An Application of Data Mining Algorithms for Shipbuilding Cost Estimation

Bohdan L. Kaluzny*
Defence Research & Development Canada Centre for Operational Research & Analysis
Bohdan.Kaluzny@drdc-rddc.gc.ca

April 27, 2011

Acknowledgments

Sorin Barbici† Göran Berg† Renzo Chiomento‡
Dimitrios Derpanis§ Ulf Jonsson¶ R.H.A. David Shaw‖
Marcel C. Smit** Franck Ramaroson††

*Data requests are subject to approval of the nations participating in the NATO RTO SAS Panel 076 Task Group.
†Swedish Defence Materiel Administration (FMV)
‡Direction générale de l'armement (DGA), France
§Greek Defence Planning, Programming Division
¶Swedish Defence Research Agency (FOI)
‖Defence Research & Development Canada Centre for Operational Research & Analysis
**Netherlands Organization for Applied Scientific Research (TNO) Defence, Security and Safety
††Organisation conjointe de coopération en matière d'armement (OCCAR)—European Armament Organization

An Application of Data Mining Algorithms for Shipbuilding Cost Estimation

Abstract

This paper presents a novel application of known data mining algorithms to the problem of estimating the cost of ship development and construction. The work is a product of North Atlantic Treaty Organization Research and Technology Organization Systems Analysis and Studies 076 Task Group "NATO Independent Cost Estimating and its Role in Capability Portfolio Analysis". In a blind, ex post exercise, the Task Group set out to estimate the cost of a class of Netherlands' amphibious assault ships, and then compare the estimates to the actual costs (the Netherlands Royal Navy withheld the actual ship costs until the exercise was completed).

Two cost estimating approaches were taken: parametric analysis and costing by analogy. For the parametric approach, the M5 system (a combination of decision trees and linear regression models) of Quinlan (1992) for learning models that predict numeric values was employed. Agglomerative hierarchical cluster analysis and non-linear optimization was used for a cost estimation by analogy approach void of subjectivity.

INTRODUCTION

Background

A goal of the North Atlantic Treaty Organization (NATO) Research and Technology Organization (RTO) Systems Analysis and Studies (SAS) 076 Task Group, titled "NATO Independent Cost Estimating and its Role in Capability Portfolio Analysis", is to demonstrate the practicality of guidelines for cost estimation of defence systems described by the NATO RTO SAS-054 Panel (2007), titled "Methods and Models for Life Cycle Costing".

This paper documents the SAS-076 Task Group efforts in generating independent estimates of the development and construction costs of the Royal Netherlands Navy Rotterdam class ships in a blind, ex post exercise—only after the independent cost estimates were completed did the Task Group obtain information on the actual cost of Rotterdam class ships. The differences between the actual and estimated costs were then analyzed.

The Rotterdam class is a Landing Platform Dock (LPD) or amphibious warfare ship of the Royal Netherlands Navy. The lead ship, Her Netherlands Majesty's Ship (HNLMS) Rotterdam, pennant number L800, was launched in 1997. The second ship of the class, HNLMS Johan de Witt (L801), was launched in 2007. The ships have a large helicopter deck and a well deck for large landing craft. The class was a joint design between the Netherlands and Spain[1]. HNLMS Rotterdam is pictured in Figure 1.

Figure 1: HNLMS Rotterdam L800 Landing Platform Dock Ship

Methodology

Two independent methods were used in generating a cost estimate for HNLMS Rotterdam and Johan de Witt. The first method, classified as a parametric approach, employs the M5 model

tree algorithm of Quinlan (1992), a system that combines features of decision trees with linear regression models. The second method employed is an analogy approach based on hierarchical cluster analysis and non-linear optimization.

Parametric Cost Estimation

Parametric approaches to cost estimation use regression or other statistical methods to develop Cost Estimating Relationships (CERs). Strengths in parametric approaches are their potential to capture major portions of an estimate quickly and with limited information, their basis on consistent and quantitative inputs, and standard tests of validity, including a coefficient of correlation indicating the strength of association between the independent variables and the dependent variables in the CER. A major disadvantage of typical parametric approaches is that they may not provide low level visibility (cost breakdown) and changes in sub-systems are not reflected in the estimate if they are not quantified via an independent variable.

Traditional ship building CERs are often mathematically simple (e.g., a simple ratio) or involve linear regression analysis (of historical systems or subsystems) on a single parameter (weight, length, density, etc.). However, Miroyannis (2006) noted that this is often insufficient and that other cost driving factors must be incorporated to develop estimates of sufficient quality at the preliminary design phase. Furthermore, the relationship between the parameter(s) and cost may not be best expressed in linear form. While the field of regression analysis offers a multitude of alternative approaches (see Ryan (1997)), linear regression is the most popular and easiest to understand.

Using the M5 model tree algorithm, this paper describes a novel parametric approach for ship cost estimation that incorporates a multitude of cost driving factors, while remaining a "top down" approach applicable in early design phases of the procurement cycle. It combines features of decision trees with linear regression models to both classify similar ships (based on attributes) and build piece-wise multivariate linear regression models.

Costing by Analogy

Cost estimation by analogy is typically accomplished by forecasting the cost of a new system based on the historical cost of similar or analogous systems. This requires a reasonable correlation between the new and historical system. The cost of the historical system is adjusted by undertaking a technical evaluation of the differences between the systems, deducting the cost of components that are not comparable to the new design and adding estimated costs of the new components. Usually subject matter experts are required to make a subjective evaluation of the differences between the new system of interest and the historical system, and subjectively chosen complexity factors are often used to adjust the analogous system's cost to produce an estimate. This subjectivity is a disadvantage of traditional analogy methods.

This paper describes how a combination of hierarchical cluster analysis, principal component analysis, and non-linear optimization can be used for a novel cost estimation by analogy approach that is void of subjective adjustment factors. The approach also considers multiple analogous systems rather than just one. Hierarchical cluster analysis identifies the historical systems that are the "nearest neighbours" to the new system. A hierarchy of clusters grouping similar items together is produced: from small clusters of very similar items to large clusters that include more

dissimilar items. A matrix of distances (measure quantifying the similarity of two ships) among systems is calculated expressing all possible pairwise distances among them. These distances are then used to predict the cost of a new instance.

Both data mining approaches are data intensive. A database of 57 ships in 16 classes from 6 nations was compiled. 136 descriptive, technical, and cost attributes were gathered for each of the ships.

Outline

The paper starts by detailing the multinational data that was gathered to facilitate data mining. Subsequent sections present the parametric cost estimation method and analogy cost estimation method, and their application to the data set. The results for the predicted cost of HNLMS Rotterdam and Johan de Witt LPDs are then presented and are compared to the actual costs.

DATA

The SAS-076 Task Group compiled a database of 57 ships in 16 classes from 6 nations. The sources of data were culled from SAS-076 participants and publicly-available sources such as Jane's Fighting Warships[2], Federation of American Scientists[3], Navy Matters[4], Forecast International[5], U.S. Naval Institute sources (e.g., Friedman (2005)), and Wikipedia: The Free Encyclopedia[6]. The ships included cover a span of years (commissioned) from 1954 to 2010. Table 1 lists the ships in the data set. The first three columns indicate the name, pennant number, and type of ship. The subsequent three columns indicate the rank (in class), year of commission, and nationality. The data set includes seven ship categories:

- Amphibious Assault Ship (AAS);
- Auxiliary Oiler Replenishment (AOR);
- Landing Platform Dock (LPD);
- Landing Platform Helicopter (LPH);
- Landing Ship Dock (LSD); and,
- Icebreaker.

Forty of the ships are from the U.S., seven from the United Kingdom, four from France, three from Sweden, two from Canada, and one from Norway.

The SAS-076 ship data set contains military and civilian auxiliary (coast guard or similar) vessels that were judged to be analogous to HNLMS Rotterdam LPD. Representatives of the SAS-076 Task Group were solicited to provide technical and cost information for their nation's ships most closely resembling the role or size of a LPD. The selection of ships for inclusion was primarily driven by the accessibility of costing information (for example, detailed technical information was available for the Italian San Giorgio class and the Spanish Galicia class LPDs, but cost information was unobtainable). Ships such as Sweden's Carlskrona LPD, and Atle and Oden icebreakers were

Table 1: Description of analogous ships

Name	Number	Type	Rank	Commissioned	Country
Thomaston	LSD 28	LSD	1	1954	United States
Plymouth Rock	LSD 29	LSD	2	1954	United States
Fort Snelling	LSD 30	LSD	3	1955	United States
Point Defiance	LSD 31	LSD	4	1955	United States
Spiegel Grove	LSD 32	LSD	5	1956	United States
Alamo	LSD 33	LSD	6	1956	United States
Hermitage	LSD 34	LSD	7	1956	United States
Monticello	LSD 35	LSD	8	1957	United States
Anchorage	LSD 36	LSD	1	1969	United States
Portland	LSD 37	LSD	2	1970	United States
Pensacola	LSD 38	LSD	3	1971	United States
Mount Vernon	LSD 39	LSD	4	1972	United States
Fort Fisher	LSD 40	LSD	5	1972	United States
Whidbey Island	LSD 41	LSD	1	1985	United States
Germantown	LSD 42	LSD	2	1986	United States
Fort McHenry	LSD 43	LSD	3	1987	United States
Gunston Hall	LSD 44	LSD	4	1989	United States
Comstock	LSD 45	LSD	5	1990	United States
Tortuga	LSD 46	LSD	6	1990	United States
Rushmore	LSD 47	LSD	7	1991	United States
Ashland	LSD 48	LSD	8	1992	United States
Harpers Ferry	LSD 49	LSD	1	1995	United States
Carter Hall	LSD 50	LSD	2	1995	United States
Oak Hill	LSD 51	LSD	3	1996	United States
Pearl Harbour	LSD 52	LSD	4	1998	United States
Raleigh	LPD 1	LPD	1	1962	United States
Vancouver	LPD 2	LPD	2	1963	United States
La Salle	LPD 3	LPD	3	1964	United States
Austin	LPD 4	LPD	1	1965	United States
Ogden	LPD 5	LPD	2	1965	United States
Duluth	LPD 6	LPD	3	1965	United States
Cleveland	LPD 7	LPD	4	1967	United States
Dubuque	LPD 8	LPD	5	1967	United States
Denver	LPD 9	LPD	6	1968	United States
Juneau	LPD 10	LPD	7	1969	United States
Coronado	LPD 11	LPD	8	1970	United States
Shreveport	LPD 12	LPD	9	1970	United States
Nashville	LPD 13	LPD	10	1970	United States
Trenton	LPD 14	LPD	11	1971	United States
Ponce	LPD 15	LPD	12	1971	United States
Svalbard	W303	Icebreaker	1	2001	Norway
Carlskrona	M04	LPD	1	1982	Sweden
Atle	—	Icebreaker	1	1985	Sweden
Oden	—	Icebreaker	1	1989	Sweden
Protecteur	AOR 509	AOR	1	1969	Canada
Preserver	AOR 510	AOR	2	1970	Canada
Albion	L14	LPD	1	2003	United Kingdom
Bulwark	L15	LPD	2	2005	United Kingdom
Largs Bay	L3006	LSD	1	2006	United Kingdom
Lyme Bay	L3007	LSD	2	2007	United Kingdom
Mounts Bay	L3008	LSD	3	2006	United Kingdom
Cardigan Bay	L3009	LSD	4	2006	United Kingdom
Ocean	L12	LPH	1	1998	United Kingdom
Siroco	L9012	LSD	2	1998	France
Mistral	L9013	AAS	1	2006	France
Tonnerre	L9014	AAS	2	2007	France
Dixmude (BPC3)	L9015	AAS	3	2010	France

included—leading to the potentially useful data combination of ships of the "right purpose, wrong size" and "wrong purpose, right size".

The selection of technical specifications included was also driven by the availability of public information.

Technical Data

Descriptive, technical, and cost data was gathered for each of the ships in the SAS-076 data set. The list of these ship attributes were broken down into the categories as per Table 2. Table 3 details the attributes as well as the input for HNLMS Rotterdam and Johan de Witt ships Attribute units are expressed by either nominal values (e.g., "fixed pitch" (fp), "controlled pitch" (cp), "yes" (Y), "no" (N)), or by numerical units such as meters (m), millimeters (mm) megawatts (MW), knots (kts), hours (hrs), nautical miles (nmi), etc. Unknown or missing data was denoted by "?" entries.

Although considered the sister ship to HNLMS Rotterdam, HNLMS Johan de Witt has significant technical differences including longer length, larger displacement, podded propulsion (as oppose to shaft propulsion), increased lift capacity (e.g., fuel, vehicles, etc.), higher superstructure (by one deck), and larger crew capacity. For this reason HNLMS Johan de Witt's rank in class is set to one.

Table 2: Categories of ship data

	Category	Number of Attributes
I	DESCRIPTION	6
II	CONSTRUCTION	8
III	DIMENSIONS	5
IV	PERFORMANCE	8
V	PROPULSION	9
VI	ELECTRICAL POWER GENERATION	3
VII	LIFT CAPACITY	35
VIII	FLIGHT DECK	19
IX	ARMAMENT	13
X	COUNTERMEASURES	5
XI	RADARS / TACAN / IFF / SONARS	13
XII	COMBAT DATA SYSTEMS	1
XIII	WEAPONS CONTROL SYSTEMS	1
XIV	OTHER CAPABILITIES	7
XV	COST DATA	3
	Total:	136

Table 3: Complete list of ship data for the Rotterdam and Johan de Witt LPDs

Data Category & Element	Rotterdam LPD	Johan de Witt LPD
I. DESCRIPTION		
a. Name	Rotterdam	Johan de Witt
b. Nation	Netherlands	Netherlands
c. Number	L800	L801
d. Type	Landing Platform Dock	Landing Platform Dock
e. Rank in class	1	1
f. Vessel type (Military / Civilian)	Military	Military
II. CONSTRUCTION		
a. Laid down	1/25/1996	6/18/2003
b. Launched	2/22/1997	5/13/2006
c. Commissioned	4/18/1998	11/30/2007
d. Shipyard	Royal Schelde	Royal Schelde
e. City	Vlissingen	Vlissingen
f. Country	Netherlands	Netherlands
g. Continent	Europe	Europe
h. Built to civilian classification society standards?	Y	Y
III. DIMENSIONS		
a. Length (m)	162.2	176.35
b. Beam (m)	25	25
c. Draught (m)	5.9	5.9
d. Displacement (tonnes) Light load	8410	11560
e. Displacement (tonnes) Full load	12750	16680
IV. PERFORMANCE		
a. Top speed (kts)	19	19
i. Range: Total distance (nmi)	6000	10000
ii. Range: Economical speed (kts)	12	12
iii. Range: Sailing time (hours)	500	833
b. Endurance (days)	42	42
c. Crew: complement	113	146
i. Officers	13	17
ii. Non-officers	100	129
V. PROPULSION		
a. Propulsion technology	Electric	Electric
b. Propeller shafts	2	0
i. Shaft propulsion power (MW)	12	0
ii. Propeller type	fixed pitch	N/A
c. Propulsion pods	0	2
i. Total podded propulsion power (MW)	0	11
d. Net propulsion power (MW)	12	11
e. Bow Thrusters	1	2
i. Total thruster power (MW)	1.15	1.8
VI. ELECTRICAL POWER GENERATION		
a. Generators	4	4
i. Total power generation capacity (MW)	14.6	14.4
ii. Generator technology	Diesel	Diesel
VII. LIFT CAPACITY		
a. Vehicle fuel (litres)	9000	14500
b. Aviation fuel (litres)	284400	306600
c. Fresh water (litres)	263100	329900
d. Bulk cargo space (m^3)	3680	4170
e. Vehicle space (m^2)	720	1770
f. Well deck (Y/N)	Y	Y
i. Length (m)	50	35
ii. Width (m)	14	15
iii. Capacity (m^2)	700	525
iv. LCAC	0	0
v. LCM 6	?	?
vi. LCM 8	4	4
vii. LCU	4	4
viii. LVT	?	?
ix. LCVP	6	6
x. LCPL	?	?
xi. EFV	?	?
g. Cargo/Aircraft Elevator/Lifts	0	0
i. Capacity $\leq$ 5 tonnes	0	0
ii. 5 < capacity $\leq$ 10 tonnes	0	0
iii. 10 < capacity $\leq$ 15 tonnes	0	0
iv. Capacity $\geq$ 15 tonnes	0	0
h. Cranes	1	1
i. Capacity $\leq$ 5 tonnes	0	0
ii. 5 < capacity $\leq$ 10 tonnes	0	0
iii. 10 < capacity $\leq$ 15 tonnes	0	0
iv. 15 < capacity $\leq$ 20 tonnes	0	0
v. 20 < capacity $\leq$ 25 tonnes	1	1
vi. 25 < capacity $\leq$ 30 tonnes	0	0

continued on next page

continued from previous page

Data Category & Element	Rotterdam LPD	Johan de Witt LPD
vii. 30 < capacity $\leq$ 40 tonnes	0	0
viii. 40 < capacity $\leq$ 50 tonnes	0	0
ix. 50 < capacity $\leq$ 60 tonnes	0	0
x. Capacity > 60 tonnes	0	0
i. Berthing (troop capacity): Baseline	611	555
j. Berthing (troop capacity): Surge	100	100
VIII. FLIGHT DECK		
a. Equipped with flight deck? (Y/N)	Y	Y
b. Flight deck length (m)	56	58
c. Flight deck width (m)	25	25
d. Flight deck area (m^2)	1400	1450
e. Helicopter landing spots (maximum number)	2	2
i. Merlin / Sea King	2	2
ii. NH 90 / Lynx / Puma / Cougar	2	2
iii. CH-46E Sea Knight	?	?
iv. CH-53 Sea Stallion	?	?
v. MV-22 Osprey	?	?
f. Chinook capable (Yes/No)	N	Y
g. Equipped with hangar? (Y/N)	Y	Y
h. Hangar size (m^2)	475	560
i. Helicopters supported (largest total number)	6	6
i. Merlin / Sea King	4	4
ii. NH 90 / Lynx / Puma / Cougar	6	6
iii. CH-46E Sea Knight	?	?
iv. CH-53 Sea Stallion	?	?
v. MV-22 Osprey	?	?
IX. ARMAMENT		
a. Guns (calibre $\geq$ 75mm)	0	0
b. Guns (50mm $\leq$ calibre < 75mm)	0	0
c. Guns (30mm $\leq$ calibre < 50mm)	0	0
d. Guns (20mm $\leq$ calibre < 30mm)	0	0
e. 30mm CIWS emplacements (Goalkeeper)	2	2
f. 20mm CIWS emplacements (Phalanx)	0	0
g. Machine guns (12.7mm)	8	4
h. Machine guns (7.62mm)	0	0
i. SSM launchers	0	0
j. SAM launchers	0	0
k. Number of torpedoes carried	0	0
l. Torpedo tubes	0	0
m. Torpedo launchers	0	0
X. COUNTERMEASURES		
a. Chaff launchers	4	4
b. Torpedo decoys	1	1
c. Other systems	0	0
d. Number of ESM systems	1	1
e. Number of ECM systems	1	1
XI. RADARS / TACAN / IFF / SONARS		
a. Total radar systems mounted	4	5
i. A-band	0	0
ii. B-band	0	0
iii. C-band	0	0
iv. D-band	0	0
v. E-band	1	2
vi. F-band	1	2
vii. G-band	0	1
viii. H-band	0	0
ix. I-band	3	3
x. J-band	0	0
b. Number of TACAN/IFF systems mounted	1	1
c. Number of distinct sonar systems mounted	0	0
XII. COMBAT DATA SYSTEMS		
a. Number of distinct systems	2	2
XIII. WEAPONS CONTROL SYSTEMS		
a. Number of distinct systems	1	1
XIV. OTHER CAPABILITIES		
a. Equipped with hospital? (Y/N)	Y	Y
i. Number of beds	10	7
ii. Operating rooms	1	1
iii. X-Ray facility (Y/N)	Y	Y
b. Dental capability (Y/N)	Y	Y
c. Command/Control facility (Y/N)	Y	Y
d. NBCD Facilities (Y/N)	Y	Y
XV. COST DATA		
a. Base year	?	?
b. Currency	EUR	EUR
c. Development and Production Cost	?	?

Cost Data

The original ship cost data gathered by the SAS-076 Task Group expressed costs in various currencies: Great Britain pound sterling (GBP), U.S. dollars (USD), Canadian dollars (CAD), Norwegian krone (NOK), Swedish kronor (SEK), and Euros (EUR). Costs were also expressed relative to var-

ious then-years (amounts that include the effects of inflation or escalation, and/or reflect the price levels expected to prevail during the year at issue), ranging from 1952 to 2009. Following SAS-054 guidelines for cost normalization, the ship costs were converted to a common currency and then-year. Respecting the anonymity request of some of the nations, the ship costs cannot be made explicit. For the same reason, the common currency and then-year are not disclosed—all subsequent cost figures are presented using a fictitious notional common currency (abbreviated NCC). Figure 2 illustrates the histogram of the SAS-076 data set costs normalized to NCC.

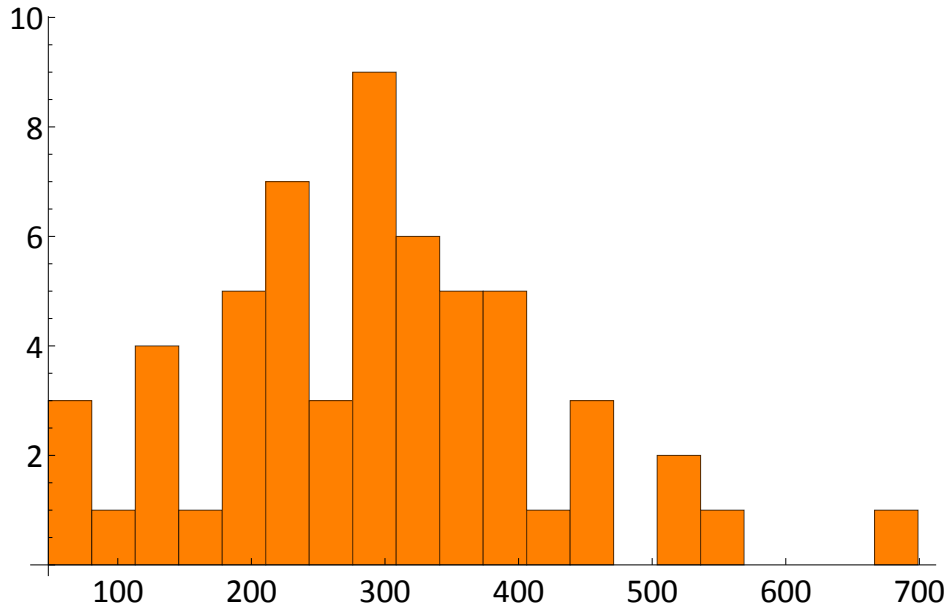

Figure 2: Histogram of normalized costs (millions NCC).

The cost of each ship was log-transformed so that the estimation models output a single predicted ship cost in log-space. This prediction is considered to center a normal distribution whose standard deviation is the standard deviation of the model (also in log space) in estimating the costs of the known ships. As a result of the initial log-transformation of the costs, the uncertainty in the prediction of a ship's cost is presented by a log-normal distribution.

The logarithmic transformation is commonly used for positive data; the log-normal distribution domain of zero to infinity is more suitable for modelling ship costs than a normal distribution which includes the negative domain. Log-transformation is also commonly applied when the data ranges over several orders of magnitude—the SAS-076 data set cost range from 48.8 to 698.7 million NCC. Total cost estimates for weapon-system acquisition programs are typically log-normally distributed and often skewed right.

DATA MINING FOR PARAMETRIC COST ESTIMATION

This section describes a novel parametric approach for ship cost estimation that incorporates a multitude of cost driving factors, while remaining a "top down" approach applicable in early design phases of the procurement cycle. It combines features of decision trees with linear regression models to both classify similar ships (based on attributes) and build piece-wise multivariate linear regression models.

The M5 Model Tree System

Quinlan (1992) pioneered the M5 system for learning models that predict numeric values. The M5 system combines features of decision trees with linear regression models. Whereas the nodes of a regression tree each contain a constant value (prediction), each model tree node is a multivariate linear regression model. This difference is the reason why regression trees will not predict values lying outside the range of learned training cases, while M5 model trees can extrapolate. M5 model trees have an advantage over regression trees with respect to compactness and prediction accuracy due to the ability of model trees to exploit local linearity in the data. M5 model trees are also smaller, easier-to-understand, and Wang and Witten (1997) show that their average error values on the training data are lower.

The M5 model tree algorithm has four parts to it. In the first part, a decision tree is constructed using the same procedure as for regression trees. The tree is constructed recursively by splitting the training set per attribute value chosen to minimize error in estimation. In the second part, the algorithm constructs multivariate linear models at each node of the model tree using only the attributes that are referenced by tests somewhere in the subtree of this node. These linear models are further simplified by eliminating parameters to minimize the estimated error (accuracy of the model on unseen cases). Part three of the algorithm applies a tree pruning routine which eliminates subtrees of a node if the estimation error is higher in the lower branches than the estimation error when using the node's internal regression model. Finally, a smoothing process, with the goal of improving prediction accuracy, is employed to ensure that the linear regression models of adjacent leaves are continuous and smooth. This process is particularly effective when some of the linear regression models are constructed from few training cases.

Figure 3 depicts a simple example of a M5 model tree. The sub-figure on the left shows a two-dimensional space of independent variables x_1 and x_2. The M5 model tree algorithm splits up the space into regions corresponding to decisions in the tree shown on the right. Linear regression models (LMs) are fitted to the data in each region.

To predict a value for a new instance, the M5 model tree is followed down to a leaf using the instance's attribute values to make routing decisions at each node. The leaf contains a linear regression model based on a subset of the attributes, and this is evaluated for the new instance to output a predicted value.

The use of M5 model trees for numeric prediction has increased since comprehensive descriptions, implementations, and refinements of Quinlan's method became available (Wang and Witten; 1997; Malerba et al.; 2004; Torgo; 2000, 2002; Dobra; 2002). Recently Chen (2006) discussed the benefits of the system for estimating the cost of software development.

Wang and Witten have shown that M5 model trees can excel when data is limited, and can learn efficiently (computationally) from large data sets. They can handle data sets which include

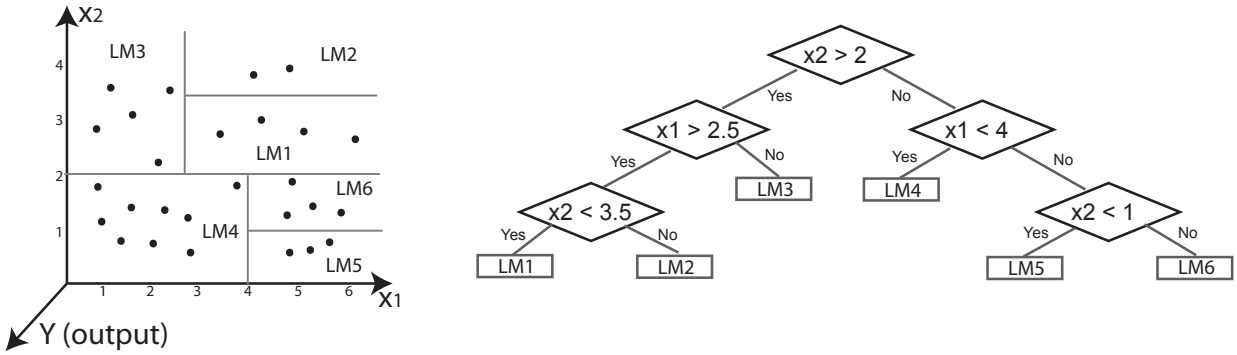

Figure 3: Example M5 model tree.

systems with notable differences, missing data, and noise (as is the case for the SAS-076 ship data set). The decision tree can branch on any variable type: nominal (e.g., military vs. non-military) or numeric (e.g., tonnage less than 15000 or greater than 15000).

Application to SAS-076 Ship Data Set

An easy-to-use implementation of Quinlan's M5 model tree is available as part of the WEKA project[7]. Witten and Frank (2005) provide documentation on both the theory and implementation (data formatting, execution, etc.) of the algorithm. The descriptive attributes (I.e), (I.f), (II.h) (as per Table 2), the complete set of technical attributes, and a normalized cost attribute for each of the ships in the SAS-076 data set was used as input to the M5 model tree algorithm. Using WEKA[8], the construction of the M5 model tree for the ship database of 57 ships each with 123 attributes (7011 elements) took less than a second of computation on an Intel(R) 2.4GHz computer with 4GB RAM.

Figure 4 shows the resulting M5 model tree. The root of the decision tree splits the ships in two based on the number of air-cushioned landing craft (LCAC) that the ship is designed to carry (it should be noted that this split also groups all ships void of a well deck). Internal nodes of the tree further split the data set on attributes such as the number (#) of torpedo decoy systems on board, the ship's rank in class, the maximum number of helicopters supported, the ship's length, and the ship's range in terms of total distance in nautical miles. The tree branches out to nine leaves where nine corresponding linear regression models are fitted. The regression models are presented in Table 4. In addition to the ship attributes used in the decision tree for branching, the linear regression models use the ship's range in terms of total sailing time in hours. The linear regression models output the log-transformed (decadic) cost of a ship. The individual linear regression models are mostly intuitive: the cost of a ship increases as the length or the number of LCAC supported or number of torpedo decoy systems increase(s). The regression models also predict a shipbuilding learning curve as the cost of constructing a ship decreases as a function of the ship's rank in class. The negative coefficient for a ship's range (sailing time) in the regression models is counter-intuitive. It seems unlikely that a ship will cost less as its range increases. The SAS-076 ship data explains this anomaly: the median ship range (sailing time) of the ships captured in the SAS-076 ship data set is 444 hours and the mean is 616 hours. Only 6 of the 57 ships have a range greater

than 770 hours, these are the U.S. Anchorage class LSDs and Sweden's Oden icebreaker—their sailing time range is between 3-5 times the median. The Anchorage class LSD costs and the Oden icebreaker cost are relatively low (in comparison to the other SAS-076 ships). The combination of these low costing ships and outlying sailing time ranges provides a mathematical explanation for the negative coefficient of the sailing time range in the regression models. The M5 model can be potentially adjusted in an attempt to remove such anomalies by disabling the particular attribute (sailing time), however there is no guarantee that the regenerated model will not substitute this attribute with another, also allocated a negative coefficient. Similarly, removing the instances (e.g., Oden and Anchorage class LPDs) from the data set provides no guarantees. Rather than subjectively diminishing the data set, anomalies are noted and discussed as part of the results.

Table 5 provides the minimum, median, mean, and maximum values found in the SAS-076 ship data set for the attributes used by the M5 model. Table 6 shows the classification of the SAS-076 ships by the M5 model tree.

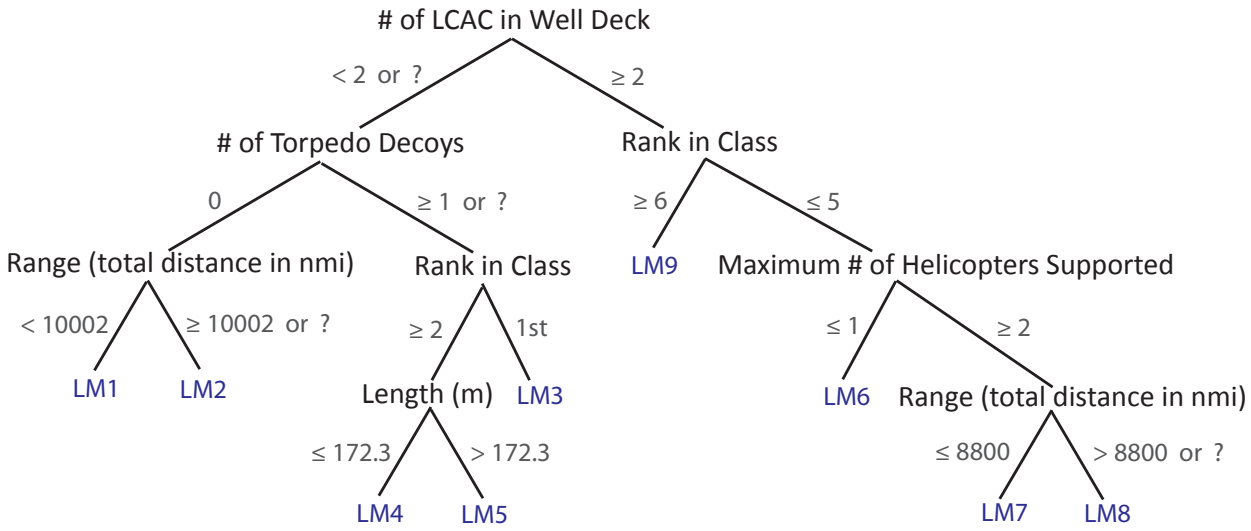

Figure 4: M5 model tree applied to the NATO RTO SAS 076 ship data set.

Figure 5 plots the actual ship costs vs. the costs predicted by the M5 model tree. The worth of a regression-based model is measured by the coefficient of correlation, R, the quantity that gives the quality of a least squares fitting to the original data. Similarly, the value R^2, known as the coefficient of determination, is a measure of how well the regression line represents the data—it is a measure determining how certain one can be in making predictions from the model. R^2 is also the ratio of the explained variation (by the model) to the total variation of the data set. The measures of worth of the M5 model tree for predicting ship costs are a strong $R = 0.96$ and $R^2 = 0.92$ (a correlation greater than 0.8 is generally described as strong (Ryan; 1997)). The coefficient of determination indicates that 92% of the total variation in the ship costs can be explained by the linear relationships described by the M5 model tree linear regression equations. The remaining 8% of the total variation remains unexplained.

The mean absolute percent error for the M5 model tree applied to the SAS-076 ship data set is 12%. The standard deviation is 46.4 million NCC. Mean absolute percent errors and standard deviations specific to the individual M5 model tree linear regression models are shown in Table 7.

Table 4: M5 model tree linear regression models

LM1	
Log(Cost) =	7.4297
	- 0.0112 × rank in class
	+ 0.0045 × length (m)
	- 0.0002 × range (sailing time in hrs)
	+ 0.0445 × # of LCAC in well deck
	+ 0.1104 × # of torpedo decoys
LM2	
Log(Cost) =	7.4208
	- 0.0112 × rank in class
	+ 0.0045 × length (m)
	- 0.0002 × range (sailing time in hrs)
	+ 0.0445 × # of LCAC in well deck
	+ 0.1104 × # of torpedo decoys
LM3	
Log(Cost) =	7.6222
	- 0.0167 × rank in class
	+ 0.0041 × length (m)
	- 0.0002 × range (sailing time in hrs)
	+ 0.0445 × # of LCAC in well deck
	+ 0.0659 × # of torpedo decoys
LM4	
Log(Cost) =	7.7567
	- 0.0172 × rank in class
	+ 0.0032 × length (m)
	- 0.0002 × range (sailing time in hrs)
	+ 0.0445 × # of LCAC in well deck
	+ 0.0659 × # of torpedo decoys
LM5	
Log(Cost) =	7.7912
	- 0.0170 × rank in class
	+ 0.0030 × length (m)
	- 0.0002 × range (sailing time in hrs)
	+ 0.0445 × # of LCAC in well deck
	+ 0.0659 × # of torpedo decoys
LM6	
Log(Cost) =	7.9846
	- 0.0245 × rank in class
	+ 0.0038 × length (m)
	- 0.0003 × range (sailing time in hrs)
	+ 0.0300 × # of LCAC in well deck
	+ 0.0343 × # of torpedo decoys
LM7	
Log(Cost) =	8.1461
	- 0.0361 × rank in class
	+ 0.0035 × length (m)
	- 0.0003 × range (sailing time in hrs)
	+ 0.0300 × # of LCAC in well deck
	+ 0.0343 × # of torpedo decoys
LM8	
Log(Cost) =	8.3575
	- 0.0312 × rank in class
	+ 0.0024 × length (m)
	- 0.0003 × range (sailing time in hrs)
	+ 0.0300 × # of LCAC in well deck
	+ 0.0343 × # of torpedo decoys
LM9	
Log(Cost) =	8.3001
	- 0.0221 × rank in class
	+ 0.0020 × length (m)
	- 0.0003 × range (sailing time in hrs)
	+ 0.0300 × # of LCAC in well deck
	+ 0.0343 × # of torpedo decoys

For the purpose of determining the standard deviation of a M5 model tree output prediction, the standard deviation over all the training data is more reflective of the M5 system than just the standard deviation of the training cases reaching the particular leaf node used for the prediction. By M5 model tree construction, each of the training cases influence the structure of the final model tree.

The results presented in Table 7 show that linear regression model LM7 contributes greatest to the standard deviation. Further analysis revealed that LM7 models the nine highest costing ships, all over 400 million NCC. In eight of these cases, LM7 underestimates the actual cost. By the piece-wise linear M5 model tree construction, LM7 is influenced by the adjacent linear

Table 5: Statistics of attributes used in the M5 model tree linear regression models.

Attribute	Minimum	Median	Mean	Maximum
Rank	1	3	3.68	12
Length	103.7	173.8	170.3	203.4
Range (sailing time)	385	444	616	2308
Range (total distance)	7500	10003	8000	30000
# LCAC	0	2	2	4
# torpedo decoys	0	0	2	8
# of helicopters supported	0	5	6	18

Table 6: M5 model tree classification of SAS-076 ships.

LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9
Svalbard	Carlskrona	Thomaston	Plymouth Rock	Lyme Bay	Anchorage	Whidbey Island	Raleigh	Tortuga
Protecteur	Atle	Largs Bay	Fort Snelling	Mounts Bay	Portland	Germantown	Vancouver	Rushmore
Preserver	Oden	Ocean	Point Defiance	Cardigan Bay	Pensacola	Fort McHenry	La Salle	Ashland
			Spiegel Grove		Mount Vernon	Gunston Hall	Mistral	Denver
			Alamo		Fort Fisher	Comstock	Tonnerre	Juneau
			Hermitage		Harpers Ferry	Austin	Dixmude (BPC3)	Coronado
			Monticello		Carter Hall	Ogden		Shreveport
			Siroco		Oak Hill	Duluth		Nashville
					Pearl Harbour	Cleveland		Trenton
						Dubuque		Ponce
						Albion		
						Bulwark		

Table 7: Mean absolute percent errors of known instances and standard deviations per individual M5 model tree linear models.

	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9
Mean % error:	22%	27%	17%	3%	33%	12%	14%	8%	6%
Standard deviation	24.3M	16.9M	53.0M	6.4M	45.6M	43.4M	78.0M	39.3M	24.3M
# of instances:	3	3	3	8	3	9	12	6	10

regression models LM6 and LM8. To evaluate the degree of this influence, a separate multiple linear regression model was fitted to the ships reaching the LM7 leaf using the same five ship attributes as LM7. The resulting CER is as follows:

$$\begin{aligned} Log_{10}(cost) = 8.6376 &- 0.0651 \times \text{ rank in class} \\ &+ 0.0637 \times \text{ number of LCAC.} \end{aligned} \quad (1)$$

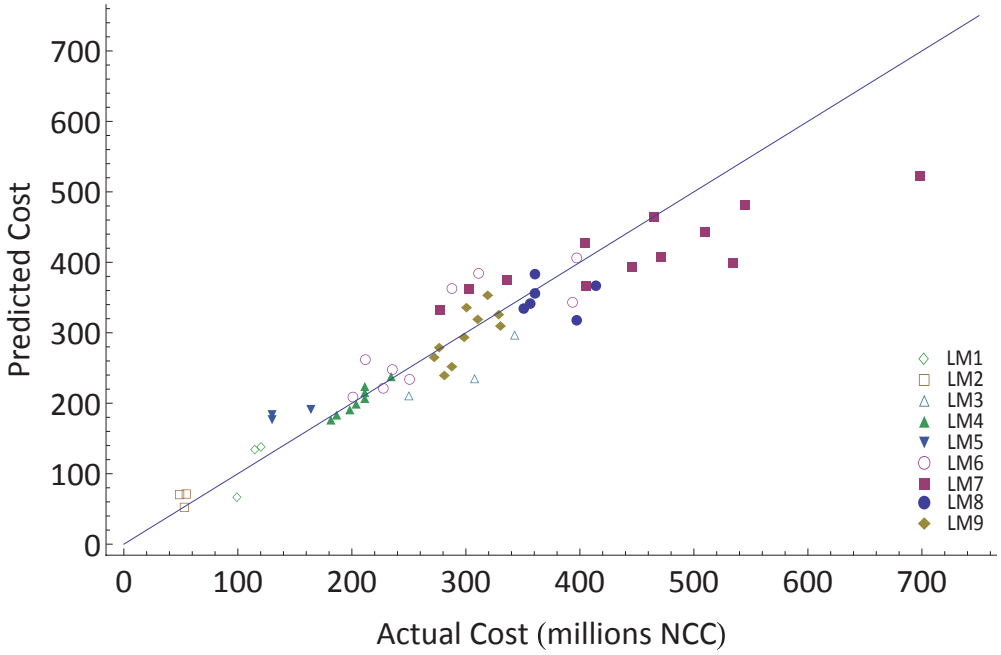

Figure 5: M5 model tree: correlation plot of actual vs. predicted ship costs (millions NCC).

with an R value of 0.95 ($R^2 = 0.90$), mean absolute percent error of 7% and a standard deviation of 34.8 million NCC. This result indicates that it is indeed possible to derive better models for subsets of the data. This does not depreciate the M5 model tree algorithm whose strength is its optimization in predicting unknown cases (rather than simply memorizing the known). Using the linear regression model of equation (1) in place of LM7 would leave adjacent linear models LM6 and LM8 sharply discontinuous. The M5 system applies a smoothing process to construct piece-wise linear regression models. Experiments by Wang and Witten show that this smoothing substantially increases the accuracy of predictions of unseen cases.

Comparison to Linear Regression Models

The best CER returned by applying simple linear regression on the SAS-076 data set (nominal attributes and attributes for which data was missing was omitted) is

$$Log_{10}(cost) = 6.95 + 0.01 \times \text{ length of ship (in meters)}, \tag{2}$$

with an R value of 0.75 ($R^2 = 0.56$). Applying multiple linear regression with a greedy attribute selection method (step through the attributes removing the one with the smallest standardized coefficient until no improvement is observed in the estimate of the error given by the Akaike (1980) information criterion), yields the CER

$$\begin{aligned} Log_{10}(cost) = 5.7368 &- 0.0224 \times \text{ rank in class} \\ &+ 0.0121 \times \text{ length (in meters)} \\ &+ 0.0338 \times \text{ beam (in meters)} \\ &+ 0.1071 \times \text{ draught (in meters)} \\ &- 0.0001 \times \text{ full load displacement (in tonnes)} \\ &+ 0.0012 \times \text{ crew size} \\ &- 0.0876 \times \text{ number of propeller shafts} \\ &- 0.0239 \times \text{ number of guns of calibre } \geq 75, \end{aligned} \tag{3}$$

with an R value of 0.92 ($R^2 = 0.85$). While the CER produced by applying simple linear regression is straightforward to understand, the negative coefficient signs of the multiple linear regression CER makes its interpretation non-trivial—detailed analysis of the input data is required. For comparison to the M5 model tree, the straightforward linear regression estimates for HNLMS Rotterdam and Johan de Witt LPDs are presented in the results section.

DATA MINING FOR COST ESTIMATION BY ANALOGY

This section describes how hierarchical cluster analysis is used for a novel cost estimation by analogy approach that is void of the subjectivity inherent (of the traditional approach) in quantifying the cost of the technical and other differences between the historical system and the new system. The approach also considers multiple analogous systems rather than just one.

Hierarchical cluster analysis

Hierarchical cluster analysis is a data mining approach that facilitates cost estimation by analogy by identifying the systems that are the "nearest neighbours" to the new system. Hierarchical cluster methods produce a hierarchy of clusters grouping similar items together: from small clusters of very similar items to large clusters that include more dissimilar items. In particular, agglomerative hierarchical methods work by first finding the clusters of the most similar items and progressively adding less similar items until all items have been included into a single large cluster. Hierarchical agglomerative cluster analysis begins by calculating a matrix of distances among systems expressing all possible pairwise distances among them. Initially each system is considered a group, albeit of a single item. Clustering begins by finding the two systems that are most similar, based on the distance matrix, and merging them into a single group. The characteristics of this new group are based on a combination of the systems in that group. This procedure of combining two groups and merging their characteristics is repeated until all the systems have been joined into a single large cluster.

Hierarchical cluster analysis is a useful means of observing the structure of the data set. The results of the cluster analysis are shown by a dendrogram (tree), which lists all of the samples and indicates at what level of similarity any two clusters were joined. The x-axis is a measure of the

similarity or distance at which clusters join. The resulting clustering can be used to estimate the cost of a new system by taking a weighted average of the cost of historical systems based on the relative distances between the new system and the historical systems.

Application to SAS-076 Ship Data Set

Using the SAS-076 ship data set, hierarchical clustering is used to define a ship distance function which takes as input ship attributes for a pair of ships and outputs a single value indicating the distance, or similarity, between the two ships. Formally, define

$$d_{ijk} = \text{ distance between ship } i \text{ and } j \text{ with respect to attribute } k. \tag{4}$$

For numeric attributes, d_{ijk} is normalized to lie in the $[0,1]$ range with $d_{ijk} = 1$ indicating that ships i and j lie at opposite ends of the observed spectrum for attribute k (e.g., shortest and longest length ships), and $d_{ijk} = 0$ indicating that the ships are the same with respect to attribute k. For nominal attributes, d_{ijk} is binary—set to 0 if ship i and j are the same with respect to attribute k, and 1 if they are not.

A variety of distance metrics can be used to calculate similarity of two ships based on the attribute distances d_{ijk}. Using a simple Euclidean distance metric, the aggregate distance between two ships i and j, is expressed as

$$d_{ij} = \sqrt{\sum_{k \in A} d_{ijk}^2} \tag{5}$$

where A is the subset of attributes considered.

Computing the distances between all pairs of ships using equation (5), the cost of a ship i can be estimated by computing the weighted-average cost of the other ships. Let C_j be the known cost of ship j, then

$$\tilde{C}_i = \sum_{j \neq i} \frac{C_j}{d_{ij}^2} \cdot \frac{1}{\sum_{j \neq i} \frac{1}{d_{ij}^2}} \tag{6}$$

is the estimated cost of ship i. Figure 6 plots the actual ship costs, C_i, vs. the costs predicted, $\tilde{C}_i$, by the analogy method using hierarchical clustering based on a simple distance metric. The measures of worth of the analogy method via hierarchical clustering analysis (simple distance metric) for predicting ship costs are $R = 0.48$ and $R^2 = 0.23$. The mean absolute percent error is 49% and the standard deviation of 112 million NCC. This approach does poorly in learning the known ship costs.

An assumption in the above approach is that all attributes are of equal importance; the (normalized) differences or similarities for each attribute contribute equally to the measure of similarity between ships. To potentially improve the predictive capability of the method, attribute weights are defined. Let

$$w_k = \text{ weight of attribute } k. \tag{7}$$

Using a weighted Euclidean distance metric, the aggregate distance between two ships i and j, is expressed as

$$\hat{d}_{ij} = \sqrt{\sum_{k \in A} (w_k \cdot d_{ijk})^2}, \tag{8}$$

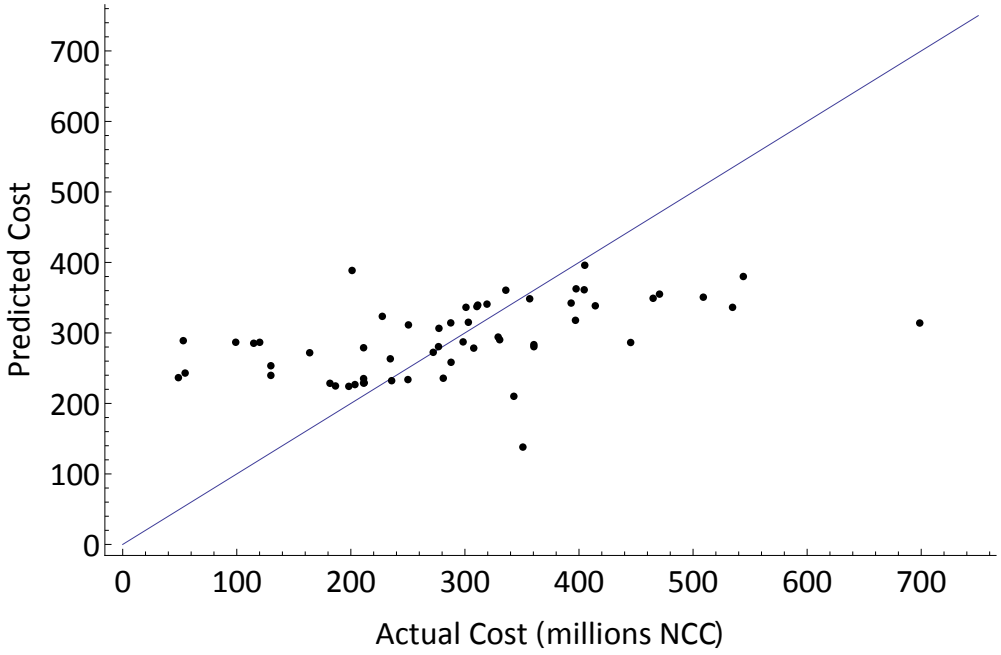

Figure 6: Hierarchical clustering with simple distance function: correlation plot of actual vs. predicted ship costs (millions NCC).

where $\sum_{k\in A} w_k = 1$ and $w_k \geq 0$ for all k. As before, let C_j be the known cost of ship j, then

$$\hat{C}_i = \sum_{j\neq i} \frac{C_j}{\hat{d}_{ij}^2} \cdot \frac{1}{\sum_{j\neq i} \frac{1}{\hat{d}_{ij}^2}} \tag{9}$$

is the estimated cost of ship i using weighted attributes. The optimal allocation of weights is determined by minimizing the prediction error for the known ships,

$$\text{minimize} \sum_{i=1}^{57} \left(C_i - \hat{C}_i\right)^2. \tag{10}$$

The resulting mathematical optimization is a non-linear convex program. With the full set of attributes used previously, $|A| = 123$, the mathematical program was too computationally intensive to solve in reasonable time using Wolfram *Mathematica*© on a Intel(R) 2.4GHz computer with 4GB RAM. To reduce the dimensionality of the problem, a smaller subset of attributes had to be selected. While there are numerous attribute selection algorithms (see Witten and Frank), principal component analysis (PCA), a tool in exploratory data analysis and for making predictive models, was used. PCA involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The reader is referred to Jolliffe (2002) for mathematical details of PCA.

The WEKA software tool was used to perform PCA. The ship data set consisting of 123 attributes (as used with the simple distance function) was reduced to a set of 16 macro-attributes accounting for 95% of the original set's variability. Each macro-attribute is a linear combination of the original attributes. For example, macro-attribute A2 is

$$\begin{aligned} \text{A2} = {} & 0.204 \times \text{length} \\ & + 0.196 \times \text{beam width} \\ & + 0.183 \times \text{vehicle space} \\ & + 0.18 \times \text{\# of expeditionary fighting vehicles} \\ & + 0.165 \times 1 \text{ if has a well deck, otherwise 0} \\ & + 0.165 \times \text{width of the well deck} \\ & + 0.164 \times \text{length of the well deck} \\ & + 0.159 \times \text{\# of large personnel landing craft} \\ & + 0.156 \times \text{\# of Chinook helicopters supported} \\ & + 0.155 \times \text{full load displacement} \\ & + 0.153 \times \text{\# of combat data systems} \\ & + 0.150 \times \text{light load displacement} \\ & + 0.144 \times \text{well deck capacity} \\ & + 0.143 \times \text{\# of elevators} \\ & + 0.142 \times \text{vehicle fuel capacity} \\ & \text{etc.} \end{aligned} \tag{11}$$

(Only the top 15 attributes—in terms of PCA coefficient size—are enumerated in equation (11).)

Table 8 lists the macro-attributes and the respective percentage of data variability (cumulative) each accounts for.

Attempting to find solutions to the mathematical program (10) with $|A| = 16$ macro-attributes still resulted in memory overflow. Using the top ten macro-attributes, accounting for 80% of the original data set's variance, an optimal solution for the macro-attribute weights was determined. The weights are listed in Table 9. Only macro-attributes A1, A2, A4, and A8 have non-zero weights. This is a typical extreme output of mathematical optimization software. There may exist other optimal solutions with other non-zero macro-attributes weights.

The hierarchical cluster analysis using the weighted distance function on the top ten macro-attributes determined by PCA is visualized in Figure 7. The figure illustrates the resulting dendrogram indicating that HNLMS Rotterdam LPD is grouped with the United Kingdom's Albion class LPDs and Largs Bay class LSDs, followed by France's Siroco LSD, etc. HNLMS Johan de Witt LPD is clustered with Sweden's Svalbard icebreaker, France's Mistral class AAS, United Kingdom's Ocean LPH, and so on. As expected, ships within the same class (but different rank) are closely grouped together.

Table 8: Principal component analysis results

Macro-Attribute	% of Data Variability Accounted for Proportion	Cumulative
A1	17%	17%
A2	12%	29%
A3	11%	41%
A4	9%	49%
A5	8%	57%
A6	7%	64%
A7	6%	69%
A8	5%	74%
A9	4%	78%
A10	3%	81%
A11	3%	85%
A12	3%	88%
A13	3%	90%
A14	2%	92%
A15	2%	94%
A16	1%	95%

Table 9: Optimal macro-attribute weights for cost estimation by hierarchical clustering

Attribute	Weight
A1	0
A2	0.452
A3	0
A4	0
A5	0.334
A6	0
A7	0
A8	0
A9	0
A10	0.214

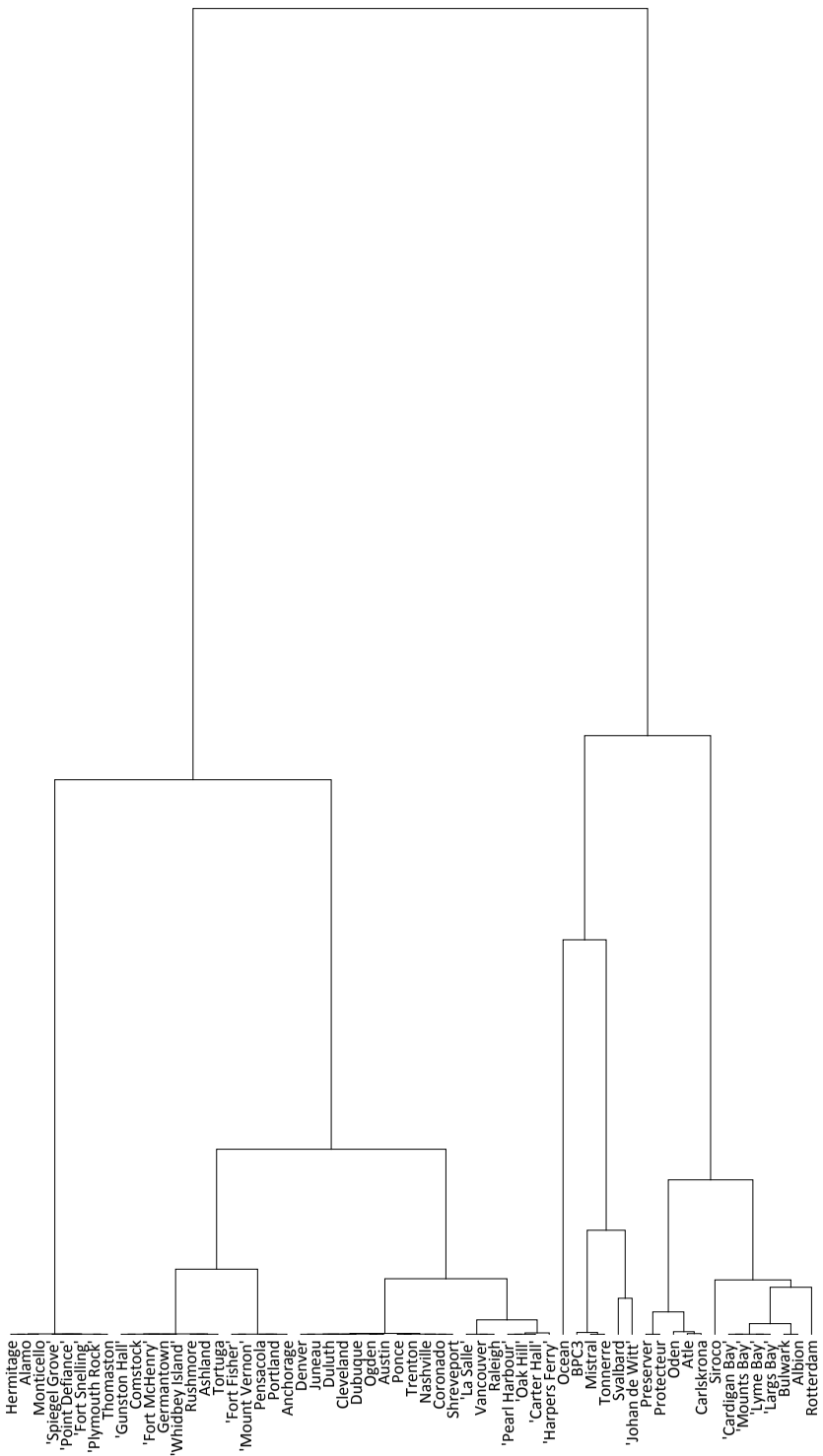

Figure 7: Dendrogram illustrating the arrangement of the clusters produced by the hierarchical clustering of ships (weighted distance function).

Figure 8 plots the actual ship costs, C_i, vs. the costs predicted, $\hat{C}_i$, by the analogy method using hierarchical clustering based on a weighted distance metric. The measures of worth of the analogy method via hierarchical clustering analysis (weighted distance matrix) for predicting ship costs are $R = 0.93$ and $R^2 = 0.86$. The latter coefficient of determination indicates that 86% of the total variation in the ship costs can be explained by an average cost of the known ships weighted by an optimized distance metric. The mean absolute percent error is 16% and the standard deviation is 55.9 million NCC—improvements over the hierarchical clustering based on the simple distance metric.

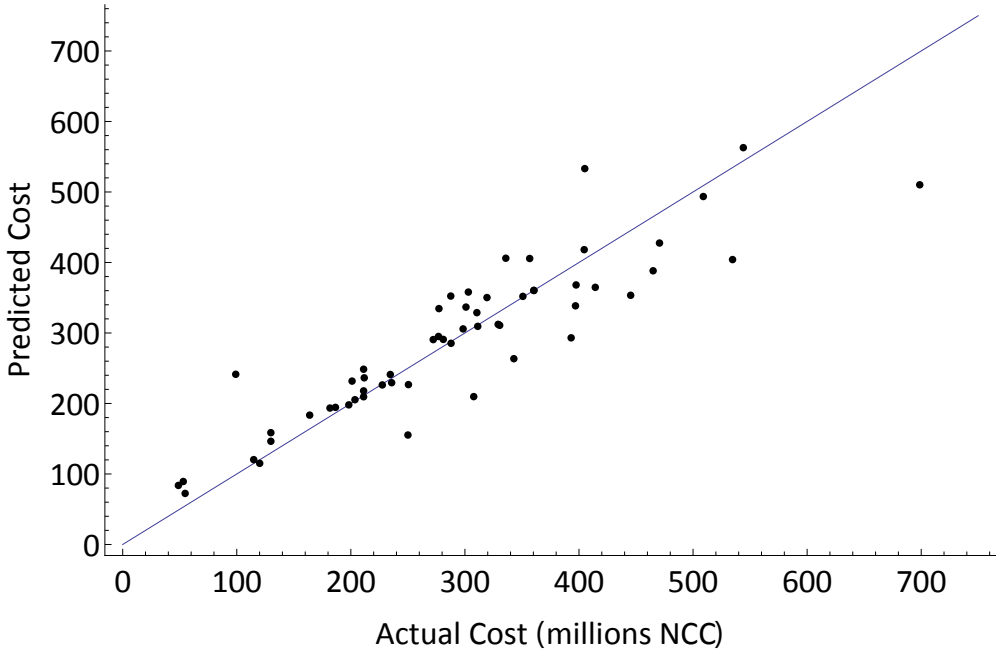

Figure 8: Weighted hierarchical clustering: correlation plot of actual vs. predicted ship costs (millions NCC).

Discussion

Figure 8 shows that the analogy method using hierarchical clustering based on a weighted distance metric underestimates the cost of seven of the eight most expensive ships (all over 406 million NCC). This was also a characteristic of the parametric estimation presented in a preceding section. In the latter it was conjectured that the underestimation was likely a result of the smoothing and pruning functions of the M5 model tree algorithm, designed to optimize the predictive capability of the method. However, the underestimation of expensive ships in the second, independent method, are potentially an indication that the attributes (and their values) of the SAS-076 ship data set do not provide enough information to help distinguish the highest costing ships from their peers.

RESULTS

M5 Model Tree Results

The technical specifications of HNLMS Rotterdam LPD (Table 3) were used to trace down the M5 model tree depicted in Figure 4. In particular, the input data indicates that HNLMS Rotterdam LPD does not have the capacity to carry LCAC in its well deck, carries one torpedo decoy system, and is ranked first in class. This results in linear regression model LM3 as per Table 4. Similarly, the attributes of HNLMS Johan de Witt LPD are used to follow the same path in the M5 model tree to linear regression model LM3. The linear regression model LM3 has the ship's rank in class, length, range in terms of sailing time in hours, number of LCAC, and number of torpedo decoy systems as independent variables. Using the LM3 model, the predicted development and production cost of HNLMS Rotterdam LPD is 197.7 million NCC. The predicted cost of HNLMS Johan de Witt LPD is 212.3 million NCC. The LM3 model outputs different predictions since the HNLMS Rotterdam LPD and Johan de Witt LPD differ in range (sailing time in hours) and length—two of the independent variables in LM3.

Figures 9 and 10 illustrate the log-normal probability density and cumulative distribution functions for the M5 model tree estimates for the cost of HNLMS Rotterdam LPD and Johan de Witt LPD. The predicted costs coincide to the 50th percentile of the respective log-normal probability distribution functions. This effect is explained by Goldberger (1968): when a power-function form is used for a CER, attention shifts from the mean to the median as a measure of central tendency; the CER yields an estimate of the median value of Y rather than the mean. The mean of the presented log-normal distributions are 200.2 million NCC and 215.0 million NCC for HNLMS Rotterdam LPD and Johan de Witt LPD respectively.

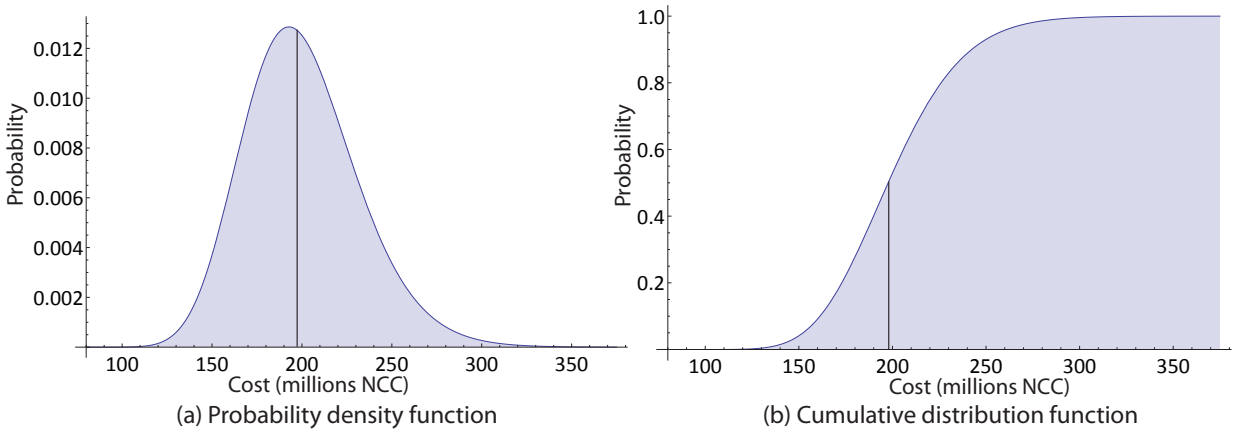

Figure 9: Probability density function (a) and cumulative distribution function (b) of the M5 model tree estimate of HNLMS Rotterdam LPD cost.

Linear Regression Results

Using simple linear regression CER, equation (2), the estimate for HNLMS Rotterdam LPD is 219.2 million NCC. The estimate for HNLMS Johan de Witt LPD is 289.7 million NCC. Using the

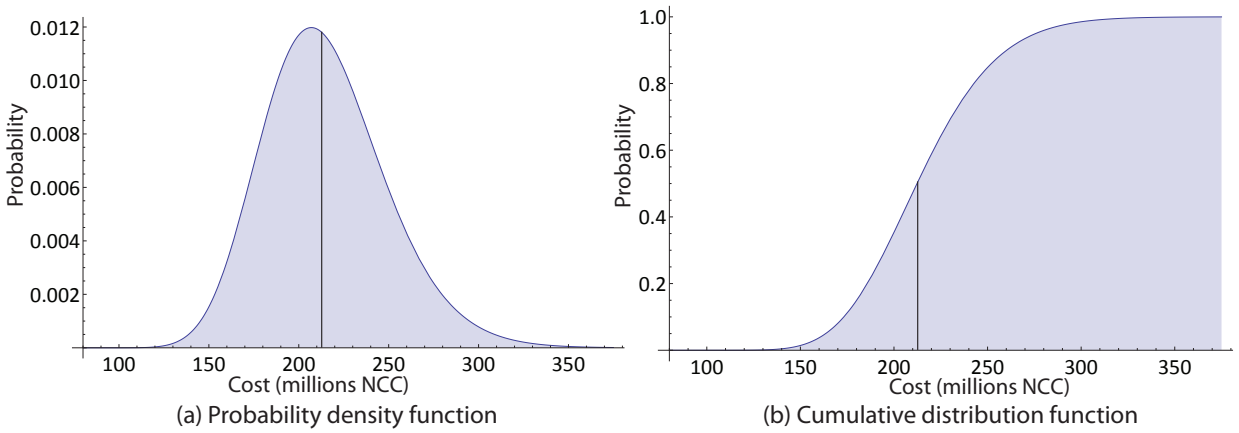

Figure 10: Probability density function (a) and cumulative distribution function (b) of the M5 model tree estimate of HNLMS Johan de Witt LPD cost.

multiple linear regression CER, equation (3), the estimate for HNLMS Rotterdam LPD is 158.9 million NCC. The estimate for HNLMS Johan de Witt LPD is 201.4 million NCC.

Hierarchical Cluster Analysis Results

The technical specifications of HNLMS Rotterdam LPD (Table 3) were mapped to the ten attributes selected by the principal component analysis. Using the optimized macro-attribute weights, the normalized relative distances of HNLMS Rotterdam LPD to the other ships are listed in Table 10 (the distances have been normalized so that the furthest ship has a distance of 1). The resulting hierarchical clustering cost estimate for HNLMS Rotterdam LPD is 214.6 million NCC, and 243.9 million NCC for HNLMS Johan de Witt LPD.

Figures 11 and 12 illustrate the log-normal probability density and cumulative distribution functions for the hierarchical cluster estimates for the cost of HNLMS Rotterdam LPD and Johan de Witt LPD. The mean of the presented log-normal distributions are 219.8 million NCC and 249.8 million NCC for HNLMS Rotterdam LPD and Johan de Witt LPD respectively.

Discussion

The estimates generated by the M5 model tree were considered to be the primary estimates for HNLMS Rotterdam and Johan de Witt, the hierarchical clustering estimates were considered as secondary estimates. This decision was driven by the following:

- The M5 model tree algorithms are optimized to both learn known cases and predict unknown cases. The attribute weights used in the hierarchical clustering method are optimized learn the known cases.
- The hierarchical clustering approach uses principal component analysis to reduce the dimensionality of the attribute space. Due to computational limitations, the weight optimization method could only be applied on the top ten macro-attributes, accounting for 80% of the

Table 10: Weighted distance of the Rotterdam LPD to ships in the Rotterdam data set.

Name	Distance	Name	Distance	Name	Distance
Rotterdam	0.000	Vancouver	0.312	Nashville	0.639
Largs Bay	0.034	La Salle	0.316	Trenton	0.646
Lyme Bay	0.034	Harpers Ferry	0.405	Ponce	0.650
Mounts Bay	0.035	Carter Hall	0.431	Whidbey Island	0.655
Cardigan Bay	0.035	Oak Hill	0.435	Germantown	0.659
Oden	0.039	Pearl Harbour	0.439	Fort McHenry	0.664
Carlskrona	0.044	Anchorage	0.546	Gunston Hall	0.668
Johan de Witt	0.046	Portland	0.550	Comstock	0.673
Atle	0.052	Pensacola	0.553	Tortuga	0.678
Albion	0.057	Mount Vernon	0.557	Rushmore	0.682
Bulwark	0.058	Fort Fisher	0.561	Ashland	0.687
Siroco	0.067	Austin	0.601	Thomaston	0.971
Svalbard	0.068	Ogden	0.606	Plymouth Rock	0.975
Protecteur	0.128	Duluth	0.610	Fort Snelling	0.979
Preserver	0.129	Cleveland	0.612	Point Defiance	0.983
Ocean	0.227	Dubuque	0.617	Spiegel Grove	0.987
Tonnerre	0.244	Denver	0.621	Alamo	0.992
Mistral	0.246	Juneau	0.626	Hermitage	0.996
BPC3	0.266	Coronado	0.630	Monticello	1.000
Raleigh	0.309	Shreveport	0.634		

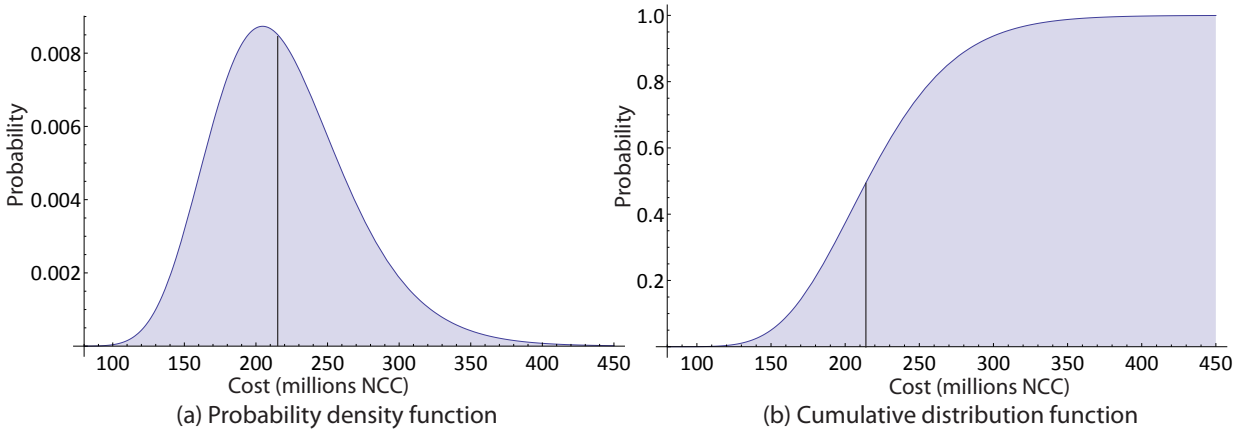

Figure 11: Probability density function (a) and cumulative distribution function (b) of the hierarchical clustering estimate of HNLMS Rotterdam LPD cost.

original data set's variability. In comparison, the M5 model tree algorithm is computationally superior as it efficiently learns from large data sets.

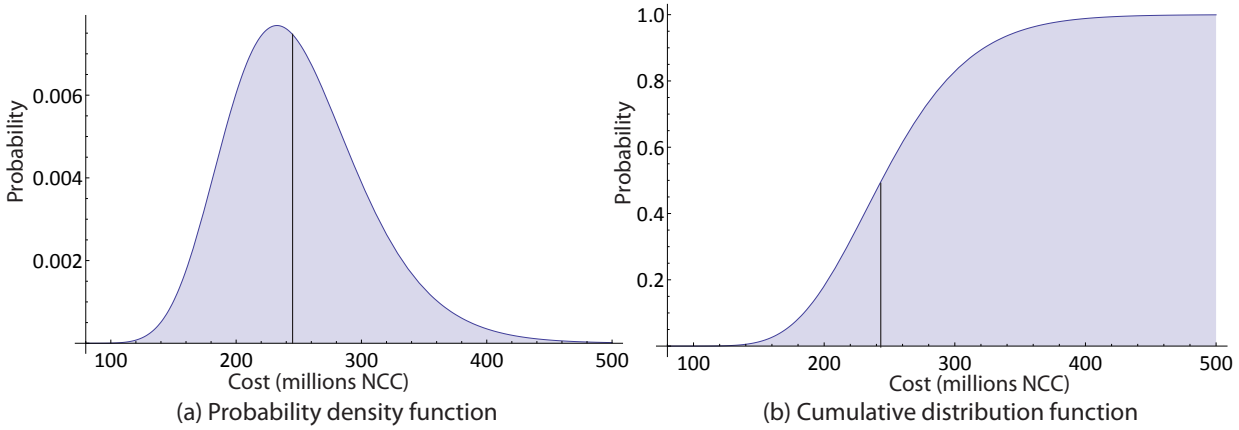

Figure 12: Probability density function (a) and cumulative distribution function (b) of the hierarchical clustering estimate of HNLMS Johan de Witt LPD cost.

- The M5 model tree results in a better correlation measure, lower mean absolute percent error, and smaller standard deviation in estimating the known cases.

Table 11 synthesizes the predictions and compares properties of the M5 model tree and hierarchical clustering methods.

Table 11: Comparison of the M5 model tree and hierarchical clustering methods and their estimates.

	M5 model tree	Hierarchical clustering
HNLMS Rotterdam estimate	197.7M NCC	214.6M NCC
HNLMS Johan de Witt estimate	221.3M NCC	243.9M NCC
Coefficient of correlation	0.96	0.93
Coefficient of determination	0.92	0.86
Standard deviation	46.4M NCC	55.9M NCC
Mean absolute % error	11%	16%
Ability to learn known cases	✓	✓
Optimized to predict unknown cases	✓	×
Uses entire data set	✓	×

Ex Post Revelation

The Royal Netherlands Navy revealed the actual development and production costs of HNLMS Rotterdam and Johan de Witt LPDs to the NATO RTO SAS 076 Task Group once the cost estimates were established. After normalization to the fictitious notional common currency, the HNLMS

Rotterdam LPD development and production costs totaled 202.2 million NCC. HNLMS Johan de Witt LPD development and production costs totaled 253.7 million NCC.

Figures 13 and 14 illustrate where the actual costs (thick vertical lines) fall with respect to the log-normal probability density and cumulative distribution functions for the M5 model tree and hierarchical cluster cost estimates for HNLMS Rotterdam LPD and Johan de Witt LPD.

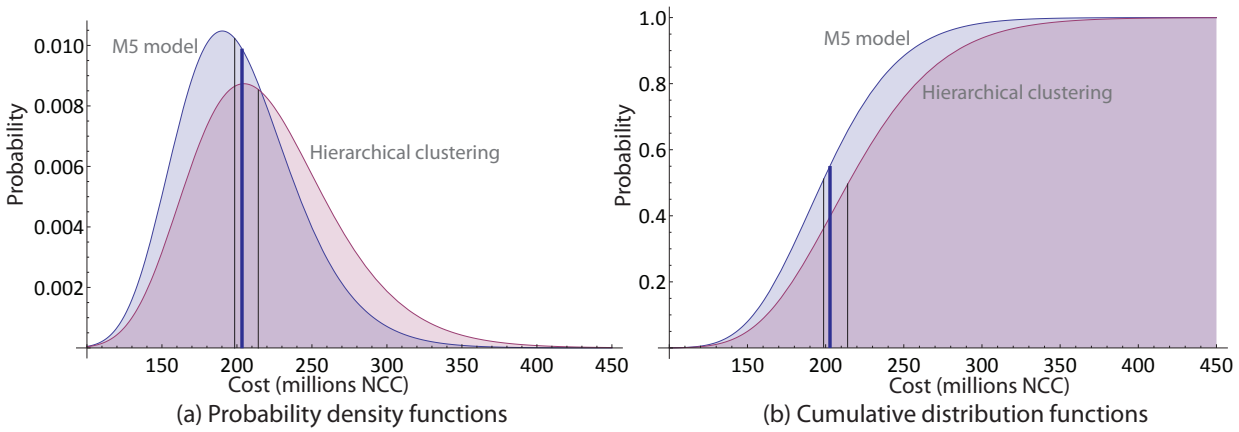

Figure 13: Probability density function (a) and cumulative distribution function (b) of the M5 model tree (blue) and hierarchical clustering (red) estimates of HNLMS Rotterdam LPD cost.

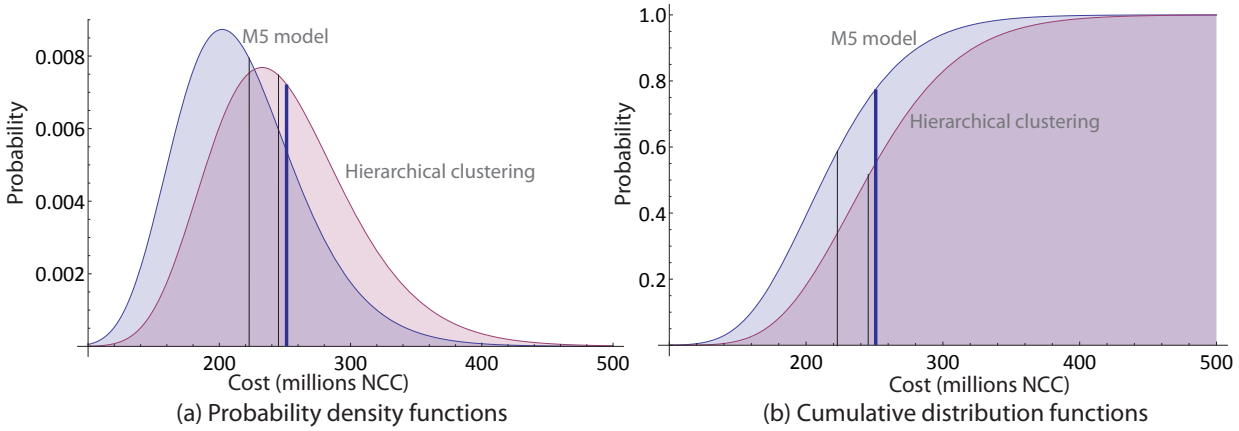

Figure 14: Probability density function (a) and cumulative distribution function (b) of the M5 model tree (blue) and hierarchical clustering (red) estimates of HNLMS Johan de Witt LPD cost.

Table 12 compares the actual development and production costs of HNLMS Rotterdam and Johan de Witt to the estimates generated by the M5 model tree and hierarchical clustering methods.

The percent error of the M5 model tree estimates (50th percentile) relative to the actual costs are -2% (under estimated) for HNLMS Rotterdam LPD and -16% (under estimated) for HNLMS Johan de Witt LPD. The percent error of the hierarchical clustering estimates (50th percentile) relative to the actual costs are 6% (over estimated) for HNLMS Rotterdam LPD and -4% (under

Table 12: Comparison of actual to estimated costs (millions NCC).

	M5 model tree estimate	Hierarchical clustering estimate	Actual
HNLMS Rotterdam	197.7	214.6	202.2
HNLMS Johan de Witt	212.3	243.9	253.7

estimated) for HNLMS Johan de Witt LPD. With respect to the fitted log-normal distributions, the actual costs lie at the 55th (HNLMS Rotterdam LPD) and 80th (HNLMS Johan de Witt LPD) percentiles for the M5 model tree results, and at the 39th (HNLMS Rotterdam LPD) and 57th (HNLMS Johan de Witt LPD) percentiles for the hierarchical clustering results. Table 13 lists the predicted costs for incremental percentiles of all the fitted log-normal probability density functions.

Table 13: Percentiles of the fitted log-normal density functions for the M5 model tree and hierarchical clustering estimated HNLMS Rotterdam LPD and Johan de Witt LPD costs (millions NCC).

	M5 Model Tree Distribution		Hierarchical Clustering Distribution	
Percentile	Rotterdam	Johan de Witt	Rotterdam	Johan de Witt
0.05	152.2	163.5	143.1	170.4
0.1	161.3	173.2	153.7	184.5
0.15	167.7	180.1	161.3	194.6
0.2	173.0	185.7	167.6	203.0
0.25	177.6	190.7	173.2	210.6
0.3	181.9	195.3	178.4	217.6
0.35	186.0	199.7	183.3	224.3
0.4	189.9	203.9	188.1	230.8
0.45	193.8	208.1	192.9	237.3
0.5	197.7	212.3	197.7	243.9
0.55	201.7	216.6	202.6	250.7
0.6	205.8	221	207.8	257.7
0.65	210.2	225.7	213.2	265.3
0.7	214.9	230.7	219.1	273.4
0.75	220.1	236.3	225.7	282.5
0.8	226.0	242.7	233.2	293.0
0.85	233.1	250.3	242.3	305.7
0.9	242.3	260.2	254.3	322.5
0.95	256.7	275.7	273.1	349.1

In retrospect, it is interesting to recall the earlier discussion on the negative coefficient of a ship's sailing range in the M5 model tree's linear regression models. It was noted that the majority

of the ships had a range under 770 hours, the impact on the estimated cost of these ships would be minimal while lowering estimates of ships with outlying sailing ranges (over 1200 hours). HNLMS Johan de Witt LPD's sailing range is listed as 833 hours. Substituting the median sailing range (444 hours) of the SAS-076 data set, effectively neutralizing the attribute, would result in a revised estimate of 253.9 million NCC, virtually matching the actual figure of 253.7 million. The M5 model tree estimate of HNLMS Rotterdam LPD less sensitive to this factor as its sailing range is 500 hours, already quite close to the median of 444 hours, neutralizing the attribute by substituting the median sailing range results in a revised estimate of 202.8 million.

CONCLUSION

This paper describes two novel approaches to cost estimation using known data mining algorithms. As a proof of concept, the approaches were applied in a blind ex post cost estimation exercise of the Netherlands' landing platform dock ships.

Both methods incorporate a multitude of cost driving factors that required the compilation of a multinational data set of dozens of somewhat similar ships. The data mining approaches allow for a greater variability in the input data set—variability that could be questioned when using traditional approaches. As with other parametric and analogy approaches, the fidelity of the estimation models are very dependent the data set, especially if the size of the data set is small. Both are "top down" approaches applicable in early design phases of the procurement cycle.

The parametric approach combined features of decision trees with linear regression models to both classify similar ships (based on attributes) and build piece-wise multivariate linear regression models. The attributes of HNLMS Rotterdam class ships were use to trace down the tree and as input to the resulting regression models which outputted a prediction.

As an analogy costing approach, hierarchical agglomerative cluster analysis, principal component analysis, and non-linear optimization was used to calculate a matrix of distances among the data set ships. These distances were then used to predict the cost of HNLMS Rotterdam class ships.

Despite a limited data set, the proof of concept results provide evidence that the methods can provide accurate estimates. The methods should be considered by the United States Office of the Assistant Secretary of the Navy (Research, Development and Acquisition) for generating cost estimates for the new America class large-deck amphibious assault (LHA) ships.

Notes

[1]The two ships of the Spanish belonging to the Galicia class are the Galicia (commissioned in 1998) and the Castilla (2001).

[2]Jane's Fighting Warships: http://jfs.janes.com

[3]Federation of American Scientists: http://www.fas.org/

[4]Navy Matters: http://www.navy-matters.beedall.com

[5]Forecast International: http://www.forecastinternational.com

[6]Wikipedia: The Free Encyclopedia: http://www.wikipedia.com

[7]WEKA project: http://www.cs.waikato.ac.nz/ml/weka/

[8]The program's default parameters for the M5 model tree algorithm were used.

References

Akaike, H. (1980). Likelihood and the Bayes procedure, *Trabajos de Estadística y de Investigación Operativa* **31**: 143–166.

Chen, Z. (2006). *Reduced-Parameter Modeling for Cost Estimation Models*, PhD thesis, Faculty of the Graduate School, University of Southern California.

Dobra, A. (2002). Secret: A scalable linear regression tree algorithm, *In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, pp. 481–487.

Friedman, N. (2005). *U.S. Amphibious Ships and Illustrated Design History*, 2nd edn, Naval Institute Press, Maryland, USA.

Goldberger, A. (1968). The interpretation and estimation of Cobb-Douglas functions, *Econometrica* **35**: 464–472.

Jolliffe, I. (2002). *Principal component analysis*, Springer New York.

Malerba, D., Esposito, F., Ceci, M. and Appice, A. (2004). Top-down induction of model trees with regression and splitting nodes, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(5): 612–625.

Miroyannis, A. (2006). *Estimation of Ship Construction Costs*, PhD thesis, Department of Mechanical Engineering, Massachusetts Institute of Technology.

NATO RTO SAS-054 Panel (2007). Methods and models for life cycle costing, *Technical Report RTO-TR-SAS-054*, NATO Research & Technology Organization.

Quinlan, J. (1992). Learning with continuous classes, *Proceedings AI'92 Singapore: World Scientific (Adams & Sterling Eds)*, pp. 343–348.

Ryan, T. (1997). *Modern Regression Methods*, 2 edn, John Wiley & Sons, Inc., New York, NY, USA.

Torgo, L. (2000). *Inductive Learning of Tree-based Regression Models*, PhD thesis, Universidade do Porto.

Torgo, L. (2002). Computationally efficient linear regression trees, *In Proceedings of International Federation of Classification Societies (IFCS) 2002: Classification, Clustering and Data Analysis: Recent Advances and Applications*.

Wang, Y. and Witten, I. (1997). Inducing model trees for continuous classes, *In Proceeding of the 9th European Conference on Machine Learning Poster Papers*, pp. 128–137.

Witten, I. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn, Morgan Kaufmann, Massachusetts, NE, USA.