2008-06

Probability prediction of a nation's internal conflict based on instability

Wann, Shian-Kuen

Monterey California. Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

PROBABILITY PREDICTION OF A NATION'S INTERNAL CONFLICT BASED ON INSTABILITY

by

Shian-Kuen Wann

June 2008

Thesis Advisor: Kyle Lin
Second Reader: Yu-Chu Shen

Approved for public release; distribution is unlimited

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** June 2008	**3. REPORT TYPE AND DATES COVERED** Master's Thesis	
4. TITLE AND SUBTITLE Probability Prediction of a Nation's Internal Conflict Based on Instability		**5. FUNDING NUMBERS**	
6. AUTHOR(S) Shian-kuen Wann			
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 is unlimited		**12b. DISTRIBUTION CODE**	

13. ABSTRACT (maximum 200 words)

Since the end of cold war, predicting a nation state's instability has been a challenging national security issue for the United States. This thesis presents several methods to predict the conflict potential for failed nation states by comparing their social, economic, political, and military statistics with those in the past. This study uses the Brier scoring rule to evaluate the performances of these probability prediction methods. The study provides insights into situations where one method expects to outperform the others.

14. SUBJECT TERMS Probability Prediction, K-Nearest-Neighbor Algorithm, Brier Scoring Rule			**15. NUMBER OF PAGES** 65
			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** UU

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

PROBABILITY PREDICTION OF A NATION'S INTERNAL CONFLICT BASED ON INSTABILITY

Shian-Kuen Wann
Major, Army of Republic of China (Taiwan)
B.S., Virginia Military Institute, 1996

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 2008

Author: Shian-Kuen Wann

Approved by: Kyle Lin
Thesis Advisor

Yu-Chu Shen
Second Reader

James N. Eagle
Chairman, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Since the end of cold war, predicting a nation state's instability has been a challenging national security issue for the United States. This thesis presents several methods to predict the conflict potential for failed nation states by comparing their social, economic, political, and military statistics with those in the past. This study uses the Brier scoring rule to evaluate the performances of these probability prediction methods. The study provides insights into situations where one method expects to outperform the others.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

For the United States, nation states' instabilities, or internal conflicts, have accounted for a significant proportion of military operations. These instabilities, or internal conflicts, result in human tragedies, clashing of international interests, and in disturbance of global peace. To mitigate the negative effect of a nation's internal conflict, it is essential to have knowledge, prior to their occurrences, of where and when these conflicts may happen. In the last few decades, prediction of internal conflicts of nation states has been an ongoing effort.

This thesis extends a recent study conducted by LTC Robert Shearer from the Center for Army Analysis. Several modifications and extensions to Shearer's work are proposed in this thesis. First, more careful treatment to missing data and data set rescaling is given. Second, rather than projecting a nation state's various statistics for the next year, their current year statistics are utilized to directly predict the conflict potential. Third, when computing a probability prediction, a weighted average is used instead of an arithmetic average.

With the data set provided by the Center for Army Analysis, this study experiments with proposed methods and evaluates their performance by the Brier scoring rule. This study provides insights into situations where one method expects to outperform the others.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

My deepest gratitude goes first and foremost to Professor Lin, my thesis advisor, for his patience and guidance. He walked me through all the stages of writing this thesis. Without his consistent and illuminating instruction, this thesis could not have reached its present form.

Second, I express my thankfulness to LTC Shearer, Center for Army Analysis, who generously provided me with his study. Without his study and help, I would not have been able to explore other methods and contribute to this field of study.

Last, my thanks and appreciation goes to my beloved family for their support, great confidence, and loving considerations throughout these years.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

This thesis proposes several methods to predict the probability that, in the near future, a nation will develop an internal conflict. In 2007, the Center for Army Analysis (CAA) conducted a study on various statistics of a nation state as they relate to internal conflicts [1]. The importance of predicting the internal conflict of a nation is evident in a quote from this study:

> Since the end of cold war, economic dislocations, civil war, famine, and ancient ethnic and religious animosities have contributed to conflict and political instability in states extending from Haiti to the vast archipelago of Indonesia. These conflict and instabilities frequently challenge national security interests; at other times, the human rights atrocities that often accompany these dislocations offend the moral imperatives of individual states as well as the international community [2]. Increasingly, the United State's military has found itself executing a range of operations as a direct result of these conflicts. Understanding where and when these conflicts could occur is essential in developing a sound military strategic plan [1].

In this CAA study, Shearer [1] identified a vector of 13 features that correlate with whether an internal conflict will occur in a nation. These 13 features can be put into four categories as follows:

- **Economic features**: male unemployment, GDP per capita, and trade openness.
- **Military features**: conflict history.
- **Political features**: civil liberties, democracy, and political rights.

- **Social features**: adult male literacy, caloric intake, ethnic diversity, infant mortality rate, life expectancy, and religious diversity.

The method used in Shearer's study can be described briefly as follows:

Step1: Put each country-year into a 13 dimensional space based on the 13 feature values. If the country has internal conflict in that year, paint the point red; otherwise, paint it blue.

Step2: Use a weighted moving average on the 13 feature values to project a country's movement in this space in the future.

Step3: Predict, based on the projected location of a country in the future and the colors of its neighbors, whether that country will develop an internal conflict in the given year.

Shearer's method consists of extending statistical extrapolations indefinitely. This is a significant strength: for each prediction, his method allows choice of how far into the future one wants.

Shearer's [1] method, however, has room for improvement. First, in step 1, a point's color is painted based on whether, for the same year as the 13 feature values, a country has internal conflict. Yet the actual problem is to use feature values of the current year to predict conflict in the next year. Second, in step 3, the neighbors' colors are weighted equally without considering distance between neighbors and the location under consideration. This thesis explores possible extensions to Shearer's method [1] to improve prediction results.

A. RESEARCH OBJECTIVE

This thesis' objective is to develop new methods to improve the quality of predictions from Shearer's study. The most significant differences of the new methods include the following:

- In step 1, points are painted red or blue based on whether the country has internal conflict in the next year. There is no need to project the movement of a country on the 13-dimensional map. Rather than forecasting the future features, the 13 feature values from the previous years are utilized to directly predict the future's conflict outcome.
- In step 3, to predict the probability of internal conflict, the distance between the location under consideration and its neighbors are taken into account.

To compare this study's methods with those studied by Shearer [1], the same data set that Shearer used, which contained 13 macro-structural features of 155 nations observed from 1993 to 2003, is used. Further, the Brier scoring rule is utilized to evaluate each prediction method and to compare their performances.

B. LITERATURE REVIEW

Organizations and scholars studied the previous works about interstate instability or conflict. Each of these works represents a unique contribution to the development of conflict prediction or crisis early warning. According to O'Brien [2], one research conducted by the State Failure Task Force (SFTF) that had used logistic regression, neural networks, and genetic algorithm with some key feature variables associated with nation's internal instability to provide an early warning of state failures.

O'Brien [2] built on SFTF's study to forecast the conflict by presenting a macro-structural approach. O'Brien's results suggest that predictions for countries experiencing a certain level of intensity can be accomplished based on their similarities. Shearer [1] also presented a macro-structural approach to predict the conflict potential of nations. The O'Brien and Shearer works are similar. However, they used different pattern classification algorithms. O'Brien used fuzzy analysis of statistical evidence (FASE); whereas, Shearer used Nearest Centroid (NC) and K-Nearest-Neighbor (KNN) algorithms. Both results can forecast nations' levels of conflict, out to 5 years, with about 80% overall accuracy.

There are other interesting related studies on interstate instability or conflict. For example, Beck, King, and Zeng [3] presented a version of a neural network model that revealed interesting structural features of international conflict. Pevehouse [4] discovered that increased trade dependence can stimulate conflict simultaneously. Kilgour and Zagare [5] used a discrete game model to analyze a problem of limited conflict. Robst, Polachek, and Chang [6] showed how geography and trade work influence international conflict. These approaches and studies can generate forecasts that provide the strategic decision maker with good knowledge of when and where a nation will likely experience instability or conflict.

This thesis can be viewed as an extension to the works by O'Brien [2] and Shearer [1]. Further, it uses macro-structural factors to predict a nation's future conflict level. The goal is to explore other prediction methods to provide a good mechanism to support the decision maker prior to the occurrences of future conflicts.

C. THESIS ORGANIZATION

After this introductory chapter, there are three chapters remaining in this thesis. Chapter II discusses and explains this study's data set and prediction methodologies. The data set is an internal conflict data set containing 13 feature variables of 155 nations from 1993 to 2003. The proposed methodologies are designed to satisfy thesis objective and to improve prediction accuracy. Each methodology contains several alternatives that will be discussed in detail.

Chapter III analyzes the results of prediction from each prediction method. Like most statistical analysis, each of this study's designed methods will have both a training set and a test set to validate the prediction values. Because these methods offer a probability prediction, the Brier score rule is used to compare different prediction methods. The goal is to discover if any of this study's methods can provide a remarkable improvement over the existing Shearer's method [1].

Chapter IV concludes this thesis, discusses findings, and provides ideas for further study.

THIS PAGE INTENTIONALLY LEFT BLANK

II. DATA AND METHODOLOGY

This section presents the data and methodology used to predict a nation's internal conflict. Section A presents the data set. Section B describes the variables from the data set. Section C introduces the prediction methodology.

A. DATA SETS

The data set contains 155 nations observed from 1993 to 2003 (11 years). For each year, each nation contains one conflict indicator and thirteen feature variables. These feature variables are used to identify patterns of nation state instability or conflict. These include three political features (civil liberties, democracy, and political rights); three economic features (male unemployment, GDP per capita, and trade openness); six social features (adult male literacy, caloric intake, ethnic diversity, infant mortality rate, lifer expectancy, and religious diversity); and one military feature (conflict history). Each nation-year pair is plotted using the 13 feature variables as a point in 13-dimensional space. Thus, 155 × 11 = 1705 data points.

In the raw data set, each conflict indicator indicates the nation's intensity level of conflict in that year. There are four levels of intensity: latent crisis, crisis, severe crisis, and war. Latent crisis is in level 1; crisis is in level 2; and so on. The feature variables were obtained from multiple studies by the CAA in the late 1990s and 2000s; these variables were identified as key macro-structural features that affect nation state stability in Shearer [1].

This thesis uses this same international conflict data set. For further analysis, the data is divided into two different groups: independent variables and dependent variables. The next section discusses these two groups of variables in detail.

B. VARIABLES DESCRIPTION

1. Independent Variables

The independent variables include the 13 feature variables. These feature variables were collected from different sources and measured in different scale. Some are continuous variables and others are discrete variables. Table 1 describes definitions and sources of the 13 feature variables. In the table, it is observed that, in the data set, some features have missing values. This is especially significant in the Male Unemployment and Adult Literacy.

Category	Feature	Definition	Source[1]	Percent Missing
Political	Civil Liberty	A measure of the freedom of a country's people to develop views, institutions, and personal autonomy apart from the state.	Freedom House	0.41
	Democracy	A measure of the degree of democracy.	Polity IV Project	2.50
	Political Rights	A measure of the rights to participate meaningfully in the political process.	Freedom House	0.41
Economic	Male Unemployment	The percentage of the male labor force that is unemployed.	World Bank	79.40
	GDP	The annual gross domestic product per person measured in	World Bank	4.60

Category	Feature	Definition	Source[1]	Percent Missing
		constant 1998 U.S. dollars.		
	Trade Openness	The ratio of a country's total imports and exports to GDP.	World Bank	4.10
Social	Adult Male Literacy	The percentage of males who can read or write -- ages 15 and above.	World Bank	28.20
	Caloric Intake	An estimate of the average number of calories consumed per person per day.	FAOUN[2]	1.90
	Ethnic Diversity	The population of the largest ethnic group in the country as a percentage of the total population.	CIA WFB & CIFPP[3]	0.65
	Infant Mortality Rate	The number of deaths of children under 1 year of age per 1,000 live births.	U.S. Bureau of the Census	0.00
	Life Expectancy	The average life expectancy (males and females combined)	U.S. Bureau of the Census	0.00
	Religious Diversity	The population of the largest religious group in the country as a percentage of the total population.	CIA WFB & CIFPP	2.50
Military	Conflict History	The percentage of time (in years) spent in a state of conflict (war or severe crisis).	HIIK[4]	0.00

Table 1. Definitions and Sources of Independent Variables.

1. All data were collected by the Center for Army Analysis to produce Shearer [1].
2. FAOUN = Food and Agriculture Organization of the United Nations.
3. CIA WFB&CIFPP = CIA World Fact Book and Country Indicators of Foreign Policy Project.
4. HIIK = Heidelberg Institute of Conflict Research Male.

Figure 1 shows the histogram of the 13 feature variables. In Figure 1, it is observed that some variable distributions looks normal (i.e., Calories Intake and Civil Rights); some variables distribute in a wide range of values (i.e., Democracy and Political Rights); and some variable distributions are skewed (i.e., Religion Diversion, Life Expectance, etc.). The distributions of all 13 variables are quite distinctive. Since these variables are scaled in different measurements, these variables could be rescaled to the same measurement for further analysis. Section C discusses three such methods in detail.

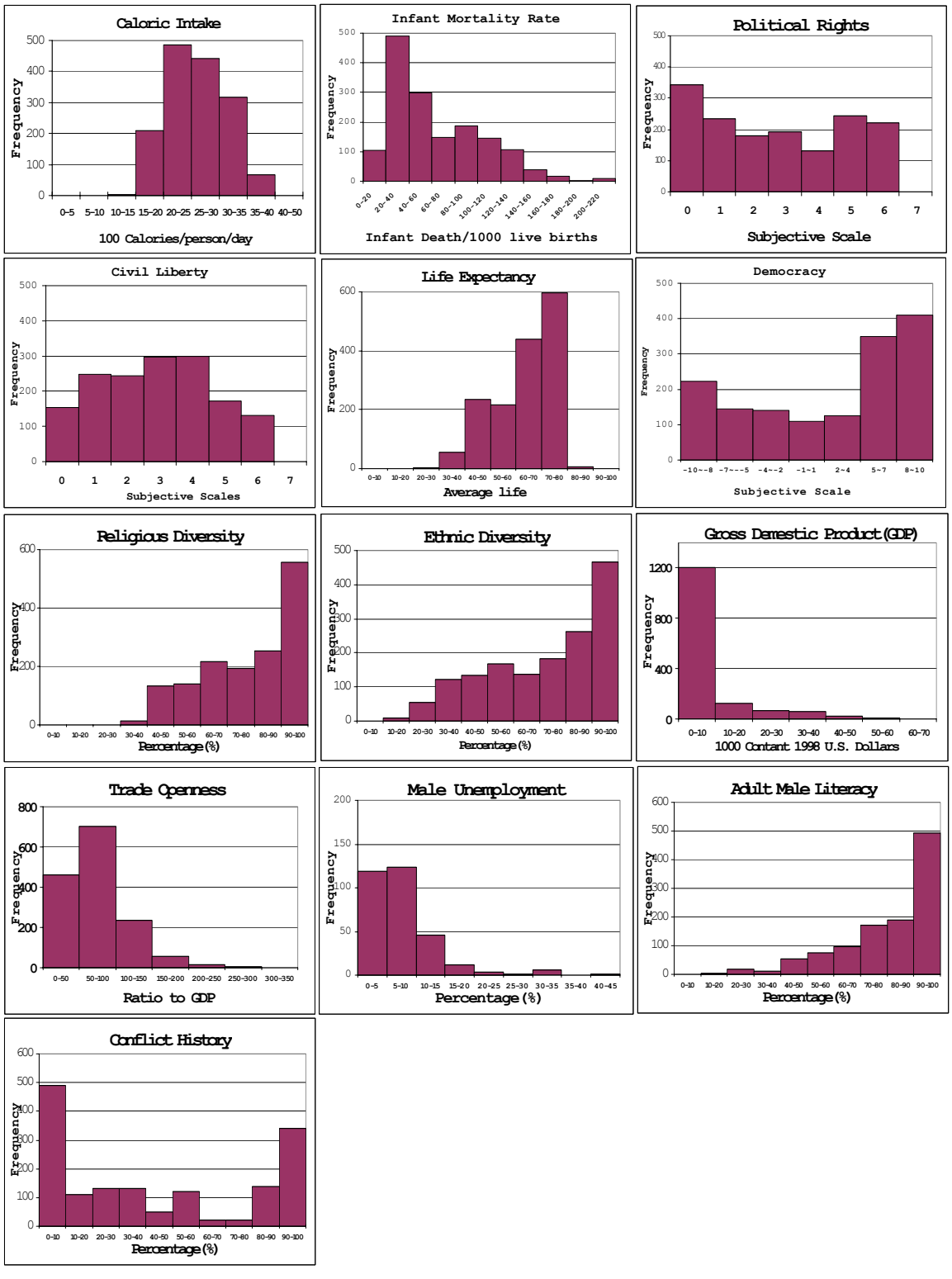

Figure 1. Variable Distribution.

Table 2 describes the data summary of the 13 feature variables. This table summarizes minimum, maximum, mean and standard deviations of the feature variables -- excluding the missing value in the data set. In this table, it is noticed that replacing a missing value by the mean of feature variable can cause a large error if this missing value is actually an extreme value. This is especially evident for those feature variables with high proportions of missing values.

Category	Feature	Min	Max	Mean	StDev
Political	Civil Liberty	1	7	3.87	1.76
	Democracy	-10	10	2.85	6.61
	Political Rights	1	7	3.72	1.16
Economic	Male Unemployment	0%	42%	7.89%	5.86%
	GDP	100	59000	6007.5	10348
	Trade Openness	0	290	78.7	40.70
Social	Adult Male Literacy	19%	100%	81.7%	18%
	Caloric Intake	1500	3800	2648.3	519.80
	Ethnic Diversity	17%	100%	73.1%	22.5%
	Infant Mortality Rate	0	200	48.4	40.60
	Life Expectancy	10	81	64.2	12.20
	Religious Diversity	30	100	79.4	17.20
Military	Conflict History	0%	92.9%	36%	37.00

Table 2. Data Summary of Independent Variables.

2. Dependent Variable

The dependent variable used is the level, or intensity, of conflict experienced by a country in a given year. This variable was collected from Heidelberg Institute of International Conflict Research (HIIK) by the CAA. According to HIIK's definition [7], conflicts are defined as the clashing of interests on national values and issues of some duration and magnitude between at least two parties (states, groups of states, organizations or organized groups). These conflicts include territory, secession, decolonization, autonomy, system/ideology, national power, regional predominance, international power, resources, and others. There are four levels of intensity: latent conflict, crisis, severe crisis, and war. In the CAA study, Shearer [1] classifies the four levels of intensity into two categories: peace and conflict. This thesis adopts the same classification. Table 3 summarizes the definition of conflict classification of the four levels given in Shearer [1]. The level of intensities is defined by HIIK [7] as follows:

Latent Conflict:

> A latent conflict is a positional difference over definable values of national meaning – only if demands are articulated by one of the parties and perceived by the other as such.

Crisis:

> A crisis is a tense situation in which at least one of the parties uses violence forces repeatedly in sporadic incidents.

Severe Crisis:

A severe crisis is defined as a state of high tension between the parties: either they threaten to resort to the use of force or they actually use physical, or military, force in an organized way.

War:

A war is a violent conflict in which violent force is used with certain continuity in an organized and systematic way. Depending on the situation, the conflict parties exercise extensive measures. The extent of destruction is massive and of long duration.

Table 3 describes how the four levels of intensity are further put into two levels of conflict.

HIIK's Level of Intensity	Name of Intensity	CAA's Level of Conflict	Classification
1	Latent Conflict	0	Peace
2	Crisis		
3	Severe Crisis	1	Conflict
4	War		

Table 3. Conflict Classification.

C. METHODOLOGY

This thesis asserts that in Shearer's [1] study of conflict pattern prediction there is room for improvement. The first issue is that Shearer [1] colored a point (13-dimensional coordinated by independent variables) based upon whether a nation had internal conflict in the same year. In predicting the level of conflict, it is advantageous to use feature values of the current year to predict conflict in the next year. The second issue in Shearer's [1] method is that when predicting the level of

an unknown conflict, the colors of its neighbors are weighted equally without considering the distance between each neighbor and the location under consideration. In this thesis, those nearest neighbors are weighed either by their distances or by their ranks. This study's method can be summarized in four steps as follows:

Step 1: Replace the missing feature data.

Step 2: Rescale the feature variables.

Step 3: Rearrange the data set.

Step 4: Predict the conflict probability.

These four steps are explained below in detail.

1. Replace the Missing Feature Data

In the histograms of independent feature variables distribution in Section A, it is observed that, for some features, there is much missing data -- especially adult male literacy and male unemployment. In CAA's study, the missing data is replaced by the mean of the overall feature variable. In this thesis, the following rules are used to replace the missing data:

a. For a country that misses all values for one feature variable (i.e., Unemployment in 1993-1997 for Burma), for each year, replace with yearly mean of the entire sample of that variable across all other countries in the same year.

b. For a country that misses one or more (but not all) data points, replace the missing data with the value that was last available for the same country. For example, unemployment was only available for Cameron in 1996; thus, missing value (1997-2003) was replaced with the value from 1996.

2. Rescale the Feature Variables

In the raw data set, the 13 feature variables are measured in different scales. To calculate the Euclidean distance between two points, each feature must be scaled into the same measurement. This thesis introduces three different scaling methods: normalization, standardization, and principle components. Each method is explained below.

a. Normalization

Normalizing a feature (variable) refers to scaling by the minimum and range of the variable, which makes the variable score between 0 and 1. Let $x_{i,j,k}$ be *ith* feature variable which is created for *jth* nation state for *kth* year. By normalization, each indicator was scaled into a new variable *yi,j,k* between 0 and 1, where

$$y_{i,j,k} = \frac{x_{i,j,k} - \min_i x_{i,j,k}}{\max_i x_{i,j,k} - \min_i x_{i,j,k}} .$$

$$\forall\, i,j,k$$

b. Standardization

To standardize a variable is to subtract a variable with its mean and then divided this difference by the variable's standard deviation; thus, the standardized variable has mean 0 and standard deviation 1. Let $x_{i,j,k}$ be the *ith* feature variable which was created for *jth* nation state for *kth* past year. By standardization, each variable is scaled into a new variable with mean 0 and standard deviation 1, where

$$y_{i,j,k} = \frac{x_{i,j,k} - \underset{i}{\text{mean}}\ x_{i,j,k}}{\underset{i}{\text{std}}\ x_{i,j,k}} .$$

$$\forall\ i, j, k$$

c. Principle Component Analysis (PCA)

A PCA is a non-parametric method used to identify the correlation among the original variables and to reduce the dimension of data set that still captures the essence of the data. In PCA, extraction is performed from a set of *m* variables to a set of *n* factors (*m* > *n*). By definition, these factors are inferred from the correlations among the *m* variables and each factor is estimated as a weighted sum of the *m* variables. Interested readers can refer to Montgomery, Peck and Vining [8] for a discussion on PCA.

To proceed with PCA, the missing values are replaced with additional steps after normalizing (or standardizing) the data set.

Step 1: For feature variables that have more than 10% missing.

- Generate a binary indicator for each one of them, where the indicator equals 1 if the variable is missing and, otherwise, 0.
- Replace the missing feature values with yearly average of the entire sample.

Step 2: For feature variables with less than 10% missing, replace the value with the missing code. When running PCA, the statistical software will automatically drop those missing values.

Step 3: For a country's missing values in the predict year, replace it with the first existing data from the previous year.

Step 4: Run the principle component model in S-plus.

After missing value is replaced, the new data set consists of 21 variables (13 original variables plus 8 binary variables). From a set of 21 variables, extraction is used to obtain a set of underlying variables with fewer dimensions.

This study would have liked to see the correlation matrix (Table 4) for a new data set. In the correlation matrix, the variables are ALL correlated, but there are only a few variables with significant correlation. Examples are Calories and Infant Mortality Rate (IMR) and, also, IMR and Life Expectance. PCA is useful when there are strong correlations among feature variables. In Table 4, however, it does not appear to be the case.

	Calories	*IMR*	*Pol Rights*	*CivLibers*	*LifeExp*	*Democ*	*ReliDiv*	*EthnicDiv*	*GDP*	*TradeOp*	*Unemp*	*AdultLit*	*TimeConf*	*MissUemp*	*MissAdultLit*	*MisCal*	*MisDemoc*	*MissReliDiv*	*MissEthnicDiv*	*MissGDP*	*MissTradeOp*
Calories	1.0	**-0.7**	-0.4	-0.5	**0.7**	0.3	0.3	0.2	0.6	0.2	0.0	0.4	-0.3	0.1	0.1	0.0	0.1	0.1	-0.1	0.0	0.0
IMR	**-0.7**	1.0	0.5	0.5	**-0.9**	-0.4	-0.3	-0.3	-0.5	-0.2	0.0	-0.5	0.3	-0.1	0.0	0.0	-0.1	-0.1	0.1	0.0	0.0
PolRights	-0.4	0.5	1.0	**0.9**	-0.5	**-0.9**	-0.1	-0.2	-0.5	-0.1	0.0	-0.2	0.4	-0.1	0.0	0.1	0.0	-0.1	0.1	0.3	0.1
CivLiberty	-0.5	0.5	**0.9**	1.0	-0.5	**-0.8**	0.0	-0.2	-0.5	-0.2	0.0	-0.2	0.5	0.0	-0.1	0.1	-0.1	-0.1	0.0	0.3	0.1
LifeExp	**0.7**	**-0.9**	-0.5	-0.5	1.0	0.4	0.4	0.4	0.5	0.2	0.0	0.5	-0.2	0.1	0.0	0.0	0.1	0.1	-0.1	0.0	0.0
Democ	0.3	-0.4	**-0.9**	**-0.8**	0.4	1.0	0.1	0.2	0.3	0.0	0.0	0.2	-0.3	0.0	0.0	-0.2	0.0	0.1	-0.1	-0.2	-0.1
ReliDiv	0.3	-0.3	-0.1	0.0	0.4	0.1	1.0	0.3	0.2	-0.1	0.0	0.1	0.1	0.1	0.0	0.0	0.1	0.0	-0.1	0.1	-0.1
EthnicDiv	0.2	-0.3	-0.2	-0.2	0.4	0.2	0.3	1.0	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1
GDP	0.6	-0.5	-0.5	-0.5	0.5	0.3	0.2	0.2	1.0	0.1	0.0	0.1	-0.3	0.1	0.2	0.1	0.2	-0.1	0.0	0.0	0.0
TradeOp	0.2	-0.2	-0.1	-0.2	0.2	0.0	-0.1	0.0	0.1	1.0	0.1	0.2	-0.4	0.0	0.0	0.0	0.3	0.2	0.0	-0.1	0.0
Unemp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	1.0	0.0	0.0	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.1
AdultLit	0.4	-0.5	-0.2	-0.2	0.5	0.2	0.1	0.1	0.1	0.2	0.0	1.0	-0.2	0.1	0.0	0.0	0.0	0.2	-0.2	0.1	0.0
TimeConf	-0.3	0.3	0.4	0.5	-0.2	-0.3	0.1	0.0	-0.3	-0.4	0.0	-0.2	1.0	0.0	0.2	-0.1	-0.1	-0.1	0.1	0.1	0.0
MissUemp	0.1	-0.1	-0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.3	0.1	0.0	1.0	0.0	0.0	0.1	0.0	-0.1	0.0	0.1
MissAdultLit	0.1	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	0.2	0.0	0.1	0.0	0.2	0.0	1.0	0.0	0.0	-0.1	0.0	0.1	0.0
MisCal	0.0	0.0	0.1	0.1	0.0	-0.2	0.0	0.0	0.1	0.0	0.0	0.0	-0.1	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.2
MisDemoc	0.1	-0.1	0.0	-0.1	0.1	0.0	0.1	0.0	0.2	0.3	0.0	0.0	-0.1	0.1	0.0	0.0	1.0	0.0	0.0	0.2	0.2
MissReliDiv	0.1	-0.1	-0.1	-0.1	0.1	0.1	0.0	0.0	-0.1	0.2	0.0	0.2	-0.1	0.0	-0.1	0.0	0.0	1.0	0.0	0.0	0.0
MissEthnicDiv	-0.1	0.1	0.1	0.0	-0.1	-0.1	-0.1	0.0	0.0	0.0	0.0	-0.2	0.1	-0.1	0.0	0.0	0.0	0.0	1.0	0.0	0.0
MissGDP	0.0	0.0	0.3	0.3	0.0	-0.2	0.1	0.1	0.0	-0.1	0.0	0.1	0.1	0.0	0.1	0.0	0.2	0.0	0.0	1.0	0.5
MissTradeOp	0.0	0.0	0.1	0.1	0.0	-0.1	-0.1	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.2	0.2	0.0	0.0	0.5	1.0

Table 4. Correlation Matrix. The reds indicate that the two variables are highly correlated (≥0.7).

Next, the principal components need to be extracted. S-plus is used to extract 21 components. This will involve solving 21 equations with 21 unknowns. The variance in the correlation matrix is transformed into 21 eigenvalues. Each eigenvalue represents the variance that had been captured by one component. Each component is a linear combination of the 21 variables. Further, each principal component can be viewed as 21-dimensional space where each dimension is perpendicular to each other dimension. For the conflict data, the importance of components is summarized in Table 5.

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
Standard deviation	2.2654297	1.4652305	1.26751906	1.19777504	1.14205619	1.07280493	1.03954849	1.01622953	0.98247768	0.90558493
Proportion of Variance	0.2443891	0.1022334	0.07650498	0.06831738	0.06210916	0.05480526	0.05146005	0.04917726	0.04596488	0.03905162
Cumulative Proportion	0.2443891	0.3466225	0.42312747	0.49144486	0.55355402	0.60835927	0.65981932	0.70899658	0.75496146	0.79401308

	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15	Comp.16	Comp.17	Comp.18	Comp.19	Comp.20
Standard deviation	0.89313246	0.82793736	0.79579748	0.71600737	0.70135554	0.60536109	0.57675691	0.51053986	0.334315582	0.271984790
Proportion of Variance	0.03798503	0.03264192	0.03015684	0.02441269	0.02342379	0.01745057	0.01584041	0.01241195	0.005322234	0.003522654
Cumulative Proportion	0.83199811	0.86464003	0.89479687	0.91920956	0.94263335	0.96008393	0.97592433	0.98833628	0.993658516	0.997181170

	Comp.21
Standard deviation	0.24330112
Proportion of Variance	0.00281883
Cumulative Proportion	1.00000000

Table 5. Importance of Components.

After extracting from the original 21 features, there are 21 important components, which contain 13 feature variables and 8 binary variables.

Thus far in this study, 21 correlated variables have been mapped to 21 uncorrelated components by linear transformation. A decision is needed to determine how many components are required. A Rule of thumb is that there is a need for components that capture at least 90% of variance,

which, in this study's case, leaves 13 components. To decide on the number of components to retain, another device, the scree plot (Figure 2), is used. The plot provides a visual aid to decide what components are necessary to retain.

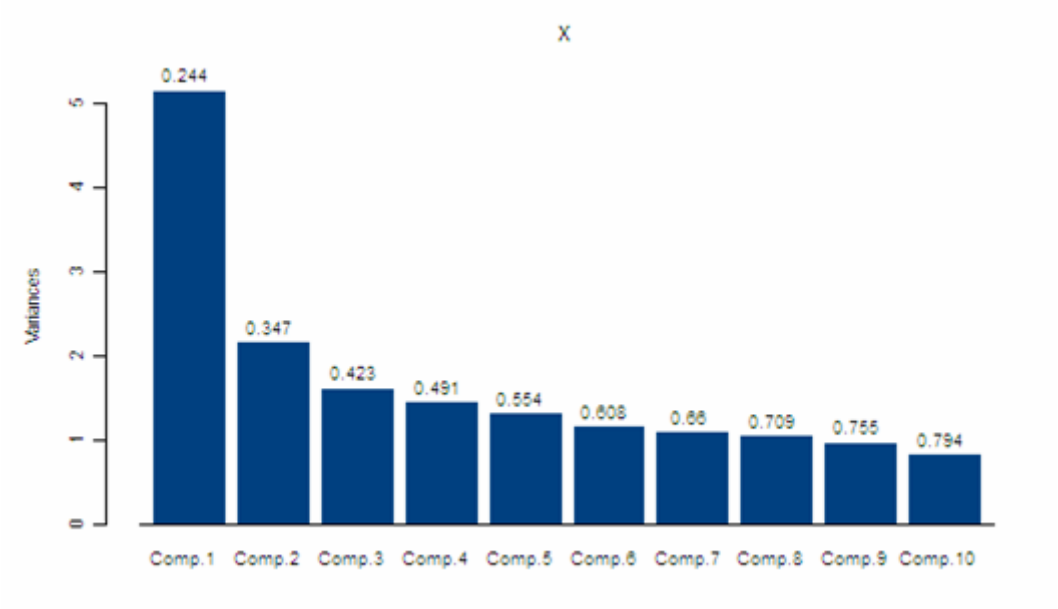

Figure 2. Scree Plot of principal components. Each number on the top of bar in the plot indicates the proportion of variance. Here, there are only 10 components to view, but this plot gives the same information as the important components indicated above.

Next, the first three components are used to plot the conflict data (Figure 3). The plot seems to be clustered for both peace and conflict data. The unknown is how it would look in 13-dimensional. Although there is not a high correlation among feature variables, the conflict cluster plot appears to indicate that it is reasonable to use these principal components as the new data set to predict the future conflict level.

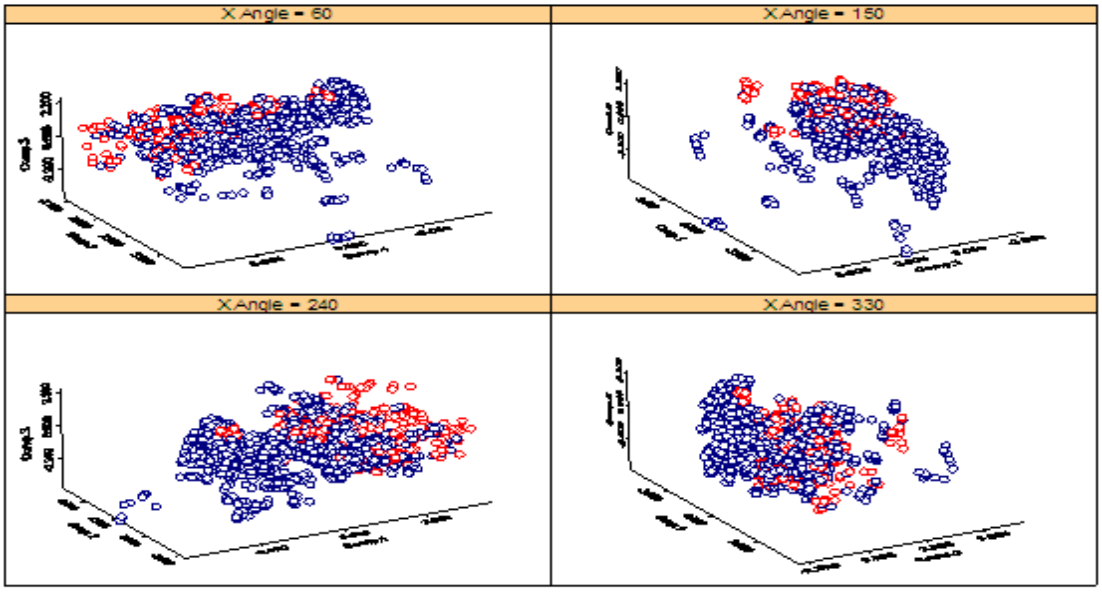

Figure 3. Conflict Cluster Maps. Red indicates conflict; blue indicates peace.

3. Rearrange the Raw Data Set

In the original raw data set, the 13 feature values of a country are used to predict whether that country will develop an internal conflict in the same year. To illustrate the idea, figures (Figures 4-10) are plotted with two feature variables: Political Right and Infant Mortality Rate. The location of each point is determined by the values of these two feature variables from the year written inside the point. The red point indicates conflict; whereas, the blue point indicates peace. This is based on the conflict-peace status of the year written outside the point. In those figures (Figures 4-10), the numbers inside the point and outside the point are identical.

Since the internal conflict and the feature variable are taken from the same year in the raw data set, there is a need to project the future feature variables from the

past prior to predicting the future conflict level. For example, in Figure 4, if one wanted to predict the conflict level in Chile in 2000, then it is necessary to project the feature values for Chile in 2000 based on data from 1995 to 1999. In CAA's study, they use a statistical extrapolation -- Weight Moving Average -- to project the future feature variables. Because the future features are projected from the past, this method can be used to predict the future as far as it is wanted.

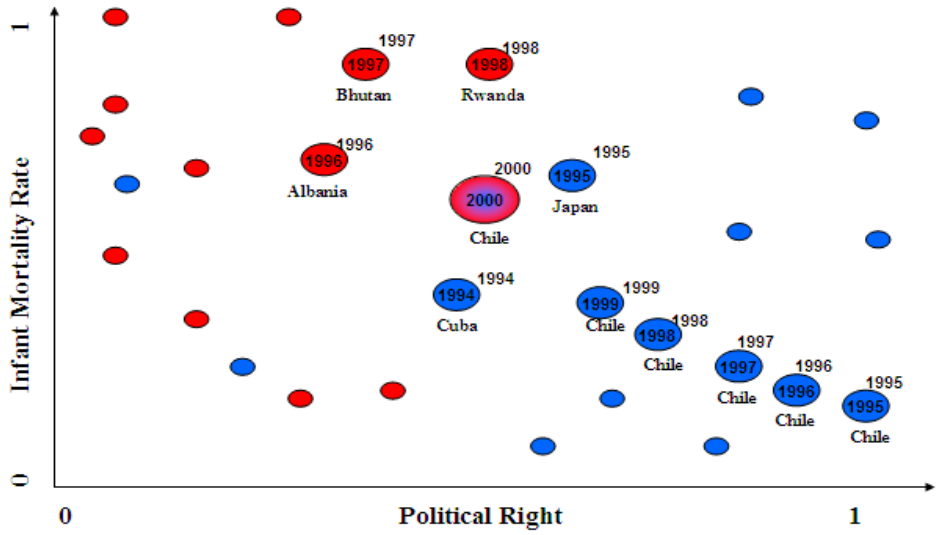

Figure 4. Prediction from the raw data set. The two feature variables (Political Right and Infant Mortality Rate) had been scaled between 0 and 1.

What is needed is to use feature variables of the current year to predict conflict in the next year. This can be accomplished directly without projecting a country's future feature values. To do that, the data set must be rearranged. The idea is to paint the point with the color based on the conflict-peace status from one year later. Figure 5 conceptually illustrates how the one-year-out

prediction looks two dimensionally. The same concept can be applied where there is a desire to predict a country's conflict-peace status for two and three years later (Figures 6 and 7).

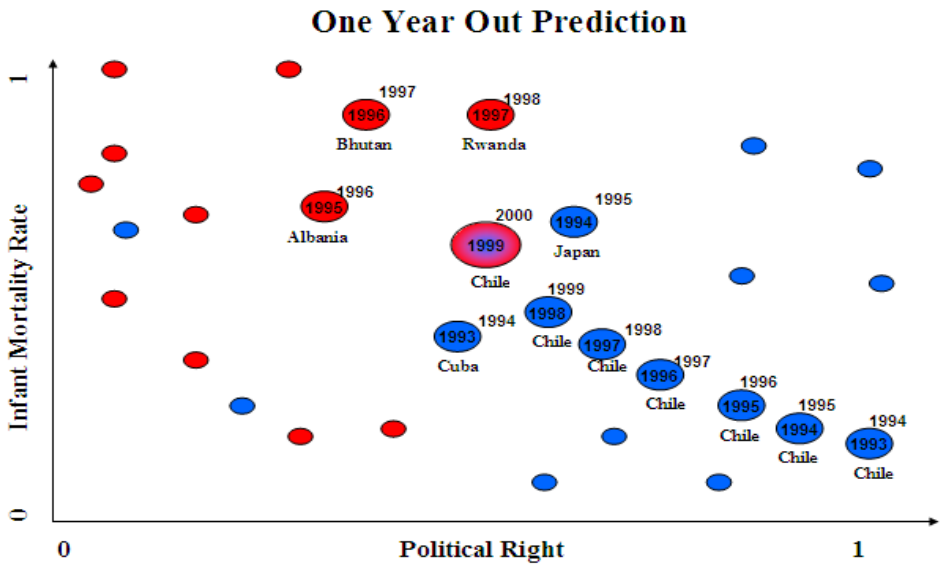

Figure 5. One-Year-Out-Prediction. The points are painted with the color based on the conflict-peace status from one year later.

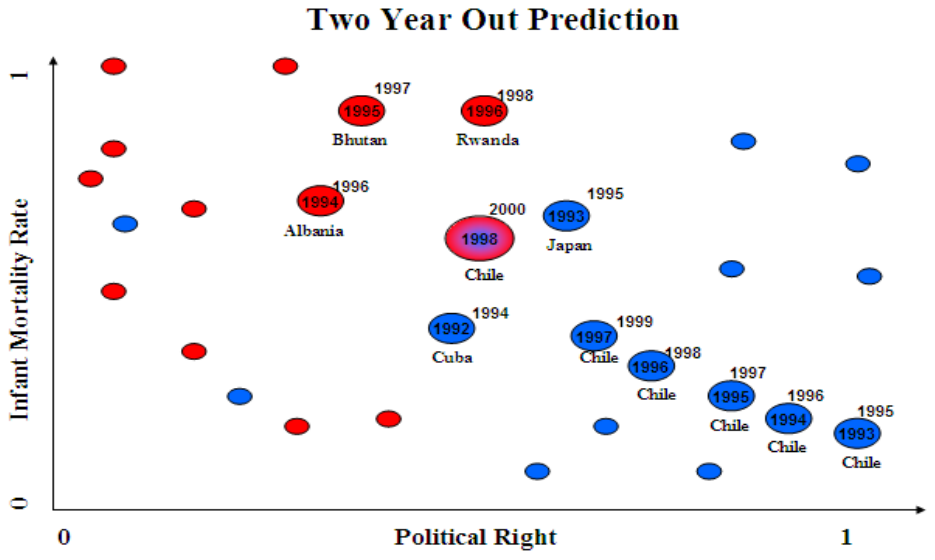

Figure 6. Two-Year-Out-Prediction. The points are painted with the color based on the conflict-peace status from two years later.

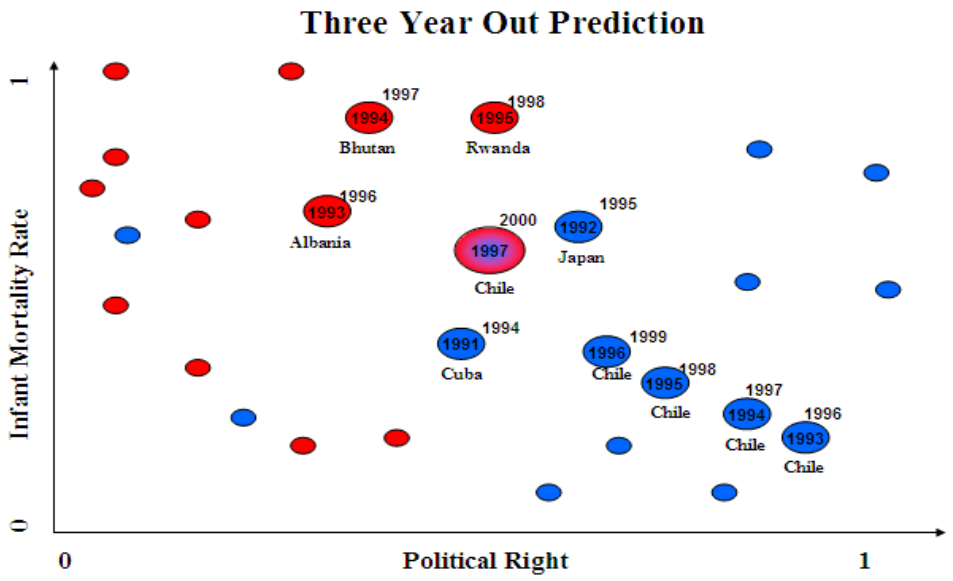

Figure 7. Three-Year-Out-Prediction. The points are painted with the color based on the conflict-peace status from three years later.

By using this data set arrangement, there is no need for a statistical extrapolation to predict the future feature values. Due to the elimination of the step of statistical extrapolation, this method is easy to conduct and efficient to obtain the conflict prediction.

4. Predict the Conflict Probability

This study uses K-Nearest-Neighbor (KNN) algorithm to predict the probability of future conflict for each nation. KNN is used to classify a future point according to its Euclidean distance from all other past points in the 13-dimensional spaces. KNN classifies the future point as a function of the *n* closest past point of one class (peace or conflict). In Shearer's [1] study, he colored the future point by equally weighting k closest past points, where k is the number of closest points to be chosen. For instance,

in Figure 8, the k is set to k=5 (five closest neighbors). Thus, three out of five points are red (conflict) and the chance to be red for that future point in 2000 is

$$\frac{(0)+(1)+(1)+(1)+(0)}{5} = 0.6.$$

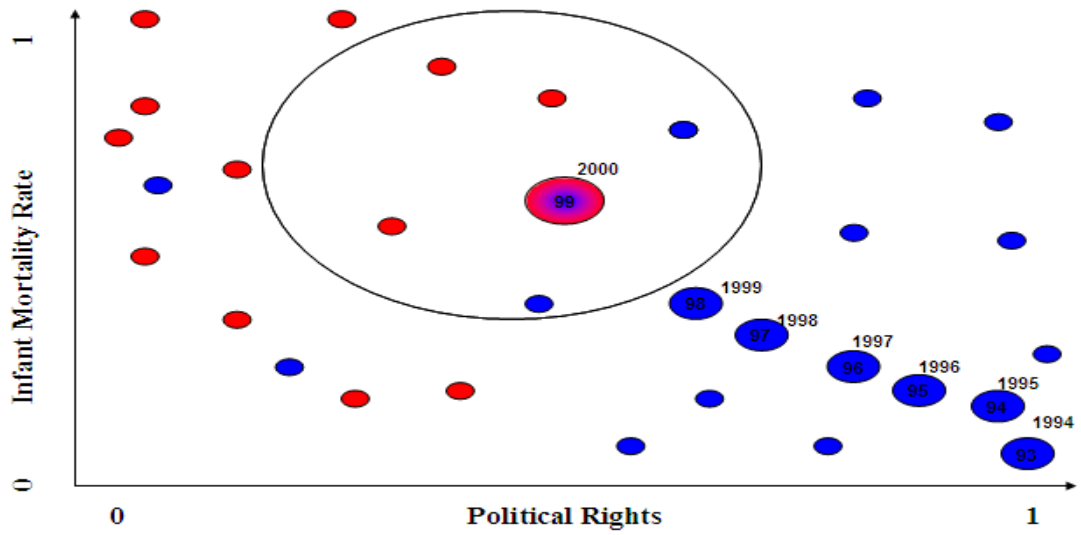

Figure 8. Predict conflict potential by equally weighting the closest neighbors.

The general mathematical form of this classification rule can be expressed by

$$P(\text{Conflict}) = \frac{\sum_{i=1}^{k} C_i}{k}.$$

$$C_i = 0,1$$

C is the level of conflict of ith's nearest neighbor (0 indicates peace, where 1 indicated conflict).

k = number of the nearest neighbors.

In this thesis, two variations of the KNN method are studied to predict the conflict probability. In the first variation, a country's future status is predicted by a weighted average over its neighbors with each neighbor's weight proportional to the inverse of its distance. For instance, in Figure 9, k=5 is set to predict the conflict potential in 2000. The status of the closet neighbors (1 or 0) is summed and weighted by the inverse of distance and divided by the totaled weights. Therefore the chance to be red for that future point in 2000 is

$$\frac{(0)\times(1/6)+(1)\times(1/4)+(1)\times(1/8)+(1)\times(1/7)+(0)\times(1/5)}{(1/6)+(1/4)+(1/8)+(1/7)+(1/5)} = 0.59.$$

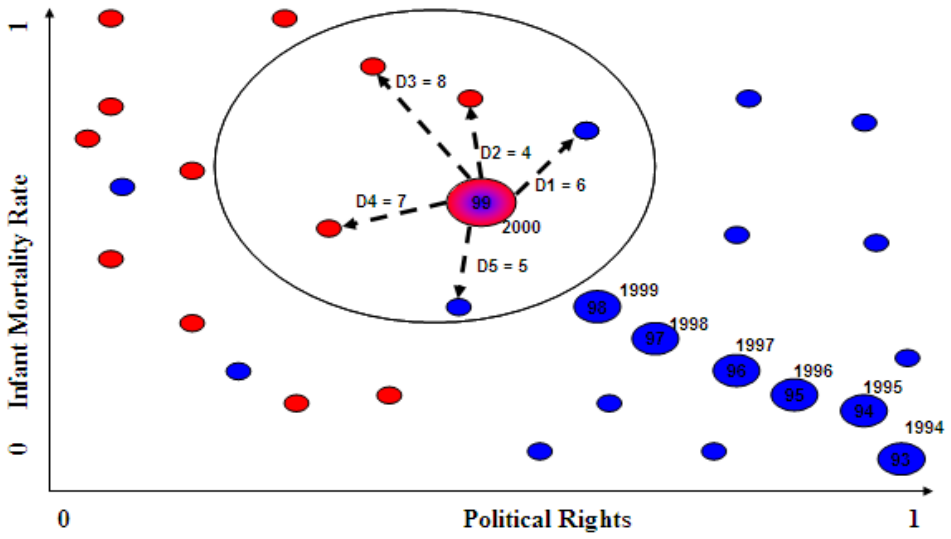

Figure 9. Predict conflict potential by the closest neighbors weighted by the inverse of distance. D# refers to the distance between point # and the predicted point.

The general mathematical form of this classification rule, the *k* nearest neighbors weighted by inverse of *distance*, can be expressed by

$$W_i = \frac{1}{D_i}$$

$$C_i = 0,1 \qquad .$$

$$P(\text{Conflict}) = \frac{\sum_{i=1}^{k}(W_i \times C_i)}{\sum_{i=1}^{k} W_i}$$

D_i is the Euclidean distance between the $i^{th's}$ nearest neighbor and the future feature vector.

W_i is the weight of the ith nearest neighbor.

In the second variation of the KNN method, a country's future status is predicted by a weighted average over its neighbors with each neighbor's weight proportional to the inverse of its rank. The rank is determined by the distance: the closer the distance, the lower the rank. For instance, in Figure 10, k=5 is set to the rank of the closest past points, from 1 to 5. 1 is the closest point and 5 is the farthest point. To predict the conflict potential in 2000, the status of the closet neighbors (1 or 0) is summed and weighted by the inverse of rank and divide by the totaled weights. Thus, the chance to be red for that future point in 2000 is

$$\frac{(0)\times(1/3)+(1)\times(1/1)+(1)\times(1/5)+(1)\times(1/4)+(0)\times(1/2)}{(1/1)+(1/2)+(1/3)+(1/4)+(1/5)} = 0.64 .$$

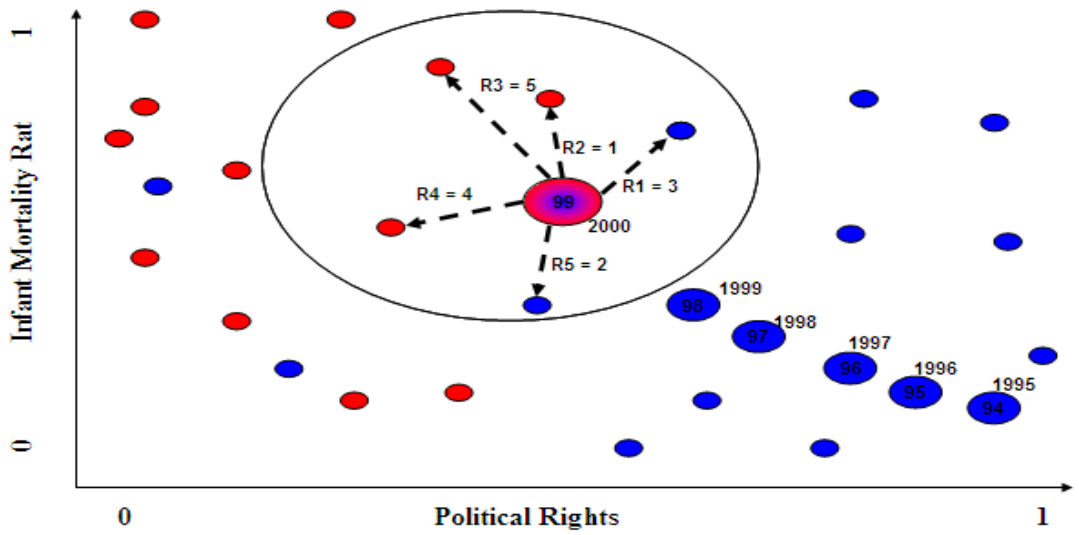

Figure 10. Predict conflict potential by the closest neighbors weighted by the inverse of rank. R# refers to the rank between point # and the predicted point.

In this classification rule, the *k* nearest neighbors are weighted by inverse of rank. The general mathematical form can be expressed by

$$W_i = \frac{1}{R_i}$$

$$C_i = 0,1 \qquad .$$

$$P(\text{Conflict}) = \frac{\sum_{i=1}^{k}(W_i \times C_i)}{\sum_{i=1}^{k} W_i}$$

Ri is the rank of the ith nearest neighbor among k's nearest neighbor based upon the distance with the future feature vector.

Wi is the weight of the ith's nearest neighbor.

THIS PAGE INTENTIONALLY LEFT BLANK

III. ANALYSIS

Based on this study's proposed methodology, there are three different ways to rescale the raw data and three KNN variations to classify the future conflict potential. This gives a total of nine methods (see Table 6).

Method	**Classification weighted Equally**	**Classification weighted by 1/Distance**	**Classification weighted by 1/Rank**
Normalized Data	NE Method	ND Method	NR Method
Standardized Data	SE Method	SD Method	SR Method
Principle Components	PE Method	PD Method	PR Method

Table 6. Method table. The names of methods are the combination of the bold letters of methodology

Due to the availability of the data set, this study will conduct conflict predictions in 155 nations up to three years. The predicted results will be evaluated by the Brier scoring rule. The data set used to validate predicted results is from 1998 to 2003. This is, then, used to determine which method performs the best according to the Brier scoring rule.

The rest of this chapter is organized as follows. Section A discusses the results of one-year-out predictions. The data up to the current year is used to predict the conflict probability of the next year. Section B conducts a comparison of different prediction methods in one-year-out prediction. Section C discusses the results in two- and three-year-out predictions. Finally, the result of

overall performances for these three predictions and the determination of which method gives the best prediction result is discussed.

A. ONE-YEAR-OUT PREDICTION

In one-year-out prediction, this study predicts the conflict potential of next year based on the data available in the current year. The data set is divided into both a training set and a test set, depending on the predicted year. The training set is the data set before the predicted year and is used to predict the conflict potential in the predicted year. The test set is the data set used to validate the predicted results. For example, to predict the conflict potential in 1998, the data set from 1994 to 1997 is used as the training set; the data in 1998 is used as the test set. The conflict occurrences in 1998 is either 1 (conflict) or 0 (peace). The predicted results in 1998, however, are measured in term of probability. The same idea applies in predicting the conflict potential in 1999, 2000, and so on.

In section II.C, the KNN algorithm variations were discussed. These variations are used to predict the conflict potential in the future. The k-value of KNN refers to the number of nearest neighbors in the past. These neighbors' distances are measured by Euclidean distance from the predicted point. This thesis varies k for 11 different values: 1, 2… 10, and 15. In each prediction method, each k-value offers a set of probability assessments of 155 nation's future conflict potential.

To evaluate a probability prediction when there are two possible outcomes, there are two commonly used proper

scoring rules: the Brier scoring rule and the logarithm scoring rule [9]. With this study's prediction methods, as well as with those in the Shearer's [1] report, sometimes a conflict probability of 0 is predicted. In these instances, using the logarithm scoring rule will result in a score of negative infinity. For that reason, this study chooses the Brier scoring rule. The Brier scoring rule is defined as follows:

$$\text{Brier score} = \begin{cases} (1-P(\text{Conflict}))^2 & \text{if conflict} \\ P(\text{Conflict})^2 & \text{if peace} \end{cases} .$$

Since Brier score is a penalty score, a lower score indicates a better prediction. This study wants to use the Brier score rule to conduct two validations. First, to determine which prediction method provides the lowest score (highest accuracy). Second, from the selected method, to determine which k-value of KNN variations provides the lowest score the prediction of 155 nations' future conflict potential is assessed in all methods.

B. COMPARISON OF DIFFERENT PREDICTION METHODS

Figure 11 gives the results of one-year-out prediction from 1998 to 2003. In the plot, each curve represents the average Brier score over 155 nations between 1998 and 2003 using one prediction method.

Recall that this study proposes nine prediction methods (Table 6). In addition to the method in Shearer [1], there are 10 curves in Figure 11. Shearer's [1] score pattern starts from the lowest score at k=1 and, then, as k

increases, the score increases dramatically. The result shows that the more neighbors set to predict the conflict potential, the less accurate the result. The patterns of this study's proposed methods start from higher scores at k=1, decreases at k=2, and, then, jumps up and down irregularly.

From Figure 11, there are two observations. First, among the three rescaling methods, the best methods based on the Brier score appear to be those using normalized data (0.071), followed by those using standardized (0.077) data, and, then, by those using principal components (0.077). Thus, the methods using normalized data appear to be the best prediction methods because they provide the lowest Brier scores.

Second, for each rescaling method, the lowest score is obtained by those methods using the classification rule weighted by the inverse of distance. Those methods using equally weighted classification always have the highest score for each k value. Hence, using the classification rule weighted unequally does improve the prediction accuracy. Further, according to the score pattern of each method, to increase closest neighbors to predict the conflict potential of a nation does not provide better prediction. It is probably due to density of cluster map in 13-dimensional space, there are more alien points exist in clusters, and the conflict and peace clusters are not identical to each other. Overall, the performance of the 10 methods based on the Brier score can be ordered, from the best to the worst, as follows: Shearer [1], ND, NR, NN, SD, SR, SN, PD, PR, and PN methods.

According to the Brier score, the methods using principal components (PN, PD, and PR methods) are the worst. This is because they need at least k=8 to obtain the lowest score which is still not the lowest score of all. Using principal components to rescale the feature values is also a least efficient method. This is because the extra step is time-consuming. Due to these drawbacks, this study drops the PN, PD, and PR methods in the next two- and three-year-out predictions.

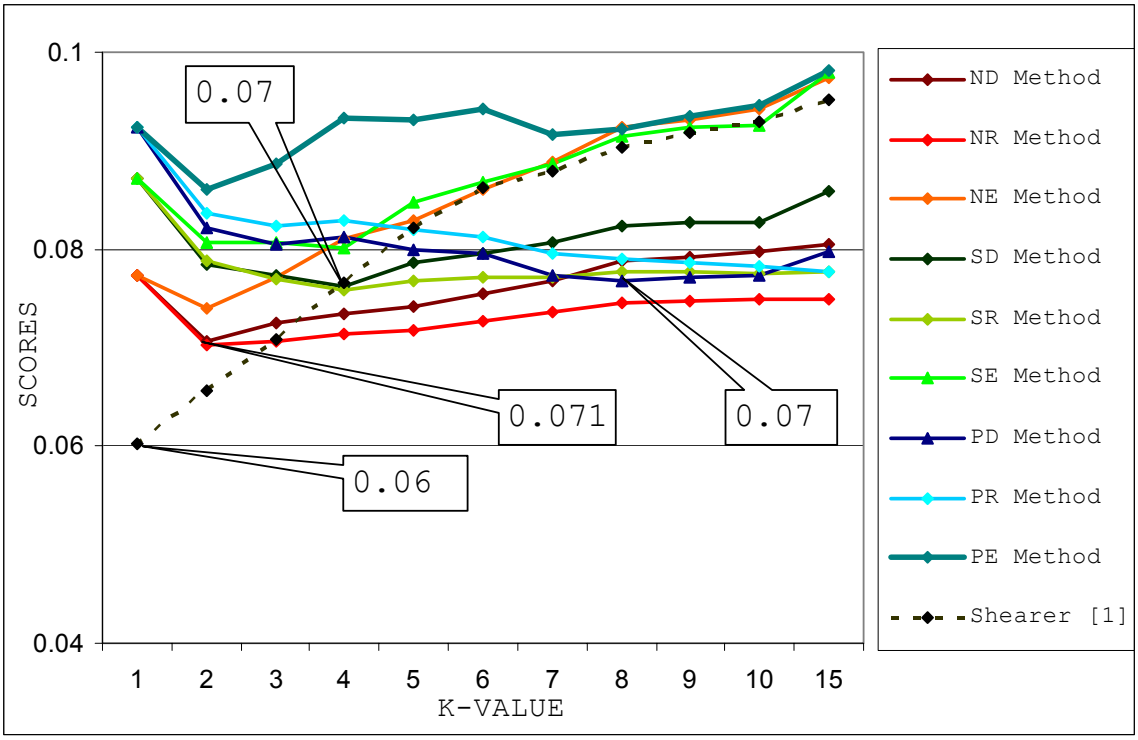

Figure 11. Result: One-Year-Out Prediction (1998-2003). The lowest score of all methods is given by the Shearer method. The lowest score of this study's proposed methods is given by ND method. The difference is 0.011.

C. TWO- AND THREE-YEAR-OUT PREDICTIONS

In one-year-out prediction, the Shearer [1] method outperforms this study's proposed methods by at least 0.011 average Brier score for each nation, and provides the best prediction only using one closest neighbor to predict a nation's conflict status. In this section, this study continues to look at how each of the proposed methods, excluding those using principal components, performs in the two- and three-year-out predictions.

In two-year-out prediction, the conflict potential is predicted based on the conflict-peace status from two years later. Figure 12 shows the predicted results of two-year-out prediction. The lowest score of the Shearer [1] method is obtained at k=1; whereas, the lowest scores of all this study's methods are provided at k=2. The pattern of Shearer's [1] method increases as k value increases and decreases as k>10. In this study's methods, the score patterns start at higher score at k=1; decrease dramatically at k=2; and, then, increase very slowly except for those using equally weighted classification rules. Briefly, the score patterns of this study's methods show that increasing k value does not improve the prediction, but does show a potential improvement in the Shearer [1] method.

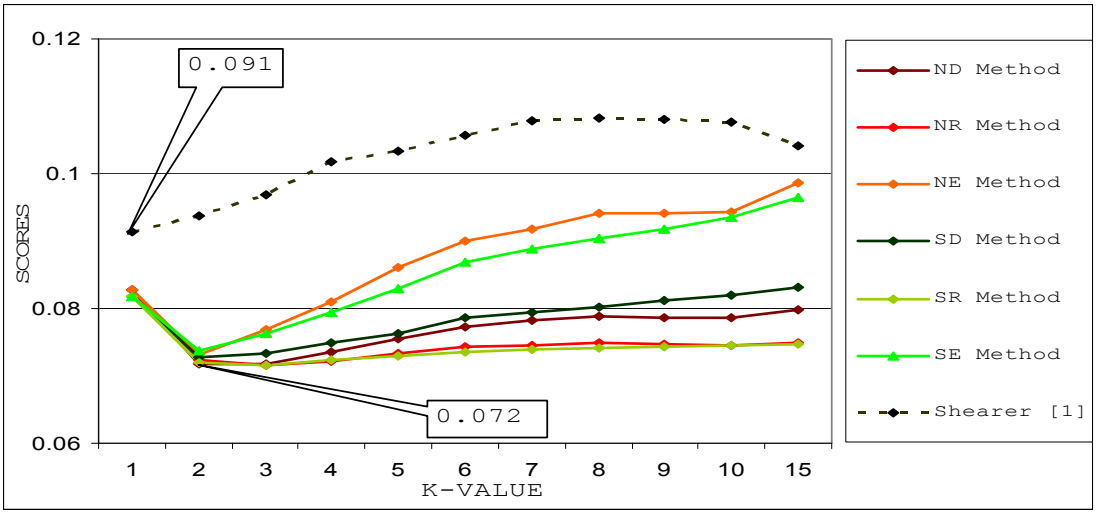

Figure 12. Result: Two-Year-Out Prediction (1998-2003). The best predicted result is give by ND method at k =2. This method improves by 0.019 from the Shearer method.

In three-year-out prediction, this study predicts the conflict potential based on the conflict-peace status from three years later. Figure 13 shows the prediction results. The Shearer [1] method does not work as well as this study's methods. Comparing this study's methods in the two-year-out prediction with that of Shearer, the approximate same lowest score and the score pattern for each method is similar -- except for the lowest score, which is provided by SD method at k=3. The difference, however, from ND method's lowest score is marginal. In Shearer's [1] method, its lowest score is not obtained at k=1; instead at k=15. It will continue to decrease as k increases. When comparing the pattern of SD method with the Shearer [1] method, increasing k value in this study's method does not improve SD performance at all, but shows a potential improvement in Shearer's [1] method.

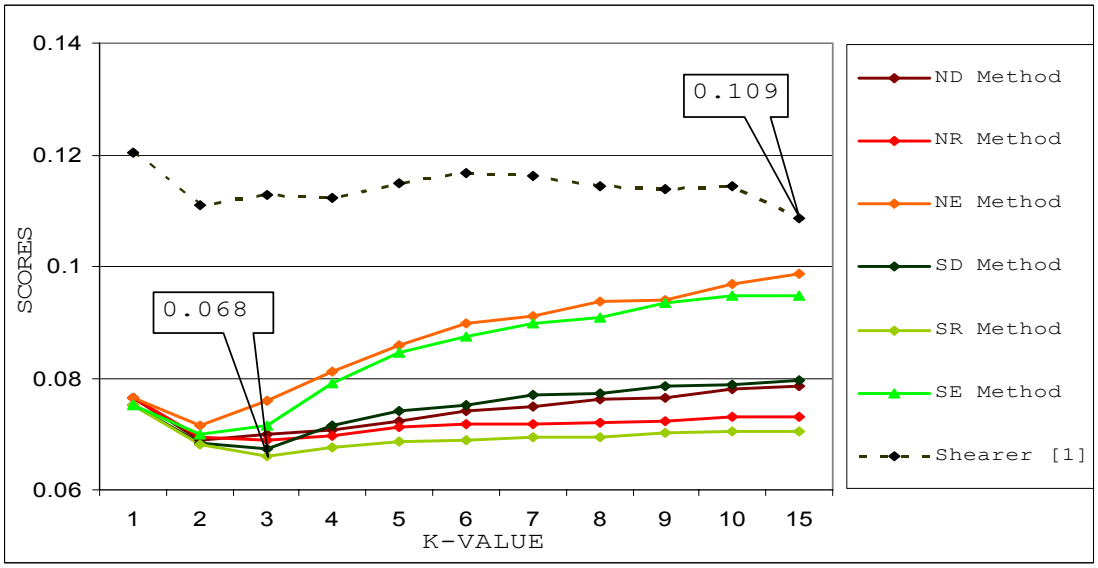

Figure 13. Result: Three-Year-Out Prediction (1993-2003). The best predicted result is given by SD method at k =2. This method improves by 0.019 from Shearer's method.

Next, this study wants to know which method, by utilizing the k value, provides the lowest score as the prediction of 155 nations' conflict potential up to three years (validated from 1998 to 2003). To obtain the overall performance of each method, this study combines three results regarding one- two- and three-year-out predictions. The combined result in predicting 155 nation's conflict potential from 1998 to 2003 at k=1, 2…10, and 15 is the average score of three results. The results show in Figure 14.

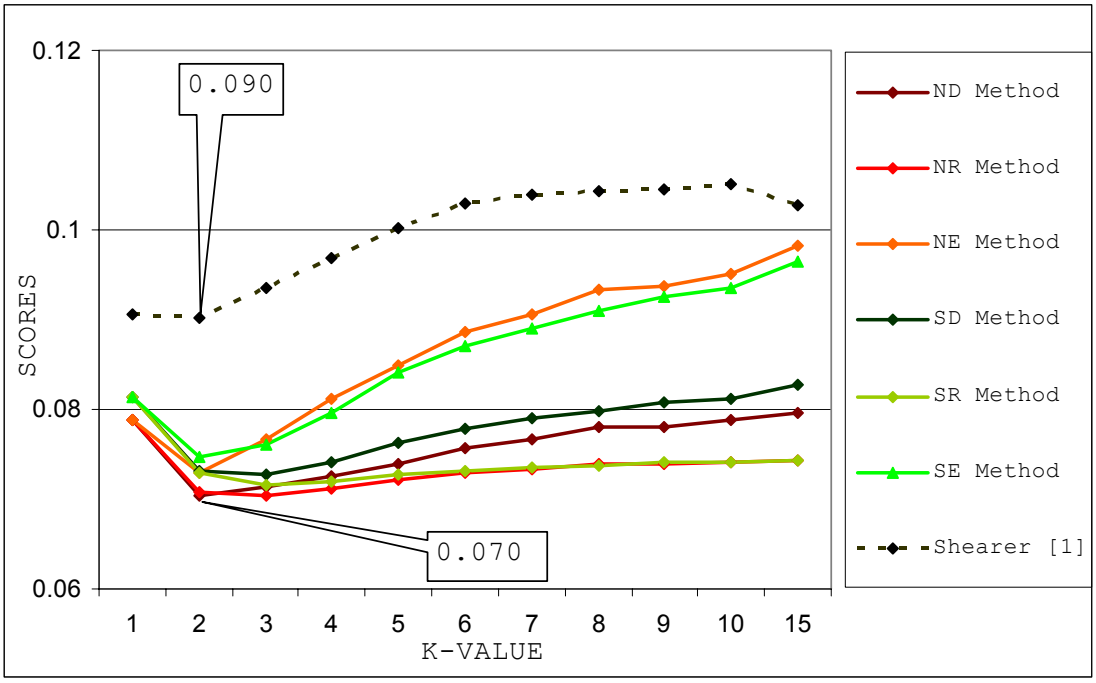

Figure 14. Result: Overall Prediction (1993-2003). The best predicted result is given by ND method at k=2. This method improves by 0.019 from the Shearer method.

D. DISCUSSION

Up to now, this study used its proposed methods to predict the conflict status in 155 nations from 1998 to 2003. Three different predictions, one-, two-, and three-year-out predictions, were applied to compare the prediction results. In each year, k was set for 11 different values to predict the conflict probability and Brier scoring rule was used to evaluate accuracy. This study compared its proposed methods with the Shearer's [1] method and determined which method is the most accurate and efficient method.

In one-year-out prediction, this study's proposed methods did not have any improvement over the Shearer's method. Prediction results in this study, however, had

some improvement in the two- and three-year-out predictions. The overall prediction performance for each method is the combined result of three predictions: one-year-out, two-year-out, and three-year-out. The result suggests that this study's methods provide better prediction. Among these prediction results, the three-year-out prediction gives the largest improvement by dropping the average Brier score by 0.055. This is followed by the two-year-out prediction by 0.02 and, then, followed by the one-year-out prediction by -0.01 (negative improvement). It appears that this study's method is more suitable to predict the conflict probability further into the future. One possible reason is that the moving average method Shearer [1] used to project a nation's 13 statistics does not work well to project their values far into the future.

In one- two- and three-year-out predictions, this study validates the predicted results from 1998 to 2003. The lowest scores of Shearer's [1] are 0.06, 0.091, and 0.109 in three predictions; their responding k values are 1, 1, and 15. Shearer's method results show that as more year-out prediction is conducted, the predicted errors of Shearer's [1] method increases. It implies that the conflict and peace clusters in cluster map are less identical to each other, and a larger k value is needed to obtain a lower score. Note that in Shearer method, the cluster densities in one-, two- and three-year-out predictions are the same. The cluster maps for these three predictions have the same data points, but their pictures are not the same due to the difference of projected feature

values. Once the clusters are not identical to each other (more alien points in each cluster), it is necessary to use more neighbors to obtain a satisfactory accuracy.

In this study's methods, the score pattern shows the biggest improvement at k=2. This k-value also provides the lowest score in the overall prediction. In each year-out prediction, the methods which provide the lowest score are ND, ND, and SD; their responding k value are 2, 2, and 3; and their scores are 0.071, 0.071, and 0.068. This shows that as the conflict cluster density decreases, there may be a need for a larger k value to get the lowest score, but this k value is going to increase very slowly and the improvement is not significant. When the data points are plotted in the 13-dimensional spaces, the densest conflict cluster map is provided by the one-year-out prediction; then two- and three-year-out predictions, respectively. Thus, this study's outcomes suggest that even the cluster map are getting sparser, the conflict and peace clusters are still identical to each other, and a small k value still provides a good prediction result. In all, this observation shows that k=2 is good if the density of the cluster plot is high; whereas, if the density is low, a larger value k may be better.

Among this study's proposed methods, each method has it own score pattern. These score patterns provide good knowledge as to which methodology is the best to apply. For instance, comparing the prediction result by the rescaling method, this study identified that the prediction using the normalized data outperforms other rescaling methods. Also, this study identifies that those methods

using classification rule by the inverse of distance provide the highest accuracy. As result from the overall prediction, the ND provides the most accurate prediction (lowest score).

IV. CONCLUSION AND RECOMMENDATION

In this thesis, several probability prediction methods are proposed to predict the future conflict potential in 155 nations. This study applies these methods to predict the conflict potential for up to three years into the future. The probability predictions are evaluated using the Brier score. The results suggest that, overall, this study's proposed methods give lower scores (higher accuracy) in comparison with the Shearer [1] method.

Nevertheless, different methods are more applicable in different scenarios. For instance, the results show that Shearer's method is better for one-year-out prediction; however, this study's methods are better for two- and three-year-out predictions. It appears that the statistical projections do not work well to project future feature variables: the more year-out predictions give more projecting errors. Since this study's methods do not apply an extrapolated projecting, when more than two-year-out prediction is conducted, the trend pattern for each feature variable is well maintained; therefore, the prediction error is smaller.

In addition, the selection of k value also plays an essential role to predict the conflict potential, and it affects prediction results in different setting. For instance, when the density of the cluster is high, Shearer's method provides the best prediction result with a small k in one-year-out prediction, but as more years out predictions are applied, a larger k may be better. When the density is low, such as more years out predictions, this

study's method provides a more accurate prediction with a small k value. The objective is to obtain the lowest score and k=2 is always a good choice.

The overall result suggests that the best method to obtain the highest accuracy is to use normalized data and predict future conflict potential weighted by the inverse of distance between the two closest neighbors. On average, this method, using the Brier score, improves about 0.02 in predicting the future conflict potential for 155 nations validating from 1998 to 2003. This best method is also an efficient way to predict the future conflict potential. This is because there is no need to forecast future features or to use many neighbors (k-value) to predict the conflict probability. The result shows the improvement.

Unfortunately, using this study's proposed methods has one disadvantage. Without using a trend function to predict the future feature variables, there is prediction limitation due to the unavailability of data sets. For instance, the current data set cannot be used to predict the conflict probability in 2015, like what Shearer did in his method. It is because no data is available across 12 years in this study.

In closing, what was observed in this thesis study can be built upon, explored, and verified with more year-out predictions. Further, it is expected that other methods can be explored to minimize prediction error and shed significant light on the possibility to predict conflict potential by using different concepts and methodologies and to provide a more efficient and accurate early warning of state failure in the future.

LIST OF REFERENCES

[1] Lieutenant Colonel Robert Shearer, Recognizing Patterns of Nation State Instability that Lead to Conflict, *Center for Army Analysis,* 2007.

[2] Sean P. O'Brien, Anticipating the Good, the Bad, and the Ugly: An Early Warning Approach to Conflict and Instability Analysis, *Journal of Conflict Resolution* vol. 46, pp. 791-811, 2002.

[3] Nathaniel Beck, Gary King, and Langche Zeng, Improving Quantitative Studies of International Conflict: A Conjecture, *American Political Science Review*, vol. 94, pp. 21-36, 2000.

[4] Jon C. Pevehouse, Interdependence Theory and the Measurement of International Conflict, *The Journal of Politics*, vol. 66, pp. 247-266, 2004.

[5] D. Marc Kilgour, and Frank C. Zagare, Explaining Limited Conflicts, *Conflict Management and Peace Science,* vol. 24, pp. 65-68, 2007.

[6] John Robst, Solomon Polachek, and Yuan-ching Chang. Geographic Proximity, Trade, and International Conflict/Cooperation, *Conflict Management and Pease Science*, vol. 24, pp. 1-24, 2007.

[7] Heidelberg Institute of International Conflict Research at the Department of Political Science, University of Heidelberg, *Conflict Barometer 2007*, *16th Annual Conflict Analysis*, 2007.

[8] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining, *Introduction to Linear Regression Analysis*, Academic Press, 4th Edition, 2006.

[9] Tilmann Gneiting, *Strictly Proper Scoring Rules: Assessing Predictions for an Uncertain World*, INFORMS Annual Meeting, 2007.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Kyle Lin
   Naval Postgraduate School
   Monterey, California

4. Yu-Chu Shen
   Naval Postgraduate School
   Monterey, California

5. Shian-Kuen Wann
   Naval Postgraduate School
   Monterey, California