



**FORECASTING COUNTRY CONFLICT
WITHIN MODIFIED COMBATANT
COMMAND REGIONS USING STATISTICAL
LEARNING METHODS**

THESIS

Sarah Neumann, 2d Lt, USAF
AFIT-ENS-MS-18-M-149

**DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY**

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

DISTRIBUTION STATEMENT A
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views expressed in this document are those of the author and do not reflect the official policy or position of the United States Air Force, the United States Department of Defense or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT-ENS-MS-18-M-149

FORECASTING COUNTRY CONFLICT WITHIN MODIFIED COMBATANT
COMMAND REGIONS USING STATISTICAL LEARNING METHODS

THESIS

Presented to the Faculty
Department of Operational Sciences
Graduate School of Engineering and Management
Air Force Institute of Technology
Air University
Air Education and Training Command
in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Operations Research

Sarah Neumann, B.S.

March 2018

DISTRIBUTION STATEMENT A
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT-ENS-MS-18-M-149

FORECASTING COUNTRY CONFLICT WITHIN MODIFIED COMBATANT
COMMAND REGIONS USING STATISTICAL LEARNING METHODS

THESIS

Sarah Neumann, B.S.

Committee Membership:

Dr. Darryl K. Ahner
Chair

Dr. Raymond R. Hill
Member

Abstract

Conflict forecasts are crucial to Combatant Commanders' understanding of the dynamic environment encompassing countries within their area of responsibility. The current structure of the Combatant Commands (COCOMs) is rooted in geography by grouping nations in geographic proximity to the same regional command. However, leaders today question the effectiveness of the current structure. A novel modified k-means clustering algorithm is developed and implemented that groups countries based on data similarities and geographic proximity resulting in new COCOM groupings that improve conflict forecasts. The data spans various political, military, economic, and social characteristics of countries, and is used to develop conditional logistic regression models that predict the likelihood of a country to transition into or out of conflict. Predictions from models using the original COCOM regions are compared to predictions from models using the new, data and geography based, regions. Reorganizing the COCOMs based on data similarity and geography improves conflict forecasting models and prediction capabilities significantly. The new models achieved training data classification accuracies exceeding 89% for each new COCOM and improved predicting the validation data with classification accuracies increasing up to 2% as compared to the current COCOM models and up to 2.5% in comparison to previous conflict studies. This methodology of grouping countries based on data similarities and geographic proximity leads to newly defined COCOMs which improve overall forecasting ability of conflict transitions, the best found in the literature to date. These results can help Combatant Commanders better understand the dynamic conflict environment of their area of responsibility and lead to the development of more effective operational and strategic plans.

This work is dedicated my family and friends who have encouraged and supported me unconditionally through the course of this work.

Table of Contents

	Page
Abstract	iv
List of Figures	viii
List of Tables	ix
I. Introduction	1
1.1 Background	1
1.2 Problem Statement	3
1.3 Research Objectives	3
1.4 Research Questions	3
1.5 Assumptions and Limitations	4
1.6 Overview	5
II. Literature Review	6
2.1 Overview	6
2.2 Previous Conflict Modeling Efforts	6
2.3 Influential Variables	11
2.4 Country Groupings	14
2.5 Overview of Methods	16
2.6 Summary	19
III. Methodology	20
3.1 Overview	20
3.2 Dataset	21
3.2.1 Database Development	21
3.2.2 Dependent Variable	21
3.2.3 Independent Variables	22
3.2.4 Challenges working with the Data	25
3.3 Country Grouping Development	29
3.3.1 Data for New COCOM Development	29
3.3.2 Principal Component Analysis	30
3.3.3 Modified K-means Algorithm	34
3.4 Model Building Procedure	37
3.5 Assessing Model Adequacy	42
3.6 Summary	46

	Page
IV. Analysis and Results	47
4.1 Overview	47
4.2 New Combatant Command Groupings Formation	47
4.3 Model Results	58
4.3.1 Model Assessment and Validation	62
4.3.2 Model Comparisons	69
V. Conclusions	74
5.1 Overall Conclusions	74
5.2 Implications of Research	77
5.3 Future Research	78
Appendix A. Independent Variables Information	80
Appendix B. Final Models	81
Bibliography	90

List of Figures

Figure		Page
1.	Current Combatant Command Structure	16
2.	Methodology Overview	20
3.	Variables used in PCA and Country Grouping	30
4.	Example Horn's Curve Plot	32
5.	Purposeful Selection of Covariates Summary	39
6.	Horn's Curve Plot	48
7.	Heat Map of Correlations between Variables and Principal Components	49
8.	Military and Government (PC2) vs Quality of Life (PC1)	50
9.	Data Similarity Only Groupings	52
10.	Score Plot with Data Similarity Only	54
11.	Data Similarity and Geography Groupings	55
12.	Score Plot with Data Similarity and Geography Equally Weighted	56
13.	Map of Final Groupings	57
14.	Score Plot for Final Groupings	58

List of Tables

Table	Page
1. Mapping of the Dependent Variable	22
2. Government Type Mapping from Polity	24
3. Mapping of Regime Type to Fill in Missing Polity Values	28
4. Example Classification Table	44
5. Discrimination Rating Guidelines	46
6. Principal Components Descriptions and Variance	50
7. Overview of Frequency of Variables in the Models	61
8. Individual Model Significance and Hosmer-Lemeshow Test Results	63
9. Individual Model AUC Results	65
10. Overall Classification Accuracies for Cutoff Point=0.5	66
11. New COCOM Classification Accuracy Results with Optimal Cutoff Point	67
12. New COCOM Classification Accuracy Results with Best Cutoff Point	68
13. Suite HL Test Results	70
14. Suite AUC Results	71
15. Suite Classification Accuracy Results with Cutoff Point=0.5	71
16. Study Classification Accuracy Comparison	73
17. USCENTCOM In Conflict	81
18. USSOUTHCOM In Conflict	81
19. USAFRICOM In Conflict	82
20. Modified USEUCOM In Conflict	82

Table	Page
21. USPACOM In Conflict	83
22. Modified USNORTHCOM In Conflict	83
23. NewCOCOM 1 In Conflict	83
24. NewCOCOM 2 In Conflict	84
25. NewCOCOM 3 In Conflict	84
26. NewCOCOM 4 In Conflict	84
27. NewCOCOM 5 In Conflict	85
28. NewCOCOM 6 In Conflict	85
29. USCENTCOM Not In Conflict	85
30. USSOUTHCOM Not In Conflict	86
31. USAFRICOM Not in Conflict	86
32. Modified USEUCOM Not in Conflict	86
33. USPACOM Not in Conflict	87
34. Modified USNORTHCOM Not in Conflict.....	87
35. NewCOCOM 1 Not in Conflict	87
36. NewCOCOM 2 Not in Conflict	87
37. NewCOCOM 3 Not in Conflict	88
38. New COCOM 4 Not in Conflict.....	88
39. NewCOCOM 5 Not in Conflict	88
40. New COCOM 6 Not in Conflict.....	89

FORECASTING COUNTRY CONFLICT WITHIN MODIFIED COMBATANT COMMAND REGIONS USING STATISTICAL LEARNING METHODS

I. Introduction

1.1 Background

Violent conflict between competing nations and within a given nation is not a new problem, but is an ever evolving one. The factors that influence and indicators that signal conflict for nations is an expanding field of study that relies on the ever expanding availability of quality data. The ability to predict when nations will enter into conflict can help leaders to identify threats and potential risks within regions and gives the ability to better understand and possibly mitigate the impending conflict. In 2016, the Heidelberg Institute for International Conflict Research (HIIK) tracked conflicts around the world and observed 402 conflicts globally [1]. The level of intensity, nations involved, and factors that influence these conflicts vary between conflicts and create a unique conflict environment for each nation. Because there is no set way to model conflict, there are many different ways in which conflict has been studied and predicted.

From global models encompassing all conflicts around the world to modeling disputes within a single country, these conflict models aim to identify potential indicators of a nation at risk of entering conflict and identify factors that influence a nation's likelihood to leave a state of conflict. Previous regional models of conflict have shown success in predicting conflict and have provided insight into how geographically similar countries transition into and out of conflict [2, 3]. Although these models have

provided insight into predicting conflict, these models may not give the best insight from a military perspective. These previous models have largely been based on groups of nations with geographic similarities, but do not necessarily capture the nature of conflict explicitly from the perspective of the Combatant Commands. Understanding conflict through the lens of the military in each of the Combatant Commands is key for leaders to best develop the most effective strategic and operational plans. These models can help better inform leaders of the dynamic conflict environment they oversee and are intended to enhance, not replace analysis of country-level experts.

After experiences in World War II, the United States developed the Combatant Command (COCOM) structure to more effectively handle multiple conflicts in different regions of the world [4]. The basis of the COCOM structure has largely remained unchanged despite changes in many nations' development and the overall global environment. Top leaders of the military have questioned the current structure of COCOMs by considering if our COCOMs are as effective as they can be or if the current structure could be improved upon [5]. While questions have been raised in regards to how best to group nations together, there have been few answers on how they should be grouped together, specifically from a military perspective. Questions on how to group similar countries together effectively and how to best predict conflict for the different regions of the world are the questions that motivated this research. Specifically, this research seeks to answer if the COCOM structure can be improved to best serve the military and Combatant Commanders in understanding the nature of conflict and predicting future conflict in their area of responsibility.

This research examines models based on nations currently assigned to each COCOM and considers if the COCOM structure can be improved to yield more accurate conflict predictions. This study considers improving the COCOM structure by grouping nations based on location and data similarity. By developing COCOM specific

models to best provide insight and predict regional conflicts around the world, a new perspective is added to the conflict prediction field of study and can give more accurate insight into nations vulnerable to conflict. Models developed from a military perspective may better allow identification of countries susceptible to rogue non-state actors, and more efficiently allocate resources for future conflicts and conflict prevention.

1.2 Problem Statement

Define new Combatant Command groupings for nation-states based on geography and open source political, military, economic, social, information and infrastructure (PMESII) data. Develop statistical models to compare current and new Combatant Command groupings' conflict prediction capabilities and gain insight into nations' tendencies to transition in and out of a state of conflict.

1.3 Research Objectives

The objectives of this study are to develop new groupings of countries within geographic Combatant Commands based on data similarity and geographic proximity and to develop and compare prediction models to predict the probability of a nation-state to transition into or out of a state of conflict from a military perspective.

1.4 Research Questions

This study seeks to answer the following research questions pertaining to the structure of the Combatant Commands and nation-states' tendencies to transition into or out of conflict.

Question 1

How should we group countries to develop statistical models for conflict prediction?

Question 2

Can we obtain better models with different groupings other than strictly regional groupings?

Question 3

Do neighboring countries affect a nation's likelihood of transitioning into or out of a state of conflict?

Question 4

What is the best measure of border conflict in predicting transitions of countries into or out of a state of conflict?

1.5 Assumptions and Limitations

This study is based on five underlying assumptions. The first assumption is that the data used in the study are accurate and important to describe the similarities between countries and their likelihood of conflict. The second assumption is that Shallcross's methods for data selection in his study were appropriate [2]. The third assumption is that the HIIK conflict intensity variable is an accurate measure of the level of conflict a country experiences and that we have appropriately defined the cut off point to distinguish between in conflict and not in conflict states of a country. The fourth assumption is that the current Combatant Command structure is a sufficient starting point for grouping countries together based on data similarity and geography

and that there should be 6 geographic Combatant Commands. The final assumption is that the prediction models developed in this study represent the base case situation for countries' likelihood to transition into or out of conflict and does not account for future data changes or unforeseen shocks to the conflict environment.

The main limitation in this study is data availability. This study had access to several open source datasets that have information for many of the countries included in the study, but there are gaps in this data; not every country has all the information for every variable. Thus, this study was limited in what variables could be included and used imputation methods only when necessary.

1.6 Overview

This thesis is organized into 5 chapters including this introduction chapter, Chapter 1. Chapter 2 contains a literature review of prominent studies and background information pertaining to this study. Chapter 3 discusses the methodology of the study to group nations and form the prediction models. Chapter 4 details the results and analyses and chapter 5 offers final conclusions, implications, and possibilities for future research.

II. Literature Review

2.1 Overview

The purpose of this chapter is to provide a background of important and influential works in the nation-state conflict prediction field of study as well as background on statistical methods used in previous studies. This chapter is divided into 5 sections each providing a background on the extent of the research and performance of the models. Section 2.2 will discuss previous models that have been developed to predict conflict and their performance. Section 2.3 will discuss important variables that have been used in previous studies to indicate the likelihood of conflict or peace for countries. Section 2.4 provides the background on how nation-states have been grouped in terms of location and similarities between nation-states in previous studies. Finally, section 2.5 will briefly discuss the background of some of the methods used to analyze important factors in the data and prediction models.

2.2 Previous Conflict Modeling Efforts

There have been numerous studies aimed to predict when nations will experience conflict. Within these models there are varying definitions of conflict and how it should be measured, methodologies to produce predictive models, ranges of predictions, and accuracies of the models. This section serves to detail some of the major contributing works to the field of forecasting conflict.

The work, *Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction*, argued the utility of creating models for predicting political conflicts in a diverse global environment, specifically for policy purposes [6]. After defending the utility behind the research and efforts put into predicting conflicts, the authors discuss the problems with previous analyses. Ward's study expanded

previous work by calculating the cumulative probability of civil war onset in various countries [6]. While all the variables are statistically significant in this model, the predictive accuracy was not high. Ward [6] developed a new predictive model that incorporates behavioral and institutional variables to predict escalation of conflict in the form of a hierarchical logit model to predict conflict six months prior to the onset of conflict. This hierarchical model allowed the intercept and slope to vary depending on different groups of nations modeled. Using a validation set of observations from the data, the models were able to accurately predict the onset of civil war with 94% accuracy, but also under predicted the actual number of civil wars [6]. Ward's [6] work is significant in that it expanded previous work to focus on increased prediction accuracy instead of only the significance of variables in a regression analysis.

The study, *A Global Forecasting Model of Political Instability*, led by Dr. Jack Goldstone and funded by the U.S. Central Intelligence Agency's Directorate of Intelligence, predicted the onset of political instability two years prior to the beginning of a conflict [7]. Using open source global data from 1955 to 2003, the study employed a series of techniques to model and predict the onset of 141 incidences of political instability over time [7]. Their definition of conflict included events that caused major political instability and included the four types of events: Revolutionary Wars, Ethnic Wars, Adverse Regime Changes, and Genocides and Politicides [7]. The dependent variable for the study was the status of a country's stability. A country was considered to be in a state of instability in a certain year if any one of the four major political instability events occurred during that year. The independent variables for the study were chosen through extensive research in consideration of the variables' impact on political instability and reasonably available data [7]. The study tested a series of models including stepwise logistic regression and neural networks to predict conflict. The final best performing model was a simple logistic regression model that

included 4 independent variables, was able to predict the onset of conflict two years prior with 80% accuracy. The study also found that the most powerful predictive factor was regime type [7].

The Goldstone study made significant contributions to conflict prediction research in that it produced a model that could predict conflict instead of only identifying significant factors in a model. This study was limited to a global model and did not account for differences of nations' geographic locations. The study did consider an individual model for Africa and saw a change in what variables mattered in predicting conflict, but it was the only region modeled individually [7].

Another study conducted by the Center for Army Analysis sought to classify patterns of instability to inform what factors lead to conflict [8]. This study sought to use more intuitive methods to explain how nations move in and out of conflict. The dependent variable for the model was developed from the HIIK data which classifies national and international political conflicts into six categories relating to the intensity of the conflict [8]. In Shearer's [8] study, the top 4 levels of the HIIK variable were mapped into two categories: peace and conflict. The independent variables consisted of thirteen features from unclassified data sources relating to political, military, economic, and social characteristics of a country [8]. Principle Component Analysis (PCA) was conducted to reduce the feature space into three dimensions which helps to visualize the data [8]. The study utilized a smoothing algorithm to forecast future feature vectors and employed two classifiers, k-Nearest Neighbors and Nearest Centroid algorithm, to classify the stability of a nation over time with 85% accuracy [8]. This study provided an intuitive way to visualize and interpret a nation's likelihood of conflict and was able to predict conflict further into the future with accuracies similar to other prediction models. Only global models were considered for this study and some of the prediction techniques led to unknown classifications [8].

In 2009, the study *Predicting Armed Conflict, 2010-2050* proposed a dynamic multinomial logit model for predicting long-term conflict trends [9]. The data used for this model included data from 1970-2008 and predictions were obtained through simulating the behavior of conflict variable using the estimates from the logit model [9]. The logit model estimated the likelihood of a nation's transition into one of three states of conflict: no armed conflict, minor conflict, and major conflict [9]. The definition of conflict used for this study divided conflict into three levels based on the number of battle related deaths in a given year [9]. The study also divided the world into eight regions based on a condensed version of the United Nations' definition of regions and incorporated predictor variables based on these regions in their models [9]. Different combinations of independent variables were used to develop six separate models that emphasized different influences on the onset of conflict: conflict history, country development, and neighborhood behavior [9]. The simulated predictions for this study were based on proposed scenarios and projections of demographic trends over time including changes in infant mortality, education and fertility rates [9].

Building on previous theses and work in the field, Shallcross [2] combined several methods that had been used in the field to produce a new method of modeling conflict transitions over time. Shallcross [2] developed logistic regression models for six geographic regions of the world to calculate the probability of each nation transitioning into and out of a state of conflict. The data Shallcross [2] used was from several open-source data repositories from which he determined thirty predictor variables that were used in different combinations for each conflict model for each region. The dependent variable for the models was a binary variable indicating a change in conflict status. The conflict status of a nation was based off of a mapping from the HIIK conflict intensity data. Instead of using only the top four levels of conflict like some previous works had, Shallcross [2] used all levels of the HIIK variable and considered states

not in conflict for HIIK levels 0 to 2 and in conflict for levels 3 to 5. The models produced the probability of a nation transitioning into conflict and transitioning out of conflict. After obtaining the probabilities from the models for each country, the probabilities were applied to a nation specific Markov Chain model with two states, “in conflict” and “not in conflict” to forecast nations’ tendencies to move in and out of conflict. Shallcross’s [2] models were focused on prediction accuracy and sought to predict regional conflict trends over time. Shallcross [2] achieved overall prediction accuracies above the 80% benchmark for his models and achieved the best results in the literature to date.

Expanding on the work of Shallcross [2], Leiby’s [3] study specifically focused on incorporating environmental factors, specifically water and neighboring country conflict, into prediction models and analyzing the effect they had on prediction accuracy. Leiby [3] used conditional logistic regression to predict transitions of nation conflicts from 2004-2014 in each geographic region consistent with the regions developed in Shallcross [2]. Leiby [3] used the same open source data in Shallcross [2] with the addition of two independent variables. The additional independent variables were the percentage of the total number of bordering countries in conflict and a binary variable indicating if at least one bordering nation was in a state of conflict. Leiby’s [3] choice of model building strategy was bidirectional stepwise selection in which he began with a model with only the intercept and added or subtracted variables one at a time depending on each variable’s G statistic. The G statistic is defined later in this chapter. Five suites of models were created depending on the focus of the model. The five suites of models included building the most significant model based on G statistics and forcing the water variable and/or one of the border conflict variables into the model and then completing stepwise regression. Overall, Leiby [3] found that incorporating environmental factors into the models only marginally im-

proved model parsimony and predictive accuracy, but was able to achieve a training classification accuracy of 92%.

This section described the recent methodologies and models used to predict conflict. Overall, many of the models used logistic regression as a way to estimate conflict and conflict transition probabilities. Simulation methods, including Markov Chains, utilized probabilities from the logistic regression models to model conflict trends over time. Across the board, the prediction models achieved classification accuracies of at least 80%.

2.3 Influential Variables

There has been a plethora of research on investigating factors and indicators that influence conflict in a nation. One factor that is being investigated is the presence of new technology and its effect on conflict within a nation. In the article *New Technology and the Prevention of Violence and Conflict*, Mancini and O'Reilly [10] discuss research into the possibility of conflict prevention through new technology, specifically through mobile cell phone devices and the Internet. They cite that the rapid growing number of new technology especially in developing countries has brought new levels of global interconnectivity as well as dramatic cultural, social, economic, and political changes in societies around the world [10]. Not only does the increase in technology increase interconnectivity among people, it also produces a plethora of data that could be harnessed to better understand the influence of technology as well as indicators of possible conflict trends in regions [10]. Their study is based on specific instances in various regions throughout the world and the role that technology may have played in each instance. While technology is being emphasized as a potential conflict prevention method, their study comments on the possibility of secondary effects that come from an increase in technology in a country. Efforts can be made to use the technology

in a positive way such as disseminating information on potential threats to people, but there is also the potential for bad actors such as drug cartels or rebel groups to use the technology to further communicate and promote their agendas which may increase conflict [10]. Overall, the study concluded that increasing technology in a country, especially developing nations, does influence conflict within nations though not always in a positive or negative way. New technology such as cell phones and the Internet have the potential for both positive effects through communicating potential threats to the people of the nation and negative influences on conflict through increased connectivity between bad actors in a nation. This article provided specific instances of evidence of the effect technology has on conflict in specific regions of the world, but there was no overarching statistical analysis or evidence to support the claims of the mixed effect technology has on country conflict.

Another factor that comes up in conflict analyses and prediction studies is border conflict. Almost all of the previous studies mentioned had some type of effect modeled for the influence of bordering nations in conflict. The definition on what is considered a bordering nation and what that influence is on its surrounding nations varies between studies.

Several studies have modeled border conflict effects through some type of geographically based variable that considers if another bordering nation within a certain distance is in conflict. Shallcross [2] modeled this effect by considering the conflict intensity of nations directly sharing a border and the percent of the total border shared border with a given nation. Leiby [3] also considered border conflict with two additional variables that measured the percent of surrounding nations in a state of conflict and the presence of at least one bordering nation in a state of conflict. Other studies considered the distance to another nation and not necessarily if they share a common border. Hegre [9] considered nations as bordering if the two nations had

less than 100km between any points in their territories. These studies only consider the proximity of a nation to other nations in conflict to model the effect of conflict spillover across borders.

Other studies focus less on geographical definitions and argue that a border conflict effect has more to do with specific demographic influences that also define borders. *Bordering on Peace: Democracy, Territorial Issues, and Conflict* expands the definition of borders beyond geography. In this study, Gibler [11] includes power differences, colonial heritage, border history, ethnicity, and international institutions as pieces of the definition of a border. Gibler [11] also points to certain indicators that signal unstable borders and increased risk of conflict. These indicators include increased militarization, centralization, and ethnic groups that straddle a border [11]. This definition of borders and what indicates a stable border goes beyond a simple geographic definition and can be more insightful in what causes conflicts to move across borders.

In the study, *Refugees and the Spread of Civil War*, it is argued that population movements, specifically refugees, are “an important mechanism through which conflict spreads across regions” and across borders [12]. While refugees themselves do not necessarily engage directly in violence, the study found that the presence of refugees increases the risk of conflict in the host or origin countries [12]. This effect was modeled by an independent variable in a logistic regression which related to the number of refugees from a neighboring country with the term neighbor including countries that were within 950 km from each other’s territories [12]. Another study conducted in 2013, argued transborder ethnic kin (TEK) groups were influential in driving conflicts across borders [13]. While many have assumed that larger TEK groups would increase the risk of conflict, this study found there to be a curvilinear effect of the size of the TEK on conflict with large TEKs decreasing the risk of conflict across borders [13]. The study tested these effects on a series of logit models and found

the curvilinear effect significant in predicting conflict transitions [13]. In his 2002 study, Goldstone [14] poses a broader argument of growing evidence that demographic changes, specifically migrations of people into other countries, significantly influence the start of conflict in a nation. He specifically states that it is not necessarily the migration of people, but rather the clash of national identities when one distinct group migrates into an area that is considered to be the homeland of another ethnic group [14].

Across studies, there is agreement that border conflicts play a role in influencing other countries' susceptibility to transition into or out of conflict, but there is no consensus as to the best way to model this effect. The problem with modeling the border effect with only geographical considerations is that it assumes that island nations, who have no direct neighbors, do not have any spillover effects from other nations. However, definitions which include more demographic and social factors are not as intuitive to model and it may be difficult to model the true effect of these factors and how they influence other countries' tendencies to enter into conflict.

2.4 Country Groupings

Another area in which global conflict analyses and prediction studies have differed is in the level at which to model conflict. In the works of Hegre [9], Shearer [8], Goldstone [7], and Bokestein [15], the conflict was modeled at a global level in which the data for all countries were combined into one model. Bokestein [15] was able to increase the overall prediction accuracy by 3% by including a region variable to account for the differences in conflict in different regions of the world. More recent works modeled regions independently of one another in separate models. Most regions analyzed in conflict studies have largely been defined by geographic proximity, economic, ethnic, religious, and linguistic similarities. Nations sharing

commonalities in these factors are often grouped together. Shallcross [2] and Leiby [3] built their models off the work of Boekstein [15] and developed independent regional models based off regional groupings of nation-states suggested by the statistician Hans Rosling. Hans Rosling was a statistician well known for his work in using statistics to analyze world development trends and was a proponent of open source data [15]. In his TED Talk “The best stats you’ve ever seen”, Rosling proposes groupings of countries based on location and similarities in economic development [16]. The groupings he proposed inspired the regions used in the Boekstein [15], Shallcross [2], and Leiby [3] works to account for regional differences and develop region specific models.

Another way to view the different regions of the world is from a military perspective through the lens of the United States Combatant Commands. The COCOMs were developed after the United States’ experience in World War II and were established as the military’s way of dividing up the world to more effectively engage in different affairs and conflicts around the world [4]. The COCOMs have a long standing history of their geographic nature in terms of how they are structured and which countries fall under each COCOMs responsibilities. Since their development after World War II, they have largely remained unchanged in structure. Figure 1 shows the current structure of the COCOMs. It is apparent from the map that the current COCOM structure is heavily based on geography.

Military leaders now question the current structure of the COCOMs and considered if there is a more effective alignment of the COCOMs. General Dunford commented in his Senate confirmation hearing on the need to reevaluate the Combatant Commands and consider if there is a more effective way to structure them [5]. The questioning of the current COCOM structure ultimately leads to the question, how should the COCOMs be structured?

Gupta, Hanges, and Dorfman [17] presented another way to cluster countries to-

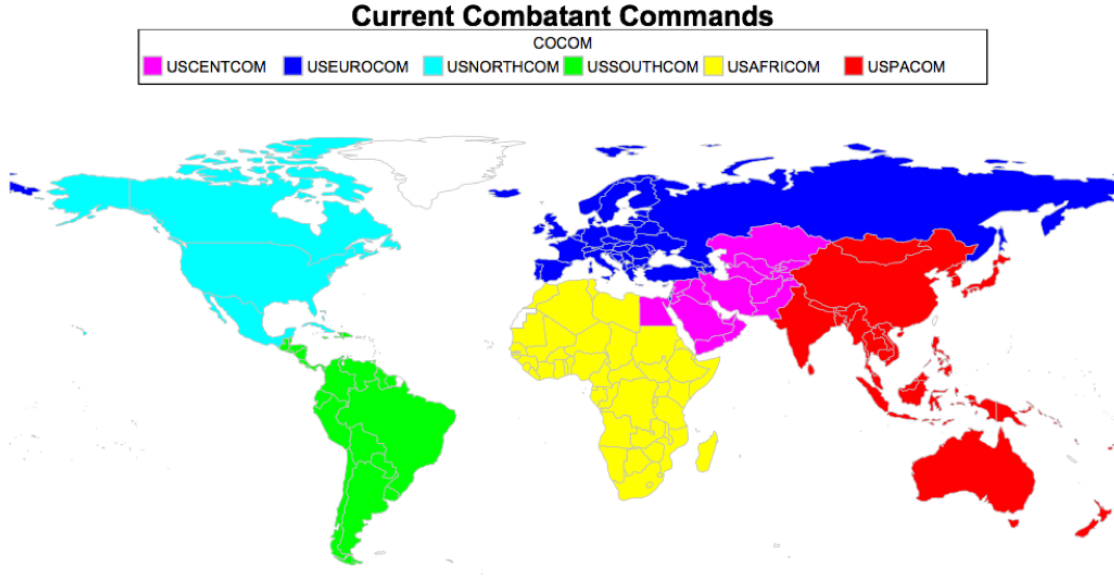


Figure 1: Current Combatant Command Structure

gether based on similarities beyond location and economic status. Their study predicted clusters of countries based on psychological, sociological, demographic, and economic characteristics (a priori) and used discriminant analysis to confirm clusters in a split half sample [17]. Their results provide strong evidence to support existence of 10 cultural clusters and provides the possibility for new definitions of regions within the world based on factors outside of geography.

2.5 Overview of Methods

This final section reviews the theory of logistic regression. The most recent and most successful prediction models have used forms of logistic regression as a way to calculate the probability of a nation to transition into or out of conflict. The goal in developing a logistic regression model is the same as in developing a linear model; find the best fitting and most parsimonious model to describe the true relationship between the independent variables, also known as covariates, and the dependent variable [18, p. 1]. While the goals are the same, logistic regression does differ in several ways

from linear regression. Logistic regression is used in situations in which the response variable y is dichotomous, such as in the instance of conflict transitions where either a transition was observed or not in a given year [18]. The parameters in the logistic regression model are estimated through maximum likelihood estimation. Estimations of parameters made through the maximum likelihood estimation technique produces estimated parameters that maximize the probability of obtaining the observed set of data [18, p. 8]. Because of the dichotomous dependent variable, the conditional mean is greater than or equal 0 and less than or equal to 1. The conditional mean is defined in Equation 1.

$$\pi(x) = E(Y|x) = \frac{e^{\beta_o + \beta_1 + \dots + \beta_p}}{e^{1 + \beta_o + \beta_1 + \dots + \beta_p}} \quad (1)$$

$\pi(x)$ = estimated probability of success ($Y = 1$) for given x

Given the nonlinear nature of the equation, a logit transformation is performed on $\pi(x)$ to obtain more desirable properties that are seen in linear regression. The logit transformation is defined in Equation 2.

$$g(x) = \ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] \quad (2)$$

The resulting equation after applying the logit transformation, $g(x)$, results in continuous linear parameters, over the range $[-\infty, \infty]$. The antilog of the parameters are interpreted as the estimated increase in the probability of success.

The methods to determine the overall significance of logistic regression models and the significance of parameters is another difference between linear regression and logistic regression. In logistic regression, the overall significance of a model is determined by the Likelihood Ratio Test (LRT) which compares the observed values of the response to the predicted values obtained from the model. The overall significance of the model is obtained by comparing the likelihood of the fitted model with only

the intercept and the likelihood of the saturated model. This comparison is otherwise known as the G statistic and is represented by the Equation 3.

$$G = -2\ln \left[\frac{\text{likelihood of fitted model}}{\text{likelihood of saturated model}} \right] \quad (3)$$

The G statistic follows a chi-square distribution with p degrees of freedom where p is the number of independent variable parameters in the model [18]. The null hypothesis for the LRT is that the p coefficients for the covariates in the model are equal to 0 and the null is rejected if the G statistic is less than the chi-square statistic with p degrees of freedom [18].

The LRT can also be used to determine the significance of parameters in the model. The LRT in this case compares the model without the variable and the model with the variable, assuming all other variables stay in the model. The G statistic in the single variable significance test follows a chi-square distribution with 1 degree of freedom.

One other way that the significance of variables in the model can be determined is by the Wald Test. The Wald Statistic for testing the significance of a single variable compares the estimated parameter to the estimated standard error of the parameter. The equation for a Wald statistic for a single variable i can be represented as seen in Equation 4.

$$W_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \quad (4)$$

The Wald Statistic, under the null hypothesis that an individual coefficient is zero, follows a standard normal distribution [18]. The LRT and Wald Statistics can be used in combination to determine the significance of the overall model, compare models with or without a certain subset of variables, and determine the significance of individual parameters.

2.6 Summary

This literature review provides background on previous models and techniques used to forecast conflict transitions of nations, how border conflicts affect a nation's likelihood of entering into a state of conflict, and the regional considerations of conflict. Logistic regression and simulation are effective ways to predict and model conflict trends over time and region specific models have produced the most accurate conflict predictions. Prediction models have been studied mainly based on geographic similarities between nations, but other methods have been proposed to group nations together outside of geographic factors. Furthermore, many of the variables used to predict the onset of conflict overlap throughout the studies. One variable that is modeled differently throughout the studies is the influence of border conflict. Geographic proximity measures and demographic variables such as number of refugees have also been used to determine the influence of a neighboring country on another country, but there is no one measure that has consistently been used to model the influence of border conflicts in conflict prediction models.

III. Methodology

3.1 Overview

This research explores the techniques used to potentially group countries into new COCOM groupings and develop models to forecast nation-state conflict. Section 3.2 describes the methodology used to develop the dataset used in this research in regards to data sources, variable selection, development of the dependent variable, and imputation procedures. Section 3.3 discusses the development of new COCOMs from each country’s current COCOM assignment. Section 3.4 covers the development of the conditional regional models to include discussions on logistic regression and the purposeful selection of covariates. Finally, Section 3.5 examines the model adequacy performance measures used in this research to validate and compare models.

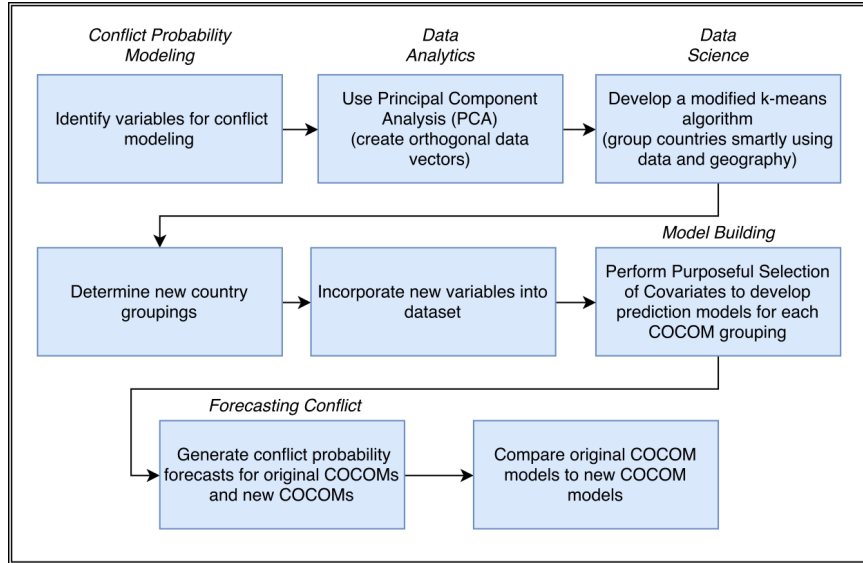


Figure 2: Methodology Overview

3.2 Dataset

3.2.1 Database Development

The data for this research is similar to and built off of the data used in the Shallcross [2] and Leiby [3] studies. This study considers the same 182 countries from the Shallcross [2] and Leiby [3] studies from 2004-2015 and includes 29 of the 30 independent variables studied in Shallcross [2]. The 182 countries consist of 181 of the 193 United Nations members as well as Palestine which is referred to as West Bank. The variable that was included in the Shallcross [2] study but not in this study was the Military Expenditure as a percent of government spending. This variable was not included in this study because it was missing almost 50% of the data. The other differences between the data used for this research and the Shallcross [2] study data is the addition of 2015 data for all variables, 2 additional technology related variables, 2 border conflict variables inspired by the variables developed in Leiby's [3] work, and a minor change in the definition of the dependent variable. These differences are discussed in greater detail in the following sections.

3.2.2 Dependent Variable

The dependent variable, Conflict Transition, is modeled as a binary variable indicating a change in conflict status for a country from the previous year. The conflict status of each country in a given year is determined by mapping the country's highest level of conflict intensity, also known as its HIIK value, in a given year. A HIIK level of 0, 1, or 2 is mapped to the conflict status of 0 indicating that a country is not in a state of conflict. HIIK levels of 3, 4, and 5 are mapped to a conflict status of 1 indicating that a country is in a state of conflict in that given year. The dependent variable Conflict Transition is equal to 0 if there is no change between the conflict

status of the current year i and the conflict status of the previous year $i - 1$. Conflict Transition is equal to 1 if the conflict status in a given year i is not equal to the conflict status in the previous year $i - 1$. Table 1 shows the combinations of values for conflict status in year $i - 1$ and year i and the respective values for the dependent variable Conflict Transition in year i .

Table 1: Mapping of the Dependent Variable

Conflict Status Yr $i - 1$	Conflict Status Yr i	Conflict Transition Yr i
0=Not In Conflict 1=In Conflict	0=Not In Conflict 1=In Conflict	0=No Transition 0= No Transition
0=Not In Conflict 1=In Conflict	1=In Conflict 0= Not In Conflict	1= Transition 1= Transition

3.2.3 Independent Variables

This research uses many variables included in previous conflict prediction studies, but also incorporates several new variables. The variables unique to this study and discussions on specific variable development are discussed in this section. The sources and units of measurement for all independent variables are found in Appendix A.

Internet Users

The Internet Users variable is from the World DataBank World Development Indicators and describes the number of Internet users per 100 people in a country in a given year.

Mobile Cell Subscriptions

The Mobile Cell Subscription variable is from the World DataBank World Development Indicators and gives the total number of mobile cellular phone subscriptions in a country in a given year.

Democratic Government Type (Binary Indicator)

Based on previous research, a country having a democratic government or not has been a significant variable in previous conflict modeling efforts. While the govern-

ment type variable does account for a democratic government type, the Democratic Government Type (Binary Indicator) variable served as a more general variable to differentiate between democratic governments and non-democratic governments. This variable was created by mapping the government type variable. The Democratic Government variable is equal 1 to if government type is democratic. Otherwise, the variable is equal to 0.

Transformed Polity

Both Shallcross [2] and Leiby [3] used the Polity IV variable which originates from the Center for Systemic Peace. The original Polity variable scored the government level based on a -10 to 10 scale with a -10 indicating fully autocratic and a 10 fully democratic [19]. Additionally, the variable indicates if a country experienced a foreign interruption, anarchy, or was in a transitional state with values of -66, -77, and -88 respectively. The Polity Score spectrum can be divided into three regime categories: autocracies (-10 to -6), anocracies (-5 to +5 and values -66, -77, and -88), and democracies (+6 to +10) [19]. The governments encompassed by the extreme negative values were similar to the types of governments with a value of zero and included in the same regime category. Because these extreme negative values are associated with similar types of governments as those governments included in the middle of the Polity score spectrum, the original polity variable was transformed so that those instances of extreme negative values, -66, -77, and -88, were mapped to zero. This ensured that all instances were in the -10 to 10 scale and could help better identify those countries experiencing more unstable or transitional governments.

Government Type

The Government Type variable was used in both Shallcross [2] and Leiby [3] studies as a dynamic variable to indicate a country's government type. It is a categorical variable which stems from mapping the original Polity Variable into 6 categories as

shown in Table 2.

Table 2: Government Type Mapping from Polity

Original Polity Value	Government Type Number	Government Type
-10 to -6	0	Autocratic
-5 to 5	1	Emerging Democratic
6 to 10	2	Democratic
-66	3	Foreign Interruption
-77	4	Anarchy
-88	5	Transitional

Percent Border Conflict

The Percent Border Conflict variable is consistent with the border conflict variable in Shallcross [2] and Boekestein [15]. The Percent Border variable is calculated by summing the percent of the perimeter of the border shared with a neighboring country multiplied by the HIIK level of conflict intensity for that country for all countries that share a physical border with a given country. Islands were assumed to not have any neighboring countries. Equation 5 defines the Percent Border Conflict variable.

$$PctBC_{ij} = \sum_{k=1}^n H_{kj} p_k \text{ where}$$

n = number of bordering countries for country i

H_{kj} = HIIK conflict intensity level for country k in year j

p_k = percent of border country i shares with country k (5)

i = Country $\in \{1, 2, \dots, 182\}$

j = Year $\in \{2004, \dots, 2015\}$

k = Bordering country

Average Border Conflict

The Average Border Conflict variable measures the average level of conflict surrounding a given nation in a given year. Islands are assumed to not have any neigh-

boring countries. Average Border Conflict score for a country is defined in Equation 6.

$$\begin{aligned}
 AvgBC_{ij} &= \frac{\sum_{k=1}^n H_{kj}}{n} \text{ where} \\
 n &= \text{number of bordering countries for country } i \\
 H_{kj} &= \text{HIIK conflict intensity level for country } k \text{ in year } j \\
 i &= \text{Country} \in \{1, 2, \dots, 182\} \\
 j &= \text{Year} \in \{2004, \dots, 2015\} \\
 k &= \text{Bordering country}
 \end{aligned} \tag{6}$$

Binary Border Conflict

The Binary Border Conflict variable is consistent with the binary border conflict variable in the Leiby [3] study and indicates if at least one bordering country is in a state of conflict in a given year. Islands are assumed to not have any neighboring countries. Binary Border Conflict score for a country is defined in Equation 7.

$$\begin{aligned}
 BinBC_{ij} &= \begin{cases} 1 & \text{if } H_{kj} \geq 3 \text{ for any country bordering country } i \\ 0 & \text{if otherwise} \end{cases} \\
 H_{kj} &= \text{HIIK conflict intensity level for country } k \text{ in year } j \\
 i &= \text{Country} \in \{1, 2, \dots, 182\} \\
 j &= \text{Year} \in \{2004, \dots, 2015\} \\
 k &= \text{Bordering country}
 \end{aligned} \tag{7}$$

3.2.4 Challenges working with the Data

One of the challenges in working with a large open source data set was determining which variables to include in the study and how to handle the missing values in the

dataset. The base set of variables are those 30 variables identified in the Shallcross [2] study. The amount of missing data in each of those variables led to removing the Military Expenditure as a percent of government spending; almost 50% of its values were missing. To compare different border variables in our study two additional border conflict variables similar to the ones used in the Leiby [3] study in addition to the variables from the Shallcross [2] study were added. The Mobile Cell Subscription and Internet users variables helped investigate the effect technology access has on conflict.

After determining which variables to include in the study, missing data gaps in the full dataset needed consideration. There are a number of ways in which missing data can be handled with varying levels of complexity, validity, and the nature of randomness of the missing data. This study assumed all data are missing at random meaning that the probability of a observation having a missing value for a variable may depend on the known values, but not on the value of the missing data itself [20].

One of the simplest ways to handle missing data is to only consider complete cases which have no missing values within all the fields for an observation. Only considering those complete cases that had no missing values means losing over 50% of the observations in the dataset. Therefore, imputation methods were used to fill in the missing data with appropriate values that would allow for analysis with as much information as possible, but maintain the integrity of the original data.

Before using general imputation methods on the whole dataset, a preliminary analysis of the missing values in the dataset looked for any patterns within the missing data and considered using prior knowledge of the data to fill in missing values. There were many values missing for South Sudan from 2004-2010. South Sudan gained independence from Sudan in 2011 [21]. Using this knowledge, South Sudan's missing values for 2004-2010 were imputed with the available and corresponding Sudan values.

Arable land was another variable that had a pattern of missing data. The majority of missing values for the arable land variable were 2015 values. Because the amount of arable land does not change very quickly in the span of a year or two, the missing values were imputed by carrying forward the last original value of arable land data for a given country.

The population variable only had missing values for Eritrea from 2012-2015. The missing values were imputed using a time series forecast of the population based on the population in Eritrea in 2011 and assuming a 2% growth rate. The 2% growth rate is consistent with the average growth rate of Eritrea’s population from 2004-2011.

The last variable using a specific imputation method based on prior knowledge was the Polity variable. There are 3 variables included in the dataset that provide information on the governing structure within a country: regime type, polity, and government type. Regime type was first cited in the CIA study by Goldstone as a significant predictor of political instability [7]. Boekestein [15] used a simplified, mapped version of the original regime type variable which originally had 57 government descriptions. The simplified 3 level indicator variable developed in Boekestein [15] was also used in Shallcross [2] and Leiby [3] and is static for every year for every country in the study. There were no missing values for this variable. The original Polity variable indicated the government level based on a -10 to 10 scale with a -10 indicating fully autocratic and a 10 fully democratic and had indicators for countries with anarchies, transitioning governments, or governments experiencing a foreign interruption. Polity did contain some missing values which were necessary to fill in before creating government type which is a categorical variable that maps the Polity values. The missing values for Polity were imputed based on the countries regime type using the following mapping scheme from regime type’s three levels as seen in Table 3.

Table 3: Mapping of Regime Type to Fill in Missing Polity Values

Regime Type	Corresponding Imputed Polity Value
Central Ruling Party	-10
Emerging, Transitional, Recent Change, Disputed	0
Democratic	10

After applying the specific imputation methods to the variables previously mentioned, there were two main methods used to impute the remaining missing data values: multivariate normal imputation and univariate imputation. These two methods were employed using JMP Missing Values Utility software. The multivariate normal imputation method fills in missing values based on the multivariate normal distribution and uses least squares imputation [22]. This method only works with continuous data and is the preferred method of imputation compared to univariate imputation because it provides an estimate of the value based on the information in other rows and columns of the data rather than only considering one variable. A problem encountered when using this method was getting estimates for missing data that were not within the bounds of the original variable. For each variable requiring imputation, multivariate imputation was used first. If the imputed data did not resemble the original data, the univariate imputation method was used instead.

The univariate imputation method only considers a single column of data to impute and fills in the missing data points for that column with the mean of the original values for the variable. A summary of the imputation methods used for each variable is found in Appendix A.

After imputing all missing values with the specified imputation techniques the Kolmogorov-Smirnov (KS) Test tests how similar the imputed data is to the original data. The KS test is a goodness of fit test used to compare the distributions of a sample to some specific distribution or another sample [23]. This test was conducted in R to compare the distributions of the original data to the imputed data for each

continuous variable. The test statistic for the KS Test is based on the maximum difference between the cumulative distribution function (CDF) of each sample. Under the null hypothesis, the distributions are the same. An alpha of 0.05 was used as the significance level for this test. All but two variables, Refugee Asylum and Caloric Intake, passed the KS Test meaning that there is no statistical difference between the distributions of the of the non-imputed data and the imputed data.

3.3 Country Grouping Development

This section details how new Combatant Command groupings were developed using principal component analysis and the Modified K-means Algorithm. The current COCOMs and the new COCOM groupings developed from the methodology discussed in this section were the groups used to build conflict prediction models in the model building portion of the study.

3.3.1 Data for New COCOM Development

The data used for developing new COCOM groupings was based off of the data used in the Shallcross study [2]. The data consisted of 30 data elements which describe various political, military, economic, social, information and infrastructure characteristics of 182 countries from the years 2004-2014. For this portion of the analysis only one year of data for each of the 182 countries was used. The year 2014 was chosen as the dataset to analyze because it was the most recent year that had the most complete data. Additionally, regime type and government type were the only categorical variables in the dataset and were transformed into binary indicator variables for each level of the respective variable. Thus, the final set of variables used for principal component analysis consisted of 35 numeric variables for the year 2014 for 182 countries. The list of variables for this portion of the analysis can be found in Figure 3.

Arable Land	Fresh Water per Capita
Birth Rate	Trade (% GDP)
Death Rate	Unemployment
Fertility Rate	Polity IV
GDP Per Capita	Government Type
Improved Water	Caloric Intake
Life Expectancy	Freedom Score
Military Expend (% Gov Spending)	2 Yr Freedom Trend
Military Expend (% GDP)	3 Yr Freedom Trend
Infant Mortality rate	5 Yr Freedom Trend
Youth Bulge	Regime Type
Population density	Ethnic Diversity
Population Growth	Religious Diversity
Refugee (Asylum)	Border Conflict Score
Refugee (Origin)	2 Yr Conflict Intensity Trend

Figure 3: Variables used in PCA and Country Grouping

3.3.2 Principal Component Analysis

Principal component analysis (PCA) is a data reduction technique used to reduce the dimensionality of a larger dataset and investigate underlying relationships of the original variables. PCA uses orthogonal linear combinations of the original variables, also known as principal components, to produce a lower dimensionality set of variables that describe commonalities within the data and account for as much of the variation in the original data as possible [24, p. 20]. The principal components (PC) are extracted from the data such that the first principal component accounts for the largest amount of the total variation in the data and each subsequent principal component accounts for less variation than the previous. The equation for the m^{th} PC is found in Equation 8 [24, p. 25].

$$PC_{(m)} = w_{(m)1}X_1 + w_{(m)2}X_2 + \dots + w_{(m)p}X_p \quad (8)$$

$$X_j = \text{Observed variable } j, j = 1, 2, \dots, p$$

The weights w are chosen to maximize the ratio of variance of $PC_{(m)}$ to the total variation such that the sum of the squared weights is equal to 1 and the principal components are orthogonal [24, p. 24]. The principal components are based on the eigenvalues and eigenvectors of the correlation matrix R of the full standardized data matrix. PCA can be performed on unscaled, unstandardized data, but may give misleading results as to which variables have the most impact on the variation in the data. Because the variables in the dataset have different scales, the data was standardized at the beginning of the PCA to ensure that all the data was measured on the same scale and could be compared on a common scale. The standard procedure for calculating PCs is as follows:

1. Calculate the correlation matrix R of the data matrix X where X is an $n \times p$ matrix where n = number of observations and p = number of variables.
2. Calculate the eigenvalues and eigenvectors from R .
3. Use eigenvectors to calculate the Loadings Matrix, L , and determine number of components to retain.
4. Calculate scores for each observation for each component using Loadings matrix.

The maximum number of components calculated from the data is equal to the number of original variables in data. Since the goal of PCA is to reduce the number of variables needed to explain the variance in the data, a subset of principal components are used while still maximizing the variance in the original data. A dimensionality assessment is conducted to reduce the number of components used in the analysis while still retaining as much original variation as possible in the data. Horn's Test is

one such dimensionality assessment where the calculated eigenvalues from the data are plotted in descending order of magnitude along with a theoretical line based on normally and independently distributed random variates from a population whose correlation structure is characterized by an identity matrix [24, p. 49]. A Horn's Curve plot consists of plotting the theoretical and actual eigenvalues versus the number of components. In Horn's Test, the criterion for the number of principal components to retain is the number of components at the point where the theoretical and actual lines intersect. An example of a Horn's Curve plot is shown in Figure 4.

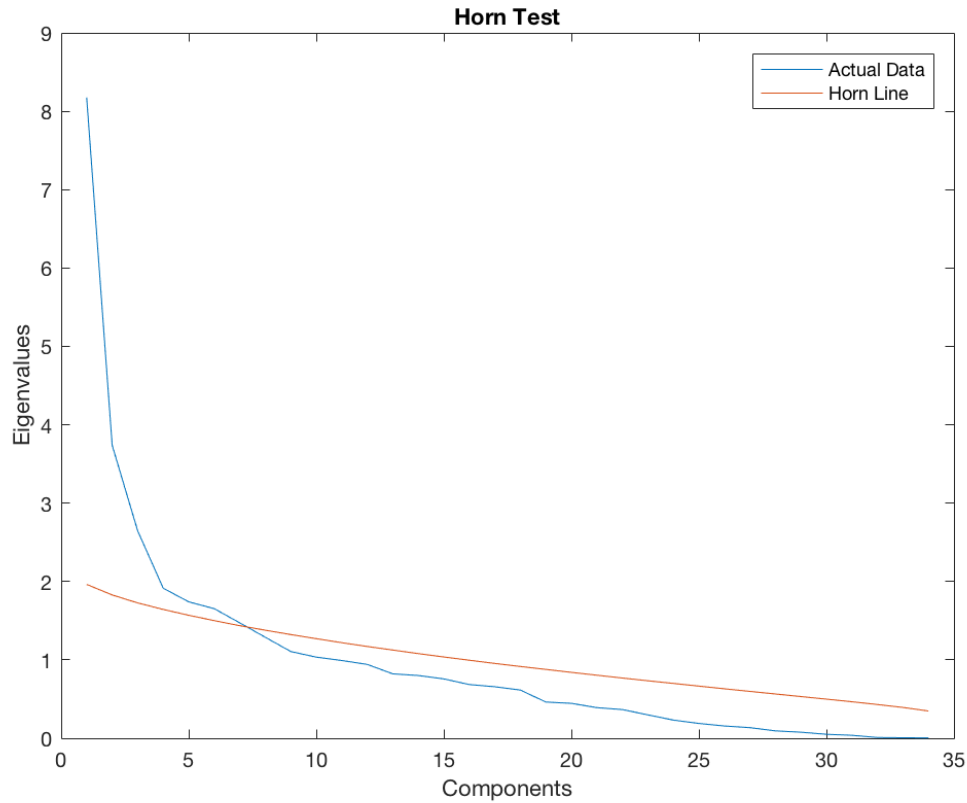


Figure 4: Example Horn's Curve Plot

The red line in the plot represents the theoretical data and the blue line represents the actual observed data. The point at which they cross is around 7 components. Because the actual data line remains close to the theoretical line until about 9 com-

ponents, additional components can arguably be retained. Nine principal components are retained.

Next, the Loadings matrix is calculated. The Loadings matrix describes the correlation between the variables and each component [24]. The Loadings matrix is calculated through the following equations shown in Equation 9.

$$\begin{aligned}
 L &= D^{-1/2} A_R \Lambda_R^{1/2} \\
 D &= \text{Diagonal matrix of } R \\
 A_R &= \text{Eigenvectors of } R \\
 \Lambda_R &= \text{Eigenvalues of } R
 \end{aligned}
 \tag{9}$$

The Loadings matrix explains the correlation between the variables within each component and gives a sense as to which variables have the greatest influence on each component [24]. This technique can be informative in explaining what is common among groups of observations and explaining the variance within large datasets. By highlighting the variables that have the highest correlation in each of the components, insight can be gained as to what aspect of the data each principal component describes.

The final portion of PCA consists of calculating and plotting the principal component scores for each observation. A score is calculated for each observation for each of the retained principal components. These scores are obtained by multiplying the linear combination of the original variables to the appropriate data for each observation. The score matrix S contains the score for each observation for each principal component. The calculation of the score matrix S is shown in Equation 10.

$$S = X_s D^{-1/2} A_R$$

$$X_s = \text{Standardized data matrix} \tag{10}$$

$$D = \text{Diagonal matrix of R}$$

$$A_R = \text{Eigenvectors of R}$$

The component scores encompass most of the information from the original dataset and can be used to compare how similar observations are based on the data. Plotting the scores in different pairings allows a visual comparison of observations. A score plot is produced by plotting two component scores for each observation as the x and y coordinates in a 2-dimensional plot.

Even though this technique is not explicitly a classification technique, sometimes patterns exist within the data which can be visualized in the plots of principal component scores. Plots of combinations of two principal components scores for each observation are created to analyze similarities between countries under the current Combatant Command structure. The color and shape of the data points represent the group a country is a part of, in this case COCOM. Visually inspecting the score plot, provides an indication of any patterns or similarities between countries based on data. The results from the principal component analysis were used in the implementation of the Modified K-means Algorithm.

3.3.3 Modified K-means Algorithm

Conducting principal component analysis reduces a highly-dimensional data set into a smaller and easier to visualize dataset and compares observations on a common scale. The score plots present clusters of similar observations based on the groupings and the ability to perform cluster analysis on the data. There are several different ways to conduct cluster analysis depending on what assumptions are made about the

data and the desired outcome from the analysis. K-means clustering was the chosen method for cluster analysis and inspired the development of a Modified K-means Algorithm.

K-means clustering partitions observations within a data set into K distinct and mutually exclusive clusters [25, p. 387]. The objective of K-means clustering is to partition the data to minimize the within cluster variation [25, p. 387]. In K-means clustering the number of clusters, K, is assumed known and predefined at the beginning of the cluster analysis.

There are several ways to define and compare variation between observations. The most common way suggested in *An Introduction to Statistical Learning* is using the squared Euclidean distance between observations [25]. The scores from principal component analysis give a means to compare the observations based on data characteristics on a common scale. The difference between component scores between two observations can be compared to determine how similar the observations are to one another. The distance between the scores can be thought of as a measure of how similar the two observations are and a distance can be calculated between the two points. The squared Euclidean distance between observations was chosen as the way to calculate the distance between two points with x and y coordinates (one principal component score for the x and one for the y) for each observation. The Euclidean distance between observations on a score plot was the comparison metric chosen to determine how similar observations were to each other. The objective function for solving the K-means clustering problem using the Euclidean distance between observations as the comparison metric is shown in Equation 11.

$$\min_{C_1 \dots C_k} \left\{ \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \right\} \quad (11)$$

In Equation 11, $|C_k|$ is the number of observations in the k th cluster and the

Euclidean distance between an observation i and every other observation i' that is in the cluster is calculated by $(x_{ij} - x_{i'j})^2$ [25].

The large number of possible ways to partition n observations into K clusters makes this a very large and difficult problem to solve. *An Introduction to Statistical Learning* proposes a simple and elegant approach to solve this problem for a local optimum with the K-means algorithm [25, p. 388]. A summary of the K-means clustering algorithm proposed in the text can be seen in Algorithm 1.

Algorithm 1 Pseudocode for the K-means Clustering Algorithm

Determine k number of clusters in data

Assign each observation a number 1 to k to serve as initial cluster assignment for each observation

while cluster assignments change **do**

- a: Calculate the centroid of each cluster 1 to k . The centroid of the k th cluster is defined as the vector of means for each feature for observations in the k th cluster
 - b: Assign each observation to the closest centroid as determined using the Euclidean distance between each observation and each cluster centroid.
-

This algorithm is based on the assumption that there exist K mutually exclusive clusters in the data, with unknown cluster membership. Another assumption is that the closer the distance between two observations, the more similar the two observations are to one another. A third assumption is that all observations are assigned to a cluster. One limitation in this algorithm is that the algorithm finds a local optimum, not a global optimum, and the results from the algorithm depend on the initial setting of the clusters.

The K-means algorithm provides a way to group observations based on data similarities. In this study the observations are individual countries and the data represent various political, military, economic, social, infrastructure, and information character-

istics of a country. The analysis needs to compare countries based on data similarity and location. To develop new COCOM groupings of countries location is used in a clustering algorithm to ensure that the new COCOMs would be mostly geographically clustered and somewhat similar to the current structure. A modified k-means clustering algorithm, called the Modified K-means Algorithm, incorporates both data characteristics and geographic proximity in grouping similar countries together. Comparing countries' similarities involved PC scores 1 and 2 for each country. Comparing the geographic proximity of each country to one another used the Latitude and Longitudinal coordinates for each country's capital city. The geographic similarity metric was calculated using the Great Circle distance between countries' capital cities. This distance was normalized and scaled to coincide with the scale of the principal component scores to combine the data and geographic similarity metrics into a single, total distance or similarity metric. The pseudo-code for the Modified K-means Algorithm, which clusters similar countries together based on data similarity and geography, is described in Algorithm 2.

Applying the Modified K-means Algorithm to the 2014 data for the selected 182 countries resulted in groups of countries that were similar based on data and geography. The groupings for the conflict probability models were developed from the current COCOM structure and the results obtained from the Modified K-means Algorithm.

3.4 Model Building Procedure

The goal of many model building procedures is to build the most parsimonious model that accurately reflects the true outcome experience of the data [18]. This study uses the purposeful selection of covariates (PSC) model building method as proposed by Hosmer, Lemeshow, and Sturdivant in *Applied Logistic Regression* [18].

Algorithm 2 Pseudocode for the Modified K-means Algorithm

```
Perform PCA on Data
Plot PC1 vs. PC2 Scores for each country
Define Initial Groupings (Current Combatant Commands)
procedure MODIFIED K-MEANS
  while total number of cluster changes > 0 do
    Calculate Data Centroid for each cluster 1 to  $k$ 
    Calculate Location Centroid for each cluster 1 to  $k$ 
    Calculate Euclidean Distance between each observation and current cluster's
    Data Centroid
    Calculate normalized, scaled Great Circle distance between each country's
    capital city and current cluster's Location Centroid
    Calculate total weighted distance from Location and Data centroids for each
    observation
     $maxdist \leftarrow$  max distance any one observation in a cluster is from current
    cluster's centroids
     $candidatepoints \leftarrow$  observations  $\geq 2/3 * maxdist$  away from current cluster's
    centroids
    for each candidatepoint do
      Calculate total weighted distance to each cluster centroid
      Assign candidatepoint to nearest cluster
    Sum total number cluster changes
```

The purposeful selection of covariates method is a seven-step model building process which starts very broad with possible variables to be included in the model and takes steps to narrow down the most important variables to be included in the model. Figure 5 gives an overview of each step in PSC.

This study developed a general script in R as an outline for performing each step of the PSC model building procedure. Each model was built individually in a unique application of PSC, but in general the models were built following the same general 7 step process outlined in Figure 5.

Step one of the PSC model building strategy is to select variables significant in a univariate analysis. Each variable is fitted in a univariate model with the dependent variable. Any variable whose LRT p-value is less than 0.25 is selected as a candidate variable for step 2. Categorical variables with k levels must be analyzed together by

Purposeful Selection of Covariates Overview	
Step 1: Univariate Analysis	Fit univariate models for each potential variable, identify variables whose Chis-squared Test Statistic p-value is <0.25
Step 2: Multivariate Model	Fit all candidate variables identified in step 1 to a multivariate model(full model). Take out variables with the largest insignificant p-value one at a time until all variables are significant in the multivariate model to obtain the reduced model
Step 3: Check Change in Coefficients	Compare values of estimated coefficients in the reduced model to the corresponding coefficient in the ful multivariate model. Identify any change in coefficients that are greater than 20% and add in variables that were taken out in step 2
Step 4: Add in variables from step 1	Add variables not selected in step 1 back into the model one at a time and see if the variable is significant
Step 5: Check Linearity Assumptions	Check that the logit function increases/decrease linearly as a function of the covariates. Regress predicted log odds vs a covariate for all covariates
Step 6: Interaction Terms	Determine if interaction terms between variables in the model should be included based on statistical and practical evidence. Add interaction term if significant at 0.01 level and practically makes sense.
Step 7: Model Adequacy Assessment	Check adequacy and performance of the model with the Hosmer-Lemeshow Test, Classification Tables, and Reciever Operator Curves/Area Under the Curve

Figure 5: Purposeful Selection of Covariates Summary

fitting a model with $k - 1$ categories and determining the overall significance of the all the variables with the LRT. The 0.25 p-value significance threshold can be adjusted either higher or lower, but Hosmer, Lemeshow, and Sturdivant suggest keeping the threshold between 0.20 and 0.25. Using higher levels of a p-value threshold produce the risk of including variables that may not be important in the analysis [18].

Step 2 of PSC begins by fitting a multivariate model, referred to as the full model, with all the covariates that were identified as significant in step 1. After fitting this full model, each variable is assessed for its significance in the model. Individual variables are removed one at a time starting with the variable that is the most insignificant (largest Wald Statistic p-value) until all variables in the model are significant at a 0.05 level. The model produced after taking out all insignificant variables is referred to as the reduced model. The reduced model is then compared to the full model with a partial likelihood ratio test. The results of the partial likelihood ratio test determine if there is a statistically significant difference between the full model and the reduced model. Failing to reject the null hypothesis means there is no significant difference between the models and therefore favor the more parsimonious model of the two.

After fitting an appropriate reduced multivariate model, step 3 compares the estimated coefficients of each covariate in the reduced model to their estimated coefficient in the full multivariate model. The percent change in the coefficients, referred to as the $\Delta\hat{\beta}\%$, is calculated according to Equation 12 for a given parameter p which was presented by Hosmer, Lemeshow, and Sturdivant [18].

$$\Delta\hat{\beta}\% = 100 \frac{\hat{\theta}_p - \hat{\beta}_p}{\hat{\beta}_p} \text{ where} \quad (12)$$

$\hat{\theta}_p$ = Coefficient for variable p from the reduced multivariate model

$\hat{\beta}_p$ = Coefficient for variable p from the full multivariate model

Of concern is $\Delta\hat{\beta}\% \geq 25\%$. A $\Delta\hat{\beta}\% \geq 25\%$ indicates that one or more of the variables that were taken out during step 2 are important in the sense of providing an adjustment to the effect of the variables in the model. During this step variables that were taken out during step 2 are added back into the model to provide the necessary adjustment to the significant covariates' estimated parameters. Variables were added one at a time and $\Delta\hat{\beta}\%$ recalculated with each change in variables included in the

model. Sometimes the only way all $\Delta\hat{\beta}\%$ would be less than the 25% threshold is to add back in all the variables from step 2. To balance parsimony of the model and proper adjustments of the parameters, consider adding in variables that significantly reduced the the change in $\hat{\beta}$. If more than 5 variables were added and the $\Delta\hat{\beta}\%$ were still greater than 25%, continue analysis with the reduced model obtained at the end of step 2 in favor of a more parsimonious model. Compare the models with additional variables to the reduced model from step 2 using a partial likelihood ratio test and if there was no statistically significant difference between the models, take the model with less variables as our model from step 3.

In step 4, add in each variable not significant in the step 1 univariate analysis back into the model obtained at the conclusion of step 3. The variables are added into the model one at a time and evaluated on their significance using the Wald statistic p-value. This step is meant to identify variables that are not significant by themselves, but are influential in the the presence of other covariates [18]. The model produced at the end of this step is referred to as the preliminary main effects model.

Having obtained the preliminary main effects model, check the linearity assumptions of the variables in the model. In step 5, check the assumption that the logit function increases or decreases linearly as a function of the covariate. In this study, the linearity assumption was checked using smoothed scatter plots. In these scatter smooth plots, the log-odds of the outcome Y are plotted against the values of covariate X and should show a linear trend if the linearity assumption is held. The smooth scatter plots help to identify when the linearity assumptions are not held and identify when a covariate may require a transformation. The scatter plots can also be used to identify potential outliers in the data that may be skewing the plots. The model obtained at the end of step 5 is referred to as the main effects model.

From the main effects model, check for interactions between variables in the model.

An interaction between two variables implies that the effect is not constant over levels of the other variable [18]. The final determination on whether or not a variable was included in the model was based on both statistical significance and plausibility of the interaction term. The interaction terms were identified as statistically significant if their Wald statistic p-value was below 0.01. Any statistically significant interactions are also reviewed as plausible interactions. Each significant interaction's plausibility is based on prior knowledge and a realistic possibility of the variables interacting with one another. If an interaction term was found to be both statistically significant and a realistic interaction possibility, the interaction term was added to the model. If more than one interaction term was identified as significant then all interaction terms were added into the model and possibly removed one at a time using the same procedure in step 2. At this point, only insignificant interaction terms were removed from the model and all other terms were considered locked regardless of their p-value.

The final step before establishing a model as a final model is to assess the adequacy of the model. This study assessed model adequacy using several methods.

3.5 Assessing Model Adequacy

Assessing the adequacy of a model before using the model for interpretation is the final step in the purposeful selection of covariates model building process. This study uses two tests to assess the adequacy of the models. The first test we utilize is the Hosmer-Lemeshow test. In this test, the range of outcomes are grouped into a number of bins [18]. The actual number of observations in each bin are compared to the expected number of observations in each bin. The bins can be determined either by the percentiles of the estimated probabilities or by fixed intervals of the estimated probabilities [18]. The Hosmer-Lemeshow test statistic $\hat{C}$ is calculated by summing the differences between the number of observed values and expected

values for each level 0 and 1 over the total number of bins, g . Equation 13a shows the full calculation of the Hosmer-Lemeshow statistic and the algebraically simplified version. In these equations o_{ik} = the number of observed values i in group k , e_{ik} = the number of expected observations for value i in group k , and n_k is the number of observations in group k . Finally, c_k is the number of unique covariate patterns in the k th group and m_j is the total number of observations that have covariate pattern j . In the data, there are as many covariate patterns as there are observations and therefore, m_j is set equal to 1.

$$\hat{C} = \sum_{k=1}^g \left[\frac{(o_{1k} - \hat{e}_{1k})^2}{\hat{e}_{1k}} + \frac{(o_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}} \right] = \sum_{k=1}^g \left[\frac{o_{1k} - n_k \bar{\pi}_j}{n_k \bar{\pi}_j (i - \bar{\pi}_j)} \right] \quad (13a)$$

$$o_{1k} = \sum_{j=1}^{c_k} [y_j] \quad (13b)$$

$$o_{0k} = \sum_{j=1}^{c_k} [m_j - y_j] \quad (13c)$$

$$e_{1k} = \sum_{j=1}^{c_k} [m_j \hat{\pi}_j] \quad (13d)$$

$$e_{0k} = \sum_{j=1}^{c_k} [m_j (1 - \hat{\pi}_j)] \quad (13e)$$

$$\bar{\pi}_j = \frac{1}{n_k} \sum_{j=1}^{c_k} [m_j \hat{\pi}_j] \quad (13f)$$

The null hypothesis in this test is that the model fits adequately. Through simulations, Hosmer and Lemeshow demonstrated that when the number of covariate patterns, J , is equal to the number of observations, n , the Hosmer-Lemeshow statistic is well approximated by the chi-square distribution with $g - 2$ degrees of freedom [18]. In this study, a model was deemed to fit the data adequately if the p-value was greater than 0.05.

The second method used to check model fit and adequacy is through the analysis of the classification capabilities of our models. The two approaches in this study are classification tables and receiver operating characteristic (ROC) curves. Classification tables are built by comparing the actual outcomes of each observation to the predicted outcomes. Classification tables are created by cross classifying the actual outcome y with a binary representation of the estimated probabilities [18]. The binary representation of the estimated probabilities is obtained by selecting a cutoff point, c , and assigning a predicted outcome according to the Equation 14.

$$\text{predicted outcome} = \hat{p} = \begin{cases} 0 & \text{if } \hat{y} < c \\ 1 & \text{if } \hat{y} \geq c \end{cases} \quad (14)$$

A typical choice for a cutoff point is $c = 0.5$, but the cutoff point can be varied depending on the focus of the discrimination. Table 4 gives a representation of a classification table and the location of where to find the number of observations that were correctly and incorrectly classified based on a specified cutoff point.

Table 4: Example Classification Table

		Actual Outcome	
		0	1
Predicted Outcome	0	True Negatives	False Negatives
	1	False Positives	True Positives

A single classification table provides the sensitivity, specificity, and overall accuracy of the model given a set cutoff point. Sensitivity measures how well the model can detect the true positives in the data, those observations that are predicted to be 1 and are actually observed to be 1. Specificity measures how well the model can detect true negatives, those observations predicted as a 0 and are actually a 0. The overall accuracy calculates the total number of observations that were correctly predicted.

Equation 15 shows how each of these values are calculated using the values in the classification table.

$$\text{Sensitivity} = \frac{\# \text{ of true positives}}{\text{total } \# \text{ actual positive observations}} \quad (15a)$$

$$\text{Specificity} = \frac{\# \text{ of true negatives}}{\text{total } \# \text{ actual negative observations}} \quad (15b)$$

$$\text{Overall Accuracy} = \frac{\text{true negatives} + \text{true positives}}{\text{total observations}} \quad (15c)$$

Using the information from classification tables considers how well the model discriminates over a range of cutoff points. Plotting sensitivity and specificity over a range of cutoff points gives an idea as to how well the model discriminates outcomes, in this case transitions and non-transitions. The point at which the sensitivity and specificity intersect is sometimes referred to as the “optimal” cut point for classification because it is the point at which sensitivity and specificity are maximized [18].

The receiver operating characteristic (ROC) curve shows how the model detects the true signal in the presence of noise [18]. Plotting sensitivity versus 1-specificity for a range of possible cutoffs produces the ROC curve. The area under the ROC curve, or AUC, helps determine how well the model is able to discriminate across the range of cutoff values. There is no real measure of how well a model discriminates, but Hosmer, Lemeshow, and Sturdivant outline a guideline for how well a model discriminates based on its AUC as seen in Table 5 [18].

Once a model adequacy assessment is complete the models help gain insight into important factors that influence conflict transitions and predict when conflict transitions may occur.

Table 5: Discrimination Rating Guidelines[18]

AUC	Discrimination Rating
AUC=0.5	No discrimination
$0.5 < \text{AUC} < 0.7$	Poor discrimination
$0.7 \leq \text{AUC} < 0.8$	Acceptable discrimination
$0.8 \leq \text{AUC} < 0.9$	Excellent discrimination
$\text{AUC} \geq 0.9$	Outstanding discrimination

3.6 Summary

This chapter covered the methodology employed to develop the dataset used in the study, the development of 6 new COCOM groupings, and the model building strategies and assessments used to create and test the prediction models for each COCOM grouping.

IV. Analysis and Results

4.1 Overview

This chapter discusses the results and analysis of applying the methodology outlined in Chapter 3. Section 4.2 discusses the results from the principal component analysis and applying the Modified K-means Algorithm to obtain new COCOM groupings. Section 4.3 discusses the models built for each COCOM and new COCOM and compares the suites of models.

4.2 New Combatant Command Groupings Formation

The new COCOM groupings were formed through the application of principal component analysis and the Modified K-means Algorithm. The results from PCA provided a measure of similarity between countries based on the data elements and were applied in the Modified K-means algorithm to cluster similar countries together based on both data and geography. The results from both the PCA and clustering analysis are detailed in the remainder of this section.

Principal Component Analysis Results

Principal component analysis was conducted on the 30 data elements shown in Figure 3. The goal with running PCA on the country data was to elicit possible similarities in the data and produce a smaller set of data that encompasses the same information from the larger dataset, but in a way that can be visualized and used to compare countries to one another.

The first step in conducting PCA was to calculate the eigenvalues and eigenvectors of the correlation matrix of the standardized data matrix and determine the number of components to retain in the analysis. Examining Horn's Curve in Figure 6 suggests

retaining at least 7 components. Because of uncertainty with this choice, 9 principal

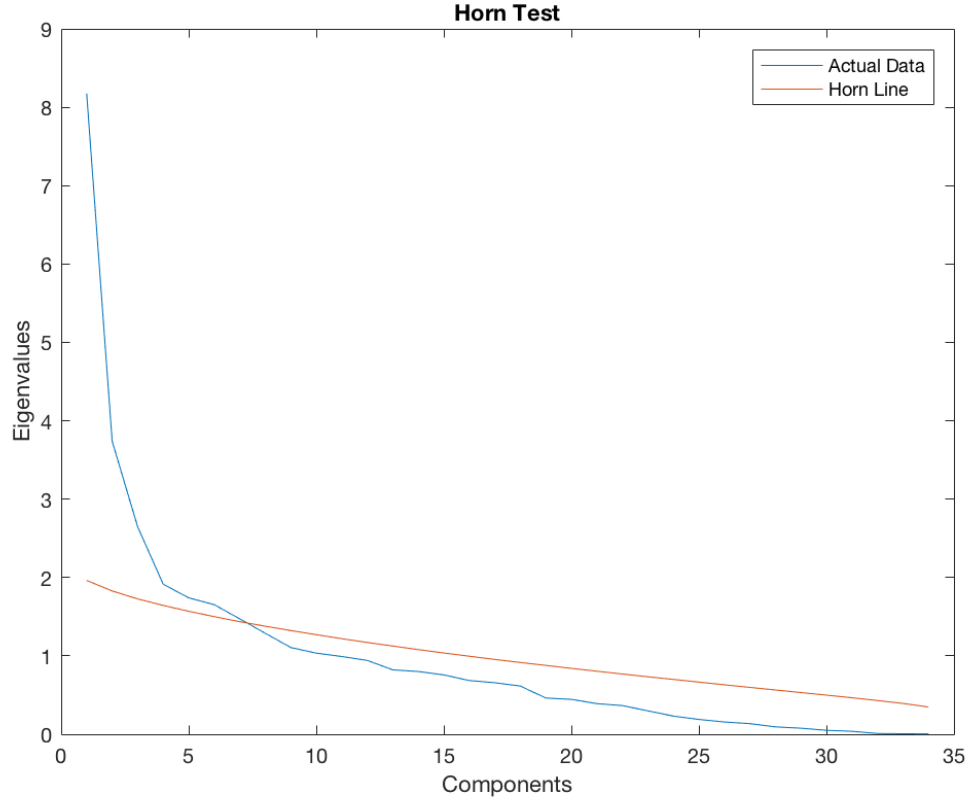


Figure 6: Horn's Curve Plot

components were used. These 9 principal components account for about 70% of total variation in the original dataset.

Variables highly correlated with each principal component provide insight regarding what aspect of the data the principal component describes. The loadings matrix for the principal components calculates the correlation between each variable and each principal component. The heat map of the correlation between each variable and each principal component is shown in Figure 7.

The most positive correlations are indicated by dark red and the most negative correlations are indicated by the dark blue. For each variable, the principal component with the greatest absolute value correlation best describes the data. The description

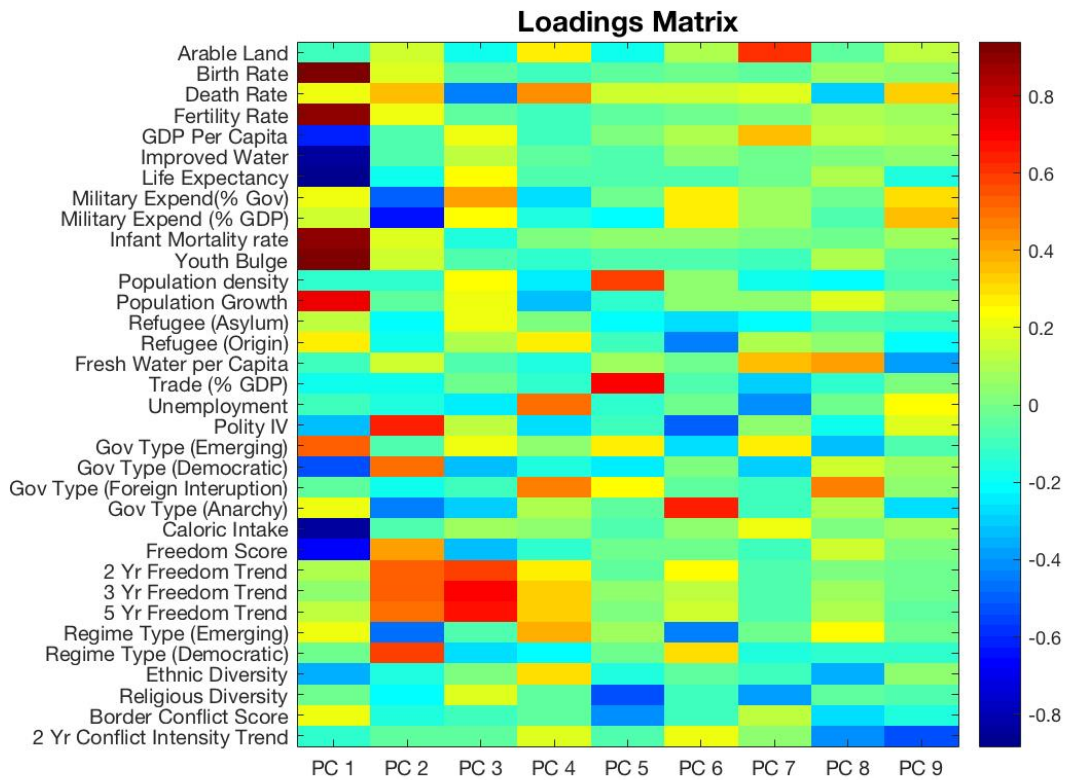


Figure 7: Heat Map of Correlations between Variables and Principal Components

for each principal component and the amount of variation it accounts for in the data can be found in Table 6. Principal components 1 and 2 account each account for more than 10% of the variation in the data individually and together account for the most variation in the data with 35% total variation of the original data.

The next step in the PCA study is to plot combinations of principal component scores to better understand how similar the countries are to one another based on data. Each country received a score for each principal component and these scores were plotted in combinations of two to produce score plots. Additionally, each country was color and shape coded in the plots according to the Combatant Command they are currently assigned to. Color coding the points allows a comparison on how clustered the groups of countries were based on their principal component scores. In other

Table 6: Principal Components Descriptions and Variance

Principal Component	Description	Percent Variation
PC1	Quality of Life	24.0%
PC2	Military and Government	11.0%
PC3	Freedom	7.8%
PC4	Unemployment	5.6%
PC5	Trade and Religious Diversity	5.1%
PC6	Anarchy Government	4.9%
PC7	Arable Land	4.3%
PC8	Fresh Water	3.8%
PC9	Conflict Intensity	3.3%
Total Variation		69.8%

words, these plots depict how similar countries are within each COCOM based on data.

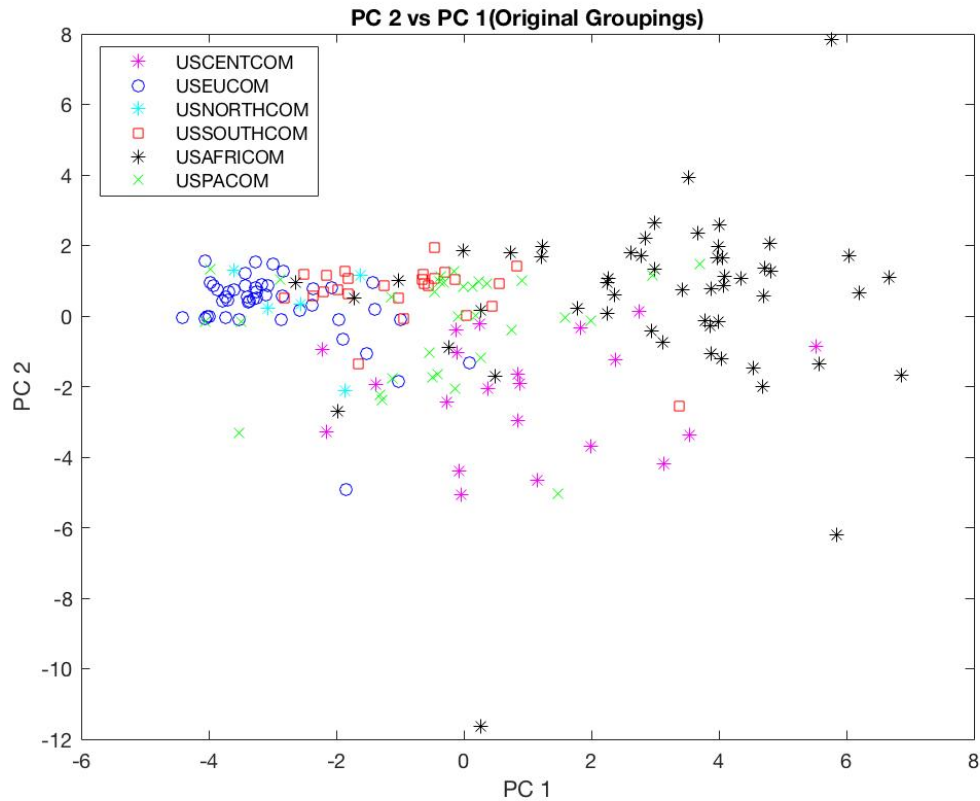


Figure 8: Military and Government (PC2) vs Quality of Life (PC1)

Plotting and considering all combinations of principal component scores for principal components 1 through 9, the plot of PC1, quality of life principal component, versus PC2, the military and government principal component, shows the most distinct clusters of countries. This plot ultimately shows how similar countries are to one another based on the data with more focus on quality of life factors and military and government characteristics of the countries.

From the plot, most of the countries in USAFRICOM are clustered together in the top right section of the plot while most of the countries in USEUCOM are in the top left of the plot. There is evidence of clusters of COCOMs based on data similarities and there is overlap between the clusters of COCOMs. This overlap indicates that some countries may be more similar to countries of a different COCOM. Because principal component 1 and principal component 2 accounted for the most variation in the data and the clusters of countries were the most distinct as compared to other PC score plots, principal component 1 and principal component 2 were chosen as the means to measure data similarity between countries. Principal component 1 scores and principal component 2 scores of each country were plotted on the X and Y axis, respectively. The Euclidean distance between each country on the score plot was used as the data similarity metric in the Modified K-means Algorithm.

Modified Grouping Results

After completing PCA and determining the data similarity metric to be used in the Modified K-means Algorithm for cluster analysis, the Modified K-means Algorithm was run to determine the new COCOM groupings for the subsequent model building portion of the study. Principal component scores for PCs 1 and 2 were used as the X and Y coordinates in calculating the Euclidean distance between each country for the data similarity metric. The geographic proximity metric used the Latitude and

Longitude of the capital city of each country and calculated the Great Circle distance between Latitude and Longitude coordinates. The distance metric was normalized and scaled to coincide with the ranges of the principal component scores. This created a total distance metric to measure how similar each country was to every other country in terms of both data and geography. The total distance metric was the weighted sum of the PC scores distances and the normalized, scaled geographic distance. The Modified K-means Algorithm was run weighting the amount that data similarity and geographic proximity influenced the total distance with weights w and $1 - w$ respectively, varying w from 1 to 0 in increments of 0.1. A weight $w = 1$ is interpreted as considering only the data aspect of countries in the cluster analysis, while $w = 0$ is a geographic proximity only cluster analysis. The results from a data only cluster analysis are displayed in Figure 9.

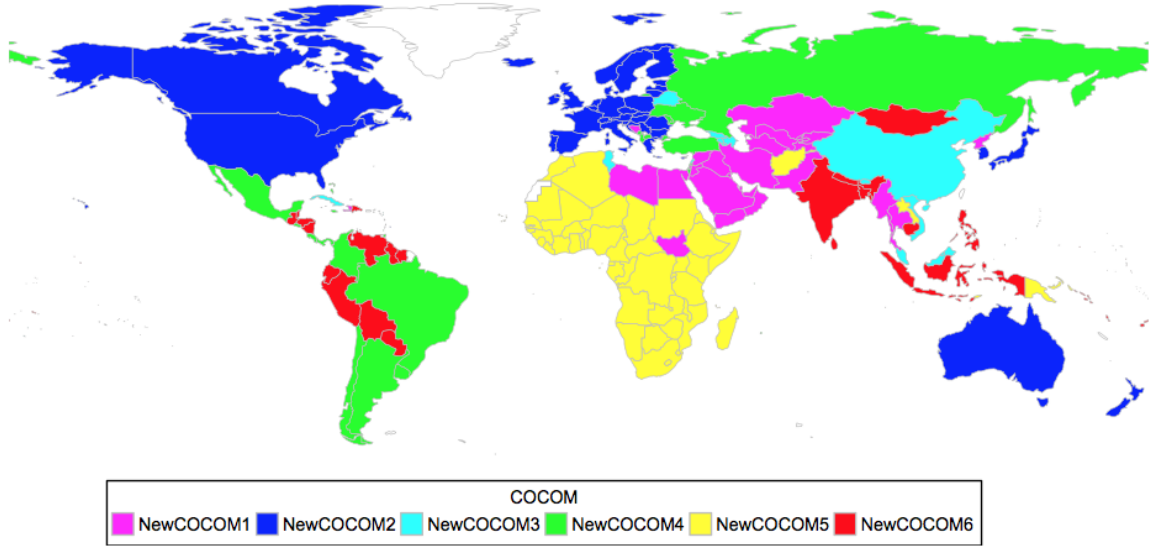


Figure 9: Data Similarity Only Groupings

The resulting cluster assignments account for data similarity only and do not consider geographic proximity as part of the calculation in how close each country is to every other country. As seen in Figure 9 the resulting groupings are scattered

throughout the world and do not cluster geographically. From a military operations perspective, these groupings do not make sense and would be difficult to reasonably conduct military operations in the different regions of the world in terms of logistics and general understanding of each Combatant Commander's area of responsibility. One important insight gained from the data only cluster analysis was that when only considering data similarities, the United States, Canada, Australia, South Korea, and the Western European countries group together. These results are consistent with previous analyses that have grouped these countries together as they are all members of the Organisation for Economic Co-operation and Development (OECD) and identified as countries who share similarities in terms of economic status and overall development.

While the data only results are scattered and not clustered together geographically, the PC score plots show how they are clustered according to data similarity. Considering only data in the Modified K-means Algorithm resulted in the most distinct groupings of countries with the least amount of overlap shown in Figure 10. This result is not surprising as the goal in the Modified K-means Algorithm is to minimize the distances between countries within in cluster based on each country's PC 1 and PC 2 scores. Without the geographic element, the results suggest that the current structure of the COCOMs does not always group similar countries together based on characteristics observed in data.

Clustering solely on data similarity will not group countries together in a mostly contiguous fashion. The next step is to determine how to balance incorporating data similarities and maintaining a reasonable geographic structure for the proposed new groupings. Decreasing the weight of the data similarity metric and increasing the weight of the geographic location metric in the total distance metric determined how much influence each metric had in clustering countries together in the Modified

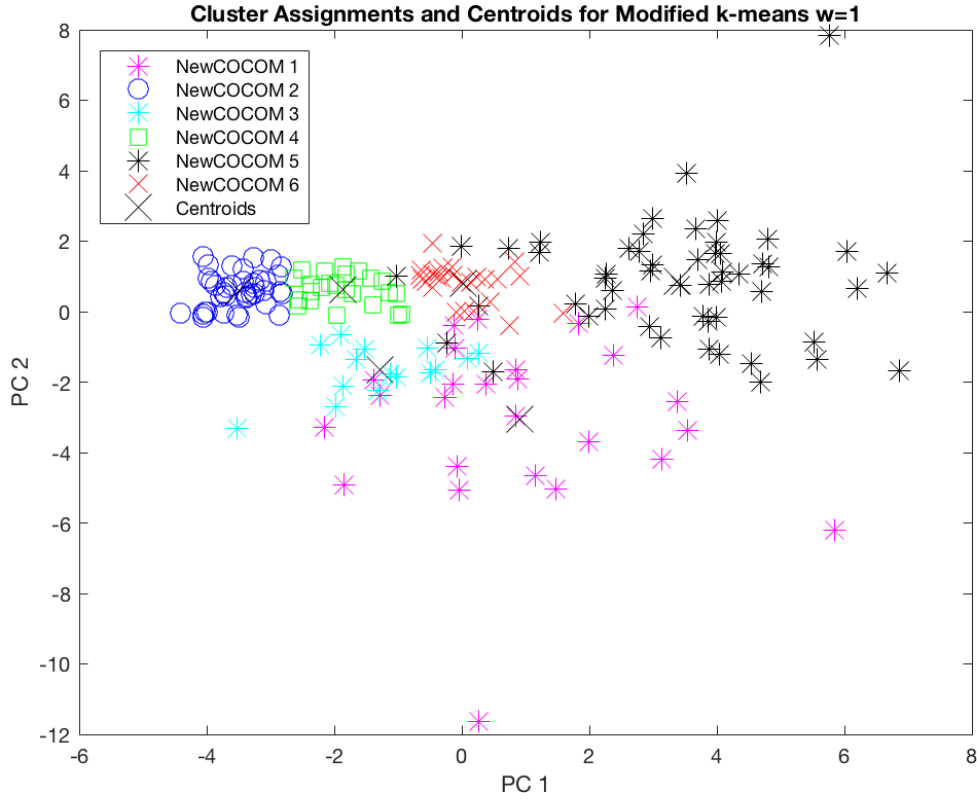


Figure 10: Score Plot with Data Similarity Only

K-means Algorithm. The next weighting structure considered equally weighted the influence of the data similarity metric and the geography metric. Figure 11 and Figure 12 show the results of equally weighting data similarity and geographic proximity in the Modified K-means Algorithm.

In Figure 11 the groupings of countries are more geographically clustered and there exist more contiguous groups of countries within each cluster. These groupings make more sense from a military perspective because they are more geographically clustered and the groups could more reasonably be designated as areas of responsibility for Combatant Commanders. While these new groupings appear similar to the current structure, they do differ from the current Combatant Command structure. Furthermore, while the countries are more geographically clustered, there is a

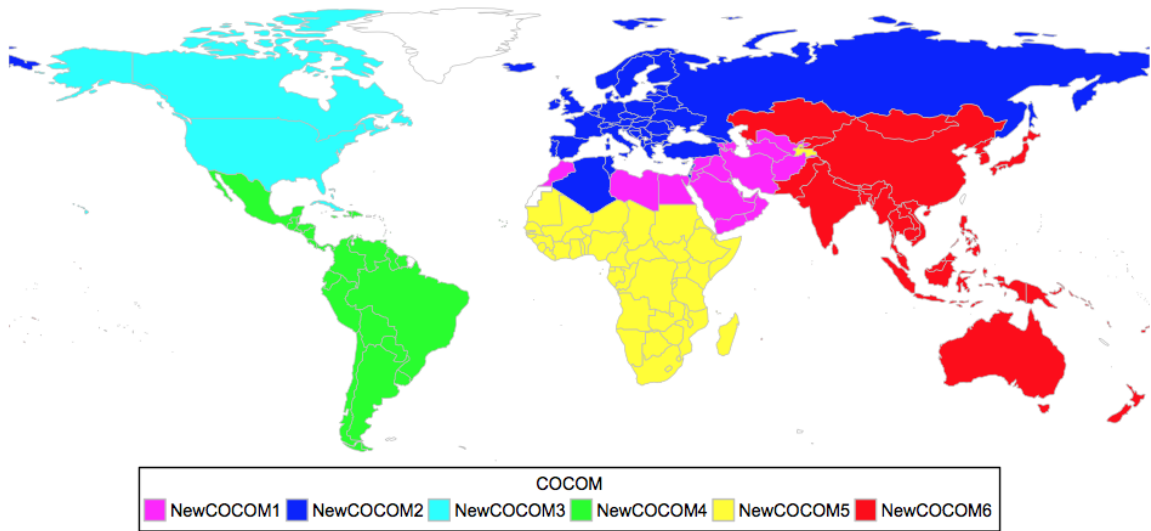


Figure 11: Data Similarity and Geography Groupings

trade-off in incorporating an element of geographic similarity into consideration for clustering countries together. This trade-off is observed in the PC score plots. From Figure 12 it is apparent that the clusters are less distinct than the data only score plot. There still exists large clusters of similar countries, but there is more overlap between the groupings than in the data only score plot. The results are as expected and confirm that adding in an element of geographic proximity into the total similarity metric does influence the clustering of countries and not all countries that are close in terms of geography are necessarily similar in terms of the data characteristics associated with a country. While these groupings make more sense from a military perspective and could be more reasonably considered as a possible way to structure the COCOMs, one final element of distinction between countries was added to the final groupings.

In previous models, OECD countries have been grouped and modeled together based on their similarity in economic status and overall development. As seen in the data only clusters, these countries prove to be similar from a data perspective, but are

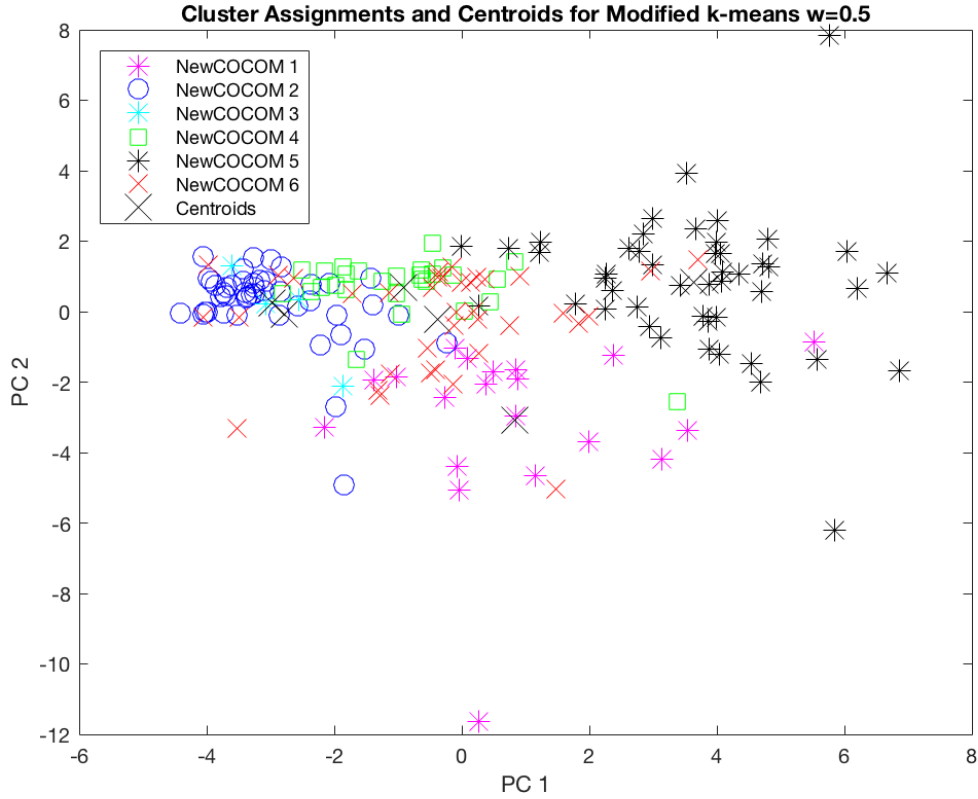


Figure 12: Score Plot with Data Similarity and Geography Equally Weighted

in different regions of the world. Final groupings should incorporate the similarity of OECD countries despite their geographic distances between the countries. The final groupings used the groupings from the Modified K-means Algorithm with equal weights on data similarity and geographic proximity as the baseline assignments. Any USEUCOM country that was a member of the OECD was assigned to the same grouping as the United States and Canada. The final groupings that are referred to as the New COCOMs in the modeling section are seen in Figure 13.

These groupings maintain a relatively strong clustering in terms of contiguous groups of nations within each new COCOM grouping. Furthermore, the PC score plot in Figure 14 still shows relatively distinct groupings and shows some distinction between the Western European OECD countries incorporated in New COCOM 3

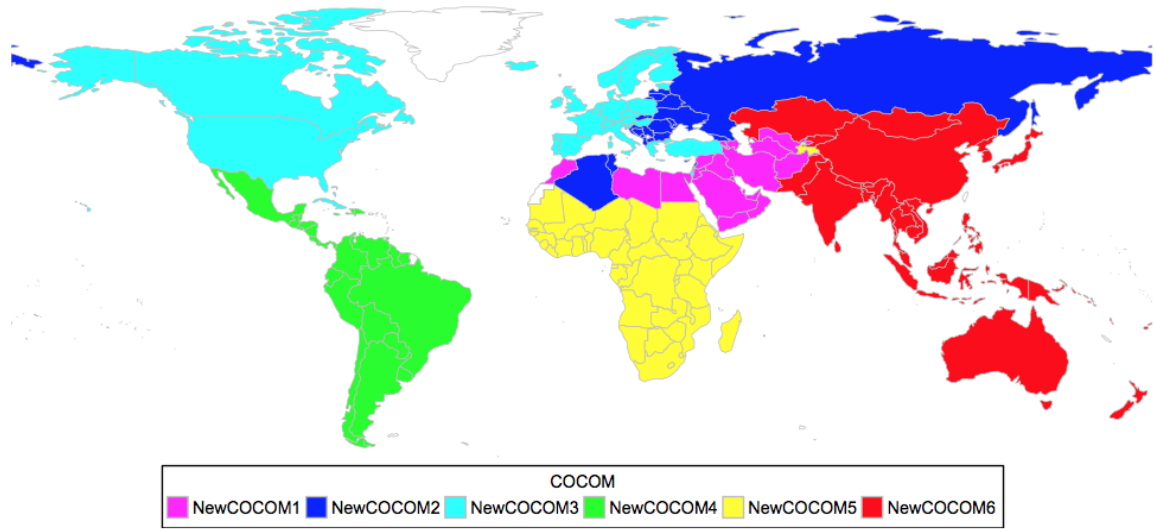


Figure 13: Map of Final Groupings

and the Eastern European countries in New COCOM 2. The countries that are not necessarily incorporated into contiguous groupings are Algeria, Tunisia, Morocco, and Tajikistan. Even though the groupings are not consistently adjoined together in one grouping, those countries were kept in the new COCOM that they were assigned so as to minimize any changes from the results of the Modified K-means Algorithm.

The results from employing the Modified K-means Algorithm with varied weights on the influence of data similarity and geographic proximity metrics and considering the similarity of OECD countries despite geographic distance resulted in our final new COCOM groupings. These new COCOM groupings were the groupings used to develop the new conflict prediction models. The models for the new COCOM groupings which incorporate data similarity, geographic proximity, and OECD countries, were compared to the models of the current COCOM groupings.

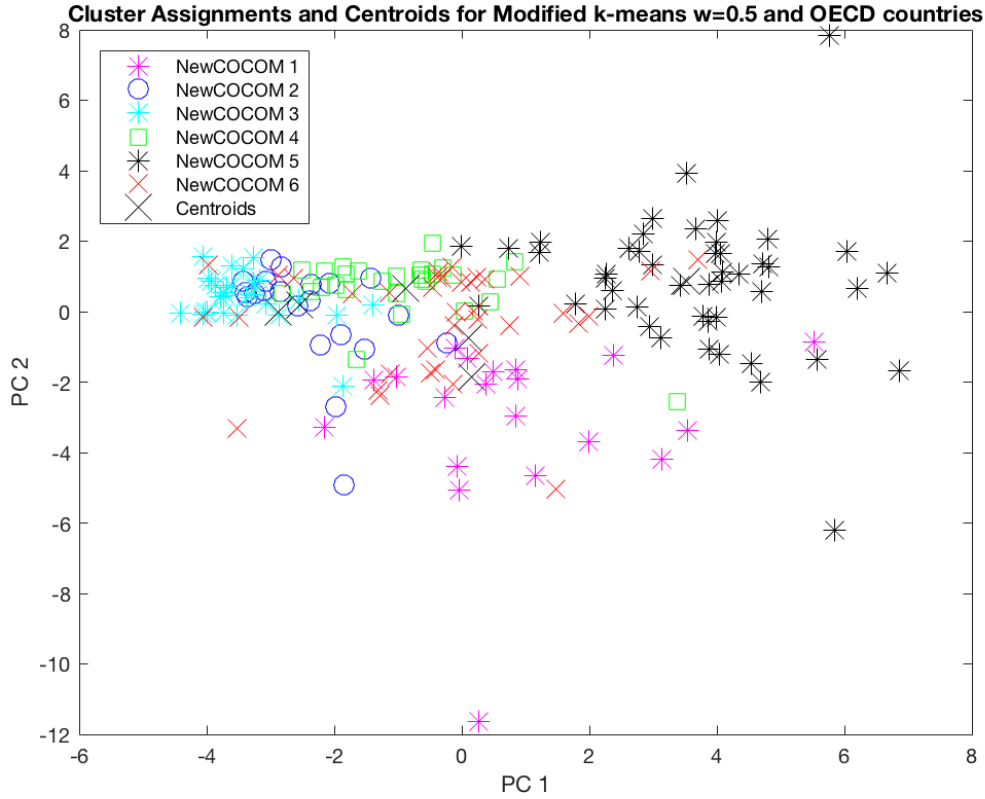


Figure 14: Score Plot for Final Groupings

4.3 Model Results

Two suites of models were built and compared in this study. For each suite, 12 conditional logistic regression models were built to predict the probability of a country transitioning into or out of a state of conflict given its current conflict status. A total of 24 models were built using the Purposeful Selection of Covariates method previously discussed in the methodology section. The data used to build the models was the data for the years 2004-2008 and 2010-2014. The data for the years 2009 and 2015 was used as the validation data.

The first suite of models built in this study were conditional logistic regression models for the current COCOM structure with one change to the USNORTHCOM

and USEUCOM Combatant Commands. For the current COCOM models those Western European countries that are members of the OECD were included into US-NORTHCOM instead of USEUCOM. This was based off of the necessity to make the groups more even in terms of the number of observations in each COCOM and to hold the assumption that members of the OECD are similar to one another and can be modeled together. These models were conditionally based on if a country in a given year was in a state of conflict or not. 12 total models were created in the current COCOM suite, with an in conflict and not in conflict model for each COCOM.

The second suite of models built was for the final groupings discussed in the previous section obtained through PCA and the Modified K-means Algorithm. The final groupings represent a new COCOM structure with each group representing a new COCOM. 12 total models were created in the new COCOM suite, with an in conflict and not in conflict model for each new COCOM grouping. A total of 24 unique models were built between the two suites of models and each of the 24 models was created following the PSC model building strategy and contained a subset of the original set of variables. The details of the covariates in each model and their parameters estimates are found in Appendix B.

Model Building Issues

There were several issues encountered in the model building process. The first major problem encountered was that a model would be fit with a set of variables and was determined to be unstable and parameter estimates for the covariates could not be calculated. This problem stems from the fact that the variables considered in this study are highly correlated with each other. This problem mainly occurred in step 2 of PSC. The model becomes unstable when the some of the parameters are theoretically infinite and the model perfectly predicts the response or if there

are more parameters in the model than can be estimated by the data [22]. If a full multivariate model could not be estimated in the beginning of step 2 due to this issue, checking the correlation of the independent variables included in step 2 to the dependent and independent variables helped determine which variables were most correlated with one another and the dependent variable. The variable that was most correlated with the dependent variable was selected between similar variables that were correlated with each other. The variables taken out at this step were considered again, one at a time in step 4 to ensure keeping potentially significant variables. Once a variable was removed, the multivariate model was refit. This process was repeated until the parameters for the covariates were able to be estimated. This model then was designated as the full multivariate model in step 2 and the model building process was continued as normal from this point.

The second problem encountered in the model building portion of the analysis was that the model was unstable, but parameters for the covariates could be estimated. This issue also stems from the fact that many variables are correlated with one another, but it is a less severe problem than the first problem discussed. These models with unstable but estimable parameters are good at classifying but any inferences about the parameters themselves should be avoided [22]. The models that contained unstable estimated parameters were kept as they were and no specific inferences about the parameters themselves were made.

General Observations of Models

The final results of using the PSC method to build models for each COCOM and new COCOM grouping for the in conflict observations and not in conflict observations resulted in 24 unique models. Table 7 details the frequency of the number of times a variable was included in each suite of conflict models for all models (both current

and new COCOMs) and only new COCOM models.

Table 7: Overview of Frequency of Variables in the Models

Variable	In Conflict Models		Not In Conflict Models	
	Number of Total Models Variable is Included In	Number of New COCOM models Variable is Included In	Number of Total Models Variable is Included In	Number of New COCOM models Variable is Included In
BirthRate	1	0	1	0
DeathRate	1	1	3	1
FertilityRate	2	1	6	3
InternetUsers	4	2	0	0
LifeExpectancy	2	1	0	0
MobileCellSubscriptions	7	4	3	2
InfantMortalityRate	3	1	1	0
YouthPopulation	5	2	1	1
PopulationGrowth	4	2	3	1
RefugeeOrigin	3	2	0	0
UrbanPopulationGrowth	0	0	0	0
ArableLand	4	2	0	0
ImprovedWater	2	0	0	0
MilitaryExpendGDP	4	1	2	1
PopulationDensity	2	1	3	1
Trade	2	0	4	2
Unemployment	1	0	5	2
Transformed Polity	0	0	1	0
PctBC	1	1	0	0
AvgBC	1	1	3	2
BinBC	2	1	1	1
EthnicDiversity	3	2	2	1
ReligiousDiversity	1	0	4	2
FreshWaterPerCapita	5	2	2	0
GDPperCapita	3	2	3	2
RefugeeAsylum	0	0	0	0
CaloricIntake	1	0	3	2
2YrConflictIntensity	10	5	10	5
FreedomScore	4	3	5	4
FreedomTrend2Yr	2	1	1	0
FreedomTrend3Yr	0	0	1	0
FreedomTrend5Yr	0	0	0	0
DemGovType	1	1	0	0
GovernmentType	4	2	2	1
RegimeType	2	1	4	3
PopulationGrowth*ReligiousDiversity	N/A	N/A	1	0
PopulationDensity*Unemployment	N/A	N/A	1	0

The variables in bold are highlighted as variables that appeared in more than a third of the new COCOM models either in the in conflict conditional models, the not in conflict conditional models, or are border conflict variables. Variables of particular note that tended to appear in multiple models were fertility rate, mobile cell subscriptions, 2 year conflict intensity trend, freedom score and regime type. The two year conflict intensity trend was included as a significant variable in 10 of the 12 new COCOM models, which indicates it is an important indicator for transitions into and out of a state of conflict.

There was no single border conflict variable that stood out as more significant in the models than the others. Each of the border conflict variables was included in at least one of the models further suggesting that there is no single best indicator for border conflict which is important to conflict prediction. Furthermore, a border conflict variable was not included or significant in every model, but was included in several. This observation leads to the conclusion that border conflict is a significant factor in conflict prediction, but depends on the region and conflict status of a country. Different border conflict variables were significant in different models and not necessarily included in every model. Thus, border conflict is a significant factor in predicting conflict transitions, but a single best measure of border conflict cannot be determined.

4.3.1 Model Assessment and Validation

Every model required a check of model assumptions and employed several validation techniques to ensure the models did not violate any assumptions and were adequate models. The first validation metric considered was the overall significance of the model. This was determined by using a LRT to determine if the model was better than the model with only the intercept, or the mean response. Each of the

models was significant at the 0.001 level. The next validation metric employed was the Hosmer-Lemeshow test. The Hosmer-Lemeshow (HL) test compares the actual and estimated frequencies of observations into a number of bins between 0 and 1 based on the predicted probabilities. The idea is that if that correct model is fit there should be sufficient agreement between the observed and estimated frequencies regardless of the strategy used to group the observations into a number of bins [18]. This study conducted a HL test on each model and assumed that there were 10 equally spaced bins to place the number of observed and estimated expected number of observations. A p-value greater than an $\alpha = 0.05$ leads to a fail to reject the null hypothesis and conclude that the model appears to fit the data sufficiently. A p-value less than α indicates that there is little evidence that the model fits correctly. The results of each model's overall significance and HL test performance are summarized in Table 8.

Table 8: Individual Model Significance and Hosmer-Lemeshow Test Results

Model Significance and Hosmer-Lemeshow Test Results given alpha=0.05				
Model	Overall Model Significance	HL Stat (C)	HL p-value	HL Test Conclusion
USCENTCOM In Conflict	<0.0001	2.07	0.979	Model Appears to Fit Data Well
USSOUTHCOM In Conflict	<0.0001	7.04	0.532	Model Appears to Fit Data Well
USAFRICOM In Conflict	<0.0001	4.04	0.853	Model Appears to Fit Data Well
USEUCOM In Conflict	<0.0001	3.42	0.905	Model Appears to Fit Data Well
USPACOM In Conflict	<0.0001	12.71	0.122	Model Appears to Fit Data Well
USNORTHCOM In Conflict	<0.0001	2.02	0.980	Model Appears to Fit Data Well
NewCOCOM1 In Conflict	0.0001	3.31	0.913	Model Appears to Fit Data Well
NewCOCOM2 In Conflict	<0.0001	1.40	0.994	Model Appears to Fit Data Well
NewCOCOM3 In Conflict	0.0001	2.89	0.941	Model Appears to Fit Data Well
NewCOCOM4 In Conflict	<0.0001	11.43	0.179	Model Appears to Fit Data Well
NewCOCOM5 In Conflict	<0.0001	3.50	0.899	Model Appears to Fit Data Well
NewCOCOM6 In Conflict	<0.0001	12.79	0.119	Model Appears to Fit Data Well
USCENTCOM Not In Conflict	<0.0001	5.52	0.701	Model Appears to Fit Data Well
USSOUTHCOM Not In Conflict	<0.0001	2.94	0.938	Model Appears to Fit Data Well
USAFRICOM Not In Conflict	<0.0001	15.76	0.046	Model Does Not Fit Data Well
USEUCOM Not In Conflict	<0.0001	1.64	0.990	Model Appears to Fit Data Well
USPACOM Not In Conflict	<0.0001	0.88	0.999	Model Appears to Fit Data Well
USNORTHCOM Not In Conflict	<0.0001	0.64	1.000	Model Appears to Fit Data Well
NewCOCOM1 Not In Conflict	<0.0001	8.35	0.400	Model Appears to Fit Data Well
NewCOCOM2 Not In Conflict	<0.0001	1.54	0.992	Model Appears to Fit Data Well
NewCOCOM3 Not In Conflict	<0.0001	0.62	1.000	Model Appears to Fit Data Well
NewCOCOM4 Not In Conflict	<0.0001	1.82	0.986	Model Appears to Fit Data Well
NewCOCOM5 Not In Conflict	<0.0001	13.53	0.095	Model Appears to Fit Data Well
NewCOCOM6 Not In Conflict	<0.0001	14.33	0.073	Model Appears to Fit Data Well

All models are overall significant at the 0.001 level. Furthermore, all models except for one pass the Hosmer-Lemeshow Test indicating that the models appear to fit the data well. The only model that failed the HL test was the USAFRICOM Not in Conflict Model and had a p-value 0.046 which was just below the threshold for alpha of 0.05. This model was still overall significant and performed well throughout the remainder of the validation assessment portion of the study. Therefore, the model was kept as it was for the remainder of the analysis.

The next validation metric used was the Area Under the Receiver Operating Characteristic Curve (AUC). This metric considers how well the model classifies over a range of cutoff points and compares the sensitivity and specificity of the model. This metric gives an overall indication of how well a model discriminates between the two outcomes 0 or 1 over a range of cutoff points. The results of the AUC analysis are in Table 9.

For the training data, all of the models achieve an AUC of at least 0.80 which is an excellent discrimination rating. Additionally, 21 of the 24 models achieve a AUC of at least 0.9 and an outstanding discrimination rating in the training data. As expected, the validation AUC's are less than the training data AUC, but the AUCs are still within an acceptable range. 11 of the 24 models have a validation data AUC above 0.8 and an excellent discrimination rating. Six of the 24 models received poor discrimination ratings. Overall, the models perform adequately in terms of AUC.

The final validation metric used to assess and validate the models was classification accuracy. The baseline accuracy used in previous studies for a minimum acceptable classification accuracy was 80% [2]. For this study, the minimum baseline classification accuracy was based on a naive prediction accuracy. The naive prediction assumes that given a country's conflict transition status in a given year, that country would experience the same conflict transition status in the next year. For example, if a country

Table 9: Individual Model AUC Results

AUC Results				
Model	Training AUC	Training AUC Discrimination Rating	Validation AUC	Validation AUC Discrimination Rating
USCENTCOM In Conflict	0.915	Outstanding	0.526	Poor
USSOUTHCOM In Conflict	0.895	Excellent	0.891	Excellent
USAFRICOM In Conflict	0.940	Outstanding	0.884	Excellent
USEUCOM In Conflict	0.980	Outstanding	0.933	Outstanding
USPACOM In Conflict	0.961	Outstanding	0.589	Poor
USNORTHCOM In Conflict	0.972	Outstanding	0.673	Acceptable
NewCOCOM1 In Conflict	0.877	Excellent	0.759	Acceptable
NewCOCOM2 In Conflict	0.954	Outstanding	0.697	Poor
NewCOCOM3 In Conflict	0.954	Outstanding	0.909	Outstanding
NewCOCOM4 In Conflict	0.908	Outstanding	0.908	Outstanding
NewCOCOM5 In Conflict	0.919	Outstanding	0.929	Outstanding
NewCOCOM6 In Conflict	0.932	Outstanding	0.778	Acceptable
USCENTCOM Not In Conflict	0.954	Outstanding	0.571	Poor
USSOUTHCOM Not In Conflict	0.940	Outstanding	0.783	Acceptable
USAFRICOM Not In Conflict	0.940	Outstanding	0.854	Excellent
USEUCOM Not In Conflict	0.977	Outstanding	0.923	Outstanding
USPACOM Not In Conflict	0.976	Outstanding	0.708	Acceptable
USNORTHCOM Not In Conflict	0.968	Outstanding	1.000	Outstanding
NewCOCOM1 Not In Conflict	0.954	Outstanding	0.694	Poor
NewCOCOM2 Not In Conflict	0.954	Outstanding	0.926	Outstanding
NewCOCOM3 Not In Conflict	0.954	Outstanding	1.000	Outstanding
NewCOCOM4 Not In Conflict	0.942	Outstanding	0.760	Acceptable
NewCOCOM5 Not In Conflict	0.864	Excellent	0.777	Acceptable
NewCOCOM6 Not In Conflict	0.968	Outstanding	0.752	Acceptable

did not experience a conflict transition in 2004, the naive predicted conflict transition status for 2005 would be that it did not experience a conflict transition. The naive predicted values were then compared to the actual conflict transitions from the years 2005-2015 and the naive prediction accuracy calculated from the total number correct predictions under the naive assumption over the total number of observations. The overall naive prediction accuracy for the whole dataset was 83%. This is a minimum classification accuracy needed from the models. If the classification accuracy for a given model was below that threshold then the model does not perform better than making a naive prediction of each country's conflict transition. The overall classifi-

cation accuracies for each model with the cutoff point set at 0.5 are shown in Table 10.

Table 10: Overall Classification Accuracies for Cutoff Point=0.5

Classification Accuracies at c=0.5					
Model	Number of Transitions in Training Data	Training Classification Accuracy	Number of Transitions in Validation Data	Validation Classification Accuracy	Training and Validation Classification Accuracy
USCENTCOM In Conflict	16	86.5%	6	76.0%	84.8%
USSOUTHCOM In Conflict	25	85.7%	5	85.2%	85.6%
USAFRICOM In Conflict	47	87.4%	9	84.2%	86.8%
USEUCOM In Conflict	12	90.9%	3	92.3%	91.1%
USPACOM In Conflict	21	93.7%	9	72.4%	89.7%
USNORTHCOM In Conflict	8	95.3%	2	86.7%	93.7%
NewCOCOM1 In Conflict	14	91.5%	6	68.0%	87.3%
NewCOCOM2 In Conflict	11	94.4%	3	78.6%	91.9%
NewCOCOM3 In Conflict	7	92.9%	2	92.3%	92.8%
NewCOCOM4 In Conflict	26	89.4%	5	93.1%	90.1%
NewCOCOM5 In Conflict	44	89.2%	9	90.6%	89.5%
NewCOCOM6 In Conflict	27	89.7%	9	71.9%	86.4%
USCENTCOM Not In Conflict	14	92.9%	3	64.7%	88.1%
USSOUTHCOM Not In Conflict	21	93.5%	2	88.0%	92.8%
USAFRICOM Not In Conflict	38	92.6%	7	93.9%	92.8%
USEUCOM Not In Conflict	11	93.3%	1	88.9%	92.5%
USPACOM Not In Conflict	24	94.3%	3	91.4%	93.9%
USNORTHCOM Not In Conflict	7	97.5%	1	95.6%	97.2%
NewCOCOM1 Not In Conflict	9	94.0%	3	80.0%	91.8%
NewCOCOM2 Not In Conflict	11	94.2%	1	92.9%	94.0%
NewCOCOM3 Not In Conflict	7	97.4%	1	95.6%	97.1%
NewCOCOM4 Not In Conflict	21	93.4%	2	88.9%	92.8%
NewCOCOM5 Not In Conflict	38	91.6%	7	85.4%	90.7%
NewCOCOM6 Not In Conflict	29	94.2%	3	92.9%	94.0%

With a cutoff point set at 0.5, predicted probabilities for an observation less than 0.5 were predicted to not experience a transition while those observations with a predicted probability of a transition greater than 0.5 were classified as a 1, meaning it would experience a transition. The predicted classifications were compared to the actual conflict transition values in calculating the overall classification accuracy. The number of correctly predicted transitions and non transitions over the total number of observations gives the overall classification accuracy for a model. Each of the individual models achieve a total classification accuracy above the naive prediction baseline accuracy of 83%. Thus, the models perform better than making a naive prediction of the country's likelihood to transition states of conflict and are sufficiently

able to predict conflict transitions within a given region.

Varying the cutoff point was considered to further analyze the performance of the new COCOM models and achieve the best settings to maximize overall accuracy. One suggested point to consider in choosing an optimal cutoff point for classification purposes is the point that maximizes both sensitivity and specificity [18]. This cutoff point is called the optimal cutoff point as it is the optimal point for simultaneously maximizing sensitivity and specificity in the training data. Table 11 details the optimal cutoff point for each model as well as the classification performance at the optimal cutoff point for each model.

Table 11: New COCOM Classification Accuracy Results with Optimal Cutoff Point

Classification Accuracies at c =Optimal Cutoff Point						
Model	Optimal Cutoff Point	Training Classification Accuracy	Validation Classification Accuracy (Optimal Cutoff Point from Training)	Total Transitions Correctly Classified	Total Non-Transitions Correctly Classified	Training and Validation Classification Accuracy
NewCOCOM1 In Conflict	0.103	72.6%	40.0%	13	82	66.9%
NewCOCOM2 In Conflict	0.137	94.4%	71.4%	13	65	90.7%
NewCOCOM3 In Conflict	0.304	96.4%	92.3%	7	59	95.7%
NewCOCOM4 In Conflict	0.547	91.2%	93.1%	22	108	91.5%
NewCOCOM5 In Conflict	0.180	84.9%	86.8%	47	196	85.3%
NewCOCOM6 In Conflict	0.184	84.1%	68.8%	29	115	81.4%
NewCOCOM1 Not In Conflict	0.033	80.7%	73.3%	10	68	79.6%
NewCOCOM2 Not In Conflict	0.172	94.2%	92.9%	11	145	94.0%
NewCOCOM3 Not In Conflict	0.040	91.5%	91.1%	8	247	91.4%
NewCOCOM4 Not In Conflict	0.088	80.8%	74.1%	22	133	79.9%
NewCOCOM5 Not In Conflict	0.187	83.2%	82.9%	33	199	83.2%
NewCOCOM6 Not In Conflict	0.170	92.4%	78.6%	29	212	90.3%

The optimal cutoff point did not increase overall accuracy of the models. This is because to increase the sensitivity of the model, there tends to be a decrease in specificity. As both sensitivity and specificity fluctuate with the optimal cutoff point, overall accuracy is not typically maximized. Furthermore, with the optimal cutoff points there was a decrease in the prediction accuracy of the validation data compared to the accuracies with a cutoff point of 0.5. There were some instances in which the optimal cutoff point improved overall accuracy and maintained a validation accuracy at least as good as when the cutoff point was 0.5. The best setting based on comparing

the models with cutoff points equal to 0.5 and the optimal cutoff point for each model was used. The default cutoff point for each of the models was 0.5. If changing the cutoff point to the optimal cutoff point value resulted in an increase in either the sensitivity (correct transitions classified) or the specificity (correct non-transitions) but did not decrease the other below the corresponding value at the default cut off point, the optimal cutoff point was designated as the best cutoff point for the model. The best settings for each of the models and the classification performances are shown in Table 12.

Table 12: New COCOM Classification Accuracy Results with Best Cutoff Point

Classification Accuracy at Best Setting				
Model	Best Cutoff Point	Training Classification Accuracy	Validation Classification Accuracy	Training and Validation Classification Accuracy
NewCOCOM1 In Conflict	0.5	91.5%	68.0%	87.3%
NewCOCOM2 In Conflict	0.5	94.4%	78.6%	91.9%
NewCOCOM3 In Conflict	0.304	96.4%	92.3%	95.7%
NewCOCOM4 In Conflict	0.547	91.2%	93.1%	91.5%
NewCOCOM5 In Conflict	0.5	89.2%	90.6%	89.5%
NewCOCOM6 In Conflict	0.5	89.7%	71.9%	86.4%
NewCOCOM1 Not In Conflict	0.5	94.0%	80.0%	91.8%
NewCOCOM2 Not In Conflict	0.5	94.2%	92.9%	94.0%
NewCOCOM3 Not In Conflict	0.5	97.4%	95.6%	97.1%
NewCOCOM4 Not In Conflict	0.5	93.4%	88.9%	92.8%
NewCOCOM5 Not In Conflict	0.5	91.6%	85.4%	90.7%
NewCOCOM6 Not In Conflict	0.5	94.2%	92.9%	94.0%

By setting the cutoff point to the best setting the overall classification accuracies of the new COCOM models only marginally increased the in conflict models and did not change the not in conflict models. The overall classification accuracy for the in conflict models at the best setting is 89.6% and the not in conflict models is 93.6% and an overall accuracy for both suites of models of 91.9%. The in conflict model accuracy and overall suite classification accuracy do improve when best cutoff point setting is applied to each model, but the improvements are marginal and only increase

the accuracies by less than 1%. The classification accuracies at the best settings outperform the respective current COCOM model classification accuracies. Because the increase in classification accuracies was only marginal, the final cutoff point settings for the new COCOM suite of models were kept at 0.5.

Based on the validation assessment, the models are acceptable and thus the final models for the study. The models are compared to the current COCOM suite of models and the new COCOM models to determine which suite of models performed best.

4.3.2 Model Comparisons

Current COCOM and New COCOM Model Suites Comparisons

Answering the second research question required comparing the current COCOM suite of models to the New COCOM suite of models. A one-to-one comparison between the current COCOM models within the suites and the New COCOMs cannot be made. Because each model was estimated using a different set of observations, a direct comparison between individual model performances cannot be made. The suites of models can be compared to each other because all the same observations are encompassed within the each suite of models. This section compares the current COCOM models to the New COCOM models for both the in conflict and not in conflict conditions.

The first test used to compare the models was the Hosmer-Lemeshow test. Table 13 compares the minimum, maximum, and average p-values within each suite of models. For both the in conflict and not in conflict suite of models, the New COCOMs have greater minimum and at least as great maximum p-values compared to their corresponding current COCOM suite of models. A greater p-value suggests that there is greater evidence to suggest that the models fit the data well. The average

p-value for the New COCOM models is less than their current COCOM counter parts, but all models in the New COCOM suites appear to fit the data well with no model's p-value less than 0.05. There was one model in the current COCOM not in conflict suite of models that had a p-value less than 0.05 suggesting that the model does not fit the data well, but in the New COCOM suite of models, all models fail to reject the null hypothesis for the HL test and all models appear to fit the data well. In comparing the current COCOM models to the new COCOM models in the Hosmer-Lemeshow test, the new COCOM models are at least as good as the corresponding current COCOM models. For the not in conflict models, the New COCOM models are better than the current COCOM models because all models appear to fit the data well and pass the HL test.

Table 13: Suite HL Test Results

Hosmer-Lemeshow Test p-value Comparison				
Suite of Models	HL Min	HL Max	HL Average	Number of Models with Poor Fit
COCOMs In Conflict	0.122	0.980	0.729	0
New COCOMs In Conflict	0.119	0.994	0.674	0
COCOMs Not in Conflict	0.046	1.000	0.779	1
New COCOMs Not In Conflict	0.073	1.000	0.591	0

The next metric used to compare model suites was the Area Under the Receiver Operating Characteristic Curve (AUC). Table 14 shows the minimum, maximum, and average AUC for both the training and validation data for each suite of models. The New COCOM models have slightly lower AUC's in all categories of the training data. While the new COCOM models perform slightly worse than the current COCOM models in the training data, the new COCOM models show improvements in the validation data. For both the in conflict and not in conflict suites of models, the new COCOM models have greater minimum AUC's with the minimum AUC being at least 0.12 greater than the current COCOM models. AUC values closer to 1 are desired

and reflect the model’s ability to discriminate outcomes over a range of cutoff points. Thus, the higher AUCs in the new COCOM models suggest the new COCOM models are able to discriminate better than their current COCOM counterparts. Overall, in comparing the current COCOM model suites to the new COCOM model suites, the new COCOM models are able to classify approximately as well for the training data and perform better than the current COCOM models for the validation data.

Table 14: Suite AUC Results

Suite of Models	Training				Validation			
	Min AUC	Max AUC	Average AUC	Average AUC Rating	Min AUC	Max AUC	Average AUC	Average AUC Rating
COCOMs In Conflict	0.89	0.98	0.94	Outstanding	0.53	0.93	0.75	Acceptable
New COCOMs In Conflict	0.88	0.95	0.92	Outstanding	0.70	0.93	0.83	Excellent
COCOMs Not in Conflict	0.94	0.98	0.96	Outstanding	0.57	1.00	0.81	Excellent
New COCOMs Not In Conflict	0.86	0.97	0.94	Outstanding	0.69	1.00	0.82	Excellent

The final metric used to compare suites of models was the classification accuracies of each suite of models, using the minimum, maximum, average, and weighted average for the overall classification accuracy of the suite of models. The overall classification accuracy for a given model is calculated as the percentage of true negatives (actual non transition observations predicted and classified as not transitioning) and true positives (actual transition observations predicted and classified as transitions) over the total number of observations in both the training and validation data sets with a cutoff point set to 0.5. The average accuracy metric is strictly the average of the individual models’ overall classification accuracies with a cutoff point set to 0.5, while the weighted average weights the classification accuracies according to the number of observations in each model.

Table 15: Suite Classification Accuracy Results with Cutoff Point=0.5

Total Classification Accuracies at c=0.5				
Suite of Models	Min Accuracy	Max Accuracy	Average Accuracy	Weighted Average
COCOMs In Conflict	84.8%	93.7%	88.6%	87.8%
New COCOMs In Conflict	86.4%	92.8%	89.7%	89.1%
COCOMs Not in Conflict	88.1%	97.2%	92.9%	93.5%
New COCOMs Not In Conflict	90.7%	97.1%	93.4%	93.6%

With the cutoff point set to 0.5, the new COCOM models overall classify conflict transitions with greater accuracy than the current COCOM models. The minimum classification accuracy for a model within a given suite is greater for the new COCOM models than in the current COCOM models. The strict average and the weighted averages for overall classification accuracy for the new COCOM models are also greater than the current COCOM models. This evidence leads to the conclusion that in terms of overall classification accuracy the new COCOM models perform better than the current COCOM models.

Overall, in comparing the models based on fit, discrimination, and overall classification accuracy at a cutoff point of $c=0.5$, the new COCOM suite of models performs better than the corresponding current COCOM suite of models for conflict prediction. Thus, the New COCOM models are superior to the current COCOM models and grouping similar countries together based on data and geography improves prediction capabilities for a given region.

Study Model Suites Comparisons

The final comparison made in this study was between the New COCOM suite of models and the performance of models from other studies. Models from previous studies used different sets of training and validation data and grouped countries into regional groupings, not COCOMs. Each set of the models was compared based on their classification accuracy of the training data, the data that was used to fit the models, and the validation data, the data unseen by the model. The classification accuracy for the validation data speaks to how well the model classifies new observations or unseen data and the prediction capabilities of the model. Because the data used in each study was different and grouped into different models, the models are compared with a weighted average in overall classification accuracy of all the models for the training data and validation data.

Table 16: Study Classification Accuracy Comparison

Comparison of Overall Classification Accuracy between Studies				
Data	Boekestein	Shallcross	Leiby	New COCOM Models
Overall World Results (Training Data)	86.63%	88.76%	91.99%	92.69%
Overall World Results (Validation Data)	78.30%	84.67%	82.56%	87.08%

The classification accuracies listed for each study’s set of training and validation data reflects the overall classification accuracies for all observations in each set of data from the best models in each study. The new COCOM models achieve the highest classification accuracies of the three previous studies. There is a slight increase in classification accuracy in terms of classifying the training data, but the most notable increase is found in comparing the new COCOM models overall validation classification accuracy which increases by about 3% compared to previous models. The increase in overall classification accuracy supports the claim that grouping countries together based on data similarities and geography improves overall prediction capabilities and achieve the best results found in the literature to date as compared to geographically based groupings.

V. Conclusions

5.1 Overall Conclusions

This study developed a methodology to group countries together into new COCOM groupings based on data similarity and geographic proximity and compared the new COCOM prediction models to current COCOM prediction models developed. Conditional logistic regression models were developed for in conflict and not in conflict conditions for each of the 6 COCOMs in the new and current groupings for a total of 24 models. This study considered 35 data elements describing various political, military, economic, social, information, and infrastructure characteristics of 182 countries from the years 2004-2015. The results of the study led to several conclusions which are summarized in answering the original research questions.

Question 1: How should we group countries to develop statistical models for conflict prediction?

Grouping countries together strictly by geography assumes that nations near one another behave in the same way and are overall similar, which is not necessarily the case as was shown through PCA. On the other hand, grouping countries strictly based on data similarity leads to disjointed COCOMs which is neither intuitive nor feasible for Combatant Commanders to effectively create plans for their area of responsibility. Grouping countries into 6 mutually exclusive groups equally based on data similarity and geographic proximity led to near contiguous groupings. These new COCOM groupings differ from the current COCOMs and suggest that the current structure is more heavily based on geography rather than similar characteristics. By grouping countries based on data similarity and geographic proximity, conflict transition predictions were improved. The predictions from the new COCOM models had higher overall classification accuracies compared to the current COCOM prediction mod-

els. The overall weighted classification accuracy for the suite of in conflict prediction models were 87.8% and 89.1% for the current and new COCOM models respectively. The new COCOM groupings improved overall classification accuracy by nearly 2% as compared to predictions from the current COCOM groupings. The new COCOM models included in the not in conflict suite also had a slightly higher overall classification accuracy with an increase of overall weighted classification accuracy of 0.1%. The methodology used to group countries together in this study can be applied as a possible alternate way to restructure COCOMs to reflect greater similarity in terms of characteristics observed in data.

Question 2: Can we obtain better models with different groupings other than strictly regional groupings?

Grouping nations into new COCOMs based on data and geography overall improved forecasting capabilities. The most notable increases in classification accuracy were seen in the in conflict suites of models and the prediction of unseen data or the validation data. The new COCOM models outperformed the current COCOM models in terms of discrimination and classification. Furthermore, not only do the new COCOM suite of models retain at least 83% classification accuracies across the models, they also obtain the highest classification accuracies compared to Boekestein [15], Shallcross [2], and Leiby [3] studies. The new COCOM models increased overall classification accuracies by 2.5% for the validation data and approximately 1% for the training as compared to the best results of the previous studies. The nearly 3% increase in prediction accuracy of the validation data shows that the new COCOM models were better able to predict new observations than the models developed in previous studies. The models obtained from the new groupings based on data and geography improved forecasting capabilities of country conflicts and achieved greater overall classification accuracies compared to models for the current COCOM struc-

ture and models developed in previous conflict prediction studies which considered geographic groupings of nations.

Question 3: Do neighboring countries affect a nation's transition into or out of a state of conflict?

There were three border conflict variables investigated in this study: percent border conflict, average border conflict, and binary border conflict. No single border conflict variable stood out amongst the others in terms of significance and inclusion in the models. Each border conflict variable was included in at least one of the final new COCOM models, but not every final model included a border conflict variable. During the univariate analysis step in the model building process, border conflict variables were often significant, but were usually taken out during step 2 of PSC and not included in the final models. At least one border conflict variable was significant in new COCOM 1 and 6 models for countries currently in conflict. These two new COCOM groupings are associated with the Middle East and Asia nations respectively. At least one border conflict variable was significant in the new COCOM 2 and 6 models for nations not in conflict. New COCOMs 2 and 6 represent the Eastern European and Asia nations respectively. Border conflict variables were included in several of the models developed in this study, indicating border conflict does influence the likelihood of a nation to transition into or out of a state of conflict, but it is not necessarily influential in every region. The significance of border conflict in predicting a country's likelihood to transition into or out of conflict is dependent on the region and current conflict status of a given country.

Question 4: What is the best measure of border conflict in predicting transitions of countries into or out of a state of conflict?

As was previously stated in the answer to question 3, no single border conflict variable stood out amongst the others. None of the border conflict variables was

superior to the others. Overall, in terms of border conflict effects on other nations, it depends on the region and current conflict state of a nation. No one border conflict variable proved more significant than the others and border conflict was not significant in every model, but border conflict does seem to have an effect on the likelihood of countries' transitions into or out of conflict in some regions and should continued to be studied in future conflict prediction models.

5.2 Implications of Research

There are several implications from the results of this study. The first implication is the methodology developed allows grouping countries together based on data factors in addition to geography. The Combatant Commands' current structure has remained largely unchanged for many years and even though top military leaders may question the current structure, it is difficult to determine how countries should be grouped together and where the borders should be drawn for the geographic COCOMs. This new methodology provides a process for comparing countries based on data similarity and geography. Grouping similar countries together in terms of data and geography, retains mostly contiguous groups of nations and improves conflict prediction capabilities.

The implications of improved prediction capabilities have far reaching effects from regional experts to top leaders of the military. Having the ability and foresight of future threats in an area can serve as a basis for region and country experts to further analyze the potential risks within a region or country. More accurate predictions and identification of potential risks can further assist country experts and leaders to develop more effective strategic and operational plans. Improved conflict prediction capabilities also allows Combatant Commanders to have a better understanding of their area of responsibility. This improved capability directly benefits leaders by

helping to avoid future conflicts and identifying countries that are susceptible to insurgent groups. This allows them to make more accurate assessments of the threats and vulnerabilities in their region and develop strategic and operational plans that best suit the needs of their area of responsibility. More effective planning ultimately leads to more prepared and effective military operations and a more efficient allocation of resources. Furthermore, having the means to accurately predict conflict enables leaders to identify potential risks in different regions and the nations within those regions. The ability to identify vulnerabilities early can help in the development of effective plans to either prepare for increased conflict in a nation or mitigate possible threats before conflict ensues.

5.3 Future Research

There are several areas in which this research could be continued to further the field of country conflict prediction and yield meaningful results. The first suggested area of study is to conduct further analysis on how nations should be grouped together. This study assumed that 6 mutually exclusive groups of nations exist that could serve as COCOMs. Further cluster analysis could be conducted to determine if 6 is the appropriate and most effective number of COCOMs or if some other number of COCOMs should be considered. Additionally, there could be further sensitivity analysis conducted on the Modified K-means Algorithm in terms of how to identify candidate points and how to measure data similarity. This study considered only one year of data to compare how similar countries are to one another. This idea could be expanded to incorporate trends over several years of data and consider similarity over time.

Another area that could be explored in future studies is expanding the definition of neighboring or bordering nations. The most recent studies in this field of country

conflict prediction have included numerous border conflict variables in the studies, but have assumed that the definition of a neighboring country is one that explicitly shares a direct border with another nation. These border conflict variables also assume that island nations do not have any neighbors. To expand this area of research, border conflict variables could be developed to expand the definition of a neighboring country to include interactions and relations that are not strictly border related. This could include information about alliances, trade interactions, or differences between surrounding nations in terms of economics or demographics.

One final area of research is expanding predictions into alternative futures of countries. This study focused on predictions for country conflict transitions based on one year of data for a country. The models developed in this study give the probability of a country to transition into or out of a state of conflict given a current state of conflict and one year of data for a country. Forecasts could be expanded by exploring alternate futures for countries using the information obtained from the models in a Markov Chain model or simulation analysis. The improved predictions of countries' likelihoods to transition into or out of a state of conflict could be applied in future studies to explore possible future states of conflict of a country and the tendency of a country to transition conflict states over a number of years. Applying a simulation type of analysis with the results of this study could provide more dynamic and long term insights of possible threats or vulnerable countries within each COCOM and further enhance the development of strategic and operational plans for the COCOMs.

Appendix A. Independent Variables Information

Variable Name	Source	Description	Number of Missing Values	Imputation Method
BirthRate	World DataBank: World Development Indicators	Birth rate, crude (per 1,000 people)	182	Multivariate Normal Imputation
DeathRate	World DataBank: World Development Indicators	Death rate, crude (per 1,000 people)	182	Multivariate Normal Imputation
FertilityRate	World DataBank: World Development Indicators	Fertility rate, total (births per woman)	182	Multivariate Normal Imputation
InternetUsers	World DataBank: World Development Indicators	Internet users (per 100 people)	36	Univariate Imputation
LifeExpectancy	World DataBank: World Development Indicators	Life expectancy at birth, total (years)	182	Multivariate Normal Imputation
MobileCellSubscriptions	World DataBank: World Development Indicators	Total Mobile cellular subscriptions	15	Univariate Imputation
InfantMortalityRate	World DataBank: World Development Indicators	Mortality rate, infant (per 1,000 live births)	0	No Missing Values
YouthPopulation	World DataBank: World Development Indicators	Population ages 0-14 (% of total)	4	Multivariate Normal Imputation
PopulationGrowth	World DataBank: World Development Indicators	Population growth (annual %)	6	Multivariate Normal Imputation
RefugeeOrigin	World DataBank: World Development Indicators	Refugee population by country or territory of origin	34	Univariate Imputation
UrbanPopulationGrowth	World DataBank: World Development Indicators	Urban population growth (annual %)	4	Multivariate Normal Imputation
ArableLand	World DataBank: World Development Indicators	Arable land (hectares per person)	198	Last Value Forward
ImprovedWater	World DataBank: World Development Indicators	Improved water source (% of population with access)	54	Univariate Imputation
PopulationDensity	World DataBank: World Development Indicators	Population density (people per sq. km of land area)	12	Multivariate Normal Imputation
Trade	World DataBank: World Development Indicators	Trade (% of GDP)	124	Multivariate Normal Imputation
Unemployment	World DataBank: World Development Indicators	Unemployment, total (% of total labor force) (national estimate)	303	Multivariate Normal Imputation
Transformed Polity	Center for Systemic Peace	Revised Combined Polity Score for Time-Series Analysis	137	Based off Regime Type
PctBC	Developed from HIIK intensity levels and border information	Border conflict measure considering the percent of perimeter of border shared with a country and the HIIK intensity level of border sharing country	0	No Missing Values
AvgBC	Developed from HIIK intensity levels and border information	Border conflict measure for the average HIIK conflict intensity level surrounding a country	0	No Missing Values
BinBC	Developed from HIIK intensity levels and border information	Border conflict indicator which indicates if at least one bordering nation is in a state of conflict	0	No Missing Values
EthnicDiversity	Shallcross Database	Percent of Dominant Ethnic Group	0	No Missing Values
ReligiousDiversity	Shallcross Database	Percent of Dominant Religious Group	0	No Missing Values
FreshWaterPerCapita	Shallcross Database	Cubic meters, average of 2007,2012,2013 data	0	No Missing Values
GDPperCapita	World DataBank: World Development Indicators	GDP per capita growth (annual %)	48	Univariate Imputation
MilitaryExpendGDP	World DataBank: World Development Indicators	Military expenditure (% of GDP)	380	Modified South Sudan Values w/ Sudan's values, then Multivariate Normal Imputation
RefugeeAsylum	World DataBank: World Development Indicators	Refugee population by country or territory of asylum	289	Modified South Sudan Values w/ Sudan's values, then Univariate Imputation
CaloricIntake	UN Food and Agriculture Organization	Caloric Intake (kcal/capita/day)	768	Univariate Imputation
2YrConflictIntensity	Developed from HIIK intensity levels	Change in HIIK conflict intensity between current year of observation and HIIK intensity level of current year -2	0	No Missing Values
FreedomScore	Developed, Freedom House	Combined average of normalized Civil Liberties and normalized Political Rights	31	Modified South Sudan Values w/ Sudan's values, then Multivariate Normal Imputation
FreedomTrend2Yr	Developed from freedom score	Difference between freedom scores of current year -2 and current year -1	0	No Missing Values
FreedomTrend3Yr	Developed from freedom score	Difference between freedom scores of current year -3 and current year -1	0	No Missing Values
FreedomTrend5Yr	Developed from freedom score	Difference between freedom scores of current year -5 and current year -1	0	No Missing Values
DemGovType	Developed from Government type	Indicator variable denoting if observation has democratic government or not	0	No Missing Values
GovernmentType	Developed from original polity variable	Mapping of Original Polity variable into 6 categories	0	Based off Regime Type
RegimeType	Shallcross Database	Static variable indicating the type of regime for a country	0	No Missing Values

Appendix B. Final Models

Table 17: USCENTCOM In Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-7.4320	5.9506	-1.25	0.2117
MobileCellSubscriptions	-0.0000	0.0000	-2.45	0.0142
ImprovedWater	0.1360	0.0580	2.34	0.0190
PopulationDensity	-0.0118	0.0060	-1.95	0.0514
YouthPopulation	-0.1976	0.1278	-1.55	0.1222
BirthRate	0.2160	0.1624	1.33	0.1836
Trade	-0.0255	0.0188	-1.36	0.1741
ArableLand	1.2022	1.1434	1.05	0.2930

Table 18: USSOUTHCOM In Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-7.4374	1.9433	-3.83	0.0001
2YrConflictIntensity	-13.9289	3.6112	-3.86	0.0001
FreedomScore	13.2731	3.7117	3.58	0.0003
FreedomTrend2Yr	-6.5546	7.3352	-0.89	0.3715
InternetUsers	-0.0607	0.0303	-2.01	0.0447
FreshWaterPerCapita	-0.0001	0.0000	-2.27	0.0230

Table 19: USAFRICOM In Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	10.0406	4.7616	2.11	0.0350
MobileCellSubscriptions	-0.0000	0.0000	-2.90	0.0037
YouthPopulation	-0.4078	0.1295	-3.15	0.0016
PopulationGrowth	2.2217	0.7387	3.01	0.0026
RefugeeOrigin	-0.0000	0.0000	-3.47	0.0005
MilitaryExpendGov	-0.7353	0.3000	-2.45	0.0143
2YrConflictIntensity	-12.5468	2.5654	-4.89	0.0000
GovernmentType1	-3.7742	1.3513	-2.79	0.0052
GovernmentType2	-4.4482	1.5327	-2.90	0.0037
GovernmentType3	-1.9262	1455.4042	-0.00	0.9989
GovernmentType4	-4.1811	1.6273	-2.57	0.0102
GovernmentType5	-7.1685	2.5919	-2.77	0.0057
RegimeTypeDemocratic	1.7999	1.4575	1.23	0.2169
RegimeTypeTransitional	4.2796	1.5673	2.73	0.0063
Unemployment	0.0679	0.0513	1.32	0.1860
EthnicDiversity	1.9549	1.1349	1.72	0.0850
FreshWaterPerCapita	0.0000	0.0000	0.90	0.3665
InfantMortalityRate	0.0464	0.0164	2.83	0.0047

Table 20: Modified USEUCOM In Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	44.0042	18.8434	2.34	0.0195
YouthPopulation	-1.1534	0.5565	-2.07	0.0382
MilitaryExpendGDP	-4.9848	1.8467	-2.70	0.0069
GDPperCapita	-0.0010	0.0005	-2.11	0.0349
2YrConflictIntensity	-22.1817	9.8003	-2.26	0.0236
GovernmentType1	-3.7078	2.9957	-1.24	0.2158
GovernmentType2	-11.8928	5.3767	-2.21	0.0270
GovernmentType3	-14.2295	6.8446	-2.08	0.0376

Table 21: USPACOM In Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-147.7998	43.9129	-3.37	0.0008
FertilityRate	6.9893	2.6095	2.68	0.0074
LifeExpectancy	1.9645	0.5923	3.32	0.0009
PopulationGrowth	7.9487	2.4011	3.31	0.0009
ArableLand	57.4942	18.3715	3.13	0.0018
ReligiousDiversity	-14.2189	4.2827	-3.32	0.0009
2YrConflictIntensity	-8.9989	4.0008	-2.25	0.0245
Trade	-0.0693	0.0255	-2.72	0.0066
MobileCellSubscriptions	-0.0000	0.0000	-1.84	0.0658
MilitaryExpendGDP	-3.2575	1.4188	-2.30	0.0217
InfantMortalityRate	-0.2655	0.1280	-2.07	0.0381
CaloricIntake	0.0016	0.0011	1.44	0.1510
InternetUsers	-0.0786	0.0484	-1.62	0.1045
BinBC	2.6895	1.9539	1.38	0.1687

Table 22: Modified USNORTHCOM In Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	59.1605	31.9524	1.85	0.0641
FreshWaterPerCapita	0.0010	0.0003	2.92	0.0035
2YrConflictIntensity	-18.8757	7.2839	-2.59	0.0096
ImprovedWater	-0.6786	0.3415	-1.99	0.0469

Table 23: NewCOCOM 1 In Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-5.5600	2.2340	-2.49	0.0128
MobileCellSubscriptions	-0.0000	0.0000	-2.60	0.0093
PopulationDensity	-0.0304	0.0121	-2.51	0.0121
PctBC	2.5943	0.8891	2.92	0.0035
GovernmentType1	-1.4569	0.7558	-1.93	0.0539
GovernmentType3	-20.0807	2662.8527	-0.01	0.9940
GovernmentType4	-3.5082	1.7299	-2.03	0.0426
GovernmentType5	-5.4795	1812.3045	-0.00	0.9976

Table 24: NewCOCOM 2 In Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	57.7740	24.3671	2.37	0.0177
DeathRate	-1.2450	0.6407	-1.94	0.0520
YouthPopulation	-2.2304	0.8997	-2.48	0.0132
FreshWaterPerCapita	-0.0011	0.0006	-1.94	0.0519
2YrConflictIntensity	-20.1980	8.2696	-2.44	0.0146
DemGovType	-3.9581	1.8363	-2.16	0.0311

Table 25: NewCOCOM 3 In Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-39.2947	23.9329	-1.64	0.1006
ArableLand	61.4264	45.6843	1.34	0.1788
2YrConflictIntensity	-42.1526	28.2778	-1.49	0.1361
PopulationGrowth	-2.7332	2.1520	-1.27	0.2040
FreedomScore	25.7389	19.9295	1.29	0.1965
MobileCellSubscriptions	-0.0000	0.0000	-1.11	0.2651

Table 26: NewCOCOM 4 In Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-7.2947	2.8455	-2.56	0.0104
X2YrConflictIntensity	-15.9035	4.3260	-3.68	0.0002
FreedomScore	9.7842	3.5112	2.79	0.0053
FreedomTrend2Yr	-11.9471	7.6242	-1.57	0.1171
RefugeeOrigin	-0.0001	0.0000	-1.74	0.0819
EthnicDiversity	2.1344	1.6842	1.27	0.2050
InternetUsers	-0.0659	0.0297	-2.22	0.0266

Table 27: NewCOCOM 5 In Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	4.4608	4.9069	0.91	0.3633
MobileCellSubscriptions	-0.0000	0.0000	-3.69	0.0002
YouthPopulation	-0.3049	0.1642	-1.86	0.0633
RefugeeOrigin	-0.0000	0.0000	-2.80	0.0051
2YrConflictIntensity	-12.1568	2.4461	-4.97	0.0000
FreedomScore	7.3977	2.2677	3.26	0.0011
EthnicDiversity	2.8668	1.0070	2.85	0.0044
GovernmentType1	-1.5639	1.3146	-1.19	0.2342
GovernmentType2	-4.1384	1.7225	-2.40	0.0163
GovernmentType3	-7.6529	3956.1822	-0.00	0.9985
GovernmentType4	-2.6223	1.7001	-1.54	0.1230
GovernmentType5	-14.7486	1356.4210	-0.01	0.9913
FertilityRate	1.0459	0.6253	1.67	0.0944
GDPperCapita	0.0004	0.0002	2.35	0.0190

Table 28: NewCOCOM 6 In Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-79.0581	27.0954	-2.92	0.0035
InternetUsers	-0.0676	0.0354	-1.91	0.0562
LifeExpectancy	1.0474	0.3646	2.87	0.0041
MobileCellSubscriptions	-0.0000	0.0000	-2.03	0.0427
InfantMortalityRate	0.0890	0.0483	1.85	0.0650
PopulationGrowth	1.9660	0.9138	2.15	0.0314
ArableLand	3.7880	1.3064	2.90	0.0037
AvgBC	-2.1159	0.7494	-2.82	0.0047
BinBC	4.9459	1.9387	2.55	0.0107
FreshWaterPerCapita	0.0001	0.0000	3.48	0.0005
2YrConflictIntensity	-10.2960	3.2099	-3.21	0.0013
MilitaryExpendGDP	0.8437	0.4707	1.79	0.0731
GDPperCapita	-0.0001	0.0001	-1.14	0.2543
RegimeTypeDemocratic	-1.3044	0.9971	-1.31	0.1908
RegimeTypeEmergingTransitional	-15.7118	1775.0073	-0.01	0.9929

Table 29: USCENCOM Not In Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.9293	2.9265	-1.00	0.3169
FertilityRate	2.1772	1.2371	1.76	0.0784
Trade	-0.0715	0.0279	-2.57	0.0103
FreshWaterPerCapita	0.0007	0.0002	2.94	0.0032
2YrConflictIntensity	16.6487	4.8960	3.40	0.0007

Table 30: USSOUTHCOM Not In Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-25.1406	8.7983	-2.86	0.0043
DeathRate	2.9623	1.1213	2.64	0.0082
FertilityRate	-6.7033	2.6102	-2.57	0.0102
InfantMortalityRate	0.3034	0.1298	2.34	0.0194
PopulationGrowth	8.6022	3.2177	2.67	0.0075
PopulationDensity	-0.0278	0.0113	-2.46	0.0139
EthnicDiversity	-16.3478	5.4934	-2.98	0.0029
ReligiousDiversity	22.5481	6.9847	3.23	0.0012

Table 31: USAFRICOM Not in Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	11.8818	4.2836	2.77	0.0055
BirthRate	0.3484	0.1005	3.47	0.0005
DeathRate	-0.2462	0.1076	-2.29	0.0222
MobileCellSubscriptions	0.0000	0.0000	3.31	0.0009
PopulationGrowth	-5.2930	1.3534	-3.91	0.0001
TransformedPolity	0.3108	0.0991	3.14	0.0017
ReligiousDiversity	-15.7390	5.0820	-3.10	0.0020
2YrConflictIntensity	10.4475	2.0287	5.15	0.0000
FreedomScore	-11.8117	2.7908	-4.23	0.0000
RegimeTypeDemocratic	-4.3693	1.1286	-3.87	0.0001
RegimeTypeEmergingTransitional	-6.4842	1.6469	-3.94	0.0001
PopulationReligiousDiversity	4.7516	1.8422	2.58	0.0099

Table 32: Modified USEUCOM Not in Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-18.3853	5.9568	-3.09	0.0020
Unemployment	0.3863	0.1343	2.88	0.0040
CaloricIntake	0.0175	0.0073	2.42	0.0157
2YrConflictIntensity	40.1615	14.1792	2.83	0.0046
GovernmentType1	-21.7841	3249.5581	-0.01	0.9947
GovernmentType2	-1.3162	1.6382	-0.80	0.4217
GovernmentType3	-1.5056	2.6642	-0.57	0.5720
AvgBC	3.1422	1.3201	2.38	0.0173
FreedomTrend2Yr	84.4707	36.4636	2.32	0.0205

Table 33: USPACOM Not in Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	9.4096	2.8605	3.29	0.0010
Trade	-0.0547	0.0188	-2.91	0.0036
Unemployment	-1.0826	0.2735	-3.96	0.0001
FreshWaterPerCapita	-0.0000	0.0000	-2.72	0.0064
GDPperCapita	-0.0002	0.0001	-3.09	0.0020
2YrConflictIntensity	14.6152	3.2621	4.48	0.0000
FreedomTrend3Yr	26.4023	13.5410	1.95	0.0512
PopulationDensity	-0.0001	0.0013	-0.08	0.9329
PopulationDensityUnemploy	0.0008	0.0003	2.86	0.0043

Table 34: Modified USNORTHCOM Not in Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-23.0795	6.8407	-3.37	0.0007
FertilityRate	6.0554	2.4777	2.44	0.0145
MilitaryExpendGDP	2.1880	0.7616	2.87	0.0041
Unemployment	0.4280	0.1551	2.76	0.0058
2YrConflictIntensity	13.6051	4.2518	3.20	0.0014

Table 35: NewCOCOM 1 Not in Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8759	2.9677	-0.30	0.7679
FertilityRate	2.2693	1.3469	1.68	0.0920
Trade	-0.0978	0.0379	-2.58	0.0098
2YrConflictIntensity	17.3657	5.4186	3.20	0.0014

Table 36: NewCOCOM 2 Not in Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.6934	6.3152	0.11	0.9126
Unemployment	0.3158	0.1204	2.62	0.0087
CaloricIntake	0.0125	0.0060	2.07	0.0382
2YrConflictIntensity	26.8896	10.2726	2.62	0.0089
AvgBC	1.5533	0.9188	1.69	0.0909
FreedomScore	-11.4452	5.8953	-1.94	0.0522
YouthPopulation	-0.6140	0.3478	-1.77	0.0775
RegimeTypeDemocratic	4.4379	3.4258	1.30	0.1952
RegimeTypeEmergingTransitional	-0.4030	3.1439	-0.13	0.8980

Table 37: NewCOCOM 3 Not in Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-23.0427	6.8234	-3.38	0.0007
FertilityRate	6.0723	2.4819	2.45	0.0144
MilitaryExpendGDP	2.1715	0.7715	2.81	0.0049
Unemployment	0.4259	0.1556	2.74	0.0062
2YrConflictIntensity	13.5531	4.2644	3.18	0.0015

Table 38: New COCOM 4 Not in Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-11.8593	8.6854	-1.37	0.1721
DeathRate	2.9100	1.3604	2.14	0.0324
FertilityRate	-8.2471	3.7303	-2.21	0.0270
PopulationGrowth	11.1420	4.8447	2.30	0.0215
EthnicDiversity	-15.4985	6.1604	-2.52	0.0119
ReligiousDiversity	22.1544	8.7369	2.54	0.0112
FreedomScore	-11.9570	4.8524	-2.46	0.0137

Table 39: NewCOCOM 5 Not in Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.9092	1.4355	2.72	0.0065
ReligiousDiversity	-2.4763	1.1808	-2.10	0.0360
GDPperCapita	-0.0002	0.0001	-2.19	0.0282
CaloricIntake	0.0015	0.0008	1.76	0.0791
2YrConflictIntensity	9.7110	1.8136	5.35	0.0000
FreedomScore	-5.6011	1.8450	-3.04	0.0024
RegimeTypeDemocratic	-2.2049	0.8653	-2.55	0.0108
RegimeTypeEmergingTransitional	-4.7996	1.4369	-3.34	0.0008
GovernmentType1	-0.2817	1.1168	-0.25	0.8008
GovernmentType2	0.3748	1.4641	0.26	0.7979
GovernmentType5	18.8924	1455.3981	0.01	0.9896
MobileCellSubscriptions	0.0000	0.0000	1.49	0.1355

Table 40: New COCOM 6 Not in Conflict

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.3436	2.3561	2.27	0.0233
MobileCellSubscriptions	0.0000	0.0000	2.78	0.0054
Trade	-0.0632	0.0186	-3.39	0.0007
BinBC	5.9973	2.2442	2.67	0.0075
2YrConflictIntensity	16.5140	3.6093	4.58	0.0000
FreedomScore	-9.8479	2.8880	-3.41	0.0006
RegimeTypeDemocratic	3.9917	1.3999	2.85	0.0044
AvgBC	-1.5551	0.7725	-2.01	0.0441
GDPperCapita	-0.0001	0.0001	-2.23	0.0258
PopulationDensity	0.0026	0.0011	2.41	0.0159

Bibliography

1. "Conflict Barometer 2016," tech. rep., Heidelberg Institute for International Research, Heidelberg, Germany, 2016.
2. N. Shallcross, "A Logistic Regression and Markov Model for the Prediction of Nation-State Violent Conflicts and Transitions," Master's thesis, Wright-Patterson AFB, Ohio: Air Force Institute of Technology, 2016.
3. B. Leiby, "A Conditional Logistic Regression Predictive Model of World Conflict Considering Neighboring Conflict and Environmental Security," Master's thesis, Wright-Patterson AFB, Ohio: Air Force Institute of Technology, 2017.
4. E. Drea, R. Cole, W. Poole, J. Schnabel, R. Watson, and W. Webb, *History of the Unified Command Plan*. Joint History Office, Chairman of the Joint Chiefs of Staff, 2013.
5. J. Serbu, "Dunford plans to trim Joint Staff workforce after New Year," 2015.
6. M. D. Ward, N. W. Metternich, C. L. Dorff, M. Gallop, F. M. Hollenbach, A. Schultz, and S. Weschle, "Learning from the past and stepping into the future: Toward a new generation of conflict prediction," *International Studies Review*, vol. 15, no. 4, pp. 473–490, 2013.
7. J. A. Goldstone, R. H. Bates, T. R. Gurr, M. Lustik, M. G. Marshall, J. Ulfelder, and M. Woodward, "A Global Forecasting Model of Political Instability," 2005.
8. R. Shearer, "Recognizing Patterns of Nation-State Instability that Lead to Conflict," *Military Operations Research*, vol. 15, no. 3, pp. 17–30, 2010.
9. H. Hegre, J. Karlsen, H. M. Nygard, H. Strand, and H. Urdal, "Predicting Armed Conflict 2010-2050," *International Studies Quarterly*, vol. 57, no. 163115, pp. 250–270, 2013.
10. F. Mancini, E. F. Letouze, P. Meier, P. Vinck, G. M. Musila, R. Muggah, G. Diniz, H. Puig Larrauri, A. Matveeva, and M. O'Reilly, "New Technology and the Prevention of Violence and Conflict," *Stability: International Journal of Security & Development*, vol. 2, no. 3, 2013.
11. D. M. Gibler, "Bordering on peace: Democracy, territorial issues, and conflict," *International Studies Quarterly*, vol. 51, no. 3, pp. 509–532, 2007.
12. I. Salehyan and K. S. Gleditsch, "Refugees and the Spread of Civil War," *International Organization*, vol. 60, no. 02, pp. 335–366, 2006.
13. L.-E. Cederman, K. S. Gleditsch, I. Salehyan, and J. Wucherpfennig, "Trans-border Ethnic Kin and Civil War," *International Organization*, vol. 67, no. 02, pp. 389–410, 2013.

14. J. A. Goldstone, "Population and Security: How Demographic Change Can Lead to Violent Conflict," *Journal of International Affairs*, vol. 56, no. 1, pp. 3–21, 2002.
15. B. Boekestein, "A Predictive Logistic Regression Model of World Conflict Using Open Source Data," Master's thesis, Wright-Patterson AFB, Ohio: Air Force Institute of Technology, 2015.
16. H. Rosling, "Hans Rosling: The best stats you've ever seen," 2006.
17. V. Gupta, P. J. Hanges, and P. Dorfman, "Cultural clusters: methodology and findings," *Journal of World Business*, vol. 37, no. 1, pp. 11–15, 2002.
18. D. Hosmer, S. Lemeshow, and R. Sturdivant, *Applied Logistic Regression*. Hoboken, New Jersey: John Wiley & Sons, 3rd ed., 2013.
19. Center for Systemic Peace, "The Polity Project."
20. S. P. Mandel J, "A Comparison of Six Methods for Missing Data Imputation," *Journal of Biometrics & Biostatistics*, vol. 06, no. 01, pp. 1–6, 2015.
21. A. Sikainga, J. Spaulding, R. Collins, and M. Din Sabr, "South Sudan," 2017.
22. SAS Institute Inc., "Explore Missing Values Utility — JMP 12."
23. J. Banks, J. I. Carson, B. Nelson, and D. Nicol, *Discrete-Event System Simulation*. Pearson Education Inc., 5th ed., 2010.
24. W. Dillion and M. Goldstein, *Multivariate Analysis: Methods and Applications*. New York: John Wiley & Sons, 1984.
25. G. James, D. Witten, T. Hastie, and T. Rober, *An Introduction to Statistical Learning with Applications in R*. New York: Springer, 6th ed., 2015.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, 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 Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE (DD-MM-YYYY) 22-03-2018			2. REPORT TYPE Master's Thesis		3. DATES COVERED (From — To) Sept 2016 — Mar 2018	
4. TITLE AND SUBTITLE Forecasting Country Conflict within Modified Combatant Command Regions using Statistical Learning Methods					5a. CONTRACT NUMBER	
					5b. GRANT NUMBER	
					5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Neumann, Sarah, E, 2LT, USAF					5d. PROJECT NUMBER	
					5e. TASK NUMBER	
					5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Way WPAFB OH 45433-7765					8. PERFORMING ORGANIZATION REPORT NUMBER AFIT-ENS-MS-18-M-149	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Joint Chiefs of Staff J-7: Lessons Learned 116 Lake View Parkway Suffolk, VA 23435-5226 DSN:668-7618 js.dsc.j7.room.dd-fjfd-jlld@mail.mil					10. SPONSOR/MONITOR'S ACRONYM(S) JCS J-7	
					11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT DISTRIBUTION STATEMENT A: APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.						
13. SUPPLEMENTARY NOTES This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.						
14. ABSTRACT Conflict forecasts are crucial to Combatant Commanders' understanding of the dynamic environment encompassing countries within their area of responsibility. The current structure of the Combatant Commands (COCOMs) is rooted in geography by grouping nations in geographic proximity to the same regional command. However, leaders today question the effectiveness of the current structure. A novel modified k-means clustering algorithm is developed and implemented that groups countries based on data similarities and geographic proximity resulting in new COCOM groupings that improve conflict forecasts. The data spans various political, military, economic, and social characteristics of countries, and is used to develop conditional logistic regression models that predict the likelihood of a country to transition into or out of conflict. Predictions from models using the original COCOM regions are compared to predictions from models using the new, data and geography based, regions. Reorganizing the COCOMs based on data similarity and geography improves conflict forecasting models and prediction capabilities significantly. The new models achieved training data classification accuracies exceeding 89% for each new COCOM and improved predicting the validation data with classification accuracies increasing up to 2% as compared to the current COCOM models and up to 2.5% in comparison to previous conflict studies. This methodology of grouping countries based on data similarities and geographic proximity leads to newly defined COCOMs which improve overall forecasting ability of conflict transitions, the best found in the literature to date. These results can help Combatant Commanders better understand the dynamic conflict environment of their area of responsibility and lead to the development of more effective operational and strategic plans.						
15. SUBJECT TERMS Country Conflict Forecasting, Combatant Commands, Clustering Analysis						
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON	
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Darryl K. Ahner, AFIT/ENS	
U	U	U	UU	103	19b. TELEPHONE NUMBER (include area code) (937) 255-3636, x4708; darryl.ahner@afit.edu	