AN IMPUTATION APPROACH TO DEVELOPING ALTERNATIVE FUTURES OF COUNTRY CONFLICT

THESIS

Zachary J Kane, 2d Lt

AFIT-ENS-MS-19-M-128

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-19-M-128

AN IMPUTATION APPROACH TO DEVELOPING ALTERNATIVE FUTURES OF COUNTRY CONFLICT

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

Zachary J Kane, B.S.

2d Lt, USAF

21 March 2019

AFIT-ENS-MS-19-M-128

AN IMPUTATION APPROACH TO DEVELOPING ALTERNATIVE FUTURES OF COUNTRY CONFLICT

THESIS

Zachary J Kane, B.S.
2d Lt, USAF

Committee Membership:

PhD Darryl K. Ahner,
Chair

PhD Seong-Jong Joo
Member

AFIT-ENS-MS-19-M-128

Abstract

Understanding what causes countries to be in a state of violent conflict is of vital importance to developing realistic national strategies on both a regional and global scale. Given these causes, it is important to understand the effects of missing data, how to impute that data, and the interrelation between data elements. Utilizing both open source data and previously generated equations that predict a country's likelihood to transition conflict statuses, this research projects data into the future and predicts each nations' subsequent conflict statuses. Future data is populated using a novel approach inspired by stochastic regression imputation. The replicated future data and predictions were interpreted as alternative futures of regional conflict in both the Arab world and Southeast Asia. The conflict occurrences in the Arab world region were projected to trend upward compared to the region's historic behavior. In Southeast Asia, the next ten years forecasted a decline in total violent conflicts. Regional scenarios where the elements of national power influenced a data element were implemented to learn how alternative futures might be effected. These results can inform military and political leadership on the ever changing conflict landscapes in two world regions of immense political and strategic importance.

This work is dedicated to my family and friends for their love and support.

Acknowledgements

I would like to express my gratitude to my AFIT research advisor, Dr. Darryl Ahner, for his guidance, patience, and mentorship throughout the course of this research.

Table of Contents

List of Tables

List of Figures

AN IMPUTATION APPROACH TO DEVELOPING ALTERNATIVE FUTURES OF COUNTRY CONFLICT

I. Introduction

1.1 Background

Over the course of human history, the world's countries have continually transitioned in and out of states of violent conflict. These conflicts range from small unarmed bouts to deadly world wars and have become a major area of study for nations trying to understand and mitigate the threats brought on by potential conflict and regional instability. The Heidelberg Institute for International Conflict Research (HIIK) identified 385 conflicts globally in 2017 [1]. Each nation in a defined state of conflict experienced varying levels of intensity, involved nations, and influential factors. Due to the uniqueness of each individual conflict, there have been multiple research efforts to build accurate predictive models of armed conflict. These models have ranged from incorporating every nation in the world down to instances within a single nation. Rooted in a United States Combatant Commands approach to grouping nations, previous predictive models of country conflict have achieved highly accurate results [2]. Recent studies largely focused on finding the most influential variables for predicting observed conflicts, while this still begs the question of how to project future data and predict conflicts that will emerge or stagnate over the years to come.

This research examines predicting a country's conflict based on generated future data of two world regions where their groupings are based on both geographical proximity and data

similarity. Alternative futures of conflict transitions were calculated for the nations in two historically warring world regions of national interest. With a forecasted outline of a region's conflict transitions, a nation can improve their resource allocation and strategic planning to address changing future conflict intensities.

1.2 Problem Statement

Fill the missing observations in the data set utilized by Neumann [3] through testing and identifying each variable's optimal imputation method. Develop regression models for each variable of interest for the two selected world regions. For those selected regions, iteratively create alternative futures for conflict transitions until 2030 and solve for each nation's yearly conflict status. Perform region specific what-if analysis to test the robustness of each of their future conflict environments.

1.3 Research Objectives

The objective of this study is to implement defensible imputation techniques that addresses each region's data missingness and to develop alternative futures of each regional conflict landscape using an iterative imputation style explanatory forecasting method.

1.4 Research Questions

This study seeks to answer the following research questions on alternate futures of conflict transitions in the Arab and Southeast Asia world regions.

Question 1

How should the data set utilized in the Neumann study be imputed?

Question 2

How can this research develop regression models for each region's variables of interest?

Question 3

What insights, nations susceptible or impervious to transitioning in or out violent conflict are identified by the projected conflict alternative futures?

Question 4

How robust are the alternative futures of conflict transitions when subject to regional what-if scenarios?

1.5 Assumptions and Limitations

This study is based on three underlying assumptions. The first assumption, similar to other conflict prediction research, is that the data analyzed is accurate and describes all commonalities between countries and their conflicts. The second assumption is that Neumann's [3] geographic groupings into regions were assumed to have suitable commonalities in terms of economies, locations, ethnicities, and religious demographics to develop alternative futures. The third assumption is that the variables identified as significant for Neumann's [3] conflict logistic regression models are the only relevant variables for predicting conflict during the duration of future years. This assumption means that Neumann's models remain significant predictors of country conflict transitions despite the regional conflict environments and data changing with time.

Data availability or completeness was addressed in this research by applying multiple imputation techniques to the unobserved gaps in the data. After combining multiple open sources, over half of the data's variables required imputation. A completed data set was then used to generate future data. The level of uncertainty behind the imputed values limited the fits of the regression models. Additionally, computing power and time limited the variables extrapolated. Only those variables found to be significant in each region's conflict prediction models were projected into the future as opposed to all of the variables

available. A greater degree of fidelity could have been achieved by informing the alternative futures with more data. Despite the inherent limitations of this research, it remains to provide national leadership with complete data, a tested forecasting method, and future, realistic country conflict landscapes to consider when developing foreign policy and security strategies.

1.6 Overview

This thesis is organized into 5 chapters including chapter 1, this introduction chapter. Chapter 2 contains a literature review of prominent studies and methods relating to this research. Chapter 3 discusses the study's methodology to impute the data and develop the alternative futures of conflict transitions. Chapter 4 details the results and analyses, and chapter 5 offers final conclusions and possibilities for future research.

II. Literature Review

2.1 Overview

The purpose of this chapter is to provide a background of previous influential research on country conflict prediction. It forms the basis of this research as well as explains the existing contributions to imputation and conflict prediction. This chapter is broken down into three main sections: existing nation-state conflict modeling, applied imputation techniques, and related forecasting applications. The first section focuses on previously developed predictive models for nation-state conflict that lead to the very one employed by this research. The second section identifies similar imputation methods and comparison strategies which inform this research's handling of missing data. The final section provides a summary of related applications of explanatory forecasting efforts. This chapter ultimately aims to cover literature about the main models, methods, and applications related to this research.

2.2 Previous Nation-State Conflict Models

Predicting world conflict has been a problem addressed by multiple studies. There have been different prediction methodologies where researchers have defined conflict, the influential variables, and how best to group countries prior to predicting their conflicts. There have been varying accuracies achieved by these predictive modeling efforts. This section outlines the progress already made in predicting conflict which is a key component of this research's alternative futures of violent conflict.

The 2013 research, *Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction*, focused on the benefits that prediction models of political

conflict in a complicated geopolitical landscape could provide [4]. The authors defended the importance of country conflict prediction and identified limitations in previous research. Previous research had fixated on including statistically significant variables while overlooking a variable's impact on the model's overall predictive accuracy. Ward [4] developed a model of behavioral and institutional variables to predict various countries cumulative probability of civil war, six months prior to it's onset. This hierarchical logit model's slope and intercept differed depending on groups of nations, and it accurately predicted civil wars 95% of the time. Ward's [4] research was the first to put a greater deal of importance on prediction accuracy in contrast to just building models with significant variables. Until the Ward [4] study and Goldstone's [5] efforts, conflict models' prediction accuracies were limited to around 50% [2].

The study, *A Global Forecasting Model of Political Instability*, conducted by Dr. Jack Goldstone [5] for the U.S. Central Intelligence Agency's Directorate of Intelligence predicted the onset of political instability two years prior to a conflict's start . The research took open source global data from 1955 to 2003 and performed a variety of predictive analyses to achieve 80% accuracy determining between countries that experienced political instability from those in a constant state [5]. Goldstone's [5] study found a nonlinear five-category measure of regime type to be the most powerful predictor rather than economic conditions, demography, or geography. Country conflict, the dependent variable, was defined as a revolutionary war, ethnic war, adverse regime changes, or genocides and politicides [5]. Of the predictive event history models, logistic regression, neural networks, and Markov processes tested, the simple logistic regression model performed best with only four independent variables [5]. These results informed future researchers of the importance of fewer, major variables to reduce the unexplained variance of the conflict models as well as logistic regression being the favorable method for political instability prediction. Goldstone's model was built on a world scale and ignored the differences between nations' locations and underlying cultures. In fact, the sole region modeled separately, Africa, included different variables for predicting conflict which

indicates the pertinence of regional differences [5].

The Center for Army Analysis's research, *Recognizing Patterns of Nation-State Instability that Lead to Conflict* identified the influential factors for predicting country conflict for Army operations. Shearer's [6] work initially mapped the top four intensity levels of an older, slightly different version of the HIIK conflict intensity score into two categories: peace and conflict [6]. This indicator of a nation's conflict status was the model's dependent variable. Thirteen variables from unclassified data on diplomatic, social, economic, and military factors of each country acted as the model's regressors [6]. Principal component analysis was applied to better visualize the data in a reduced three dimensional feature space. The study further used a smoothing algorithm to forecast future vectors. Using k-Nearest Neighbors and Nearest Centroid algorithms, Shearer [6] obtained a classification accuracy of 85% for a nation's stability over time. The research introduced an understandable way to view the data and define, on a global scale, a nation's likelihood of conflict while predicting further into the future with comparable precision.

Predicting Armed Conflict, 2010-2050 explained global and regional incidences of armed conflict using a multinominal logit model trained on cross-sectional data for 169 countries from 1970-2009 [7]. The model made predictions on the likelihood of a nation transitioning between no conflict, minor conflict, and major conflict. The three transition states were determined by the combat related deaths per year, and predictions were calculated by simulating the behavior of the conflict variable implied by the model's estimates [7]. The research's regional based model building strategy accounted for countries' geographically driven differences. The world was divided into eight regions from a compressed version of the United Nations' groupings and included unique regressors for each. Six separate models were developed using varying combinations of independent variables that emphasized the influence of conflict history, country development, and neighboring behavior on conflict status [7]. The simulated future data was based on proposed scenarios and projections of demographic trends over time with the the multinominal logit coefficients changing yearly until

converging [7]. With a unique prediction horizon between 7-9 years, an average postdictive accuracy of 79% with a false positive rate of 8.5% was found across all nine regions analyzed [7]. This study set a precedent of a multi level conflict matrix that allowed for the inclusion of predicting ongoing as well as initialized conflicts. The study simultaneously predicts escalation, onset, and the termination of conflicts which expands the work's application to global and regional prediction.

Building on previous theses and leading to the very predictive models featured in this research, Boekestein [8] first adapted the Heidelberg Institute for International Conflict Research "Levels of Conflict" as the dependent variable for logistic regression. Violent conflicts were defined by the highest three HIIK scores (3-violent crisis, 4-limited war, 5-war) while a country not in conflict took on one of the lower three HIIK levels (0-no conflict, 1-dispute, 2-non-violent conflict) [8]. Boekestein [8] data compiled various open sources to encompass twenty six variables. These were used to build individual parsimonious models for the six regions shaped from insights by credible statistician Hans Rosling. The study found influential variables for each region's model through variance inflation factor screening and correlation testing that produced models with a maximum prediction accuracy of 76% [8]. With a reduced logistic regression cutoff of 0.28 as opposed to the default 0.5, Boekestein [8] constructed separate regional models with varying subsets of variables that attained a postdictive accuracy around 80%. This studies resulting conflict prediction capabilities were comparable to the previously mentioned Center for Army Analysis research.

Shallcross's [2] work expanded on the Boekestein [8] analysis by introducing a nation specific Markov Chain model to forecast nations' tendencies to transition in or out of conflict. The Markovian models were based on the country's current conflict status to then determine how it would behave the following year. The model comprised of two states mapped from the HIIK conflict intensity scoring: in conflict and not in conflict [2]. The regional logistic regression models' dependent variables were the transition of conflict status, and two times the models were created to account for the two possible conflict states of country could be in

the year prior. Contingent upon a country's conflict status the year prior, an "in conflict" or "not in conflict" predictive model would be executed to find the probability of that a country would change or remain conflict statuses [2]. With the field being increasingly curious how nearby conflicts affect a nation's environment, Shallcross [2] considered independent variables for the product of the conflict intensity of nations directly sharing borders and the percent of that total border shared. Shallcross [2] achieved overall prediction accuracies above the 80% benchmark for his models and established the two state predictive modeling methodology practiced by this research.

Expanding on the work of Shallcross [2], Leiby's [9] study focused on incorporating environmental factors, specifically water and neighboring country conflict, into prediction models and analyzing their effects. For the same regions as Shallcross [2] and Boekestein [8], Leiby [9] introduced two additional independent variables into the conditional logistic regression equations that predicted each nation's conflict. The introduced 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 [9]. Bidirectional stepwise selection based on each variables' G statistics informed the variable reduction methods used for each region's models [9]. The environmental factors were forced into the models to identify their impact, but incorporating them only marginally improved model parsimony and predictive accuracy. Leiby's [9] models were still able to achieve a training classification accuracy of 92%.

Taking advantage of previous research's progress, *Forecasting Country Conflict within Modified Combatant Command Regions using Statistical Learning Methods* built the most recent conflict transition predictive models used by this research [3]. Neumann's [3] models were tailored to the regions identified by a modified k-means clustering algorithm. The regional groupings differentiated countries based on data similarities and geographic proximity. This grouping of nations improved each conditional logistic regression models that predicted the likelihood of a country to transition into or out of conflict [3]. These new regional models

were compared to others built for the current Combatant Command World Structure and yielded training data classification accuracies exceeding 89% [3]. Neumann's [3] methodology of grouping countries into a modified version of the Combatant Commands improved the overall forecasts of conflict transitions. These were the best results found in literature to date, and the models utilized to predict country conflict transitions in this research.

2.3 Applied Imputation Techniques

Researchers have taken varying degrees of complexity to fill the missing gaps in their data. This section aims to explore a few of the imputation approaches and methodologies recently practiced. There is no consensus on how to perfectly impute data, but by understanding related imputation efforts, this research is better informed to decide on a proven technique to perform on any and all foreseen data missingness.

In the study, *Overview of Missing Physical Commodity Trade Data and Its Imputation Using Data Augmentation*, incomplete physical commodity trade databases are imputed using simpler, traditional approaches and computationally complex stochastic methods [10]. The incomplete commodity trade data impedes the proper analysis of trade flow between countries, and its missingness stemmed from non-compliance of reporter countries, confidentiality issues, delays in data processing, or erroneous reporting [10]. The imputation methods tested were categorized into deterministic, single imputation approaches and stochastic imputation. The deterministic methods included imputation by mean, interpolation, and regression while the stochastic driven approaches included stochastic regression and more complex iterative processes. The study identified a key advantage of the stochastic approach is that instead of using a point estimate as the imputed value, a distribution of missing data through multiple imputations is obtained to reflect uncertainty and maintain the variability in the original data. These are both overlooked by the deterministic methods [10]. A case study considered imports of ten primary commodities from China, France, the United

Kingdom, and the United States between 1978-2010 [10]. The percent missingness for the ten variables ranged between 3.72% and 15.69%, and initially the data was transformed to be normal despite most of the more complex imputation techniques being robust to any violations of the normality assumption [10]. Auxiliary variables were included in the multiple imputations based on the improved quality and reduced bias of their correlations with incomplete variables [10]. This was a driving factor that inspired this research's decision to impute data with all available variables in a region before reducing the data set to just each region's variables of interest. The multiple imputation method applied in this study is explained in full detail within this research's methodology chapter. Synthesized estimates were compared against actual observed observations using the normalized root mean square error. The study found that the multiple imputation out-performed the substitution by mean, interpolation, and regression methods [10]. The mean imputation method was found to generate imputed values with the highest amount of deviations from the observed values, and the proposed multiple imputations yielded the smallest errors [10].

On Multivariate Imputation and Forecasting of Decadal Wind Speed Missing Data applies multiple imputations by chained equations and time series forecasting on the Department of Meteorology's daily wind speed data from 1995 through 2008 [11]. Markov Chain Monte Carlo (MCMC) imputation generated random draws from multidimensional probability distributions via Markov chains, a sequence of random variables in which the distribution of each element depends on the value of the previous one [11]. Through MCMC, the research simulated the entire joint posterior distribution of the unknown quantities and obtained simulation based estimates of posterior parameters of interest. 28% of the months of wind observations were missing [11]. MCMC was applied in a variable by variable fashion to fill in the unobserved instances, and a time series analysis on the imputed data was performed in order to then make forecasts. Forecasts were calculated using exponential smoothing by an additive Holt-Winters prediction function with constant level and no seasonality assumptions [11]. When analyzing the differences between the imputed and original wind speed

data, the study tested the standard of the multiple imputations to preserve the structure and probability functions of the imputed data. A t test was completed and found no significant difference between the original and imputed wind speed data sets, confirming the high level of reliability provided by multiple imputations [11]. A similar test comparing the original to imputed data was completed by this study to determine the superior imputation method for a given variable.

In the article, *Imputation for Multisource Data with Comparison and Assessment Techniques*, ridge regression and a state-space model, both of which take advantage of potential correlations between data are tested to impute multisource data [12]. The data comes from an experimental facility for non-regularly occurring events that collect information from four sensors: seismic, acoustic, surveillance video, and domain-name system log data [12]. The imputation by ridge regression is a constrained version of least squares regression that shrinks coefficient estimates towards zero until reaching a fit. The ridge regression then predicts the unobserved value of a certain feature. Dynamic linear models were the state-space technique that allows relationships between features to vary with time [12]. The imputation methods were compared by the mean absolute deviation between the imputed and observed values. The study found that imputation using a dynamic linear model achieved the highest accuracy and most precise confidence intervals around the imputed values. These intervals were an additional way the study assessed imputation techniques as opposed to the previously seen root mean square error statistics [12].

2.4 Related Forecasting Applications

Making predictions on future data is a complicated field involving inherent uncertainties. This section covers a few existing forecasting efforts similar to the alternative futures developed in this research. By understanding the related applications of forecasting, context is provided into the departures this research makes from the current forecasting field.

For country conflict prediction, Hegre [7] had generated predictions for future data from the years 2010 to 2050 using simulated projections of predictor variables, as provided by the United Nations World Population Prospects and the International Institute of Applied Systems Analysis. Consulting expert opinion to develop future data with Hegre's conflict status predictive models showed an overall decline in global incidences of violent crimes. The analysis attributed this decline to the improved country developmental factors: infant mortality rate, education, and youth bulges [7]. It should be noted that Hegre's long term predictions are built upon projections and should be interpreted as long term global or possibly regional conflict trends rather than specific national level predictions [2].

Non-parametric regression for space-time forecasting under missing data analyzes real time spatio-temporal data sets experiencing missingness due to long periods of unobserved sensors. The study tries to forecast future journey times of road links in central London, UK using two non-parametric regression models: kernel regression and K-nearest neighbors [13]. Traffic monitoring networks present a real time setting where imputed data are immediately required for long term forecasting that informs road users of future traffic conditions [13]. Kernel regression is a non-parametric regression technique that is used to estimate the conditional expectation of a random variable. Forecasts are produced as a combination of historical data points and weighted accordingly by the kernel function [13]. The London Congestion Analysis Project network experiences missing data which is imputed using a process called patching, replacing the missing values with estimates which vary according to the number of points that are missing in succession [13]. Both models performed well for forecasting spatio-temporal data sets that exhibit high levels of missing data. The forecasts were simple too by only having a single parameter to train and using their single upstream and downstream neighbors.

Taking a different approach, *A Stepwise Regression Method for Forecasting Net Interchange Schedule* presents a stepwise regression method for forecasting net interchange schedules [14]. These power grid operational schedules are the sum of the electric power

exchanges between an Independent System Operator/Regional Transmission Organization and its neighbors. The paper proposed using stepwise regression which iteratively adds and removes explanatory variables according to their significance in the training data to find a reduced order model that is computationally effective and can forecast the future net interchange schedule [14]. Akaike Information Criterion was used to evaluate each explanatory variable's ability to increase the goodness of fit of the statistical model. Stepwise regression was modified by adding a set of random variables whose values are drawn from normal and uniform distributions. The number of random variables was the same as the observed explanatory variables and acted as empirical stopping criterion of the stepwise regression. After some number of iterations, if the regression added more than three random explanatory variables to the regression model, the process stopped. A sliding window approach was followed to create training and test data sets, and the study found that the statistical significance of the parameters depended on the net interchange schedule forecasting horizon [14]. The regression based on the reduced explanatory variable set produced a model with smaller forecasting error in less computational time versus using the full explanatory variable set. Similarly, this research will develop alternative futures using a reduced set of variables that has been identified by Neumann [3] to be significant for predicting conflict transition in a given region.

Pedroza's [15] research takes a Bayesian approach to forecast mortality rates for the period 1990-1999 based on U.S. male mortality data from 1959-1989. Forecasts of mortality rates are vital to government agencies' development of health policies and allocation of government services' funds. Mortality rates are the ratio of deaths to mid-year population size for a given interval of age and time. A previously developed method by Lee and Carter forecasts age-specific log-mortality rates with a multivariate normal model and estimates the parameters using singular value decomposition with a random walk model with drift to forecast their vector of future levels of mortality index [15]. The Bayesian model reformulated Lee-Carter's method as a state-space model and incorporated a Markov chain Monte Carlo

method to draw samples from the joint posterior distribution of the parameters and to form the posterior predictive distribution of the log-mortality rates [15]. The model iteratively completes two steps, draw parameters from their respective conditional distributions and simulates the level of mortality state vector. The Bayesian formulation of the Lee-Carter model created wider prediction intervals which more accurately reflected the forecasting error associated with the model and underlying uncertainty of the data and technique [15]. This research doesn't go to such extensive modeling lengths but still aims to capture the uncertainties present when forecasting.

Zhang's [16] research tested a combined autoregressive integrated moving average (ARIMA) and artificial neural networks (ANN) forecasting model to take advantage of the unique strengths of each model. The study performs time series forecasting in which past observations of the same variable are collected and analyzed to develop a model describing the underlying relationship from which the model extrapolates the time series into the future [16]. ARIMA models can represent several different types of time series but are limited by the pre-assumed linear form of the model. In an ARIMA model, the future value of a variable is assumed to be a linear function of several past observations and random errors. ANN have the flexible data-driven capability of modeling nonlinearity thus erasing the need to specify a particular model form. An ANN model performs a nonlinear functional mapping from the past observations of a variable to the future value. This research proposed hybrid approach. It considers a time series composed of a linear autocorrelation structure and a nonlinear component using both models and applies ARIMA to the linear component and ANN to model the nonlinear related residuals from the ARIMA application. Three data sets were chosen to demonstrate the proposed hybrid method: nonlinear sunspot data, Canadian lynx trapping data, and British pound/US dollar exchange rate data [16]. The study found with the three data sets that the hybrid model out performed each component model used in isolation [16]. The research emphasized understanding the initial structure of the data to apply the appropriate model and accurately forecast data. The forecasting techniques

used focused on projecting single streams of data while this research investigates projecting multiple, interwoven streams of data all at once.

2.5 Summary

This literature review provides background on previous models and techniques used to develop a data set that adequately explains country conflict, how border conflicts affect a nation's likelihood of entering into a state of conflict, and what are the best ways to group countries considering their conflict statuses. Logistic regression appears to be the most effective way to predict and model conflict which is the method utilized by this research. Imputing data has been studied in many different fields that applied simple to complex strategies that helped influence the imputation technique chosen by this research. Making predictions on future data sets or forecasting has been attempted and applied with varying success. Uncertainty about the future drove most of the reviewed forecasting efforts to account for varying levels of uncertainty in their models.

III. Methodology

3.1 Overview

This research explores imputation techniques to first fill the missing gaps in the data and develop linear regression models to develop alternative futures of nation-state conflict for two major world regions. Section 3.2 describes the methodology used to develop the data set. Section 3.3 explains each imputation method tested and used to fill in those missing observations. Section 3.4 outlines the development and building procedures for the linear regression models for each variable of interest. Section 3.5 outlines the single step iterative method to create the alternate futures of each country and region. Figure 1 displays this methodology flow with the dashed line representing the repeated, iterative loops that the alternative futures generation took to produced multiple future data sets and obtain operationally feasible results.

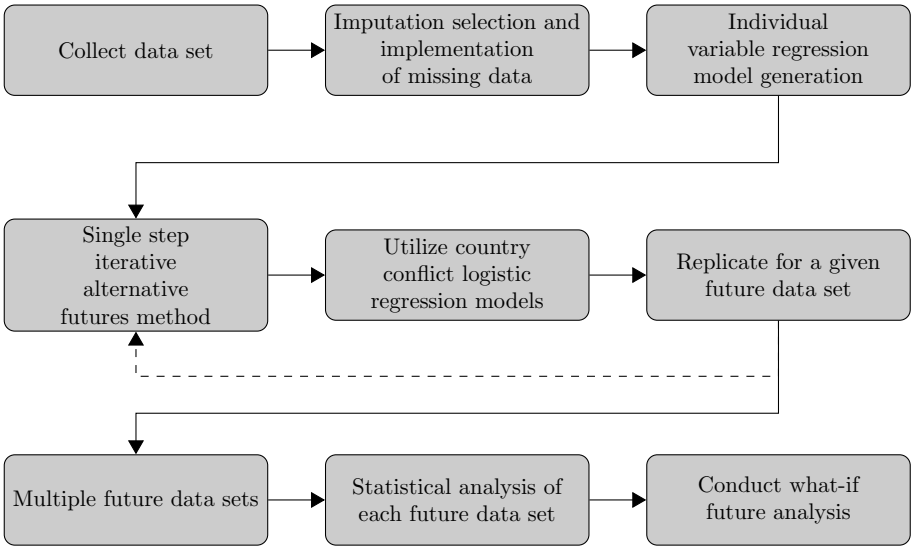

Figure 1. Methodology Overview.

3.2 Collect Data Set

The data utilized by this study recreates the Neumann [3] study's data set which built upon Leiby [9] and Shallcross's [2] insights into compiling relevant data to predict country conflict. The data initially consisted of 182 countries between 2004-2014 and included the same independent, influential variables identified by the Neumann [3]. This study decided to include military expenditure as a percentage of government spending despite Neumann [3] omitting it due to a large portion of the variable being unobserved. For reasons later discussed in Section 3.3, military expenditure as a percentage of government spending was incorporated in the multivariate imputation of the data. This research added the same two additional technology variables and two derived border conflict variables inspired by Leiby's [9] research. The dependent conflict transition variable was defined by the Neumann [3] study, and all of these generated variables are explained fully in Section 3.4 where linear regression is performed to explain the relationships present within the data.

This study assumed that all of the unobserved data are missing at random. This means that the probability of being unobserved is not the same for all cases across the data, but within an individual group of data's observed values, there is an equal probability of an observation missing [17]. The probability that a data point is missing within a group or variable is the same. Assuming the data are missing at random is the foundation for this research's applied imputation methods that take advantage of the relationships between the observed variables. The original data of all 182 countries was missing observations for 21 of the 32 variables over the past ten years. The missing observations accounted for about 6.79% of the data's total observations.

Before applying more complex imputation methods, the variable polity IV's missing observations were filled using information provided by the fully observed regime type variable. Goldstone's CIA study created the regime type variable as an indicator of political instability [5]. Regime type originally had 57 descriptors of a country's government, and Boekestein [8]

simplified the variable down to just a three level indicator. Regime type is fully observable and the same each year for every country analyzed. The polity IV is an integer variable ranging from -10 to 10 where a -10 means a country's government is fully autocratic and 10 fully democratic. Specific polity IV indicators were given to those country's with anarchies, transitioning governments, or governments experiencing a foreign interruption. The missing observations for polity IV were filled based on the regime type of a country by mapping the three levels of regime type to an appropriate polity IV value as seen in Table 1.

Table 1. Mapping of Regime Type to Fill Missing Polity Values

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

Neumann's [3] research identified new groups of countries based on applying a Modified K-Means Algorithm to the 2014 data of 182 selected countries. The groupings found by the algorithm were noticeably based on similarities between countries' data and locations. Neumann's [3] new groupings are how this study simplifies the original 182 countries down to two regions that together accounted for 45 countries. Neumann's [3] new COCOM 1 and 6 are the primary regions analyzed by this study. New COCOM 1 and 6 contain countries with historically volatile conflict statuses which makes them interesting candidates for analyzing future conflict transitions. The missingness of the first new region, largely located in the Arab world region, was 7.21% while the second region of focus, Neumann's sixth new COCOM located largely in Southeast Asia had 9.74% of its observations missing. These regions were missing more observations than the average of all the original regions which is to be expected that in countries with higher conflict occurrences, they would be more likely to have their data collection process obstructed at some point in the past ten years.

3.3 Imputation Selection and Implementation

3.3.1 Mulitple Imputation by Chained Equations

Multiple imputation by chained equations (MICE) was the method used to impute the remaining missing observations. The MICE [18] package in R [19] allowed this method to be applied to each multivariate data set of Neumann's six world regions. Multiple imputation creates $m > 1$ complete versions of the data by filling the missing observations with plausible data values [17]. Using multiple imputed data sets helps address the statistical uncertainty involved with imputing data. The MICE algorithm is a fully conditional specification imputation method which means it imputes multivariate missing data in a variable-by-variable manner [17]. MICE predicts a column of missing data as a the target variable in a regression equation with the all the other variables as the predictors unless other specified [18]. If a predictor is missing an observation, then the most recent iteration's imputation value is used to impute the target variable [18]. m chains are calculated in parallel, and after around 15-20 iterations for one regression chain, the regression coefficients for each missing variables' models are likely to converge [17]. MICE is a Markov chain Monte Carlo method in which the state space consists of of all imputed values. The MICE algorithm must satisfy three properties for its regression parameters to converge, just as any Markov chain would converge to a stationary distribution.

- irreducible, the chain must be able to reach all interesting parts of the state space
- aperiodic, the chain should not oscillate between different states
- recurrence, all interesting parts can be reached infinitely often at least from almost all starting points

The first step to fill in the initial missing data observations was checking to see if the MICE algorithm was converging for each region. Van Buuren [17] identified there being no clear-cut method for determining whether the MICE algorithm has converged but that suit-

able imputations can be spotted from plotting one or more parameters versus the iteration number. The means and standard deviations were plotted for each variable's imputation streams. Healthy converges were categorized by freely intermingled different streams, without showing any trends and the variance between different sequences not being larger than the variance within each individual sequence. No variables requiring imputation indicated any notable unhealthy levels of convergence.

After looking at MICE applied to this analysis's data and seeing previous research's MICE convergences [17], around 20 iterations was sufficient for the algorithm to converge. Each imputation conducted by this research was run for 20 iterations. MICE allows for several different univariate imputation techniques to be specifically applied to each variable every pass. MICE utilizes the columns of fully observed data in these chains. Imputing with a subset of only the missing data or simplified data set may deprive the MICE algorithm of information from the fully observed columns of data. For this reason, each imputation was performed with the missing and complete variables to include even variables deemed insignificant by Neumann's [3] models in a certain region. Multiple imputation procedurally imputes the data, analyzes each imputed data set separately, and pools the results. The initial imputation began with investigating which MICE method resulted in the data most similar to the distribution of observed data. The five mice methods tested were inspired by Brantely's [20] imputation testing methodology and displayed in Table 2.

Table 2. Imputation Methods Tested

Method	Description
cart	Classification and regression trees
pmm	Predictive mean matching
norm	Bayesian linear regression
rf	Random forrest
mean	Unconditional mean imputation

Between all of the six regions, there were on average about twelve variables per region

that required imputation. MICE has the ability to impute and utilize categorical variables. Thirteen variables in Arab region of nations were missing observations while fourteen of the Southeast Asia region's variables required imputing. MICE creates dummy variables for the categorical variables and generates their regressions and resulting imputations from these [18]. None of the variables requiring imputation were categorical, but there were categorical variables such as regime type included in the prediction of other missing variables. The MICE imputation methods investigated were selected based on there not being any missing categorical variables and trying to test a wide variety of methods. Imputation using classification and regression trees (CART) seek predictors and cut points in the predictors used to divide up the sample of data. The data are split up repeatedly until a binary tree is built to determine a target, imputed value [17]. CART methods for imputation are robust against outliers, can handle multicollinearity and skewed distributions, and are able to fit interactions and nonlinear relationships [17]. Predictive mean matching (pmm) is an imputation technique which utilizes the observed data to calculate the predicted target value. It takes a random draw from the candidate donors of complete cases with predicted values closest to the predicted missing entry value [17]. The norm method applies Bayesian linear regression that uses parameter uncertainty from random draws from a posterior probability distribution based on the observed data [17]. The random forest method uses Breiman's random forest algorithm which is essentially a combination of CART, tree predictions where the splits to arrive at classified value are found randomly [18].

3.3.2 Testing Imputation Methods

Each of the MICE methods were run to develop five different imputed data sets per region and then compared using the Kolmogorov-Smirnov (K-S) and non-parametric, 2-sample Anderson-Darling (A-D) tests. The K-S test looks at if the imputed data values are similar to the observed data values. This test makes the assumption that the imputed data should follow the same distribution as the observed data [21]. Engmann and Cousineau

[22] compared both the A-D and K-S tests and found that the A-D performed better when analyzing moments and small differences in the tails of distributions. Based on these findings, the A-D test for this analysis will be the main differentiator between imputation methods [20]. The null hypothesis behind both of these tests is that the distributions come from the same parent distribution. A small p value indicates that the imputed data and original data are significantly different and can be interpreted as a poor imputation method. For some variables missing a high percentage of observations such as freshwater per capita which was missing for ~74% of observations, no imputation method was able to find statistical similarity between that method's imputed and observed distributions.

In some literature the mean absolute error and root mean square error have been used to assess the performance of an imputation method. This analysis didn't utilize these measures of accuracy based on the difference between the true and imputed data. Due to there being so few complete cases for certain variables, evaluation metrics that were based on knowing the true values for missing data weren't implemented. Van Buuren [17] also detailed the shortcomings of treating imputation as a prediction problem geared towards finding the best value. The goal of multiple imputations is "to obtain statistically valid inferences from incomplete data" [17]. Also treating imputations as methods to enhance the classification accuracy may favor strange imputation methods [17]. For these reasons, evaluating and choosing an imputation method becomes an increasingly complex problem. Van Buuren's [17] warning as well as the conflicting results and ties between the K-S and A-D statistical tests begged for another imputation evaluation metric be applied to decide on a method. Diagnostic plots were implemented lastly to asses the plausibility of each imputation method: kernel density and box and whisker plots. The box and whisker plot was chosen over strip plots based on Van Buuren's [17] recommendation that the box and whisker plot is more appropriate for large data sets. Both these plots compare the discrepancies between the observed and imputed data. Dramatic differences would signify a possibility that something with the imputed data be further investigated [17]. The diagnostic plots were used to validate

the results of the A-D and K-S tests as well as break any potential ties from seeing which imputed values from a given method were more realistic. For example, the variable freshwater per capita from the Arab region had all five imputed values significantly different from the observed data when tested using the A-D and K-S tests. The box and whisker and kernel density plots were then examined to observe differences between imputation techniques. A suitable imputation technique would produce values that could be observed if the data had not been missing at all [18]. The best performing plots had imputed data visually closest to the observed data. After applying the K-S and A-D statistical tests and inspecting the diagnostic plots, individual imputation methods were determined for each variable in a given region and are shown in Table 3.

Table 3. Imputation Methods Used for Each Variable by Region

Variable	Arab	Southeast Asia
Caloric.Intake	cart	rf
Freedom.Score	pmm	
Freshwater.per.Capita	cart	cart
GDP.Per.Capita	cart	rf
Improved.Water.Source	pmm	pmm
Internet.Users	pmm	pmm
Military.Expend.GDP	pmm	rf
Military.Expend.Gov.Spending	cart	rf
Mobile.Cell.Subs		pmm
Population.Growth		pmm
Refugee.Asylum	pmm	pmm
Refugee.Origin		pmm
Religious.Diversity		pmm
Trade.percent.GDP	pmm	pmm
Unemployment		pmm
X2.Year.Freedom.Trend	pmm	
X3.Year.Freedom.Trend	pmm	
X5.Year.Freedom.Trend	pmm	

3.3.3 Imputation Challenges

A problem that arose specifically within the Southeast Asia region is some variables are highly correlated. When the MICE algorithm builds the regression equation to predict a missing value, a variable that is a linear combination of another will result in a singularity error that halts the MICE algorithm. To fix this, redundant variables must be excluded from the set of predictor variables used by MICE. The dependent variables can be identified by the last eigenvector of the covariance matrix of the data after performing listwise deletion [18]. The variable mobile cell subscriptions was highly correlated (> 0.5) with multiple variables. Mobile cell subscriptions had by far the smallest loading ($2.012761e-11$) on the last eigenvector of the covariance matrix. For these reasons, mobile cell subscriptions was imputed individually using MICE univariate imputation techniques, and when imputed separately, the singularity errors ceased to disrupt the MICE algorithm.

3.4 Individual Variable Regression Model Generation

With the complete data from 2004-2014, linear regression models were built for each variable of interest. Neumann's [3] model's for nations in and out of conflict defined the variables that would be of interest in each region. The variables required in Neumann's [3] models will be pertinent to predict future conflict transitions. The regression equations define each variable based on the rest of the data. Each regression model was built with the goal of achieving the most parsimonious model from the other variables in the data. Before the regression models were built and reduced, the data set required certain variables generated by Neumann's research be created using the complete data. Van Buuren [17] stressed the importance of completing all imputations before generating any additional variables. Some variables that are derivatives of other data didn't require their own regression models since their values in future scenarios would be calculated after each iteration.

3.4.1 Variable Development

After filling all the missing gaps in the data, variables of interest were developed to match the data set of previous works by Neumann [3] and Shallcross [2]. A government variable was created based on the values of the polity IV variable to indicate a nation's government type. The six categories of this variable are mapped accordingly in Table 4.

Table 4. Goverment Type Mapping from Polity

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

The percent border conflict is consistent with the percent border conflict variable in the Neumann [3] data and border conflict variable from Shallcross [2] and Boekestein [8] works. The percent border conflict is calculated by summing the product of all the percentages that a neighboring nation borders a country of interest and the neighboring countries HIIK intensity level for a given year. Islands were assumed to have no neighboring countries and a zero percent border conflict. The equation below defines the percent border conflict variable.

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

$$n = \text{number of bordering countries for country } i$$

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

$$p_k = \text{percent of border country } i \text{ shares with county } k \tag{1}$$

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

$$j = \text{Year} \in \{2004, ..., 2015\}$$

$$k = \text{Bordering country}$$

The average border conflict measures the average HIIK conflict intensity around a given country in a given year and is consistent with Neumann's [3] average border conflict variable. Islands are treated as having no bordering countries. The calculation for the average border conflict is defined as follows.

$$AvgBC_{ij} = \frac{\sum_{k=1}^{n} H_{kj}}{n} \text{where}$$

$$n = \text{number of bordering countries for country } i$$

$$H_{kj} = \text{HIIK conflict intensity level for country } k \text{ in year } j \tag{2}$$

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

$$j = \text{Year} \in \{2004, ..., 2015\}$$

$$k = \text{Bordering country}$$

The binary border conflict variable is consistent with the binary border conflict variable in both the Neumann [3] and Leiby [9] studies. It is a binary representation for a given country if one of their neighboring nations meets or exceeds a certain conflict intensity in a

given year. Islands are again assumed to have a zero score as they are not neighbored by any nations. Binary border conflict score is defined for a given country as follows.

$$BinBC_{ij} = \begin{cases} 1 & \text{if } H_{kj} \geq 3 \text{ for any country bordering country } i \\ 0 & \text{otherwise} \end{cases}$$

$$H_{kj} = \text{HIIK conflict intensity level for country } k \text{ in year } j \tag{3}$$

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

$$j = \text{Year} \in \{2004, ..., 2015\}$$

$$k = \text{Bordering country}$$

The dependent variable, conflict transition, is a binary representation if a country has changed conflict statuses since the previous year. Conflict status is defined by mapping a country's HIIK conflict intensity level in a given year. HIIK scores of 0, 1, and 2 are mapped to a 0 for that country's conflict status and indicate a year that country is not in conflict. HIIK scores of 3, 4, and 5 are mapped to a conflict status of 1 and represent a year that country is in state of conflict. The conflict transition binary variable for a country in a given year depends on the current and previous years conflict statuses. Conflict transition is equal to 1 if the conflict status of a given year, *i* is not equal to the conflict status of the previous year, *i-1*. Table 5 represents this mapping from conflict statuses for years *i-1* and *i* to the binary conflict transition variable in year *i*.

Table 5. Mapping of Conflict Transition

Conflict Status Yr $i-1$	**Conflict Status Yr** i	**Conflict Transition Yr** i
0 = Not In Conflict	0 = Not In Conflict	0 = Not In Conflict
1 = In Conflict	1 = In Conflict	0 = Not In Conflict
0 = Not In Conflict	1 = In Conflict	1 = In Conflict
1 = In Conflict	0 = Not In Conflict	1 = In Conflict

3.4.2 Individual Variable Model Building Procedure

Linear regression models were built for variables that were significant in Neumann's [3] logistic regression models that predict a country's conflict transition. Multiple linear regression models were built to statistically define each individual variable of interest based on the predictor variables in a given region. These relationships are assumed to be approximated by a straight line [23]. For multiple linear regression with k variables, the regression equation takes the form [24]:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon \tag{4}$$

The coefficients are the estimated impact that variable has on the dependent variable with an error term (ϵ). For example, the coefficient β_i can be interpreted as the estimated change in the response variable for a one unit increase in variable x_i, when all other predictor variables are held constant [24]. The regression parameters are unknown, so a point prediction of an observed value of the dependent variable is then calculated where $b_0, b_1, b_2, ..., b_n$ denote point estimates of the unknown parameters [23].

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k \tag{5}$$

The regression parameters are estimated in multiple linear regression by minimizing the sum of squared differences between the observed, y_i and predicted, $\hat{y}_i$ values of the dependent variable for the ith observation [23]. This is known as the sum of squared residuals (for $i = 1, 2, .., n$) and written as follows.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{6}$$

Using the least squares to calculate the coefficients' point estimates, higher order terms and interaction terms were tested in these models, but due to the higher initially achieved R^2_{adj} values and desire of parsimony, models were first only build using the main effects. The complete data set used to build the models consists of all five of the imputed data sets stacked on top of each other. This goes against Van Buuren's [17] advice to run regressions on each imputed data set and then pool the resulting models, but due to the desire for point estimates defining each variable of interest, the simpler method of stacking the imputations was preferred. With less than 10% of the observations missing from the two new world regions of interest, the research is impacted less from the influence of missing data. The stacked imputed data results in m x n complete records where m was still the number of imputed data sets and n were the influential, observed and imputed variables. The statistical analysis thus becomes a weighted linear regression with a weighted factor of $1/m$ applied to each record. Based on relatively lower percent missingness and unbiased point estimates, it is sufficient to build individual variable regressions from the imputed data treated as a stacked, long data set for the purposes of this research[17]. By generating five imputed data sets, analyzing them separately to decide each univariate method, and stacking them to develop a single model,

the imputation scope of this research is narrowed, but due to relatively smaller differences between imputed data sets and taking a more efficient approach, individual regression models was built and not five, one for each imputed data sets and pooling the resulting models' regression parameters.

Stepwise regression is the primary method used to generate parsimonious linear regression models for each variable. This method iteratively removes predictor variables while computing a linear regression model each time. Each time a regression is rerun, each predictor variable's associated p value is assessed to see if it is within a specified and acceptable range. By deleting variables from the model, the precision of point estimates remaining in the model are improved [25]. The stepwise method was run using JMP with a chosen p value of 0.05 for entering and removing variables. Stepwise regression operates in a forward or backward direction to either enter or remove a term with the smallest or largest acceptable p value. The mixed direction option in JMP alternates between a forward and backward selection to include the most significant terms that satisfy the probability for entering or removing terms [26]. The iterative stepwise model calculates the p value of each variable with an F test where the hypothesis test is whether the regression coefficient of that variable is equal to zero. A significant variable's p value would indicate that variable's regression coefficient has a value other than zero and actually influences the dependent variable. Models were built by stepping from a null and full model using the mixed selection method. Most times the reduced models were the same, but if there were any differences between starting with a full or null model, the model that achieved the higher R^2_{adj} was chosen. The coefficient of multiple determination, R^2, is a measure of model adequacy that represents the proportion of variance explained by the regression [25]. Adding regressors to the model will improve R^2 but can still produce a worse model and a larger mean square error from losing one degree of freedom for error [25]. A low value of R^2 for this research indicates a poorly specified model [25]. R^2 will never decrease from adding a variable to the model, so it is important in variable selection to include another evaluation statistic. R^2_{adj} will only improve if the

variable added reduces the residual mean square to prevent overfitting [25]. R^2 relates to R^2_{adj} as follows [23].

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \tag{7}$$

$$R^2_{adj} = \left(R^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-(k+1)}\right) \tag{8}$$

This research will reference R^2_{adj} when evaluating models in stepwise regression. Interaction terms were also selectively included in models with insufficient R^2_{adj} (<0.5). In an effort to develop parsimonious models, only second order interaction terms were considered. With the interaction terms factored into the model, the Arab region produced 74 possible variables while there were 209 possible variables in the Southeast Asia region. Stepwise regression reduced these numbers for the models requiring higher R^2_{adj}. All of the Arab region models were built with interactions, but only population growth's regression equation required interaction terms in the Southeast Asia region.

3.4.3 Assessing Model Adequacy

The final stepwise models were analyzed to ensure they met all the assumptions of linear regression. Linear Regression assumes the error terms or residuals must be independent, normal, and random variables with mean of zero and constant variance σ^2 [24]. Graphically, these linear regression assumptions can be evaluated using a normal probability plot of the residuals and a plot of the standardized residuals against the predicted values. A normal probability plot of the residuals is a quantile-quantile plot where quantiles of a particular distribution are plotted against the quantiles of the standard normal distribution to identify deviations from normality [24]. If a distribution is normal then a majority of the points in the

graph should fall close to the diagonal reference line. Statistically, there are lot of different ways to test normality with each depending on the data at hand. The deviations from normality were calculated using the Cramer-Von Mises test where the null hypothesis is that the distribution of the errors follows a normal distribution. It is a simplified version of the Anderson-Darling test that does not provide as much weight to the tails of the distribution. It is not the most powerful empirical distribution test, but this research choose to have more slightly relaxed normality standards. The Cramer-Von Mises statistic is calculated as,

$$CVM = \frac{1}{12n} + \sum_{i=1}^{n}[F_0(x_{(i)}) - \frac{2i-1}{2n}]^2 \tag{9}$$

where n is the sample size and x_i's are the ordered data [27]. P value scores lower than 0.05 indicate a regression model whose errors do not follow a normal distribution. The package `oslrr` [28] in R was used to test the normality of each variable of interests' residuals.

A plot of the standardized residuals against the predicted values helps identify patterns in the variance of a model's residuals. Linear regression assumes homoscedasticity, constant variance, and independence of a model's error terms [25]. These plots tests both of these assumptions and should not represent any clear funnel, linear, u-shape, or any patterns. If the variance of the errors is increasing or decreasing over time, confidence intervals for new predictions will tend to be unrealistic. The plots should be evenly spread out and distanced from the x-axis [24]. Statistically, these assumptions are tested using the Breusch-Pagan test for homogeneity of variances from the `olsrr` [28] package in R. Breusch-Pagan is a chi-squared test where the null hypothesis is that the error variances are all equal versus the alternative that the error variances are a multiplicative function of one or more variables [29]. The goal is for a model to fail to reject the Bresuch-Pagan's test which was rarely the case for the regression models developed and explained fully in the Analysis and Results Section 4.2.

3.4.4 Transformations

Certain variables regression models were improved by transforming their dependent variables which either linearized the model or stabilized its variance. These variable transformations were chosen heuristally by looking for model's with poor R^2_{adj} values or insignificant normality or homoscedasticity statistics. If a transformation improved any of these these categories then the model's dependent variable was transformed moving forward. The plots of the residuals versus fitted values of the dependent variable and actual versus predicted dependent variable values were also examined for any patterns or non linearity. The square root was taken for the population density variable in the Arab region ($\sqrt{PopulationDensity}$). The transformation yielded a higher R^2_{adj} but didn't improve the model's homoscedasticity or normality violations. The log was taken for a few variables which helped some of those models to fulfill the assumptions of linear regression ($log(y_i) = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + e_i$). Mobile cell subscriptions in the Southeast Asia region was transformed initially just by the square root, but the model continued to improve further by then taking the fourth root ($\sqrt[4]{MobileCellSubscriptions}$). JMP offered other transformations that were tested but didn't improve any other models or the three transformed (log, square root, and fourth root). Even after the transformations were performed, the regression models didn't statistically meet all of the assumptions of linear regression. This research recognizes the inadequacies of the models, but accepts them, and will proceed with the individual variables' regression models to develop alternative futures of world conflict.

3.5 Single Step Iterative Alternative Futures Method

With the data properly imputed and regression models built for each variable of interest, the alternative futures were then developed. An iterative explanatory approach inspired by univariate imputation technique, stochastic regression was primarily applied to generate the

complete future data sets. For each region individually, Neumann's [3] conflict prediction models were applied each year to find a probability of a country transitioning in or out of a state of conflict. Each variable value would be calculated using the previous years values. This was done by using the regression equations created based on the other variables of interest in their respective region. To improve on the predicted value of these equations, random noise was added to the predicted value. This noise introduces variability that reflects the inherent inaccuracy associated with predicting variables from an insufficient subset of variables. Van Buuren [17] injects noise from a random draw from the normal distribution based on the assumption that the observed data are normally distributed around the regression line. Due to the regression equations largely violating this assumption, the noise instead came from a random draw from the empirical distribution of each model's residuals. Before Neumann's [3] conflict transition logistic regressions could be applied, the complete predicted data for the next year were assessed for operational feasibility. This varied by region and variable, but the variables' values were restricted to not exceed two times the region's largest value and one half of the region's smallest value recorded in the last ten years. Exceptions to this rule were made for trade (% of GDP) and life expectancy. Two times the maximum regional values for these variables were unrealistic ceilings, so these two variables maximums were just set to the region's historic largest observed values. Negative and infeasible resulting variable values from the individual regression equations prompted each region's model being run many times before operationally feasible results were achieved with the previously stated limits. These conservative limits placed on the prediction values plus the noise eliminated variables reaching inconceivable highs or lows and maintained the operational relevancy of this research.

Certain variables identified as significant to predicting conflict transitioning were calculated based on the new variables' values rather than by their own regression equations with noise projections. Some variables were deemed difficult to impact and assumed not to be changing over the years. The two year conflict intensity trend of a country was calculated

manually based on the previous year's HIIK conflict intensity score minus the country's HIIK conflict intensity score from two years ago all divided by six for the different HIIK levels. The regime and government type variables were assumed not to change year to year because they were believed to be impervious to conflict changes. All 45 of the analyzed nations' government and regime types never changed over the past ten years. For the hard to change variables, the iterative forecast method treated a nation's political system as fixed through the course of any conflict transitions. The HIIK conflict intensity level each year was calculated based on the probability that a nation would transition in or out of conflict. The new year's data would be plugged into Neumann's logistic regression equations depending on the conflict status of the previous year. From there the predicted probability values of the logistic regression would be converted into a binary indicator of remaining or transitioning conflict states. Recall that the logit transformed logistic regression is is defined as follows [24].

$$g(x) = ln(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{10}$$

The formula for logistic regression contains p, the probability and k, the number of variables. To solve for p, the exponential of the predicted value of logistic regression was divided by one plus the same exponential of the predicted value from the logistic regression. The antilog of the predicted value of Neumann's [3] equations can be interpreted as the estimated probability of changing conflict statuses.

$$p(x) = \frac{exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)} \tag{11}$$

A higher probability indicates that a nation has a greater chance of transitioning conflict statuses while a smaller probability means a country's conflict status will remain at the same level of conflict for that year. A 0.5 probability cutoff was used to decided which direction a country would behave: transition or stationary. A random draw was taken between 0 and 1

to compare to the calculated transition probability. The random draw comparison represents the uncertainty that given a country's probability to transition conflict statuses, a country may not actually transition in the direction the logistic regression equations found to have the highest probability. If the random draw, in the direction of transition or remaining the same exceeds the logistic regression equation's probability of conflict transition then the country's conflict will in fact not do what the model predicts. For example, if a country was in conflict last year and the in conflict equation for that region found a transition probability of 0.91, the country is likely to transition to being out of conflict. A random draw would be taken between 0 and 1, and say the random draw was 0.95, than due to the random draw exceeding the transition probability in the likely transition direction, than the country would remain in conflict.

HIIK conflict intensity was mapped each year based on the conflict transition probability and random draw comparison. If a country, after taking and comparing the random draw, transitions in the direction indicated by the transition probability than the HIIK conflict intensity was just mapped following the below mapping.

Table 6. Previous Year In Conflict Mapping HIIK Conflict Intensity

Random Draw Comparison	**Conflict Status Year i**	**Year i HIIK Mapping**
Transition and Does	Not In Conflict	(1-5/6) = 0
		(5/6-4/6) = 1
		(4/6-3/6) = 2
Transition and Doesn't	In Conflict	3
Remain and Does	In Conflict	(3/6-2/6) = 3
		(2/6-1/6) = 4
		(1/6-0) = 5
Remain and Doesn't	Not In Conflict	2

For a country that was in conflict the previous year, the random draw comparison column reads transition when the transition probability calculated using Nuemann's [3] in conflict

logistic regression equations yield a probability over 0.5. If the random draw is larger than the transition probability, the column will read "transition and doesn't". Since the country's conflict is behaving contradictory to how the equations predict, the HIIK conflict intensity is mapped to being the lowest in conflict score of 3. Otherwise, the column reads "transition and does"", meaning the country for that year transitions to not in conflict and it's HIIK conflict intensity is inversely mapped based on an even division between 0.5 and 1 for the out of conflict HIIK conflict intensity scores. The same approach is taken when a country in conflict is supposed to stay in conflict (transition probability < 0.5) except now the country will only do the opposite of the predicted transition probability if the random draw is less than the transition probability. In the case that the random draw is that small and the country doesn't stay in conflict, HIIK conflict intensity is set as the lowest not in conflict score of 2. When the previously in conflict country should remain in conflict and does, the transition probability is directly mapped to the top three HIIK in conflict scores based on an even division of 0.5 to 0 as seen above. The proceeding mapping of a country that was not in conflict the previous year is mapped in the same manor but with some of the mappings flipped accordingly.

Table 7. Previous Year Not In Conflict Mapping HIIK Conflict Intensity

Random Draw Comparison	**Conflict Status Year i**	**Year i HIIK Mapping**
Transition and Does	In Conflict	(1-5/6) = 5 (5/6-4/6) = 4 (4/6-3/6) = 3
Transition and Doesn't	Not In Conflict	2
Remain and Does	Not In Conflict	(3/6-2/6) = 2 (2/6-1/6) = 1 (1/6-0) = 0
Remain and Doesn't	In Conflict	3

HIIK conflict intensity was an important variable to map as it was included in multiple

variables' individual regression equations previously built. It depends on the conflict status of a country, so it was calculated iteratively each year after the transition probability equations were already applied. With the HIIK conflict properly mapped and conflict transition logic developed, alternative futures were generated based on the trends from each variables' regression equations. The alternative futures were calculated from projections of the first of the five imputed sets. Each imputed data set could be interpreted as another set of alternative futures, but due to the scope of this research and slight differences between imputed data sets, only projections were developed initially and for the proceeding regional what-if scenarios from one of the complete, imputed data sets.

3.5.1 Conduct What-If Alternative Futures Analysis

The alternative futures created from unrestricted flow of the data into the future based on their individual regression equations were tested by scenarios a region may possibly face. These scenarios were implemented by manipulating how a certain variable behaves in the projected future yearly observations. Each region, having different predictive conflict models, was subject to slightly different variable trends in order to test a certain scenario. The three conflict what-if scenarios that were tested are listed as follows.

1. Does peace beget peace?
2. Is democracy the most peaceful form of governance?
3. How does trade impact the conflict environment of a region?

Answering each of these questions differed between the two regions of focus. For the Arab nations, the first question was tested simply by forcing the future years of projected percent of border conflict variable to be zero. The idea behind this trend is to understand how nation's conflict will change depending on if their neighboring nations become nonviolent for the foreseeable future. Does peace surrounding nations in this region permeate across borders? The second questions was aimed to understand the impact that a democratic form

of government would have on a nation's conflict status and trends. The categorical variable government type used in the Neumann [3], Shallcross [2], and Leiby [9] studies represents a nation's government ranging from autocratic to democratic. The dummy variable for an emerging democratic government type was the only variable of interest in the Arab region that could be manipulated to invoke democracy in the region. This was the closest variable in the Arab region's Neumann [3] logistic regression model which would simulate prescribed democracy's effect.

The first question in the Southeast Asian region was addressed by altering both the binary variable of border conflict and continuous variable average border conflict for each nation. The binary border conflict indicator was forced to zero for the proceeding sixteen years of projected data to represent bordering nations not being in conflict; not a single bordering nation would score above a two on the HIIK conflict intensity barometer. The average border conflict variable was restricted to not exceed a value of two which meant on average, no bordering nations would be in conflict. The second question's scenario was set by fixing every future regime type variable developed by Goldstone [5] to be democracies along with creating a floor for the freedom score variable. Freedom score is calculated by averaging a nation's civil liberties and political rights. Boekestein [8] created the variable to model a nation's political climate and oppression. This variable relates to government type, so the minimum freedom score was increased to that of the countries in the region with democratic governments. In conjunction with the preset regime type, this simulated a trend of democratic governance in the region.

The third question was addressed in the same way for both regions even thought there were vast differences between the regions' average trade as a percentage of a nation's GDP. Explained by Boekestein [8], the variable for trade is calculated from the summation of two World Bank statistics: imports of goods and services (% of GDP) added to exports of goods and services (% of GDP). The question posses the possibility of extreme trade decreases due to possible isolationistic, uncertain trade behaviors in a region where trade was disrupted

by political instability or outside pressures. The trade variable's maximum capacity was amended to be the minimum historical observed trade that a nation in that region ever operated under. This limited each nation's trading engagement to test how a scenario of shrunken trade may impact a nation's future conflicts.

IV. Analysis and Results

4.1 Overview

This chapter discusses the results and analysis of applying the methodology outlined in Chapter 3. Each variable of interest's regression models are analyzed in Section 4.2. Section 4.3 discusses the results from the data driven alternative futures of country conflict. Section 4.3 discusses the specified what-if scenarios and resulting impact on conflict in the Arab region. Section 4.5 covers the alternative futures of what-if scenarios in the Southeast Asia grouping of countries.

4.2 Regional Individual Regression Model Evaluation

Certain variables were related to other variables of interest which made them easier to explain using linear regression, but some variables were hard to change and had weaker relationships to the rest of that region's data. Generally, linear regression is able to capture the relationship between variables. Regression results in an ability to predict variable levels given the most recent value, each variable's relationships, and analysis of error terms. These individuals regression models based on historical data are evaluated using the information in Table 8.

Table 8. Regression Model Adequacy Check

	R^2	R^2_{adj}	F-Test	Normality	Homoscedasticity	Transformation
Arab Mobile Cell Subs	0.6017	0.5961	<0.0001	0.0027	0.5125	Log
Arab Population Density	0.8569	0.8471	<0.0001	0.0016	<0.0001	Square Root
Arab Percent Border Conflict	0.8363	0.8255	<0.0001	0.0132	0.0304	None
Arab Fertility Rate	0.8797	0.8725	<0.0001	0.0294	<0.0001	None
Arab Trade (% of GDP)	0.7261	0.7138	<0.0001	0.0018	0.8278	None
SE Asia Internet Users	0.8045	0.8020	<0.0001	0.0126	<0.0001	None
SE Asia Life Expectancy	0.9006	0.8994	<0.0001	0.0070	0.0125	None
SE Asia Mobile Cell Subs	0.7176	0.7148	<0.0001	0.0131	<0.0001	Fourth Root
SE Asia Infant Mortality Rate	0.8421	0.8404	<0.0001	0.0127	<0.0001	None
SE Asia Population Growth	0.9359	0.9284	<0.0001	0.0858	<0.0001	None
SE Asia Arable Lands	0.8210	0.8192	<0.0001	0.0149	0.0201	Log
SE Asia Avg Border Conflict	0.9351	0.9343	<0.0001	0.0567	<0.0001	None
SE Asia Freshwater per Capita	0.6633	0.6588	<0.0001	0.0046	<0.0001	Log

The Arab region's models achieved at worst, a R^2_{adj} of 59.61% for mobile cell subscriptions. This indicates the models, on a whole, sufficiently explain the variation of the predicted variables adjusted for the number of predictors in each model. The Cramer-Von Mises test found no variables' models for the Arab region that significantly satisfied the assumption of normally distributed residuals. On the other hand, mobile cell subscriptions and trade (% of GDP) regression models produced insignificant Breusch-Pagan tests where the null hypothesis was that the variance of the model's residuals is constant. To learn how well each variable was explained in a given country, the residuals were broken down for each nation's defined linear regression models. Within the Arab region, the variables, population density and trade (% of GDP) were the two with the highest residuals or least well defined. Based on the Breusch-Pagan statistic, they both scored the highest of the Arab variables of interest for being the least homoscedastic. Morocco and Syria each had the highest, abnormally large residuals within population density and trade (% of GDP) respectively, and the Untied Arab Emirates had large residuals for both of these variables of interest. The studentized

residuals, scaled by the exact standard deviation and used to identify outliers, found there to be outliers within four variables' models: percent border conflict, trade (% of GDP), population density, and fertility rate [23]. There was some overlap with the larger residuals and those studentized residuals greater than three which is the threshold used to identify an outlier [23]. Syria contained outlier residuals in percent border conflict and trade (% of GDP) while Tunisia included an instance of abnormal studentized residual in its model of percent border conflict. The United Arab Emirate's population density linear regression predicted an outlier within its studentized residuals. Yemen's fertility rate model was the only other instance of a studentized residual greater than three. It is difficult to attribute these poor regression fits to any direct conflict trends as they, along with other countries, all had varying likelihoods to enter conflict when the alternative futures were generated based on each variables of interest's regressions.

In Table 8, the Southeast Asia region's regression models are evaluated on model performance and ability to maintain the assumptions of normally distributed variance with constant variance. The models or each variable of interest achieved at worst, a R^2_{adj} of 65.88% for the freshwater per capita. Of the Cramer-Von Mises tests' results, only two variables, population growth and average border conflict didn't violate the linear regression assumption that the model's residuals followed a normal distribution. The Breusch-Pagan test found no models of the Southeast Asia countries' variables of interest to have significantly constant variance of their errors. Taking a closer look at the each individual model's residuals by country, various variables performed especially poor for a few nations. Based on the ranges of each models' residuals, GDP per capita stands out as disproportionately having the largest residuals compared to the other variables of interests' models. It is difficult to tell, with the greater number of variables of interest for this region, which specific countries contain poorly performing models, so the studentized residuals were calculated to identify outlying values that would indicate poor fits. Table 9 identifies each of the variables models that contained an instance of a studentized residual greater than three and the corresponding

country.

Table 9. Studentized Residual Identified Weak SE Asia Individual Regression Models

Country	Variables
Bangladesh	Avg Border Conflict
Brunei	GDP per Capita
China	Mobile Cell Subs
Papua New Guinea	Trade (% of GDP)
Philippines	Internet Users
Singapore	Population Growth and Population Density
Solomon Islands	Population Growth
Timor-Leste	Population Growth and Military Expend GDP

Table 9 shows that Singapore and Timor-Leste both contained studentized residuals greater than the standard cutoff for outlier analysis. Both of these nations, despite being poorly explained by the models for two variables each, had different conflict trends in the original data driven alternative futures. Singapore experienced an average absolute changes in all three conflict measurements, but Timor-Leste had similar conflict rates of transitions and likelihoods as historically observed.

4.3 Alternative Futures Models' Results and Evaluation

The regional, generic imputation style alternative futures model was first run with only operational bounds constraining the values of each variable of interest. The data's relationships and injected noise was able to determine the directions of each nation's future conflict transitions. To evaluate the model's statistical possibility, three metrics were calculated for each alternative future and averaged across all five repetitions. These statistics helped to understand the differences between future conflict behaviors and the observed historic data for each nation. Each evaluation metric was calculated for the past ten years of observed

data and future sixteen years of projected alternative futures. The rates of transitions were found over each past or future time period by counting the times a nation transitioned conflict status divided by the number of years in that period. For example, if a nation remained in conflict for the entirety of the projected future years then it would have transitioned zero times over sixteen years. Its rate of transitions for that nation's alternative future would be zero. The rates of transitions give insight into how closely the alternative futures are from the historic rates of conflict beginnings and stoppages in a specific nation. Conflict likelihood was calculated similarly; a nation's total number of years in conflict were divided by the total years in that period of time. By finding the likelihood of each nation's alternative future, comparing the models' conflict likelihoods provides the feasibility of that alternative future. Expanding on this idea, the most recent conflict likelihoods were calculated for the three past years (2012-2014) of observed data and the first three future, projected conflict statuses (2015-2017). Comparing the most recent conflict likelihood gives knowledge into the more short term similarities and differences between the alternative futures predictions and reality.

The Arab region in the past ten years has experienced 98 violent conflicts across seventeen countries: Algeria, Bahrain, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Qatar, Saudi Arabia, Syria, Tunisia, United Arab Emirates, West Bank, and Yemen. The generic imputation driven forecasting model projected country conflict from 2015 to 2030 based on the past ten years of completed data. The model projected future conflict replications multiple times to create five individual alternative futures for each country. The average predicted country conflicts across all five futures was 134.2 violent conflicts for the first ten years of future data, 2015-2025. The associated minimum and maximum total conflicts for a single future's over ten years were 125 and 145 respectively. Even on the lower end of projected future conflicts, there appears to be larger estimates of total conflicts that the data driven model predicts for the Arab region. Figure 2 plots the yearly average conflicts observed across all five repetitions of the future Arab regional data versus years.

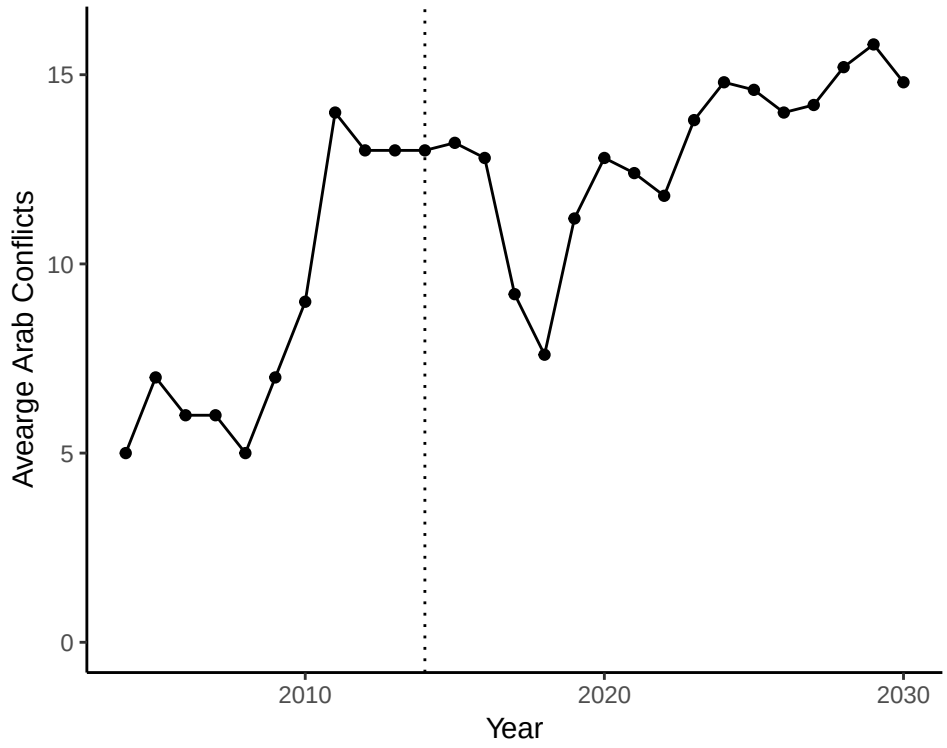

Figure 2. Arab Region's Original Alternative Futures' Average Yearly Conflicts

Historical data on and to the left of the dashed line represents the region's observations from 2004 to 2014. Over those years there was a large jump in average yearly conflicts around 2010 followed by a short couple year period of no trend. By year, there appears to be an initial dip in the average Arab conflicts once the projected, future years begin, but as the data are projected further into the future, average yearly conflicts are positively growing at a slightly decreasing rate. Therefore, number of conflicts is predicted to trend slightly upward.

The variable behaviors that appeared to indicate a tendency for conflict were low fertility rate, low trade (% of GDP), high mobile cell subscriptions, high population density, and high percent border conflict. These characteristics seemed common between the countries that experienced longer periods in conflict. Only one of the data driven model's five alternative futures predicted that Algeria, a country historically in a state of conflict for the past ten years would actually remain in conflict. The other four futures generally predicted that

Algeria would transfer out of conflict right away (2015) for about three to four years and then return to a state of conflict. These trends are seen in Figure 3 where each repetition's HIIK conflict score from the past and future are plotted over the years. Four alternative futures experienced nonviolence in Algeria immediately following the end of observed, past data.

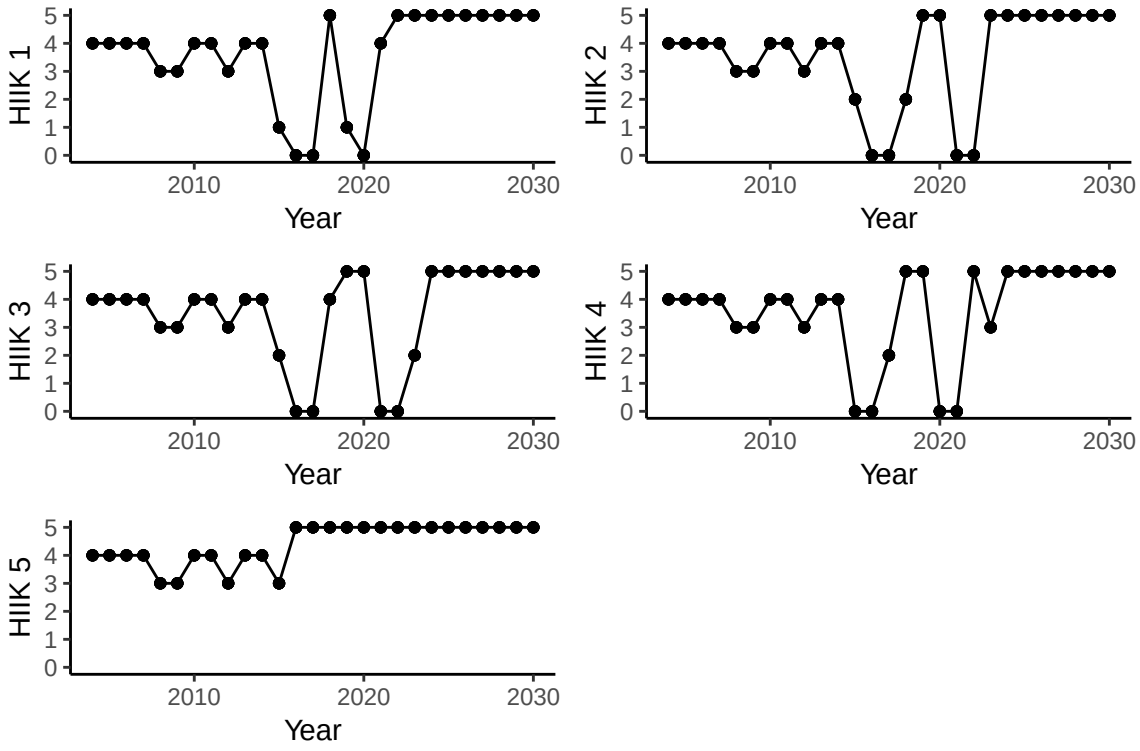

Figure 3. Algeria's Original Alternative Futures' Repeated HIIK Conflict Intensities

The average absolute changes for the three alternative futures evaluation measures showed the rate of transitions only changing by 0.138 across all the countries and replications. The conflict likelihood changed by an average of 0.387 from the past ten years of country conflict likelihoods. The more recent (three year) conflict likelihoods were slightly closer to the recent, historic regional conflict likelihood. In the short term and sixteen years of future conflicts the likelihood of violent conflicts deviated the farthest from the historic conflict likelihoods, but the future rates of transitions were relatively similar to the observed historical data's rates of conflict transitioning. The likelihood of conflict in the region changed, generally

trending upwards. The alternative futures on average experienced a higher likelihood of nation's being in conflict and remaining in that state.

The Southeast Asia regional alternative futures model was applied to iteratively project each variable into the future and shape possible conflict trends. In the past ten years, the Southeast Asia region has experienced 136 violent conflicts throughout its 28 nations. Allowing five replications of alternative futures and letting the data's relationships dictate their own trends over time, an average of 131.4 violent conflicts were predicted for the first projected, future ten years. There was a minimum of 126 and maximum of 138 violent conflicts predicted by an individual alternate future replication, and interestingly, the data driven alternative futures method applied in a different region outputted an opposite trend for a decade's overall conflict occurrences. Figure 4 plots the yearly average conflicts observed across all five repetitions of the future Southeast Asia regional data versus years.

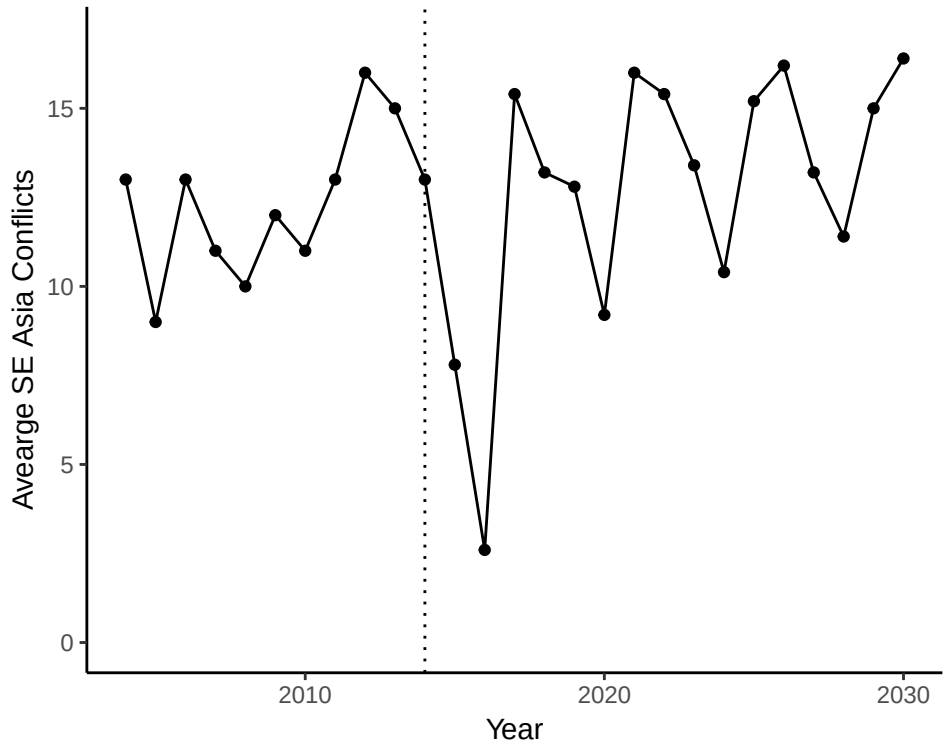

Figure 4. Southeast Asia Region's Original Alternative Futures' Average Yearly Conflicts

The past ten years of conflicts in Southeast Asia, on and to the left of the dashed line,

averaged around twelve conflicts per year. The first five years of projected, future ten years' average conflicts dropped to about ten average yearly conflicts. The first five years observed a lot of nations exit conflict, but despite the initial regional peace, the alternative futures once again saw conflicts rebound to a similar range. Although over ten years, the conflicts in Southeast Asia are forecasted to decline based of this model's results. The Southeast Asia's average yearly conflicts behave similarly to those in Arab region. There is a much greater relative decrease in conflict occurrences initially in Southeast Asia, but both region's average yearly conflicts declined at first, followed by growth at a decreasing rate.

There weren't any recognizable variables' behaviors that were common to country's with futures of prolonged non violence. The region seemed to experience few sustained years of peace or hostilities. Rather than experiencing years of peace or violence, most nations transitioned in and out of conflict at much higher rates than historically experienced. This resulted in an 0.347 absolute change in the rate of transitions, meaning countries were changing conflict statuses relatively quicker over the course of the alternative futures. There were no replicated countries' alternative futures that experienced a likelihood of conflict greater than 0.5 which is a vast regional decrease from the Arab nations' likelihoods of being in conflict. Of all the replicated alternative futures, North Korea had the highest average and max rates of transitions as well as one of the highest average future conflict likelihood. North Korea stood out in the region and experienced average absolute changes in all three of its individual alternative futures metrics. Figure 5 presents the HIIK conflict intensity scores of North Korea for each repetition against years. It's clear that North Korea, although being one of the countries to transition the fastest in Southeast Asia, transitioned between the far extremes of the HIIK conflict intensity barometer multiple times in all five repetitions. It trended towards the furthestmost environment of peace, HIIK conflict score of zero or the other extreme of war, HIIK conflict score of five. With the data driven model, nations in the Southeast Asia region underwent serious instability as nations' seemed more likely to be in either no conflict or war with few instances of intermediate HIIK conflict levels. This

may be an instance where analysis of a particular country may be better served outside of the alternative futures modeling construct given its uniqueness when compared to other countries in the region.

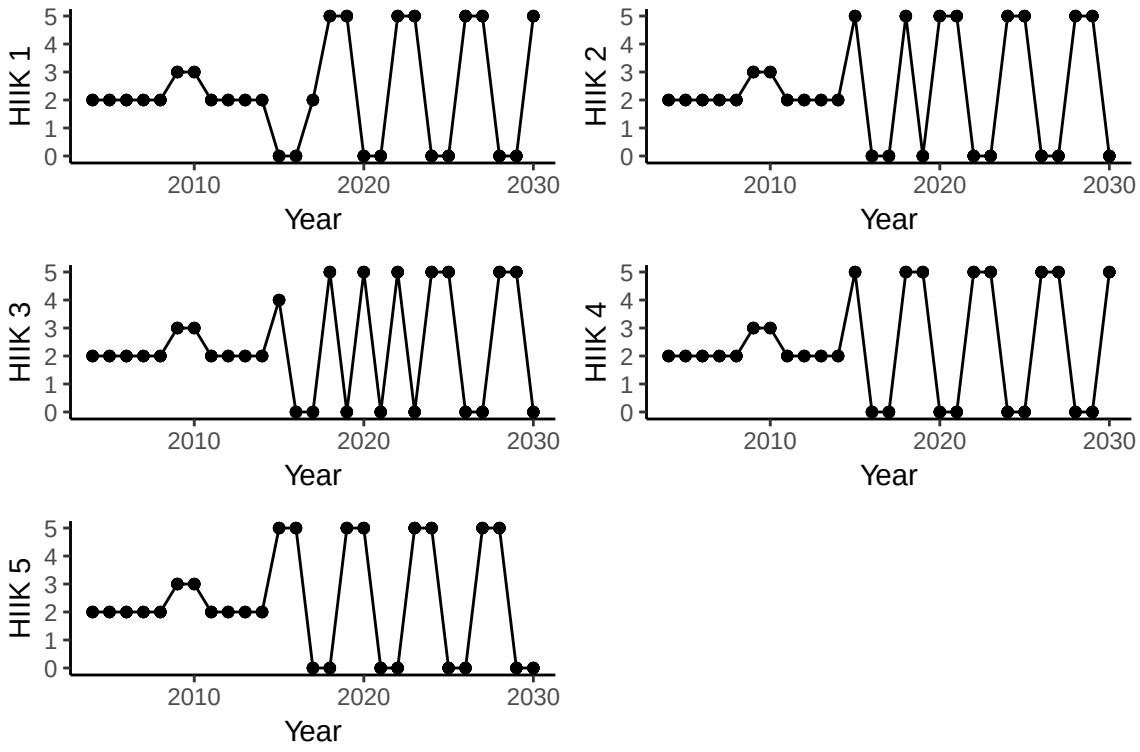

Figure 5. North Korea's Original Alternative Futures' Repeated HIIK Conflict Intensities

The model compared to the original ten years of observed data in the region on average across all countries and repetitions predicted. The most recent conflict likelihood comparison saw an absolute change of 0.417. The Southeast Asia region, overall will experience more countries transitioning in and out of conflict faster, and despite the initial slight decrease in conflict occurrences, the region's average likelihood of conflict rose over the forecast horizon. Although the two regions are largely different, they both predicted the smallest changes from their region's historic conflict behavior in rates of transitions and the greatest absolute differences in the likelihood of longer term conflict.

4.4 Arab Nations' Regional Scenarios

The three regional scenarios were tested using modified versions of the purely data driven alternative futures model. With the previously mentioned scripted behavior of the percentage of border conflicts of a given nation, the first regional scenario was tested to understand the regional impact of peaceful bordering nations. Despite the logical belief that peace besets peace, the results of this regional scenario were that an overwhelming majority of countries would enter and stay in conflict for the duration of the future sixteen years. The peaceful borders' alternative futures, averaged across all five replications forecasted 184.6 yearly instances of violent conflicts in countries over the first ten future years. This was nearly double the 98 observed instances of countries in violent conflict for the most recent ten years. The future conflicts with enacted peaceful bordering nations were about 50 conflicts more than the original data driven alternative futures. Figure 6 illustrates the region's propensity towards conflict when nation's borders become entirely peaceful. Within less than five years after the scenario begins, the entire region of seventeen nations, enter and remain in conflict for the remaining projected years.

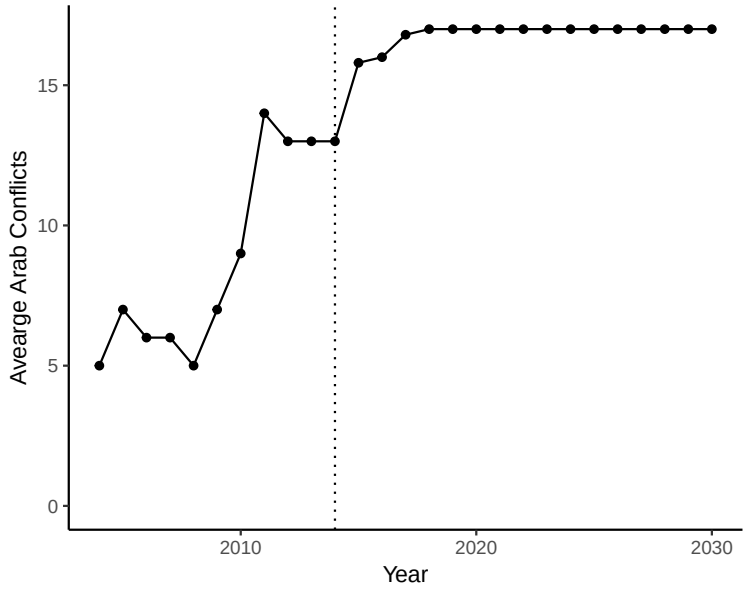

Figure 6. Arab Region's Scenario 1 Alternative Futures' Avearge Yearly Conflicts

Looking closer into each individual country, only three countries across all five repetitions experience a year not in a state of conflict. Oman, Qatar, and United Arab Emirates entered states of peace only within the first three years of future data while all other countries' conflict statuses were in conflict. The noticeable variable trends associated with all of the repeated states of conflict despite supposed peaceful bordering nations were low fertility rates, high population densities, and high mobile cell subscribers. Conflict transitions were influenced only when moderate levels of population density, fertility rate, trade (% of GDP), and mobile cell subscribers existed. The peace in bordering countries does not appear to bring peace to the Arab nations themselves. The peaceful border nations scenario had a average absolute change of 0.108 in the rate of transitions, a 0.454 absolute change in conflict likelihood, and the more recent conflict transitions changed by 0.188. Compared to the purely data driven alternative futures, preset peaceful borders predicted greater changes in only the long term likelihood of conflict. Otherwise, nation's when surrounded by peace were likely to be in conflict and transitioned between conflict statuses at rates more similar to the region's historic performance. For the most part, increased peace in bordering Arab countries generated alternative futures of increased violent conflicts and minimal conflict transitions back to peace.

The second scenario of complete regional democratic governance showed an improvement in the overall Arab conflict environment. There were an average of 130.8 yearly instances of nation's being in a state of conflict over the first ten years of projected data. This was a slightly smaller estimate than the data driven model, but the influence of democracies was still higher than the total number of conflicts in the most recent ten years of observed data. However, the all democratic governments produced average conflicts with a smaller range than the original data driven model indicating a smaller range of ten year conflict estimates. Figure 7 displays the region's average yearly conflicts over time. The first ten years of the future trended less positively than the original alternative futures. Over the future sixteen years, the democratic alternative futures experienced an average of 12.7 yearly conflicts

while the original, data driven futures averaged 13.0 conflicts per year. Democracy trended the region slightly more towards peace but still prompted an increase from the historically experienced average yearly conflicts (8.9).

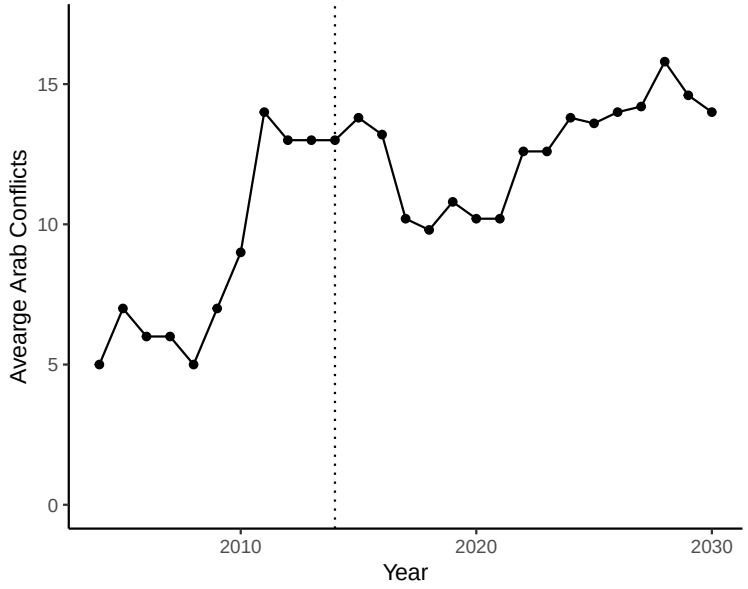

Figure 7. Arab Region's Scenario 2 Alternative Futures' Avearge Yearly Conflicts

Oman, Qatar, Bahrain, Kuwait, Lebanon, and Morocco all experienced alternative futures completely in conflict with no transitions. The years with continued violent conflicts and no transitions out of conflict were generally explained by high mobile cell subscribers, high population density, high percent of bordering conflicts, low fertility rates, and low trade (% of GDP). It appeared that there was no significant impact that prescribed democracies had on the overall conflicts within a region. This research's three conflict measures had similar average absolute changes as the original imputed style forecasting method: rates of transitions, conflict likelihoods, and more recent conflict likelihoods. The only average statistic that didn't change from the dictated democratic regimes were the recent, three year conflict likelihoods which projected an identical absolute change of 0.341 in conflict likelihoods. The imposed state of democracy only improved the model's ability to predict conflict likelihoods more closely to the regions previous three years of conflict likelihoods. Bahrain, Lebanon, and Morocco all saw no changes in recent conflict likelihoods across all

five repetitions from the historic conflict transitions. This appears that with the influenced democracies, the model produced recent (three year) conflict transitions identical to the recent previous observed conflict transitions when a given nations had spent all their previous recent years in a state of conflict. The countries that experienced the largest differences from the historic conflict behaviors in transitions were Oman, Qatar, United Arab Emirates, and West Bank. Two of these countries, not surprisingly received predictions of at least one alternative future that was in conflict all sixteen years. This reassured the model's tendency towards predicting statuses of conflict over peace region and impact of democracies within the Arab. Neither of these nations that in future years remained solely in conflict had previously been democratic.

The third regional scenario simulated the reduction of trade as a percentage of GDP for each country in the Arab region. Future trade projects were set to the minimum historic trade levels within the Arab region to force each country to trade as infrequent as operationally feasible. Compared to the instances of conflict in the future ten years of original, data driven alternative futures, the decreased trade model predicted very similar violent conflict occurrences. There were 132.4 average conflicts in the first ten years of future data predictions which was only on average about two conflicts less than the original alternative future model. Figure 8 plots the region's average yearly conflicts to understand the future conflict trends. The average yearly conflicts across all future years was only a downtick from the original alternative futures model.

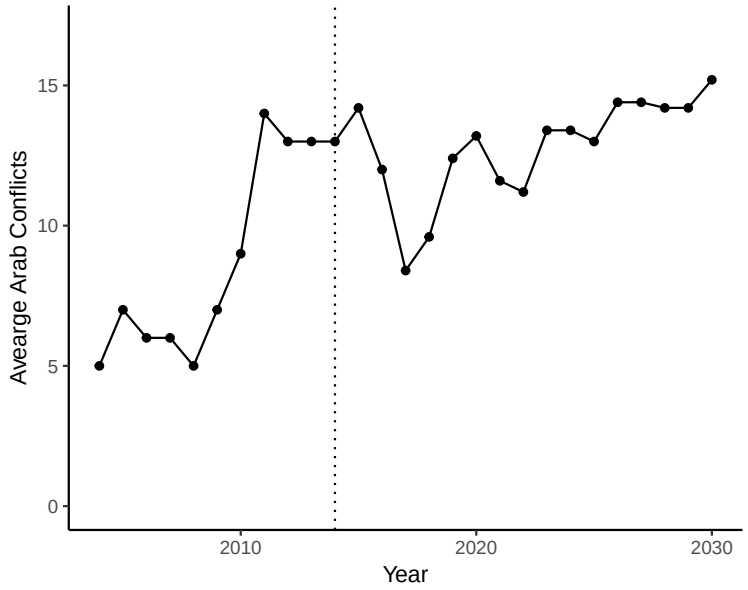

Figure 8. Arab Region's Scenario 3 Alternative Futures' Avearge Yearly Conflicts

Bahrain, Libya, and Yemen were all predicted for all five repetitions to be in a state of conflict over the course of the entire forecast horizon. Of these three countries' observed historic conflict trends, only Yemen had experienced a perfect historic conflict likelihood, meaning it was in conflict the entire past ten years. Bahrain and Libya both have observed conflict likelihoods of 54.545% and 36.364%. This indicated no clear relationship between the historic conflict likelihood and set decreased trade within each country that would lead to consistent future years in conflict without transitions. The preset decline in future trade (% of GDP) generated results closer to the region's historic average rates of transitions, conflict likelihoods, and recent (three year) conflict likelihoods. All of the average absolute changes were the same as the original data driven alternative futures while having higher differences between the rates of transitions and recent conflict likelihoods. Only the overall forecast conflict likelihoods changed a few points from the historic conflict likelihood. This is likely due to the decline in trade predicting a few less conflicts and historically the Arab region having lower conflict occurrences and likelihoods than any alternative futures developed. Despite a marginally less ten year conflict average, there weren't any identifiable differences that closed trade had on the region's original data driven conflict statuses and transitions.

Compared to the Southeast Asia region, the Arab nation's didn't on average spend as much of their GDP's on trade to begin with, which may make their conflict statuses robust to any cessation of trade.

4.5 Southeast Asia's Regional Scenarios

The three regional scenarios aimed to test for the same three possible political and economic realities on the Southeast Asia region. Peaceful bordering nations were fixed for the future data projections in order to see if peace would likely inspire more peace within a region. The binary border conflict and average bordering country conflict variables were hard coded to not allow any bordering nation exceed a HIIK score of three which means no violent conflicts would surround a given nation during the sixteen years of projected data. The peaceful border scenario predicted an average of 113.2 violent conflicts in Southeast Asia region's first ten forecasted years of data. Of the five repetitions, there was a high estimate of 124 and low of 108 total conflicts. These summary statistics all shifted towards much less average conflicts versus the free flowing original conflict predictions, but compared to the total, 136 violent conflicts in the past ten years of observed data, peaceful borders estimated much less future conflicts would arise in the next decade. The data, with its own relationships seems to already predict future conflicts slightly trending downwards even without peace abutting the region. Figure 9 highlights the large decrease in average yearly conflicts brought about by peaceful border nations. The historic data observed 12.4 average yearly conflicts while over the course of the sixteen future years, peaceful borders guaranteed only 10.8 average conflicts occurred each year. This was also a decrease from the original alternative futures' 12.7 average yearly conflicts over the entire forecast horizon.

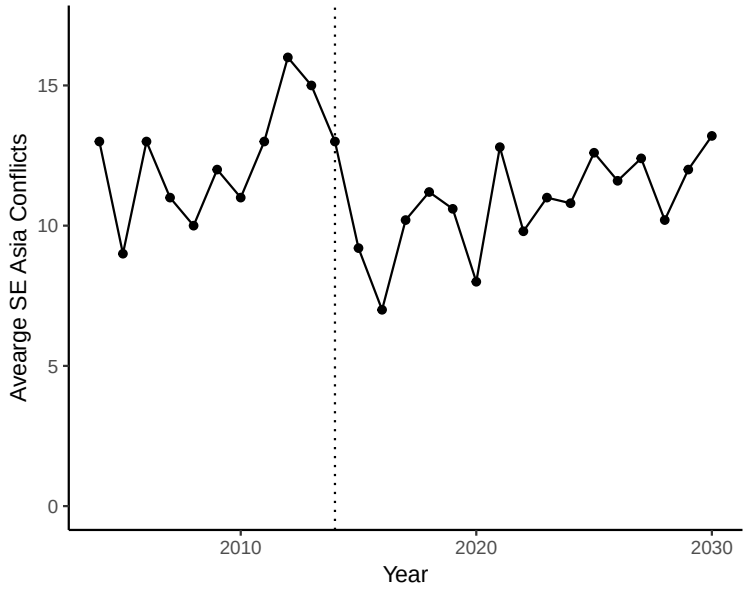

Figure 9. Southeast Asia Region's Scenario 1 Alternative Futures' Average Yearly Conflicts

Between the rates of transitions, conflict likelihoods, and recent conflict likelihoods, the first scenario didn't majorly alter any of the average absolute changes found by the original forecasting technique. The first scenario predicted very similar rates of transitions and conflict likelihoods to the original, historic ten years of observed data. Only the most recent, three years data conflict likelihoods were less similar to the original data's most recent likelihoods than the novel imputation style alternative futures. This means that in the short term, when peace surrounds a Southeast Asian nation, its conflict likelihoods are less similar to recent historic conflict likelihoods, but the further into the future those conflict likelihood's trend to be similar to the novel imputation forecasts. Kiribati and Mongolia were the nations with the highest conflict likelihoods, but even these two nations had less than a 0.5 likelihood of conflict. Compared to the Arab region, these low conflict likelihoods indicate that Southeast Asia's conflicts are generally less probable. Kiribati, as an island would in theory be unaffected by peaceful bordering nations, and this rise was confirmed by its conflict likelihoods only being marginally different from the original data driven alternative futures. The main insights from implementing peace in bordering nations were countries are less likely to be in states of conflict in the short term and less total average conflicts occur in the

next ten years.

The second scenario aimed to understand how set democratic governments would shift the conflict environment of the region. The regime type for each nation was set to democratic for the future years and freedom scores restricted to not be lower than that of any observed democratic nation's in the region. The prescribed democracies produced a downward ten year shift in total conflicts compared to the historic count. Across all repetitions, an average of 134.4 instances of nations in states of conflict were estimated which is less than the historic 136 observations. This is a slightly higher prediction than the original data driven forecasted method meaning with democracies in place, nation's conflicts are still going down but slightly less than if the alternative futures were just driven by current data relationships. Figure 10 brings to light the nearly cyclical trend of average yearly conflict occurrences caused by the rise of democracy. The democratic alternative futures averaged 13.0 conflicts per year which wasn't an extreme rise from the original alternative futures 12.7 average yearly conflicts. There were however much more drastic spikes and falls which was represented by the calculated regional higher rates of transition between conflict states.

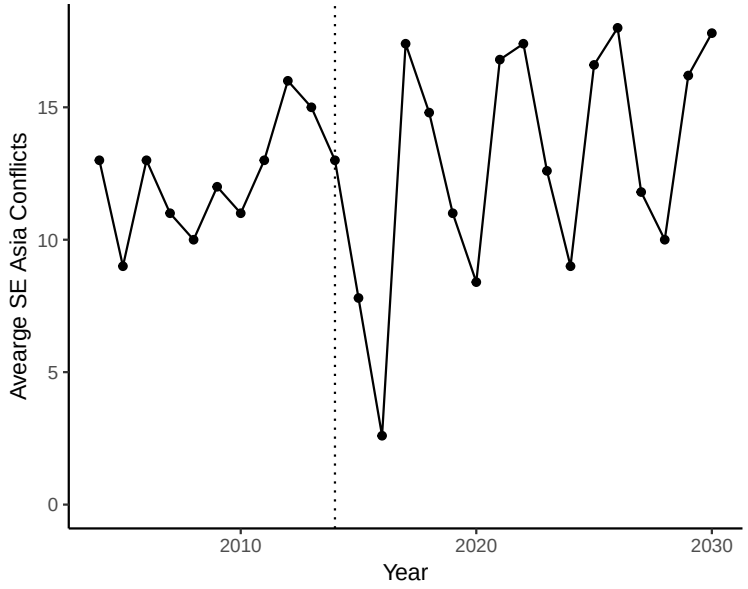

Figure 10. Southeast Asia Region's Scenario 2 Alternative Futures' Average Yearly Conflicts

The democratic scenario predicted very similar, within a one or two tenths, absolute changes of rates of transitions and recent and longer term conflict likelihoods to that of the original alternative futures. Brunei, The Philippines, and Laos all had the highest average likelihoods of conflict which was interesting since except for The Philippines, none of these nations were previously governed by democracies. It is possible that the influence of a regime change destabilized these nations rather than fostering peace there. Higher changes in the rates of transitions were another possible outcome of the disruptive nature that may spur from implementing democracies in a region that's historically been more autocratic. North Korea, Mongolia, and Samoa, of which only North Korea was previously an autocracy, observed future rates of transitions higher than any other nation's rates. With the shift towards democracy, the region experienced only a few less conflicts in the ten years of alternative futures and consistent conflict measures with the original forecasting results.

The third regional scenario imitated the impact that huge declines in trade as a percentage of GDP would have on the Southeast Asia region of the world. Trade (% of GDP) was limited to the lowest, historical trade levels historically practiced in the region to learn about the resulting conflicts trends. With a serious downturn in trade, the region experienced an average, estimated 136.6 total conflicts in the first ten years of the future; this predicted average included counts ranging from 131 to 142 between the replications. This was the first scenario to predict conflict to actually rise in the region compared to the 136 instances of violent conflict observed in the past ten years. It is expected that over the course of ten years, tensions in the region and countries conflicts would rise due to a regional cessation of trade. Compared to the alternative futures produced from unrestricted data relationships, the decline in trade resulted in higher average absolute changes of rates of transitions and conflict likelihoods from the past ten years of conflicts. Nations are transitioning in and out of conflict more and have a higher chance of being in conflict when their trade ceases to a near halt. Figure 11 visualizes the similar average yearly conflicts between the historic and reduced trade alternative futures. Compared to the historic average yearly conflicts (12.4), a

cessation of trade produced a slight uptick to 12.6 average yearly conflicts which is induced by the positively trending later, future years.

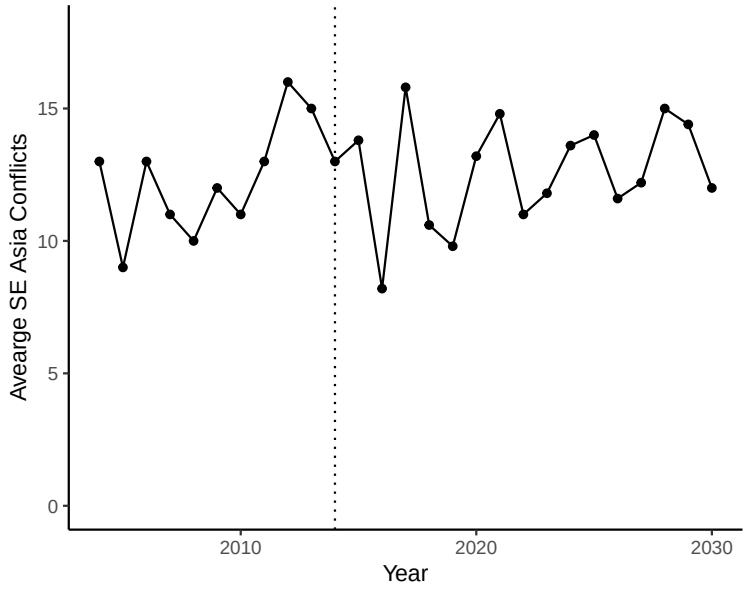

Figure 11. Southeast Asia Region's Scenario 3 Alternative Futures' Average Yearly Conflicts

None of the three statistics used for evaluating alternative futures estimated the decreased trade would be more similar to the historic data than the novel imputation approach's results. This could be attributed to ceased trade having a sweeping effect on a nation's conflict status due to Southeast Asia relatively allocating more of its GDP on trade. Bhutan, The Philippines, and Singapore were the nations with the highest average likelihoods of conflict in the alternative futures with closed off trade. Singapore historically had spent more of their GDP on trade than any other nation, so is expected that this regional scenario would make them a likely candidate to be heavily impacted. It was hard to determine an exact cause for the higher conflict likelihoods, but even all of the six nations that had historically observed no status of conflicts in the past ten years each entered at least one state of conflict during the sixteen years of the future. That could be due to the model's propensity to predict conflict over non conflict or the actual negative effects of nearly closing off trade in the region.

V. Conclusions

5.1 Overall Conclusions

This study developed a methodology to impute open source data previously found to be influential for predicting country conflict in two regional country groupings and found multiple alternative futures based on an explanatory forecasting approach. Individual linear regression models were built and evaluated for each of the variables of interest in the Arab and Southeast Asia regions. In each region, sixteen years of future data was projected and iteratively predicted upon to find the conflict transitions of every nation. Three regional scenarios were posed to understand shifts in nations' behaviors and impact on the future conflict statuses. The results of the study led to several conclusions which are summarized in answering the original research questions.

Question 1: How should the data set utilized in the Neumann study be imputed?

Utilizing and testing the `MICE` package in R on the data yielded interesting results. The main tests used to compare the multiple imputed data sets by each MICE method were used to investigate how closely the data followed the distribution of the observed data. The polity IV variable's missing observations were filled with knowledge provided by the regime type of a country. The MICE algorithm was applied to each region individually with all of the possible variables. This was done due to the assumed similarities between countries and information provided to the imputed values from all of the variable relationships within the a given region. Five data sets of imputed values were generated for each of the five MICE algorithms and analyzed separately. Kolmogorov-Smirnov, non-parametric, 2-sample Anderson-Darling tests, and diagnostic plots were used to decide the individual MICE algorithms to address

each variable's missingness. It seemed that generally variables containing smaller levels of missingness were imputed using predictive mean matching while the variables with higher percentages of missingness tended to be imputed using either classification and regression trees or random forest algorithms. Few variables were close to following a normal distribution or had imputed values similar to the observed distribution, so Bayesian Linear Regression was the MICE algorithm least chosen to impute the data. Multiple imputations by chained equations showed to be a useful method of imputation for data with many variables stricken with varying levels of missingness.

Question 2: How can this research develop regression models for each region's variables of interest?

Stepwise regression was the main technique used to develop parsimonious models for each region's individual variables of interest. Interaction terms were tentatively applied to better define a variable's relationship while avoiding overfitting and creating exhaustive models. All of the models within the Arab region included interaction terms while not every one of the Southeast Asia's models benefited from including variable interactions. The R^2_{adj} was the main metric used to evaluate the reduced regression models, and between both regions the R^2_{adj} achieved were acceptable to explain the relationships within the data. The lowest R^2_{adj}'s between both regions' models was 59.61%. The assumptions of linear regression were violated by a majority of the models in both regions. None of the five models in the Arab region satisfied the assumption of normally distributed residuals while only two of the eight models in the Southeast Asia region fulfilled it. The assumption of constant variance of a model's residuals was only met by two of the five Arab regional models, but none of the Southeast Asia variables' models maintained homoscedasticity. Multiple transformations were tested and applied depending their impact to a model's R^2_{adj} and assumptions of linear regression. This research relaxes the assumptions of linear regression in order to proceed with explaining each variable of interest based on others in the region and predicting point estimates for the alternative futures. Outlier analysis based on studentized residuals found

a few countries in each region that were poorly fit by certain variables' models, but these countries' alternative futures of conflict behavior didn't stand out among the other countries acceptably explained by the variables' models.

Question 3: What insights, nations susceptible or impervious to transitioning in or out conflict are identified by the generated regional conflict alternative futures?

In the Arab region, the data alternative futures driven by the data's relationships predicted an overall ten year increase in conflicts compared to the observed past ten years of data. On average the alternative futures subject to no regional scenarios predicted 134.2 conflicts to arise in the next ten years opposed to 98 instances of conflict observed across the Arab region. Conflict in the region was predicted to rise overall, and Bahrain, Kuwait, Lebanon, and Tunisia were the four nations of the sixteen in the region that had the highest likelihoods of conflict. Kuwait, Oman, Qatar, United Arab Emirate, and West Bank were the countries with the least similar likelihoods of conflict compared to the nations' previous behaviors. The only measure that the purely data driven alternative futures didn't predict would have a large absolute change across the region was a smaller predicted change (0.138) in the rates of transitions compared to the historic rates. The Arab region was transitioning at a similar rate in and out of conflict but there was a trend of increased nations' likelihoods towards being in conflict over time.

The Southeast Asia region experienced different conflict trends than the Arab region's general rise in violent conflict. The alternative futures driven by unrestricted relationships between the variables of interest predicted that across all countries and repetitions there would be 131.4 conflicts in the next ten years. This is a downturn from the past ten years of observed 136 conflicts in the region. Along with having a regional long term average absolute change in conflict likelihood of 0.351, the model produced alternative futures' with short term conflict likelihoods even less than the historic likelihoods of conflict (0.417). The average absolute change of rates of transitions in the region rose (0.351) as well, meaning

nations are transitioning conflict statuses much more often than historically recorded. Based on comparisons to the regional rates of transitions and long term conflict likelihood, North Korea scored the highest making it a concerning location for future conflict. The region's conflicts driven by the novel imputation alternative futures appeared to decrease in total regional conflicts for the next ten years, and during the entirety of the forecast horizon, nations behaved more erratic than they have in the past.

Question 4: How robust are the alternatives futures of conflict transitions when tested by regional what-if analysis scenarios?

The three scenarios generally tested on regional variables of interest were the impact of peaceful neighboring nations, democratic regime type, and a decline in trade (% of GDP). With peaceful bordering countries, the region's conflicts over a ten year period nearly doubled which still, as the original alternative futures predicted, indicates a rise in conflicts except at an expedited rate. Contrary to the logic that peace would spread across borders, this what-if scenario had adverse effects on the Arab region's conflict landscape. The impact of democracy spreading through the region didn't have a real impact on the future conflict predictions compared to the 134.2 conflicts forecasted by the original alternative futures model. Democracy slightly lowered the ten year average conflict future estimate to 130.8 but still predicted a stark rise in total conflicts from the historic count. The declined trade (% of GDP) in the Arab region's nations had little impact on average conflicts. The scenario produced 132.4 conflicts to occur in the next ten years which is higher than historically shown but similar to the original alternative futures approximate. In general, the Arab region's future ten year's of conflict wasn't too sensitive to the regional scenarios besides peaceful bordering nations driving up the violent conflicts. The region's conflicts were still predicted to increase by a minimum of about thirty violent conflicts regardless of the what-if scenario.

The Southeast Asia region was slightly more responsive to the preset scenarios affecting bordering nations, government, and trade. Except for the implementation of peaceful

bordering nations, the region experienced an average increase in conflict occurrences over ten years compared to the original alternative futures' estimates. The peaceful bordering nations produced a future ten year average of 113.2 conflicts which was the only scenario to decrease from the historically observed 136 conflicts. The influence of democracy caused 134.4 conflicts on average over the future ten years while the decline in trade produced the highest future total conflicts. The cessation of trade (% of GDP) scenario estimated 136.6 average conflicts over the next ten years which was the only scenario to exceed the historic ten years of total conflict. All of the regional scenarios resulted in greater average absolute changes from the historic data. The most recent, three years of conflict likelihoods and those over the course of the future data were all larger when compared to the original alternatives futures' scorings. Only the scenario of peaceful border nations generated regional rates of transitions more similar to the rates seen in the past ten years. The Southeast Asia region saw the greatest conflict likelihood changes when each nation's trade was restricted. This result made sense considering that compared to the Arab region, Southeast Asia's countries historically spent a much higher average percentage of their GDP on trade, making it a vulnerability of the region. The installation of democracies seemed to not have an overly negative or positive impact on the region's predicted conflicts. The Southeast Asia region was largely predicted to decrease in conflict over the next ten years especially if neighboring nation's become more peaceful. The only outcome of conflicts surpassing the historic averages occurred when the region's trade (% of GDP) nearly collapsed.

5.2 Significance of Research

In Issac Asimov's science fiction novel, *Foundations*, a new branch of mathematics is able to predict the future, but only on a large scale. The premise of the book is that the characters analyze their galaxies alternative futures, calculate that they are likely doomed, and take necessary actions to ensure an alternative path that stalls their society's destruction.

The alternative futures inform the characters' decisions to ensure a different statistically identified reality emanates. *Foundations* celebrates a fictitious, futuristic form of mathematics while in today's age predicting the future involves much greater uncertainty. There is a finite power to calculating the future and an inherent unknowing of whether the future should even resemble the past. This research develops sixteen years of alternative futures for two world regions based on past data from 2004-2014, encompassing 45 countries. These alternate futures provide some regional conflict trends for consideration by military strategists and future regional operations. The resulting conflict landscapes are meaningful, but with the constantly changing data patterns and regional relationships, the resultant conflict transitions provide information for sensible cogitation. The regional and global military policies' grave impact mean that deeper forecasting analyses and expert opinion must be considered to more convincingly inform decision making.

5.3 Future Research

As part of continuing research, this study recommends multiple areas for deeper analysis into the alternative futures of country conflict. The first area for continuing this research is to expand the projected variables considered for each region. The scale of this entire project could be broadened to include all of the available data. This research only developed regression models and projected the variables identified as influential in the Neumann [3] models, but all of the available variables could be used to define and project the entire data set into the future. There is a greater chance of developing better fitting regression models when more explanatory variables are considered, and a variety of regressors could raise the level of fidelity associated with these alternative futures.

In addition to expanding the variables considered to explain each region, an alternative model reduction technique could be applied to find parsimonious models for a region. Stepwise regression was the foremost model reduction method used in this study, but there

are alternate techniques that may produce different, simpler regional variable relationships. Lasso regression is one technique which could be explored to explain each region differently and project data into the future. Any improvements that can be made to the individual variables' linear regression models would produce more convincing projected data and alternative futures. This research's imputed alternative futures method could still be developed upon in the same manner that univariate missing data imputation methods have been progressed. This research took a predicted point estimate plus noise approach to defining the yearly estimates of each variable, but Van Buuren [17] identified two other univariate imputation methods that incorporate more uncertainty: Bayesian multiple imputation and bootstrap multiple imputation. Both of these methods take the prediction plus noise formulation one step further by adding more parameter uncertainty. Expanding the imputation theory behind the alternative futures may be a possible area of future research. Branching off from an imputation approach, future research could analyze an entirely different supervised learning method's ability to iteratively predict yearly estimates. Neural Networks are a powerful supervised learning method that has been used to impute data and could provide an alternate way of projecting data.

Social Sciences have established theories that relate violent conflicts to predominant economic or political factors. This research could be vectored to explore the social sciences theories, and how the social sciences' beliefs on violent conflicts are supported or rejected by analytical techniques. Incorporating more uncertainty and information into the alternate future generation would only enhance decision makers knowledge towards developing defendable strategic and operational plans.

Appendix A

5.4 Variable Information

Variable	Source	Description
BirthRate	World DataBank: World Development Indicators	Birth rate, crude (per 1,000 people)
DeathRate	World DataBank: World Development Indicators	Death rate, crude (per 1,000 people)
FertilityRate	World DataBank: World Development Indicators	Fertility rate, total (births per woman)
InternetUsers	World DataBank: World Development Indicators	Internet users (per 100 people)
LifeExpectancy	World DataBank: World Development Indicators	Life expantancy at birth, total (years)
MobileCellSubscriptions	World DataBank: World Development Indicators	Total mobile celluular subscriptions
InfantMortalityRate	World DataBank: World Development Indicators	Mortality rate, infant (per 1,000 live births)
YouthPopulation	World DataBank: World Development Indicators	Population ages 0-14 (% of total)
PopulationGrowth	World DataBank: World Development Indicators	Population growth (annual %)
RefugeeOrigin	World DataBank: World Development Indicators	Refugee population by country or territory of origin
ArableLand	World DataBank: World Development Indicators	Arable land (hectares per person)
ImprovedWater	World DataBank: World Development Indicators	Improved water source (% of population with access)
PopulationDensity	World DataBank: World Development Indicators	Population density (people per sq. km of land area
Trade(% of GDP)	World DataBank: World Development Indicators	Trade as a percentage of a country's GDP
Unemployment	World DataBank: World Development Indicators	Unemployment, total (% of total labor force)(national estimate)
PctBC	Developed from HIIK intensity levels and border information	Border conflict measure averages the % of border shared with a country times that nation's associated HIIK score
AvgBC	Developed from HIIK intensity levels and border information	Border conflict measure for the average HIIK conflict intensity level surrounding a country
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
EthnicDiversity	Shallcross Database	Percent of dominant ethnic group
ReligiousDiversity	Shallcross Database	Percent of dominant religious group
FreshWaterperCapita	Shallcross Database	Cubic meters, average of 2007,2012,2013 data
GDPperCapita	World DataBank: World Development Indicators	GDP per capita growth (annual %)
MilitaryExpendGDP	World DataBank: World Development Indicators	Military expenditure (% of GDP)
RefugeeAsylum	World DataBank: World Development Indicators	Refugee population by country or territory of asylum
CaloricIntake	UN Food and Agriculture Organization	Caloric intake (kcal/capita/day)
2YrConflictIntensity	Developed from HIIK intensity levels	Change in HIIK conflict intensity between current year of observation and HIIK intensity level of current year -2
FreedomScore	Developed, Freedom House	Combined average of normalized civil liberties and normalized political rights
FreedomTrend2Yr	Developed from freedom score	Difference between freedom scores of current year -2 and current year -1
FreedomTrend3Yr	Developed from freedom score	Difference between freedom scores of current year -3 and current year -1
FreedomTrend5Yr	Developed from freedom score	Difference between freedom score of current year -5 and current year -1
DemGovType	Developed from Government Type	Indicator variable denoting if an observation has a democratic government or not
GovernmentType	Developed from polity variable	Mapping of original polity variable into six categories
RegimeType	Shallcross Database	Static variable indicating the type of regime for a country

5.5 Regional Logistic Regression Models

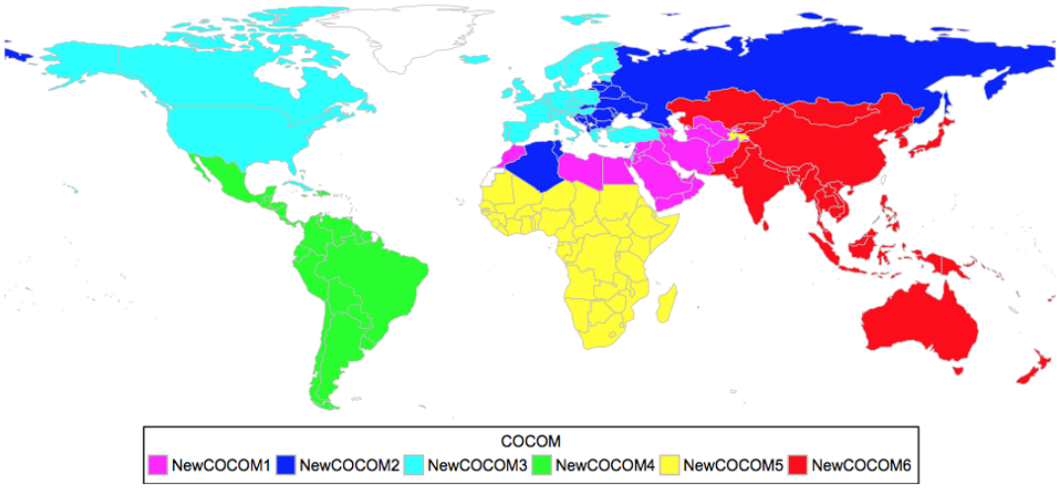

Figure 12. Regional Country Grouping

Table 10. Arab Region In Conflict Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.5600	2.2340	-2.29	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 11. Arab Region Not In Conflict Model

	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 12. Southeast Asia Region In Conflict Model

	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 13. Southeast Asia Region Not In Conflict Model

	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	3.9917	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

Appendix B

Table 14. Arab Region Average Alternate Futures' Performances

	FROT*			PROT†	FCL‡			PCL§
Country	Min	Average	Max		Min	Average	Max	
Algeria	0	0.15	0.1875	0	0.625	0.725	1	1
Bahrain	0	0	0	0.0909	1	1	1	0.5455
Egypt	0.0625	0.2125	0.5	0.2727	0.5	0.7875	0.9375	0.7272
Iraq	0.0625	0.275	0.5625	0	0.4375	0.7	0.875	1
Jordan	0.125	0.3	0.5625	0.2727	0.4375	0.6875	0.875	0.5455
Kuwait	0	0.025	0.125	0.0909	0.875	0.975	1	0.3636
Lebanon	0	0.075	0.25	0	0.625	0.9	1	1
Libya	0.125	0.3125	0.4375	0.0909	0.5	0.6625	875	0.3636
Morocco	0	0.2375	0.4375	0.2727	0.2625	0.7375	1	0.5455
Oman	0.125	0.25	0.375	0.1818	0.5625	0.7375	0.875	0.0909
Qatar	0.1875	0.2375	0.3125	0	0.6875	0.7375	0.8125	0
Saudi Arabia	0	0.175	0.375	0.1818	0.625	0.7625	1	0.0909
Syria	0.0625	0.2	0.375	0.2727	0.5625	0.7375	0.875	0.4545
Tunisia	0	0.125	0.25	0.0909	0.75	0.875	1	0.3636
United Arab Emirates	0.1575	0.2125	0.3125	0	0.625	0.675	0.6875	0
West Bank	0.125	0.3	0.4375	0	0.5625	0.725	0.875	0
Yemen	0.375	0.4125	0.4375	0	0.5625	0.5875	0.625	1

* Future average rate of transitions
† Past ten years rate of transitions
‡ Future conflict likelihood
§ Past ten years conflict likelihood

Table 15. Arab Region Average Alternate Futures' Recent Conflict Likelihoods

	FRCL*			PRCL†
Country	Min	Average	Max	
Algeria	0	0.2	1	1
Bahrain	1	1	1	1
Egypt	0.6667	0.8667	1	1
Iraq	0	0.4667	1	1
Jordan	0.6667	0.6667	0.6667	1
Kuwait	0.6667	0.9333	1	1
Lebanon	0.6667	0.9333	1	1
Libya	0.6667	0.8667	1	1
Morocco	0.6667	0.7333	1	1
Oman	0.3333	0.6	0.6667	0
Qatar	0.3333	0.6	0.6667	0
Saudi Arabia	0	0.4	1	1
Syria	0	0.6	1	1
Tunisia	0.6667	0.8	1	1
United Arab Emirates	0.3333	0.3333	0.3333	0
West Bank	0.3333	0.7333	1	0
Yemen	1	1	1	1

* Most recent future three years conflict likelihood
† Most recent past three years conflict likelihood

Table 16. SE Asia Region Average Alternate Futures' Performances

	FROT*			PROT†	FCL‡			PCL§
Country	Min	Average	Max		Min	Average	Max	
Bangladesh	0.375	0.4792	0.5	0.1818	0.375	0.3958	0.4375	0.9091
Bhutan	0.25	0.3229	0.4375	0.0909	0.1875	0.2917	0.4375	0.0909
Brunei Darussalam	0.125	0.375	0.5	0.0909	0.0625	0.2813	0.5	0.1818
Cambodia	0.375	0.4896	0.6875	0.1818	0.3125	0.3958	0.4375	0.7273
China	0.3125	0.3854	0.4375	0.1818	0.3125	0.3542	0.375	0.9091
North Korea	0.25	0.3125	0.375	1818	0.1875	0.2396	0.3125	0.1818
Fiji	0.375	0.4896	0.5625	0	0.375	0.4375	0.5	0
India	0.375	0.4479	0.5	0	0.3125	0.3854	0.4375	1
Indonesia	0.4375	0.5104	0.625	0	0.3125	0.4271	0.5	0
Kiribati	0.0625	0.2604	0.5625	0.4545	0.0625	0.1667	0.375	0.6364
Laos	0.3125	0.3333	0.375	0.0909	0.1875	0.3021	0.375	0.3636
Malaysia	0.3125	0.5104	0.875	0.2727	0.1875	0.3958	0.5	0.2727
Maldives	0.375	0.5104	0.9375	0	0.375	0.4375	0.5	0
Micronesia	0.375	0.4479	0.5	0	0.3125	0.3646	0.375	0
Mongolia	0.0625	0.3542	0.625	0	0.0625	0.2813	0.5	1
Myanmar	0.375	0.4271	0.4375	0.0909	0.3125	0.4063	0.4375	0.9091
Nepal	0.4375	0.4896	0.5	0.0909	0.375	0.4063	0.4375	0.3636
Papua New Guinea	0.5	0.6875	0.875	0	0.375	0.4688	0.5	1
Philippines	0.25	0.375	0.5	0.1818	0.125	0.3021	0.5	0.0909
Samoa	0.4375	0.4583	0.5625	0	0.25	0.3542	0.4375	0
Singapore	0.4365	0.4792	0.5	0.1818	0.25	0.4271	0.5	0.0909
Solomon Islands	0.3125	0.4375	0.5	0.3636	0.1875	0.3333	0.5	0.8182
Sri Lanka	0.3125	0.3438	0.4375	0	0.1875	0.3021	0.4375	1
Thailand	0.4375	0.4792	0.5	0.3636	0.3125	0.4063	0.5	0.3636
Timor-Leste	0	0.0936	0.25	0.1818	0	0.09375	0.25	0.0909
Tonga	0.25	0.3646	0.5	0	0.25	0.3542	0.5	0
Vanuatu	0.375	0.4063	0.5	0.2727	0.25	0.3646	0.4375	0.3636
Vietnam	0.375	0.4792	0.5	0.1818	0.375	0.3958	0.4375	0.9091

* Future average rate of transitions
† Past ten years rate of transitions
‡ Future conflict likelihood
§ Past ten years conflict likelihood

Table 17. SE Asia Region Average Alternate Futures' Recent Conflict Likelihoods

	FRCL*			PRCL†
Country	Min	Average	Max	
Bangladesh	0	0.2778	0.3333	1
Bhutan	0	0.0556	0.3333	0
Brunei Darussalam	0.3333	0.3889	0.6667	0.6667
Cambodia	0	0.1111	0.3333	1
China	0	0.1111	0.3333	1
North Korea	0	0	0	0
Fiji	0	0.3333	0.6667	0
India	0	0.1667	0.3333	1
Indonesia	0	0.1667	0.3333	1
Kiribati	0.3333	0.3889	0.6667	0
Laos	0	0.0556	0.3333	0.3333
Malaysia	0	0.0556	0.3333	1
Maldives	0.3333	0.3889	0.6667	0.6667
Micronesia	0	0.1667	0.3333	0
Mongolia	0	0.1667	0.3333	0
Myanmar	0	0.2778	0.3333	1
Nepal	0	0.0556	0.3333	0.6667
Papua New Guinea	0	0.2778	0.3333	1
Philippines	0.3333	0.3333	0.3333	1
Samoa	0	0.2222	0.3333	0.3333
Singapore	0.3333	0.3333	0.3333	0
Solomon Islands	0	0.3333	0.6667	0
Sri Lanka	0	0.1667	0.6667	0.6667
Thailand	0	0.1111	0.3333	1
Timor-Leste	0.3333	0.3889	0.6667	0.3333
Tonga	0	0	0	0
Vanuatu	0	0.1111	0.6667	0
Vietnam	0	0.0556	0.3333	1

* Most recent future three years conflict likelihood
† Most recent past three years conflict likelihood

Bibliography

[1] Heidelberg Institute for International Research. "Conflict Barometer 2017". Technical report, Heidelberg, Germany, 2017.

[2] Nicholas J. Shallcross. "A Logistic Regression and Markov Chain 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] Sarah Neumann. "Forecasting Country Conflict Within Modified Combatant Command Regions Using Statistical Learning Methods". Master's thesis, Wright-Patterson AFB, Ohio: Air Force Institute of Technology, 2018.

[4] Michael D. Ward, Nils W. Metternich, Cassy L. Dorff, Max Gallop, Florian M. Hollenbach, Anna Schultz, and Simon Weschle. "Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction". *International Studies Review*, 15(4):473–490, 2013. doi: 10.1111/misr.12072.

[5] Jack A. Goldstone, Robert H. Bates, David L. Epstein, Ted Robert Gurr, Michael B. Lustik, Monty G. Marshall, Jay Ulfelder, and Mark Woodward. "A Global Model for Forecasting Political Instability". *American Journal of Political Science*, 54(1):190–208, 2010. doi: 10.1111/j.1540-5907.2009.00426.x.

[6] Robert Shearer. "Recognizing Patterns of Nation-State Instability that Lead to Conflict". *Military Operations Research*, 15(3):17–30, 2010.

[7] Havard Hegre, Joakim Karlsen, Havard M. Nygard, Havard Strand, and Henrik Urdal. "Predicting Armed Conflict, 2010-2050". *International Studies Quarterly*, 57:250–270, 2013. doi: 10.ini/isqu.l2007.

[8] Benjamin 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.

[9] Benjamin 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.

[10] J. Farhan. "Overview of Missing Physical Commodity Trade data and Its Imputation Using Data Augmentation". *Transportation Research Part C: Emerging Technologies*, 54:1–14, 2015. doi: 10.1016/j.trc.2015.02.021.

[11] Ronald Wesonga. "On Multivariate Imputation and Forecasting of Decadal Wind Speed Missing Data". *SpringerPlus*, 4(1), 2015. doi: 10.1186/s40064-014-0774-9.

[12] Emily Casleton, Dave Osthus, and Kendra Van Buren. "Imputation for Multisource Data with Comparison and Assessment Techniques". *Applied Stochastic Models in Business and Industry*, 34(1):44–60, 2018. doi: 10.1002/asmb.2299.

[13] James Haworth and Tao Cheng. "Non-Parametric Regression for Space-Time Forecasting Under Missing Data". *Computers, Environment and Urban Systems*, 36(6):538–550, 2012. doi: 10.1016/j.compenvurbsys.2012.08.005.

[14] Maria Vlachopoulou, Tom Ferryman, Ning Zhou, and Jianzhong Tong. "A Stepwise Regression Method for Forecasting Net Interchange Schedule". *2013 IEEE Power and Energy Society General Meeting*, pages 1–5, 2013.

[15] Claudia Pedroza. "A Bayesian Forecasting Model: Predicting U.S. Male Mortality". *Biostatistics*, 2006. doi: 10.1093/biostatistics/kxj024.

[16] Peter G. Zhang. "Time Series Forecasting using a Hybrid ARIMA and Neural Network Model". *Neurocomputing*, 50:159 – 175, 2003. ISSN 0925-2312. doi: https://doi.org/10.1016/S0925-2312(01)00702-0.

[17] Stef van Buuren. *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis Group, 2012. ISBN 9781439868249.

[18] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. URL https://www.jstatsoft.org/v45/i03/.

[19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/.

[20] Luke Brantley. "Looking Past the Spark to Find the Fuel of the Arab Spring Fire". Master's thesis, Wright-Patterson AFB, Ohio: Air Force Institute of Technology, 2018.

[21] Kobi Abayomi, A Gelman, and Marc Levy. "Diagnostics for Multivariate Imputations". *Journal of the Royal Statistical Society*, 57(3):273–291, 2008.

[22] Sonja Engmann and Denis Cousineau. "Comparing Distributions: The Two-Sample Anderson-Darling Test as an Alternative to the Kolmogorov-Smirnoff Test". *Journalof Applied Quantitative Methods*, 6(3):1–17, 2011.

[23] Bruce L. Bowerman, Richard T. O'Connell, and Anne B. Koehler. *Forecasting, Time Series, and Regression: An Applied Approach*. Thomson Brooks/Cole., 4th edition, 2005.

[24] Daniel T. Larose and Chantal D. Larose. *Data Mining and Predictive Analytics*. Wiley Publishing, 2nd edition, 2015. ISBN 1118116194, 9781118116197.

[25] Douglas C. Montgomery, Elizabeth A. Peck, and Geoffrey G. Vining. *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, 5th edition, 2012.

[26] Stepwise Regression Control Panel. https://www.jmp.com/support/help/14-2/stepwise-regression-control-panel.shtml#240750.

[27] Nornadiah Mohd Razali and Bee Yap. "Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests". *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.

[28] Aravind Hebbali. *olsrr: Tools for Building OLS Regression Models*, 2018. URL https://CRAN.R-project.org/package=olsrr. R package version 0.5.2.

[29] Richard Williams. "Heteroskedasticity". https://www3.nd.edu/~rwilliam/stats2/l25.pdf.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
21-03-2019	Master's Thesis	Jun 2018 – Mar 2019

4. TITLE AND SUBTITLE
An Imputation Approach to Developing Alternative Futures of Country Conflict

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Kane, Zachary, J, 2d Lt, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
AFIT-ENS-MS-19-M-128

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/Department of Operational Sciences)
2950 Hobson Way, Building 640
WPAFB OH 45433-7201

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT-ENS-MS-19-M-128

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Joint Chiefs of Staff J-7: Lessons Learned
11. SPONSOR/MONITOR'S REPORT
116 Lake View Parkway Suffolk, VA 23435-5226
DSN:668-7618 js.dsc.j7.room.dd-fjfd-jlld@mail.mil

10. SPONSOR/MONITOR ACRONYM(S)
JCS J-7

11. SPONSOR/MONITOR 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
Understanding what causes countries to be in a state of violent conflict is of vital importance to developing realistic national strategies on both a regional and global scale. Given these causes, it is important to understand the effects of missing data, how to impute that data, and the interrelation between data elements. Utilizing both open source data and previously generated equations that predict a country's likelihood to transition conflict statuses, this research projects data into the future and predicts each nations' subsequent conflict statuses. Future data is populated using a novel approach inspired by stochastic regression imputation. The replicated future data and predictions were interpreted as alternative futures of regional conflict in both the Arab world and Southeast Asia. The conflict occurrences in the Arab world region were projected to trend upward compared to the region's historic behavior. In Southeast Asia, the next ten years forecasted a decline in total violent conflicts. Regional scenarios where the elements of national power influenced a data element were implemented to learn how alternative futures might be

15. SUBJECT TERMS
Data Imputation, Country Conflict Forecasting, Stepwise Regression

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	90	19b. PHONE NUMBER (Include area code) (937) 255-3636, x4708; darryl.ahner@afit.edu