adapted from "Aircraft Airframe Cost Estimating Relationships: All Mission Types," by R. Hess and H. Romanoff, 1987, p.11, 80. Retrieved from <http://www.rand.org/pubs/notes/N2283z1.html>.

Appendix B

Meta-Analysis Data

Research				Methodology						
#	Author	Cost Estimating Focus Area	Dataset Description	n		Nearest N.	Tree	ANN	Accuracy Measure	
				Train	Validate					
1	Chiu et al. (2007)	Software	Canadian Financial	14	7 ^b	88.0	49.0	89.0	70.0	MAPE
2	Chiu et al. (2007)	Software	IBM DP Services	15	8 ^b	72.0	36.0	77.0	90.0	MAPE
3	Cirilovic et al. (2014)	Construction	World Bank Asphalt	106	1 ^a	0.68			0.75	R ²
4	Cirilovic et al. (2014)	Construction	World Bank Road Rehab.	94	1 ^a	0.58			0.71	R ²
5	Dejaeger et al. (2012)	Software	ISBSG	773	387	58.5	46.9	56.4	56.7	MdAPE
6	Dejaeger et al. (2012)	Software	Experience	418	209	48.8	42.6	41.0	44.8	MdAPE
7	Dejaeger et al. (2012)	Software	ESA	87	44	58.3	48.4	53.3	57.1	MdAPE
8	Dejaeger et al. (2012)	Software	USP 05	129	64	51.2	31.8	38.9	48.1	MdAPE

9	Dejaeger et al. (2012)	Software	Euroclear	90	1 ^a	64.4	48.0	51.2	51.7	MdAPE
10	Dejaeger et al. (2012)	Software	COCNASA	94	1 ^a	51.3	44.0	45.0	38.5	MdAPE
11	Dejaeger et al. (2012)	Software	COC81	63	1 ^a	177.0	74.8	65.3	79.0	MdAPE
12	Dejaeger et al. (2012)	Software	Deshairnais	81	1 ^a	29.4	34.6	30.4	25.4	MdAPE
13	Dejaeger et al. (2012)	Software	Maxwell	61	1 ^a	48.2	36.2	45.5	44.1	MdAPE
14	Finnie et al. (1997)	Software	Desharnais	249	50	0.62	0.36		0.35	MARE
15	Heiat (2002)	Software	IBM DP Services, Hallmark	60	7	40.4			32.0	MAPE
16	Huang et al. (2008)	Software	COC81	42	21 ^b		446.0	244.0	143.0	MAPE
17	Huang et al. (2008)	Software	IBM DP Services	22	11 ^b		58.0	76.0	86.0	MAPE
18	Kaluzny et al. (2011)	Shipbuilding	NATO Task Gp. ('54-'10)	57	2		16.00	11.00		MAPE
19	Kim et al. (2004)	Construction	S. Korean Residential ('97-'00)	490	40	7.0	4.8		3.0	MAER
20	Kumar et al. (2008)	Software	Canadian Financial	36	8	158.3			14.7	MAPE

Appendix B, continued

Research			Methodology				
#	Author	Cost Estimating Focus Area	Dataset Description	n		Accuracy Measure	
				Train	Validate		
21	Park et al. (2008)	Software	S. Korean IT Service Vendors	104	44	150.4	MRE
22	Shehab et al. (2010)	Construction	Sewer Rehab. ('00-'04)	44	10	37.9	MAPE
23	Shepperd et al. (1997)	Software	Albrecht	24	1 ^a	90.0	MAPE
24	Shepperd et al. (1997)	Software	Atkinson	21	1 ^a	40.0	MAPE
25	Shepperd et al. (1997)	Software	Desharnais	77	1 ^a	66.0	MAPE
26	Shepperd et al. (1997)	Software	Finnish	38	1 ^a	128.0	MAPE
27	Shepperd et al. (1997)	Software	Kemerer	15	1 ^a	107.0	MAPE
28	Shepperd et al. (1997)	Software	Mermaid	28	1 ^a	226.0	MAPE

29	Shepperd et al. (1997)	Software	Telecom 1	18	1 ^a	86.0	39.0	MAPE
30	Shepperd et al. (1997)	Software	Telecom 2	33	1 ^a	72.0	37.0	MAPE
31	Shin (2015)	Construction	S. Korean Schools ('04-'07)	204	30	5.8	6.1	MAER
32	Zhang et al. (1996)	Manufacturing	Product Packaging	60	20	13.2	5.2	MAPE

LEGEND	
a	leave-one-out cross validation
b	three-fold cross validation
MAPE	Mean Absolute Percent Error
MdAPE	Median Absolute Percent Error
MAER	Mean Absolute Error Rate
MARE	Mean Absolute Relative Error
MRE	Mean Relative Error
R ²	coefficient of determination

Appendix C

Meta-Analysis Win-Loss Results

#	OLS	Nearest N.	OLS	Tree	OLS	ANN	Nearest N.	Tree	Nearest N.	ANN	Tree	ANN
1	Lose	Win	Win	Lose	Lose	Win	Win	Lose	Win	Lose	Lose	Win
2	Lose	Win	Win	Lose	Win	Lose	Win	Lose	Win	Lose	Win	Lose
3					Lose	Win						
4					Lose	Win						
5	Lose	Win	Lose	Win	Lose	Win	Lose	Win	Win	Lose	Win	Lose
6	Lose	Win	Lose	Win	Lose	Win	Lose	Win	Win	Lose	Win	Lose
7	Lose	Win	Lose	Win	Lose	Win	Win	Lose	Win	Lose	Win	Lose
8	Lose	Win	Lose	Win	Lose	Win	Win	Lose	Win	Lose	Win	Lose
9	Lose	Win	Lose	Win	Lose	Win	Win	Lose	Win	Lose	Win	Lose
10	Lose	Win	Lose	Win	Lose	Win	Win	Lose	Lose	Win	Lose	Win
11	Lose	Win	Lose	Win	Lose	Win	Lose	Win	Win	Lose	Win	Lose
12	Win	Lose	Lose	Win	Lose	Win	Lose	Win	Lose	Win	Lose	Win
13	Lose	Win	Lose	Win	Lose	Win	Win	Lose	Win	Lose	Lose	Win
14	Lose	Win			Lose	Win			Lose	Win		
15					Lose	Win						
16							Lose	Win	Lose	Win	Lose	Win
17							Win	Lose	Win	Lose	Win	Lose
18							Lose	Win				
19	Lose	Win			Lose	Win			Lose	Win		

Appendix C, continued

#	OLS	Nearest N.	OLS	Tree	OLS	ANN	Nearest N.	Tree	Nearest N.	ANN	Tree	ANN
20					Lose	Win						
21					Lose	Win						
22					Lose	Win						
23	Lose	Win										
24	Lose	Win										
25	Lose	Win										
26	Lose	Win										
27	Lose	Win										
28	Lose	Win										
29	Lose	Win										
30	Lose	Win										
31											Win	Lose
32					Lose	Win						
Wins	1	20	2	9	1	19	8	6	10	5	9	5
Losses	20	1	9	2	19	1	6	8	5	10	5	9

Author Biographies



Capt Gregory E. Brown, USAF, is the cost chief for Special Operations Forces and Personnel Recovery Division, Air Force Life Cycle Management Center, Wright-Patterson Air Force Base, Ohio. He received a BA in Economics and a BS in Business-Finance from Colorado State University and an MS in Cost Analysis from the Air Force Institute of Technology. Capt Brown is currently enrolled in graduate coursework in Applied Statistics through Pennsylvania State University.

(E-mail address: Gregory.Brown.34@us.af.mil)



Dr. Edward D. White is a professor of statistics in the Department of Mathematics and Statistics at the Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio. He received his MAS from Ohio State University and his PhD in Statistics from Texas A&M University. Dr. White's primary research interests include statistical modeling, simulation, and data analytics.

(E-mail address: Edward.White@afit.edu)

This content is in the Public Domain.