NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

PREDICTING BATTLE OUTCOMES WITH CLASSIFICATION TREES

by

Muzaffer Coban

December 2001

Thesis Advisor: Thomas W. Lucas
Second Reader: Samuel E. Buttrey

Approved for public release; distribution is unlimited.

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
19 Dec 2001	N/A	-

Title and Subtitle Predicting Battle Outcomes with Classification Trees	**Contract Number**
	Grant Number
	Program Element Number
Author(s) Coban, Muzaffer	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Naval Postgraduate School Monterey, California	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es)	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)
Distribution/Availability Statement Approved for public release, distribution unlimited	
Supplementary Notes	
Abstract	
Subject Terms	
Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU
Number of Pages 126	

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** December 2001	**3. REPORT TYPE AND DATES COVERED** Master's Thesis	
4. TITLE AND SUBTITLE: Title (Mix case letters) Predicting Battle Outcomes with Classification Trees			**5. FUNDING NUMBERS**
6. AUTHOR(S) Muzaffer Coban			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A			**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution unlimited			**12b. DISTRIBUTION CODE**

14. ABSTRACT ***(maximum 200 words)***

Historical combat data analysis is a way of understanding the factors affecting battle outcomes. Current studies mostly prefer simulations that are based on mathematical abstractions of battles. However, these abstractions emphasize objective variables, such as force ratio. Models have very limited abilities of modeling important intangible factors like morale, leadership, and luck. Historical combat analysis provides a way to understand battles with the data taken from the actual battlefield. The models built by using classification trees reveal that the objective variables alone cannot explain the outcome of battles. Relative factors, such as leadership, have deep impacts on success. This result suggests that combat simulations will have a difficult time predicting combat outcomes unless we can better account for these intangible factors. Historical combat analysis helps us comprehend these factors. The classification model predictions on test sets reveal correct classification rates as high as 79 percent. Considering the variability in the data set this outcome is satisfying. Classification models also reveal that the factors affecting outcome of battles have changed throughout history. The leadership advantage played an important role for hundreds of years. However, in the 20th century, air sorties, tanks, and intelligence showed a higher importance.

14. SUBJECT TERMS Predicting battle outcomes, historical combat data, CAA, what relates to winning, classification and regression trees, important factors in battles, combat modeling			**15. NUMBER OF PAGES** 126
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UL

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

PREDICTING BATTLE OUTCOMES WITH CLASSIFICATION TREES

Muzaffer Coban
First Lieutenant, Turkish Army
B.S., Turkish Army Academy, 1996

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
December 2001

Author: _____
Muzaffer Coban

Approved by: _____
Thomas W. Lucas, Thesis Advisor

_____
Samuel E. Buttrey, Second Reader

_____
James Eagle, Chairman
Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Historical combat data analysis is a way of understanding the factors affecting battle outcomes. Current studies mostly prefer simulations that are based on mathematical abstractions of battles. However, these abstractions emphasize objective variables, such as force ratio. Models have very limited abilities of modeling important intangible factors like morale, leadership, and luck. Historical combat analysis provides a way to understand battles with the data taken from the actual battlefield. The models built by using classification trees reveal that the objective variables alone cannot explain the outcome of battles. Relative factors, such as leadership, have deep impacts on success. This result suggests that combat simulations will have a difficult time predicting combat outcomes unless we can better account for these intangible factors. Historical combat analysis helps us comprehend these factors. The classification model predictions on test sets reveal correct classification rates as high as 79 percent. Considering the variability in the data set this outcome is satisfying. Classification models also reveal that the factors affecting outcome of battles have changed throughout history. The leadership advantage played an important role for hundreds of years. However, in the 20^{th} century, air sorties, tanks, and intelligence showed a higher importance.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to Dr. Helmbold, who has made the historical data set available to us, for his guidance and help in my understanding of historical combat data analysis. I would also thank to my advisor, Prof. Tom Lucas, and my second reader, Prof. Samuel Buttrey, for their guidance, patience and help throughout the course of this thesis.

Special thanks extended to my wife, Mecra for her invaluable patience and understanding and to my son, Akif Umut, who made life happier for us.

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Predicting outcomes of battles has been a main concern of analysts throughout history. Analysts have tried to answer many questions relating to principles of war such as rates of advance, attrition, force ratios, and battle termination rules. These analyses were intended to give commanders a better understanding of battles, thus helping them make better decisions.

While making decisions about battles, commanders need a solid basis for their decisions. The chaotic and unpredictable nature of a war makes the decision process extremely difficult. Given this, can we find trends, models and guidelines that explain some of the phenomena of war? Are these consistent over time?

Simulations are widely used to understand and to predict battles. Simulations are built on mathematical models, like Lanchester equations and Stochastic Lanchester Models. However, these models are only simplifications of combat. Modeling intangible factors like fear, leadership, morale and surprise is complex and has proven difficult to do.

Another way of understanding and predicting wars is to use historical combat data. The data are collected from historical military archives and field reports. The statistical exploration of the data reveals historical trends, quantifies the importance of different variables, and suggests models. However, few historical data sets are available. In 1983, the U.S. Concepts Analysis Agency (CAA) contracted the Historical Evaluation and Research Organization (HERO) to prepare a detailed historical combat data set of 601 battles and engagements [Ref.5: p 1-1]. An updated version of the data set, namely the CDB90G, is the main data set used in this thesis.

This data set was first analyzed by the U.S. Concepts Analysis Agency (CAA) under the Combat History Analysis Study Effort, CHASE, which was initiated in 1984. Through this effort, they tried to find historically based quantitative results for use in "military operations research, concepts formulation, war gaming, and studies and analysis" [Ref.5: p 1-5]. The guiding elements of the effort, listed under the essential elements of analysis and answers are

1. Can the factors associated with victory in battle be identified?
2. What long-term trends can be detected in historical combat data?
3. Can the historical influence of air support on the outcome of the battles be quantified?
4. What can be said about the factors influencing the rates of advance in land combat?

As to the relationship of factors to victory, a logistic regression analysis revealed that the advantage parameter (ADV) had the strongest relationship. The probability of an attacker's victory was related to a logistic function of (the defender's) empirical advantage parameter (ADV).

In another effort, 427 non-WWII battles were used for the logistic regression [Ref.6]. By using this model, the authors tried to predict the outcomes of the 62 wars fought between 1823 and 1979. The best prediction for the combined data set was 72.5 percent correct [Ref.6: p 4-3].

In another CAA study [Ref.7], McQuie calculated 28 ratios and rates from the data set. The main purpose was to compare war game results with the data from historical battles. He wanted to set some criteria and standards for war games. McQuie pointed out that while future battles would not be replications of past ones, the most credible comparisons were with past battles.

Faruk explored CAA's revised version of the HERO database, the CDB90FT [Ref.3]. The data set consist of 660 battles and engagements with 140 different attributes. Faruk analyzed the 3-1 force ratio rule of thumb, the dispersion rates and the daily casualty rate. He divided the data into chronological subsets and analyzed each subset. He concluded that the force ratio was a reasonable predictor of outcomes, even though it is probabilistic.

In this research, we use the latest version of the data set, namely the CDB90G. The data set is detailed and contains many variables that cannot be used in classification purposes. In this study, some variables are pre-selected from the data set and they are used to build classification trees. Classification trees are more human readable and easier

to understand than the multiple logistic regression models. They don't need distributional assumptions, so transformations are not needed. Classification trees have also some capabilities to include variables with missing values.

The pre-selected variables are analyzed to show descriptive statistics, and conditional plots. The pre-selected variables are

1. Objective variables: force ratio, tank ratio, artillery ratio, cavalry ratio, the attacker's primary tactical scheme, and the defender's primary defensive posture.
2. Relative variables: relative surprise, relative air superiority in the theater, relative combat effectiveness, relative leadership advantage, relative training advantage, relative morale advantage, relative logistics advantage, relative momentum advantage, relative intelligence advantage, relative technology advantage, relative initiative advantage.
3. Terrain and weather variables: three terrain factors and five weather factors.

The descriptive statistics and conditional plots reveal the importance of the variables in determining the outcome of battles. The descriptive statistics reveal that the objective variables are not highly correlated with victory. However, some of the relative variables, such as leadership, have a strong relationship with the battle outcome, see Figure 1.

Using these variables, three models are considered. Model 1 uses only the objective variables in building classification trees. The predictions made with this model produced high misclassification rates. This result is parallel to the findings with descriptive statistics. Objective variables alone are not sufficient to classify battle outcomes. Model 2 uses both objective and relative variables. The resulting classification models have relatively low misclassification rates. The best of these makes 87 percent correct predictions (for subset 3, years 1920-1945). This result may be close to the limit of what we can predict, when we consider the variability in the data set and the role of luck in battles. Model 3 includes the terrain and weather variables, as well as the objective and relative variables. However, the resulting classification trees did not

include the terrain and weather variables. Moreover, the misclassification rates were no better then those of Model 2.

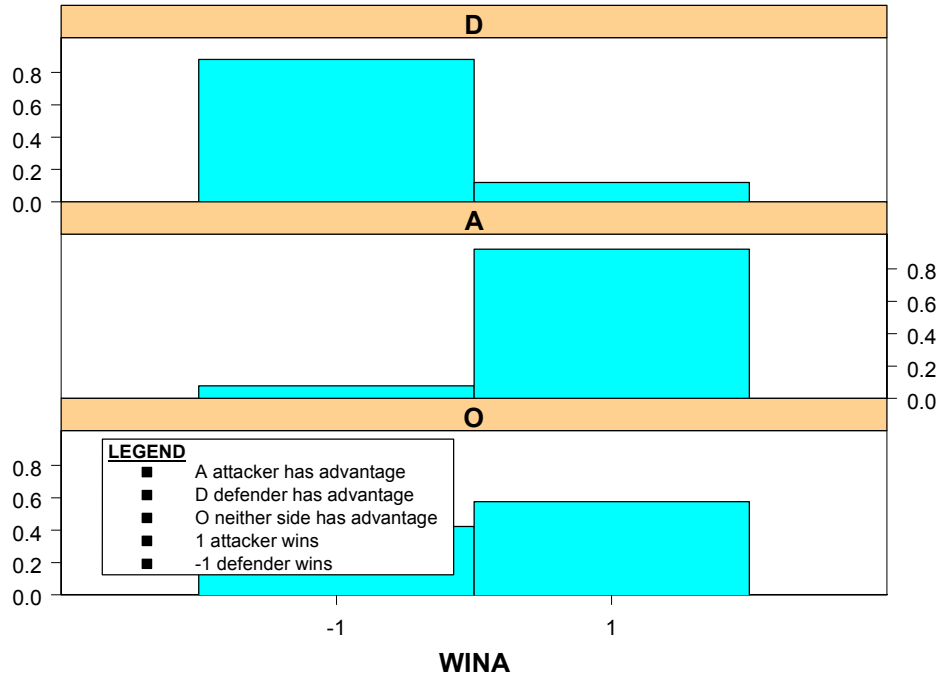

Figure 1. The Effect of the Leadership Advantage on Battle Outcome

When the defenders had a relative leadership advantage, they won 88 percent of the time, an increase of 48 percent compared to the initial battle outcome distribution. When the attackers had a relative leadership advantage, they won 92 percent of the time, an increase of 32 percent. When neither side had the advantage, the attacker won 58 percent of the time and the defender won 42 percent of the time, a two percent change. These results show that relative leadership advantage is highly correlated with the outcome of a battle.

In another analysis to understand the historical trends in battles, multiple classification trees are built by using objective and relative variables with training test sizes of 125. Each classification tree is built with a training set size of 125 and the battle after the 125 battles in the data set is predicted. Then, another classification tree is built with the next 125 battles, with an overlap of 124 battles. At the end, 658-125=533 classification trees are built and 533 predictions are made. This analysis revealed some important results. First, the importance of variables has changed throughout history. Second, the misclassification rates show that past battles failed to predict the battles of

World War II, in which new tactics and weapons were introduced to fighters. Figure 2 shows the variables that appeared in the first split of the classification trees. This figure reveals that the variables affecting the outcome of a battle have changed throughout history.

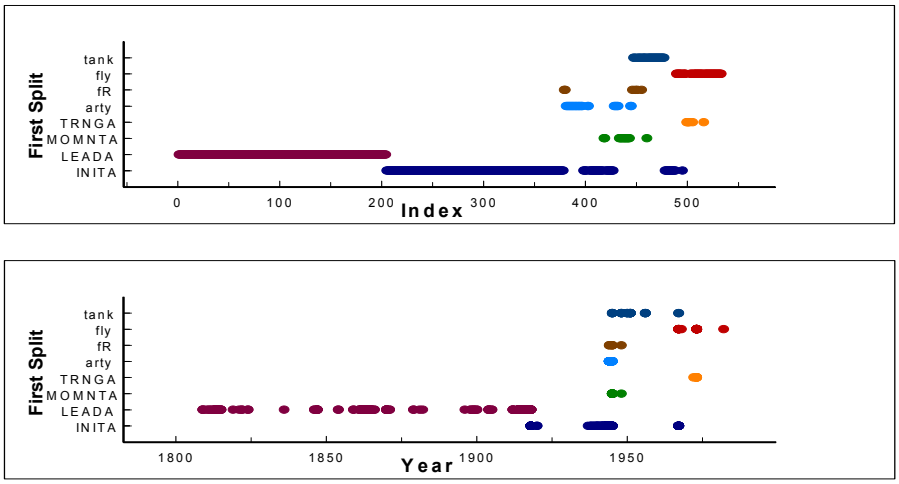

Figure 2. First Split Criteria of Classification Models

The relative leadership advantage appeared as the first split criterion in the first battles of the data set. The relative initiative advantage got importance in World War I. In World War II and after, objective variables, such as tank and CAS sorties ratio got precedence over other variables. The "index" shows the number of the classification tree in which the first split is taken. The "year" shows the last year of the 125 battles in the training set.

However, the validity of these results is directly related to the validity of the data set. Dr. Helmbold, who has supervised the CBD90G data set, pointed out that there may be some errors in the data set, and the models are affected by the diverse definitions and categories. He also added that despite its shortcomings, this data set was far and away the best available historical combat data base for statistical analyses.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION AND BACKGROUND

A. GENERAL INTRODUCTION

Predicting outcomes of battles has been a main concern of military analysts throughout history. Analysts have tried to answer many questions relating to principles of war such as rates of advance, attrition, force ratios, and battle termination rules. Helmbold [Ref.1] cites Vegetius 380 AD, whose work was published in 1944, as the first historian who worked on the rates of advance. Moreover, Helmbold compared the works of 34 analysts who studied the rates of advance. These analyses were intended to give commanders a better understanding of battles, thus helping them make better decisions.

While making decisions about battles, commanders need a solid basis for their decisions. The chaotic and unpredictable nature of a battle makes the decision process extremely difficult. For this purpose, many rules of thumb have been proposed to understand and to forecast battles. To name two [Ref.2: p. 5] "God is on the side of heavier battalions (Napoleon)," and "a successful attacker should be three times as strong as the opposing defender." For a long time, attackers used the 3 to 1 ratio to decide whether to attack. However, Faruk's analyses [Ref.3] shows that attackers won only 68 percent of the time when they had a three to one or greater advantage. In addition, Dupuy [Ref.2:p.5] contradicts some other "forecasting propositions from military history," such as "the numerically inferior force is usually victorious." Given this, can we find trends, models and guidelines that explain some of the phenomena of war? Are these consistent over time?

Simulations are widely used to understand and to predict battles. Simulations are built on mathematical models, like Lanchester equations and Stochastic Lanchester Models. However, these models are only simplifications of combat. Modeling intangible factors like fear, leadership, morale and surprise is complex. Rowland [Ref.4: p 46] shows that the degradation can be nine-tenth in real combat. In other words, a soldier's achievement in real combat can be one-tenth of his or her success in a simulation or even in a training exercise.

Another way of understanding and predicting battles is to use historical combat data. The data are collected from historical military archives and field reports. The statistical exploration of the data reveals historical trends, quantifies the importance of different variables, and suggests models. However, few historical data sets are available. In 1983, the U.S. Concepts Analysis Agency (CAA) contracted the Historical Evaluation and Research Organization (HERO) to prepare a detailed historical combat data set of 601 battles and engagements [Ref.5: p 1-1]. An analysis of the data set appeared in several papers [Ref.5, Ref.6, and Ref.7]. In addition, Dupuy [Ref.2] formed the Quantified Judgment Model by using historical combat data. Faruk did an exploratory analysis on the revised version of the CAA's database, called CDB90FT, which consists of 660 battles [Ref.3]. An updated version of the CDB90FT data set, namely the CDB90G, is the main data set used in this thesis.

B. BACKGROUND

Historical combat data sets are likely to contain errors and uncertainties in them. However, the CDB90G is regarded as the best available historical combat data. This thesis does not address data errors.

The literature about the analysis of the historical combat data is reviewed. The papers relating to the CAA's data set are a good source about the collection, sources, and veracity of the data set. They also provide information about the battles, factors related to victory, and rates of advance.

1. CHASE Study

The U.S. Concepts Analysis Agency initiated the Combat History Analysis Study Effort in 1984. Through this effort, they tried to find historically based quantitative results for use in "military operations research, concepts formulation, war gaming, and studies and analysis" [Ref.5: p 1-5]. The guiding elements of the effort, listed under the essential elements of analysis and answers are

1. Can the factors associated with victory in battle be identified?
2. What long-term trends can be detected in historical combat data?
3. Can the historical influence of air support on the outcome of the battles be quantified?

4. What can be said about the factors influencing the rates of advance in land combat?

The historical combat database has been analyzed for descriptive statistics, factors associated with victory, redundancy analysis, breakpoint analysis, rates of advance, air support and long-term trends. Of the factors associated with victory, six variables were considered [Ref.5: p 4-1]. Force ratio (FR)

1. Bitterness (EPS)
2. Casualty Exchange Ratio (CER)
3. Fractional Exchange Ratio (FER)
4. Advantage (ADV)
5. Residual Advantage (RESADV)

Detailed definitions of these variables are provided at the Appendix I.

A logistic regression analysis revealed that the advantage parameter had the strongest relationship with victory. The probability of an attacker's victory was related to a logistic function of (the defender's) empirical advantage parameter (ADV). The relationship between ADV and the battle outcome is given at Figure 3.

With respect to the dependence of victory on ADV, post-1940 era battles differed significantly from pre-1940 era battles. This was called the World War II anomaly. A number of hypotheses to locate the source of this anomaly were tested and some new steps were suggested, such as omitting Italian and Okinawan campaigns and analyzing the logistic regression fit after the outliers are eliminated.

Following the CAA work, our analysis will begin with exploring this relationship by using a classification tree.

2. Relationship between Casualties and Victory

In a following CAA study [Ref.6], similarities between battles and wars in accordance with casualties and victory were analyzed.

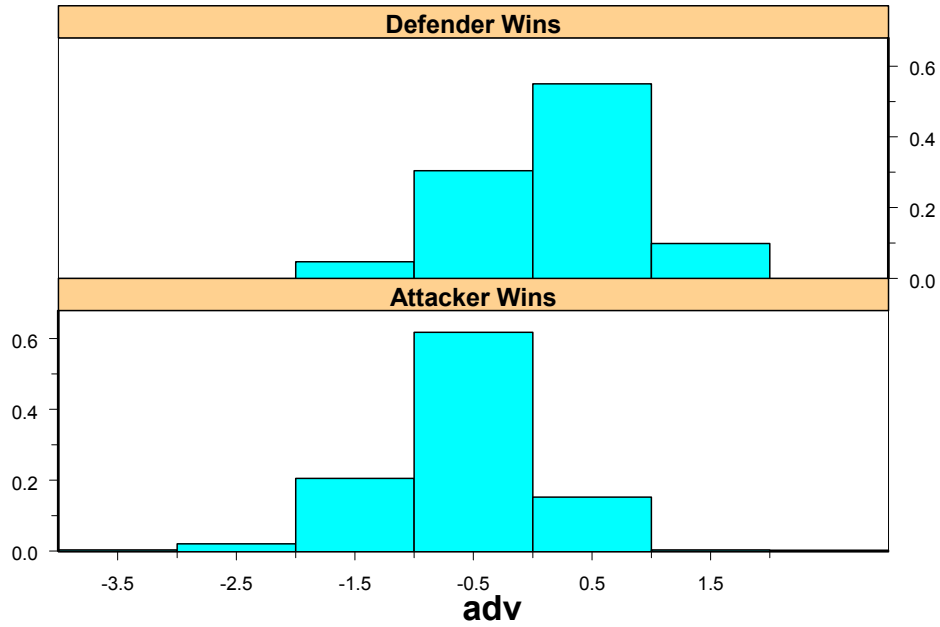

Figure 3. The Histogram of ADV

Defenders won the majority of battles that had high ADV values. The overlap in two histograms will be the cause of misclassified outcomes.

The relationship between victory and the defender's empirical advantage was proposed as:

$$P(ADV) = EXP (a + b * ADV) / (1 + EXP (a + b * ADV)) \quad (1)$$

where P(ADV) is the probability that the attacker wins a battle in which the defender's advantage relative to the attacker is ADV. The parameter's a and b are called the logistic regression intercept and slope. The relationship between the probability of victory and the ADV is given in Figure 4.

The (defender's) empirical advantage parameter was calculated from the casualties and personnel strengths.

$$ADV = (1/2) * LOG (FER) \quad (2)$$

with

$$FER = FX / FY \quad (3)$$

where FX and FY are the attacker's and defender's fractional losses.

In the paper, "Do Battles and Wars Have a Common Relationship between Casualties and Victory," [Ref.6], 427 non-WWII battles were used for the logistic

regression. Figure 4 shows the probability of the attacker's and the defender's winning as a function of ADV. The values obtained in the logistic regression are

Logistic regression intercept = a = −0.02017

Logistic regression slope = b = −4.87764

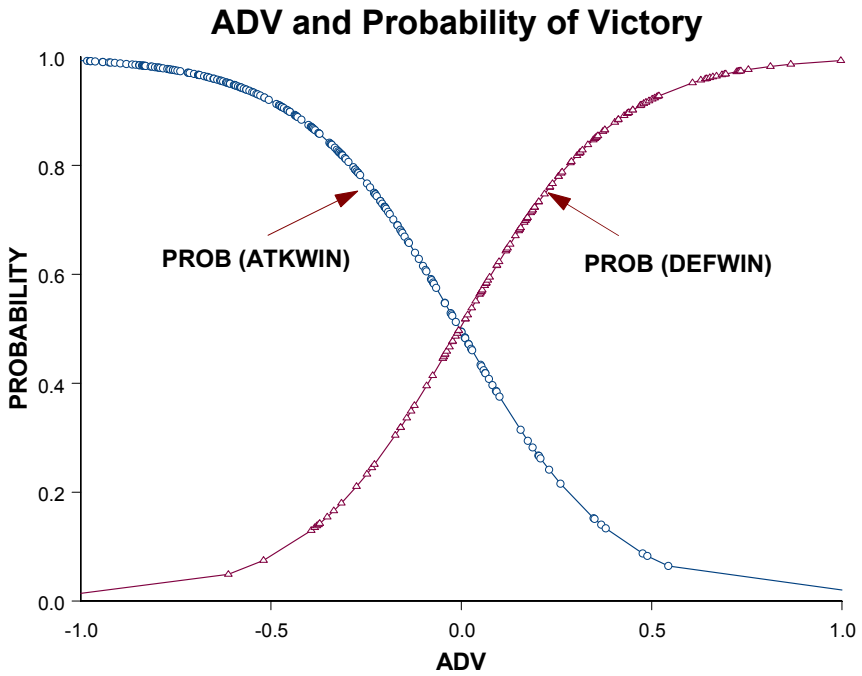

Figure 4. Probability of Battle Outcome for Non WW II Battles versus ADV. After [Ref.5: p 4-16]

By using this model, the authors tried to predict the outcomes of the 62 wars fought between 1823 and 1979. These wars were listed in another database called the "inter-state." This database was divided according to observations with high and low confidence levels with regard to the accuracy of the data. For the high confidence level, the best prediction was 85 percent correct, for the low confidence level 64.7 percent correct, and for the combined data set 72.5 percent correct [Ref.6: p 4-3].

3. Benchmarks

In another CAA study [Ref.7], McQuie calculated 28 ratios and rates from the historical combat database. The main purpose was to compare war game results with the data from historical battles. He wanted to set some criteria and standards for war games. If some of the criteria were violated, more runs would be needed to understand why that

happened in a war game. He pointed out that while future battles would not be replications of past ones, the most credible comparisons were with past battles. From the Army's database of historical battles of 601 wars, he selected 260 battles since 1937. McQuie chose only the characteristics "that could have been obtained from a war correspondent at the scene of the conflict." [Ref.7: p 4]. Some information about battles was lost, but military historians estimated that information after the battle. The missing values in the data set were another concern. These values could not be obtained because some historical records were destroyed or inaccessible.

In addition, the reliability of the data was analyzed. According to the military historians, the data from the Western European and Italian campaigns of WWII were the most reliable, whereas that from the Korean front was the least reliable. The reliability of the information from the Middle East wars stood in the middle, with data on the 1956 and 1973 wars being better than the data on the 1948 and 1967 wars.

McQuie gave measures of variability in the data; see Table 1. The variability is generally high because of the nature of combat. For example, during the Sinai campaign, the Israeli army moved at a rate of 45km per day, while US troops moved 100m per day at one of the engagements in Okinawa.

4. Faruk Yigit's Thesis

Yigit [Ref.3] explored CAA's revised version of the HERO database, the CDB90FT. The dataset consist of 660 battles and engagements with 140 different attributes. Faruk analyzed the 3-1 force ratio rule of thumb, the dispersion rates and the daily casualty rate. He divided the data into chronological subsets and analyzed each subset. He concluded that the force ratio was a reasonable predictor of outcomes. In that, for example, force ratio of 3 to 1 or greater lead to victory 68 percent of the time. Figure 5 shows the attacker's and defender's probability of winning as a function of the force ratio.

C. TREATMENT OF THE DATA

The main goal of this thesis will be classifying battle outcomes according to the response variable, WINA, which shows who wins (1 attacker wins, 0 draw, −1 defender

wins). Two battles in the data had an unknown outcome, so they were discarded. In addition, 6.5 percent of the battles resulted in draws. Following the CAA analysis [Ref.5: p 4-13] and [Ref.6: p 2-2], the draws were recorded as wins on the defenders' side. By regrouping into two categories, the response variable became binominal. The resulting data set had 658 rows, with 398 outcomes favoring the attacker and 260 favoring the defender.

Characteristic	Type of Charac-teristic	Year	Attacker & Defender	Value of Characteristic	Ratio of High to Low Values
Force Ratio Men	High :	1967	Egypt: Israel	17:1	57:1
(atkr:dfdr)	Low :	1945	Japan:USA	0.3:1	
Force Ratio Artillery	High :	1945	USA:Japan	50:1	450:1
(atkr:dfdr)	Low :	1948	Israel:Syria	0.11:1	
Mortar Density dfdr	High :	1943	Britain:Germany	132	730:1
(wpns/km)	Low :	1973	Egypt:Israel	0.19	
Artillery Density atkr	High :	1944	USA:Japan	444	2200:1
(wpns/km)	Low :	1948	Israel:Jordan	0.2	
Casualty Rate atkr	High :	1945	USA:Japan	96%	740:1
(% per day)	Low :	1944	Britain:Germany	0.13%	
Tank Loss Rate atkr	High :	1967	Israel:Syria	92%	150:1
(% per day)	Low :	1944	USA:Germany	0.63%	
Advance Rate	High :	1967	Israel:Egypt	45	450:1
(km per day)	Low :	1945	USA:Japan	0.1	

Table 1. Variability of Characteristics from [Ref.7: p.13]
All the Ratios Vary Considerably.

In the data set, the column names are recorded in capital letters. In order to differentiate the original columns and our calculations, such as the force ratio, calculations are added to the data set with names starting with lower-case letters.

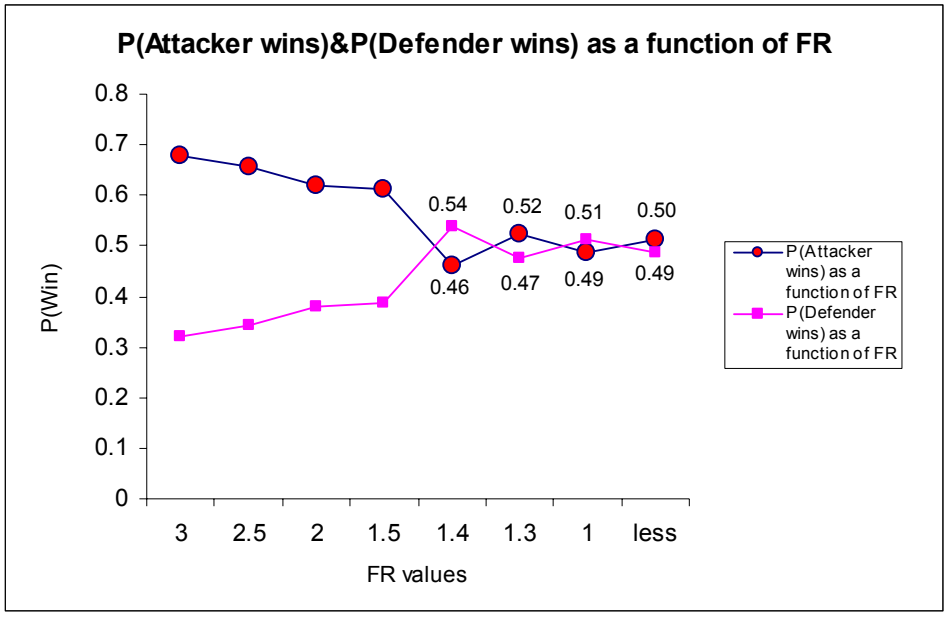

Figure 5. P (Attacker wins) and P (Defender wins) Values as a Function of Force Ratio from [Ref.3: p. 67]

The missing values in the data set will be handled by a special classification tree library, rpart (recursive partitioning), in S-PLUS [Ref.8 and Ref.10]. During the recursive partitioning, an alternative split (surrogate) is selected for the missing values. The procedure is best described in [Ref.8: p. 318] as:

> …The surrogates, alternative splits, are used to handle missing cases both in training and in prediction (including cross-validation to choose complexity). Each of the surrogate splits is examined in turn, and if the variable is available, that split is used to decide whether to send the case to the left or right. If no surrogate is available or none can be used, the case is sent to the majority….
>
> …
>
> When missing values are encountered in considering a split they are ignored and the probabilities and impurity measured are calculated from the non-missing values of that variable. Surrogate splits are then used to allocate the missing cases to the daughter nodes.

This method of handling missing values is important to our data. In the data, every row has one or more missing values. Other classification models omit rows with missing values, so they allow the user to include only those columns that have less missing values in the model. However, with the `rpart` library we can use all related columns in the data set to build our model and to make predictions. In this way, we are using as much data as possible.

The data are divided into five subsets; see Table 2. We know that weapons and tactics have changed tremendously during history. The weapons of the 1600's cannot be compared to the weapons of World War II. Looking for trends that have proved invariant over time, we divided the data set into subsets. The five subsets were selected so as to have an approximately equal number of battles in each.

Each subset will be divided into two parts. The first two-thirds of the subset will be the "training set." The tree models will be built by using the training set. The last one-third will be the "test set." This test set will be used in prediction and cross-validation. The predictions will help us answer the question, "Can we predict the outcomes of future battles?"

SUBSET NO	SUBSET	SIZE	TRAINING SET	SIZE	TEST SET	SIZE
1	1600 -1847	164	1600 - 1799	109	1799 - 1847	55
2	1805 - 1918	260	1805 - 1915	178	1916 - 1918	82
3	1920 - 1945	202	1920 - 1944	131	1944 - 1945	71
4	1940 - 1982	223	1940 - 1948	150	1950 - 1982	73
5	1600 - 1982	658	1600 - 1944	435	1940 - 1982	223

Table 2. The Division of Data into Subsets

The first subset consists of battles before 1847. The training set will be the battles before 1799. Using the model built from the training set, we will predict the outcomes of battles between 1799 and 1847. The second subset consists of battles between 1805 and 1918. The outcomes of later battles of WWI will be predicted by the model built from the battles after 1805 and WWI battles through 1915. The third subset consists of battles between 1920 and 1945. There are nine battles between 1920 and 1939, so this subset is mainly from World War II. The battles of 1944 and 1945 are predicted from the model trained by the early World War II battles. The fourth subset consists of battles between 1940 and 1982. The outcomes of battles from the Arab-Israeli wars are predicted using the model of late World War II battles. The fifth subset is the entire data set. The

training set is made up by the wars before 1940. The wars after 1943 will be the test set. This subset will reveal historical trends.

D. TREE-BASED MODELS

One efficient way of classifying the outcome of the battle is using classification trees [Ref.9]. Classification trees are more human-readable and easier to understand than multiple logistic regression models. Trees simply show the structure in the data. In addition, the output can be coded as a rule set and this can be implemented in other software, like Visual Basic, Java or C. For example, the output of a tree can be put into pseudo code as [Ref.9]:

if (force ratio>2) and (SURPA=A)

then WINA is most likely to be in level 1(Attacker wins)

Trees do not need distributional assumptions, so transformations are not needed. Any interactions between variables are automatically included in the tree structure.

Trees are arranged hierarchically. Until a terminal node is reached, the data flowing down the tree encounters one decision at a time. Special cases can be used without affecting other decisions.

The tree is constructed in an iterative process using the `rpart` package [Ref. 10] in the S-Plus data analysis system [Ref. 11]. All the observations start in a single group or "node." Then every possible unique split of the form "$x < x_0$" (for a continuous x) or "$x \in$ subset i" (for a categorical x) is examined. The split that reduces the multinomial log likelihood the most is chosen. This split separates the data into two pieces. Then each of those pieces in turn is split, and so on, until no more splitting is sensible or possible.

The tree model formed in this way is generally "over-fit," that is, too tightly bound to the peculiarities of the training set. In order for the model to generalize well to future data, the tree is "pruned" using cross-validation. In this process sub-trees of different sizes are constructed with 90% of the data set and their misclassification rates on the remaining 10% are computed. This is done ten times with each item in the data held out one time. Then, the misclassification rates are aggregated over the replications.

The “optimal” tree size is the one whose aggregated misclassification rate is smallest. Finally, the tree we choose for prediction follows the “one SE” rule of [Ref. 8], in which we choose the smallest tree whose cross-validated error rate is no larger than one standard error above the optimal one. Occasionally the one SE rule chooses a one-node tree; in these cases we have chosen the entire tree.

Classification trees have special advantages over other classification methods [Ref.9: p.378]:

1. In certain applications, especially where the set of predictors contains a mix of numeric variables and factors, tree-based models are sometimes easier to interpret and to discuss than linear models.
2. Tree based models are invariant to monotone re-expressions of predictor variables, so the precise form in which these appear in a model formula is irrelevant.
3. The treatment of missing values (NA) is more satisfactory for tree-based models than for linear models.
4. Tree-based models are more adept at capturing nonadditive behavior; the standard linear model does not allow interactions between variables unless they are pre-specified and of a particular multiplicative form.

Graphs of classification trees are easy to understand. The oval and rectangular boxes represent non-terminated nodes and terminal nodes respectively. The levels inside the boxes show the majority level of that node. The boxes also contain the number of battles the defender won and the number of battles the attacker won in that node. For example, in the root node of Figure 6, the majority level is “1”, "attacker wins;" the attacker won 71 battles and the defender won 38. A split criterion is written on the branches. If the criterion is satisfied, the observation is sent into the following node on that branch. For instance,

if cavalry ratio, “cav,” is less than 1.060 and

if force ratio, “fR,” is less than 0.848 and

if the attacker’s primary tactics, “PRIA,” is frontal assault, “FF,” and

if force ratio, “fR,” is greater than 0.704

then the battle is classified as “1”, attacker wins.

There are 11 battles that satisfy all the criteria in the training set. Of these battles the defender won four battles and the attacker won seven battles. The attacker won more than the defender, so this leaf is assigned as the attacker wins.

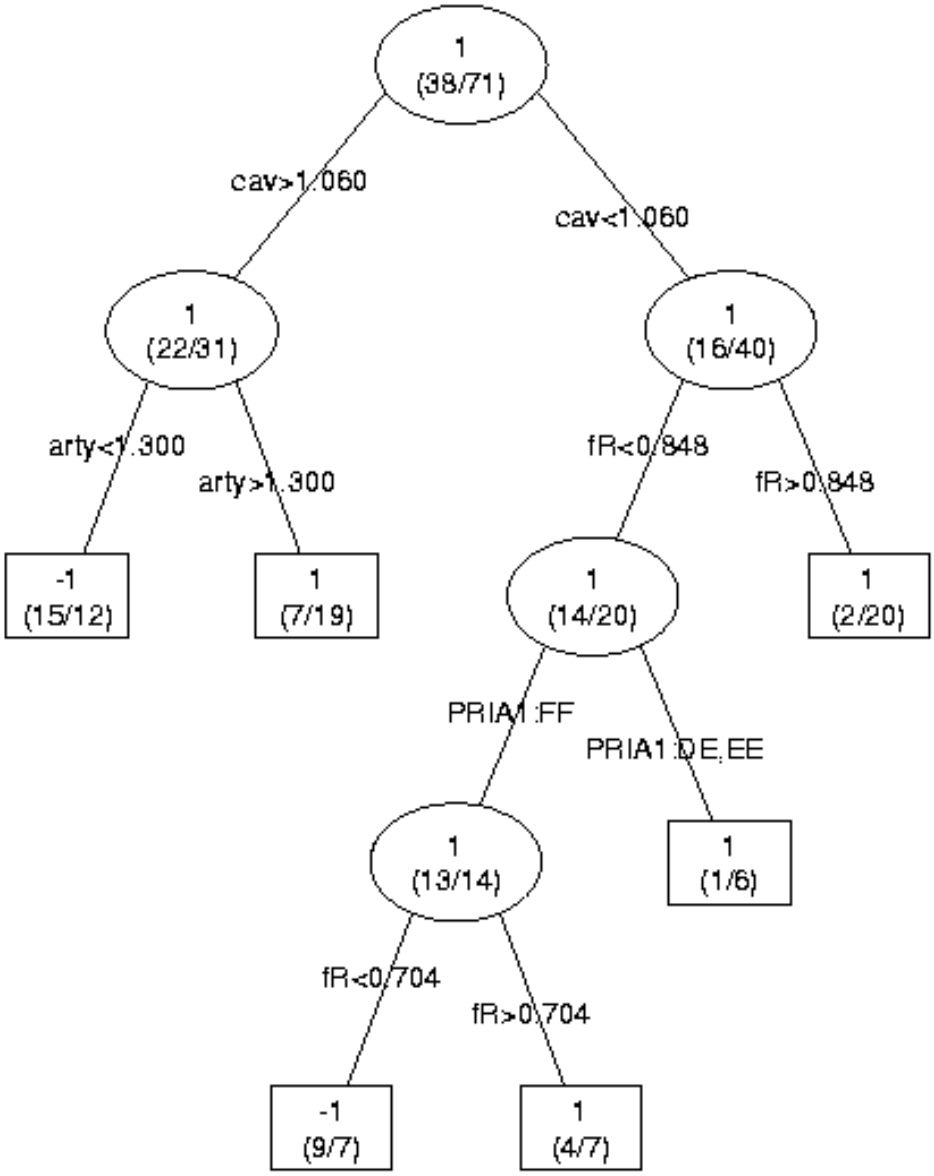

Figure 6. A Sample Classification Tree

E. REVISITING THE CAA STUDY

In the first analyses of the data set by CAA [Ref.5], a simple relation between the probability of winning and the defender's empirical advantage parameter (ADV) was found. The relation is formulated in Equation 1 and Equation 2. However, CAA used the ADV to show the relationship between casualties and victory. Neither the ADV nor the force exchange ratio can be used as predictor variables of battles because each is a function of battle casualties, and casualties are known only after the battle. Our analysis will begin by exploring the relationship between victory and force exchange ratio, but our main concern will be building models by using variables that can be known before a battle.

Since classification trees do not require transformation of the variables, the relationship between winning and the force exchange ratio (FER) can be modeled in natural units, i.e. no logarithmic transformations are needed.

In order to understand the relationship between the battle outcome and the force exchange ratio, the following tree, Figure 7, is built by using the entire data set and cross-validating within the data set. This model correctly explains 78 percent of the battle outcomes in the entire data set of 658 battles. This model fits the data set, so in order to show whether this trend changed over history, we need to build models in training sets composed of earlier battles and to predict on the test sets that are composed of later battles.

The results of splitting the data into five subsets are summarized in Table 3. The test sets consisted of about one third of the data.

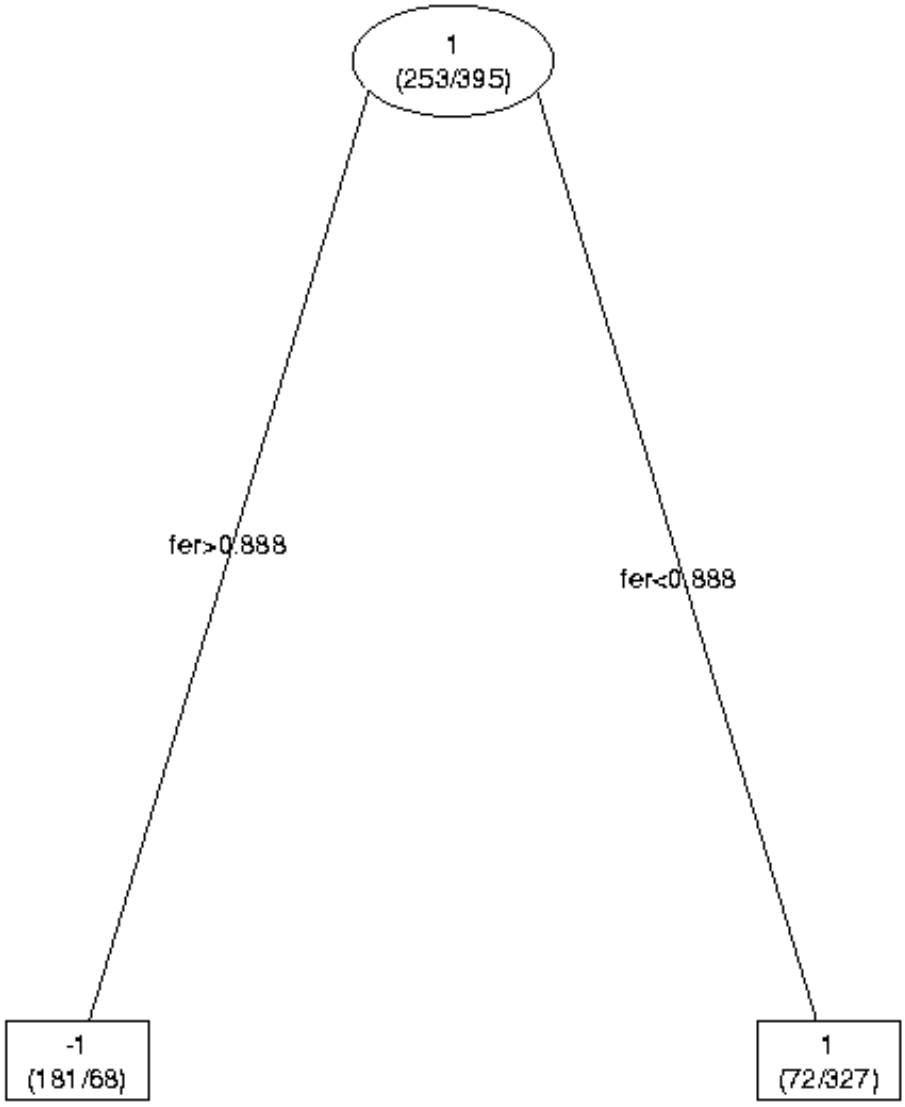

Figure 7. The Classification Tree Model

On the leaves, "1" indicates that the outcome favored the attacker and "–1" shows that the outcome favored the defender.

SUBSETS	ERROR RATE ON THE TRAINING SET	ERROR RATE OF PREDICTIONS ON THE TEST SET
SUBSET 1	0.14	0.09
SUBSET 2	0.20	0.28
SUBSET 3	0.12	0.22
SUBSET 4	0.26	0.20
SUBSET 5	0.21	0.28

Table 3. Error Rates of Models

The models build by using the training sets can correctly predict battles in the test sets more than 72 percent of the time by using only one factor (fer).

F. RESEARCH QUESTIONS

1) Can we build classification trees to classify the outcomes of battles?

2) What factors contribute to winning?

3) Do these factors change over time?

4) What are the effects of the tactics of the defender and the attacker?

5) What are the effects of weapon systems?

6) Is surprise an important factor in winning?

G. SCOPE OF THE THESIS

Descriptive statistics about factors such as weather, terrain, tactics, and force ratios will be calculated from CDB90G data by using the S-PLUS software. Furthermore, classification models will be built and used for predictions. Thus, the thesis will consist of:

1) Literature review about historical combat data analysis;

2) Summary statistics and plots;

3) Building classification models to predict the outcome of the battles;

4) Predicting new data from another data set or within the CDB90G dataset;

5) Conclusions about the findings and results of the analysis.

THIS PAGE INTENTIONALLY LEFT BLANK

II. SUMMARY STATISTICS

A. INTRODUCTION

> In your deliberations, when seeking to determine the military conditions, let them be made the basis of a comparison, in this wise:
>
> (1) Which of the two sovereigns is imbued with the Moral law?
>
> (2) Which of the two generals has most ability?
>
> (3) With whom lie the advantages derived from Heaven and Earth?
>
> (4) On which side is discipline most rigorously enforced?
>
> (5) Which army is stronger?
>
> (6) On which side are officers and men more highly trained?
>
> (7) In which army is there the greater constancy both in reward and punishment?
>
> By means of these seven considerations I can forecast victory or defeat.
>
> – Sun-Tzu, *The Art of War*

In this chapter, the CDB90G data set will be analyzed. The data set has many variables that cannot be used for predicting battle outcomes, such as degree of influence of relative factors (14 variables). Some variables have too many levels to be useful, like battle location, commanders' names, and campaign. The data set contains a great number of variables that show the dates and times of the attack's start and end (79 variables) and of times that defensive fronts become effective (15 variables). To find the variables that relate to winning, we first made a pre-selection of the variables. The pre-selection was done according to the author's military judgment. The pre-selected variables are divided into two groups, objective and relative variables.

Objective variables are those that can be collected from the battleground. These variables can be known before the battle occurs and almost everyone can agree on their value, at least within some tolerance. These variables are the most important to our models. Our data set gives total strengths, number of cavalry, artillery tubes, tanks, and air sorties for both the attacker's and defender's side. Local terrain and weather conditions, posture of the defender, and tactics of the attacker are also included in the data set.

Relative variables, such as surprise, combat effectiveness, leadership advantage, and so on, are based on the judgment of military historians. The preliminary analysis showed that objective variables alone fail to describe the outcome of most battles. In order to build a model that can be used to understand and to predict battles, relative values need to be included.

In the data set, some relative variables (relative combat effectiveness, leadership, training, morale, logistics, momentum, intelligence, technology, and initiative) are given with levels ranging from "–4" to "+4." A level of "–4" shows that the variable very strongly favors the defender, while "+4" shows that the variable very strongly favors the attacker. The other levels come between, and level "0" favors neither side. Another relative variable, surprise, is given in a scale ranging from "–2" to "+2." Negative levels show the defender's surprise and positive ones show the attacker's surprise. Before the battle, knowing the relative advantage on this scale is difficult. However, we may know which side has the advantage before the engagement. Thus, the relative variables have been modified to show only which side, if any, had the advantage, without showing the relative levels.

Weapons effects are expressed as ratios. In some battles, the attackers had no weapons of a particular type. This made the ratio zero, which gives no information about the number of the defender's weapons. In some other cases, the defender had no weapons and that makes the ratio infinity. Adding a constant to both sides avoids these two pitfalls. Therefore, in finding ratios, one is added to each side's strength. When neither side had that weapon, a missing value indicator is assigned to the ratio variable.

B. DESCRIPTIVE STATISTICS

Descriptive statistics help us understand the properties of the variables. Box plots and histograms give a graphical picture of the distribution, median and outliers. The descriptive statistics of the response variable and pre-selected variables that can affect the outcome of war follow.

1. Response Variable

a) Battle Outcome: "WINA"

> Security against defeat implies defensive tactics; ability to defeat the enemy means taking the offensive.
>
> − Sun-Tzu, *The Art of War*

This variable shows the outcome of a battle as a win for the attacker, a win for the defender or a draw. In our classification model, "WINA" will be the response variable. The distribution of "WINA" is given in Table 4.

WINA	**Unknown Outcome (–9)**	**Defender Wins (–1)**	**Draw (0)**	**Attacker Wins (1)**
Size	2	217	43	398

Table 4. The Distribution of the Battle Outcome Variable, "WINA"

The two battles with unknown battle outcome values will be discarded from the data set. In the remaining set, 43 observations (7 percent of the data) are draws. Following the CAA's approach, in order to make the response variable Bernoulli, draws will be regarded as a win for the defender. The new distribution of the WINA is given in Table 5.

WINA	**Defender Wins –1**	**Attacker Wins 1**
Size	260	398

Table 5. The Revised Distribution of the Response Variable, "WINA"

In this table, 60 percent of outcomes favor the attacker, 40 percent favor the defender. The attacker had a higher chance of winning the battle. This, by itself, suggests that our doctrine should seek offense as a primary course of action (COA).

2. Objective Variables

a) Force Ratio (Manpower Ratio): "fR"

> The superiority in numbers is the most important factor in the result of a combat, only it must be sufficiently great to be a counterpoise to all the other co-operating circumstances. The direct result of this is, that the greatest possible number of troops should be brought into action at the decisive point.
>
> − Clausewitz, *On War*

Force ratio has been considered an important predictor of battle outcome throughout history [Ref.2]. However, the conditional plots of force ratios when the attacker won and when it lost look similar. The median force ratio when the attacker won the battle is 1.6. The median when the defender won the battle is 1.3. When the attacker had a force ratio greater than 1.6, it won 65 percent of the time. When the attacker had a force ratio of 1.3 or less, the attacker won 55 percent of the time. The spread of force ratio is given in Figure 8.

In order to see whether medians of two conditional samples of force ratio (force ratio when the attacker wins and when it loses) are the same (in other words, the difference may be explained by chance), Wilcoxon's rank-sum test will be used. Wilcoxon's rank-sum test is like a two-sample t test; however, Wilcoxon's rank-sum test does not assume the sample comes from a parametric family, e.g., a normal population [Ref.14]. The assumptions of the Wilcoxon's rank-sum test are that two samples are randomly and independently chosen from continuous distributions, and they have medians $\mu 1$ and $\mu 2$ respectively [Ref.14: p.659]. The samples are assumed to have the same shape and the same spread. The only difference between them is, possibly, their medians. Note: Wilcoxon's rank-sum test is robust to these assumptions.

Our hypothesis is that the medians of two samples are equal. This hypothesis will be tested against the alternative hypothesis that the median of force ratio when the attacker wins is greater than the median of force ratio when the defender wins.

H_0: $\mu 1 - \mu 2 = 0$

H_a: $\mu 1 - \mu 2 > 0$

The Wilcoxon's rank-sum test reveals a p-value of 0.0019. "The P-value is the probability, calculated assuming H_0 is true, of obtaining a test statistic value as least as contradictory to H_0 as the value that actually resulted" [Ref.14: p.342]. At the five-percent significance level, the null hypothesis is rejected. The median of the force ratio when the attacker wins is greater than the median of the force ratio when the defender wins.

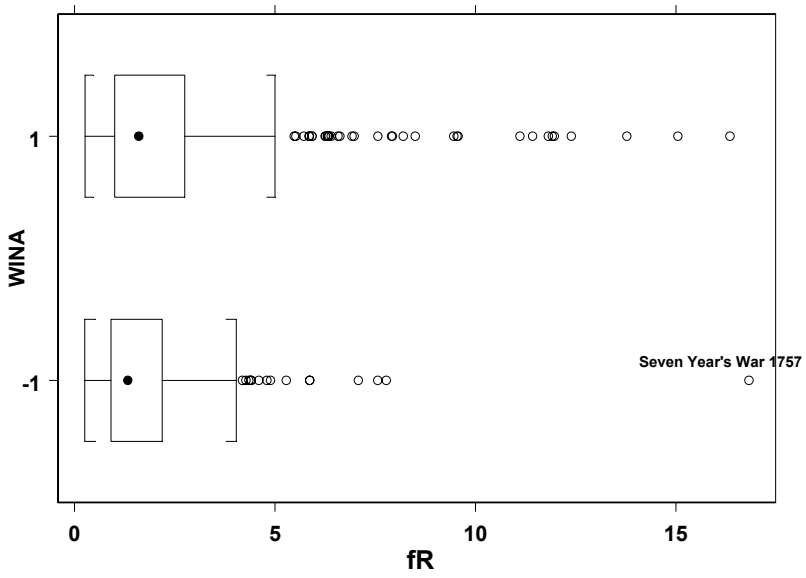

Figure 8. The Spread of Force Ratio, "fR"

The attacker had a slightly better chance of winning when it had higher force ratio values. However, the defender won the battle 31 percent of the time when the attacker had a force ratio advantage of more than three to one. In one battle, when the Bengali army attacked the British army in 1757, the attacker lost despite a force ratio advantage of more than sixteen to one.

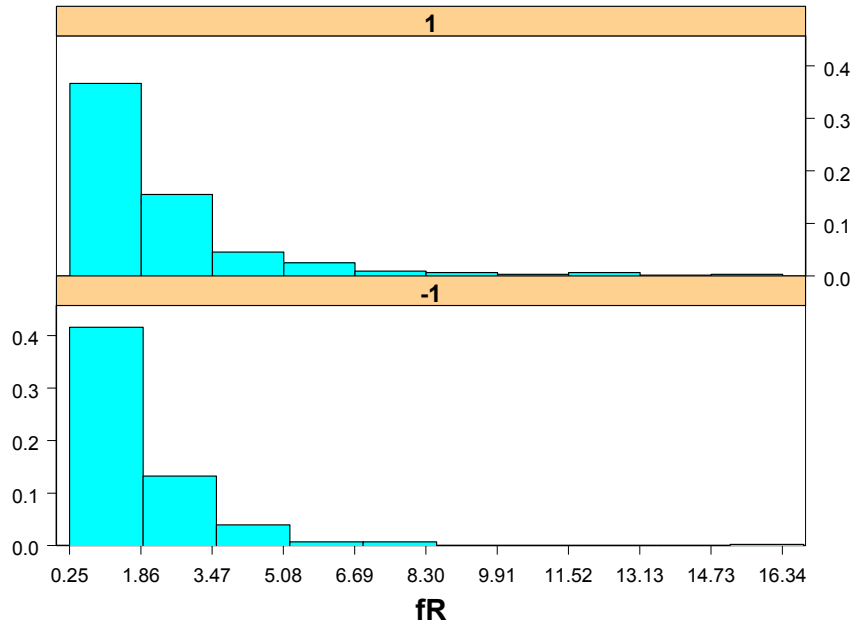

Figure 9. The Conditional Histogram of Force Ratio, "fR"

The two histograms look similar. However, the attacker's histogram has a longer tail than the defender's.

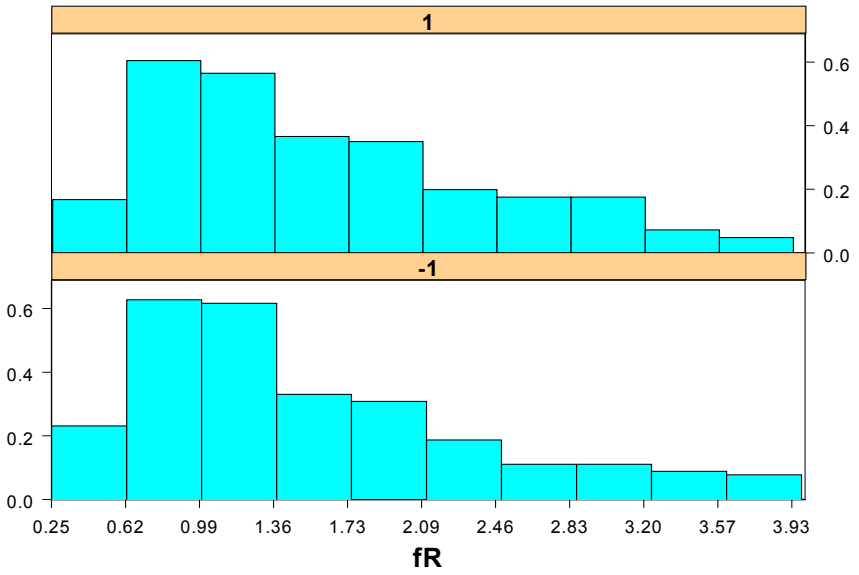

Figure 10. The Truncated Conditional Histogram of Force Ratio, "fR"

A truncated conditional histogram reveals the same result. The effect of the force ratio on the battle outcome is not very significant. We do not expect the force ratio variable to appear as an important variable in most of the classification models.

b) Artillery Ratio "arty"

Artillery is the only weapon in the data set that has been used from the first battles to the last ones. Artillery played a particularly important role during the Napoleonic campaigns. The median of artillery ratio where the attacker won the battle is 1.5 and where the attacker lost is 1.26. Conditional plots reveal that artillery cannot be the only predictor of battle outcomes. The spread of the artillery ratio is given in Figure 11. The conditional histogram of the artillery ratio is given in Figure 12.

The Wilcoxon's rank-sum test is used to see whether the median of artillery ratio when the attacker wins is the same as the median of artillery ratio when the defender wins.

H_0: $\mu 1 - \mu 2 = 0$

H_a: $\mu 1 - \mu 2 > 0$

The Wilcoxon's rank-sum test reveals a p-value of 0.0292. At the five-percent significance level, the null hypothesis is rejected. The median of the artillery ratio when the attacker wins is greater than the median of artillery ratio when the defender wins.

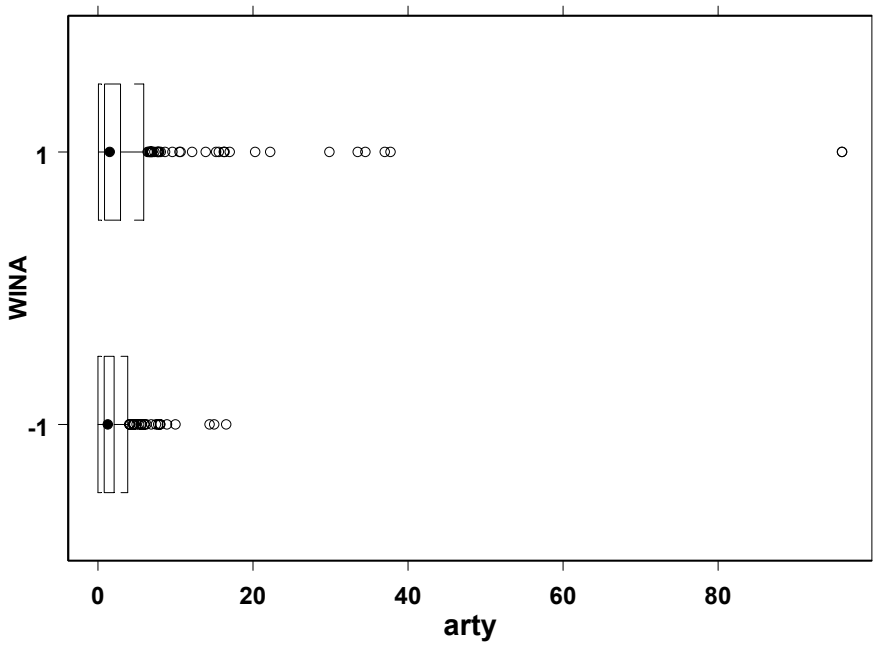

Figure 11. The Spread of Artillery Ratio, "arty"

The median and interquartile ranges of both box plots are close to each other. Huge outliers are present in both box plots. The defender won 28 percent of the time when the attacker had an artillery ratio of three to one or greater.

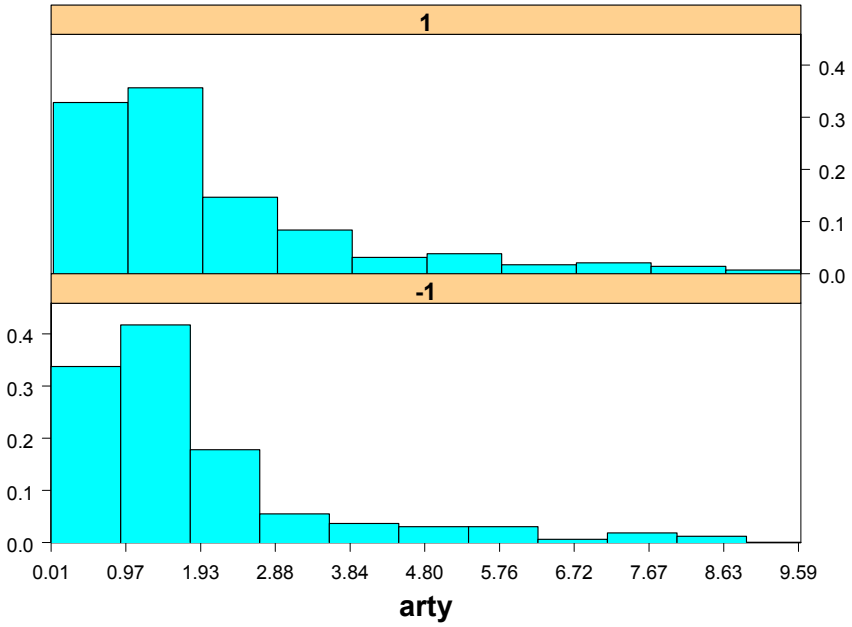

Figure 12. The Conditional Histogram of Artillery Ratio, "arty"

The two conditional histograms look similar. The attacker's histogram is slightly to the right and the defender's histogram is slightly to the left. These graphs show that there must be factors other than artillery ratio that relate to the outcome of battles.

c) Close Air Support Ratio "fly"

Air power was widely used in World War II and has been used since. A main goal in today's battles is to gain air superiority in the theater and launch the maximum number of air sorties. In the Gulf War and the Kosovo War, air power was the decisive factor.

The median of the close air support (CAS) sorties ratio in which the attacker won the battle is 14.27. The median of air sorties ratio in which the attacker lost the battle is 2.12. The difference in the medians is noticeable. The attacker won 73 percent of the battles when it had a CAS sorties ratio of 14.27 or greater. The defender won 52 percent of the battles when the attacker had a CAS sortie ratio of 2.12 or less. This result shows that a higher CAS sorties ratio gives the attacker a better chance and a low ratio gives the defender a better chance to win the battle.

The Wilcoxon's rank-sum test is used to see whether the median of CAS sorties ratio when the attacker wins is the same as the median of CAS sorties ratio when the defender wins.

H_0: $\mu 1 - \mu 2 = 0$

H_a: $\mu 1 - \mu 2 > 0$

The Wilcoxon's rank-sum test reveals a p-value of 0.0003. At the five-percent significance level, the null hypothesis is rejected. The median of CAS sorties ratio when the attacker wins is greater than the median of CAS sorties ratio when the defender wins.

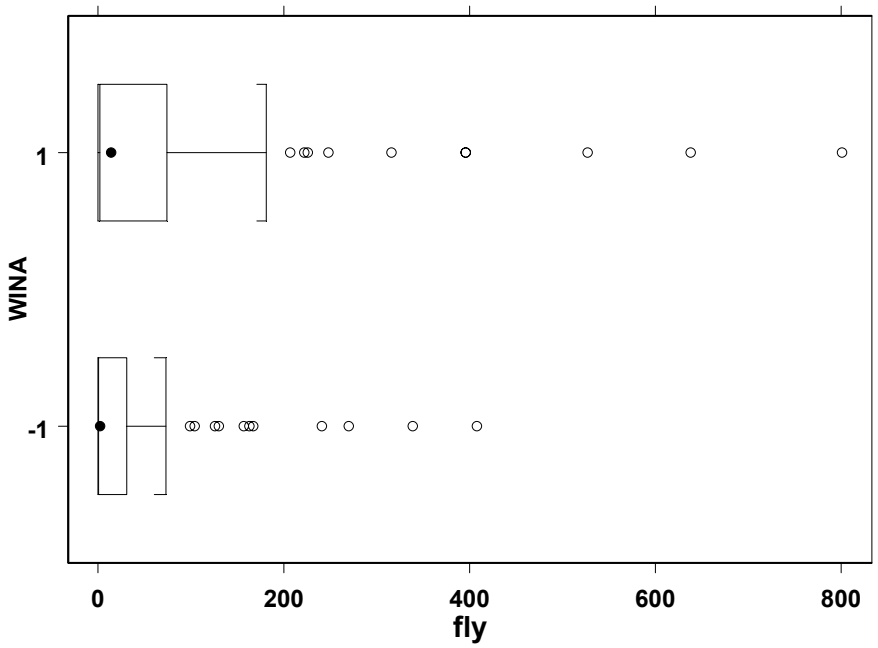

Figure 13. The Spread of CAS Sorties Ratio, "fly"

The box plots of CAS sorties ratio in which the attacker won reveals a larger interquartile range than the other box plot in which the attacker lost the battle. When the attacker had a higher CAS sorties ratio value, it had a higher chance of winning the battle. In addition, outliers as large as 800 are present in the air sorties ratio. Outliers show the need to use robust classification techniques, such as classification trees.

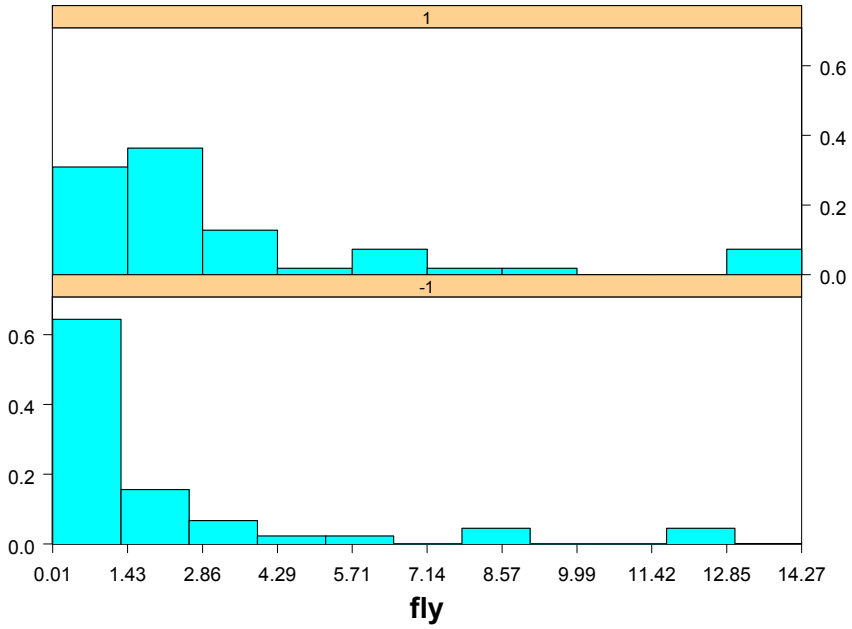

Figure 14. The Truncated Conditional Histogram of CAS Sorties Ratio, "fly"

The attacker had a greater chance of winning the war when the CAS sorties ratio value was large, and a higher change of losing the battle when the ratio was small. The defender won 27 percent of the time when the attacker had a CAS sorties ratio of three to one or less.

d) Tank Ratio "tank"

In the data set, tanks (light tanks and main battle tanks) were present from 1906 to 1982. The median of the tank ratio when the attacker won the battle is 3.7. For battles won by the defender, the median is 2.06. The attacker won 77 percent of the battles when it had a tank ratio advantage of 3.7 or greater. The defender won 42 percent of the battles when the tank ratio was less than 2.06. The spread of the tank ratio is given in Figure 15.

The Wilcoxon's rank-sum test is used to see whether the median of tank ratio when the attacker wins is the same as the median of tank ratio when the defender wins.

H_0: $\mu 1 - \mu 2 = 0$

H_a: $\mu 1 - \mu 2 > 0$

The Wilcoxon's rank-sum test reveals a p-value of 0.0092. At the five-percent significance level, the null hypothesis is rejected. The median of tank ratio when the attacker wins is greater than the median of tank ratio when the defender wins.

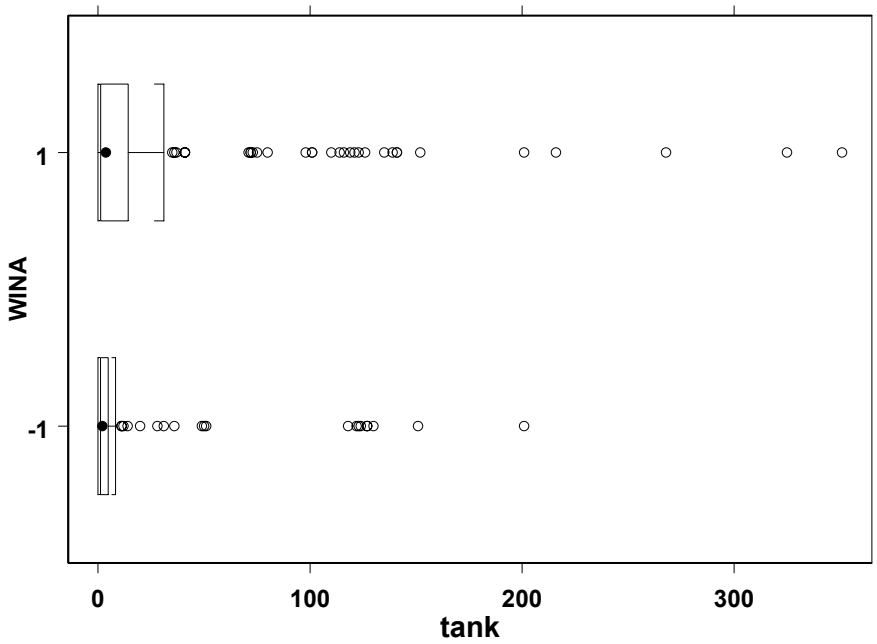

Figure 15. The Spread of the Tank Ratio, "tank"

The interquartile range of the first box plot is larger than the second, showing that the attacker had a higher chance of winning the battle when it had a higher tank ratio. However, in some battles the attacker had a ratio of more than 100 to 1 and still lost. Outliers in the figure reveal the need for robust classification techniques, like the classification trees.

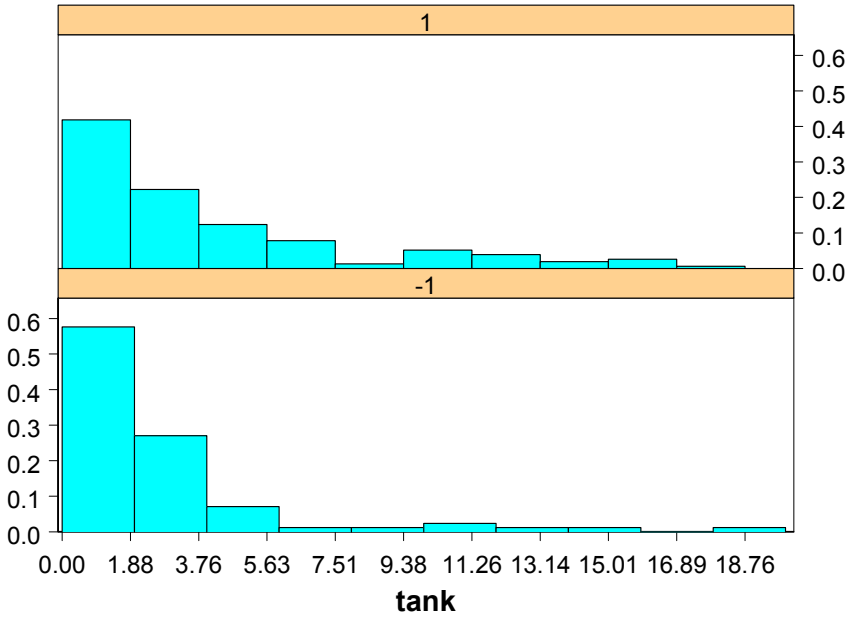

Figure 16. The Truncated Conditional Histogram of Tank Ratio

The attacker had a higher chance of winning the battle when it had higher tank ratios. The defender won 26 percent of the time when the attacker had a 3/1 or greater tank ratio. We would expect the tank ratio to be present in the classification models.

e) Cavalry Ratio "cav"

In the data set, the cavalry ratio variable is present in battles from 1600 to 1905. The spread of the cavalry ratio is given in Figure 17. The median of the cavalry ratio when the attacker won the battle is 1. The median of the cavalry ratio when the defender won the battle is 1.2. The second median is greater than the first one, meaning that the defender won more battles that had a high cavalry ratio favoring the attacker. The conditional histogram of the cavalry ratio is given in Figure 17.

The Wilcoxon's rank-sum test is used to see whether the median of cavalry ratio when the attacker wins is the same as the median of cavalry ratio when the defender wins.

H_0: $\mu 1 - \mu 2 = 0$

H_a: $\mu 1 - \mu 2 > 0$

The Wilcoxon's rank-sum test reveals a p-value of 0.947. At the five-percent significance level, the null hypothesis is not rejected. The median of the cavalry ratio when the attacker wins is not greater than the cavalry ratio when the defender wins.

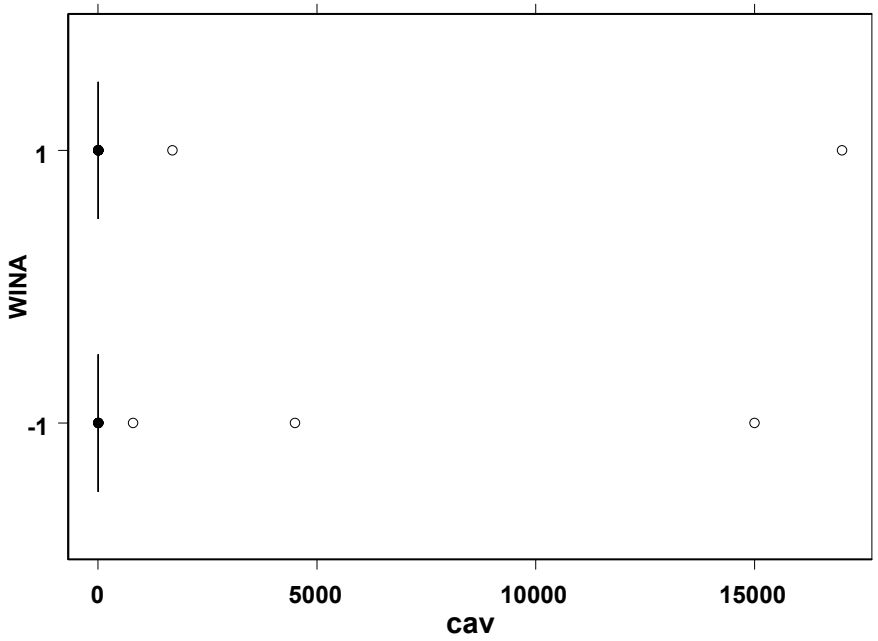

Figure 17. The Spread of the Cavalry Ratio, "cav"

In one battle, the attacker had a cavalry ratio of 15,000 to one and lost. In another, the attacker had a cavalry ratio of 5,000 to one and lost that as well. The defender won 55 percent of the battles when the attacker had a cavalry ratio of three to one or greater. These examples show that the cavalry ratio cannot be the sole predictor of a battle outcome.

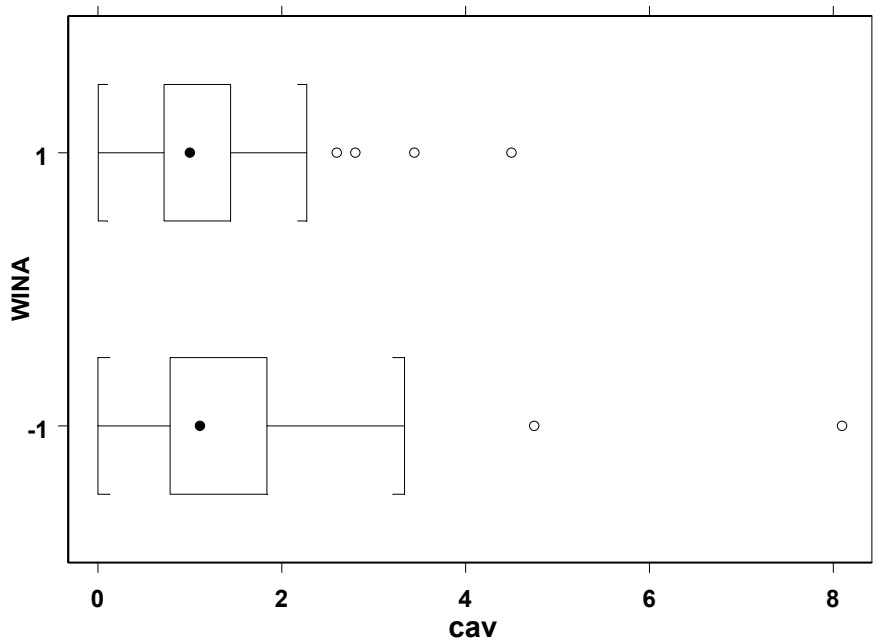

Figure 18. The Spread of the Cavalry Ratio Without Large Outliers

The interquartile range for the box plot in which the attacker wins is smaller than the box plot in which the defender wins. The defender had a higher chance of winning when the attacker had a higher cavalry ratio. This result is not intuitive. Cavalry ratio cannot be a good predictor of the battle outcome.

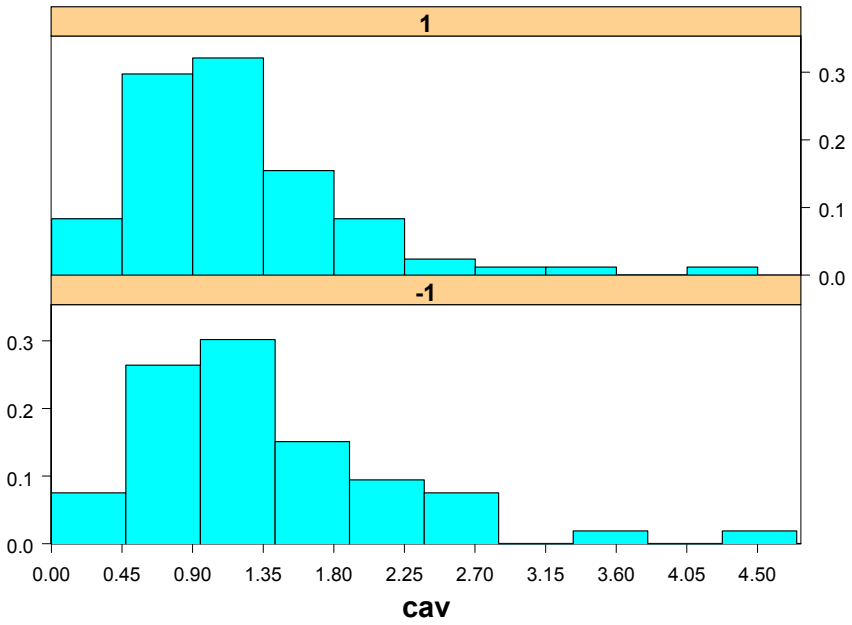

Figure 19. The Truncated Conditional Histogram of the Cavalry Ratio, "cav"

In the right tail, the defender had a higher chance of winning. In the left tail, the attacker had a higher chance. The conditional histograms reveal the same result as the box plots. We do not expect cavalry ratio to be included in most of the models.

f) Defender's Primary Defensive Posture: "POST1"

The distribution of the defender's primary defensive posture is given in Table 6.

POST1 / WINA	Delay (DL)	Fortified Defense (FD)	Hasty Defense (HD)	Prepared Defense (PD)	Withdrawal (WD)
-1	4	72	129	55	0
1	16	107	164	107	2
TOTALS	20	179	293	162	2

Table 6. The Defender's Primary Posture

The defender used "hasty defense" 45 percent, "fortified defense" 27 percent, and "prepared defense" 25 percent of the time. Moreover, the defender won 49 percent of the time when it made a "hasty defense," and this is the defender's highest success rate between other defensive postures. These results suggest that the defender should be prepared to fight without enough time for fortifications and preparations.

g) Attacker's Primary Tactical Scheme: "PRIA1"

How victory may be produced for them out of the enemy's own tactics—that is what the multitude cannot comprehend.

− Sun Tzu, *The Art of War*

The attacker's primary tactical scheme is given in Table 7.

PRIA1 / WINA	Double Envelopment (DE)	Defensive Offensive Plan (DO)	Single Envelopment (EE)	Frontal Attack (FF)	River Crossing (RC)
−1	5	0	11	238	5
1	14	1	30	324	22
Size	19	1	41	562	27

Table 7. The Attacker's Primary Tactics

Attackers used frontal attack 87 percent of the time. In World War II, "frontal attack" was used 81 percent of the time. "Frontal attack" is the least desirable offensive tactic because of the high casualty risk. However, it has been the most commonly used tactic throughout history. The other maneuvers are harder to plan and execute.

The attacker's chance of winning a battle given the defender's posture and the attacker's tactics is given in Table 8.

Posture / Tactics	Delay	Fortified Defense	Hasty Defense	Prepared Defense	Withdrawal
Double Envelopment	0/0	1/4 0.25	8/10 0.80	5/5 1.00	0/0
Offensive Defensive Plan	0/0	0/0	1/1 1.00	0/0	0/0
Single Envelopment	1/1 1.00	5/7 0.71	17/24 0.71	6/8 0.75	1/1 1.00
Frontal Attack	10/13 0.77	96/160 0.60	136/256 0.53	80/131 0.61	1/1 1.00
River Crossing	1/2 0.50	5/7 0.71	2/2 1.00	13/15 0.87	0/0

Table 8. The Attacker's Chances of Victory Given the Attacker's Tactics and the Defender's Posture

The numbers in the cells show the number of the attacker's victory, the total number of battles and the probability of the attacker's victory. Most of the time (39 percent), the attacker used a frontal attack against the defender's hasty defense, and the attacker was successful 53 percent of the time. A frontal attack against a hasty defense was less successful than a frontal attack on a prepared defense and a fortified defense. This may happen when the attacker was pursuing the defender and was not well prepared for a frontal attack. Against a hasty defense, single envelopment and double envelopment tactics reveal much higher chances of victory. According to the defender's point of view, a flexible defense in a sector with a number of hasty defenses may be a better option than defending in a fortified defense against a well-prepared attack.

3. Relative Variables

a) Relative Surprise: "SURPA"

Strike the enemy at a time or place or in a manner for which he is unprepared.

− *FM 100-5 Operations*

Surprise is considered one of the principles of war [Ref.13]. As stated in the introduction to this chapter, the "SURPA" variable has been adjusted to show only which side achieved surprise, not the level of it. The distribution of the adjusted "SURPA" variable is given in Table 9.

The defender achieved surprise only two percent of the time. The attacker achieved surprise 25 percent of the time. In other words, the attacker could not achieve surprise 75 percent of the time. This result shows that it is hard to achieve surprise even when an attacker plans to do so. The effect of surprise on the battle outcome is given in Figure 20.

SURPA / WINA	A	D	O
–1	32	14	206
1	124	0	249
TOTALS	156	14	455

Table 9. The Distribution of the Surprise Variable, "SURPA"

Attackers achieved surprise 25 percent of the time and the defenders achieved surprise two percent. The attacker's low chance of achieving surprise shows that achieving surprise is not an easy task to accomplish. The defender's percentage shows that even the defenders can achieve surprise by selecting different tactics, posture, time, and so on.

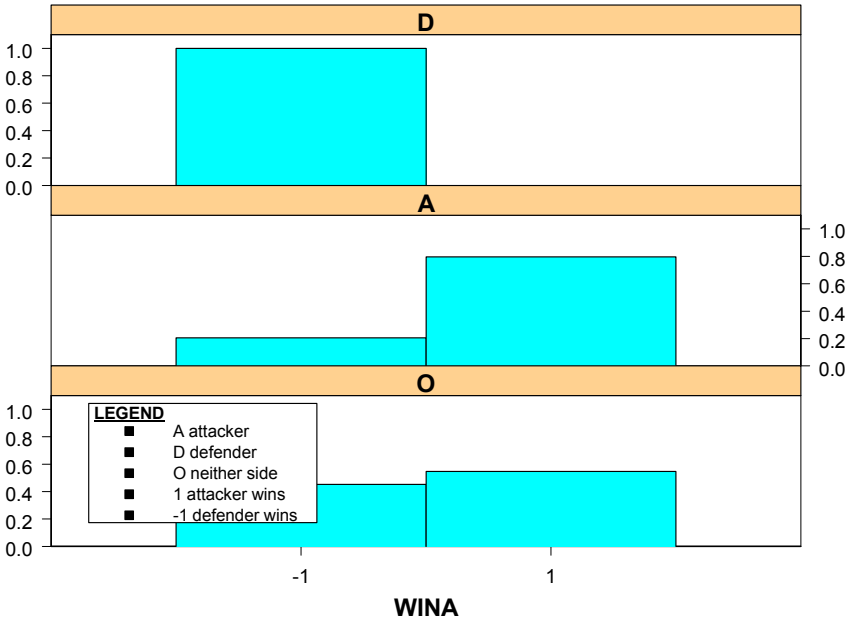

Figure 20. The Effect of Surprise on the Battle Outcome

When they achieved surprise, attackers won 80 percent of the time and defenders won 100 percent of the time. In general, surprise is related to offensive maneuvers. The defender's relative surprise means a failure on the attacker's side**.** This might be the reason for the defender's high chance of winning the battle when it achieved surprise. Compared to the initial distribution of the battle outcomes, achieving surprise increased the attacker's and the defender's chance of winning the battle by 20 percent and 60 percent respectively. We would expect this variable to be present in our classification models.

b) Relative Air Superiority in the Theater: "AEROA"

Control of air gives commanders the freedom to conduct successful attacks that can neutralize or destroy an enemy's war fighting potential.

− *FM 100-5 Operations*

Air power has been widely used since World War II. The distribution of AERO is given in Table 10.

AEROA / WINA	A	D	O
–1	67	38	147
1	162	19	192
TOTALS	229	57	339

Table 10. The Distribution of the Relative Air Superiority Variable, "AEROA"

Of the battles in which one side had air superiority, attackers had air superiority 80 percent of the time and defenders had air superiority 20 percent of the time.

The effect of relative air superiority on the battle outcome can be seen in Figure 21.

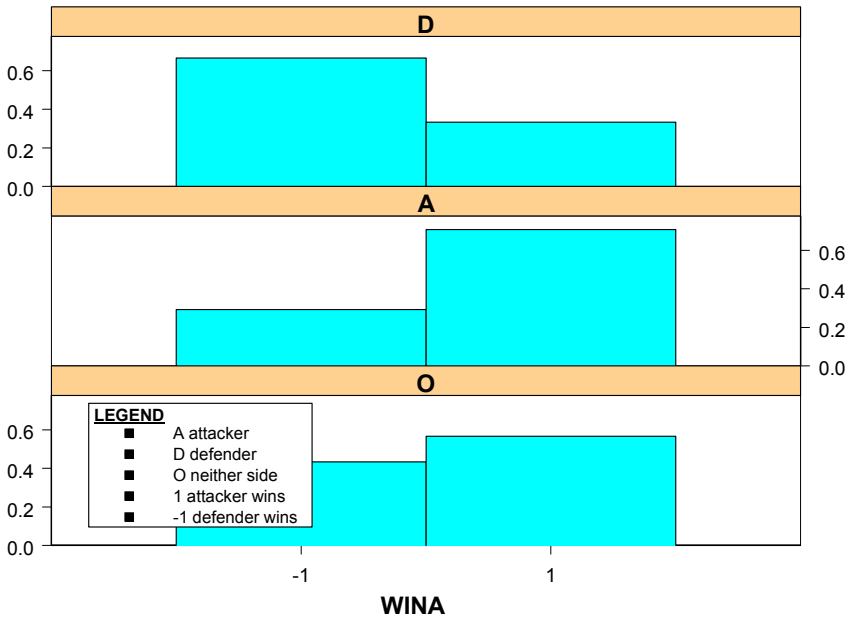

Figure 21. The Effect of Relative Air Superiority on Battle Outcome

When they had air superiority, the attackers won 71 percent of the time and the defenders won 67 percent of the time. The air superiority's effect on the battle outcome is more significant for the defender's side. Compared to the initial distribution of the battle outcomes, the defender's chance of winning the battle increased by 27 percent with the air superiority.

c) Relative Combat Effectiveness: "CEA"

The distribution of relative combat effectiveness is given in Table 11. The effect of relative combat effectiveness on the battle outcome is given in Figure 22.

CEA / WINA	A	D	O
–1	22	55	175
1	120	27	226
TOTALS	142	82	401

Table 11. The Distribution of Relative Combat Effectiveness Variable, "CEA"

The attacker had a relative combat effectiveness advantage 23 percent of the time; the defender had the advantage 13 percent of the time.

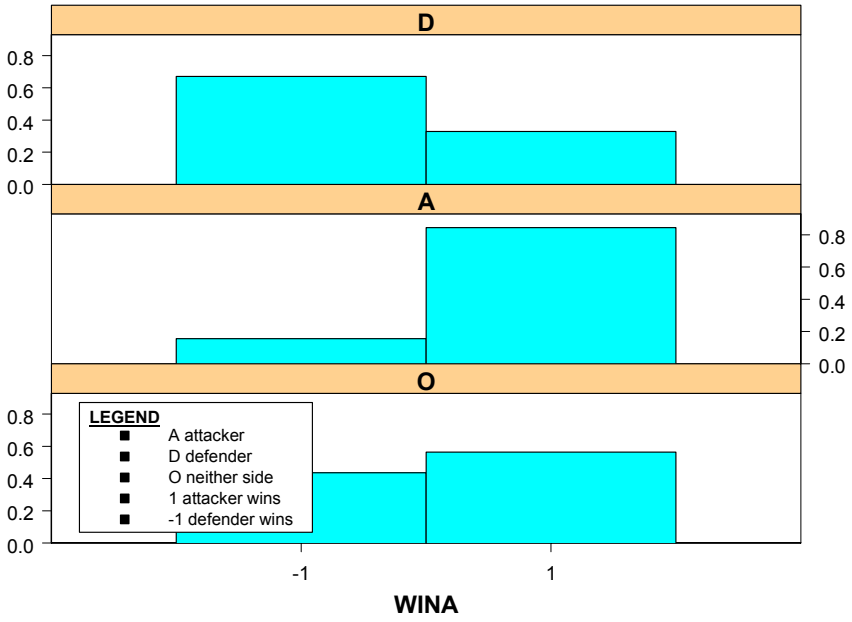

Figure 22. The Effect of Relative Combat Effectiveness on Battle Outcome

When they had a relative combat effectiveness advantage, the attackers won 89 percent of the time and the defenders won 59 percent of the time. Having a relative combat effectiveness advantage increased the attacker's chance of winning by 29 percent compared to the initial distribution of battle outcomes.

d) Relative Leadership Advantage: "LEADA"

> The most essential dynamic of combat power is competent and confident officer and noncommissioned officer leadership
>
> *-FM 100-5 Operations*

Relative leadership advantage is an important factor in a battle. The distribution of relative leadership advantage is given in Table 12. The effect of relative leadership advantage on the battle outcome is given in Figure 23.

LEADA / WINA	A	D	O
–1	13	88	151
1	155	12	206
TOTALS	168	100	357

Table 12. The Distribution of the Relative Leadership Advantage Variable, "LEADA"

Attackers had the leadership advantage 27 percent of the time; 16 percent of the time the defenders did. In the thirteen battles in which the attacker lost despite having the leadership advantage, the defender usually outnumbered the attacker. There are two exceptions, one in the 1900 Boer War, and other in the 1967 Arab-Israeli War, but the outcomes of these battles are not clear-cut. In the twelve battles in which the defender lost despite having a leadership advantage, the attacker usually outnumbered the defender. There is only one exception, a 1918 World War I battle, but the outcome of that battle is also not clear-cut.

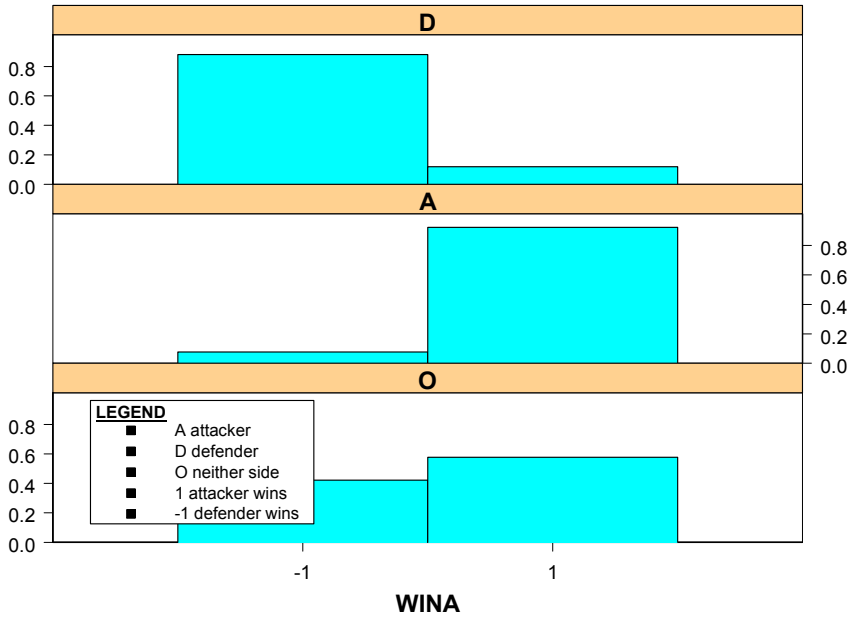

Figure 23. The Effect of the Leadership Advantage on Battle Outcome

When they had a relative leadership advantage, the attackers won 92 percent of the time and the defenders won 88 percent of the time. The attacker's chance of winning the battle increased by 32 percent compared to the initial battle outcome distribution. The defender's chance of winning the battle increased by 48 percent. Both of these results show that relative leadership advantage plays an important role in determining the outcome of a battle. This variable is expected to appear in the classification models.

e) Relative Training Advantage: "TRNGA"

The distribution of relative training advantage is given in Table 13. The effect of the relative training advantage on battle outcome is given in Figure 24.

TRNGA / WINA	A	D	O
–1	24	50	178
1	87	52	234
TOTALS	111	102	412

Table 13. The Distribution of the Relative Training Advantage Variable, "TRNGA"

The attackers and the defenders had a relative training advantage 18 percent of the time and 16 percent of the time, respectively. Thus, 66 percent of the time, neither side had the relative training advantage.

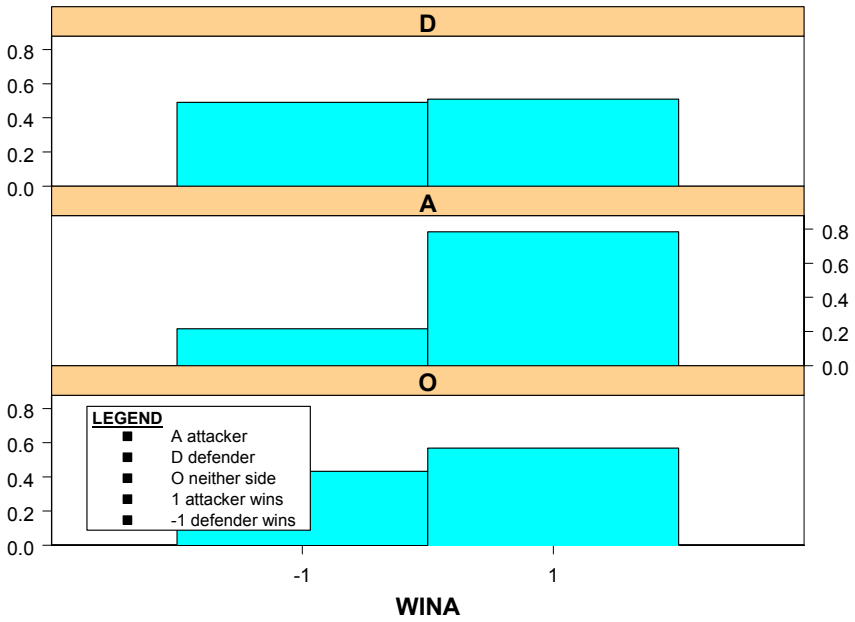

Figure 24. The Effect of Relative Training Advantage on Battle Outcome

When they had the relative training advantage, attackers won 78 percent of the time and defenders won 49 percent of the time. Compared to the initial distribution of battle outcomes, the attacker's chance of winning increased 18 percent when it had a relative training advantage. However, relative training advantage did not increase the defender's chance of winning significantly. This may result from the historian's inclination to give the relative advantage to the winner's side and to the attacker's side.

f) Relative Morale Advantage: "MORALA"

Morale is the greatest single factor in successful wars

–Dwight D. Eisenhower

The moral is to the physical as three is to one

–Napoleon

The distribution of relative morale advantage is given in Table 14. The effect of relative morale advantage on the battle outcome is given in Figure 25.

MORALA / WINA	A	D	O
–1	22	8	222
1	110	2	261
TOTALS	132	10	483

Table 14. The Distribution of the Relative Morale Advantage Variable, "MORALA"

The attacker had a relative morale advantage 21 percent of the time, and the defender only two percent of the time. The defender's low percentage of relative morale advantage may result from the specialty of defensive maneuvers or from the historians' inclination toward assigning the relative advantage to the attacker's side and the winner's side. Another explanation may be that the attacker is inclined not to attack when morale is low.

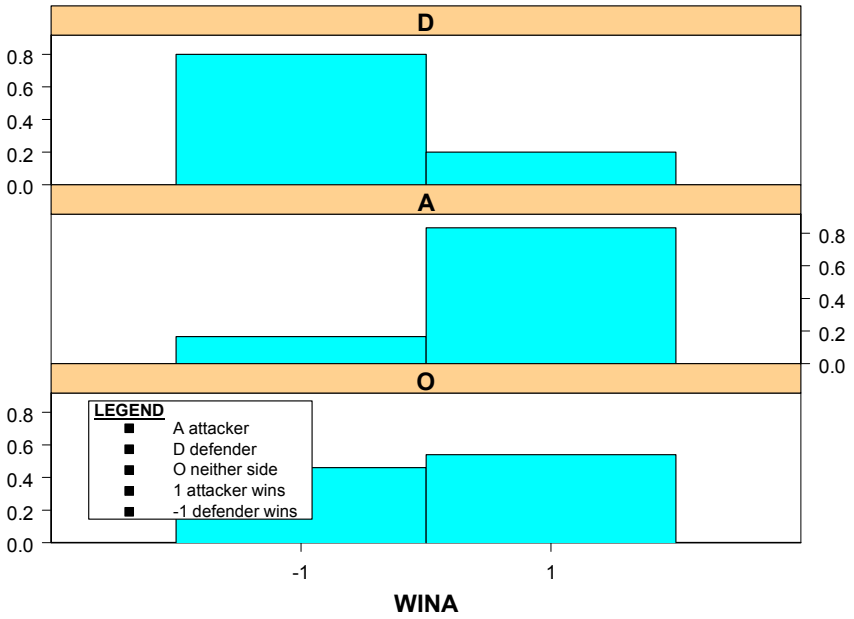

Figure 25. The Effect of Relative Morale Advantage on Battle Outcome

When they had the relative morale advantage, attackers won 83 percent of the time and defenders won 80 percent of the time. Compared to the initial distribution of the battle outcomes, the morale advantage increases the attacker's and defender's chance of winning by 23 percent and 40 percent respectively. We expect our classification models to include morale advantage.

g) Relative Logistics Advantage: "LOGSA"

Logistics cannot win a war, but its absence or inadequacy can cause defeat.

− *FM 100-5 Operations*

The distribution of relative logistics advantage is given in Table 15. The effect of a relative logistics advantage over the battle outcome is given in Figure 26.

LOGSA / WINA	A	D	O
–1	6	16	230
1	47	7	319
TOTALS	53	23	549

Table 15. The Distribution of the Relative Logistics Advantage Variable, "LOGSA"

Attackers had the logistics advantage in eight percent of the battles, and defenders had this advantage in four percent of the battles. Generally (88 percent) neither the attacker nor the defender had a relative logistics advantage. This is not an intuitive distribution. We would expect one side to have better logistics more often.

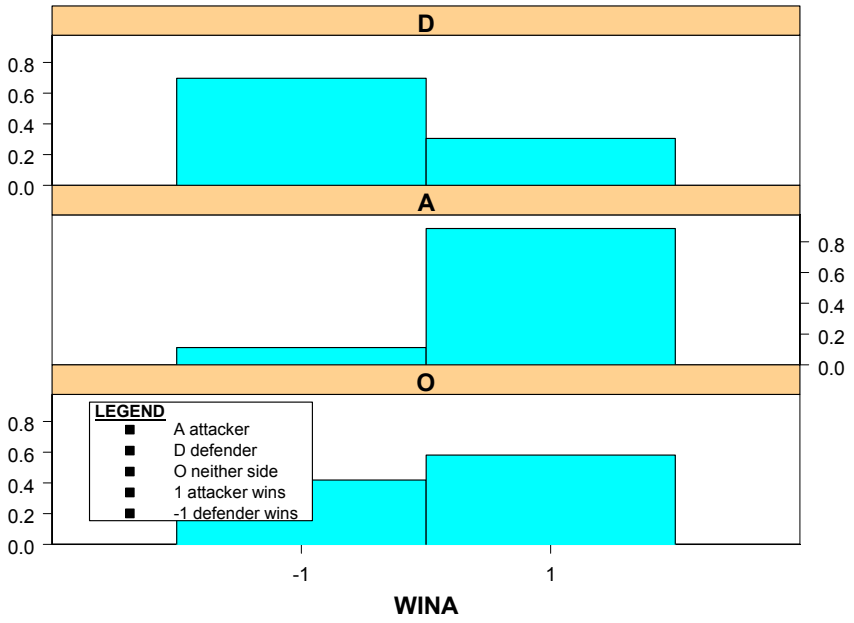

Figure 26. The Effect of a Relative Logistics Advantage on Battle Outcome

When they had a relative logistics advantage, attackers won 89 percent of the time and defenders won 70 percent of the time. This is a high increase in the attacker's and defender's chance of winning compared to the initial distribution of the battle outcomes. However, 88 percent of the time neither side had an advantage. Thus, we would not expect this variable to be included in most of the models.

h) Relative Momentum Advantage: "MOMNTA"

The energy developed by good fighting men is as the momentum of a round stone rolled down a mountain thousands of feet in height."

− Sun-Tzu, *The Art of War*

The distribution of relative momentum advantage is given in Table 16. The effect of relative momentum advantage over the battle outcome is given in Figure 27.

MOMNTA / WINA	A	D	O
–1	31	3	218
1	116	1	256
TOTALS	147	4	474

Table 16. The Distribution of the Relative Momentum Advantage, "MOMNTA"

Attackers had the relative momentum advantage 24 percent of the time and defenders had this advantage 1 percent of the time. The momentum advantage is related to force strength and movement. The defender's momentum advantage may result from offensive maneuvers in defensive tactics.

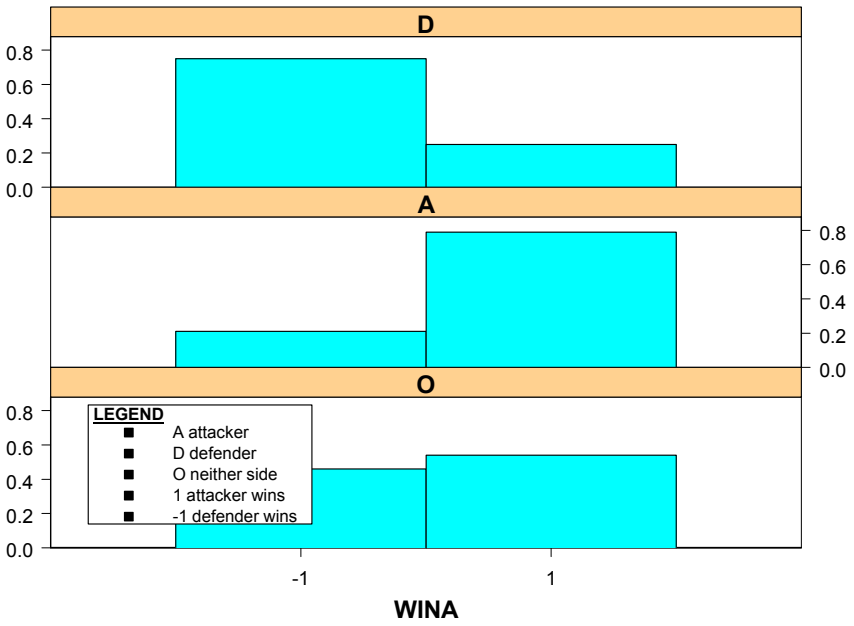

Figure 27. The Effect of Relative Momentum Advantage on Battle Outcome

When they had the relative momentum advantage, attackers won 79 percent of the time and defenders won 75 percent of the time. For the defenders' side this is out of four observations, so this result is not significant. The relative momentum advantage increased the attacker's chance of winning by 19 percent compared to the initial distribution of the battle outcomes.

i) Relative Intelligence Advantage: "INTELA"

Intelligence is fundamental to effective planning, security and deception.

– *FM 100-5 Operations*

The distribution of relative intelligence advantage is given in Table 17. The effect of relative intelligence advantage on the battle outcome is given in Figure 28.

INTELA / WINA	A	D	O
–1	7	39	206
1	73	8	292
TOTALS	80	47	498

Table 17. The Distribution of the Relative Intelligence Advantage Variable, "INTELA"

Attackers and defenders had a relative intelligence advantage 13 percent and 8 percent of the time, respectively. Neither side had the relative intelligence advantage 79 percent of the time. This result shows that it has been difficult to gain an intelligence advantage in a battle.

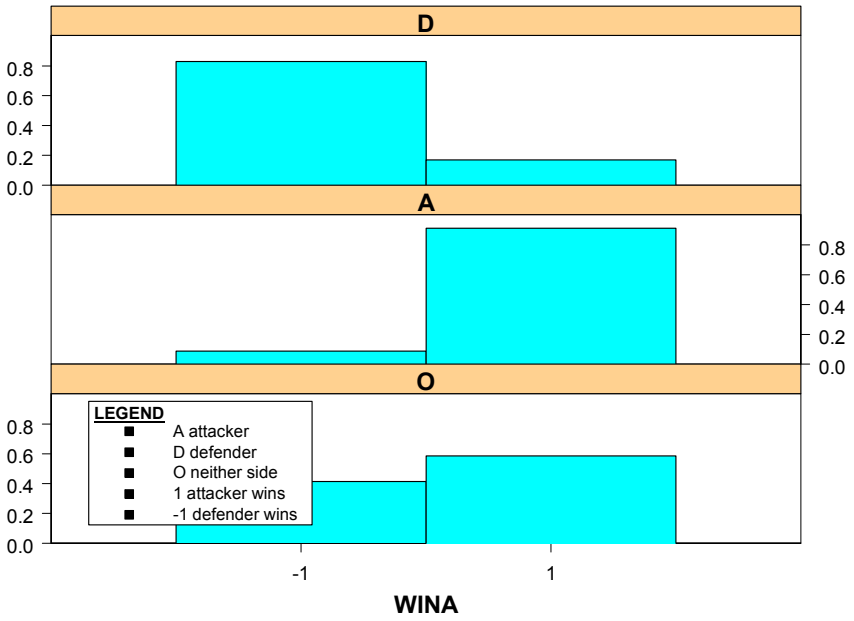

Figure 28. The Effect of Relative Intelligence Advantage on Battle Outcome

When they had a relative intelligence advantage, the attackers won 91 percent of the time and the defenders won 83 percent of the time. Compared to the initial distribution of the battle outcomes, the intelligence advantage increased the attacker's and the defender's chance of winning by 31 percent and 43 percent respectively. Thus, we expect that this variable will be present in the classification models.

j) Relative Technology Advantage: "TECHA"

The distribution of relative technology advantage is given in Table 18. The effect of relative technology advantage over the battle outcome is given in Figure 29.

TECHA / WINA	A	D	O
–1	8	5	239
1	18	1	354
TOTALS	26	6	593

Table 18. The Distribution of Relative Technology Advantage Variable, "TECHA"

Attackers had the relative technological advantage four percent of the time and defenders had this advantage one percent of the time. In general, neither the attacker nor the defender had a significant technological advantage. This result is not intuitive because we expect the attackers to have a technology advantage more often.

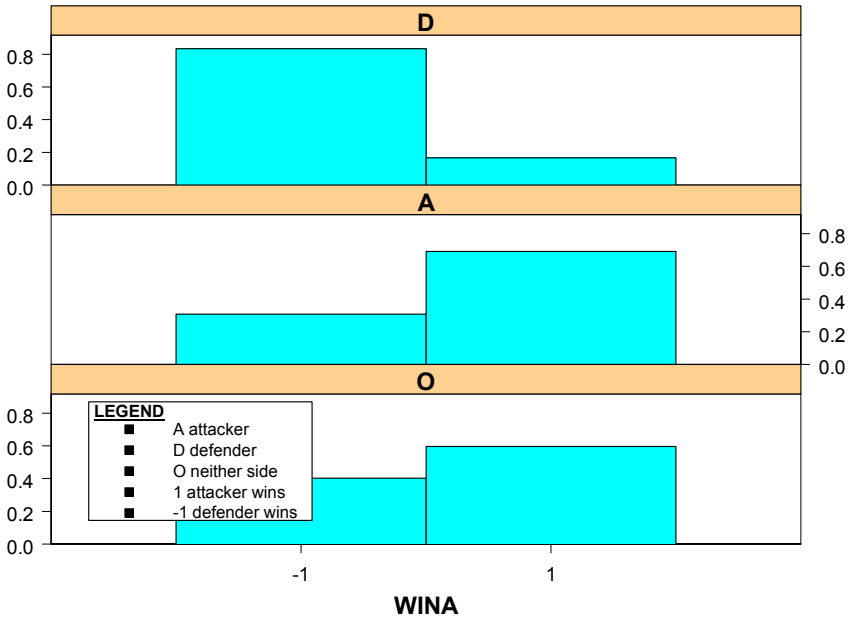

Figure 29. The Effect of Relative Technology Advantage on Battle Outcome

When they had a relative technological advantage, the attackers won 69 percent of the time and the defenders won 83 percent of the time. Due to the low percentages of the attacker's and the defender's relative advantage, we do not expect this variable to be present in most of the classification models.

k) Relative Initiative Advantage: "INITA"

Initiative sets or changes the terms of battle by action and implies an offensive spirit in the conduct of all operations

-FM 100-5 Operations

FM 100-5 Operations, the Army's keystone doctrine, includes initiative as one of the tenets of army operations [Ref.13: p. 2-6]. The distribution of relative initiative advantage is given in Table 19. The effect of relative initiative advantage on the battle outcome is given in Figure 30.

INITA / WINA	A	D	O
–1	105	24	123
1	311	0	62
TOTALS	416	24	185

Table 19. The Distribution of the Relative Initiative Advantage Variable, "INITA"

Attackers had a relative initiative advantage 67 percent of the time and defenders had this advantage four percent of the time. In 24 battles in which the defender had an initiative advantage and won the battle, other variables were also favoring the defender's side. The defender had the leadership advantage in 19 of them; in the remaining five, neither side had the leadership advantage. In 14 of them, the defender had the intelligence advantage, while the attacker had the intelligence advantage only in one battle. In all battles, generally, the attacking side had the initiative advantage, which is inherent in the nature of offense.

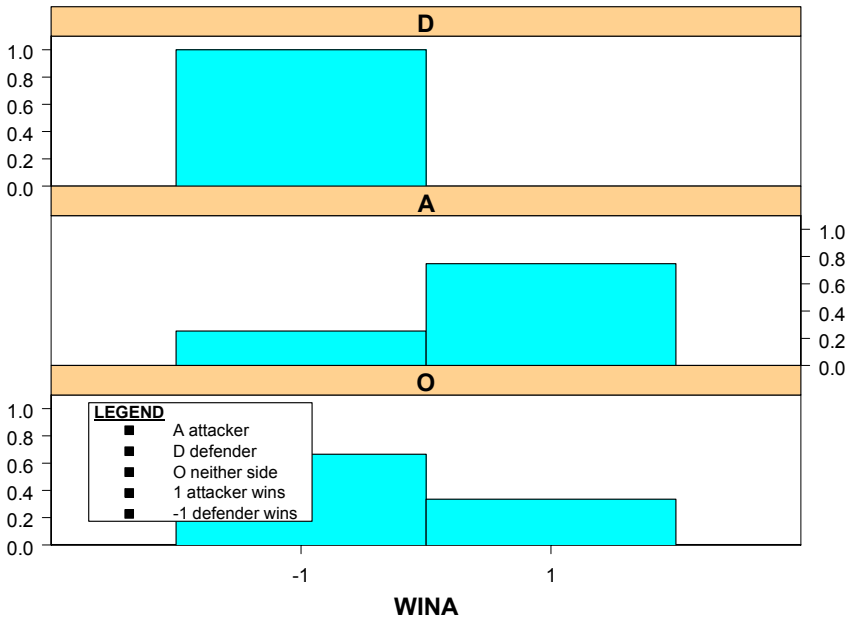

Figure 30. The Effect of Relative Initiative Advantage on the Battle Outcome

When they had the relative initiative advantage, the attackers won 75 percent of the time and the defenders won 100 percent of the time. Initiative advantage is one of the properties of offense. The defender's high success rate when they had the initiative advantage may reflect the battles in which the attacker's offense was not effective.

4. Terrain and Weather Descriptors

a) Primary Local Terrain Description: "TERRA1"

The primary local terrain description variable, "TERRA1," consists of three characters, each describing a feature of the local terrain. However, TERRA1 has 17 levels. In order to decrease the number of levels, "TERRA1" was split into three variables with one element each and with names terra1.1, terra1.2, and terra1.3.

(1) First Character: "terra1.1" (Landscape Evenness)

terra1.1 / WINA	Flat (F)	Rugged (G)	Rolling (R)	Other (O)
−1	44	48	163	3
1	69	83	235	3
TOTALS	113	131	398	6

Table 20. The First Terrain Descriptor

The terrain was rolling 61 percent of the time, rugged 20 percent, and flat 17 percent of the time. Compared to the initial distribution of battle outcomes, "terra1.1" variable did not significantly change the attacker's or defender's chance of victory. We do not expect this variable to be present in our classification models.

(2) Second Character: "terra1.2" (Vegetation)

terra1.2 / WINA	Bare (B)	Desert (D)	Mixed (M)	Wooded (W)	Other(0)
−1	**36**	**11**	**198**	**9**	**4**
1	**40**	**22**	**299**	**23**	**6**
TOTALS	76	33	497	32	10

Table 21. The Second Terrain Descriptor

The terrain was mixed 75 percent of the time and bare 12 percent of the time. Compared to the initial distribution of battle outcomes, the attacker had 0.12 higher chance of winning in wooded terrain, 0.06 higher chance in desert. The defender had a 0.07 higher chance of winning in a bare terrain. We would expect the defender to be more victorious in a wooded terrain and the attacker to be more victorious in a bare terrain. As a conclusion, this variable did not change the attackers' and defenders' chance of winning significantly.

(3) Third Character: "terra1.3" (Composition)

terra1.3 / WINA	Other (0)	Dunes (D)	Marsh (M)	Urban (U)
−1	254	1	0	3
1	385	1	1	3
TOTALS	639	2	1	6

Table 22. The Third Terrain Descriptor

The third character was not available 99 percent of the time. The battles fought in an urban environment are not significantly different from others, out of six battles the attacker and the defender won equally. This variable is not expected to appear in the classification model.

b) Primary Local Weather Descriptor: "WX1"

The local weather descriptor, "WX1," consists of five characters, each representing a characteristic of the weather. In this state, "WX1" has 49 levels. In order to decrease the number of levels,"WX1" will be split into five variables, each represented by one character.

(1) First Character: "wx1.1" (Precipitation 1)

wx1.1 / WINA	Dry (D)	Wet (W)	Other (0)
−1	194	65	1
1	313	85	0
TOTALS	507	150	1

Table 23. The First Weather Descriptor

The weather was dry 77 percent of the time and wet 23 percent of the time. Compared to the initial distribution of battle outcomes, the attacker had a 0.01 higher chance of winning in dry weather, and the defender had a 0.03 higher chance of winning in wet weather. This variable did not significantly increase the chances of winning, so we do not expect it to be present in our models.

(2) Second Character: "wx1.2" (Precipitation 2)

wx1.2 / WINA	Other (0)	Heavy Precipitation (H)	Light Precipitation (L)	Overcast (O)	Sunny (S)
−1	10	16	44	11	179
1	25	33	51	30	259
TOTALS	35	49	95	41	438

Table 24. The Second Weather Descriptor

The weather was sunny 66 percent of the time, light precipitation 14 percent, and heavy precipitation 7 percent of the time. Compared to the initial distribution of battle outcomes, the attacker had a 0.07 higher chance of winning under heavy precipitation, a 0.13 higher chance in overcast weather. The defender had a 0.06 higher chance under light precipitation.

(3) Third Character: "wx1.3" (Temperature)

wx1.3 / WINA	Cold (C)	Hot (H)	Temperate (T)	Other (0)
−1	28	27	204	1
1	41	62	295	0
Size	69	89	499	1

Table 25. The Third Weather Descriptor

The weather was temperate 76 percent of the time, hot 14 percent of the time, and cold 10 percent of the time. Compared to the initial distribution of battle outcomes, the attacker had a 0.09 higher chance of winning under hot weather.

(4) Fourth Character: "wx1.4" (Season)

wx1.4 / WINA	Spring ($)	Fall (F)	Summer (S)	Winter (W)
−1	60	78	87	35
1	67	131	149	51
Size	127	209	236	86

Table 26. The Fourth Weather Descriptor

Battles were fought 36 percent of the time in summer, 32 percent in fall, 19 percent in spring and 13 percent of the time in winter. In general, the attacker chose to fight in summer and fall. Compared to the initial distribution of battle outcomes, the attacker had a 0.07 higher chance of winning in spring. We do not expect this variable to be present in our models.

(5) Fifth Character: "wx1.5" (Tropical Condition)

wx1.5 / WINA	Desert (D)	Tropical (E)	Temperate (T)
−1	4	1	255
1	14	4	380
Size	18	5	635

Table 27. The Fifth Weather Descriptor

Battles occurred one percent in tropical conditions, three percent in desert conditions, and others were in temperate conditions. Compared to the initial distribution of battle outcomes, the attacker had a 0.17 higher chance in desert conditions; however, there are only 18 battles fought in this condition. Therefore, this result is not significant. We do not expect this variable to be present in our models.

C. DISCUSSION

This finishes the variables we will look at specifically. Due to the nature of the combat and inaccessibility of historical records, many factors have missing values. The missing values have a challenging affect on the modeling process. The numbers of missing values of the data set and subsets are given in Table 28.

Many classification models just ignore every battle that has any missing value. However, in the CDB90 data set no battle has all of the variables we have described in this chapter. Modeling with classification trees, using the `rpart` library of S-Plus, helps us to build models with the missing values. This process uses surrogate splits when the model encounters a missing value. The model has some disadvantages too. The graph of the `rpart` only gives the regular splits, not the surrogate ones. In order to understand an actual model, the long summary output should be used.

D. MISSING VALUES

Column Names	Entire Data	Subset1		Subset 2		Subset 3		Subset 4		Subset 5	
		train 1	test 1	train 2	test 2	train 3	test 3	train 4	test 4	train 5	test 5
WINA	**0**	0	0	0	0	0	0	0	0	0	0
fR	**1**	0	0	0	0	0	0	1	0	0	1
arty	**153**	39	17	79	19	48	34	5	2	146	7
tank	**362**	109	55	178	57	131	64	6	3	353	9
cav	**514**	34	21	116	82	36	46	150	73	291	223
fly	**484**	109	55	178	81	131	68	76	17	391	93
CEA	**33**	0	0	0	0	0	0	13	20	0	33
LEADA	**33**	0	0	0	0	0	0	13	20	0	33
MORALA	**33**	0	0	0	0	0	0	13	20	0	33
INTELA	**33**	0	0	0	0	0	0	13	20	0	33
TECHA	**33**	0	0	0	0	0	0	13	20	0	33
TRNGA	**33**	0	0	0	0	0	0	13	20	0	33
LOGSA	**33**	0	0	0	0	0	0	13	20	0	33
SURPA	**33**	0	0	0	0	0	0	13	20	0	33
AEROA	**33**	0	0	0	0	0	0	13	20	0	33
PRIA1	**8**	0	0	0	0	0	0	1	7	0	8
POST1	**2**	0	0	0	0	0	0	2	0	0	2
terra1.1	**10**	0	0	0	0	0	0	0	10	0	10
terra1.2	**10**	0	0	0	0	0	0	0	10	0	10
terra1.3	**10**	0	0	0	0	0	0	0	10	0	10
wx1.1	**0**	0	0	0	0	0	0	0	0	0	0
wx1.2	**0**	0	0	0	0	0	0	0	0	0	0
wx1.4	**0**	0	0	0	0	0	0	0	0	0	0
wx1.3	**0**	0	0	0	0	0	0	0	0	0	0
wx1.5	**0**	0	0	0	0	0	0	0	0	0	0

Table 28. Missing Values in the Data Set and the Subsets

Most of the calculated ratios have missing values that are assigned when the attacker or the defender did not have the specific kind of weapon or system. The cavalry ratio has 78 percent missing values, the air sorties ratio has 74 percent, the tank ratio has 53 percent, and the artillery ratio has 23 percent missing values. In the data set, none of the rows has all the ratio variables (force ratio, tank ratio, cavalry ratio, and air sorties ratio) present. In other words, every row has at least one missing value. This result shows the importance of using a classification model, such as classification trees with the `rpart` library, which can handle missing values without discarding the rows containing them.

E. CORRELATION BETWEEN VARIABLES

Most of our variables are factors. In order to find correlation between them, the ordinal ones, variables that can be ordered, are converted to integer values. In this case, level "A" is converted to "1," "D" to "–1," and "O" to "0." The cavalry ratio had no observations in common with tank and air sorties ratio and is not included. The correlation between variables is given in Table 29.

	WINA	fR	tank	arty	fly	CEA	LEADA	MORALA	INTELA	TECHA	TRNGA	LOGSA	SURPA	AEROA	INITA	MOMNTA
WINA	1.00	0.13	0.11	0.10	0.12	0.32	**0.52**	0.26	0.33	0.07	0.17	0.21	0.27	0.21	**0.45**	0.22
fR	0.13	1.00	0.18	**0.53**	**0.41**	−0.13	−0.14	0.22	0.06	0.00	−0.13	0.08	0.04	0.17	−0.02	0.11
tank	0.11	0.18	1.00	0.21	**0.50**	0.04	−0.08	0.14	−0.03	0.33	0.02	−0.07	−0.05	0.14	−0.01	−0.15
arty	0.10	**0.53**	0.21	1.00	0.31	0.02	−0.10	0.17	0.04	0.09	0.07	0.05	0.01	0.12	0.00	0.05
fly	0.12	**0.41**	**0.50**	0.31	1.00	0.03	−0.04	0.06	−0.07	0.21	0.05	0.06	−0.13	0.28	−0.02	−0.09
CEA	0.32	−0.13	0.04	0.02	0.03	1.00	**0.54**	0.10	0.06	0.18	**0.68**	0.02	0.13	0.09	0.28	0.14
LEADA	**0.52**	−0.14	−0.08	−0.10	−0.04	**0.54**	1.00	0.06	0.23	0.06	0.39	0.08	0.21	0.11	0.40	0.11
MORALA	0.26	0.22	0.14	0.17	0.06	0.10	0.06	1.00	0.09	0.15	−0.08	0.18	0.06	0.28	0.20	0.26
INTELA	0.33	0.06	−0.03	0.04	−0.07	0.06	0.23	0.09	1.00	0.06	0.06	0.19	**0.41**	−0.04	0.29	0.10
TECHA	0.07	0.00	0.33	0.09	0.21	0.18	0.06	0.15	0.06	1.00	0.14	0.10	0.04	0.18	0.12	0.12
TRNGA	0.17	−0.13	0.02	0.07	0.05	**0.68**	0.39	−0.08	0.06	0.14	1.00	0.09	0.09	0.02	0.15	0.21
LOGSA	0.21	0.08	−0.07	0.05	0.06	0.02	0.08	0.18	0.19	0.10	0.09	1.00	0.06	0.18	0.10	0.24
SURPA	0.27	0.04	−0.05	0.01	−0.13	0.13	0.21	0.06	**0.41**	0.04	0.09	0.06	1.00	−0.04	0.28	−0.03
AEROA	0.21	0.17	0.14	0.12	0.28	0.09	0.11	0.28	−0.04	0.18	0.02	0.18	−0.04	1.00	0.14	0.15
INITA	**0.45**	−0.02	−0.01	0.00	−0.02	0.28	0.40	0.20	0.29	0.12	0.15	0.10	0.28	0.14	1.00	0.27
MOMNTA	0.22	0.11	**−0.15**	0.05	−0.09	0.14	0.11	0.26	0.10	0.12	0.21	0.24	−0.03	0.15	0.27	1.00

Table 29. Correlation between Relative Variables

a. The highest correlation is between relative combat effectiveness and training (0.68).
b. The lowest negative correlation is between tank ratio and momentum advantage (−0.15). This result is not intuitive. However, when the tank ratio is available, 70 percent of the MORALA is coded as "O", meaning neither side had the advantage or it is unknown.

c. Winning is highly correlated with relative leadership advantage and relative initiative advantage. This variable is not highly correlated with weapon ratios, tank ratio, artillery ratio, CAS sorties ratio, and force ratio. No negative correlation exists between the "WINA" and other variables.
d. Force ratio is negatively correlated with training, and combat effectiveness. When they had training and combat effectiveness advantage, good leaders fought with fewer troops. Force ratio is highly correlated with the artillery ratio and the CAS sorties ratio.
e. Tank ratio is highly correlated with the air sorties ratio, and technical advantage.
f. The artillery ratio is correlated with the force ratio, tank ratio and air sorties ratio. It is negatively correlated with leadership advantage.
g. The CAS sorties ratio is correlated with force ratio, tank ratio, artillery ratio, technical advantage and air superiority, but the correlation is not very high. This is not intuitive.
h. The combat effectiveness advantage is correlated with training advantage, leadership advantage, and initiative advantage. Good leaders had combat-effective troops who are well trained. Initiative advantage is easier to gain with combat effective troops.
i. The leadership advantage is correlated with combat effectiveness advantage and training advantage and initiative advantage.
j. The morale advantage is positively correlated with the force ratio and momentum advantage and initiative advantage.
k. The technical advantage is positively correlated with tank ratio, CAS sorties ratio and combat effectiveness.
l. The training advantage is highly correlated with combat effectiveness.
m. The logistics advantage is correlated with morale advantage, intelligence advantage, and momentum advantage.
n. Surprise is correlated with intelligence advantage and initiative advantage. It is negatively correlated with the CAS sorties ratio.
o. Air superiority is correlated with the CAS sorties ratio, force ratio, morale advantage, logistics advantage and technical advantage.
p. The initiative advantage is highly correlated with winning, leadership, intelligence, surprise, combat effectiveness and momentum advantage.
q. The momentum advantage is correlated with morale, training, and initiative. It is negatively correlated with tank ratio.

F. SUMMARY OF THE VARIABLES

According to the descriptive statistics, some variables had more effect on the battle outcome than the others did. These variables will most likely appear in the classification tree models. However, we will provide all these variables as inputs to the

model. Then, the classification tree will select the variables that most affect the outcome of battles most greatly. The variables that are likely to appear in the classification model are

Objective Variables:

1. Air sorties ratio, “fly”
2. Tank ratio, “tank”

Relative Variables:

3. Surprise, “SURPA”
4. Leadership, “LEADA”
5. Intelligence, “INTELA.”

G. NOTES ON THE DATA:

1. In battles with ISEQNO 304 and 319, the attacker’s casualties are larger than their total personnel strength. The “fx” variable is assigned as “NA” to these two observations.
2. In battles with ISEQNO 304, 319, and 334 the defender’s casualties are larger than their total force strength. The “fy” variable is assigned “NA” to these three observations.
3. In battle with ISEQNO 22, the defender had one casualty while the attacker had 3,000 casualties (44 percent of its force). The variables “fer,” ”fy,” “adv” is assigned “NA” for this observation.
4. In row 652, the second and third character of TERRA1 should be switched.
5. In the data, the second character of the terrain is about the cover status. “D” represents the Desert. In the weather data, the fourth letter represents the climate. “D” is for desert. However, in terrain there are 33 places with the character “D” as Desert conditions. In these 33 observations, only one of them, the weather, is represented as Desert. In all other 32 observations, while the terrain is Desert, the

weather appears to be temperate. This is inconsistent. On the other hand, the weather appears as Desert 20 times, but one of them matches with the Desert terrain.

THIS PAGE INTENTIONALLY LEFT BLANK

III. CLASSIFICATION MODELS

A. INTRODUCTION

In respect of military method, we have, firstly, Measurement; secondly, Estimation of quantity; thirdly, Calculation; fourthly, Balancing of chances; fifthly, Victory.

–Sun-Tzu, *The Art of War*

In this chapter, classification tree models will be analyzed. The first set of models will be formed by using only the objective variables (MODEL 1). The second set of models will include objective and relative variables (MODEL 2). The third set of models will be built by using objective, relative, terrain, and weather variables (MODEL 3). The pre-selected variables, which will be used in building classification trees, are listed below:

Objective Variables:

1. Force ratio, "fR"
2. CAS sorties ratio, "fly"
3. Tank ratio, "tank"
4. Artillery ratio, "arty"
5. Cavalry ratio, "cav"
6. Attacker's primary tactical scheme, "PRIA1"
7. Defender's primary defensive posture, "POST1"

Relative Variables:

8. Relative surprise, "SURPA"
9. Relative air superiority in the theater, "AEROA"
10. Relative combat effectiveness, "CEA"
11. Relative leadership advantage, "LEADA"
12. Relative training advantage, "TRNGA"

13. Relative morale advantage, “MORALA”

14. Relative logistics advantage, “LOGSA”

15. Relative momentum advantage, “MOMNTA”

16. Relative intelligence advantage, “INTELA”

17. Relative technology advantage, “TECHA”

18. Relative initiative advantage, “INITA”

Terrain and Weather Variables

19. First terrain descriptor, “terra1.1”

20. Second terrain descriptor, “terra1.2”

21. Third terrain descriptor, “terra1.3”

22. First weather descriptor, “wx1.1”

23. Second weather descriptor, “wx1.2”

24. Third weather descriptor, “wx1.3”

25. Fourth weather descriptor, “wx1.4”

26. Fifth weather descriptor, “wx1.5.”

In the data set, there is a variable, “CRIT,” that shows how well the outcome of a battle is assigned. If the level of the variable is “1,” then the assignment is clear-cut. If it is “2,” then the assignment is not clear-cut. If it is “0,” then this variable is not available. There are 465 battles with clear-cut outcomes, 136 with outcomes that are not clear-cut, and 57 with an unknown criterion. The sizes of the test sets with clear-cut outcomes are given in Table 30. In some models and tables, test sets with only clear-cut outcomes will be referred as test.b.

The validity of the models will be checked by predicting both the test sets, and the test sets with only clear-cut outcomes.

	Test1	Test2	Test3	Test4	Test5
All outcomes	55	82	71	73	223
Clear-cut outcomes	45	60	32	50	126

Table 30. The Sizes of Test Sets, and Test Sets with Only Clear-Cut Outcomes

B. BASE MODEL

The distribution of the battle outcome variable, "WINA," reveals that of the 658 battles in the data set, the attacker was victorious 60 percent of the time. Thus, a model assigning victory to the attacker can be considered as a base model. The base model will correctly predict the outcome of the battles in the data set 60 percent of the time. If a classification model cannot predict a battle outcome better than the base model, then that model is not informative. Thus, we will compare the predictions of the classification models with the base model.

Since the data set is split into five subsets, the correctness of the base model for each subset may differ from the others. Table 31 shows the misclassification rate of the subsets, training sets, test sets, and the test sets with only clear-cut outcomes when predicted by using the base model.

Data		Misclassification Rate
Entire Data Set		0.40
Subset 1	Train 1	0.35
	Test 1	0.45
	Test 1.B	0.38
Subset 2	Train 2	0.52
	Test 2	0.32
	Test 2.B	0.27
Subset 3	Train 3	0.37
	Test 3	0.38
	Test 3.B	0.16
Subset 4	Train 4	0.35
	Test 4	0.30
	Test 4.B	0.32
Subset 5	Train 5	0.43
	Test 5	0.33
	Test 5.B	0.25

Table 31. The Misclassification Rates of the Data Set and Subsets When Predicted by the Base Model

The base model predictions can be as high as 84 percent correct. However, as in predictions of training set two, it can be as low as 48 percent. Thus, the base model is not reliable. Still, these misclassification rates may help us understand how good our models' predictions are. Note: "test n.b" refers to the test set "n" with only clear-cut outcomes.

C. MODEL 1: FORCE RATIO, WEAPON RATIOS, POSTURE, AND TACTICS

In Model 1, we will build classification trees on training sets by using only the objective variables. Battle simulations and war games mainly use objective variables in determining the outcome of engagements. In this model, if we can achieve a high correct classification rate, we can conclude that objective variables are strongly related to the

outcome of battles. Otherwise, the objective variables alone will not be sufficient in predicting battle outcomes.

In a model, six classification trees will be built, one for the entire data set and five for the subsets. In order to understand the validity of the models, the classification trees built with training sets will be used in predicting both test sets and test sets with only clear-cut outcomes. These predictions will reveal the quality of the model and its predictive power.

1. Model 1.1: Entire Data Set (Yrs 1600-1982)

The model built by using the entire data set (Figure 31) reveals that the most important objective factor is the CAS sorties ratio. This model also correctly predicts the outcome of the Gulf War and the Kosovo War.

In the data set, when neither side had CAS sorties, a missing value indicator is assigned to the CAS sorties ratio, "fly." There are 484 missing values for this variable. In the classification model, surrogate splits were used to classify the battles where the CAS sorties ratio was missing. (Here the surrogate split says: go to left branch if fR<0.4)

2. Model 1.2: Subset 1 (Yrs. 1600-1847)

The model for Subset 1 (Figure 32) shows that between years 1600 and 1799, the most important variable affecting the outcome of battles was cavalry. Here the training set included years 1600-1799 and the test set, years 1799-1847.

3. Model 1.3: Subset 2 (Yrs. 1805-1918)

The first split for the model of Subset 2 (Figure 33) is the force ratio. However, there are only nine battles for which the "fR" is greater than 4.9. The second split is cavalry, which is again revealed as the most important factor. The training set included years 1805-1915 and the test set, years 1916-1918.

4. Model 1.4: Subset 3 (Yrs. 1920-1945)

The model for Subset 3 (Figure 34) reveals that the most important variable affecting the outcome of battles between years 1920 and 1944 is artillery. The training set included years 1920-1944 and the test set, years 1944-1945.

5. Model 1.4: Subset 4 (Yrs. 1940-1982)

The model for Subset 4 (Figure 35) reveals that the most important variable between years 1940 and 1948 is tank ratio. The second most important variable is artillery ratio. The training set included years 1940-1948 and the test set, years 1950-1982.

6. Model 1.6: Subset 5 (Yrs. 1600-1982)

The model for Subset 5 (Figure 36) reveals that the most important variable affecting outcome of battles between years 1600-1944 is force ratio. The training set included years 1600-1944 and the test set, years 1940-1982.

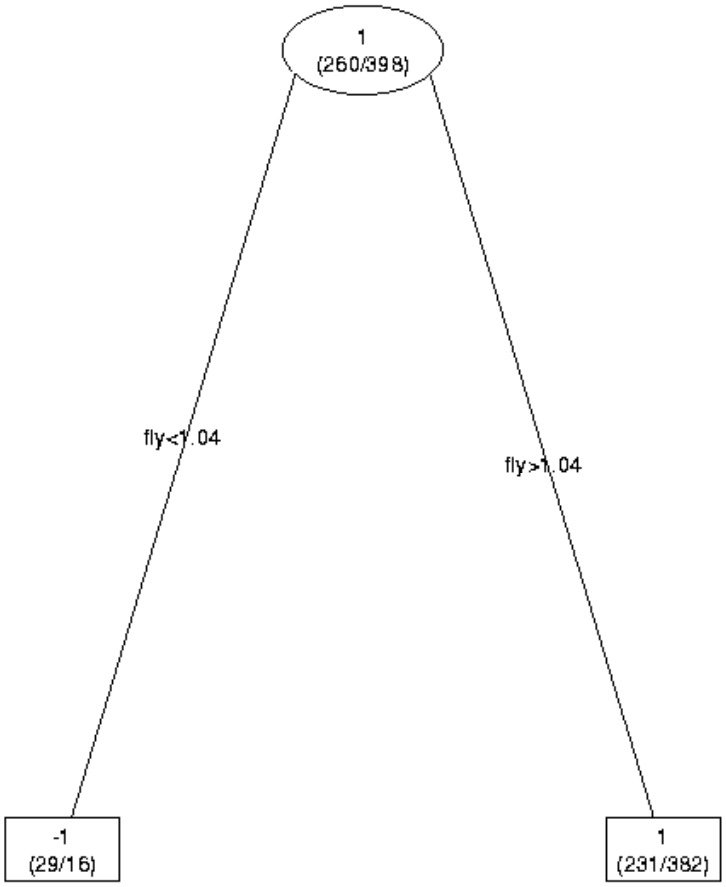

Figure 31. Model 1.1 for Entire Data Set

Model 1.1 for the entire data set (658 battles from 1600 to 1982) reveals the CAS sorties ratio as the most important variable. When the air sorties ratio is not available (in 484 battles), the split is done by a force ratio criterion (a surrogate split). If force ratio is less than 0.4, the battle outcome is classified as "defender wins," otherwise "attacker wins." However, this model can explain only 62 percent of the outcomes in 658 battles. Compared to the base model, this model makes only a 0.02 improvement in the classification. Therefore, the model is not useful.

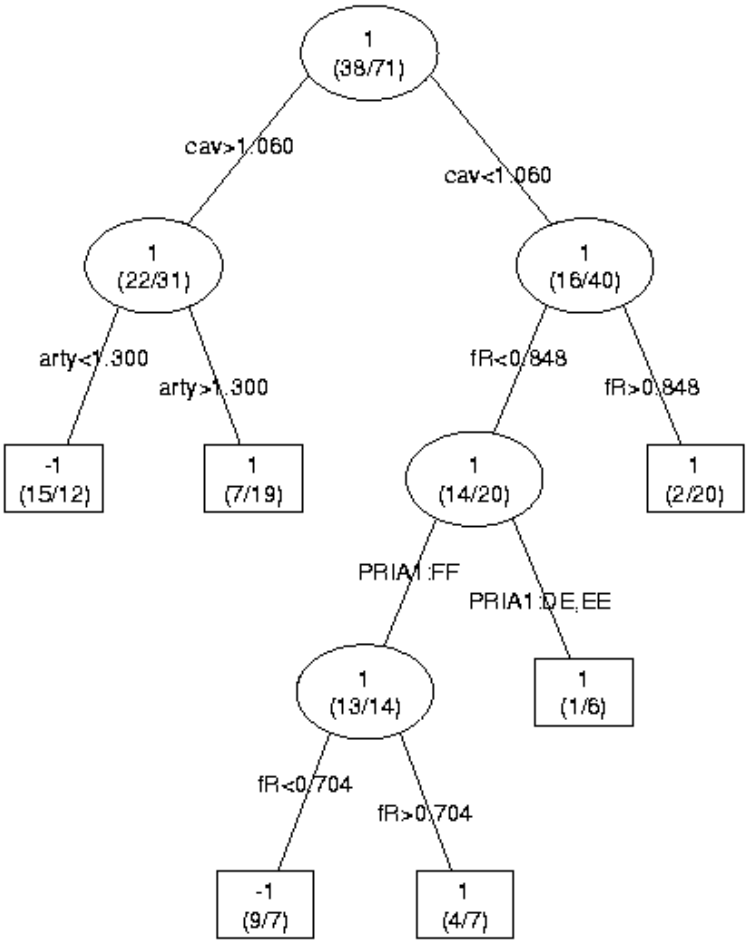

Figure 32. Model 1.2 for Subset 1

Model 1.2 can explain 69 percent of the battle outcomes in Training Set 1, which consists of 109 battles. The base model can explain 65 percent of outcomes. However, its predictions on Test Set 1 reveal a very high misclassification rate of 0.55, which is 0.10 worse than the base model prediction. On the Test Set 1.b, the model prediction is 0.22 worse than the base model. For this subset, the classification model is not informative, and it does not have predictive power. The model reveals the cavalry ratio as the most important variable. In Chapter II, our analyses show that the cavalry ratio is not expected to be a good predictor of battle outcomes. The split of attacker's primary tactics, "PRIA," is interesting. After restricting the set of battles by cavalry ratio and force ratio, the model assigns the battle as win for the attacker when the tactic is single or double envelopment. These are the preferred offensive tactics even today.

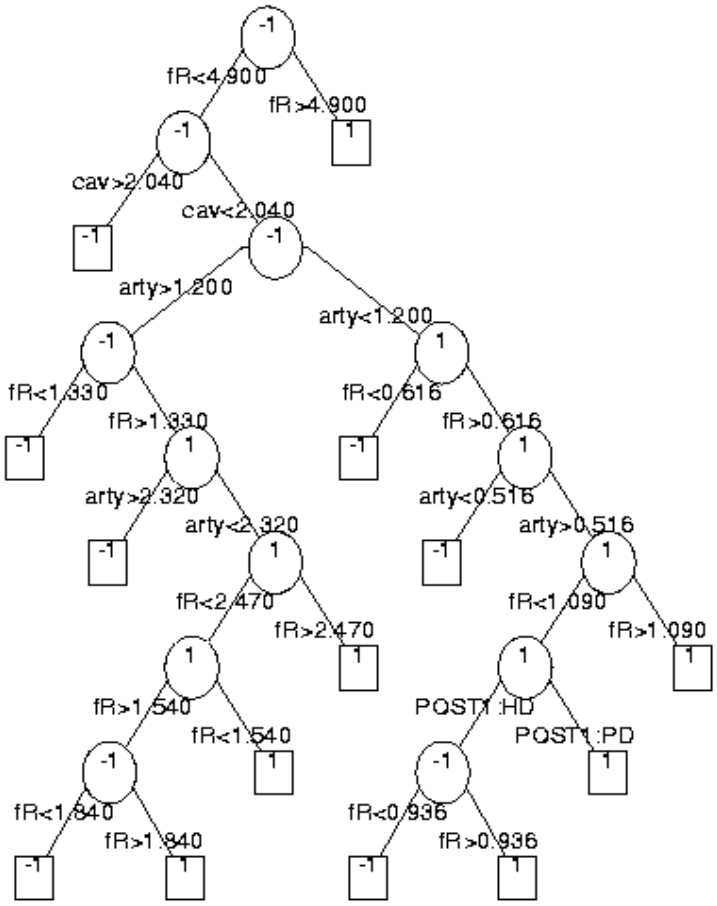

Figure 33. Model 1.3 for Subset 2

Model 1.3 for Subset 2 can explain only 69 percent of the outcomes in the training set, which consists of 178 battles. Predictions made by using this model on the Test Set 2 reveal a misclassification rate of 0.57, which is 0.25 worse than the base model predictions. On the Test Set 2.b, the predictions of the classification model are 0.31 worse than the base model predictions. As a result, this model is not informative, and it does not have predictive power. The first split criterion of the model is force ratio. However, there are only nine battles on that leaf (fR>4.9). In this training set, the attacker won all the battles when it had a force ratio of 4.9/1 and greater. The second split criterion is cavalry ratio. However, when the cavalry ratio is higher than 2.03, which means the attacker had over two times cavalry than the defender, the model assigned victory to the defender. This anomaly is parallel to our findings in Chapter II. Note: The tree picture for this model does not contain the number of battles at the leaves, because the model has a high number of splits and the graph does not appear as readable.

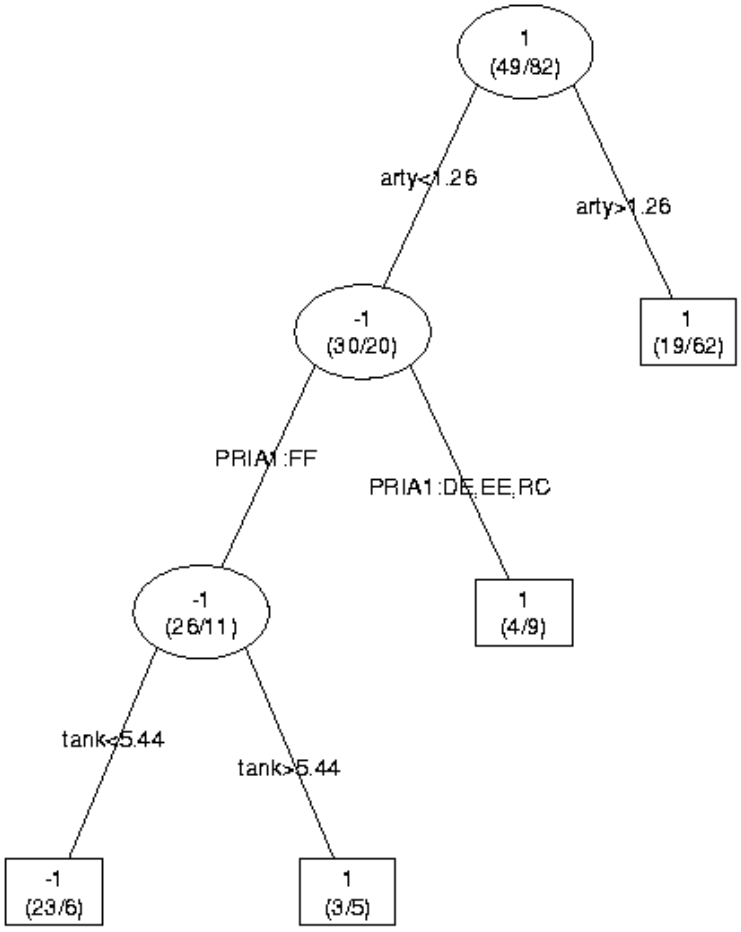

Figure 34. Model 1.4 for Subset 3

Model 1.4 can explain 69 percent of the battle outcomes in the Training Set 3, which consists of 131 battles. Predictions made with this model on Test Set 3 reveal a misclassification rate of 0.34, which is 0.04 better than the base model predictions. For Test Set 3.b, the difference between misclassification rates of the classification model and the base model is 0.03. We can conclude that the Model 1.4 has low prediction power. The first split criterion is for the model is artillery ratio. The second criterion is the attacker's tactical scheme. The model assigned attacker as "winner" when it used an offensive tactic other than the frontal attack. The third split criterion is tank ratio.

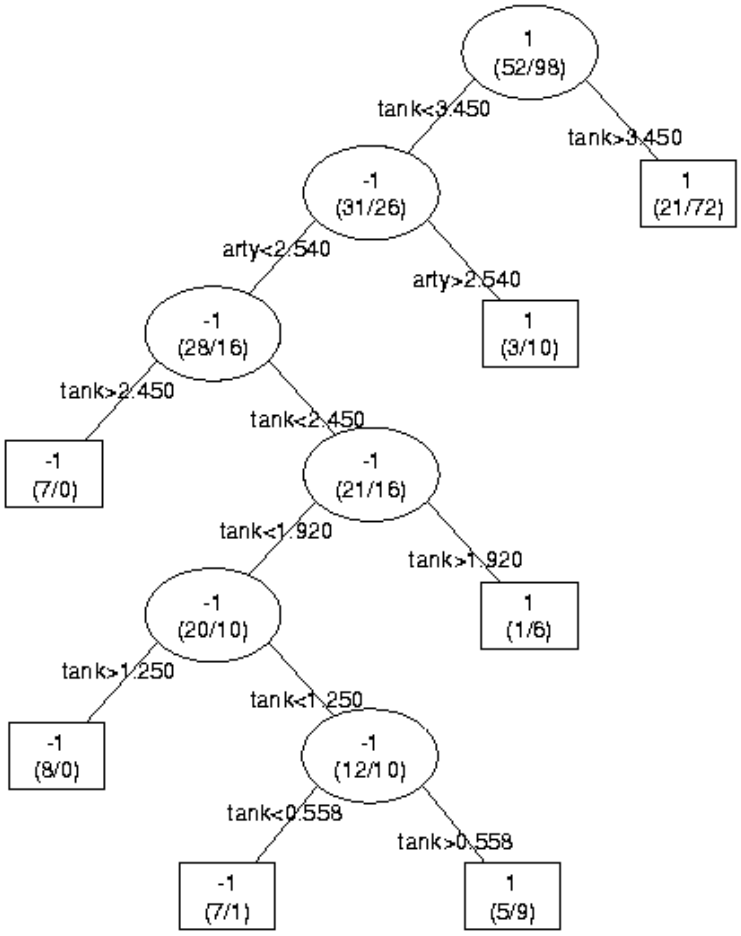

Figure 35. Model 1.5 for Subset 4

The Model 1.5 can explain the outcome of 79 percent of the battles in Training Set 4, which consists of 150 battles. Predictions of the model on Test Set 4 reveal a misclassification rate of 0.29, which is 0.01 better than the base model predictions. For Test Set 4.b, the predictions of the classification model are 0.04 worse than the base model predictions. We may conclude that the predictive power of this model is not good. The first split criterion in the model is tank ratio. The second is artillery ratio and the others are again tank ratio.

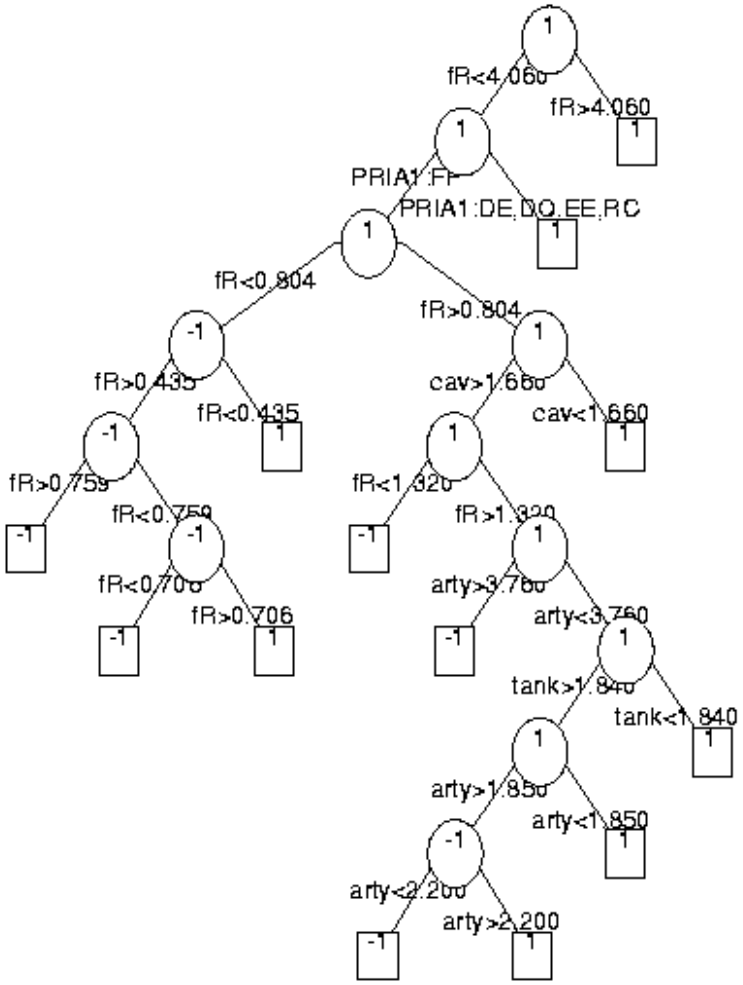

Figure 36. Model 1.6 for Subset 5

The Model 1.6 can explain 62 percent of battle outcomes in Training Set 5, which consists of 435 battles. Predictions made by using this model on Test Set 5 reveal a misclassification rate of 0.42, which is 0.09 worse than the base model predictions. Regarding the predictions on the Test Set 5.b, the base model is 12 percent better than the classification model. This model is not informative, and it does not have predictive power.

7. Conclusion

Models built by using the objective values result in high misclassification rates (see Table 32). We can conclude that numbers alone are not reliable predictors of battle outcomes. This result may cast some suspicion on large-scale simulations, which cannot simulate the intangible factors of battles very well.

Subset	Misclassification Rate of the Training Set	Misclassification Rate of the Test Set	Misclassification Rate of the Test Set with Clear-cut Outcomes
Subset 1 Yrs. 1600-1847	0.30	0.55	0.60
Subset 2 Yrs. 1805-1918	0.31	0.57	0.58
Subset 3 Yrs. 1920-1945	0.31	0.34	0.13
Subset 4 Yrs. 1940-1982	0.21	0.29	0.36
Subset 5 Yrs. 1600-1982	0.32	0.42	0.37
Data set Yrs. 1600-1982	0.38	NA	NA

Table 32. Misclassification Rates of the Models for Objective Variables

The classification models resulted in high misclassification rates in test sets and the test sets only with clear-cut outcomes. Since the models are built by using training sets, the high misclassification rates on the training sets suggest that the objective variables were insufficient to describe the outcome of battles. The highest misclassification rate of this model is 60 percent on Test Set 1.b. The lowest is on Test Set 3.b. The large differences between the misclassification rates show that some of the models are unreliable.

D. MODEL 2: OBJECTIVE AND RELATIVE VARIABLES

Model 1, using only objective values, resulted in high misclassification rates on both the training set and the test set. Objective variables are not reliable predictors of battle outcome. In Chapter II, the correlation matrix reveals the same result, namely, low correlation rates with WINA and objective variables. In Model 2, we are going to use both relative variables and objective variables.

1. Model 2.1: Entire Data Set (Yrs. 1600-1982)

The model for the entire data set, see Figure 37, reveals that the most important factor affecting the outcome of the battles throughout history is relative initiative advantage. The other important variables are relative leadership advantage and force ratio.

2. Model 2.2: Subset 1 (Yrs. 1600-1847)

The model for Subset 1, see Figure 38, shows that the most important factor affecting the outcome of the battles was the relative leadership advantage.

3. Model 2.3: Subset 2 (Yrs. 1805-1918)

The model for Subset 2, see Figure 39, shows that the most important factor affecting the outcome of the battles is again the relative leadership advantage.

4. Model 2.4: Subset 3 (Yrs. 1920-1945)

The model for Subset 3, see Figure 40, reveals relative initiative advantage as the most important variable affecting the outcome of the battles. The other important variables are artillery ratio and combat effectiveness advantage.

5. Model 2.5: Subset 4 (Yrs. 1940-1982)

The model for Subset 4, see Figure 41, reveals that the most important variable affecting the outcome of the battles is relative initiative advantage.

6. Model 2.6: Subset 5 (Yrs. 1600-1982)

The model for Subset 5, see Figure 42, reveals that the most important variable affecting the outcome of the battles is relative leadership advantage. The other important variables are relative initiative advantage and force ratio.

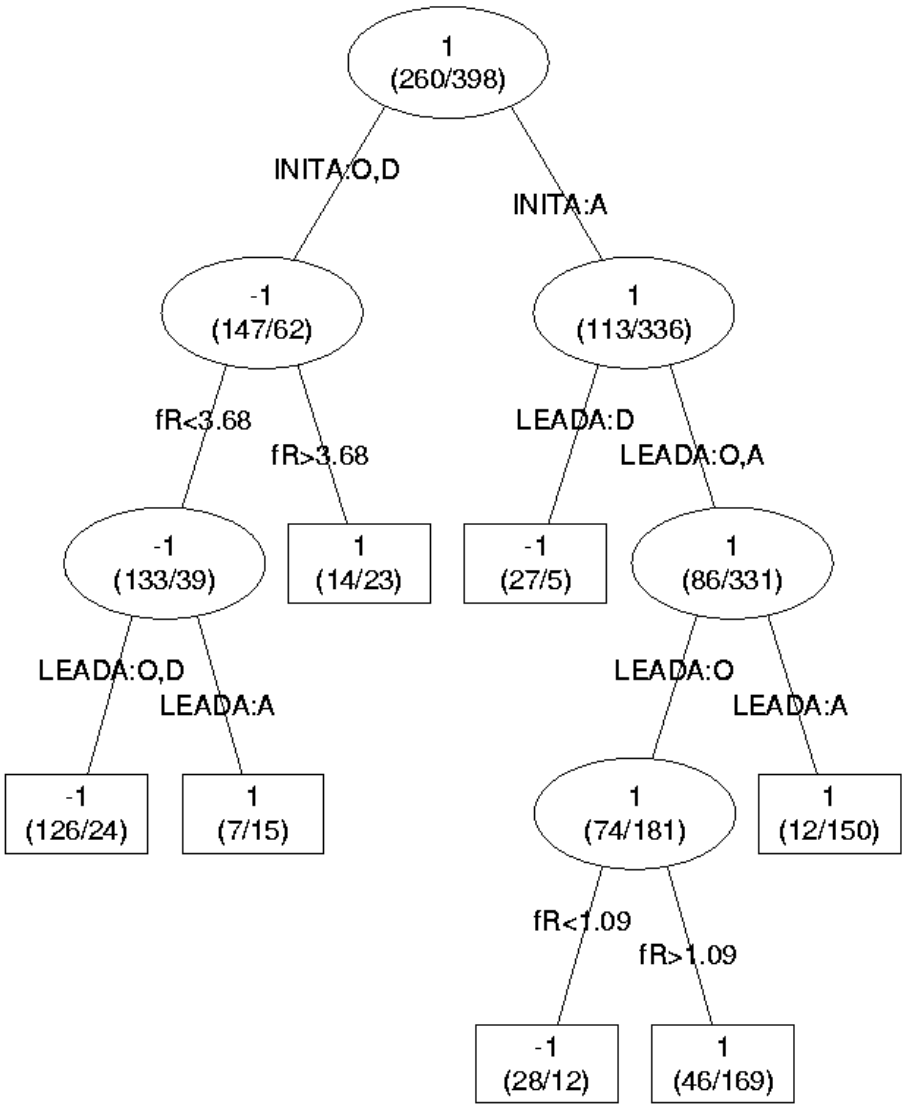

Figure 37. Model 2.1 for Entire Data Set

Model 2.1 can explain the outcomes of 82 percent of the battles in the entire data set of 658 battles. The first split criterion is relative initiative advantage. After that, force ratio and relative leadership advantage gains importance.

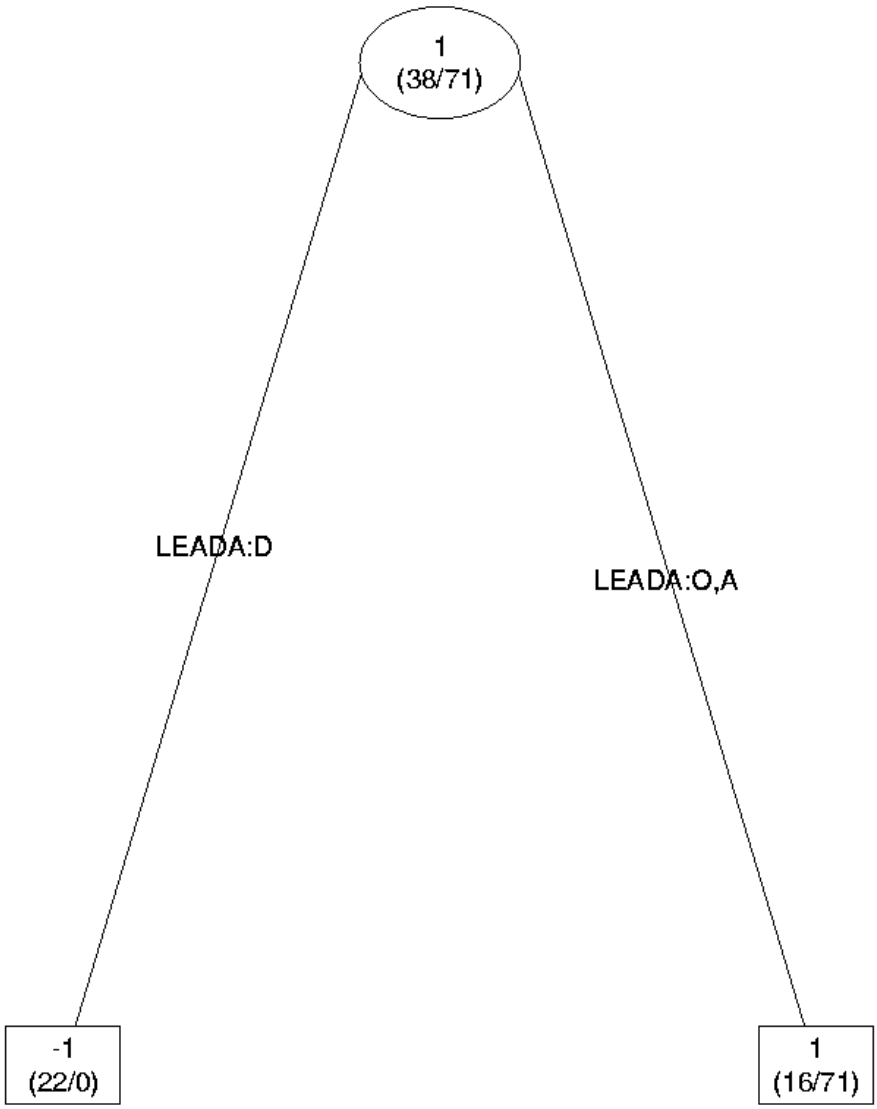

Figure 38. Model 2.2 for Subset 1

Model 2.2 can explain 85 percent of the battle outcomes in Training Set 1. Predictions of this model on Test Set 1 reveal a misclassification rate of 0.20, which is 0.25 better than the base model predictions. The prediction on Test Set 1.b is 0.22 better than the base model prediction. Battles are classified only according to the relative leadership advantage. The relative leadership advantage was more important than all other factors in determining the battle outcome between the years 1600-1799.

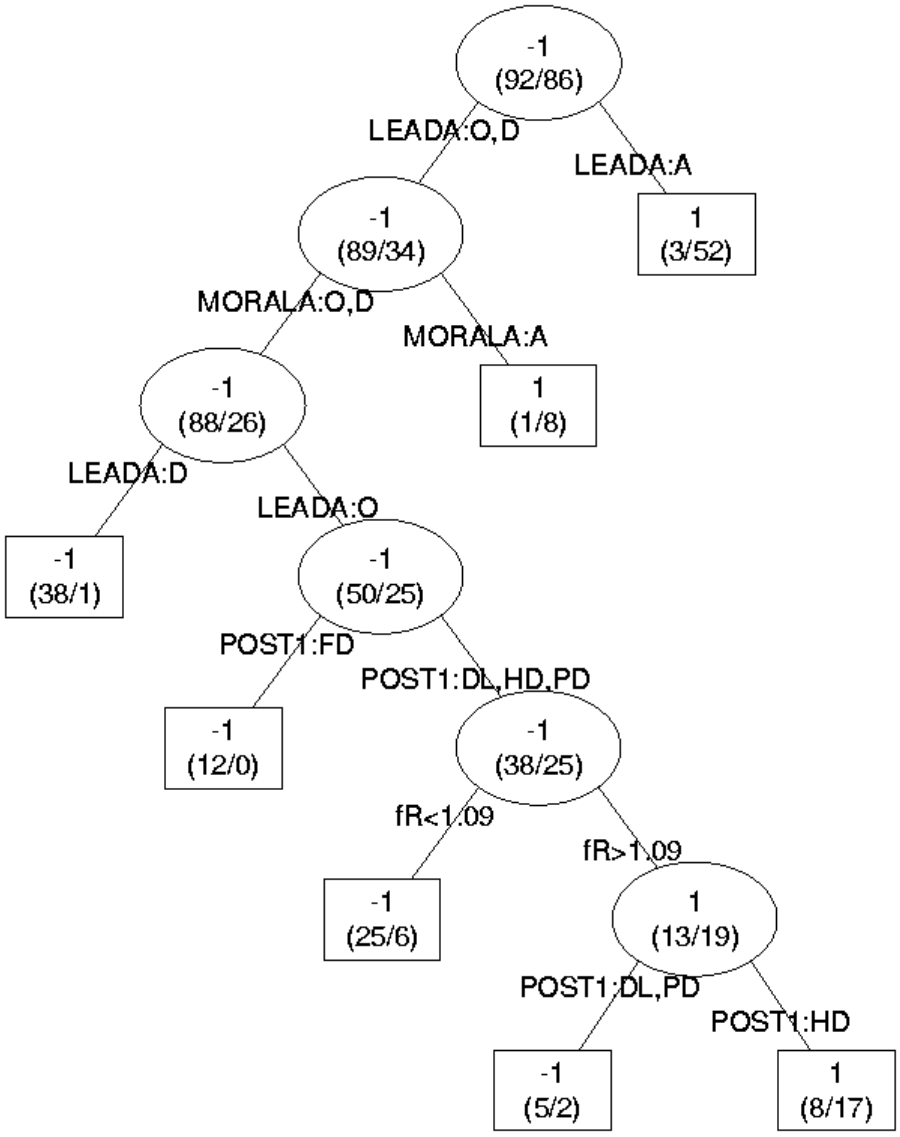

Figure 39. Model 2.3 for Subset 2

Model 2.3 can explain 88 percent of the battle outcomes in Training Set 2. Predictions made with this model on Test Set 1 reveal a misclassification rate of 0.24, which is 0.08 better than the base model predictions. Predictions on the Test Set 2.b are 0.07 better than the base model predictions. The most important variable is the relative leadership advantage. Other decisive variables are morale advantage, force ratio and defender's posture. The low misclassification rates show that this model is useful and has predictive power.

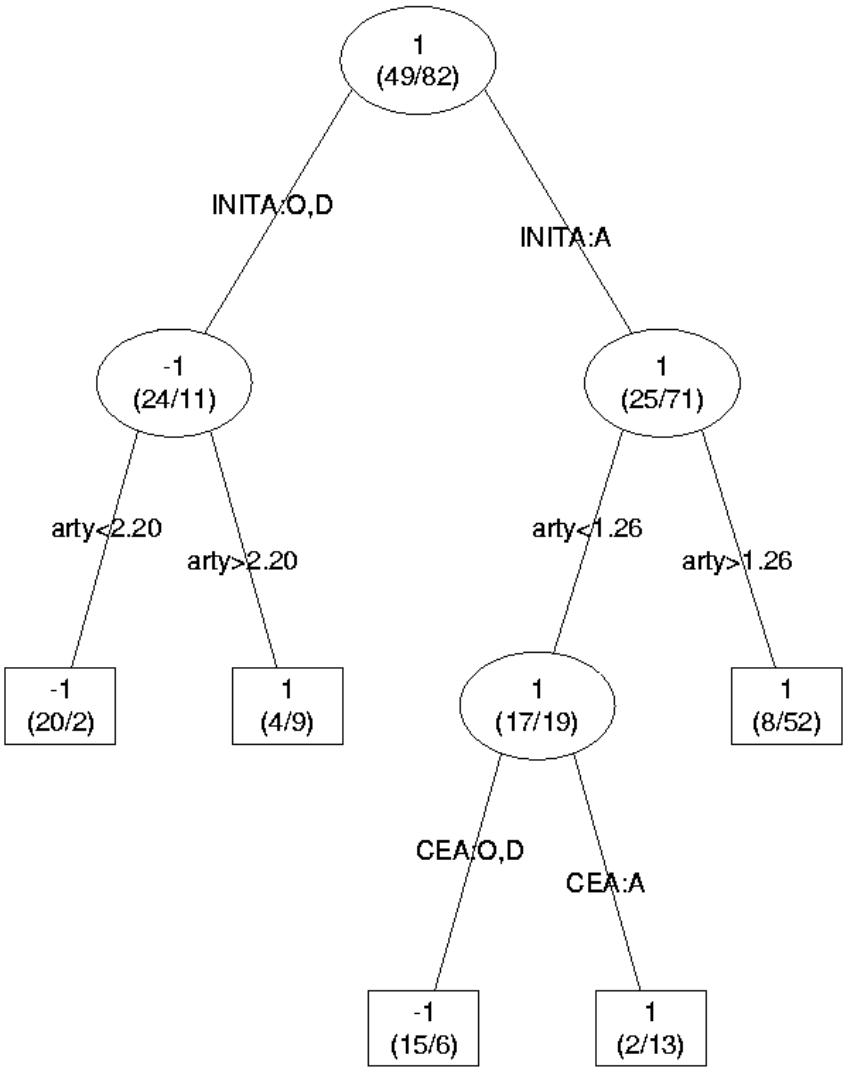

Figure 40. Model 2.4 for Subset 3

Model 2.4 can explain 83 percent of the battle outcomes in Training Set 3. Predictions of this model on Test Set 3 reveal a misclassification rate of 0.30, which is 0.08 better than the base model predictions. Predictions on Test Set 3.b are 0.04 better. The most important variable is relative initiative advantage. The other important variables are the artillery ratio and combat effectiveness advantage. The low misclassification rates show that this model is useful and have predictive power.

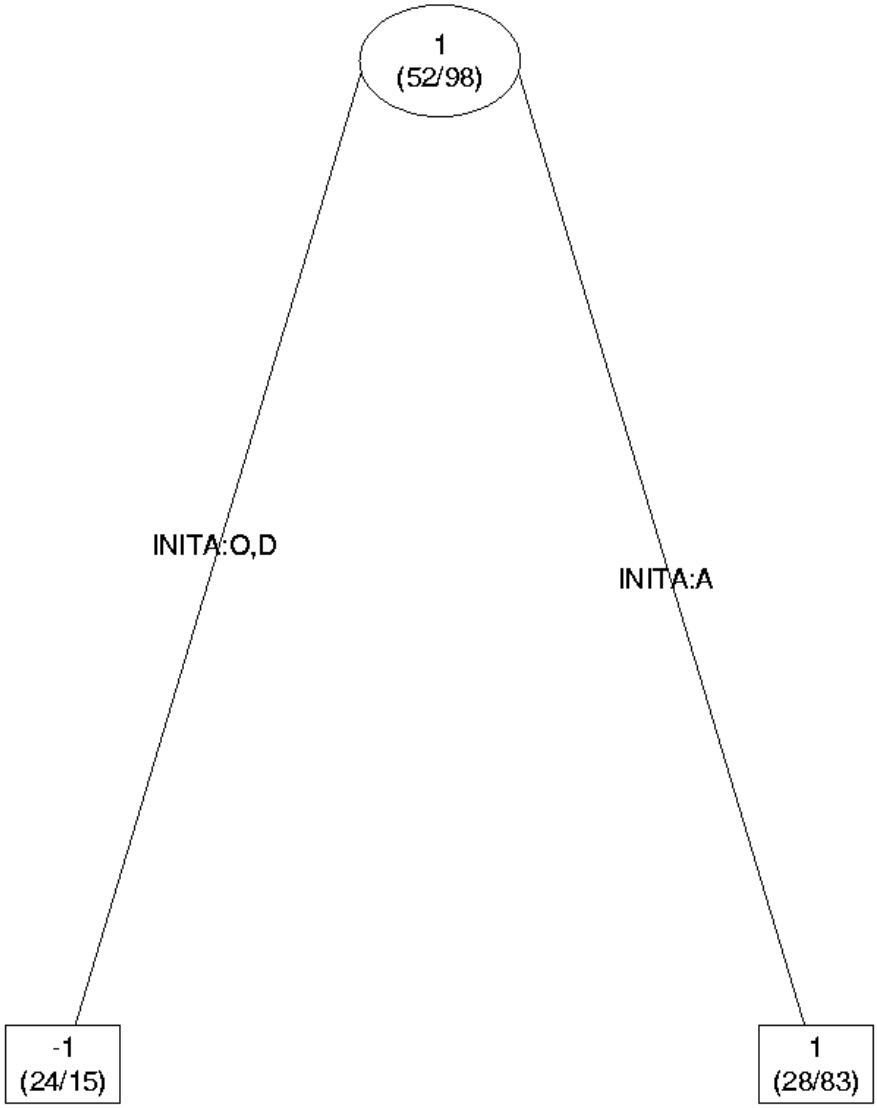

Figure 41. Model 2.5 for Subset 4

Model 2.5 can explain 71 percent of the battle outcomes in Training Set 4. Predictions of this model on Test Set 4 reveal a misclassification rate of 0.34, which is 0.04 worse than the base model predictions. Predictions on Test Set 4.b are 0.06 better than the base model predictions. The most important variable is relative initiative advantage. The training set of Subset 4 mostly contains World War II battles, in which tanks and air power played a great role. Intuitively, we would expect tank ratio and CAS sorties ratio to have importance in determining the battle outcome.

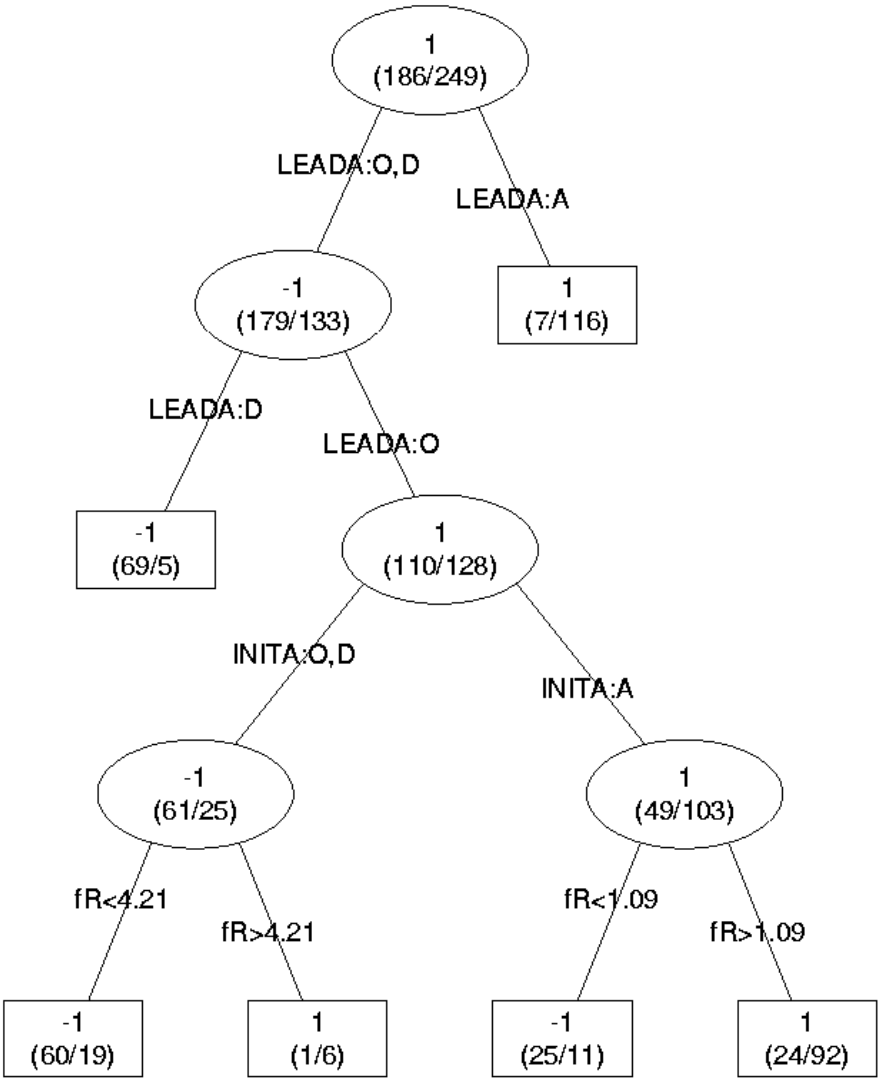

Figure 42. Model 2.6 for Subset 5

Model 2.6 can explain 85 percent of the battle outcomes in Training Set 5. Predictions of this model on Test Set 5 reveal a misclassification rate of 0.26, which is 0.17 better than the base model predictions. Predictions on Test Set 5.b are 0.10 better. The most important variable in the model is the relative leadership advantage. Other variables only gain importance when neither side has the leadership advantage. In that case, relative initiative advantage and force ratio gain importance.

7. Conclusion

The models formed by using the objective variables and relative factors have better misclassification rates than the models with objective variables. The most important factor in the models turns out to be the relative leadership advantage.

	Misclassification Rate of the Training Set	**Misclassification Rate of the Test Set**	**Misclassification Rate of the Test Set with Clear-Cut Outcomes**
Subset 1 Yrs. 1600-1847	0.15	0.20	0.16
Subset 2 Yrs. 1805-1918	0.12	0.24	0.20
Subset 3 Yrs. 1920-1945	0.17	0.30	0.13
Subset 4 Yrs. 1940-1982	0.29	0.34	0.26
Subset 5 Yrs. 1600-1982	0.15	0.26	0.15
Data Set Yrs. 1600-1982	0.18	NA	NA

Table 33. Misclassification Rates of Model 2 Subsets

The models can explain the battle outcomes in the training sets as high as 88 percent. Predictions on the test sets are better than the Model 1 predictions. Predictions on the Test Sets with Clear-cut Outcomes are even more accurate. The worse prediction was 74 percent right, and the best prediction was 87 percent right. This result may be close to the limit of what we can predict, when we consider the variability in the data set and the role of luck in battles that can change the battle outcome.

E. MODEL 3: OBJECTIVE AND RELATIVE VARIABLES; TERRAIN, AND WHETHER

None of the models built with terrain and weather variables selected terrain and weather as predictor variables. In addition, the misclassification rate of the models did not get better. Thus, we can conclude that the weather conditions and terrain did not enable us to predict the battle outcomes more correctly. Because of this, Classification trees for Model 3 are not included in the text.

F. IMPORTANT VARIABLES AND MISCLASSIFICATION RATES OVER TIME

1. Introduction

In order to answer the question of "Can we predict the outcome of battles?" we built models using a set of earlier battles to predict the next one. The process works as follows:

Take a number of battles as a training set, build a model, and predict the outcome of the next battle. Take the next training set and predict the battle after that. In other words, if we decide to make the training set size 100, then the first 100 battles will be used to build a model to predict the 101st battle. For the second prediction, take the second through 101st battles, build a model and predict the 102nd battle. In this way, we are making (658 − training set size) predictions. We also record the correct and incorrect predictions, split criteria in each model, and the year of the test set. In this way, we can see how models change over time.

One of the challenges with this method is determining the size of the training set. Does the ability of past battles to predict the next one diminish with time? In order to find the best size, we began using the model with different training set sizes. To predict the same battles, we always start predicting the 201st battle. The result shows that larger training sets usually have slightly better misclassification rates.

Size	**50**	**75**	**100**	**125**	**150**	**175**	**200**
Misclassification rate	0.32	0.29	0.27	0.27	0.28	0.28	0.38

Table 34. Misclassification Rates versus Training Set Size

Classification trees, built by using a training set size of 125, give the lowest misclassification rate of 27 percent. The resulting run produces 533 trees (658-125) and 533 predictions. Figure 43 shows the misclassification rates of the predictions. The figure is formed by taking 30 predictions and calculating the misclassification rate, in a moving window method. In order to show the trend in misclassifications, an overlap is made. The first data point in Figure 43 shows the misclassification rate of first 30 predictions. The second point shows the misclassification rate of sixth through 35th predictions and so on.

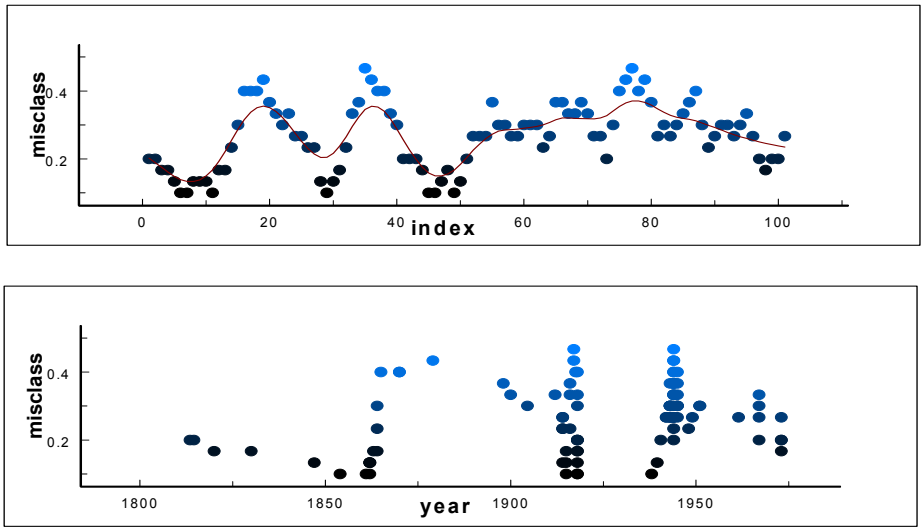

Figure 43. The Misclassification Rate

The misclassification rates ranged from 0.1 to 0.53. The first graph shows sharp increases and decreases in the misclassification rates. The second graph reveals that the sharp increases happened in the mid-19th century and in two World Wars. In the mid-19th century, gunpowder weapons were introduced, bringing about sharp increase lethality [Ref. 16:p.25]. In the two Word Wars, powerful new weapons, such as chemical weapons, tank and CAS are used. Each classification tree is built by using the previous 125 battles. These trees show the trend in the past battles simply and the prediction of the next battle is made by using this trend. However, during periods of dramatic changes in weapons and tactics, the trend in past battles fails to predict the next (very different) battle. When the trees are built by the data showing the changes, the models begin to make better predictions. This result shows that data from earlier battles may have little value in understanding modern battles. In order to use historical combat data effectively, we need more data from the latest battles.

2. Checking the Assumptions

The 533 classification trees built with size 125 revealed an overall misclassification rate of 27 percent. However, at the start we had made some assumptions. In this section, the effects of these assumptions will be checked.

First, at the beginning of our analyses, we had assumed draws as wins on the defender's side. In order to test this assumption, first, all draws are discarded from the data set. The resulting data set has 615 battles. The resulting models have a 25 percent misclassification rate. This is slightly better than the previous model.

Second, draws are assigned as wins to the attacker. The resulting models have a 26 percent misclassification rate, which is no better than the previous ones.

As a third model, the original battle outcome variable, "WINA" is used. The original battle outcome variable has three levels, "1," "0," and "−1." The models using the original battle outcome variable revealed a 30 percent misclassification rate, which is worse than all the previous ones.

A fourth model includes only the battles with clear-cut outcomes in training sets and test sets (data size=465). The models revealed a 21 percent misclassification rate, which is the best rate of all. In other words, classification models built on training sets with size 125 correctly predicted 79 percent of 340 battles (465−125=340).

Fifth, some battles are fought with two million men, like the Moscow defense in WWII, and some fought with very small numbers, like the battles of Okinawa. In order to distinguish battles with a high number of participants from small battles with brigade size troops, a "divisive clustering" analysis is made. Then, the data is assigned to three clusters with regard to year and total personnel strength on both sides. The data in the largest cluster is used to build trees. The resulting models have a 28 percent misclassification rate. This is not a great difference.

Sixth, another run is made by using only battles of World War II and after. The resulting models have a misclassification rate of 0.28.

3. Relative Importance of Variables

a) First-Split Criterion

For a split criterion, a classification tree model chooses the variable that most affects the response variable. If we look at the first split criteria in the models, we can comprehend the relative importance of the variables. With a training set size of 125, 533 classification trees are built. Table 35 shows the percentages of variables that appear in the first split. The relative initiative advantage appeared 40 percent of the time. The relative leadership advantage appeared in the first split 38 percent of the time. However, some of the variables, such as CAS sorties ratio and tank ratio were not available in battles throughout history. Figure 44 gives a better picture of the first split criterion. The relative leadership and initiative advantage appeared in the first split when more technical and powerful weapons, such as tanks and airplanes, were not available to fighters.

As a result, we can say that the importance of factors affecting the battle outcome changed through history. Relative factors yielded to objective factors.

Variable	**INITA**	**LEADA**	**fly**	**arty**	**tank**	**MOMNTA**	**fR**	**TRNGA**
Frequency	0.40	0.38	0.07	0.05	0.05	0.03	0.01	0.01

Table 35. The First Split of the Classification Models

The relative initiative advantage appeared in the first split 40 percent of the time. Second, the relative leadership advantage appeared 38 percent of the time.

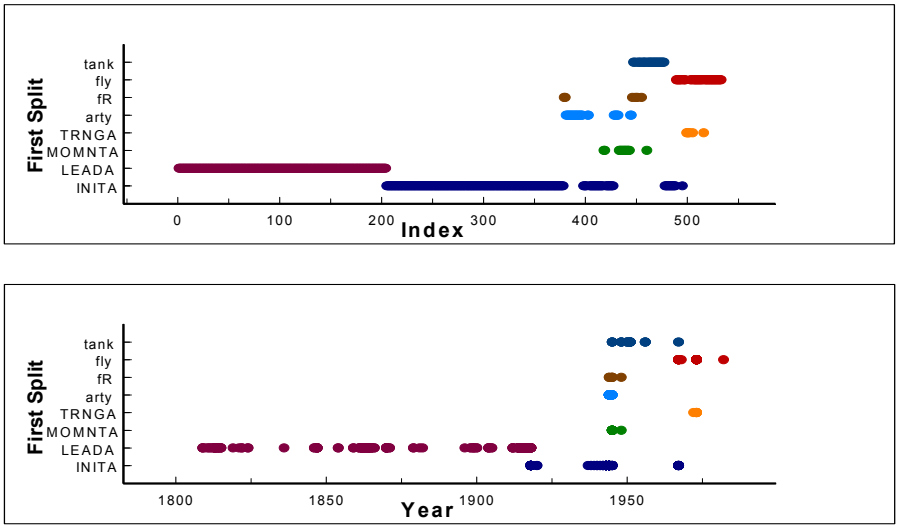

Figure 44. First-Split Criteria of Classification Models

The relative leadership advantage appeared as the first-split criterion in the first battles of the data set. The relative initiative advantage became important in World War I. In World War II and after, objective variables, such as tank and CAS sorties ratio gained precedence over other variables.

b) All-Splits Criterion

The first-split criterion gives information, but some of the variables may appear on other splits. To understand the overall importance of variables, all split criteria and their respective positions are recorded for the 533 classification trees. On the other hand, the importance of the first split is somehow greater than the importance of the last split. To account for this, a weighted sum is used. However, finding optimal weights is another challenge, and it would be beyond the scope of the thesis--if they exist. We give the weights intuitively, weighing the first split the most. The first splits are much more important than the following splits. They help classify a large portion of the data, while the later splits classify a small portion of data. Because of this property, we gave higher weights to the first splits and lower weights to the later ones. The chosen weights are 100 for the first split, 50 for the second split, 25 for the third split, 12 for the fourth split, six for the fifth split, five for the sixth split, four for the seventh split, three for the eighth split, two for the ninth split, and one for the tenth split.

From the argument in the first-split criterion, we know that the importance of variables changed throughout history. In order to understand the difference, two runs are made to build classification trees for battles before World War II and wars during and after World War II. Table 36 shows the split criteria for battles before World War II, and their respective positions and Table 37 shows the split criteria for battles of Word War II and after.

Position	1	2	3	4	5	6	7	8	9	10
AEROA	0	0	6	9	0	0	1	0	0	0
CEA	0	0	2	14	6	0	0	0	0	0
INITA	56	6	17	15	6	0	0	0	0	0
INTELA	0	0	0	2	2	3	0	1	0	0
LEADA	204	35	5	2	0	6	0	1	1	0
LOGSA	0	0	17	0	0	0	0	0	0	0
MORALA	0	15	28	6	2	2	1	0	1	2
POST1	0	0	1	0	0	0	2	0	0	0
SURPA	0	0	0	0	13	0	0	1	0	0
TRNGA	0	0	0	1	1	0	0	0	0	0
arty	0	0	3	8	1	0	0	2	0	0
cav	0	0	0	2	1	0	4	0	0	0
fR	0	58	22	15	10	1	0	1	3	0
weights	**100**	**50**	**25**	**12**	**6**	**5**	**4**	**3**	**2**	**1**

Table 36. The Split Criteria, the Respective Positions, and Weights for Battles Before WWII

Position	1	2	3	4	5	6	7	8	9	10	11	12	13
AEROA	12	0	0	4	2	6	6	2	3	3	1	0	0
CEA	0	21	18	0	1	6	4	11	1	1	0	0	0
INITA	4	19	9	12	2	4	8	3	1	1	0	0	0
INTELA	0	35	14	5	14	1	1	0	0	0	0	0	0
LEADA	0	0	0	0	1	0	0	0	0	0	0	0	0
LOGSA	0	0	0	3	0	1	1	0	0	0	0	0	0
MOMNTA	0	5	0	0	0	0	1	0	0	0	0	0	0
POST1	0	0	17	2	7	4	2	3	5	2	1	0	0
PRIA1	0	0	0	9	2	1	0	4	1	1	0	0	0
SURPA	0	10	4	5	3	0	0	0	0	0	2	0	0
TRNGA	10	1	0	0	0	0	0	0	1	0	0	0	0
arty	48	14	10	22	7	6	9	8	5	6	2	4	1
fR	0	1	1	8	12	7	12	8	10	5	6	1	2
fly	70	0	1	0	0	3	1	4	1	1	0	0	0
tank	4	3	23	15	16	20	10	6	11	6	4	4	0
Weights	**100**	**50**	**25**	**12**	**6**	**5**	**4**	**3**	**2**	**1**	**1**	**1**	**1**

Table 37. Split Criteria, Respective Positions, and Weights for Battles of World War II and After

As a second step, the number of appearances in the order of splits is multiplied by the weights. In order to normalize the values, the resulting values are divided by the sum of all. Figure 45 shows the normalized values for both the variables in both tables.

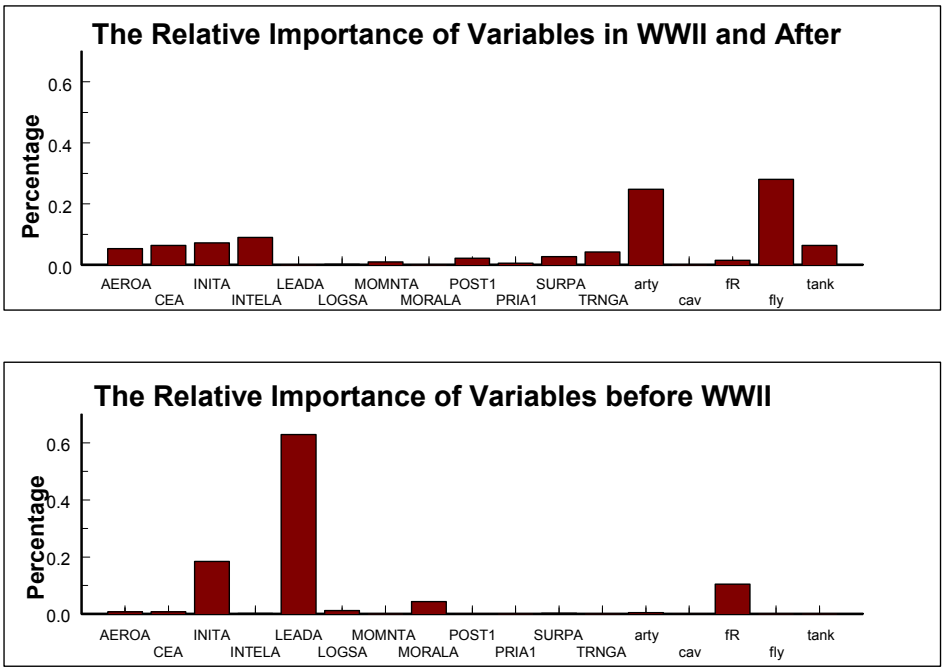

Figure 45. The Relative Importance of Variables

Before World War II, the most important variables were the relative leadership advantage and the relative initiative advantage. The force ratio was the third ranking variable. However, in World War II, the dynamics of battle had changed. After that, objective variables, such as CAS sorties ratio, artillery ratio, tank ratio became more important than the leadership advantage. The importance of relative variables, such as combat effectiveness advantage and relative intelligence advantage increased during and after World War II.

G. CONCLUSION

Models using only objective variables resulted high misclassification rates. However, models using both objective and relative variables revealed low misclassification rates. As a result, objective variables alone are not reliable predictors of battle outcomes.

Moreover, the importance of variables affecting the battle outcome has also changed through history. The relative factors such as leadership and initiative advantage gave their importance to objective factors such as tank ratio and CAS sorties ratio.

In addition, the models show that data from earlier battles have a little value in understanding the modern battles. In order to use historical combat data effectively, we need more data from the latest battles.

IV. CONCLUSION

The analyses in the previous chapters reveal some important results. In Chapter I, the literature about historical combat data analysis reveal that historical combat data analysis can help to understand different aspects of battles which cannot be well simulated. The descriptive statistics in Chapter II analyzes some of the variables that can affect the outcome of battles. The descriptive statistics show that some variables had a higher correlation with the battle outcome than others. The classification models in Chapter III reveal that the battle outcomes can be predicted by using classification tree models built by with historical combat data. In this chapter, we will evaluate the validity of our findings and their prospective uses.

The predictions of battle outcomes by using classification trees revealed as high as 79 percent correct (clear-cut outcomes). This result is satisfying when the role of luck in battles and hard to quantify factors are considered. The classification models also provided some information about the relative importance of variables in relation to the outcome of battles. In addition, these models showed how this relationship changed throughout history.

However, the validity of these results is directly related to the validity of the data set. We have contacted Dr. Helmbold, who has supervised the CBD90G data set, about the validity and variability in the data set. He pointed out that there were many uncertainties in the data set, and the models are affected by the diverse definitions and categories. However, despite its shortcomings, it was the far and away the best available data base for statistical analyses. Dr. Helmbold also pointed out that the models using this database should be confirmed by using independently compiled databases.

About the relative variables we have used in building classification trees, Dr. Helmbold proposes two approaches. One approach is to consider them as objective and reliable. In this view, these variables can be used as predictor variables, as in Model 2. The second approach is to think that these variables are “not really very sound, but instead are ‘after the fact’ appraisals contaminated by already knowing the battle

outcome." In this approach, the results of Model 2 do not carry any importance. He points out one way to solve this problem as making predictions before the prospective battles. However, this approach has many challenges, such as assessing classified information.

Descriptive statistics of the objective variables reveal that the battle outcome is not highly correlated with these variables. The conditional plots of these variables show distributions that are similar when the attacker won and when it lost the battle. Huge outliers in conditional graphs also show that some other important variables affected the outcome of a battle. In addition, Model 1, which is built by using the objective variables alone, revealed high misclassification rates on predictions. These results show that the objective variables alone are not sufficient to determine the outcome of a battle. Thus the intangibles, relative factors, must have had a high effect.

Our results are useful in providing some insights about battles. There results are not precise because of the variability in the data and inclusion of relative factors. However, classification models provide information on how the importance of variables changed thorough history and which factors have most affected the battle outcome.

LIST OF REFERENCES

1. Dr. Helmbold, Robert L, "*A Survey of Past Work on the Rates of Advance in Land Combat Operations,*" U.S. Army Concepts Agency, 1990.
2. Dupuy, Col. T. N. (U. S. Army, Ret.), "*Numbers, Predictions and War,*" Hero Books, 1985.
3. Yigit, Faruk, *"Finding the Important Factors in Battle Outcomes: A Statistical Exploration of Data from Major Battles,"* Masters Thesis, Naval Postgraduate School, 2000.
4. Speight, L.R, Rowland, D, *"Modelling the Mobile Land Battle: Combat Degradation and Criteria for Defeat,"* Military Operations Research, V4 N3, 1999.
5. Requirements and Resources Directorate, "*Combat History Analysis Study Effort (CHASE): Progress Report for the Period August 1984-June 1985,*" U.S. Army Concepts Analysis Agency, 1986.
6. Office Special Assistant for Model Validation, "*Do Battles and Wars Have a Common Relationship between Casualties and Victory,*" U.S. Army Concepts Analysis Agency, 1987.
7. McQuie, Robert, "*Historical Characteristics of Combat for War Games (Benchmarks),*" U.S. Army Concepts Analysis Agency, 1988.
8. Venables, W.N., Ripley, B.D., *"Modern Applied Statistics with S-PLUS,"* Springer, 1999.
9. Chambers, John M, Hastie, J. Trevor, "*Statistical Models in S,*" p.378, Wadsworth & Brooks/Cole Advanced Books & Software, 1992.
10. Therneau, Terry M., Atkinson, Elisabeth J., Mayo Foundation *"An Introduction to Recursive Partitioning Using the RPART Routines,"* [http://www.stats.ox.ac.uk/pub/SWin/rpartdoc.zip]. September 2001.
11. *S-PLUS 2000 Guide to Statistics Vol I,* Data Analysis Products Division, Seattle, WA: Insightful Corp., 1999.
12. Sun Tzu, "*The Art of War,* " [http://classics.mit.edu//Tzu/artwar.html]. October 2001.
13. Headquarters Department of Army, *"FM 100-5 Operations " Washington* DC, 1993.

14. Devore, Jay L., *"Probability and Statistics,"* Duxbury Thomson Learning, 2000.
15. Charlton, James, "The Military Quotation Book," St. Martin's Press, New York, 1990.
16. Dupuy, Col. T. N. (U. S. Army, Ret.), "*Attrition,*" Nova Publications, Falls Church, Virginia, 1995.

APPENDIX I. DEFINITIONS

For detailed descriptions, see [Ref.5: p.4-2].

Force Ratio (FR) is the ratio of the attacker's number of personnel (X_0) and the defenders number of personnel (Y_0).

$$FR = X_0 / Y_0 \quad (4)$$

Bitterness (EPS) and advantage (ADV) are derived from Lanchester Equations.

$$EPS = SQR (FX * FY) \quad (5)$$

where FX and FY are the attacker's and the defender's fractional losses.

The attacker's *Fractional Loss* (FX) is

$$FX = CX / X_0 \quad (6)$$

where CX is the attacker's number of casualties.

The defender's *Fractional Loss* (FX) is

$$FY = CY / Y_0 \quad (7)$$

where CY is the defender's number of casualties.

The *Casualty Exchange Ratio* (CER) is the ratio of the attacker's and the defender's casualties.

$$CER = CX / CY \quad (8)$$

The *Fractional Exchange Ratio* is the ratio of the attacker's and the defender's fractional losses. FER = FX / FY.

Advantage (ADV) is the defender's empirical advantage parameter. ADV = (1/2) * LOG (FER).

Residual Advantage is "residual value of ADV after the average effects of any differences in FR values are removed." [Ref.5: p. 4-11]

$$RESAV = ADV – a – b * LOG (FR) \quad (9)$$

where a and b are the regression coefficients.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX II. MODEL OUTPUTS

A. MODEL 1.1

```
1) root 658 300 1 ( 0.4 0.6 )
  2) fly<1.03759 45  20 -1 ( 0.6 0.4 ) *
  3) fly>1.03759 613 200 1 ( 0.4 0.6 ) *
Node number 1: 658 observations,    complexity param=0.05
  predicted class= 1  expected loss= 0.4
     class counts:  260 398
    probabilities:  0.4 0.6
  left son=2 (45 obs) right son=3 (613 obs)
  Primary splits:
      fly   < 1    to the left,  improve=10, (484 missing)
      tank  < 3    to the left,  improve= 6, (362 missing)
      fR    < 5    to the left,  improve= 6, (1 missing)
      PRIA1 splits as  RRRLR, improve= 5, (8 missing)
      arty  < 0.08 to the left,  improve= 3, (153 missing)
  Surrogate splits:
      fR   < 0.4  to the left,  agree=0.8, (483 split)
      arty < 0.07 to the left,  agree=0.8, (0 split)

Node number 2: 45 observations
  predicted class= -1  expected loss= 0.4
     class counts:  29 16
    probabilities:  0.6 0.4

Node number 3: 613 observations
  predicted class= 1  expected loss= 0.4
     class counts:  231 382
    probabilities:  0.4 0.6
```

B. MODEL 2.1

```
1) root 658 300 1 ( 0.4  0.6  )
   2) INITA:O,D 209  60 -1 ( 0.7  0.3  )
     4) fR<3.67532 172  40 -1 ( 0.8  0.2  )
       8) LEADA:O,D 150  20 -1 ( 0.8  0.2  ) *
       9) LEADA:A 22   7 1 ( 0.3  0.7  ) *
     5) fR>3.67532 37  10 1 ( 0.4  0.6  ) *
   3) INITA:A 449 100 1 ( 0.3  0.7  )
     6) LEADA:D 32   5 -1 ( 0.8  0.2  ) *
     7) LEADA:O,A 417  90 1 ( 0.2  0.8  )
      14) LEADA:O 255  70 1 ( 0.3  0.7  )
        28) fR<1.08765 40  10 -1 ( 0.7  0.3  ) *
        29) fR>1.08765 215  50 1 ( 0.2  0.8  ) *
      15) LEADA:A 162  10 1 ( 0.07 0.9  ) *

Node number 1: 658 observations,    complexity param=0.3
  predicted class= 1  expected loss= 0.4
     class counts:  260 398
    probabilities:  0.4 0.6
  left son=2 (209 obs) right son=3 (449 obs)
  Primary splits:
```

```
      INITA  splits as  LRL, improve=60, (33 missing)
      LEADA  splits as  RRL, improve=50, (33 missing)
      CEA    splits as  LRL, improve=20, (33 missing)
      MORALA splits as  LRL, improve=20, (33 missing)
      INTELA splits as  RRL, improve=20, (33 missing)
  Surrogate splits:
      LEADA  splits as  RRL, agree=0.7, (0 split)
      CEA    splits as  RRL, agree=0.7, (0 split)
      INTELA splits as  RRL, agree=0.7, (0 split)
      SURPA  splits as  RRL, agree=0.7, (0 split)
      TECHA  splits as  RRL, agree=0.7, (0 split)

Node number 2: 209 observations,    complexity param=0.03
  predicted class= -1  expected loss= 0.3
     class counts:  147  62
    probabilities:  0.7 0.3
  left son=4 (172 obs) right son=5 (37 obs)
  Primary splits:
      fR     < 4   to the left,  improve=9, (0 missing)
      LEADA  splits as  LRL, improve=8, (0 missing)
      arty   < 10  to the left,  improve=7, (54 missing)
      fly    < 1   to the left,  improve=7, (149 missing)
      INTELA splits as  LRL, improve=5, (0 missing)
  Surrogate splits:
      TECHA splits as  LLR, agree=0.8, (0 split)

Node number 3: 449 observations,    complexity param=0.08
  predicted class= 1  expected loss= 0.3
     class counts:  113 336
    probabilities:  0.3 0.7
  left son=6 (32 obs) right son=7 (417 obs)
  Primary splits:
      LEADA  splits as  RRL, improve=20, (33 missing)
      CEA    splits as  LRL, improve= 9, (33 missing)
      MORALA splits as  LRL, improve= 7, (33 missing)
      AEROA  splits as  RRL, improve= 7, (33 missing)
      INTELA splits as  LRL, improve= 7, (33 missing)

Node number 4: 172 observations,    complexity param=0.03
  predicted class= -1  expected loss= 0.2
     class counts:  133  39
    probabilities:  0.8 0.2
  left son=8 (150 obs) right son=9 (22 obs)
  Primary splits:
      LEADA  splits as  LRL, improve=10, (0 missing)
      CEA    splits as  LRL, improve= 5, (0 missing)
      INTELA splits as  LRL, improve= 4, (0 missing)
      TRNGA  splits as  LRL, improve= 4, (0 missing)
      INITA  splits as  R-L, improve= 2, (0 missing)
  Surrogate splits:
      fR     < 0.5 to the right, agree=0.9, (0 split)
      MOMNTA splits as  LLR, agree=0.9, (0 split)

Node number 5: 37 observations
  predicted class= 1  expected loss= 0.4
```

```
     class counts:  14 23
    probabilities:  0.4 0.6

Node number 6: 32 observations
  predicted class= -1  expected loss= 0.2
     class counts:  27  5
    probabilities:  0.8 0.2

Node number 7: 417 observations,    complexity param=0.03
  predicted class= 1  expected loss= 0.2
     class counts:   86 331
    probabilities:  0.2 0.8
  left son=14 (255 obs) right son=15 (162 obs)
  Primary splits:
      LEADA  splits as  LR-, improve=10, (33 missing)
      AEROA  splits as  RRL, improve= 7, (33 missing)
      CEA    splits as  LRL, improve= 4, (33 missing)
      INTELA splits as  LRL, improve= 4, (33 missing)
      tank   < 3   to the left,  improve= 4, (202 missing)
  Surrogate splits:
      CEA   splits as  LRL, agree=0.8, (0 split)
      fR    < 1   to the right, agree=0.7, (32 split)
      TRNGA splits as  LRL, agree=0.7, (0 split)
      AEROA splits as  RLL, agree=0.7, (0 split)
      PRIA1 splits as  RRRLL, agree=0.7, (1 split)

Node number 8: 150 observations
  predicted class= -1  expected loss= 0.2
     class counts:  126  24
    probabilities:  0.8 0.2

Node number 9: 22 observations
  predicted class= 1  expected loss= 0.3
     class counts:   7 15
    probabilities:  0.3 0.7

Node number 14: 255 observations,    complexity param=0.03
  predicted class= 1  expected loss= 0.3
     class counts:   74 181
    probabilities:  0.3 0.7
  left son=28 (40 obs) right son=29 (215 obs)
  Primary splits:
      fR     < 1   to the left,  improve=20, (1 missing)
      arty   < 1   to the left,  improve= 8, (49 missing)
      MORALA splits as  LRL, improve= 8, (16 missing)
      tank   < 4   to the left,  improve= 5, (98 missing)
      LOGSA  splits as  LRL, improve= 4, (16 missing)

Node number 15: 162 observations
  predicted class= 1  expected loss= 0.07
     class counts:   12 150
    probabilities:  0.1 0.9

Node number 28: 40 observations
  predicted class= -1  expected loss= 0.3
```

```
     class counts:  28 12
    probabilities:  0.7 0.3


Node number 29: 215 observations
  predicted class= 1  expected loss= 0.2
     class counts:   46 169
    probabilities:  0.2 0.8
```

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   8725 John J. Kingman Rd., STE 0944
   Ft. Belvoir, Virginia 22060-6218

2. Dudley Knox Library
   Naval Postgraduate School
   411 Dyer Rd.
   Monterey, California 93943-5101

3. Genelkurmay Baskanligi
   Personel Baskanligi
   Bakanliklar
   Ankara, TURKEY

4. Kara Kuvvetleri Komutanligi
   Personel Daire Baskanligi
   Bakanliklar
   Ankara, TURKEY

5. Kara Kuvvetleri Komutanligi
   Kara Kuvvetleri Kutuphanesi
   Bakanliklar
   Ankara, TURKEY

6. Kara Harp Okulu Komutanligi
   Cumhuriyet Sitesi
   Kara Harp Okulu Kutuphanesi
   Bakanliklar
   Ankara, TURKEY

7. The Helmbolds
   2625 E. Southern Ave.
   C-185
   Tempe, AZ 85282

8. Mr. Ed Vandiver
   U.S. Center for Army Analyses
   6001 Goethals Road
   Fort Belvoir, VA 22060

9. Professor Thomas W. Lucas, Code OR/Lt
   Department of Operations Research
   Naval Postgraduate School
   Monterey, CA 93943-5000

10. Professor Samuel E. Buttrey
    Department of Operations Research
    Naval Postgraduate School
    Monterey, CA 93943-5000

11. Utgm.Muzaffer Coban
    Genel Kurmay Baskaligi
    Bakanliklar,
    Ankara, Turkey