Annex B1

Estimating the Acquisition Cost of the Royal Netherlands Navy Landing Platform Dock Ships: an Ex Post Analysis

NATO RTO Systems Analysis and Studies Task Group-076*

November 1, 2011

*Corresponding author contact: Dr. Bohdan L. Kaluzny, Defence Research and Development Canada (Bohdan.Kaluzny@drdc-rddc.gc.ca)

1.0 INTRODUCTION

1.1 Background

This paper documents the SAS-076 Task Group efforts in generating independent estimates of the development and construction costs of the Royal Netherlands Navy landing platform dock 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 landing platform dock 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 (HMS) Rotterdam, pennant number L800, was launched in 1997. The second ship of the class, HMS Johan de Witt (L801), was launched in 2007. The ships have a large helicopter deck and a well dock for large landing craft. The class was a joint design between the Netherlands and Spain[1]. HMS Rotterdam is pictured in Figure 1. Although considered the sister ship to

Figure 1: HMS Rotterdam L800 Landing Platform Dock Ship

HMS Rotterdam, HMS 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 HMS Johan de Witt's rank in class is set to one.

Following a key recommendation of SAS-054, this report documents two independent methods that were used in generating a cost estimate for HMS Rotterdam and Johan de Witt. The first method, classified as a parametric approach, employs Quinlan's M5 model tree algorithm [1], a system that combines features of decision trees with linear regression models. The second method is an analogy approach based on hierarchical cluster analysis. It considers multiple analogous systems and is void of subjectivity that is typically inherent

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

in analogy approaches. Both data mining approaches are data intensive—the SAS-076 Task Group compiled a database of 57 ships in 16 classes from 6 nations. 136 descriptive, technical, and cost attributes were gathered for each of the ships.

1.2 Objective

The objective of this report is to present two independent methods for generating estimates for the development and production cost of HMS Rotterdam and Johan de Witt LPDs. In an ex post exercise, the actual costs are revealed and compared to the estimates. The report presents a novel application of known data mining algorithms to the problem of ship cost estimation.

1.3 Scope

This report focuses on two data mining approaches: M5 model trees for parametric cost estimation, and hierarchical cluster analysis for costing by analogy. Limitations and assumptions specific to the parametric method are presented in Sections 3.0 and 4.0.

1.4 Outline

Section 2.0 details the multinational data that was gathered (Annex **??** contains data tables continuing from Section 2.0). The parametric cost estimation method is developed in Section 3.0 and applied to the data set. The analogy cost estimation method is developed in Section 4.0 and applied to the data set. The results for the predicted cost of HMS Rotterdam and Johan de Witt LPDs are presented in Sections 5.0 and in particular are compared to the actual costs. Section 6.0 compiles lessons identified, conclusions are presented in Section 7.0, and Annex **??** provides the algorithmic details of the approach.

2.0 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., [2]), 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 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);

[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

- 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 by the SAS-076 Task Group to be analogous to HMS Rotterdam LPD. The selection of ships for inclusion was primarily driven by the accessibility of costing information, which was limited to the information gathered from national sources by the SAS-076 Task Group representatives[7]. 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 decision as to which ships are similar or analogous is subjective.

The novel parametric approach to cost estimation using data mining (to be elaborated in Section 3.0) allows for, or can excel on, a greater variability in the input data set—variability that could be questioned when using traditional parametric approaches. The fidelity of parametric approaches also depends on the size of the input data set. In this case, the availability of ship cost data limited the size of the SAS-076 ship data set. For these reasons, the SAS-076 Task Group decided to include ships such as Sweden's Carlskrona LPD and Atle and Oden icebreakers—leading to the potentially useful data combination of ships of the "right purpose, wrong size" and "wrong purpose, right size" [3].

The SAS-076 Task Group also decided to omit the U.S. San Antonio class LPDs due to its well-documented high cost of ship construction: as of July 2007, the U.S. Navy had invested over 1.75 billion U.S dollars constructing the San Antonio LPD 17, over 800 million dollars over budget [4] (including a mid-construction transfer from Avondale shipyard in New Orleans to the Northrop Grumman shipyard in Mississippi due to poor performance by the former). According to the U.S. Government Accountability Office [5]: "the U.S. Navy's reliance on an immature design tool led to problems that affected all aspects of the lead ship's design. Without a stable design, work was often delayed from early in the building cycle to later, during integration of the hull". The LPD-17 cost was a clear outlier in comparison to the costs of the other ships. It was deemed that its cost was not accurately reflective of its technical attributes.

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

[7]For example, detailed technical information was available for the Italian San Giorgio class and the Spanish Galicia class LPDs, but cost information was unobtainable

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

2.1 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 are broken down into the categories as per Table 2. Tables 15 and 16 in Annex **??** detail the technical attributes (categories III to XIV of Table 2) of the 17 ship classes. Table 3 details the attributes as well as the input for HMS 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), milimeters (mm), megawatts (MW), knots (kts), hours (hrs), nautical miles (nmi), etc. Unknown or missing data is denoted by "?" entries. For conciseness, undefined technical acronyms appearing in Tables 2 or 3 are listed at the end of the document.

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

continued on next page

continued from previous page

Data Category & Element	Rotterdam LPD	Johan de Witt LPD
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
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

continued on next page

continued from previous page

Data Category & Element	Rotterdam LPD	Johan de Witt LPD
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	?	?

2.2 Cost Data

The SAS-076 ship data set includes ship costs expressed 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 are also expressed in various then-years[8], ranging from 1952 to 2009. The SAS-054 recommendations for cost normalization of multinational programs are summarized as follows:

[8] Amounts that include the effects of inflation or escalation, and/or reflect the price levels expected to prevail during the year at issue.

1. apply escalation formulae using indices from supplier nations to model changes due to national inflation prior to conversion to other currencies;

2. each nation should apply its own cost inflation model based on applicable data and nation specific economic advice that is pertinent to the particular system; and,

3. use exchange rates between the national currencies as the currency conversion mechanism.

The SAS-076 Task Group followed the SAS-054 guidelines to normalize the ship costs to a common currency and then-year. Various inflation rates were research, and in particular the Task Group was guided by U.S. Naval Sea Systems Command (NAVSEA) representatives who were not aware of any "publicly available total ship cost index that goes back to the 1950s" [6]. The U.S. Bureau of Labour Statistics [7] has a shipbuilding labour index, and there are various materiel indices that are applicable to shipbuilding, but the high-level data gathered in the SAS-076 data set is not detailed enough to compute a weighted overall composite index. Instead, inflation rate sources were obtained as follows:

- The SAS-076 U.S. representative from the Office of the Assistant Secretary of the Navy obtained the Historical Shipbuilding and Conversion, Navy (SCN) Total Obligational Authority Index used within the U.S. NAVSEA[9].

- To inflate the cost figures of ships built within countries of the European Union, the EuroStat Labour Inflation rates [8] specific to each country were applied.

- To inflate the cost of Canadian-built ships, the Defence Research & Development Canada Centre for Operational Research & Analysis Defence Economics Team recommended the Industrial Production Index (IPI) [9].

- Upon consultation with FOI—Swedish Defence Research Agency, the IPI was also used to inflate the cost of the Scandinavian-built ships for which pre-dated the available EuroStat indices.

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.

The SAS-076 Task Group log-transformed the cost of each ship as input to the estimation model. The models, to be presented in detail in Subsections 3.0 and 4.0, 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 log-normal distribution is a probability distribution of a random variable whose logarithm is normally distributed.

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 50 to 700 million NCC. The majority of total cost estimates for weapon-system acquisition programs modeled by the United States' Deputy Assistant Secretary of the Navy

[9]Provided under the condition that the indices would not be published.

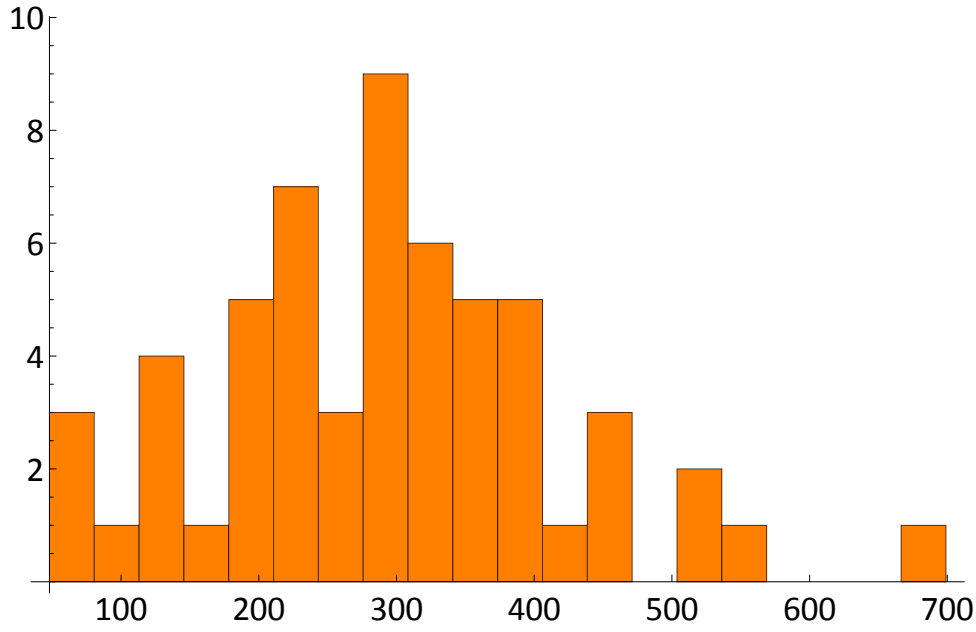

Figure 2: Histogram of normalized costs (millions NCC).

(Cost & Economics) are log-normally distributed and often skewed right [10]. The log-normal distribution is also justifiable choice when a variable is known to be the multiplicative product of many independent random variables each of which is positive (as is usually the case in shipbuilding). Beaulieu *et al.* [11] showed that the distribution of the sum of log-normal random variables can be approximated with a single log-normal distribution.

3.0 PARAMETRIC COST ESTIMATION

Parametric approaches to cost estimation use regression or other statistical methods to develop Cost Estimating Relationships (CERs). A CER is an equation used to estimate a given cost element using an established relationship with one or more independent variables.

There are several advantages of a parametric approach to cost estimation, including:

- it can capture major portions of an estimate quickly and with limited information (as is often the case in early phases of the procurement cycle);
- the CER is objective—it is based on consistent and quantitative inputs;
- once the CER is established it is easy to perform sensitivity analyses (determine the change in cost subject to changes in the independent input variables); and,
- CERs established using regression analyses include standard tests of validity, including a coefficient of correlation indicating the strength of association between the independent variables and the dependent variable in the CER.

A critical consideration in parametric cost estimating is the similarity of the systems in the underlying database, both to each other and to the system which is being estimated. 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 [12] 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 (the reader is referred to [13] for further information), linear regression is the most popular and easiest to understand.

The next 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.

3.1 The M5 Model Tree System

In 1992 Quinlan [1] pioneered the M5 system for learning models that predict numeric values. The M5 system combines features of decision trees with linear regression models. M5 model trees are similar to regression trees; a decision tree induction algorithm is used to build an initial tree, recursively splitting the data set based on the value of a chosen splitting attribute. The splitting attribute is selected to minimize the prediction error down each branch. Whereas the nodes of a regression tree each contain a constant value (prediction), each model tree node is a multivariate linear regression model. The attributes defining these regression models are the attributes that are involved in the tree's branching decisions.

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 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

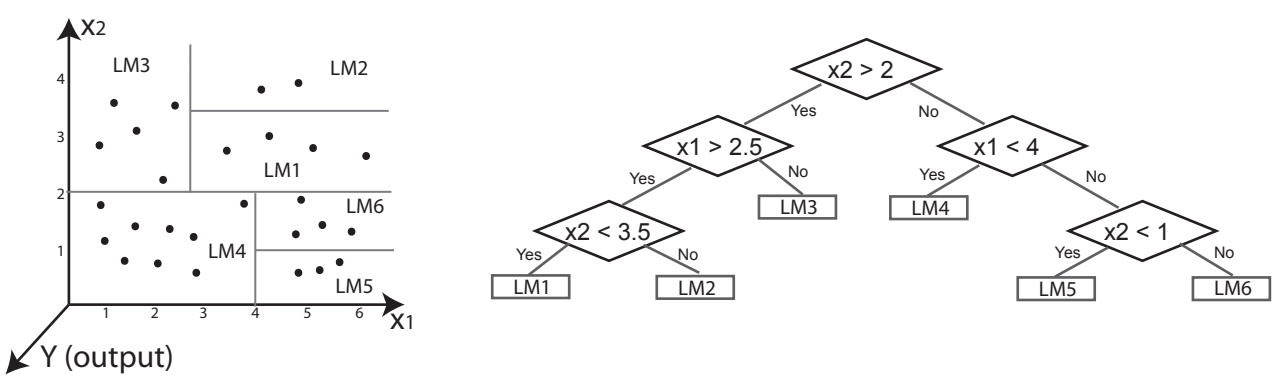

Figure 3: Example M5 model tree.

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 following paragraph synthesizes the M5 model tree algorithm. Full details are exposed in Annex **??**. 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.

3.1.1 Discussion

Regression trees are more accurate than straightforward linear regression, but the trees are often cumbersome and difficult to interpret [14]. Model trees are more sophisticated than regression trees as they approximate continuous functions by linear "patches"—M5 model trees are analogous to piecewise linear functions. 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. Regression trees will not predict values lying outside the range of learned training cases, while M5 model trees can extrapolate. The M5 model tree algorithm is optimized to both learn known cases and predict unknown cases. The trees are also smaller, easier-to-understand, and their average error values on the training data are lower.

M5 model trees tackle the difficulties in considering a multitude of cost driving factors; they determine the right set of independent variables (ship attributes) by construction. The effect of the M5 system is somewhat similar to the use of indicator variables in standard linear regression analysis, with the key difference being that appropriate indicators are identified and included based on an optimization algorithm.

M5 model trees have been shown to excel even when limited data is available [15] or learn efficiently from large data sets. They can handle data sets which include 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).

The use of M5 model trees for numeric prediction has increased since comprehensive descriptions, implementations, and refinements of Quinlan's method became available [15–19]. Recently Chen [20] discussed the benefits of the system for estimating the cost of software development.

Neural networks are commonly used for predicting numeric quantities, but suffer from the disadvantage of producing opaque structures that mask the nature of the solution. Their arbitrary internal representation means that there can be variations between networks trained on the same data. In comparison, the M5 system is transparent and the model tree construction is repeatable. The final tree and linear regression models are straightforward and clear.

3.2 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 [21]. Witten and Frank's textbook [14] provides 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[10], 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, SAS-076 recommends noting the anomalies and discussing them 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 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, the quantity that gives the quality of a least squares fitting to the original data. Let x_i be the actual cost of ship i and y_i the predicted cost (using the M5 model tree), then the coefficient of correlation, R, of the M5 system is calculated as

$$R = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}, \tag{1}$$

where $n = 57$, the number of ships in the data set, and each of the sums are over these 59 ships. By design R can range from -1 to $+1$ with $R = 1$ indicating a perfect positive linear correlation. A correlation greater than 0.8 is generally described as strong [13]. The value R^2, known as the coefficient of determination, is a

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

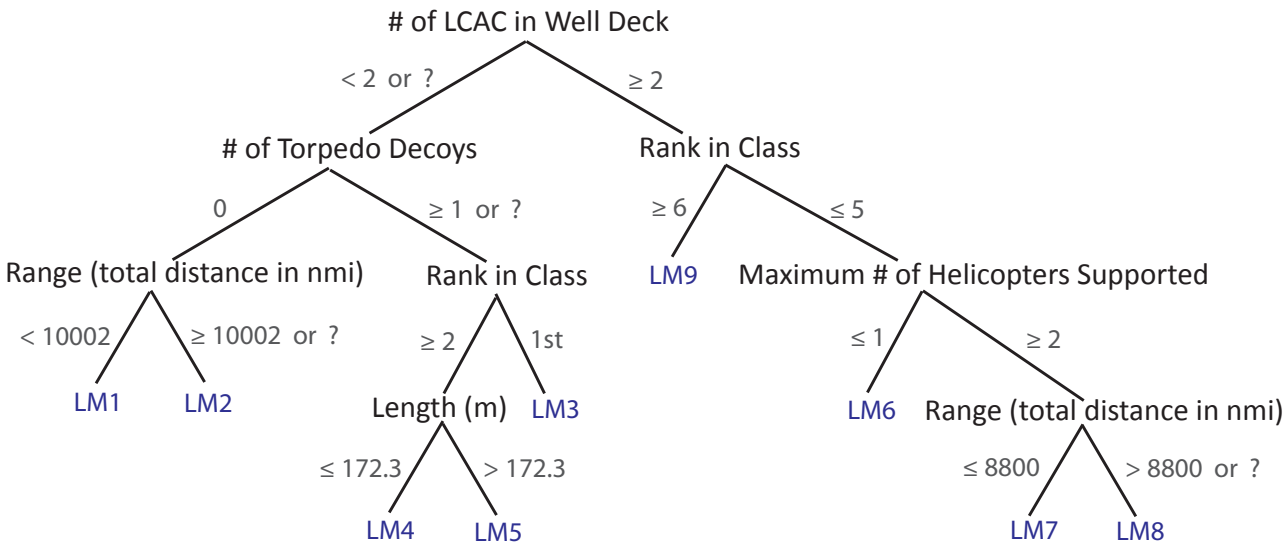

Figure 4: M5 model tree applied to the NATO RTO SAS 076 ship data set.

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 $R^2 = 0.92$. 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.

Another useful indicator is the mean absolute percent error, as it quantifies the the amount by which the estimated cost differs from the actual cost. The mean absolute percent error is computed as as follows:

$$\frac{1}{57}\sum_{i=1}^{i=57}\frac{|x_i - y_i|}{x_i} \tag{2}$$

The mean absolute percent error for the M5 model tree applied to the SAS-076 ship data set is 12%. The standard deviation, indicating the variation that M5 model tree predictions have from the actual costs, is computed as

$$\sqrt{\frac{\sum_{i=1}^{57}(x_i - y_i)^2}{57}}, \tag{3}$$

for a value of 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. 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.

Table 4: M5 model tree linear regression models

LM1		LM2	
Log(Cost) =	7.4297	Log(Cost) =	7.4208
	- 0.0112 × rank in class		- 0.0112 × rank in class
	+ 0.0045 × length (m)		+ 0.0045 × length (m)
	- 0.0002 × range (sailing time in hrs)		- 0.0002 × range (sailing time in hrs)
	+ 0.0445 × # of LCAC in well deck		+ 0.0445 × # of LCAC in well deck
	+ 0.1104 × # of torpedo decoys		+ 0.1104 × # of torpedo decoys
LM3		LM4	
Log(Cost) =	7.6222	Log(Cost) =	7.7567
	- 0.0167 × rank in class		- 0.0172 × rank in class
	+ 0.0041 × length (m)		+ 0.0032 × length (m)
	- 0.0002 × range (sailing time in hrs)		- 0.0002 × range (sailing time in hrs)
	+ 0.0445 × # of LCAC in well deck		+ 0.0445 × # of LCAC in well deck
	+ 0.0659 × # of torpedo decoys		+ 0.0659 × # of torpedo decoys
	+ 0.00001 × range (distance in nm)		
LM5		LM6	
Log(Cost) =	7.7912	Log(Cost) =	7.9846
	- 0.0170 × rank in class		- 0.0245 × rank in class
	+ 0.0030 × length (m)		+ 0.0038 × length (m)
	- 0.0002 × range (sailing time in hrs)		- 0.0003 × range (sailing time in hrs)
	+ 0.0445 × # of LCAC in well deck		+ 0.0300 × # of LCAC in well deck
	+ 0.0659 × # of torpedo decoys		+ 0.0343 × # of torpedo decoys
LM7		LM8	
Log(Cost) =	8.1461	Log(Cost) =	8.3575
	- 0.0361 × rank in class		- 0.0312 × rank in class
	+ 0.0035 × length (m)		+ 0.0024 × length (m)
	- 0.0003 × range (sailing time in hrs)		- 0.0003 × range (sailing time in hrs)
	+ 0.0300 × # of LCAC in well deck		+ 0.0300 × # of LCAC in well deck
	+ 0.0343 × # of torpedo decoys		+ 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		

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
# 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

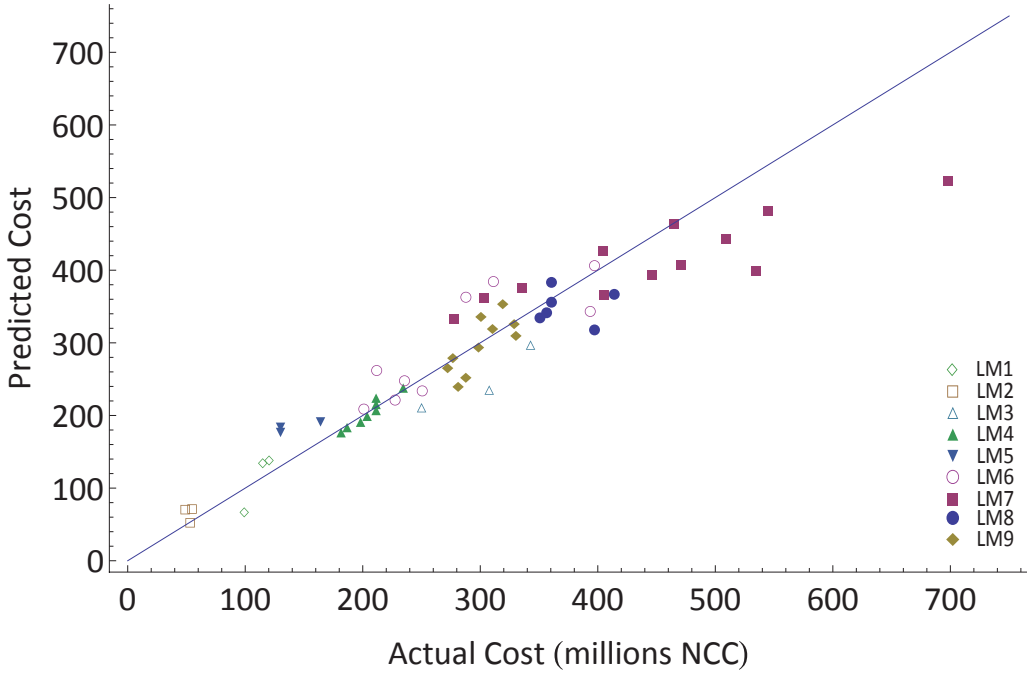

Figure 5: M5 model tree: correlation plot of actual vs. predicted ship costs (millions NCC).

3.2.1 Comparison to Linear Regression Models

The best CER returned by applying simple linear regression on the SAS-076 data set[11] is

$$Log_{10}(cost) = 6.95 + 0.01 \times \text{ length of ship (in meters)}, \tag{4}$$

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 information criterion [22]), yields the CER

$$\begin{aligned} Log_{10}(cost) = 5.7368 &- 0.0224 \times \text{ rank in class} \\ &+ 0.0121 \times \text{ length (in meters)} \\ &+ 0.0338 \times \text{ beam (in meters)} \\ &+ 0.1071 \times \text{ draught (in meters)} \\ &- 0.0001 \times \text{ full load displacement (in tonnes)} \\ &+ 0.0012 \times \text{ crew size} \\ &- 0.0876 \times \text{ number of propeller shafts} \\ &- 0.0239 \times \text{ number of guns of calibre } \geq 75, \end{aligned} \tag{5}$$

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 HMS Rotterdam and Johan de Witt LPDs are presented in Section 5.0.

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 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:

$$\begin{aligned} Log_{10}(cost) = 8.6376 &- 0.0651 \times \text{ rank in class} \\ &+ 0.0637 \times \text{ number of LCAC.} \end{aligned} \tag{6}$$

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 6 in place of LM7 would leave adjacent linear models LM6, LM7, and LM8 sharply discontinuous. The M5 system applies a smoothing process (see Annex **??**) to construct piece-wise linear regression models[12].

[11] Nominal attributes and attributes for which data was missing was omitted.

[12] Experiments by Wang and Witten [15] show that this smoothing substantially increases the accuracy of predictions of unseen cases.

4.0 COST ESTIMATION 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 system [23]. There must be 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. Subject matter experts are required to make a subjective evaluation of the differences between the new system of interest and the historical system. Subjectively chosen complexity factors are often used to adjust the analogous system's cost to produce an estimate. The credibility of the estimate for the new system may be undermined if the adjustment factors are not substantiated—this is a key disadvantage of the traditional analogy method.

In the next section 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.

4.1 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.

4.2 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 ships and outputs a single value indicating the distance, or similarity, between the two ships. Formally, define

$$d_{ijk} = \text{ distance between ship } i \text{ and } j \text{ with respect to attribute } k. \tag{7}$$

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

$$d_{ij} = \sqrt{\sum_{k \in A} d_{ijk}^2} \tag{8}$$

where A is the subset of attributes considered. Figure 7 illustrates the dendrogram resulting from the hierarchical clustering of ships using a the simple distance function (equation (8)) and a set of 123 attributes including descriptive attributes (I.e), (I.f), (II.h) and the complete set of technical attributes as per Table 2. The dendrogram indicates that HMS Rotterdam is best grouped with HMS Johan de Witt. Both Netherlands LPDs are then clustered with the French Mistral class AASs, namely the Mistral, Tonnerre, and Dixmude. The next closest ship to this agglomerative cluster is the United Kingdom's Ocean LPH, and so on.

Computing the distances between all pairs of ships using equation (8), 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

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

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 $R = 0.48$ (coefficient of correlation as calculated by Equation 1) and $R^2 = 0.23$. The mean absolute percent error is 49% and the standard deviation of 112 million NCC. The approach does poorly in learning the known ship costs.

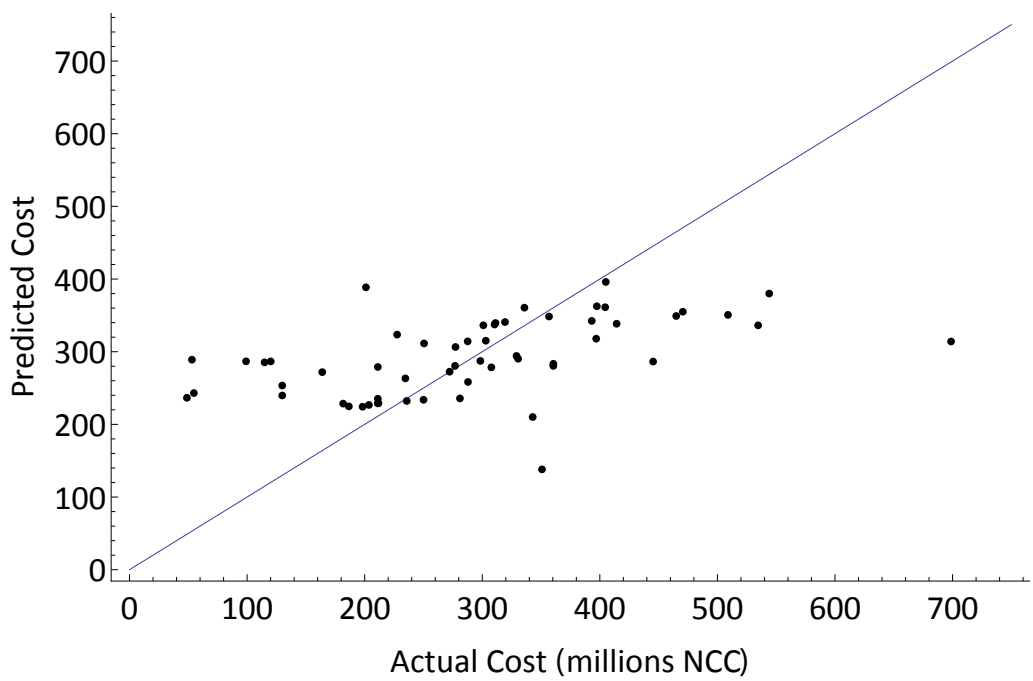

Figure 6: Hierarchical clustering with simple distance function: correlation plot of actual vs. predicted ship costs (millions NCC).

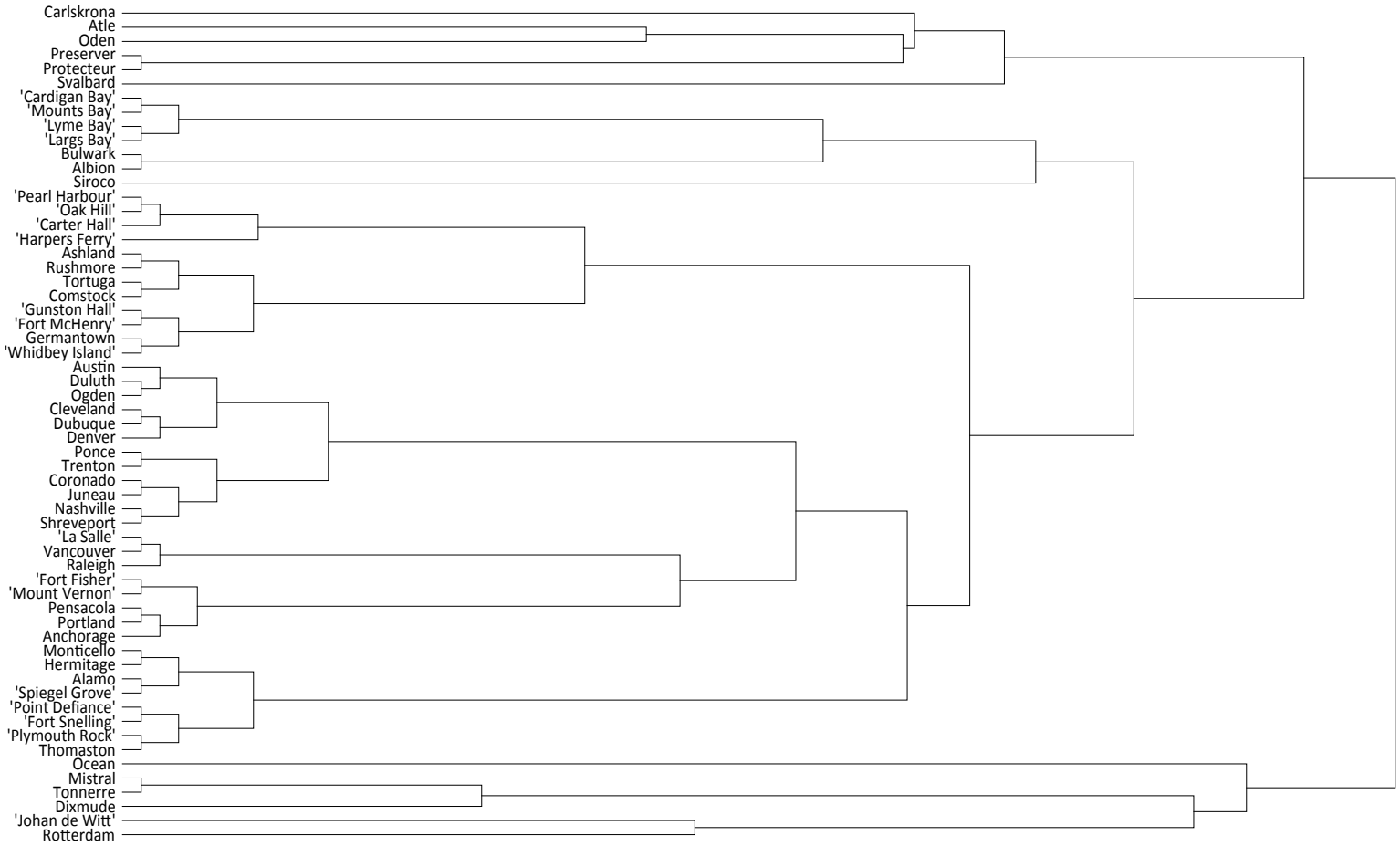

Figure 7: Dendrogram illustrating the arrangement of the clusters produced by the hierarchical clustering of ships (simple distance function).

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

$$w_k = \text{ weight of attribute } k. \tag{10}$$

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

$$\hat{d}_{ij} = \sqrt{\sum_{k \in A} (w_k \cdot d_{ijk})^2}, \tag{11}$$

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

$$\hat{C}_i = \sum_{j \neq i} \frac{C_j}{\hat{d}_{ij}^2} \cdot \frac{1}{\sum_{j \neq i} \frac{1}{\hat{d}_{ij}^2}} \tag{12}$$

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,

$$\text{minimize} \sum_{i=1}^{57} \left(C_i - \hat{C}_i\right)^2. \tag{13}$$

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 was selected. While there are numerous attribute selection algorithms (see [14]), 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[13].

The WEKA software tool [21] 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. Table 8 lists the macro-attributes and the respective percentage of data variability (cumulative) each accounts for.

Each macro-attribute is a linear combination of the original attributes. For example, macro-attribute A2 is

[13]PCA involves the calculation of the eigenvalue decomposition of a data covariance matrix. Dimensionality reduction is accomplished by choosing enough eigenvectors to account for some percentage of the variance in the original data. PCA can be thought of as revealing the internal structure of the data in a way which best explains the variance in the data. If a multivariate dataset is visualized as a set of coordinates in a high-dimensional data space (1 axis per variable), PCA supplies the user with a lower-dimensional picture, a "shadow" of this object when viewed from its (in some sense) most informative viewpoint.

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%

$$
\begin{aligned}
\text{A2} = {} & 0.204 \times \text{length} \\
& + 0.196 \times \text{beam width} \\
& + 0.183 \times \text{vehicle space} \\
& + 0.18 \times \text{\# of expeditionary fighting vehicles} \\
& + 0.165 \times 1 \text{ if has a well deck, otherwise } 0 \\
& + 0.165 \times \text{width of the well deck} \\
& + 0.164 \times \text{length of the well deck} \\
& + 0.159 \times \text{\# of large personnel landing craft} \\
& + 0.156 \times \text{\# of Chinook helicopters supported} \\
& + 0.155 \times \text{full load displacement} \\
& + 0.153 \times \text{\# of combat data systems} \\
& + 0.150 \times \text{light load displacement} \\
& + 0.144 \times \text{well deck capacity} \\
& + 0.143 \times \text{\# of elevators} \\
& + 0.142 \times \text{vehicle fuel capacity} \\
& \text{etc.}
\end{aligned}
\tag{14}
$$

(Only the top 15 attributes—in terms of PCA coefficient size—are enumerated in Equation 14.) Attempting to find solutions to the mathematical program (13) with $|A| = 16$ macro-attributes 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. Annex **??** shows the resulting top ten macro-attributes for the 57 SAS-076 ships as well for HMS Rotterdam and Johan de Witt LPDs.

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

The hierarchical cluster analysis using the weighted distance function on the top ten macro-attributes determined by PCA is visualized in Figure 8. The figure illustrates the resulting dendrogram indicating that HMS Rotterdam LPD is grouped with the United Kingdom's Albion class LPDs and Largs Bay class LSDs, followed by France's Siroco LSD, etc. HMS Johan de Witt LPD is clustered with Sweden's Svalbard icebreaker, France's Mistral class AAS, United Kingdom's Ocean LPH, and so on[14].

[14] As expected, ships within the same class (but different rank) are closely grouped together.

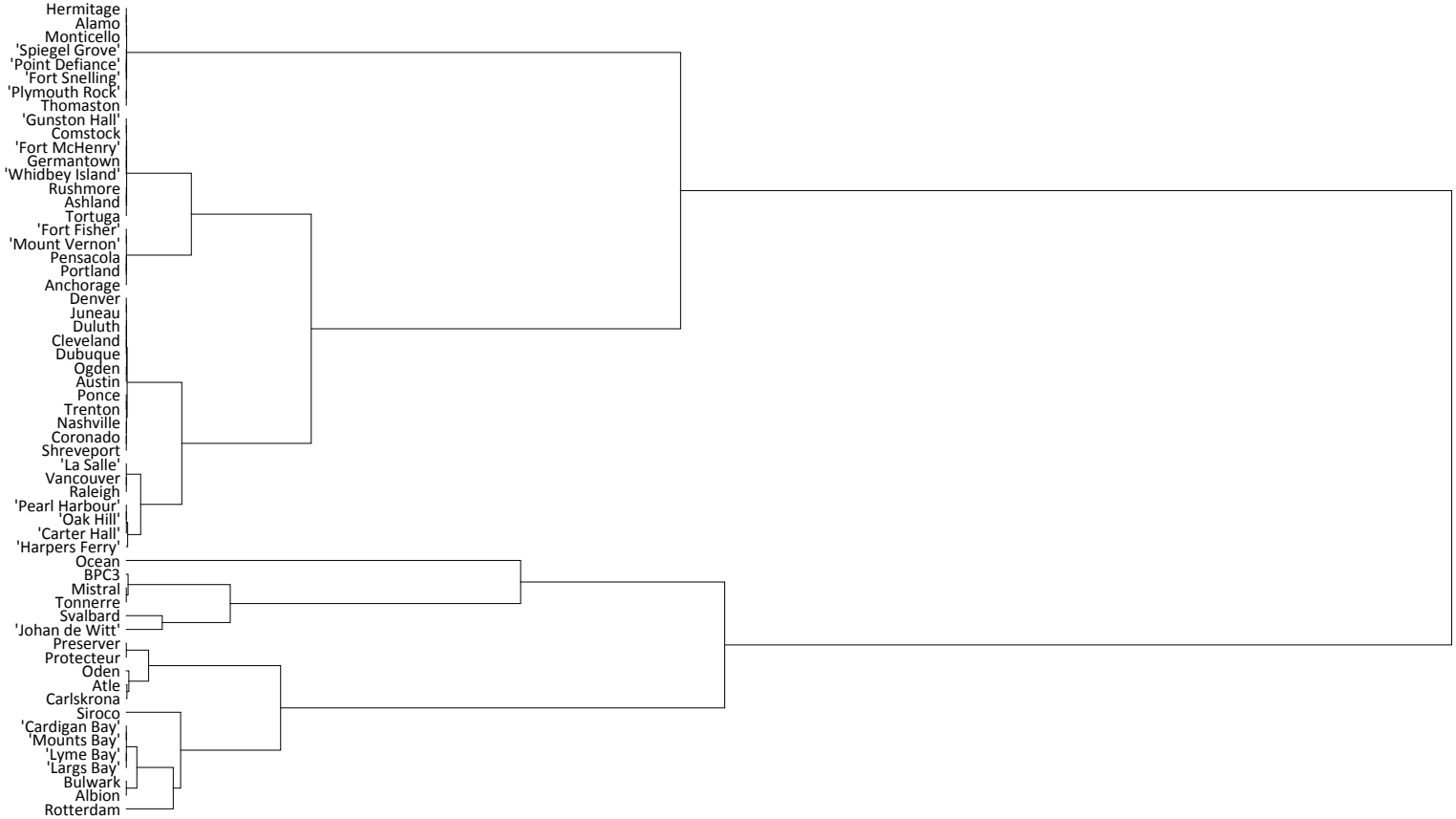

Figure 8: Dendrogram illustrating the arrangement of the clusters produced by the hierarchical clustering of ships (weighted distance function).

Figure 9 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 $R = 0.93$ (coefficient of correlation as calculated by Equation 1) 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.

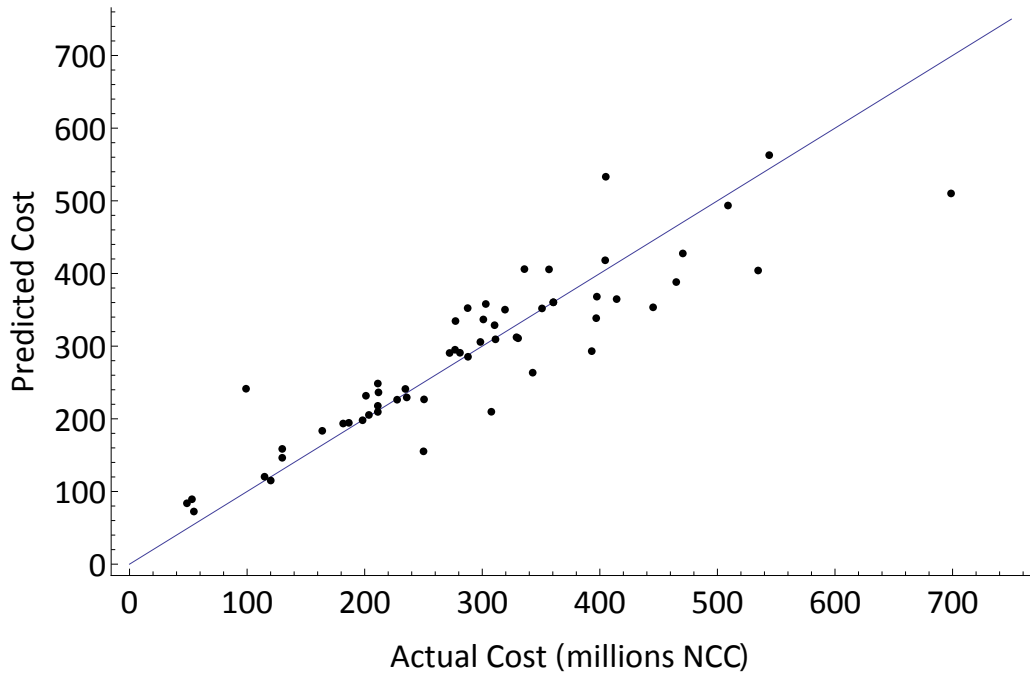

Figure 9: Weighted hierarchical clustering: correlation plot of actual vs. predicted ship costs (millions of NCC).

4.3 Discussion

Figure 9 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 Section 3.2. 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.

5.0 RESULTS

The results of the M5 model tree system are described in Subsection 5.1. The point estimates obtained using the simple linear regression (Equation 4) and multiple linear regression (Equation 5) CER models are revealed in Subsection 5.2. Subsection 5.3 details the results of the hierarchical clustering method, Subsection 5.4 discusses which result should be considered the primary estimate, and in Subsection 5.5 the results of both models are are compared with the actual production and development costs (revealed post estimation).

5.1 M5 Model Tree Results

The technical specifications of HMS Rotterdam LPD (Table 3) were used to trace down the M5 model tree depicted in Figure 4. In particular, the input data indicates that HMS 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 HMS 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 and distance (nm), number of LCAC, and number of torpedo decoy systems as independent variables. Using the LM3 model, the predicted development and production cost of HMS Rotterdam LPD is 197.7 million NCC. The predicted cost of HMS Johan de Witt LPD is 212.3 million NCC[15].

Figures 10 and 11 illustrate the log-normal probability density and cumulative distribution functions for the M5 model tree estimates for the cost of HMS Rotterdam LPD and HMS Johan de Witt LPD.

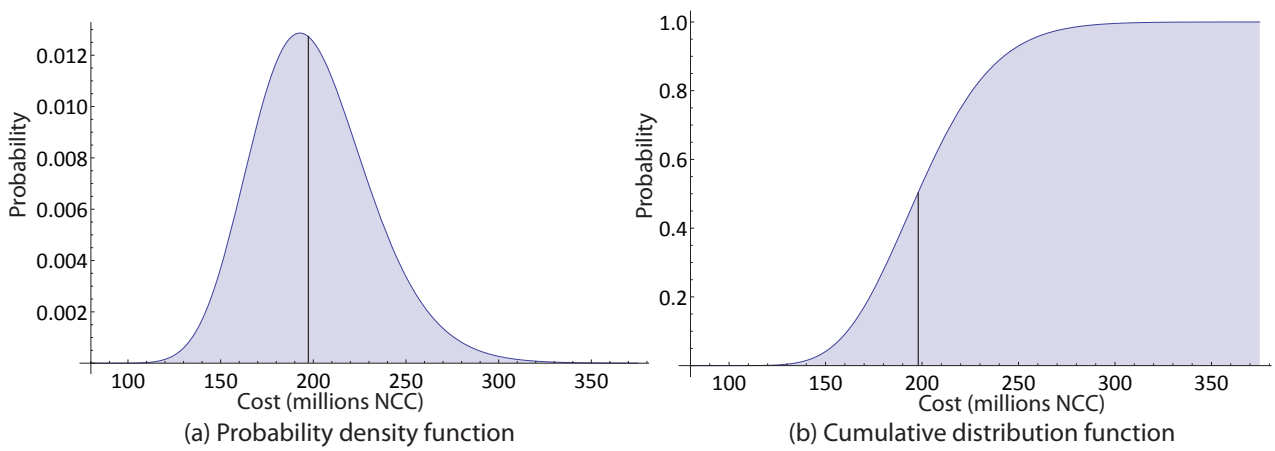

Figure 10: Probability density function (a) and cumulative distribution function (b) of the M5 model tree estimate of HMS Rotterdam LPD cost.

Table 10 lists the predicted costs for incremental percentiles of the fitted log-normal probability density functions. The predicted costs coincide to the 50th percentile of the respective log-normal probability distribution functions. This effect is explained by Goldberger [24]: 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

[15]HMS Rotterdam LPD and HMS Johan de Witt LPD differ in range (distance and sailing time in hours) and length—two of the independent variables in LM3.

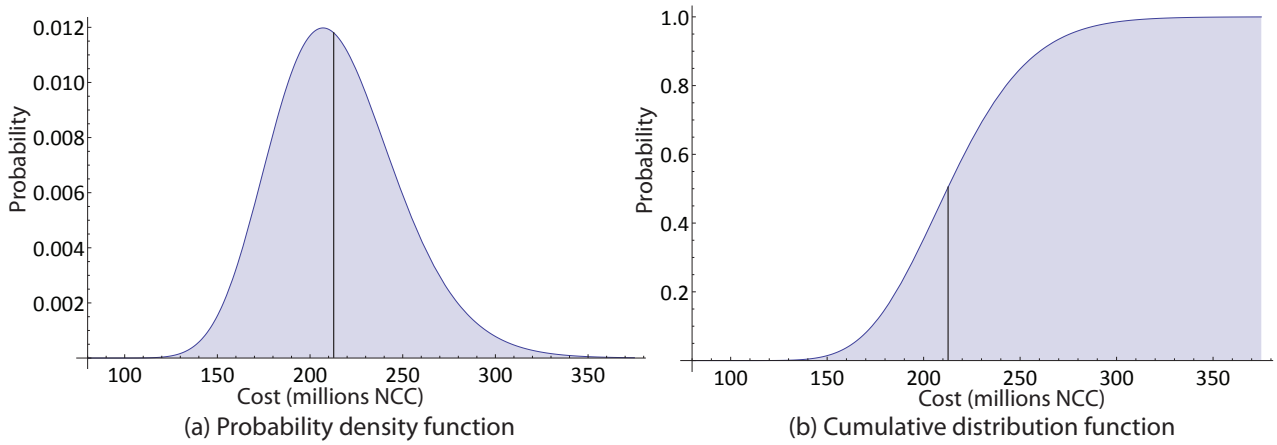

Figure 11: Probability density function (a) and cumulative distribution function (b) of the M5 model tree estimate of HMS Johan de Witt LPD cost.

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 HMS Rotterdam LPD and HMS Johan de Witt LPD respectively.

Table 10: Percentiles of the fitted log-normal density functions for the M5 model tree estimated HMS Rotterdam LPD and HMS Johan de Witt LPD costs (millions NCC).

Percentile	HMS Rotterdam	HMS Johan de Witt
0.05	152.2	163.5
0.10	161.3	173.2
0.15	167.7	180.1
0.20	173.0	185.7
0.25	177.6	190.7
0.30	181.9	195.3
0.35	186.0	199.7
0.40	189.9	203.9
0.45	193.8	208.1
0.50	197.7	212.3
0.55	201.7	216.6
0.60	205.8	221.0
0.65	210.2	225.7
0.70	214.9	230.7
0.75	220.1	236.3
0.80	226.0	242.7
0.85	233.1	250.3
0.90	242.3	260.2
0.95	256.7	275.7

5.2 Linear Regression Results

Using simple linear regression CER, Equation 4, the estimate for HMS Rotterdam LPD is 219.2 million NCC. The estimate for HMS Johan de Witt LPD is 289.7 million NCC. Using the multiple linear regression CER, Equation 5, the estimate for HMS Rotterdam LPD is 158.9 million NCC. The estimate for HMS Johan de Witt LPD is 201.4 million NCC.

5.3 Hierarchical Cluster Analysis Results

The technical specifications of HMS Rotterdam LPD (Table 3) were mapped to the ten attributes selected by the principal component analysis (see Annex **??**). Using the optimized macro-attribute weights determined in Section 4.2, the normalized relative distances of HMS Rotterdam LPD to the other ships are listed in Table 11 (the distances have been normalized so that the furthest ship has a distance of 1). The resulting hierarchical clustering cost estimate for HMS Rotterdam LPD is 214.6 million NCC, and 243.9 million NCC for HMS Johan de Witt LPD.

Figures 12 and 13 illustrate the log-normal probability density and cumulative distribution functions for the hierarchical cluster estimates for the cost of HMS Rotterdam LPD and HMS Johan de Witt LPD. Table 12 lists the predicted costs for incremental percentiles of the fitted log-normal probability density functions. The mean of the presented log-normal distributions are 219.8 million NCC and 249.8 million NCC for HMS Rotterdam

Table 11: 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		

LPD and HMS Johan de Witt LPD respectively.

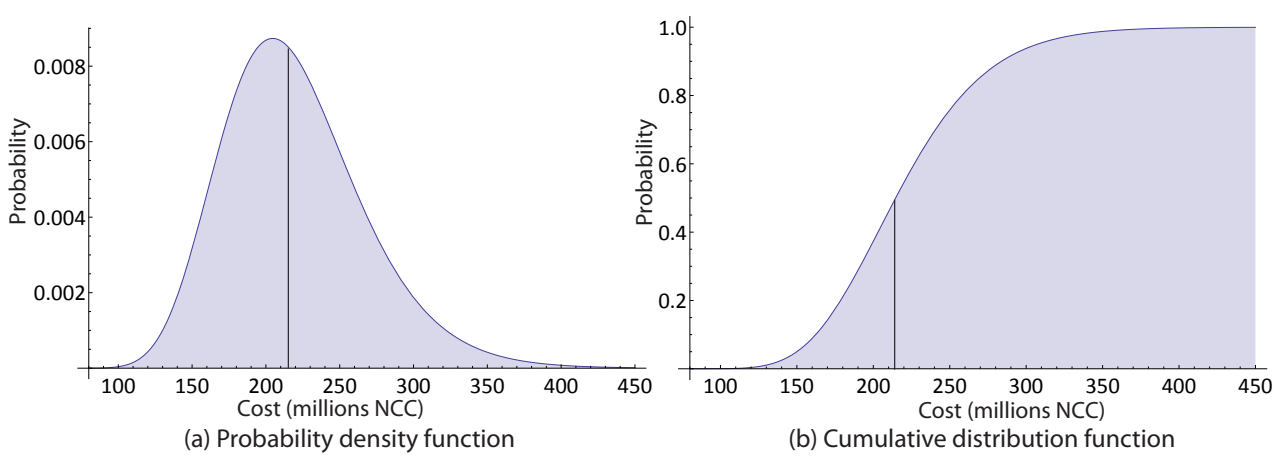

Figure 12: Probability density function (a) and cumulative distribution function (b) of the hierarchical clustering estimate of HMS Rotterdam LPD cost.

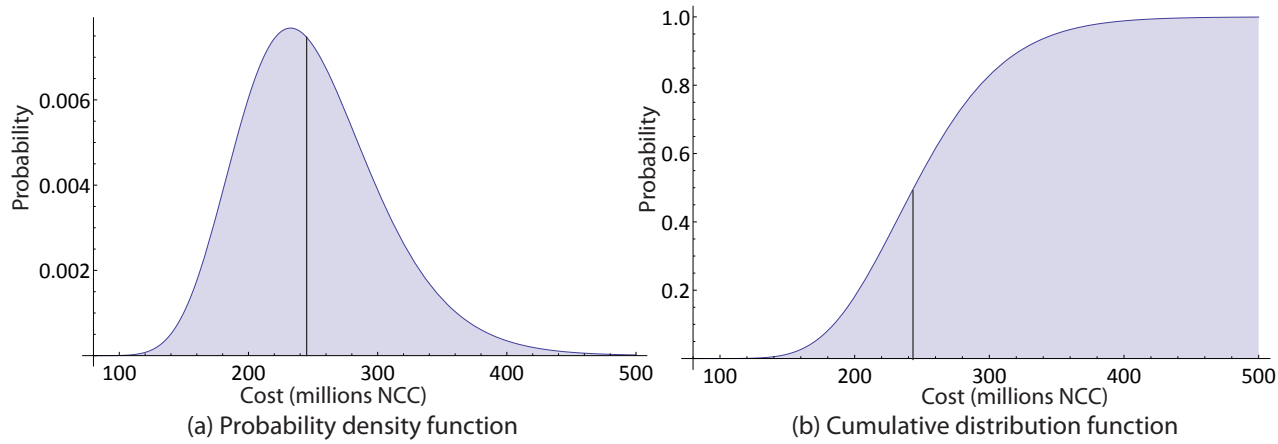

Figure 13: Probability density function (a) and cumulative distribution function (b) of the hierarchical clustering estimate of HMS Johan de Witt LPD cost.

Table 12: Percentiles of the fitted log-normal density functions for the hierarchical clustering estimated HMS Rotterdam LPD and HMS Johan de Witt LPD costs (millions NCC).

Percentile	HMS Rotterdam	HMS Johan de Witt
0.05	143.1	170.4
0.10	153.7	184.5
0.15	161.3	194.6
0.20	167.6	203.0
0.25	173.2	210.6
0.30	178.4	217.6
0.35	183.3	224.3
0.40	188.1	230.8
0.45	192.9	237.3
0.50	197.7	243.9
0.55	202.6	250.7
0.60	207.8	257.7
0.65	213.2	265.3
0.70	219.1	273.4
0.75	225.7	282.5
0.80	233.2	293.0
0.85	242.3	305.7
0.90	254.3	322.5
0.95	273.1	349.1

5.4 Discussion

The estimates generated by the M5 model tree are considered to be the primary estimates for HMS Rotterdam and Johan de Witt, the hierarchical clustering estimates are 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 to 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 original data set's variability. In comparison, the M5 model tree algorithm is computationally superior as it efficiently learns from large data sets.
- 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 13 synthesizes the predictions and compares properties of the M5 model tree and hierarchical clustering methods.

Table 13: Comparison of the M5 model tree and hierarchical clustering methods and their estimates.

	M5 model tree	Hierarchical clustering
HMS Rotterdam estimate	197.7M NCC	214.6M NCC
HMS Johan de Witt estimate	212.3M 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	✓	×

In Subsections 5.1 and 5.3, the cost estimates outputted by the M5 system and the hierarchical clustering method were fitted to log-normal distributions. The probability distributions of the cost and the associated percentile breakdowns provide decision makers the ability to choose a budget based on a risk tolerance. The risk-averse decision maker will choose a higher budget amount to minimize the budget risk—the probability that the actual cost of a weapon system acquisition program will end up exceeding a given budget.

5.5 Ex Post Revelation

The Royal Netherlands Navy revealed the actual development and production costs of HMS Rotterdam and Johan de Witt LPDs to the NATO RTO SAS 076 Task Group once the cost estimates were established. The costs were then normalized to the fictitious notional common currency. Figures 14 and 15 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 HMS Rotterdam LPD and Johan de Witt LPD. Respecting the wishes of the Netherlands, even the exact fictitious costs are not made more explicit.

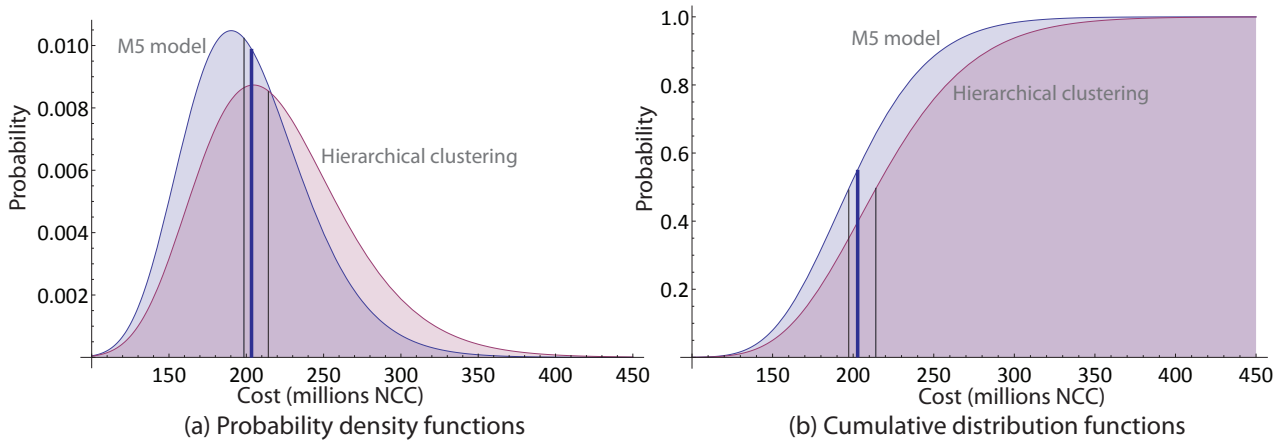

Figure 14: Probability density function (a) and cumulative distribution function (b) of the M5 model tree (blue) and hierarchical clustering (red) estimates of HMS Rotterdam LPD cost.

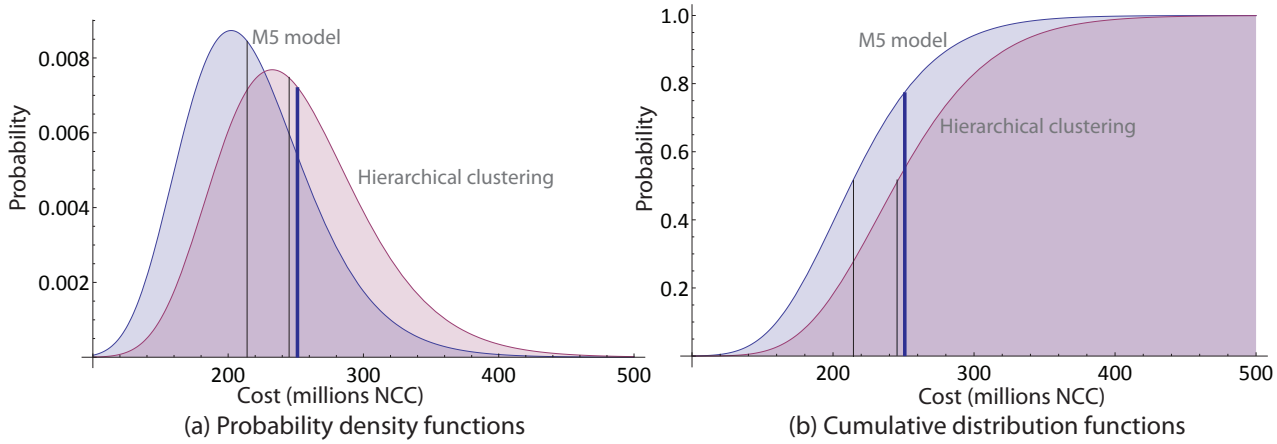

Figure 15: Probability density function (a) and cumulative distribution function (b) of the M5 model tree (blue) and hierarchical clustering (red) estimates of HMS Johan de Witt LPD cost.

Table 14 compares the actual development and production costs of HMS Rotterdam and HMS 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 HMS Rotterdam LPD and -16% (under estimated) for HMS Johan de Witt LPD. The percent error of the hierarchical clustering estimates (50th percentile) relative to the actual costs are 6% (over estimated) for HMS Rotterdam LPD and -4% (under estimated) for HMS Johan de Witt LPD. With respect to the fitted log-normal distributions,

Table 14: Comparison of actual to estimated costs (millions NCC).

	M5 model tree estimate	% error	Hierarchical clustering estimate	% error
HMS Rotterdam	197.7	-2%	214.6	6%
HMS Johan de Witt	212.3	-16%	243.9	-4%

the actual costs lie at the 55th (HMS Rotterdam LPD) and 80th (HMS Johan de Witt LPD) percentiles for the M5 model tree results, and at the 39th (HMS Rotterdam LPD) and 57th (HMS Johan de Witt LPD) percentiles for the hierarchical clustering results.

In retrospect, it is interesting to recall the discussion of Section 3.2 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). HMS 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, which is within 1% of the actual cost. The M5 model tree estimate of HMS 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.

6.0 LESSONS IDENTIFIED

This section intends to capture the feedback and lessons identified on estimating acquisitions costs of the Netherlands' Landing Platform Dock Ships Rotterdam and Johan de Witt. The section is divided in six parts: general management of the estimating process; aims and cost boundary; data and assumptions; methods, models and tools; risk and uncertainty; and analysis and presentation of results.

6.1 General Management of the Estimating Process

Initial Plan

SAS-076 started with a plan describing the activities and captured this in a Project Initiation Document (PID). Originally it was described as follows:

> *For the acquisition stage a top-down approach will be adopted, using the information in the DADD and applying the parametric method. As a second estimating method it is possible to apply the analogy method using public available data (Internet, NAO (UK) report, etc).*

Initially the cost-estimating approach consisted of two iterations. The first iteration was to estimate high-level cost elements of the cost breakdown structure. This would result in many assumptions; these assumptions would need to be checked by the Royal Netherlands (RNL) Navy. It would have been used to identify cost drivers. The second iteration was to go into more details for the costs drivers, if additional data was available. The main milestones for the first iteration were:

- Milestone 1: Agreement on cost elements to be calculated - 16/01/2009.

- Milestone 2: Release of the data and assumptions list - 26/06/2009.
- Milestone 3: Release of first report - 22/01/2010.

The main Milestones for the second iteration were:

- Milestone 4: Agreement on new cost elements to be calculated - 22/01/2010.
- Milestone 5: Release of the updated data and assumptions list - 09/07/2010.
- Milestone 6: Release of second iteration report - 01/10/2010.

Actual Implementation of the Initial Plan

The following steps were actually followed after having established the PID:

- Agreement on the Cost Breakdown Structure (CBS): follow ANEP 41 (SWBS) and the CBS from SAS-054 report;
- Developed a Data and Assumptions Document Definition (DADD) based on the data found in public sources, participants' inputs and the RNLN;
- Investigated software tools to be used; and,
- Decision on methods to be used: analogy and parametric.

Canadian participants proposed to use a data mining parametric method based on model regression trees and a new analogy method based on clustering analysis. The proposal stemmed from SAS-054 Section 4.4 recommendation that *"The best cost estimating method is one that makes the best use of the data available. It is therefore recommended to employ a method that will provide as much detail as the availability of the input data will allow. Therefore, the availability of data is a major factor in the estimator's choice of estimating method."* The team decided to follow the Canadian proposal and to deviate from the original plan. By adopting this approach the second iteration was considered not applicable any more. In addition, the SAS-076 started to write the report when the first decisions were made and when the cost estimation models were developed. During the estimating process SAS-076 decided not to further develop the DADD as a separate document. Data used and the assumptions made are included in the final report.

Lessons Identified

The group did not work continuously on this project, causing additional challenges in the estimates and in the management of some important issues. The SAS-076 group considered the estimating guidelines and the tasks described in SAS-054 of the management process to be useful. The group tried to follow the SAS 054 guidelines the work as much as possible, however the group was restricted by the limited available resources and competence available in the group or available within the group members professional network. This affected, in a certain extent, the work process. The conclusion is that following the SAS 054 guidelines could be followed, however it takes some time and resources. Therefore the resources and the budget required to do an estimate according to the SAS-054 guideline should be carefully planned before the actual work starts.

6.2 Definition of the Aims and Cost Boundary

Team's approach

The customer for the results of the study is the Royal Netherlands Navy. As this estimate is also used to apply the guideline developed by SAS-054, NATO RTO can be considered as a (secondary) customer. The aim was to practice the guidelines. For the acquisition cost estimation the objective was to compare the results from the SAS-076 analysis with the actual acquisition costs.

Lessons Identified

The aim and the objective of the study are fundamental for the way the cost analysis is conducted.

6.3 Data and Assumptions

Team's Approach

SAS-076 gathered data from ships similar to the Netherlands LPD's. The group applied some subjectivity in selecting the ships for gathering data, but the group did not limit themselves to ships very close to the Netherlands' LPDs. Civilian ships like icebreakers were also included because data related to these ships may have information relevant for the Netherlands' LPDs. The group assumed accuracy in the data. Engineers or shipbuilders were used sparingly in validating the data. Rather, participants were asked to validate their national data. SAS-076 had to deal with some limitations:

- No distinction was possible between development and production costs. Not all data elements could be found; however the methodologies proposed for acquisition handled missing values.
- Not all cost data available was subdivided into work breakdown structure items.

Technical data related to the ships has mainly been gathered from open source. Sources used were Jane's Fighting Warships, Federation of American Scientists, Navy Matters, Forecast international, Wikipedia, etc. Regarding cost normalization, the difficulty stands in the different currencies (CAD, EURO, GBP, USD, NOK, SEK) and the different base years (1952 - 2010) for the data collected. The costs were normalized to a common currency and then to a common base year (and subsequently adjusted to a notional common currency for sensitivity reasons). The methodology consists of using an index to take into account inflation, then converting to the common currency with the exchange rate in the chosen base year. Often no publicly available total ship costs index per nation exist, let alone a multinational one. SAS-076 considered a multitude of indices and approaches, for example:

- The application the U.S. Bureau of Labor Statistics index for shipbuidling labour;
- The application of Eurostat labour indices;
- The ratio of a country's Ground Domestic Product (GDP) (between original cost data year and a common year) to obtain the Then Year local currency cost and then use the market exchange rate to get cost in the common currency — abandoned because it was deemed too astray from the shipbuilding industry; and,
- Consumer Price Index taken from International Monetary Fund (IMF) International Financial Statistics (IFS) Yearbooks and IMF/World Economic Outlook database—abandoned because too far from shipbuilding industry.

The final approach decided upon is presented in Section 2.2.

Lessons Identified

- It took quite some time to gather all data required to estimate the costs for the acquisition stage. Data collection was a tedious process spanning many iterations over many months. Internet searches of publicly available databases were used to select the NATO ships that are similar in nature to the Netherlands' LPDs. Janes Fighting Warships and Federation of American Scientists were used to compose a database template of attributes. This template was sent around to SAS-076 participants to fill or validate. Cost data was obtained from national sources contacted by respective SAS-076 participants.

- The group decided it was not necessary to use a MDAL, CARD, DADD or equivalent to capture data and assumptions in a stringent way. The group used a spreadsheet containing all the parameters needed without a identification of sources for each data gathered. This simplified approach was found to be sufficient if no official endorsement is needed for data and assumptions collected.

- An important lesson is that the cost normalization procedure is very important. The choice of a certain index will have a major impact on the results. SAS-054 report is not very explicit in the approach to follow for cost normalization. A more detailed description of how to deal with cost normalization would be helpful.

6.4 Methods, Models and Tools

Team's approach

As recommended in the SAS-054 Report, two independent methods were used for estimating acquisition costs. For the acquisition costs the analogy method and the parametric method were used. For these methods two data mining techniques were applied, which is a new method for cost analysis and considered to be very useful:

- Parametric method: a model tree that is a combination of decision trees and linear regression from which a Cost Estimating Relationships (CERs) is obtained;

- Analogy method: cluster analysis that consists first of identifying the nearest neighbours to both Rotterdam and Johan de Witt ships by first defining a distance function between ships and use the distances and known costs to obtain cost estimates for the Rotterdam and Johan de Witt.

The tools investigated were mainly commercial. It was not possible to acquire one of the tools, due to the very limited resources available. Therefore the group used tools that are publicly available (e.g., the WEKA-tool).

Lesson Identified:

- The data available determined, in a very large extent, the choice of methods. Finding input data has been the criteria to make the process go forward. It was a continuous interaction between choices of methods and computer models and availability of data. For acquisition costs the group assumed that many detailed data would be publicly available; and the group assumed that for sufficient ships, cost data would be available. This allowed the data mining method to be used. The data mining method did not constrain the selection of attributes. Generally, commercial tools fix the range of attributes.

- The new methodology used, based on data mining, is very valuable as it requires only publicly available technical data of attributes and some cost data. However, this method is not mentioned in the SAS-054

report. On the other hand the complexity of the analogy method based on hierarchical clustering lead to computational difficulties when using over one hundred attributes. This issue was dealt with by applying principal component analysis and reducing the dimensionality of the input without significant loss of information.

6.5 Risk and Uncertainty Analysis

Team's Approach

SAS-076 considered the risk and uncertainty analysis task implicitly: the two cost estimating methods produce probability distributions modeling the statistical uncertainty in the prediction—this is the uncertainty of the models. SAS-076 did not consider a full range of uncertainty on the input attributes or on the indices used, however sensitivity analysis was performed. Another issue related to the uncertainty of the acquisition costs concerns the missing data. However, the methodologies used to estimate the acquisition costs were able to handle missing values. Other relevant issues related to the uncertainty of the acquisition costs regard whether the attributes collected by SAS-076 describe the ships completely. It is not clear how the uncertainty could be considered in this case (e.g., if the costs of the more expensive ships are underestimated, is it possible that an important attribute is not accounted for?). These issues are not explicitly covered by the group.

Lessons Identified

- Uncertainty analysis of the acquisition cost estimation consisted of documenting and presenting the predictive capabilities of the data mining methods used: their accuracy on historical systems and the uncertainty in the predicted cost of the Rotterdam and Johan de Witt ships. Probability distributions and cumulative distribution functions were developed and presented graphically and in tabular format to allow a decision maker to select the desired confidence level (and associated cost) of the estimates.

- Further uncertainty analysis could have helped, but was not undertaken. This could have consisted of a Monte Carlo simulation to model different inflation and exchange rates, and could substitute missing attribute values (in lieu of just the mean) - each Monte Carlo iteration would feed the revised data into the data mining algorithms to output the associated cost estimate. This effort was deemed to require substantial programming effort to automate - the value gained was not clear.

- Using cost data from different sources may be a risk. In this study it was not certain if all cost data gathered from different sources have used the same cost boundary. E.g., personnel costs are not always considered in the same way. This issue was covered by assuming that all sources use the same cost boundary.

- Risk analysis, interpreted as examining the chance of loss of injury due to an unfavorable event, was not considered.

6.6 Analysis and Presentation of Results

Team's Approach

SAS-076 did not define a strategy for analysis as such but undertook the following ad hoc analyses for acquisition costs:

- One data outlier was removed from the analysis. This was the San Antonio LPD-17 whose cost overruns are well documented. The LPD-17 cost was more than three orders of magnitude larger than the minimum ship costs.

- The analysis of the initial cluster analysis results, where attributes were given equal weighting, showed poor predictive capability (low correlation of determination, R^2). The group then decided to determine the set of attribute weights that optimized prediction of known instances. This was a non-linear convex optimization problem too large to solve in a reasonable amount of time and memory. To circumvent computation issues subjectively chosen attribute subsets (less than 15 at a time) were tested. Each of these provided a better R^2. To eliminate the subjectivity, a subsequent iteration used principal component analysis to reduce the dimensionality of the data set while accounting for a large percentage of the original data's variability.

- The reason for selecting the M5 model estimate to be the primary estimate was due to the model construction: M5 Model trees are built to optimize prediction of unseen cases; the hierarchical clustering approach is optimized on learning the known cases.

- From this choice, a closer look at the CER showed that rank in class is one of the cost-driving attributes. However, SAS-076 had a doubt regarding Johann de Witt's rank in class. Through discussions with RNLN and the analysis of technical data, SAS-076 concluded that Johann de Witt is considered a separate class. In retrospect, this decision lead to a more accurate estimate.

- SAS-076 tried to determine ANEP-41 ESWBS cost category estimates. Given that only 10 data points (all from U.S.) were available, the effectiveness of data mining techniques was limited. It was suggested to take average percentages. However, results triggered the team that ESWBS items were not interpreted in the same way. Also for not many ships this detailed data was available and forced the team to skip the method to estimate costs at the ESWBS item level.

- The results of the acquisition costs estimate were presented to the RNLN, using the probability distribution with uncertainties. It was felt that the presentation fitted the customer's expectations.

Lessons Identified

- A critical approach on the first results obtained is highly recommended. The group demonstrated that without this critical look, the results obtained would have been less satisfactory for the acquisition costs.

7.0 CONCLUSION

This report 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 HMS 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 HMS 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 for generating cost estimates for other systems for which ample data is available.

The methods herein used for cost estimating are new in the cost environment. Newly developed methods in other areas than cost analysis should always be considered for application in the cost analysis area. The cost analysis societies should be aware of new developments in methods and techniques in other related areas in order to apply them to cost analysis. This is in-line with SAS-054 recommendation O.9.7 stating "Research should be conducted continuously to enhance methods and models for life cycle costing." The methods proved to be valuable as they require only publicly available technical data of attributes and some cost data. In general, no major amendments or changes to the SAS-054 guideline are required, however some enhancements are suggested:

- Differentiate between approaches and methods: parametric, analogy, and engineering methods can each be applied in a top-down or bottom-up approach.
- The SAS-054 guidelines for cost normalization should be expanded to consider multi-national data sets where different currencies and base years are present. The approach taken by SAS-076 in this study can be used as an example.

8.0 REFERENCES

[1] Quinlan, J., "Learning with continuous classes," *Proceedings AI'92 Singapore: World Scientific (Adams & Sterling Eds)*, edited by A. Adams and L. Sterling, Singapore, 1992, pp. 343–348.

[2] Friedman, N., *U.S. Amphibious Ships and Craft: An Illustrated Design History*, Naval Institute Press, Annapolis, Maryland, USA, 2nd ed., 2005.

[3] Berg, G., Swedish Defence Materiel Administration, email exchange with R.H.A.D. Shaw, 17 March 2009.

[4] Federation of American Scientists (http://www.fas.org/), accessed 15 March 2010.

[5] Francis, P.L., Director Acquisition and Sourcing Management Team, "Government Accountability Office, Defense Acquisitions: Realistic Business Cases Needed to Execute Navy Shipbuilding Programs," Testimony Before the Subcommittee on Seapower and Expeditionary Forces, Committee on Armed Services, House of Representatives, July 24, 2007 (GAO-07-943T).

[6] Flynn, B., Office of the Assistant Secretary of the Navy (Research, Development and Acquisition), United States, "Inflation for ship building," email to B. Kaluzny, 10 March 2010 1426 PM EST.

[7] United States Bureau of Labor Statistics (http://www.bls.gov/), accessed 15 March 2010.

[8] European Commission Eurostat (http://epp.eurostat.ec.europa.eu/), accessed 22 March 2010.

[9] Solomon, B., DRDC CORA Defence Economics Team, "Inflation for ship building," email to B. Kaluzny, 09 July 2009, 11:50 EST.

[10] Flynn, B., Office of the Assistant Secretary of the Navy (Research, Development and Acquisition), United States, "Use of Mean versus median in Service Cost Positions," email to B. Kaluzny, 13 May 2010 1352 EST.

[11] Beaulieu, N., Abu-Dayya, A., and McLance, P., "Estimating the distribution of a sum of independent lognormal random variables," *IEEE Trans. Commun.*, Vol. 43, 1995, pp. 2869–2873.

[12] Miroyannis, A., *Estimation of Ship Construction Costs*, Ph.D. thesis, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA, 2006.

[13] Ryan, T., *Modern Regression Methods*, John Wiley & Sons, Inc., New York, NY, USA, 2nd ed., 1997.

[14] Witten, I. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Massachusetts, NE, USA, 2nd ed., 2005.

[15] Wang, Y. and Witten, I., "Inducing Model Trees for Continuous Classes," *In Proc. of the 9th European Conf. on Machine Learning Poster Papers*, Prague, Czech Republic, 1997, pp. 128–137.

[16] Malerba, D., Esposito, F., Ceci, M., and Appice, A., "Top-Down Induction of Model Trees with Regression and Splitting Nodes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 5, 2004, pp. 612–625.

[17] Torgo, L., *Inductive Learning of Tree-based Regression Models*, Ph.D. thesis, Universidade do Porto, Oporto, Portugal.

[18] Torgo, L., "Computationally Efficient Linear Regression Trees," *In Proc. of IFCS 2002: Classification, Clustering and Data Analysis: Recent Advances and Applications*, edited by K. Jajuga, A. Sokolowski, and H. Bock, Springer, New York, USA, 2002, pp. 409–417.

[19] Dobra, A., "SECRET: A scalable linear regression tree algorithm," *In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, Edmonton, Canada, 2002, pp. 481–487.

[20] Chen, Z., *Reduced-Parameter Modeling for Cost Estimation Models*, Ph.D. thesis, Faculty of the Graduate School, University of Southern California, Los Angeles, CA, USA, 2006.

[21] WEKA: Data Mining with Open Source Machine Learning Software in Java (http://www.cs.waikato.ac.nz/ml/weka/), accessed 20 September 2010.

[22] Akaike, H., "Likelihood and the Bayes procedure," *Trabajos de Estadística y de Investigación Operativa*, Vol. 31, 1980, pp. 143–166.

[23] NATO RTO SAS-054 Panel, "Methods and Models for Life Cycle Costing," Technical Report RTO-TR-SAS-054, NATO RTO, June 2007.

[24] Goldberger, A., "The Interpretation and Estimation of Cobb-Douglas Functions," *Econometrica*, Vol. 35, 1968, pp. 464–472.

Appendix B1-1: DATA SET

Table 15: Technical data for the United Kingdom, Swedish, Norwegian, and French ships.

Data Category & Element	Albion	Largs Bay	Ocean	Carlskrona	Atle	Oden	Svalbard	Siroco	Mistral
I. DESCRIPTION									
e. Rank in class	1	1	1	1	1	1	1	2	1
f. Vessel type	Mil	Mil	Mil	Mil	Civ	Civ	Civ	Mil	Mil
II. CONSTRUCTION									
h. Civilian classification stnds.?	N	?	Y	Y	?	Y	Y	N	Y
III. DIMENSIONS									
a. Length (m)	176	176.6	203.4	105.7	104.6	107.8	103.7	168	199
b. Beam (m)	28.9	26.4	34.4	15.2	23.8	31	19.1	23.5	32
c. Draught (m)	7.1	5.8	6.6	4	8.3	8.5	6.5	5.2	6.2
d. Displacement Light	14600	13690	?	3150	?	11000	?	8230	16529
e. Displacement Full Load	18500	16160	21758	3800	9500	13000	6300	12400	21600
IV. PERFORMANCE									
a. Top speed	18	18	19	20	19	16	17	21	19
i. Range: Total distance	8000	8000	8000	?	?	30000	10000	11000	11000
ii. Range: Econ. speed	15	15	15	?	?	13	13	15	15
iii. Range: Sailing time	533	533	533	?	?	2308	769	733	733
b. Endurance (days)	?	?	?	?	56	100	?	30	45
c. Crew: complement	325	60	491	170	20	15	48	218	177
i. Officers	?	?	?	?	?	?	?	18	20
ii. Non-officers	?	?	?	?	?	?	?	200	157
V. PROPULSION									
a. Propulsion technology	Electric	Electric	Diesel	Diesel	Electric	Diesel	Electric	Diesel	Electric
b. Propeller shafts	2	0	2	2	4	2	0	2	0
i. Shaft power (MW)	?	0	13.5	7.76	16.2	18	0	15.3	0
ii. Propeller type	LIPS	N/A	fp	cp	?	cp	N/A	LIPS cp	N/A
c. Propulsion pods	0	2	0	0	?	0	2	0	2

continued on next page

continued from previous page

Data Category & Element	Albion	Largs Bay	Ocean	Carlskrona	Atle	Oden	Svalbard	Siroco	Mistral
i. Total power (MW)	0	4.4	0	0	?	0	10	0	14
d. Net power (MW)	?	?	13.5	7.76	?	18	10	15.3	14
e. Bow Thrusters	1	1	1	?	?	0	1	1	1
i. Total power (MW)	0.865	?	0.45	?	?	0	?	0.735	1.5
VI. ELECT. POWER GEN.									
a. Generators	4	4	?	?	?	4	4	5	4
i. Total capacity (MW)	15.6	11.2	?	?	16.2	?	13.56	4.25	20.8
ii. Generator technology	Diesel	Diesel	Diesel	?	Diesel	Diesel	Diesel	Diesel	Diesel
VII. LIFT CAPACITY									
a. Vehicle fuel (litres)	?	?	?	?	?	?	?	?	?
b. Aviation fuel (litres)	?	?	?	?	?	?	?	?	1071.3
c. Fresh water (litres)	?	?	?	?	?	310000	?	?	380
d. Bulk cargo space (m^3)	?	?	?	?	?	?	?	?	?
e. Vehicle space (m^2)	?	2875	?	0	0	0	0	?	2650
f. Well deck (Y/N)	Y	Y	N	N	N	N	N	Y	Y
i. Length (m)	?	95.5	0	0	0	0	0	122	?
ii. Width (m)	?	14	0	0	0	0	0	14.2	?
iii. Capacity (m^2)	?	1440	0	0	0	0	0	1732.4	885
iv. LCAC	2	0	0	0	0	0	0	?	2
v. LCM 6	?	?	0	0	0	0	0	?	?
vi. LCM 8	?	?	0	0	0	0	0	?	?
vii. LCU	4	2	0	0	0	0	0	?	4
viii. LVT	?	?	0	0	0	0	0	?	?
ix. LCVP	?	2	0	0	0	0	0	?	?
x. LCPL	?	?	0	0	0	0	0	?	?
xi. EFV	?	?	0	0	0	0	0	?	?
g. Cargo/Aircraft Elevator/Lifts	?	?	2	0	?	?	0	1	3
i. Capacity $\leq$ 5 tonnes	?	?	?	0	?	?	0	0	?
ii. 5 < capacity $\leq$ 10 tonnes	?	?	?	0	?	?	0	0	?
iii. 10 < capacity $\leq$ 15 tonnes	?	?	?	0	?	?	0	0	?

continued on next page

continued from previous page

Data Category & Element	Albion	Largs Bay	Ocean	Carlskrona	Atle	Oden	Svalbard	Siroco	Mistral
iv. Capacity $\geq$ 15 tonnes	?	?	?	0	?	?	0	1	?
h. Cranes	?	2	?	?	?	3	0	1	?
i. Capacity $\leq$ 5 tonnes	?	0	?	?	?	1	0	0	?
ii. 5 < capacity $\leq$ 10 tonnes	?	0	?	?	?	1	0	0	?
iii. 10 < capacity $\leq$ 15 tonnes	?	0	?	?	?	1	0	0	?
iv. 15 < capacity $\leq$ 20 tonnes	?	0	?	?	?	0	0	0	?
v. 20 < capacity $\leq$ 25 tonnes	?	0	?	?	?	0	0	0	?
vi. 25 < capacity $\leq$ 30 tonnes	?	2	?	?	?	0	0	0	?
vii. 30 < capacity $\leq$ 40 tonnes	?	0	?	?	?	0	0	1	?
viii. 40 < capacity $\leq$ 50 tonnes	?	0	?	?	?	0	0	0	?
ix. 50 < capacity $\leq$ 60 tonnes	?	0	?	?	?	0	0	0	?
x. Capacity > 60 tonnes	?	0	?	?	?	0	0	0	?
i. Berthing (troop capacity): Baseline	305	356	480	?	?	65	?	470	450
j. Berthing (troop capacity): Surge	305	344	323	?	?	?	?	?	450
VIII. FLIGHT DECK									
a. Equipped with flight deck? (Y/N)	Y	Y	Y	Y	?	?	Y	Y	Y
b. Flight deck length (m)	64	?	170	?	?	?	?	?	199
c. Flight deck width (m)	?	?	31.7	?	?	?	?	?	32
d. Flight deck area (m^2)	?	?	5389	?	?	?	?	1740	6368
e. Helicopter landing spots (max. #)	2	2	6	?	?	?	1	3	6
i. Merlin / Sea King	2	2	6	?	?	?	?	?	?
ii. NH 90 / Lynx / Puma / Cougar	?	?	?	?	?	?	?	3	6
iii. CH-46E Sea Knight	?	?	?	?	?	?	?	?	?
iv. CH-53 Sea Stallion	?	?	?	?	?	?	?	?	1
v. MV-22 Osprey	?	?	?	?	?	?	?	?	1
f. Chinook capable (Yes/No)	Y	Y	Y	N	?	?	N	?	Y
g. Equipped with hangar? (Y/N)	N	N	Y	N	?	?	N	?	Y
h. Hangar size (m^2)	0	0	2338	0	?	?	0	?	1800
i. Helicopters supported (max. #)	2	?	18	0	?	?	1	4	16
i. Merlin / Sea King	?	?	12	0	?	?	?	?	?

continued on next page

continued from previous page

Data Category & Element	Albion	Largs Bay	Ocean	Carlskrona	Atle	Oden	Svalbard	Siroco	Mistral
ii. NH 90 / Lynx / Puma / Cougar	?	?	6	0	?	?	1	4	16
iii. CH-46E Sea Knight	?	?	?	0	?	?	?	?	?
iv. CH-53 Sea Stallion	?	?	?	0	?	?	?	?	?
v. MV-22 Osprey	?	?	?	0	?	?	?	?	?
IX. ARMAMENT									
a. Guns (calibre $\geq$ 75mm)	0	0	0	0	0	0	0	0	0
b. Guns (50mm $\leq$ calibre < 75mm)	0	0	0	2	0	0	1	0	0
c. Guns (30mm $\leq$ calibre < 50mm)	0	0	0	2	0	0	0	3	2
d. Guns (20mm $\leq$ calibre < 30mm)	2	0	8	0	0	0	0	0	0
e. 30mm CIWS (Goalkeeper)	2	0	0	0	0	0	0	0	0
f. 20mm CIWS (Phalanx)	0	0	3	0	0	0	0	0	0
g. Machine guns (12.7mm)	0	0	0	0	0	0	?	4	4
h. Machine guns (7.62mm)	4	6	4	0	0	0	?	0	0
i. SSM launchers	0	0	0	0	0	0	0	0	0
j. SAM launchers	0	0	0	0	0	0	0	3	2
k. # of torpedos carried	0	0	0	0	0	0	0	0	0
l. Torpedo tubes	0	0	0	0	0	0	0	0	0
m. Torpedo launchers	0	0	0	0	0	0	0	0	0
X. COUNTERMEASURES									
a. Chaff launchers	8	?	8	2	0	0	0	0	0
b. Torpedo decoys	1	?	1	0	0	0	0	1	1
c. Other systems	0	?	0	0	0	0	0	0	0
d. Number of ESM systems	1	?	1	1	0	0	0	0	1
e. Number of ECM systems	1	?	1	0	0	0	0	2	0
XI. RADARS/TACAN/IFF/SONARS									
a. Total radar systems mounted	3	1	4	5	?	?	2	4	3
i. A-band	0	0	0	0	?	?	0	0	0
ii. B-band	0	0	0	0	?	?	0	0	0
iii. C-band	0	0	0	0	?	?	0	0	0
iv. D-band	0	0	0	0	?	?	0	1	0

continued on next page

continued from previous page

Data Category & Element	Albion	Largs Bay	Ocean	Carlskrona	Atle	Oden	Svalbard	Siroco	Mistral
v. E-band	2	0	2	1	?	?	0	0	0
vi. F-band	2	0	2	1	?	?	0	0	0
vii. G-band	0	0	0	1	?	?	1	0	1
viii. H-band	0	0	0	1	?	?	0	0	0
ix. I-band	1	1	2	4	?	?	1	3	2
x. J-band	0	0	0	2	?	?	0	0	0
b. # of TACAN/IFF systems mounted	1	?	1	?	?	?	?	?	?
c. # of distinct sonar systems mounted	0	?	0	?	?	?	0	?	0
XII. COMBAT DATA SYSTEMS									
a. Number of distinct systems	5	?	6	?	?	?	?	5	7
XIII. WEAPONS CONTROL SYSTEMS									
a. Number of distinct systems	2	?	?	?	?	?	?	2	2
XIV. OTHER CAPABILITIES									
a. Equipped with hospital? (Y/N)	?	?	?	Y	?	Y	N	Y	Y
i. Number of beds	?	?	?	?	?	?	0	47	69
ii. Operating rooms	?	?	?	?	?	?	0	2	2
iii. X-Ray facility (Y/N)	?	?	?	?	?	?	N	?	Y
b. Dental capability (Y/N)	?	?	?	?	?	?	N	?	?
c. Command/Control facility (Y/N)	Y	?	Y	?	?	?	N	?	Y
d. NBCD Facilities (Y/N)	?	Y	Y	?	?	?	Y	?	?

Table 16: Technical data for the United States and Canadian ships.

Data Category & Element	Thomaston	Anchorage	Whidbey Island	Harpers Ferry	Raleigh	Austin	San Antonio	Protecteur
I. DESCRIPTION								
e. Rank in class	1	1	1	1	1	1	1	1
f. Vessel type	Mil	Mil	Mil	Mil	Mil	Mil	Mil	Mil
II. CONSTRUCTION								
h. Civilian classification stnds.?	N	N	N	N	N	N	N	N
III. DIMENSIONS								
a. Length (m)	160	168.6	185.8	185.8	159.1056	173.8	208.4	171.9
b. Beam (m)	26	25.6	25.6	25.6	30.48	30.5	31.9	23.2
c. Draught (m)	5.8	6	6.3	6.3	7.0104	7	7	10.46
d. Displacement Light	9042	8600	11125	11125	8789	9130	18477	9259
e. Displacement Full Load	11710	13700	15939	16740	14339	17244	25885	25676
IV. PERFORMANCE								
a. Top speed	21	22	22	22	21	21	22	21
i. Range: Total distance	13000	14800	8000	8000	9600	7700	?	7500
ii. Range: Econ. speed	10	12	18	18	16	20	?	11.5
iii. Range: Sailing time	?	1233	444	444	600	385	?	652
b. Endurance (days)	?	?	?	?	?	60	?	?
c. Crew: complement	341	374	413	352	429	420	396	335
i. Officers	29	24	21	24	29	24	32	38
ii. Non-officers	312	350	392	328	400	396	364	297
V. PROPULSION								
a. Propulsion technology	Steam	Steam	Diesel	Diesel	Steam	Steam	Diesel	Steam
b. Propeller shafts	2	2	2	2	2	2	2	1
i. Shaft power (MW)	17	17.9	24.6	24.6	17.9	17.9	29.84	15.7
ii. Propeller type	?	?	cp	cp	?	?	cp	?
c. Propulsion pods	0	0	0	0	0	0	0	0
i. Total power (MW)	0	0	0	0	0	0	0	0
d. Net power (MW)	17	17.9	24.6	24.6	17.9	17.9	29.84	15.7
e. Bow Thrusters	0	0	0	0	0	0	0	1

continued on next page

continued from previous page

Data Category & Element	Thomaston	Anchorage	Whidbey Island	Harpers Ferry	Raleigh	Austin	San Antonio	Protecteur
i. Total power (MW)	0	0	0	0	0	0	0	?
VI. ELECT. POWER GEN.								
a. Generators	?	?	4	4	?	4	5	?
i. Total capacity (MW)	?	?	5.2	5.2	?	?	12.5	?
ii. Generator technology	Steam	Steam	Diesel	Diesel	?	Steam	Diesel	?
VII. LIFT CAPACITY								
a. Vehicle fuel (litres)	?	?	?	?	?	?	37850	?
b. Aviation fuel (litres)	?	110200	110000	110000	?	?	1173500	?
c. Fresh water (litres)	?	?	?	?	?	?	?	?
d. Bulk cargo space (m^3)	99	?	141.6	1914	?	?	963	?
e. Vehicle space (m^2)	?	?	1161	1877	?	?	2285	?
f. Well deck (Y/N)	Y	Y	Y	Y	Y	Y	Y	N
i. Length (m)	120.7	131.1	134.1	?	51.2	120.1	?	0
ii. Width (m)	15.24	15.2	15.2	15.2	15.2	15.2	?	0
iii. Capacity (m^2)	1840	1992.72	2038.32	?	778.24	1825.52	?	0
iv. LCAC	?	3	4	2	2	2	2	0
v. LCM 6	21	18	21	9	3	9	?	0
vi. LCM 8	9	9	10	4	4	4	?	0
vii. LCU	3	3	3	1	1	?	1	0
viii. LVT	48	50	64	64	20	20	?	0
ix. LCVP	?	?	?	?	?	4	?	4
x. LCPL	?	?	2	2	?	4	?	0
xi. EFV	?	?	?	?	?	?	14	0
g. Cargo/Aircraft Elevator/Lifts	?	?	0	0	?	1	3	0
i. Capacity $\leq$ 5 tonnes	?	?	0	0	?	0	1	0
ii. 5 < capacity $\leq$ 10 tonnes	?	?	0	0	?	1	2	0
iii. 10 < capacity $\leq$ 15 tonnes	?	?	0	0	?	0	0	0
iv. Capacity $\geq$ 15 tonnes	?	?	0	0	?	0	0	0
h. Cranes	3	2	3	1	?	7	2	2
i. Capacity $\leq$ 5 tonnes	0	0	0	0	?	6	2	0

continued on next page

continued from previous page

Data Category & Element	Thomaston	Anchorage	Whidbey Island	Harpers Ferry	Raleigh	Austin	San Antonio	Protecteur
ii. 5 < capacity $\leq$ 10 tonnes	1	0	0	0	?	0	0	0
iii. 10 < capacity $\leq$ 15 tonnes	0	0	1	0	?	0	0	2
iv. 15 < capacity $\leq$ 20 tonnes	0	0	1	0	?	0	0	0
v. 20 < capacity $\leq$ 25 tonnes	0	0	0	0	?	0	0	0
vi. 25 < capacity $\leq$ 30 tonnes	0	0	0	1	?	1	0	0
vii. 30 < capacity $\leq$ 40 tonnes	0	0	0	0	?	0	0	0
viii. 40 < capacity $\leq$ 50 tonnes	2	2	0	0	?	0	0	0
ix. 50 < capacity $\leq$ 60 tonnes	0	0	1	0	?	0	0	0
x. Capacity > 60 tonnes	0	0	0	0	?	0	0	0
i. Berthing (troop capacity): Baseline	325	366	402	402	930	900	699	0
j. Berthing (troop capacity): Surge	?	?	102	102	?	?	101	0
VIII. FLIGHT DECK								
a. Equipped with flight deck? (Y/N)	Y	Y	Y	Y	Y	Y	Y	Y
b. Flight deck length (m)	?	?	64.6	?	?	51.2	?	?
c. Flight deck width (m)	?	?	25.3	?	?	?	?	?
d. Flight deck area (m^2)	?	?	1634.38	?	?	?	?	?
e. Helicopter landing spots (max. #)	1	1	2	2	?	2	4	?
i. Merlin / Sea King	?	?	?	?	?	?	?	?
ii. NH 90 / Lynx / Puma / Cougar	?	?	?	?	?	?	?	?
iii. CH-46E Sea Knight	?	?	?	?	?	2	4	?
iv. CH-53 Sea Stallion	?	1	2	2	?	2	2	?
v. MV-22 Osprey	?	?	?	?	?	0	2	0
f. Chinook capable (Yes/No)	?	?	?	?	?	?	?	N
g. Equipped with hangar? (Y/N)	?	N	N	N	N	N	Y	Y
h. Hangar size (m^2)	?	0	0	0	0	0	?	?
i. Helicopters supported (max. #)	8	1	2	0	6	6	2	3
i. Merlin / Sea King	?	?	0	0	?	?	?	3
ii. NH 90 / Lynx / Puma / Cougar	?	?	0	0	?	?	?	0
iii. CH-46E Sea Knight	?	?	0	0	6	6	2	0
iv. CH-53 Sea Stallion	?	1	2	0	?	3	1	0

continued on next page

continued from previous page

Data Category & Element	Thomaston	Anchorage	Whidbey Island	Harpers Ferry	Raleigh	Austin	San Antonio	Protecteur
v. MV-22 Osprey	?	?	0	0	?	?	1	0
IX. ARMAMENT								
a. Guns (calibre $\geq$ 75mm)	4	4	0	0	8	4	0	1
b. Guns (50mm $\leq$ calibre < 75mm)	0	0	0	0	0	0	0	0
c. Guns (30mm $\leq$ calibre < 50mm)	0	0	0	0	0	0	2	0
d. Guns (20mm $\leq$ calibre < 30mm)	6	0	2	2	0	0	0	0
e. 30mm CIWS (Goalkeeper)	0	0	0	0	0	0	0	0
f. 20mm CIWS (Phalanx)	0	0	2	2	0	0	0	0
g. Machine guns (12.7mm)	?	6	6	6	0	8	4	6
h. Machine guns (7.62mm)	?	0	0	0	0	0	0	0
i. SSM launchers	?	0	0	0	0	0	0	?
j. SAM launchers	?	0	2	2	0	0	2	?
k. # of torpedos carried	?	0	0	0	0	0	0	0
l. Torpedo tubes	?	0	0	0	0	0	0	0
m. Torpedo launchers	?	0	0	0	0	0	0	0
X. COUNTERMEASURES								
a. Chaff launchers	?	0	4	6	0	0	6	6
b. Torpedo decoys	?	0	1	1	0	0	1	0
c. Other systems	?	0	0	0	0	0	0	0
d. Number of ESM systems	?	1	2	2	1	1	1	1
e. Number of ECM systems	?	0	0	0	0	0	1	0
XI. RADARS/TACAN/IFF/SONARS								
a. Total radar systems mounted	?	3	3	3	3	3	3	3
i. A-band	?	0	0	0	0	0	0	0
ii. B-band	?	0	0	0	0	1	0	0
iii. C-band	?	0	1	1	0	0	0	0
iv. D-band	?	0	0	0	0	0	0	0
v. E-band	?	1	0	0	1	0	1	0
vi. F-band	?	1	0	0	1	0	1	0
vii. G-band	?	1	1	1	1	1	0	1

continued on next page

continued from previous page

Data Category & Element	Thomaston	Anchorage	Whidbey Island	Harpers Ferry	Raleigh	Austin	San Antonio	Protecteur
viii. H-band	?	0	0	0	0	0	0	0
ix. I-band	?	1	1	1	1	1	2	2
x. J-band	?	0	1	1	0	0	0	0
b. # of TACAN/IFF systems mounted	?	?	1	1	1	1	?	1
c. # of distinct sonar systems mounted	?	0	0	0	0	0	0	0
XII. COMBAT DATA SYSTEMS								
a. Number of distinct systems	?	2	3	3	2	3	5	2
XIII. WEAPONS CONTROL SYSTEMS								
a. Number of distinct systems	?	?	?	?	?	?	?	?
XIV. OTHER CAPABILITIES								
a. Equipped with hospital? (Y/N)	N	?	Y	Y	?	Y	Y	Y
i. Number of beds	0	?	?	?	?	12	24	8
ii. Operating rooms	0	?	?	?	?	?	2	1
iii. X-Ray facility (Y/N)	N	?	?	?	?	Y	?	Y
b. Dental capability (Y/N)	N	?	Y	Y	?	Y	Y	Y
c. Command/Control facility (Y/N)	N	?	?	?	?	Y	?	?
d. NBCD Facilities (Y/N)	?	?	?	?	?	?	?	Y

Appendix B1-2: PRINCIPAL COMPONENT DATA SET

Table 17: Principal component analysis (80% coverage): resulting attributes.

Name	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Rotterdam	-0.6206	2.4141	0.7228	-0.9292	-0.4955	0.7275	2.2589	1.8520	-0.3011	1.2791
Johan de Witt	-1.2649	2.8019	0.5701	-0.5635	-0.8626	0.4325	2.3956	1.3758	0.0415	1.4135
Thomaston	1.9276	3.1298	-0.1413	1.5096	0.7688	-0.1155	0.4850	1.1027	-0.3971	0.5587
Plymouth Rock	1.9379	3.1359	-0.1478	1.5097	0.7631	-0.1156	0.4846	1.1057	-0.3994	0.5546
Fort Snelling	1.9482	3.1419	-0.1543	1.5099	0.7574	-0.1156	0.4841	1.1087	-0.4017	0.5505
Point Defiance	1.9585	3.1480	-0.1608	1.5101	0.7517	-0.1157	0.4836	1.1117	-0.4040	0.5463
Spiegel Grove	1.9687	3.1541	-0.1673	1.5103	0.7460	-0.1158	0.4831	1.1147	-0.4063	0.5422
Alamo	1.9790	3.1602	-0.1738	1.5104	0.7403	-0.1159	0.4826	1.1176	-0.4086	0.5381
Hermitage	1.9893	3.1662	-0.1803	1.5106	0.7346	-0.1160	0.4821	1.1206	-0.4109	0.5340
Monticello	1.9995	3.1723	-0.1868	1.5108	0.7290	-0.1161	0.4816	1.1236	-0.4132	0.5298
Anchorage	1.7511	3.4262	0.0422	0.1333	0.1469	0.1039	0.4367	1.1533	0.5488	1.2309
Portland	1.7614	3.4323	0.0357	0.1335	0.1412	0.1038	0.4362	1.1562	0.5465	1.2268
Pensacola	1.7716	3.4384	0.0292	0.1337	0.1355	0.1038	0.4357	1.1592	0.5442	1.2227
Mount Vernon	1.7819	3.4444	0.0227	0.1339	0.1298	0.1037	0.4352	1.1622	0.5419	1.2185
Fort Fisher	1.7922	3.4505	0.0162	0.1340	0.1241	0.1036	0.4347	1.1652	0.5396	1.2144
Whidbey Island	2.2519	3.5869	2.5713	-0.4678	-0.9544	-0.4128	0.0767	1.0708	-0.3635	1.0883
Germantown	2.2622	3.5930	2.5648	-0.4676	-0.9600	-0.4129	0.0763	1.0738	-0.3658	1.0842
Fort McHenry	2.2724	3.5991	2.5583	-0.4674	-0.9657	-0.4130	0.0758	1.0768	-0.3681	1.0801
Gunston Hall	2.2827	3.6052	2.5518	-0.4673	-0.9714	-0.4131	0.0753	1.0798	-0.3704	1.0759
Comstock	2.2930	3.6112	2.5453	-0.4671	-0.9771	-0.4132	0.0748	1.0828	-0.3727	1.0718
Tortuga	2.3032	3.6173	2.5388	-0.4669	-0.9828	-0.4133	0.0743	1.0858	-0.3750	1.0677
Rushmore	2.3135	3.6234	2.5323	-0.4667	-0.9885	-0.4134	0.0738	1.0887	-0.3773	1.0635
Ashland	2.3238	3.6294	2.5258	-0.4666	-0.9942	-0.4134	0.0733	1.0917	-0.3796	1.0594
Harpers Ferry	1.6647	3.2549	2.1292	-0.8805	-0.5622	-0.2858	0.1686	0.7128	-0.4586	0.9417
Carter Hall	1.7338	3.3117	2.1975	-0.7664	-0.5860	-0.2634	0.2029	0.7011	-0.4489	0.9213
Oak Hill	1.7440	3.3178	2.1910	-0.7663	-0.5917	-0.2635	0.2024	0.7041	-0.4512	0.9172
Pearl Harbour	1.7543	3.3239	2.1845	-0.7661	-0.5973	-0.2636	0.2019	0.7071	-0.4535	0.9131
Raleigh	1.3761	3.3173	-0.3228	-1.0441	0.0914	0.1091	0.0184	0.8934	-0.3140	1.3569
Vancouver	1.3864	3.3233	-0.3293	-1.0439	0.0857	0.1090	0.0179	0.8964	-0.3163	1.3528
La Salle	1.3967	3.3294	-0.3358	-1.0437	0.0800	0.1089	0.0174	0.8994	-0.3186	1.3487
Austin	2.1235	4.0747	-1.0059	-1.4247	-0.6041	0.2968	0.0522	1.0929	-0.2233	0.9882
Ogden	2.1338	4.0808	-1.0124	-1.4245	-0.6098	0.2967	0.0518	1.0959	-0.2256	0.9841
Duluth	2.1441	4.0868	-1.0189	-1.4243	-0.6155	0.2966	0.0513	1.0988	-0.2279	0.9800
Cleveland	2.1509	4.0852	-1.0177	-1.4156	-0.6195	0.2900	0.0523	1.0992	-0.2269	0.9703
Dubuque	2.1612	4.0913	-1.0242	-1.4154	-0.6252	0.2899	0.0518	1.1022	-0.2292	0.9662
Denver	2.1714	4.0973	-1.0307	-1.4153	-0.6308	0.2899	0.0514	1.1052	-0.2315	0.9621
Juneau	2.1817	4.1034	-1.0372	-1.4151	-0.6365	0.2898	0.0509	1.1082	-0.2338	0.9579
Coronado	2.1920	4.1095	-1.0437	-1.4149	-0.6422	0.2897	0.0504	1.1112	-0.2361	0.9538
Shreveport	2.2022	4.1155	-1.0502	-1.4147	-0.6479	0.2896	0.0499	1.1141	-0.2384	0.9497
Nashville	2.2125	4.1216	-1.0567	-1.4146	-0.6536	0.2895	0.0494	1.1171	-0.2407	0.9456
Trenton	2.2262	4.1354	-1.0709	-1.4229	-0.6610	0.2959	0.0474	1.1227	-0.2464	0.9470
Ponce	2.2365	4.1415	-1.0774	-1.4227	-0.6667	0.2958	0.0469	1.1257	-0.2487	0.9428
Svalbard	-0.7433	1.1388	-0.4784	-0.0210	-0.1696	-0.7977	-0.3938	-0.2546	-0.0443	1.0611
Carlskrona	0.0059	0.9827	0.8272	-1.3437	0.1930	-0.0546	-1.1465	1.1058	-0.1787	2.0847
Atle	0.0467	1.9025	-0.0422	-1.3997	0.1531	-0.3840	-0.7034	0.2808	-0.2888	1.5649
Oden	0.0028	1.5587	0.0901	-1.2462	0.6961	-0.7923	-1.0531	0.4929	-0.0974	1.4152
Protecteur	0.4067	1.8877	0.6018	-1.6665	0.3229	-1.0168	-0.0801	1.1704	0.0202	0.5590
Preserver	0.4170	1.8938	0.5953	-1.6663	0.3172	-1.0168	-0.0806	1.1734	0.0179	0.5549
Albion	-0.0074	3.5859	0.8295	-0.4372	-0.1269	0.0663	1.7115	0.3668	-0.8352	2.2346
Bulwark	0.0029	3.5920	0.8230	-0.4370	-0.1326	0.0662	1.7110	0.3698	-0.8376	2.2304
Largs Bay	-0.3477	3.5454	0.2793	-0.3506	-1.0843	0.0178	0.7886	-1.0940	-0.5283	0.9502
Lyme Bay	-0.3375	3.5515	0.2728	-0.3504	-1.0900	0.0177	0.7881	-1.0910	-0.5306	0.9461
Mounts Bay	-0.3272	3.5575	0.2663	-0.3502	-1.0957	0.0177	0.7877	-1.0880	-0.5329	0.9420
Cardigan Bay	-0.3169	3.5636	0.2598	-0.3500	-1.1013	0.0176	0.7872	-1.0850	-0.5352	0.9378
Ocean	-1.3561	3.0070	0.8555	-1.2696	0.6200	-0.7988	0.0159	2.1329	-2.6844	1.2224
Siroco	-0.0044	3.5500	1.6121	-0.4833	-0.1249	2.0335	-0.2927	1.1556	0.5445	1.1961
Mistral	-2.0123	5.0343	0.5142	0.1228	-1.1827	-0.4498	-0.2610	1.3142	0.2379	0.5384
Tonnerre	-2.0020	5.0404	0.5077	0.1230	-1.1884	-0.4499	-0.2615	1.3172	0.2355	0.5343
Dixmude (BPC3)	-1.9670	5.0010	0.7382	0.2820	-1.4906	-0.5423	-0.3208	1.4085	0.2621	0.7027

Appendix B1-3: M5 MODEL TREE ALGORITHM

This Annex details the M5 model tree algorithm. It does not present original research and the algorithm itself should not be attributed to the authors of this paper. The reader is referred to Witten and Frank [14] and Wang and Witten [15] for further details.

The description of the algorithm is divided into four subsections: building the decision tree, fitting linear regression models, pruning the tree, and finally smoothing the tree. The two final subsections of this annex detail how the M5 system handles nominal attributes and missing data.

Building the Initial Decision Tree

The M5 model tree algorithm constructs a decision tree by recursively splitting the instance space (training set—e.g., the SAS-076 ship data set). Let T be the set of training data that reaches a particular node. A splitting criterion is used to determine which attribute, and what value of that attribute, is optimal to split T into sets T_L and T_R. The splitting criterion is based on minimizing the standard deviation of the known class values (e.g., ship cost) in T. The expected error reduction, or standard deviation reduction (SDR), is calculated by

$$SDR = sd(T) - \left(\frac{|T_L|}{|T|} \times sd(T_L) - \frac{|T_R|}{|T|} \times sd(T_R) \right), \tag{15}$$

where $sd(T)$ is the standard deviation of the class values in T.

The recursion stops at a particular node when the standard deviation is only a small fraction (e.g., 5%) of the standard deviation of the original instance set (at the root of the tree). Splitting also terminates when only a few instances remain ($T \leq 4$)[16].

Fitting Linear Regression Models

After the tree has been grown, the M5 system computes a linear regression model for every node N of the tree, recursively descending the tree. Each regression model has the form

$$w_0 + w_1 a_1 + w_2 a_2 + \ldots + w_k a_k, \tag{16}$$

where $a_1, a_2, \ldots, a_k$ are (numeric) attribute values. The weights $w_1, w_2, \ldots, w_k$ are calculated using standard multivariate linear regression. Only the attributes that are used for splitting decisions in the subtree below a node are used in the regression. The other attributes that affect the prediction value have been taken into account in the tests that lead to that node. Attributes are greedily dropped from a regression model if doing so improves the expected estimated error, computed as

$$Error(N) = \frac{n+v}{n-v} \times \frac{\sum_{instances} |\text{deviation from predicted class values}|}{n}, \tag{17}$$

where n is the number of instances at the node and v is the number of parameters in the node's regression model. This expected estimation error is calculated by averaging the absolute difference between the predicted value and the actual value for each of the training examples that reach that node. This results in underestimation of the expected error outside the calibrating data. As a compensation factor, the expected error is multiplied by $(n+v)/(n-v)$. Dropping a term decreases the latter multiplication factor, which may be enough to offset the increase in the average error over the training cases. This procedure is similar to the modification of the coefficient of determination, R^2, to the more representative $R^2_{adjusted}$ in regression theory [13].

[16] Experiments have shown that the results obtained are not very sensitive to the exact choice of these thresholds.

Pruning the Tree

Once the tree is built and linear regression models fitted for each node, the tree is recursively pruned from the leaves if this results in a lower expected estimated error. If T is the set of instances that reach a node, let the T_L and T_R be the split of these instances between the left (L) and right (R) children nodes. For each node N, the error calculated by equation (17) is compared to the expected error from the subtree below (branches L and R), recursively calculated as

$$subtreeError(N) = \frac{|T_L| \times subtreeError(L) + |T_R| \times subtreeError(R)}{|T|} \text{ if } N \text{ is an internal node,}$$
$$subtreeError(N) = Error(N) \text{ if } N \text{ is a leaf.} \tag{18}$$

Smoothing the M5 Model Tree

After pruning, the adjacent linear models will be sharply discontinuous at the leaves of the pruned tree. This problem is pronounced for models constructed from a small number of training instances. The M5 system applies a smoothing process combining the linear regression model at a leaf with the models on the path to the root to form the final model that is placed at the leaf. In effect, the estimated value of the leaf model is filtered along the path back to the root. At each node, that value is combined with the value predicted by the linear model for that node. The calculation is

$$p' = \frac{np + kq}{n + k}, \tag{19}$$

where p' is the prediction passed up to the next higher node, p is the prediction passed to this node from below, q is the value predicted by the model at this node, n is the number of training instances that reach the node below, and k is a constant. Experiments by Wang and Witten [15] showed that smoothing substantially increases the accuracy of predictions. After the smoothing process, the M5 model tree effectively has constructed piece-wise linear regression models.

Nominal Attributes

Nominal attributes are transformed into binary variables and then treated as numeric during the construction of the model tree. For each nominal attribute, the average class value (e.g., cost) corresponding to each possible attribute value is calculated from the training set. The attribute values are then sorted from smallest to largest average class value. A nominal attribute with k possible values is converted into $k-1$ synthetic binary variables: the ith being 0 if the original nominal value is one of the first i in the sorted ordering and 1 otherwise.

Missing Values

To handle training instances with missing attribute values, the standard deviation reduction formula is modified to compensate for missing values. For an attribute for which $m > 1$ instances are missing, the formula, (15), re-written as

$$SDR = \frac{m}{|T|} \times \left[sd(T) - \left(\frac{|T_L|}{|T|} \times sd(T_L) - \frac{|T_R|}{|T|} \times sd(T_R) \right) \right]. \tag{20}$$

For instances with missing attribute values, a surrogate value is used during the construction of the model tree. The missing value is replaced with the average value of that attribute over the training instances that reach a particular node. Similarly, at the leaf level all missing values are replaced by the average values of the corresponding attributes of the training instances reaching the leaf.