9-2022

Improving Country Conflict and Peace Modeling: Datasets, Imputations, and Hierarchical Clustering

Benjamin D. Leiby

IMPROVING COUNTRY CONFLICT AND PEACE MODELING: DATASETS, IMPUTATIONS, AND HIERARCHICAL CLUSTERING

DISSERTATION

Benjamin D. Leiby, Major, USAF

AFIT-ENS-DS-22-S-065

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-DS-22-S-065

Improving Country Conflict and Peace Modeling: Datasets, Imputations, and Hierarchical Clustering

DISSERTATION

Presented to the Faculty

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 Doctor of Philosophy in Operations Research

Benjamin D. Leiby, M.S., M.B.A

Major, USAF

September 15, 2022

AFIT-ENS-DS-22-S-065

Improving Country Conflict and Peace Modeling: Datasets, Imputations, and Hierarchical Clustering

DISSERTATION

Benjamin D. Leiby, M.S., M.B.A
Major, USAF

Committee Membership:

Darryl K. Ahner, Ph.D
Chair

Raymond R. Hill, Ph.D
Member

Phillip R. Jenkins, Ph.D
Member

Edward D. White, Ph.D
Member

AFIT-ENS-DS-22-S-065

Abstract

Many disparate datasets exist that provide country attributes covering political, economic, and social aspects. Unfortunately, this data often does not include all countries nor is the data complete for those countries included, as measured by the dataset's missingness. This research addresses these dataset shortfalls in predicting country instability by considering country attributes in all aspects as well as in greater thresholds of missingness. First, a structured summary of past research is presented framed by a developed casual taxonomy and functional ontology. Additionally, a novel imputation technique for very large datasets is presented to account for moderate missingness in the expanded dataset. This method is further extended to establish the MASS-impute algorithm, a multicollinearity applied stepwise stochastic imputation method that overcomes numerical problems present in preferred commercial packages. Finally, the imputed datasets with 932 variables are used to develop a hierarchical clustering approach that accounts for geographic and cultural influences that are desired in the practical use of modeling country conflict. These additional insights and tools provide a basis for improving future country conflict and peace research.

Table of Contents

Page

List of Figures

List of Tables

Improving Country Conflict and Peace Modeling: Datasets, Imputations, and Hierarchical Clustering

I. Introduction

1.1 Research Problem

Government officials struggle to objectively defend rationale for assessing country conflicts and allocating defense resources to manage conflict risks worldwide. The Armed Forces execute command and control through two chains of command: Combatant Commands (COCOMs) and Military Departments (MilDeps). The COCOMs are responsible for missions and forces assigned to their jurisdiction while MilDeps are responsible for purposes other than operational direction of forces, such as recruitment and readiness. The relationship between the chains can be illustrated as supply and demand. The COCOMs provide a demand signal for the resources needed to provide security to the nation while MilDeps provide the supply to meet those needs. However, constraints often create a gap between supply and demand. In 2013, Rear Admiral Thomas Moore lamented that "we're an 11-carrier Navy in a 15-carrier world" which was a prophecy of the Navy only meeting about 44% of the COCOM requests in 2015 [1]. The disconnects were also experienced in the Air Force with over 25,000 non-supported tanker flying hours in 2019 and the Army struggling to maintain a deploy-to-dwell policy ratio of 1:3 with their average ratio being 1:1.2 in 2018 [1]. The blame could easily be directed at how budgeting is conducted, but it could just as easily be considered a lack of strategic prioritization and alignment. There is little incentive for COCOMs to stop asking for more and more resources.

Some have proposed a reform of federal interagency processes to align regional commands between the now disconnected COCOMS, State Departments, and intelligence communities [2]. They argue that this reform would impact budgets, authorities, and organizational identities that would ultimately provide the United States with the necessary integration needed to maintain the security of the nation [2]. Although this reform may assist in streamlining bureaucracy, it falls short of addressing the disconnect between supply and demand. COCOMs need a capability to convincingly prompt Congress to address constraints keeping MilDeps from being unable to meet the supply and demand disconnect. This research addresses that capability. This research moves forward the accuracy of predicting country conflict to address assessing optimal regional alignments and advocacy for addressing allocation of defense resources. As accuracy for conflict prediction increase, COCOMs will better assess and defend their position for the limited resources they are requesting from MilDeps.

1.2 Research Objectives

The primary goal of this research is to increase confidence in and usability of data and parameters applied to country conflict modeling. Country conflict modeling is predicting when a country will go into instability resulting in war, either internationally or intranationally. As big data expands into many sectors, it is natural that country conflict research investigate advantages that come with increased information. Often, country conflict modeling takes a purely economic or political aspect approach, potentially excluding other important contributors in predictive accuracy. This research argues a whole of concept approach, considering all aspects of country conflict modeling, for collecting data and assists researchers in developing that data toward predictive modeling. As data collection expands, problems surface to include data missingness, multicollinearity effects, and the curse of dimensionality. The ul-

timate end point addresses the research problem of assisting COCOMs with models that advocate for the allocation of defense resources: accurately predict instability in the allocated region. These issues have been addressed on a smaller scale in prior country conflict research, but the processes do not always scale with increased data. This research provides alternative methods to scale solutions toward incorporating larger datasets for the country conflict modeling problem.

This research answers three questions that will lead to providing defendable rationale in assessing country conflict and allocating defense resources.

1. What data sources are available and what data elements provide statistical insight to country conflict modeling?

2. How can incomplete country data be addressed through imputation methods?

3. Are there defensible, analytical arguments for partitioning the world into management sectors?

The first two questions drive to overcome difficulties in analytical modeling. The third question pierce to the heart of aligning federal agencies for positive integration and delivering actionable information for policy making. From a research perspective, questions 1 expands the body of knowledge through application, while question 2 expands the body of knowledge through theoretical insight. Question 3 is a combination of both application and theory.

1.3 Document Overview

This study approaches conflict modeling accuracy from a sequential approach: data collection, repair, and implementation. This document is organized in a four paper format. Chapter II provides an overview of relevant databases for collecting modeling variables along with background information demonstrating the progress

made in the field. The survey provides a foundation for breaking outside a single aspect approach to modeling and embracing a whole of concept approach, simplifying the taxonomy of the field and proposing a functional ontology for predicting global country conflict. As more information is collected, the likelihood of missing values increases presenting a need to repair the dataset. Chapter III illustrates how prior research approaches do not scale as country conflict data increases presenting a need for a new imputation approach. The research investigates a correlation methodology that overcomes numerical problems inherent with scaling data with preferred commercial packages. This new methodology also addresses current issues of applying a tolerance parameter to stop the imputation algorithm, defending when the estimate is plausible. Chapter IV expands on the imputation methodology, culminating in a defendable multiple imputation algorithm for country conflict data named Multicollinearity Applied Stepwise Stochastic imputation (MASS-impute). The methodology alleviates concerns about multicollinearity between a large set of independent variables and provides a variable range-based guard rail systems to combat extreme outliers in imputed estimates. With the dataset complete, the research continues with approaching the COCOM problem through investigating parameters for clustering methods. Chapter V provides two lanes of insight needed to explore dividing countries into regions for country conflict modeling. First, it addresses the curse of dimensionality that scaling the dataset presented. Various dimension reduction approaches are highlighted with parameters addressed for the preferred method to retain the most amount of information that will feed clustering methods. Second, an argument for hierarchical clustering to defend grouping countries into regions is presented with prediction accuracies as the focusing metric. Due to the complexity of the research problem, the study stops short presenting a case for variable selection and model coefficients, instead focusing on the clustering approach and associated

parameters. The results highlight a recommendation that there may be a need for more COCOMs rather than a dissolution of them.

II. Datasets and Models for Globally Predicting Country Conflict and Peace, A Survey

2.1 Abstract

Many studies focus on the statistical relevancy of individual variables from local datasets rather than the overall goodness of the model for predicting global conflict. Others take a purely economic or political approach to modeling country conflict. Generally, we see logistic regression techniques relying on p-value to indicate important variables, while other contemporary approaches lean on non-parametric modeling to assess predictive goodness. Since identifying which data elements are relevant to modeling country conflict is an extremely relevant topic, this article develops an 'aspect' construct with underlying 'themes' for considering data and maps those themes to historical studies; presents a comprehensive list of potential datasets for both dependent and independent conflict prediction variables; and surveys modeling and analytical techniques for predicting country conflict. Several studies are explored that predict conflict. Within these studies, variables are identified while also indicating where they can be found for replicability, and a generous number of techniques are examined to illuminate research areas. This paper provides a taxonomy of variables for country conflict prediction along with a survey of data and empirical modeling techniques for understanding conflict through modeling at a global level.

2.2 Introduction

"Peace cannot be kept by force; it can only be achieved by understanding" is a philosophy attributed to theoretical physicist and peace activist Albert Einstein [3]. Countries require policies and decisions to move forward toward safeguarding territory and economic security for most citizens who never concern themselves with such

details. If peace and conflict can be modeled accurately, non-governmental organizations could pre-position resources to lessen suffering, governments could efficiently deploy resources to maximize stability, and intergovernmental organizations could more effectively develop treaties to find compromise. This is why "many large international organizations and governments rely on regional or global forecasts of conflict in order to address humanitarian, military and political crises" [4]. However, conflict models are often contradictory and use significantly diverse data sets and variables. To wit, researchers have yet to gain consensus on how to model conflict or which metrics best portray the risk of national conflict. Therefore, it is necessary to examine the successes and shortfalls that pave the way toward better understanding by investigating conflict models, variables used within those models, and relevant datasets to better understand the state of the art (or science) in country conflict modeling and availability of quality datasets.

Ward posited "that if you can develop models that provide an understanding, you should be able to generate predictions that will not only be accurate but may also be useful in a larger societal context" [5]. This insight premises that "we need more predictions" to generate good theories, curtail bad theories, illuminate new research areas, and discourage undiscerning methods [5]. Building forecasts requires addressing some basic research questions. What are the available datasets for model consumption and where can they be found? What techniques are appropriate for empirically modeling armed conflict at a global level? Through these better predictions, Hegre asserts that we will "not only fulfill scientific objectives; it also enables policymakers to formulate evidence-based policies regarding peace and security issues" [4]. However, this is just one step toward full understanding. More questions will arise such as the quality of the data to accommodate modeling assumptions and the inevitable task of filling gaps in data records. Each step, however, brings us closer to understanding

the path towards peace.

In order to advance globally predicting country conflict and peace, we propose a functional ontology as a generic model given in Figure 1. This functional ontology provides a means to relate the properties of conflict and peace within countries. This relation is governed by the measure by which countries are determined to be in conflict along with the variables that are correlated with being in a given state plus the stochastic error associated with predictions. We will explore well-established measures of country conflict in the literature and the literature which conjectures causal dependent variables. A taxonomy of these causal variables is developed and functions are informed from a survey of data and empirical modeling techniques used for country conflict modeling.

There are many datasets available for modeling conflict, however, many focus on a definition of conflict outside the scope of this survey. They range from deep diving into specific regions within countries to overview demographics of the world. This survey seeks modeling a definition of conflict at a global scale, to include intranational, international, and transnational conflict. Furthermore, conflict includes at least one national government entity and involves fatalities. Many trends in research seek to investigate solely intrastate conflict (civil wars), regional conflict (not global modeling), or research that is an aggregation of national data, leaving global modeling as a small niche. Disaggregation is currently considered outside the scope of this survey because disaggregated databases have not been able to encompass all nations for global modeling. Although a fantastic approach to modeling conflict, this focus

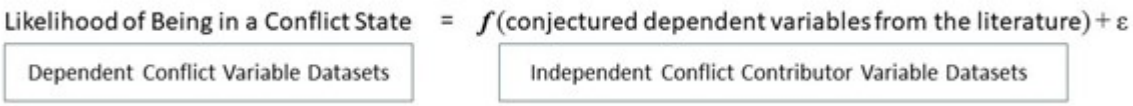

Figure 1: Conceptual functional ontology for predicting global country conflict

area requires more attention in data collection if the goal is global conflict understanding, and this approach is currently not ready for global model building. This leaves few options for identifying conflict, all of which have their own nuances for conflict identification: the Correlates of War Project *War Data*, the Uppsala Conflict Data Program *Armed Conflict Dataset*, and the Heidelberg Institute for International Conflict Research *Conflict Information and Analysis System*. Other known datasets such as the *Armed Conflict Location and Event Dataset* (ACLED), the *Georeferenced Event Dataset* (GED), and others are outside the scope of this survey due to not meeting the aforementioned definition of conflict for modeling purposes. More information on the tailoring of these other datasets has already been consolidated by Wencker [6].

As for model building, this survey references some works that possess a narrower definition of conflict (i.e. focus on civil war) since the technique may adapt itself to modeling globally. Similar works were not included if the goodness of the technique was already sufficiently covered. The primary modeling focus is identification of techniques to explain conflict at a global level with techniques critiqued in order of publication date with emphasis on assessing the robustness of significance, most notably by considering accuracy in predictions.

2.3 Theory Of Conflict

Conjecturing why conflict occurs takes many forms in the literature. Some researchers focus on innate diversities within humanity while others focus on resources that provide an element of power. This often lends itself to researchers building models to support their hypothesis that conflict is due to political instability [7, 8, 9] or economic issues [10]. Still others contend social constructs are at the heart of conflict erupting from inequalities [11, 12]. Regardless of the focus, these measures are typi-

cally a proxy from some other unmeasurable underlying variable, a latent factor. For example, some models hold population as significant, but there is typically no universal explanation as to why population changes lead to conflict. Is it from overcrowding sparking tension between people? Population may be highly correlated to the true variable, but percent of overcrowding (if the true variable) is not a measure that is typically recorded by countries. A latent factor may suffice for exploratory analysis but may limit conflict predictions accuracies if the true variable is not identified and measured. However, there are some variables that consistently surface as core proxy variables in model building such as Polity and gross domestic product (GDP). To understand the breadth of proxies, a unifying taxonomy is needed. Goldstone classified over 38 variables into three categories: political, economic, and social [9]. Building then from Goldstone, the general categorization of political, economic, and social categories creates a conflict data taxonomy which we will refer to as aspects, or the specific direction of conflict, which is often how researchers narrow their study of conflict. Later we will develop themes within these aspects for greater resolution within our taxonomy.

2.3.1 Dependent Variable – Conflict

Understanding country conflict is the current focus with desires to ultimately predict it accurately. Conflict in the context of this research is defined as violence with significant taking of human life, typically war. This contrasts with claiming conflict status between or within nations due to economic or political actions that do not directly sanction life taking. Therefore, trade embargoes or the dissolution of treaties are not considered conflict, although they may communicate non-violent conflict or signal the precursor of violent conflict. This context still allows for diverse definitions which will be explored.

The concept of conceptualizing conflict has evolved over the years from observing official war declarations, to assessing the number of casualties, to developing a mode and quality of the course of conflict. Very little modern research relies on observing official war declarations. Therefore, no datasets or models are described using this outdated methodology of declaration of war which is fraught with many inconsistencies from governments not wishing to publicize their intentions including countries conducting covert operations and non-state actors looking to topple existing political structures in civil wars.

More modern datasets classify conflicts based on overall conflict event casualties such as the Correlates of War Project *War Data* (COW) and the Uppsala Conflict Data Program *Armed Conflict Dataset* at the department of Peace and Conflict Research Institute Oslo (UCDP/PRIO). Researchers such as Fearon and Laitin [7] as well as Goldstone [9] classify civil war conflicts using COW datasets or similar metrics, which view conflicts as violent when at least 1,000 casualties occur within the event, and average at least 100 casualties per event year. Other researchers such as Celiku and Kraay [13] and others [12, 14, 15] rely on a less strict definition of conflict for civil war research using the UCDP/PRIO datasets, which set the threshold of casualties to only 25 battle-related deaths per year. Hegre takes a multinomial approach with UCDP/PRIO data setting between 25 and 999 battle-related deaths as minor conflicts while greater than 1,000 deaths is considered major conflict [16]. Meanwhile, Wallensteen and Sollenberg in their Armed Conflict report follow UCDP/PRIO data and track the trends of minor armed conflict (battle-related deaths below 1,000 during the conflict period), intermediate armed conflict (greater than 1,000 battle-related deaths during the conflict period but less than 1,000 in any given year), and war (greater than 1,000 battle-related deaths in any given year) [17].

The Heidelberg Institute for International Conflict Research (HIIK) *Conflict In-*

formation and Analysis System focuses on conflict processes rather than purely quantitative thresholds of casualties and maps conflict into five levels with the first two indicated as non-violent conflict and the top three as violent conflict. According to HIIK, incompatibility of intentions between actors emerges in the form of observable and interrelated actions and measures, and threatens state functions or the international order. Research by Boekestein [18] and others [19, 20, 21, 22] capitalize on HIIK's delineation of violent and non-violent conflict to develop models to predict violent conflict. HIIK uses a five-attribute algorithm to assign a conflict intensity level to regions within a country on a monthly basis which translates to a yearly score from the highest regional level within the country [23]. This allows for a broader view of the consequences of conflict as it incorporates not only casualties, but also the dispersion of population due to conflict, as well as the destruction of infrastructure.

Although these methodologies, summarized in Table 1 for studies listed in Appendix A, establish a consistent and observable metric for defining conflict, they do present limitations within the scope of developing models for predicting when, where and who may be involved in conflict. Conflict itself is amoral and relies on the actors to distinguish it from an evil or virtuous deed. Often, conflict is presumed an action fueled by a desire for power and control or born out of bigotry and hatred. However, intervening actors may insert themselves into conflict for the safety of a population or the preservation of an ideal leading to faults in the presumed assumptions. The latter is more consistent with developed countries as opposed to developing countries as they are more likely to have the resources to assist other nations in need.

Table 1: Study identification of conflict proxies

Data Source	Proxy	Study
COW	Casualties	1, 2, 3, 7
UCDP/PRIO	Casualties	4, 5, 6, 8, 9, 11, 13
HIIK	Conflict Mapping	10, 12, 14, 15, 16

Another attribute to be aware of when using conflict datasets is that primary classification may not always be attributed to a country where casualties are occurring. A cursory examination of the UCDP/PRIO website shows that the United States has not had any state-based violence since the 9/11 event in 2001, however, their battle-related deaths dataset shows the United States in conflict type 4 from 2002 to 2017 due to deaths tolls in Afghanistan. This phenomenon occurs from the structure of the dataset labeling conflict by primary states with another column labeled secondary states. Pettersson and Öberg highlight this in describing how intrastate conflicts become internationalized with the US being "involved in the largest number of conflicts as a secondary warring party" [24]. Similar findings were observed in HIIK datasets where no mention of conflict deaths were recorded in the United States but the US was classified as in violent conflict due to 50 US-blamed Haqqani Network deaths in Pakistan and later three Pakistani Frontier Corps soldier deaths in a US-led NATO strike near the border of Pakistan [25]. Interestingly enough, the US was not considered in violent conflict through the rationale of 710 combat-related American deaths within Afghanistan; most likely because the United States was welcomed in the country by the legitimate government to assist in police efforts to combat terrorism. The United States at that time was in a declaration of war against terrorism, however, was not declared at war with any one legitimate state government.

It is also worth noting that as database maintainers classify conflict using qualitative assessments, there is room for interpretation in labeling events. In the 2019 HIIK dataset, the United States is labeled in violent conflict due to right-wing extremists within the country [23]. Three unconnected shootings throughout the year totaled 29 dead and 32 injured [23]. The assessment classified and labeled these shootings as right-wing extremists against the government although none of the targets were against government representatives or buildings, and each of the alleged motives were

different in ethnic or religious targets. Besides involving government investigation, these incidents were mainly handled by local officials and no government counter-response occurred. The qualitative assessment could have easily been classified as domestic events and *not in conflict* rather than right-wing extremists against the government soliciting the indicator of a country *in conflict*.

Ultimately, it appears that the dependent variable of conflict may be more reliable when modeling developing countries rather than developed countries because of the nuances of primary/secondary actors and possible qualitative interpretations of certain events. Latitude is required when considering interstate conflict and the qualitative assessment of triggering events if using HIIK datasets. This latitude may influence the quality of the prediction that may be obtained through modeling. Although the counting of casualties may also incur bias depending on the openness of the country to report and the bias by the media. Eriksson and Wallensteen observed through UCDP/PRIO data an overall decline in conflict despite gloom in the media reflecting security fears [26]. Hegre supports these findings through his simulated modeling using UCDP/PRIO data concluding decreases in overall conflicts out to year 2050 [16]. This is contrary to findings using Markov modeling of HIIK data by Shallcross which shows a net increase in conflicts from 2014 to 2024 [27]. Overall, the three datasets (COW, UCDP/PRIO, and HIIK) appear to be the most comprehensive catalogue of conflict and have been the de facto standard for classifying violent and non-violent conflict.

2.3.2 Independent Variables

Before predicting conflict with any amount of accuracy, it is important to understand which variables contribute toward leading a country to conflict. Homer-Dixon identified three main perspectives on conflict type: simple scarcity, group identity,

and relative justice [28]. He proposed that environmental effects that directly or indirectly lead toward economic or political power are more prone to develop conflict at international levels while distributed justice issues are more aligned with intranational conflict [28]. Additionally, biases and discrimination between group identities foster conflict both in the international and intranational level [28]. Instead of looking at the differences between international and intranational conflict, important variables leading to conflict are better distinguished between the proposed political, economic, and social aspects. However, Homer-Dixon's perspectives provide a list of themes within these aspects, such as simple scarcity being a natural resource theme within the economic aspect and group identity being a discrimination theme within the social aspect.

Using a simple methodology of tallying significant variables from a multitude of models, concrete themes coalesce under the three aspects. As seen in Figure 2, themes naturally fall within their associated aspects. The political aspect forms two themes: Polity and liberties. Polity is a top-down view describing the type of leadership a country has while liberties is a bottom-up view describing how to people feel about their government. The economic aspect also shares two themes between economic health and natural resources. Economic health views the economic statistics of a country such as gross domestic product (GDP) per capita, trade as a percentage of GDP, and unemployment. Significant variables under the natural resource theme surface under proxies such as arable land, access to fresh water, and the export of natural resources. The social aspect rounds out the final five themes such as discrimination, conflict supports, population statistics, quality of life factors, and regionality.

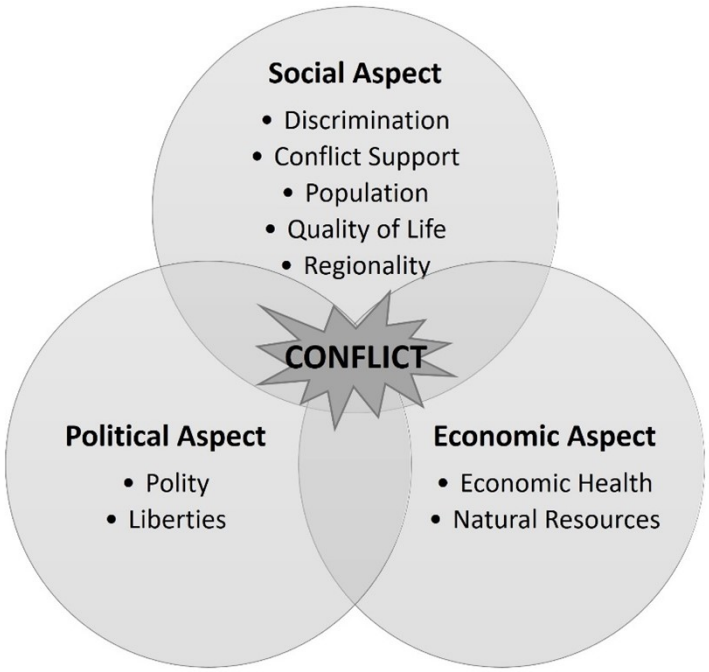

Figure 2: Main themes within each conflict aspect

2.3.2.1 Political Aspect

Aligning with HIIK's assessment, conflict is a political issue as it involves communication or action between or against state governments, which are inherently political. The main classification of governments is delineated along a spectrum of how decisions are made. At one end, autocracies consolidate state power and unlimited authority into a single person whereas at the opposite end, democracies distribute power among representatives elected by the people. Observations by Gates states democracies survive 3.6 times longer than their inconsistent counterparts and autocracies 1.9 times longer than their inconsistent counterparts [8]. The Center for Systemic Peace translates this assessment of autocracies and democracies into a variable called Polity where the spectrum diverges on a 21-point scale (-10 to 10) along with four additional values for disruptive events. This one variable appears to be the gold standard among researchers for assessing the political attribute of conflict as it

can be found in almost every research article on conflict in one fashion or another.

Polity is typically transformed into a dummy variable for use in modeling conflict. Gartzke used Polity III to distinguish between high and low democracies [10]. Fearon and Laitin capitalized on the new Polity IV to break out dummy variables such as weak and strong anocratic regimes to show a 68% increased odds of civil war outbreak for weak anocracies [7]. Goldstone later used a dummy variable for regime type and commented that the "categorical measure of political institutions was by far the most powerful factor for distinguishing stable country-years from those that soon experienced instability onsets" [9]. He continued citing the importance of modeling this variable with "once regime characteristics are taken into account, most other economic, political, social, or cultural features of the countries in our sample had no significant impact" [9]. Gates concurred that observing Polity provides insight into the stability of government and is "equally, if not more, important in terms of explaining political stability than many of the literature's standard set of explanatory variables (level of economic development, economic growth, political neighborhood)" [8]. Therefore, many conflict models following Goldstone's study include at least some transformed variable of Polity. The typical transformation of Polity is usually labeled regime type, however, Shallcross (and later Leiby) labeled the transformation government type as the database already had a variable labeled regime type based on a transformation from the CIA World Factbook [22].

The lines between Polity, regime type, and government type often get blurred when trying to categorize how researchers model the theme of Polity. The 21-point scale often is not as useful as pulling out one area of the scale for a dummy variable named Democracy or grouping sections of the scale to determine regime types. Gates went a step further to also include a variable called *Political Neighborhood*, which incorporates an average "political distance" score based on government type concerning bordering

countries [8]. As can be seen, Polity is one of the more common themes modeled in conflict even when the research hypothesis is concerned with a non-political aspect. Table 2 outlines the studies that included a Polity themed variable in their models with the bolded studies identifying if they considered the associated variable as a core variable (control variable if core was not used in syntax). The studies, which are listed in the appendix, are numbered by year published with study 1 being published in 2001 and study 16 being published in 2018.

Table 2: Study identification of significant political proxies

Themes	Proxies	Study
Polity	Polity	5, 6, 10, 14, 15
	Regime Type	**1**, 2, 4, 6, 7, **9**, 10, 12, 14, 16
	Government Type	12, 14, 16
	Political Neighborhood	**4**
Liberties	Civil Liberties	13
	Political Liberties	13
	Freedom	10, 12, 14, 16
	Voice & Accountability	15

Liberties is the other main theme within the political aspect and as Celiku and Kraay describe, measure the political accountability of government [13]. The common indicator is termed *Freedom Score*, which is a composite score between a civil liberty measure and a political rights measure from Freedom House. A similar measure from the World Bank is Voice and Accountability, which "attempts to capture the population's perception of their ability to affect their government or freedom" [19].

2.3.2.2 Economic Aspect

Many conflicts originate due to economic disparity. Table 3 lists the economic proxies under their respective theme and maps them to the relevant literature. Economic disparity does not always have to expose itself as monetary wealth but could also be access to natural resources such as land, water, or even oil. For example, the

damming of the Nile River would not only provide stored resources for the government that controls the dam but could diminish the livelihood of other governments that rely on the river to produce crops which directly impact a nation's GDP and trade. One of Homer-Dixon's main premises for conflict concerning international conflict revolved around the scarcity of resources, especially the desirability of water [28]. Similarly, Boekestein found significance in improved water when developing models for predicting conflict [18] Shallcross [22] and Leiby [20] also found that fresh water per capita were influential variables in some regions when developing prediction models with over 90% accuracy for the subsequent near-term prediction outlook. Like water in arid regions, Hegre also stressed the importance of oil in nation states whose primary commodity for trade is oil [16]. The list continues with Collier and Hoeffler highlighting "diamonds in West Africa, timber in Cambodia, and cocaine in Colombia" [29]. They suggest that natural resources used as commodities may provide an opportunity for extortion and thus make rebellion either feasible or at least attractive [29].

Table 3: Study identification of significant economic proxies

Themes	Proxies	Study
Economic Health	GDP per Capita	2, 3, **4**, **5**, **6**, **9**, 10, 11, 12, 13, 14, 16
	GDP Growth	**1**, 3, **4**, 5, 11, 12, 13, 14
	Trade	**1**, 3, 10, 11, 12, 13, 14, 16
	Unemployment	10, 14, 16
	Military Expenditure	**1**, 12, 14, 16
	Consumer Price Index	15
	Currency Pegging	1
	Financial Assistance	15
	Tourism	15
Natural Resources	Arable Land	10, 12, 14, 16
	Fresh Water per Capita	12, 14, 16
	Improved Water Source	10, 12, 14
	Natural Resource Export	**1**, 2, 3, 8
	Natural Resource Rent	13

Although these natural resources show significance in some prediction models, they are often correlated with a nation's GDP. As is the case, GDP per capita is often highlighted in conflict prediction models even when natural resources are not considered. Garzke [10] and Østby [12] included GDP as a core control variable in his models noting that it influenced how other key variables acted within his model. For example, Østby demonstrated that economic inequalities affect Polity where democracies were at a higher risk of conflict than autocracies [12]. However, Fearon and Laitin observed other influences and hypothesized that as GDP increased, risk decreased that a nation would experience conflict, at least concerning civil conflict [7]. Collier and Hoeffler agreed suggesting that increase per capita income reduced conflict risk because the good wealth provided an opportunity cost against rebellion [29]. Garzke illustrated additional rationale using decision analysis and conflict uncertainty to conclude that as a nation prospers, they become more risk adverse in choosing violent conflict over remediation through other non-violent means [10]. However, Fearon and Laitin [7] also concluded that decreased GDP may translate to weaker infrastructure, allowing rebels to gain a stronger foothold due to the government not being able to transit the population efficiently whereas Collier and Hoeffler [29] did not find any strong effects between rebellion and low income. As mentioned with ties to natural resources, trade is another variable that indicates economic prosperity that may diminish as nations build more wealth. However, trade also signals indications of conflict if issues are not resolved through other non-violent means [10].

Another economic variable that has had little consideration is a binary indicator called pegging: a term to describe that a nation's currency is being anchored to another nation's economy. Garzke explored this idea by supposing that a nation's economic health is more than just GDP and trade, and that increased global interconnectedness could decrease a nation's desire to incite violent conflict [10]. The idea

may hold promise, but the availability of data or proxy variables may be difficult to obtain for developing whole world models.

2.3.2.3 Social Aspect

Societal or cultural influences are often the main topic of concern when developing models, especially in cases of great disparity between groups. These societal and cultural proxies are identified in Table 4 and are mapped to modeling sources of use. As illustrated earlier, right-wing extremist, some of which were motivated by racial bias, were at the heart of conflict in the US, a developed country with large GDP and trade [23]. Large ethnic and religious fractionalization were at the heart of Fearon and Laitin's study where they assessed the risk of conflict increasing for populations with larger majorities, especially as GDP increased, although higher GDP reduced conflict risk overall [7]. Østby, thinking along the lines of tension between groups also included political exclusion based on data from the Minorities at Risk Project (MARP) and electoral systems inclusiveness with which she referenced a related Golder study [12]. Goldstone also included a similar variable called State-led discrimination, derived from the MARP, which he found significant in his global forecasting model [9]. Boekestein capitalized on Goldstone's findings of State-led discrimination and further used a derivative of the civil liberties and political rights score called the Freedom score. Although the Freedom score appears similar to State-led discrimination, the Freedom score is better classified under the political aspect rather than the social aspect because social discrimination is focused more on inequalities between people groups rather than citizens advocating against the government. Overall, whenever there exist haves and have-nots, perceived or justified, there is the possibility for conflict.

Perceived injustice can fuel the flames of active injustice. Many times, minori-

Table 4: Study identification of significant social proxies

Themes	Proxies	Study
Discrimination	Ethnic Diversity	8, 9, 10, 12, 13, 14, 16
	Religious Diversity	10, 12, 13, 14, 16
	Political Discrimination	6, 7
	Income Dispersion	9, 13
	Human Rights Violation	13
Conflict Supports	Border Conflict	2, 7, 10, 13, 14, 15, 16
	Conflict Intensity	15
	Conflict Lag/History	3, **5**, **6**, **8**, **9**, 12, 13, 16
	Military Alliances	**1**, 5
	Refugees	10, 12, 13, 14, 15, 16
Population	Population Density	10, 12, 16
	Population Dispersion	**1**, 3, 10
	Population Growth	10, 14, 16
	Population Size	2, 3, **5**, **6**, 8, **9**, 11, 13
	Armed Forces Personnel	15
	Deployed Troops	15
	Labor Force Participation	15
	Youth Bulge	8, 12, 15, 16
Quality of Life	Birth Rate	10, 12, 14
	Caloric Intake	10, 12, 14, 16
	CPIA[1]	13
	Death Rate	10, 12, 14, 16
	Education Level	3, 6, 8, 11
	Fertility Rate	10, 12, 14, 15, 16
	Infant Mortality	7, 8, 10, 11, 12, 14, 16
	Life Expectancy	10, 11, 12, 16
	Human Development Index	15
	Natural Disasters	13
	Technology Consumption	6, 15, 16
Regionality	Regions	**8**, **10**, **12**, **14**, **16**
	Mountainous Terrain	2, 3, 5, 11

[1] Country Policy and Institutional Assessments

ties will be emboldened to act if they hear or witness others succeeding in similar revolutionary desires. With that in mind, Goldstone considered the influence that neighboring nations 'in conflict' have on 'not in conflict' states, and he found the variable to be significant in his global forecasting models [9]. This is at the heart of the conflict supports theme. Hegre also investigated the idea of neighborhoods and found the border conflict variable to be significant [16]. Similarly, some regional models by Leiby found the border conflict score to be significant in predicting future conflict [20]. Besides bolstering courage, the destabilization of a population due to conflict often produces refugees which could change the demographics in neighboring countries. Whether they encourage others to bear arms or ignite repulsion in stable communities, Shallcross found a refugee variable to be significant in some of his regional models [22]. Interestingly, some of Leiby's *in conflict* models determined significance in the refugee-type variables, but whether it prolonged conflict or assisted in squelching it remained unclear [20].

Besides focusing on disparities of groups within a nation, general statistical demographics can also play a significant role in predicting conflict. Many models considered population statistics, such as population size, population density, or population growth. This is because larger populations appear to be associated with a nonlinear increased risk for conflict [16]. In some models, breaking the population down into subgroups enhanced prediction accuracy, for example, by focusing on the 0-14 year old group labeled *Youth Bulge* or by looking at the military population size.

Just as considering the Youth Bulge variable increased predictions, another important and correlated variable is infant mortality rate. Similarly, health indicators like infant mortality rate, life expectancy, death rate, and caloric intake often have significant value and contribute to the theme of quality of life. Moving from physical elements of quality of life, recreational elements often are significant in models via

measuring technology consumption. Brantley [19] and Neumann [21] measured cell phone subscriptions, Neumann also considered internet usage, and Østby [12] looked at typical household assets such as refrigerators as variables.

As is the case with many of the variables considered, the region variable appears to play a significant role in variable selection as well. Hegre stressed this region variable when observing conflict clustering in nine geographic regions sharing similar risk factors [16]. Boekestein consolidated regional models into six regions based upon Rosling's trendalyzer software that categorized the world into six areas [18]. By using regions as the basis for developing models, Boekestein increased prediction accuracies by 2-7% [27]. However, with the ever-changing environment, categorizations of countries should be updated periodically. Neumann, using her modified k-means algorithm to identify new demarcations for the regions, improved prediction accuracies by 2% after identifying her newly-defined six regions [21]. Regionality typically isn't viewed as a variable itself, but it is used to develop multiple conflict models. Some researchers [7, 14, 15, 29] do include a mountainous terrain variable that accounts for intranational regions.

Without a doubt, social variables play an important role in modeling conflict. However, many studies contradict each other in reference to variable significance for model inclusion, which cannot be reliably assessed by p-value alone as explained by Ward [30]. Hegre and Sambanis present one viable solution to assessing the robustness of variables while investigating 88 concept variables, many of which would be categorized as social influences. Granted, they can only be assessed when the model is sufficiently calibrated. Therefore, the inclusion of three core variables (two of which are from the social aspect), which are theoretically important, are always present in the models: the natural log of population, a decay function for the duration of peacetime, and the natural log of per capita GDP [14]. Top robust social variables

include the influence of neighboring states, regional categorization, rough terrain, ethnic fractionalization, and religion [14].

2.3.3 Core Variables

Although researchers have looked at country conflict from different aspect lenses, there is no way of getting around producing an accurate model without incorporating proxies from all aspects. As noted earlier, there are interactions between variables from different aspects such as GDP per capita contributing to conflict at different rates based on government type, or ethnic diversity contributing to conflict at different rates based on GDP growth. A count of core variables is presented in Table 5 to highlight proxies that should be considered in model building. All variables identified as core variables are listed in Table 5, with the model tally identifying how often a variable was listed in a model and the core tally identifying how often that variable was listed as a core variable in the model. The bold proxies of regime type, GDP per capita, conflict lag/history, population size, and regions share both a substantial number of instances of being significant in the model and being core variables.

Table 5: Core proxies for country conflict modeling

Aspect	Themes	Proxies	Model Tally	Core Tally
Political	Polity	**Regime Type**	10	2
		Political Neighborhood	1	1
Economic	Economic Health	**GDP per Capita**	12	4
		GDP Growth	8	2
		Trade	8	1
		Military Expenditure	4	1
Social	Conflict Supports	**Conflict Lag/History**	9	3
		Military Alliances	2	1
	Population	Population Dispersion	3	1
		Population Size	8	3
	Regionality	**Regions**	5	5

With the understanding of conflict, the conditions for country conflict prediction,

and a taxonomy of political, economic, and social aspects with their associated themes and underlying proxies, the next section will explore sources of these proxies, the available datasets, along with how they were used in various modeling efforts.

2.4 Available Datasets

2.4.1 Correlates of War Project (COW)

The COW contains over a dozen disparate datasets describing violent and non-violent conflicts between and within countries. These datasets continue to receive version changes to offer increased detail in their interpretation. Historically, the COW emphasized intranational type conflict, but the introduction of "Militarized Interstate Dispute" (MID) data, which emphasizes conflict short of war, has led to neglection of COW's use for civil war type modeling [31]. The datasets contain variables for conflicting nations, start and end date down to the day, and an eight-level categorization for number of fatalities parsed by conflict side. Goldstone used these datasets to interpret civil war conflict periods using the following criteria: 1) 1,000 deaths over the entire event period involving state forces, 2) sustained 100 deaths per year [9]. The starting year of the event was the first year to contain at least 100 deaths with the last year also containing 100 deaths along with the subsequent three following years containing less than 100 deaths per year [9]. Although the MID dataset is not necessarily a dataset for war, the MID dataset could be beneficial for modeling "near misses" that have all the modeling conditions of war, but fail to escalate into war as seen in other datasets [31].

2.4.2 Uppsala Conflict Data Program (UCDP/PRIO)

The UCDP/PRIO datasets are similar to the COW datasets in that they also contain breakouts of conflict nations, conflict time period down to the day, and fatality

level. For UCDP/PRIO, there are only two categories of fatality level called intensity level: minor being between 25 and 999 battle-related deaths and major being at least 1,000 battle-related deaths. Østby included every armed conflict between a state government and an organized opposition group that caused at least 25 battle-related deaths per year conditioned upon the conflict falling below the casualty threshold for at least two subsequent consecutive calendar years [12]. Should the casualty threshold reach 25 battle-related deaths after two previous years that failed to reach the threshold, then the conflict is considered a separate onset [12]. Additionally, sub-conflicts were merged only if they differed in conflict type [12]. Hegre leveraged the conflict intensity to further distinguish between no conflict, minor conflict and major conflict [16]. He also focused on conflict year rather than capturing only conflict event periods, including only the primary conflict country and excluding any intervening countries [16]. Celiku and Kraay returned to the binary indicator of conflict combining minor and major intensity levels, however, they also focused on conflict year rather than conflict event [13].

Both Østby and Hegre also used the datasets to factor in historical context for their models. Seemingly, conflicts have run a high risk of recurring within the first post-conflict decade [16]. Østby's variable captured this in *Peace Years* using whole number of years since the end of the last conflict [12]. Hegre used a three-level dummy variable called *Conflict History* to capture the conflict intensity of the previous year [16]. Additionally, he also included the consecutive number of non-conflict years leading up to the prior dummy history variable [16]. He noted that previous studies estimate that only a third of countries succeeded in keeping peace beyond ten years [16].

One of the significant social aspects of conflict is the idea of being influenced within an environment of conflict. Hegre expressed this as the conflict neighborhood.

He initialized his model with a dummy variable indicating if the country shared at least 100km of border where a neighboring country also experienced a conflict [16]. As the simulation model continued, the dummy variable was updated from the prior year's prediction output.

Although not a predictive model, Themnèr and Wallensteen observed in the UCDP/PRIO data that there is a shift in the types of conflicts being recorded. Interstate conflicts are becoming increasingly rarer while internationalized conflicts are seeing a rising trend [32]. They further hint that while intrastate conflict make up the majority of conflicts, peace-making between intrastate conflicts is more likely than internationalized conflicts [32].

2.4.3 Heidelberg Institute for International Conflict Research (HIIK)

HIIK is a third option for defining country conflicts. HIIK evaluates country-year pairs and maps conflict intensity level into five categories (dispute, non-violent crisis, violent crisis, limited war, and war) with violent crisis, limited war, and war further defined as violent conflicts. They assess the level of violent conflict through the application of five proxy measures: weapons, personnel, casualties, destruction, and refugees/internally displaced persons [23]. Weapons and personnel are considered conflict means while the other three proxies are considered conflict consequences. By including conflict means, it provides an alternative to merging both the COW's *War Data* and the MID data. To compare HIIK against COW and UCDP/PRIO, the casualty measure provides one proxy point for between 20-60 casualties and two points for over 60 casualties. The other measures share similar scoring breakouts between zero and two points. Aggregating the five individual measure scores according to conflict means and conflict consequences result in a total intensity level. Boekestein leveraged the binomial indicator of violent or non-violent conflict for the conflict

country-year pair [18]. Shallcross in turn incorporated a conflict history variable into the dependent variable by observing the transition between violent and non-violent conflict for the current year conditioned on the prior year for each country [22]. The conditional dependent variable transition occurred if the status changed over the course of the year [22].

To retain some information from the HIIK intensity level, Boekestein also included a 2-Year HIIK Intensity Level Trend variable. The variable has a two year lag and is calculated as the difference between two consecutive year HIIK intensity levels divided by the six degrees of intensity level [18]. For example, the 2013 trend variable would be the difference between the 2010 and 2011 HIIK intensity level divided by six.

2.4.4 Center for Systematic Peace

One of the main variables in predicting nation conflicts is the political proxy indicator Polity, provided by the Center for Systemic Peace. This indicator classifies the country's historical record subject to regime characteristics and political dynamics. It is a 21-point scale from pure autocracy to pure democracy, but it also includes a few codes outside the scale for disruptive dynamics such as interruption (foreign power influence), interregnum (internal power influence) or transition. Currently, the Polity Project contains yearly data from 1800-2018 of 194 unique country codes. The current Polity Project is version V, with version IV running from 2000-2010.

In Fearon and Laitin, instead of using the raw score from Polity IV to account for political influence in their model, they used a dummy variable indicating a 3 or greater regime index change within the three previous years [7]. Goldstone took a similar approach with capturing adverse regime changes, but used a 6 or greater regime index change within the three previous years [9]. Østby took a different categorical approach using dummy variables and binned similar scores together into three cate-

gories: autocracies (-10 to -6), semidemocracies (-5 to 5), and democracy (6 to 10) [12]. She also included another regime indicator denoting the curvilinear relationship among regime types by subtracting the score of the autocracy from that of democracy [12]. Continuing the non-linear approach to regime type scoring, Goldstone also devised his own five classifications of government derived by using two component variables from the Polity database [9].

Boekestein considered Polity IV scores when building his models, however he imputed a Polity score for countries that were classified as special cases [18]. Additionally, he categorized regime type into dummy variables to mimic Goldstone, but Boekestein's regime type is not immediately calculated from Polity as previous studies have done. Shallcross [22] and Leiby [20] also considered the Polity IV scores when building models, but found Shallcross's derived government type classification to contain more interpretable information when incorporating purposeful-selection and stepwise regression respectively. Government types were indicated by dummy variables for the following six levels of polity score: Emerging Democratic Government (Polity IV: -5 to +5), Democratic Government (Polity IV: +6 to +10), Foreign Interruption (Polity IV: -66), Anarchy (Polity IV: -77), and Transitional Government (Polity IV: -88) [22]. Regime type from the Boekestein study, also included in the Shallcross and Leiby database, along with government type are transformations of Polity. In the Leiby models, if a political indicator proved high explainability in the model, it was typically the government type proxy [20].

Hegre and Sambanis highlighted Polity as an important concept variable in model building. Through their sensitivity study on historical concept variables, they capture a decay function measuring the number of years since a three-point change in the Polity index and a coded value of a Polity change within a three-year interval; both being highly significant and robust [14]. Their study also highlighted other variants

of Polity inclusion in models, citing a variable that reflects the extent of regulated political participation as highly robust [14].

The Center for Systemic Peace also includes a few other political related databases. Goldstone used the Major Episodes of Political Violence database to indicate conflict-ridden neighborhoods. The indicator flags when a country bordered four or more states that experience major armed civil or ethnic conflicts, as described in the database for the given year [9]. This usage preceded using the independent variable, border conflict binary indicator, as found in the Leiby study. Additionally, both Boekestein [18] and Leiby [20] used a similar independent variable derived from a combination of the dependent variable 'in conflict' and the CIA Factbook 'shared land boundary' report.

2.4.5 CIA World Factbook

The CIA World Factbook contains a plethora of country information in the form of comparisons and reports. Under the Guide to Country Profiles, land boundaries for each country are discussed in total kilometers, number of bordering countries, along with the kilometers by country making up the border. The Boekestein study used the "bad neighborhood" indicator labeled *Border Conflict* constructed by multiplying the percent of bordering landmass as described in the CIA World Factbook by the binary dependent variable lagged by two years and then summing the percentages together for a score between zero and one [18]. Island nations were considered to have no bordering countries and therefore always received a score of zero [18].

Both ethnicity and religion percentages are also provided under the Guide to Country Profiles. Notes for each country are provided to describe how these percentages are determined as it varies from country to country. Fearon and Laitin initially used four different measures of fractionalization in their study of which two were derived

from the CIA World Factbook report with one of the measures being the percentage of the population that belonged to the largest ethnic group recorded for each country [7]. Another measure was an ethnolinguistic fractionalization index, a derived proxy from the country's religion percentages [7]. Alternatively, they also developed dummy variables for both ethnic and religious diversity by indicating which countries had their largest groups exceeding 49% and their second largest groups exceeding 7% [7]. They hypothesized that the larger the fractionalization between the majority and the minority, the greater risk of civil war.

Boekestein developed a similar predictor using just the percentage of the population for the dominant ethnic and religious group [18]. These same values were used in the database for the Shallcross and Leiby studies. Furthermore, Boekestein developed a regime type variable based on the government types provided by the CIA World Factbook. The government types were mapped to three categories of regimes: central ruler/ruling party, democratic, and a catch-all for transitioning regimes [18]. This is a proxy definition change from previous studies which calculated their regime type off the Polity score. In the Boekestein PEMSII database, *Regime Type* is based off government type from the CIA World Factbook while *Government Type* is based off the Polity from the Center of Systemic Peace.

2.4.6 Freedom House

Freedom House assesses the level of access to political rights and civil liberties in 210 countries and territories on an annual basis. The database captures two subcategories, political rights and civil liberties in order to provide a total freedom score. The political rights category is based on a variety of questions to maintain a 40-point scoring system. Similarly, civil liberties maintain a 60-point scoring system through a variety of questions about the country. Prior to 2020, these raw scores were nor-

malized on a 1 to 7 scale. Boekestein averaged the political rights and civil liberties scores to develop the *Freedom* variable used in his study [18]. This is different than the score now provided by Freedom House which sums the two raw scores for a rating between 0 and 100. However, Freedom House does maintain the breakout between political rights and civil liberties as well as the normalized 1-to-7 scale conversion in their database. Shallcross noted he preferred the normalized scale to Freedom House's aggregated scoring method as it removes the bias attributed to having an uneven dual scoring system [27]. Boekestein also used a 2-year, 3-year, and 5-year freedom score trend which required a 2-year lag [18]. These trend scores were the difference between the 2-year lag and the corresponding trend year divided by 7 to account for the scoring range [18].

2.4.7 Food and Agriculture Organization of the United Nations

Boekestein included a caloric intake variable obtained from the Food and Agriculture Organization of the United Nations FAOSTAT database [18]. The variables "Crops Primary Equivalent" and "Livestock and Fish Primary Equivalent" are two indicators that both have food supply (kcal/capita/day) elements that mirror Boekestein's description of the *Caloric Intake* variable. The variable had a 4-year lag, so 2011 data became a proxy for 2012 data in order to close the gap where most of the other variables were current up to 2013 for his 2015 study [18].

2.4.8 International Institute of Applied Systems Analysis

The International Institute of Applied Systems Analysis hosts a data repository where the resources are documented from referred scientific literature. Hegre used data from the Demographic Health Surveys, Labour Force Surveys and national censuses in order to obtain an education level variable [16]. The variable measured male

secondary education defined as the proportion of men aged 20-24 with secondary or higher education compared to all men of the same age [16]. It inferred that greater education level reduced the risk of civil war outbreak or shortened the conflict duration if an outbreak occurred.

2.4.9 Minorities at Risk Project

The Minorities at Risk Project is a "university-based research project that monitors and analyzes the status and conflicts of politically-active communal groups in all countries with a current population of at least 500,000" [33]. Østby measured the political exclusion of minority groups with the variable POLDIS which is a 5-point scale political discrimination index starting at 0 for no discrimination and going up to 4 for exclusion/repressive policy [12]. An index of -99 indicates no basis for judgement exists [33]. Østby multiplied the 0-4 index score with "the population share of the minority discriminated against" and in cases where there existed several minorities, an index "sum of all population-weighted discrimination" was taken [12]. Goldstone also used the Minority at Risk Project to develop his binary indicator of State-led discrimination [9]. He indicated that discrimination existed if either the political (POLDIS) or economic (ECDIS) variable was coded as a 4 [9].

2.4.10 World Bank

The World Bank hosts a multitude of databases categorizing statistics for countries around the world. The associated website hosts an analytical tool to graph trends over time by country, as well as allowing downloads of the raw data. Boekestein listed that many variables in his database were taken directly from World Bank to include arable land, birth rate, death rate, fertility rate, percent of the populations with access to improved water sources, life expectancy, and refugee population by country of both

origin and asylum [18]. These same statistics were also in the Shallcross study and Leiby study. Shallcross further added an additional water variable, renewable internal freshwater resources per capita in cubic meters, that due to limitations in the dataset was a stationary variable averaging the 2007, 2012, and 2013 statistics for his 2016 study [22].

The Boekestein [18] database used infant mortality rate as the actual rate of infant deaths per 1,000 live births per year whereas Goldstone [9] employed a log transformation and normalized the year of observation according to the global average. Modelers also varied their approach to national GDP per capita. Boekestein [18] once again chose to use the raw data converted into current US dollars whereas Fearon and Laitin [7] used 1985 US dollars lagged one year with missing values imputed with data on per capita energy consumption. Meanwhile, Østby [12] used the natural log measured in constant 1995 US dollars lagged by one year. Shallcross brought in the percent of central government's military expenditures and also considered a transformation of the nation's GDP [27]. Fearon and Laitin examined oil export figures and identified country-years where oil exports exceeded one-third of the export revenues [7]. The Hegre study also used this same variable [16]. Boekestein devised a trade variable as a percent of the GDP by summing the imports of goods and services with the exports of goods and services [18].

Population is a popularly cited variable in most studies. Fearon and Laitin [7] based their data largely on the World Bank figures and incorporated them as the log of population which Østby [12] followed. Boekestein also considered the rural population as a percent of the total national population [18]. The Boekestein PEMSII database also included the population density of people per square kilometer of land area as well as the annual percent population growth [18]. Boekestein's database also incorporated unemployment as a percent of the total male labor force [18]. Shallcross added a youth

bulge variable to the database that used the nation's population between the ages of 0 to 14 as a percentage of the total population [27].

Table 6 provides a summary of relevant databases, variables, and constructed proxies. These data sources provide a wealth of data from which to develop models to predict country conflict. The next section highlights the modeling and analytical techniques used to date.

Table 6: Database sample variables and constructed proxies

Database	Variable	Constructed Proxies
COW	Violent Conflict	
UCDP/PRIO	Violent Conflict	Conflict History Conflict Neighborhoods
HIIK	Violent Conflict	Conflict History Conflict Neighborhoods
Center for Systemic Peace	Polity	Regime type Government type Conflict neighborhood
CIA World Factbook	Ethnic diversity Religious diversity Regime type	
Freedom House	Political rights Civil liberties Freedom	Freedom Freedom trend
Food and Agriculture Organization of the United Nations	Caloric intake	
International Institute of Applied Systems Analysis	Education level	
Minorities at Risk	Political discrimination Economic discrimination	
World Bank	Arable land Birth rate Death rate Fertility rate Water access Renewable water Life expectancy Refugees Infant mortality rate GDP per capita Military expenditure Oil exporter Trade Population Population density Population growth Rural population Unemployment Youth bulge	

2.5 Modeling And Analytical Techniques

Much of the research to date on predicting conflict uses binomial logistic regression. Many researchers not only wish to obtain good accuracy at predicting, but also desire to maintain interpretability of the model as well. Therefore, logistic regression with its restrictive modeling assumptions provides a good trade-off between prediction and interpretability. The estimated parameter values in the logit transformation, from logistic regression, offer a simple interpretation for non-technical discussions assuming multicollinearity between variables as sufficiently mitigated, enough independent samples were included, and that the independent variables are linearly related to the log odds ratio. Alternatively, decisions trees may be able to overcome multicollinearity and overlook linear relationships among variables, but at the expense of having solid parameter estimates, especially as more intricate variable rules are determined further down the tree. Neural networks provide robust predictions but with the interconnected algorithms for pattern recognition, neural networks are very difficult to interpret [4]. Further in-depth explanation of these techniques can be found in our resources [34, 35, 36].

Early researchers such as Gartzke, Fearon and Laitin, and Østby all used logistic regression to identify main effect variables significant to their conflict models. Their implemented logistic regression concluded variable importance based on p-value. A p-value indicates, under a null hypothesis, the probability that the expected test statistic is as extreme or more extreme than the one calculated. Although relying on p-value is a good start into understanding country conflict, it is no longer necessarily a preferred approach in the country conflict modeling community. In 2010, Ward, Greenhill and Bakke highlighted that using p-value alone in conflict modeling only provides modest improvements in conflict prediction and many times predict poorly despite variables being statistically significant in the model [30]. Initially, they presumed p-value would

fail to prune variables and thus contribute to model overfitting, but noted that Fearon and Laitin's model did not suffer overfitting despite having aspects of poor predictive power [30]. However, their conclusions urged researchers to focus on cross-validation of modeling building to improve predictions rather than focusing solely on p-value for variable importance, as the goal should be to identify "true parameters" leading to conflict rather than just model fit [30]. Hegre agrees as he lays out best practices to intertwine interpretability and forecasting with observing both in-sample and out-of-sample predictions, evaluating multiple predictive metrics, ensuring replicability and presenting visually meaningful results [4]. Fortunately, we have seen progress in this regard as Buhaug, Cederman, and Gleditsch built upon studies, such as Østby's p-value civil war study of horizontal inequalities, and they included out-of-sample predictions alongside sensitivity analysis [11].

In 2010, Goldstone developed an unconditional logistic regression model of country conflict with his most important goal being predictive insight throughout the sample testing [9]. He noted that "prediction is not the same as explanation or hypothesis testing", but was convinced that if a model proved successful at predicting future events, the expectation is that there would be strong links to variables that are able to reliably discriminate [9]. His modeling also required parsimony and he found that simple models out-performed Fearon and Laitin, achieving as much as 40% fewer misclassified cases [9]. As an aside, Goldstone also considered more complex modeling through neural network analysis, which failed to yield substantially better predictions while also having the downside of increasing the interpretation complexity [9]. Finally, he investigated dependent variable lag and noted a slight accuracy loss between a 1-year lag and 4-year lag, and therefore concluded that conflict prediction should be fairly stable with respect to time lag [9].

Hegre used a dynamic multinomial logit model estimation and then simulated the

behavior of the conflict variable [16]. His approach deviated from his predecessors in a few areas. First, instead of predicting a binary outcome, his dependent variable had three levels: no conflict, minor conflict, and major conflict. Additionally, he combined information from his logit model with a simulation to take advantage of the complexities of "neighbors in war" and "previous conflict" variables. Typically, a threshold is set to convert logit conditional means into a binary outcome, however, he used the logit estimated probabilities for random draws in his simulations and then evened out the impact of individual realizations through multiple runs [16]. By developing new dependent variables for every simulated year, he was able to maintain neighbors in war and previous conflict independent variables for a 40-year forecasting model. Although he was able to maintain speculated values for these two independent variables in his forecasting series and noted that other independent variables had good projections for the prescribed period, he did have to make the assumption that changes in a country's conflict state would not affect the 3-level dependent variable, which is not likely to be true [16]. As for maintaining lessons learned from previous research, Hegre also used a split-sample design to incorporate cross-validation to maintain the integrity of focusing on conflict prediction rather than sole model fit.

Boekestein compared three methods of logistical regression using validation sets to assess model accuracy: correlation method, alternate correlation method, and remove the least significant variable method. The correlation method is a forward stepwise algorithm adding variables according to their correlation with the HIIK intensity levels. The alternate correlation method adds a rule to remove a variable if the variable's hypothesis testing alpha is greater than 0.10. The least significant variable method is a backwards stepwise algorithm, removing variables according to an effect likelihood ratio test. Using a one world model, the prediction accuracy rose only as high as 75%, obtained from the least significant method [18]. In order to improve the accuracy,

Boekestein experimented with five different groupings of models. He obtained greater than 80% prediction accuracy on his validation dataset by implementing regional modeling based on Rosling's country groupings [18]. Additionally, the false positives from the logistic regression models were further modeled by a Markov chain to prove that although the model made some incorrect prediction, there was a consistent set of explanatory variables identified for inciting violent conflict. A 59% chance was found of entering conflict the following year and a 93% chance was determined of entering conflict within four years [18].

Muchlinski took an approach to assess the predictive power of random forests against logistic regression methods with unbalanced civil war data. Unbalanced binary data is a key consideration when developing logistic regression models because a dependent binary variable with an unbalanced conflict target class, especially in the range of 1:100, causes the probability to be rather large of predicting no conflict when there is indeed conflict (the Type 2 error) [15]. His assessment used ten-fold cross-validation (nine training folds) to prove that random forests dominate logistic regression methods under these conditions. What is interesting in his approach is the use of the Gini index to assess variable influence in non-parametric modeling. As suspected, his key variables included GDP, infant mortality, and log of the population size [15]. Additionally, Muchlinski cited an ambiguous mountainous terrain variable as strongly significant [15]. Ultimately, the successes of his random forest modeling capitalized on challenging common assumptions such as linearity in parameters, no multicollinearity, and homoskedasticity, which are all required for unbiased estimates in logistic regression [15].

Shallcross continued the idea of Markov chains and investigated the transition to conflict rather than only the current conflict state. He implemented logistic regression following the successes of regional modeling from Boekestein, but used purposeful-

selection of variables as outlined by Hosmer [36]. According to Bursac, simulations of purposeful selection retained significant variables and confounders superior to stepwise selection methods [37]. Combining transition states and purposeful-selection, improvements were realized with 88% training set accuracies and 84% validation accuracies [22]. The model results were analyzed as discrete yearly transition probabilities in Markov chains to develop sojourn times providing insight into expected transitions with validation accuracies of 85% for 3-year forecasts [27].

Although purposeful-selection often provided superior results in obtaining robust models, Hosmer maintains that stepwise selection is still useful when important covariates are not well-known or associated outcomes are not well understood [36]. Leiby investigated impacts of environmental variables using stepwise logistic regression with a modified forced variable option for predicting conflict transitions. The stepwise process also considered multiple stopping factors to develop parsimonious models, which included chi-square G-statistic limits, movement of classification accuracy threshold, receiver operating characteristic (ROC) curve values, and observations of Hosmer-Lemeshow goodness of fit statistic, $\hat{C}$. Analogous to work done by Boekestein, regional modeling provided superior results than whole world modeling [20]. Furthermore, in eight of the twelve region-state models, forced environmental variables produced more accurate models than using only stepwise regression. Ultimately, prediction accuracy on training data were close to 92%, but prediction accuracy on validation data only approached 82% [20]. The main issue plaguing these models was small sample size for the dependent transitory value greatly influencing specificity, predicting the rare transition away from the steady-state Markov state.

Celiku and Kraay sought to minimize both Type 1 and Type 2 errors in (binary) conflict prediction by minimizing the prediction loss function rather than maximizing the likelihood function is typical in logistic regression [13]. They explored both a

weighted linear combination classifier and a probit regression model that examined the number of covariates exceeding specified thresholds. The approach generally dominated other classification techniques for in-sample predictions, however out-of-sample predictions only correctly identified 90% of conflict events and incorrectly classified 30-40% of non-conflict events [13].

Neumann returned to logistic regression with a focus on developing better regions for model building. She initially implemented principal components analysis to create orthogonal data vectors that were clustered using a modified k-means algorithm to form six new regions [38]. Modeling regions, the newly formed regions by purposeful-selection logistic regression resulted in conflict transition probabilities that were superior to prior modeling efforts by Shallcross. Since initial regions, developed without her modified approach, were not always contiguous, final regions (modified approach) were found using a weighted contiguous parameter. Training data prediction accuracies rose to 92% with validation data prediction accuracies increasing to 87% using new regions [38].

Brantley conducted an examination of four different variable selection techniques to implement logistic regression modeling for predicting a violent conflict state: purposeful-selection of covariates, logistical selection of covariates, principal components regression, and representative principal components regression. Brantley describes representative principal components regression as follow: "select either one or two of the variables highly correlated with the components to represent it as a proxy variable for the component" in an initial multivariate full model as a starting point and then remove variables until all parameter estimates have a significant Wald statistic [19]. Unlike previous studies such as Boekestein, Shallcross, Leiby and Neumann, which investigated 182 countries broken into 6 regions, Brantley only considered 12 nations linked to the Arab Spring. Ultimately, three of the four models achieved greater than

90% prediction accuracy of validation data, with representative principal components regression garnering only 76% prediction accuracy [19]. Although logistical regression of covariates did not achieve the highest validation accuracy, Brantley assesses this technique as the preferred covariate selection method due to its comparable modeling accuracy and superior interpretability.

Mueller and Rauh approached the forecasting of conflict from a text mining approach. They integrated a latent Dirichlet allocation (LDA) model using over 700,000 newspaper articles with a linear model that includes other common fixed effects to predict conflict one year out [39]. They claim their success is due to the machine-learned textual topics being able to provide within-model variation which mimics the overall model's predictive power as measured using ROC curve scores [39]. The out-of-sample verification is just one year into the future and garners ROC area under the curve values in the 80% range. As for variables identified in fixed-effects models through the implementation of the least absolute shrinkage and selection operator (LASSO) technique, there were democracy scores, neighborhoods, infant mortality, and population size as the most important ones.

As seen in Table 7, logistic regression remains the dominant method to predict country conflict used in part or whole in 15 of the 17 models.

Table 7: Conflict modeling by author

Year	Study	Conflict Dataset	Model
2001	Gartzke	COW	**Logistic Regression**
2003	Fearon & Laitin	COW	**Logistic Regression**
2004	Collier & Hoeffler	UCDP/PRIO	**Logistic Regression**
2006	Gates	UCDP/PRIO	**Logistic Regression**
2006	Hegre & Sambanis	UCDP/PRIO	**Logistic Regression**
2008	Østby	UCDP/PRIO	**Logistic Regression**
2010	Goldstone	COW	**Logistic Regression** Neural Nets
2013	Hegre	UCDP/PRIO	Simulation **Multinomial Logistic Regression**
2014	Buhaug	UCDP/PRIO	**Logistic Regression**
2015	Boekestein	HIIK	**Logistic Regression**
2016	Muchlinski	UCDP/PRIO	**Logistic Regression** Random Forest
2016	Shallcross	HIIK	**Logistic Regression** Markov Chains
2017	Celiku & Kraay	UCDP/PRIO	Probit Regression Random Forest
2017	Leiby	HIIK	**Logistic Regression**
2018	Brantley	HIIK	**Logistic Regression**
2018	Mueller & Rauh	UCDP/PRIO	LDA LASSO
2018	Neumann	HIIK	**Logistic Regression** PCA & K-Means Clustering
2019	Kane	HIIK	MICE Imputation

Note: Logistic Regression assumed Binomial unless otherwise specified

2.6 Summary

Throughout the past 20 years, global conflict modeling has steadily improved in part to wider availability of databases and computational power for advanced modeling techniques. The largest impact being the transition to evaluate the goodness of modeling through assessed predictive power. Still, investigation into the robustness of modeled variables is needed to provide confidence in understanding the larger societal impacts for various governmental, non-governmental, and intergovernmental

organizations seeking to make further impact toward peace.

Three main databases continue to evolve with nuances on the structure of models. COW mainly focuses on intrastate conflict but through its additional MID dataset also considers the propensity of war short of fatalities, providing a compelling rationale for its usage. UCDP/PRIO hosts a variety of datasets, however, its current focus is to build out disaggregated datasets through ACLED and GED, which are still limited in their number of included nation states. Although, they have a rich history of being the standard of conflict research with many researchers using their datasets for both international and intranational conflict. HIIK provides an alternative to the fatality definition of conflict using means of engaging in conflict (weapons and personnel), which most closely resembles a combination of fatalities and effects short of fatalities. One hindrance to having good comparisons among historical research studies is that the objectives of multiple dependent-variable databases are not in agreement concerning when a nation is or is not in conflict. When conducting research in this area, it is important to clearly state the definition of conflict and employ the database that is most appropriate for the model.

As far as political, economic, and social aspects, all which drive the selection of independent variables, there is a host of open-access databases that can be used for building models. The question of which variables to include, along with their possible transformations, is still highly debatable but is a significant consideration in model building. Variable inclusion should be based on a combination of both domain knowledge and statistical insight. One thing is clear; conflict cannot be segregated by aspect as it is infused with both political, economic, and social influences. This is clear by adopting a common taxonomy that categorizes proxy variables into major themes. Five proxies continue to surface in modeling either through domain knowledge or statistical significance: Polity through regime types, GDP per capita, conflict history,

population size and regions. Regional grouping plays an important role in increasing prediction accuracy as seen in many of the global studies, although no study has really uncovered what this proxy is actually measuring.

Logistic regression has dominated most of the modeling due in part to its easier interpretability, although explainable artificial intelligence such as random forests has shown promise. The main complaint against logistic regression involves whether or not assumptions have been violated [15]. Although, the inclusion of out-of-sample testing for prediction [30] and sensitivity analysis [14] of variable coefficient estimates has done much to placate or validate fears. Goldstone's ten-year-old neural network model predictions were on par with logistic regression predictions, although with less interpretability, however, they may be worth revisiting if logistic regression assumptions cannot be validated or the implementation of Gini index scores assessments to artificial intelligence do not provide enough insight. Hegre's use of simulation, though, has provided the longest forecast of predictions and adds value in decision analysis for organizations. Alternatively, implementation of Markov chains can provide similar forecasts at less computational cost.

Concerning datasets and methods, there still remain challenges in country conflict and peace research. Specifically, for reasons of conflict, there are often missingness in the datasets, which significantly undermine modeling methods if listwise deletion methods are employed. Imputation methods could mitigate the absence of such data and continue to be a topic of interest [40]. In addition to missingness, the quality of data for in-conflict countries is suspect as compared to not-in-conflict countries [22]. Naturally, not-in-conflict data is much more easily gathered with application of quality assurance methods when compared to in-conflict countries where data may be difficult to obtain, is biased, or extremely noisy. Stationarity is another data challenge since, within each statistical model, the assumption of data and/or coefficient stationary

limits the long-term forecast of conflict prediction. Simulation has sought to mitigate this uncertainty, but still has issues discriminating minor conflicts while also adding considerable complexity to the models [16, 41]. Finally, the models considered thus far often are statistical or structural models that leverage the correlation of variables for whether conflict may or may not exist, which is not necessarily the result in explained causality, therefore, causal analysis as it advances may be a better modeling paradigm to address conflict.

As depicted in Figure 3, this survey of datasets and models for predicting global country conflict considers varying definitions of conflict and associated **dependent-variable datasets**, functions ($\boldsymbol{f}$) to define different modeling and analytical techniques for predicting country conflict, and a taxonomy with associated techniques for modeling independent variables gleaned from recent analytical, experimental, and modeling efforts. All three components are vital to achieve reliable **conflict prediction** of country conflict, and this survey provides a basis from which to expand on this research area.

Through the approach presented in Figure 3, this paper provides a complete functional ontology given by the refinement of the ontological framework provided in Figure 1, along with a taxonomy of variables for country conflict prediction in Figure

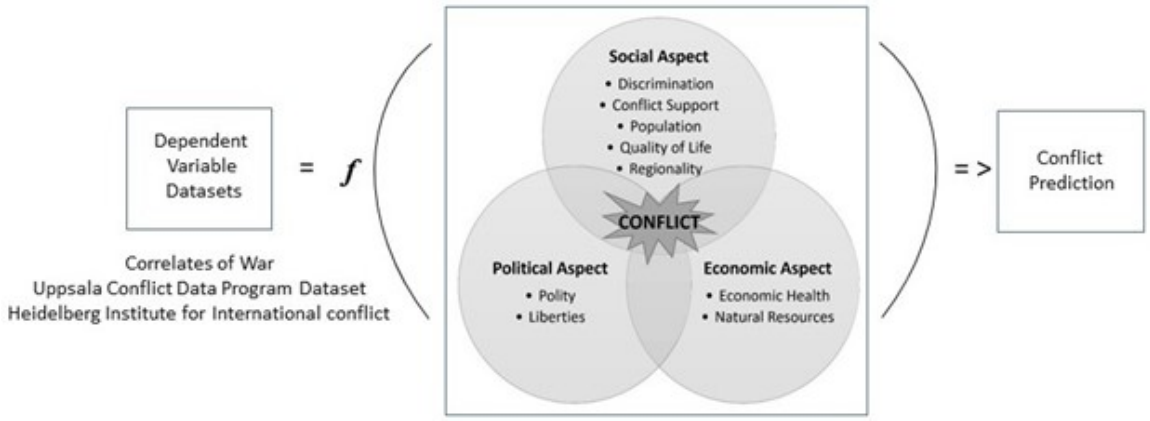

Figure 3: Completed functional ontology for predicting global country conflict

2, informed by a survey of data and empirical modeling techniques for understanding conflict through modeling at a global level.

III. A Large Dataset Imputation Approach Applied to Country Conflict Prediction Data

3.1 Abstract

This study demonstrates an alternative stochastic imputation approach for large datasets when preferred commercial packages struggle to iterate due to numerical problems. A large country conflict dataset motivates the search to impute missing values well over a common threshold of 20% missingness. The methodology capitalizes on correlation while using model residuals to provide the uncertainty in estimating unknown values. Examination of the methodology provides insight toward choosing linear or nonlinear modeling terms. Static tolerances common in most packages are replaced with tailorable tolerances that exploit residuals to fit each data element. The methodology evaluation includes observing computation time, model fit, and the comparison of known values to replaced values created through imputation. Overall, the country conflict dataset illustrates promise with modeling first-order interactions, while presenting a need for further refinement that mimics predictive mean matching.

3.2 Introduction

Imputation methods aim to estimate plausible values for gaps that may be found in datasets. Researchers have developed a large variety of methods to overcome missing values through imputation because imputation outperforms non-imputation methods and no single imputation method universally performs the best [42]. Rubin developed multiple imputation in the 1970s as a method for creating a value in a missing datum where uncertainty should be retained, and its remains the best general theory to deal with incomplete datasets [43]. The two main goals of multiple imputation are to estimate a value that is both unbiased and confidence valid [44]. However, some

popular and preferred implementations of multiple imputation struggle to deal with datasets having a large number of data elements or datasets with high missingness. Van Buuren, a pioneer in multiple imputation by chained equations (MICE), lamented that large amounts of missing data or remotely connected data will influence the time required for convergence, where the key to convergence is to achieve independence in the imputations themselves [45]. Si agrees that multiple imputation faces operational challenges concerning their 409 variable large-scale dataset, explaining that MICE cannot directly handle skip patterns and requires additional efforts to account for logical or consistency bounds [46]. Others also contend that MICE is a superior approach in special cases, but faces problems with high-dimensional data [47, 48].

The motivating case study for this research uses data from the Internal Conflict Database, which is a repository of open-source data consolidated for the purposes of peace research. The open-source data comes from various data collectors such as the Center for Systemic Peace, the CIA World Factbook, Food and Agriculture Organization of the United Nations, Freedom House, World Bank, and a variety of other organizations. From the database, 932 continuous data proxies were selected representative of all aspects of society from political to economic to social themes in preparation of future region categorization research. The scope of the observations consists of the decade between 2006 to 2015, including the 173 United Nations (UN) member countries with over 250K total population as of 2016. Of the selected data elements, 74 capture complete data leaving the remaining vectors with an average missingness of 17.5%.

Prior country conflict research by Brantley [19] and Kane [49] demonstrated the superiority of MICE as the technique of choice for imputing missing data for country conflict data. Specifically, they both agreed that the multivariate method of predictive mean matching within MICE dominated other methods for most variables. Their

assumptions rested on missing values being missing at random, which is made plausible by either limiting the country-year pair observations examined or limiting the scope of variables necessary for modeling. Brantley removed variables where entire country time-series periods were missing [19]. Kane chose only 32 significant variables from prior studies that predict country conflict, but only accounted for less than half of the percent missingness (6.79%) that is being researched in this study [49].

Attempting to apply their approach to a larger country conflict dataset resulted in algorithm computational failures. To illustrate, the R package MICE, used by both Brantley and Kane, failed to iterate one predictive mean matching pass of the 932 data elements within a 7-day computation period. Known barriers to algorithms like MICE include numerical problems from perfect prediction or collinearity, resulting in a failure to iterate [50]. A Python multiple imputation package, Iterative Imputer, also ran into computation issues, exceeding 64 GB of allocated memory after 15 iterations without converging.

In project management, it is often said that mangers must choose between only two of three constraints: time, cost, and quality. A similar sentiment may be said about analysis concerning time, computational power, and accuracy. With computational power being a fixed limiting constraint, a balancing act becomes necessary to implement an algorithm that maximizes accuracy within a reasonably defined time period. This paper presents an algorithm to impute very large datasets, outside the limits of existing packages, striking that balance between time and accuracy through a multiple imputation stepwise correlation multivariate regression approach.

The approach is similar to stochastic regression imputation where the point estimate from the regression equation is modified with a noise component to address upwards correlation bias and underestimated variability. Instead of relying on p-values to determine significant variables for the regression equation, a stepwise approach

observing correlation values is presented to determine feasible independent variables where their significance is assessed through the increasing effect of the adjusted-R^2 statistic. This methodology development study, motivated by the country conflict dataset, imputes numerous variables without running into numerical problems.

3.3 Model Implementation

Rubin describes under a Bayesian approach that creating multiple sets of repeated plausible-imputed values reflects the uncertainty for the nonresponse when the procedure properly considers the complete-data estimates and the associated variance-covariance matrices [44]. That is, the estimates require an approach that considers errors on more correlated independent variables, rather than leaving some out, to overcome biased estimates and that combinations up to some level of interactions should possibly be considered [44]. The modeling approach used in this research takes advantage of a regression model with a noise component produced from the model residuals, also known as stochastic regression. Little views parametric models, such as regressions, as a strength in imputation as the assumptions are explicit [48]. Van Buuren demonstrated that the approach provides unbiased coefficients, although the coverage for confidence validity is not as good (0.908 vs 0.951/0.941) as more computationally intensive Bayesian and bootstrap approaches [43]. However, these computationally intensive methods like MICE become overly burdensome for imputing large datasets as discussed concerning numerical problems. With the regression approach, the benefit of using the residuals to incorporate the uncertainty in the imputation estimates rests on the assumption that the residuals are mean zero and normally distributed. The assumption was visually instantiated showing adequacy for both percent missingness and convergence rate as illustrated in Figure 4.

The core component of the methodology resides in assuming correlated data ele-

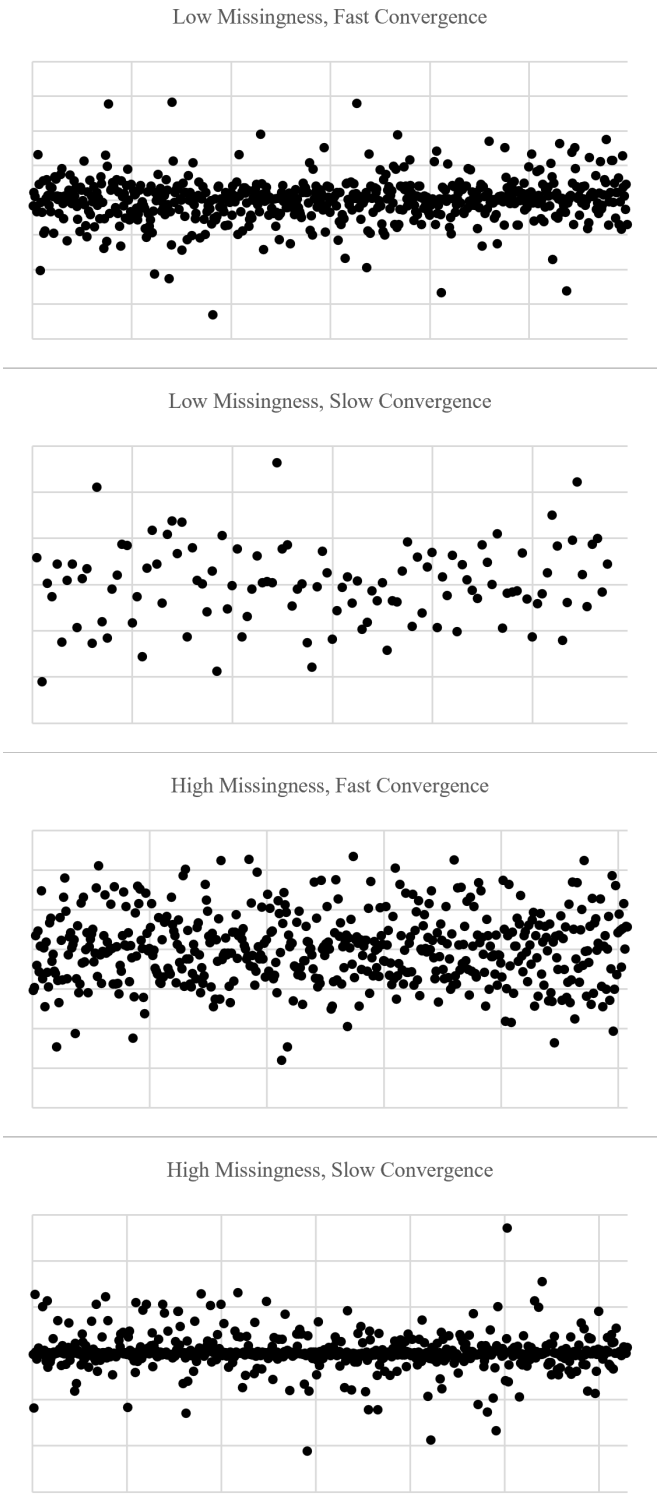

Figure 4: Residuals for model regressions

ments should assist in providing accurate estimates for the missing values in the data. For example, height and weight are often seen as highly positive correlated variables, therefore if weight is missing in a few observations, it would be reasonable to use the height variable to impute the missing data points. Statistically, this concept is represented by the p-value, where the statistic is used to reject the null hypothesis that there is no relationship between the two variables. The benefit in starting with the analysis of correlation manifests in computation time. Whereas each variable would need p-value assessment in a stepwise regression for every iteration to determining significance, only one pairwise analysis of correlation coefficients is required to provide a static ordered list to assess significance for the entire dataset. The ordered list saves thousands of computations every iteration as the process is conducted once before model building rather than every time a model attempts to add a new variable.

Positive or negative correlation is inconsequential to the evaluation of the ordered list; the usefulness is that stronger relationships are considered first. The algorithm computes the absolute value of the Pearson correlation coefficients once, using only the known values in the dataset as seen in (1), where x_i and y_i are sample pairs in two different data elements with n non-missing value pairs. This matrix, $\mathbf{Q}$, provides the foundation for discovering the strongest relationships that improve the model adjusted-R^2 within the least number of trials.

$$|r| = \left| \frac{n \sum_i^n x_i y_i - \sum_i^n x_i \sum_i^n y_i}{\sqrt{[n \sum_i^n x_i^2 - (\sum_i^n x_i)^2][n \sum_i^n y_i^2 - (\sum_i^n y_i)^2]}} \right| \tag{1}$$

Additionally, all data elements are rank ordered from the least proportion of missingness to the greatest proportion of missingness to identify the order in which the imputations will be processed. This ranking approach is similar to MICE where the least missingness is estimated first, in other words, optimizing the order of estimating the dependent variable within a regression model so subsequent imputations can

benefit from observed and currently imputed values of all the other variables in the model [51]. The algorithm dynamically updates the dataset within each iteration to minimize estimation biases presented by missingness within the independent variables. That is, the approach assists in developing complete independent variables for subsequent imputation models, however, the model for the initial dependent variables may encounter missingness requiring preliminary simple imputation such as taking the mean. The biasing mean imputation on the independent missing variables is minimized by first imputing dependent variables with less missingness. As the dependent variable order processes the data elements with more missingness, the candidate independent variables become further complete with robust imputed values rather than weaker preliminary estimates to rectify their initial missingness. Furthermore, as the algorithm iterates, the bias decreases when the mean-estimated imputed independent variable becomes the dependent variable for imputation, garnering a better estimate from its own regression model. The rectification can be observed in the increased adjusted-R^2 for subsequently iterated models as seen in Figure 5 and the quality of the normalized root mean square error discussed in the later sections.

Once these initial two processes of describing the $\mathbf{Q}$ matrix and dependent variable order are complete, the stepwise regression modeling commences. Using the data element missingness-related rank order, the data vector with the least missingness is set as the first dependent variable in need of imputation. Of the 932 data elements available, 74 already had complete data and did not require imputation, leaving 858 data vectors to impute. Using a stepwise approach, the algorithm adds independent variables to the model starting with the data element that has the strongest correlation according to matrix $\mathbf{Q}$ to the dependent variable.

While building the model, the algorithm sets aside a subset of complete data. For those instances when the dependent variable is missing, the associated independent

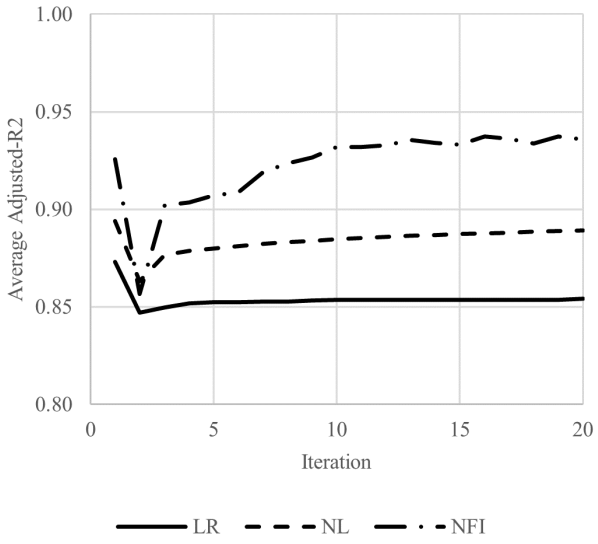

Figure 5: Model average adjusted-R^2

variable data (observation) is removed from the subset. Furthermore, as additional independent variables are considered for inclusion into the model, initial cases arise where additional observations have missing or not yet imputed values in the set. These observations are also omitted from the subset. This mechanism of list deletion could potentially cause violations of the normality assumption of residuals if the degrees of freedom are too great with respect to the number of observations. Therefore, a threshold for an adequate number of observations was assessed before including the candidate independent variable.

There are a variety of recommendations to accommodate maintaining the normality assumption of residuals. For univariate regression, a general rule of thumb maintains at least 30 observations. For multivariant regression, 10-15 observations per independent variable has been demonstrated to be an optimal ratio [52]. A final strategy maintains to keep at least a quarter to half of the observations available for the most limited independent variable in the model. The most limited variable being the variable with the least number of known observations. Five thresholds were

tested: 30 observations, 100 observations, and limiting variable observation ratios of a quarter, a third, and half.

The most conservative constraint (half of the most limiting variable) could reject the most plausible variable in the data set (highest correlation value) more often than desired inserting a less desirable variable concerning correlation value because it better suits maintaining the normality assumption of the residuals. Through testing, the most liberal constraint (at least 30 observations) was only enacted six times in the first iteration allowing the highest correlated variable to almost always enter the model, whereas the most conservative constraint forced an alternative 925 times.

No statistical difference at the 90% confidence level when observing the average, 25^{th} percentile, or 10^{th} percentile for the adjusted-R^2 of the model was identified, meaning model fit was not a factor. There was also no statistical difference when holding missingness as a factor. The most conservative constraint allowed some models to dip as low as 83 observations in the model dependent on the variables included, causing concerns about degrees of freedom and the normality assumption for a 10-variable model. Balancing maintaining a large number of observations while minimizing the number of alternative independent variables, the algorithm was set to a constraint of requiring 100 observations after listwise deleting missing values for model building. The selected cap of 10 independent variables corresponds to a minimum of 10 observations per variable, which is within the aforementioned optimal ratio. This constraint is only necessary for the first iteration as imputed values on subsequent iterations fill in any initial missing values in the data.

Next in the methodology, the candidate variable enters the model for adjusted-R^2 examination. The R^2 represents the explained variance by the independent variable toward the dependent variable. However, the R^2 continues to increase as more variables are introduced whereas adjusted-R^2 penalizes additional variables that fail to

significantly affect the dependent variable. Three different models were examined: linear (LR), nonlinear (NL), and nonlinear with first-order interactions (NFI). The LR model, as illustrated in (2), provided the baseline case of providing parsimonious terms within the regression model, where y is the estimated dependent variable, x_n is the added known independent variable, β_0 is the model intercept, and β_{n1} is the corresponding linear coefficient. The assumption includes that any potential curvilinear relationships within the variables are insignificant. The NL model makes no such assumption and includes squared variable terms, in addition to the linear terms, as seen in (3), if those terms continue to increase the adjusted-R^2 of the model, where all coefficients from the linear model are present along with β_{n2} as the corresponding squared term coefficient. Additionally, the methodology observes the strong heredity assumption, that the geometric global extremum of all variables may not be the special case of zero [53]. The NFI model assesses both squared variables and first-order interactions, in additional to the linear terms, for inclusion as long as the adjusted-R^2 continues to increase for each candidate term as seen in (4) where all coefficients from the linear model are present along with β_{n3} as the corresponding interaction coefficient. Due to the assessment of each additional term, the computation times increases exponentially from LR to NL to NFI. Although the potential exists that additional variables may increase the adjusted-R^2 past 10 modeled variables, a cap of 10 variables was implemented. When considering country conflict datasets, Ray argues that country conflict data should adhere to Achen's "rule of three" when assessing independent variables for regression while Oneal demonstrates the rule to be too strict in examples of up to 8 variables [54]. Van Buuren notes that general regression, overcoming multicollinearity and degree of freedom problems, may be suitable upwards of 25 variables, however, explained variance after 15 variables is typically negligible at best [45]. The maximum 10 variables threshold facilitates a sweet spot

to allow explained variance and manage the list deletion issue presented earlier.

$$y = \beta_0 + \beta_{11}x_1 + ... + \beta_{n1}x_n \tag{2}$$

$$y = (2) + \beta_{12}x_1^2 + ... + \beta_{n2}x_n^2 \tag{3}$$

$$y = (2) + \beta_{n3}x_1x_1 + \beta_{n3}x_1x_2... + \beta_{n3}x_nx_n \tag{4}$$

Should the candidate variable fail to increase the adjusted-R^2, the next top 9 candidate variables are evaluated for inclusion. Observations concluded that on average, three initial candidates out of the 10 allowed variables in the model would fail to increase the adjusted-R^2, however an alternate variable was found to increase the adjusted-R^2 by the third best candidate, necessitating the need to look at subsequent independent variables past the initial failure to increase the adjusted-R^2.

Once the independent variables are identified for the model, the associated data produces the linear coefficients for the model that imputes the missing dependent values according to $\hat{y} = \boldsymbol{\beta} * \boldsymbol{X}$, where $\boldsymbol{\beta}$ are the model coefficient parameters and $\boldsymbol{X}$ are the data vector values for the associated missing dependent variable. This provides a point estimate from which to develop a stochastic regression result. For the first iteration, it is possible that some of the independent values may also be missing as discussed earlier, however, with trying to impute the dependent variable, list deletion is no longer an option. In these cases, an average of the non-missing data vectors estimates a feasible point estimate for the missing data. As previously mentioned, the bias inserted into the imputation diminishes with subsequent iterations as the dependent variable converges toward a more plausible value.

Noise added to the imputed point estimate provides the stochastic element desired in multiple imputation. Using a list of residuals captured from the first iteration, residuals produced only from the original known values, the imputed point estimate

receives an adjustment from a randomly selected residual value to accommodate the uncertainty in the imputation. Seeing that the residuals are distribution normal, the uncertainty will have mean zero with standard deviation one.

Finally, the algorithm checks the stopping rule against the convergence factor to exit iterating each specific data element. The stopping rule compares each imputed before noise point estimate in the data vector from the before noise value of the previous iteration. Should all values within the data vector be less than the convergence factor, the algorithm considers the data element converged. For this study, each data element obtained a tailored convergence factor of three standard deviations of the data element's residuals to account for the different scale in values rather than rely on a static factor for the algorithm. The full pseudocode for the algorithm is provided in Figure 6.

1. Create **Q**, a matrix of absolute value Pearson correlation coefficients **r** of all **p** data vectors.
2. Rank all **p** data vectors in the dataset from least proportion of missingness to greatest proportion of missingness to identify the order in which imputation is processed. Data vectors with few missing elements are imputed first.
3. Create the stepwise regression models.
    a. Using the order from (2), select a data vector as the dependent variable requiring imputation.
    b. Add candidate data vector as independent variable based upon the maximum value in matrix **Q**.
    c. Listwise delete all observations from the model that incorporate a missing value across all variables.
    d. If the number of observations is below the threshold, go to (3b) and choose the next best candidate data vector up to 10 possible attempts. Otherwise solve model and go to (4).
    e. Solve model.
    f. If the adjusted-R^2 fails to improve, go to (3b) and choose the next best candidate data vector up to 10 possible attempts. Otherwise solve model and go to (4).
    g. If there are less than 10 variables in the model, go to (3a) to select another candidate data vector.
    h. Save the model regression coefficients.
    i. If this is the first iteration model with no imputed values, save residuals to be used as noise.
4. Impute missing values in the dependent data vector.
    a. Restore all observations removed during (3c).
    b. Using the model coefficients from (3h) produce point estimate $\hat{y}$ for missing values in the dependent data vector.
    c. Add model residual noise to the estimated $\hat{y}$, using a randomly selected residual from the first iteration model developed in (3i).
5. Assess the stopping rule for iterations against the convergence factor. If data vector has not converged, continue back to (3).

Figure 6: Methodology pseudocode

3.4 Methodology Evaluation

The analyst trade-off of time, computational power and accuracy sparked the development of this methodology due to the "numerical problems" or "breakdowns" of the multiple imputation algorithm in alternative approaches. It is acknowledged that alternative approaches may foster improved plausible accuracy should the algorithms compile an iteration or process data in an acceptable period. This approach provides a choice to analysts with large datasets to balance acceptable time and accuracy. As General Patton suggested, "A good plan violently executed now is better than a perfect plan next week" [55]. In other words, this methodology allows analysts to have good imputations quickly instead of waiting for imputations from higher acclaimed algorithms that either may deliver too late or breakdown.

The time evaluation consists of observing the quantity of data elements converged after a certain number of iterations. Computationally, building the $\mathbf{X}$ matrix takes longer as the complexity of adding squares or interactions enter the model. Furthermore, looping back in the algorithm to find alternative independent variables increases iteration time as well. However, this time addition pales in comparison to the factor of how many data elements require imputation. Each data element takes 0.95 seconds to model under LR, 1.09 seconds under NL, and 1.68 seconds under NFI, with standard error in the milliseconds. The additional time for the more complex models is attributed to evaluating additional candidate terms, namely squared and first-order interaction terms. Recognizing that all models process data elements within a second of each other, the time component can be illustrated by how many data elements still require additional iterations to converge.

Preliminary model validation typically begins with assessing model fit by observing the dependent variable variability as a function of the independent variable variability known as the R^2 statistic. Good regression models desire independent

variables that explain the variation in the dependent variable. The statistic is only useful if the residuals maintain the normal distribution assumption. Furthermore, the statistic always increases as additional independent variables are added to the model, therefore it has no stepwise assessment usefulness. The adjusted-R^2 penalizes additional variables allowing stepwise assessment. Observation of the adjusted-R^2 is twofold. First, a high value signifies that the imputations through the correlation approach may provide plausible values. Second, the initial observation of adjusted-R^2 contains only the known values in the original dataset. By the second iteration, bias was inserted into the dataset through estimating unknown values in the independent variables. Observing the adjusted-R^2 through subsequent iterations alleviates bias concerns as the value reapproaches the initial observation.

Finally, the normalized root means square error (NRMSE) functionally evaluates the goodness of the imputations to recreate known values. The NRMSE value is obtained by dividing the root mean square error by the range of the original data vector as illustrated in (5), where x_{1ip} are the known values in the test set, $\hat{x}_{1ip}$ are the imputed values corresponding to x_{1ip} with N_{1p} test set observations, x_{2p} are the known values in the original set, for the p^{th} data element of P total elements. Normalizing assists in adjusting the value to account for any scaling bias in the statistic with the common choice being range normalization [56]. A test set was created by randomly selecting 8% of the known data for imputation. Van Buuren stresses that imputation is a challenge "to obtain statistically valid inferences from incomplete data" rather than an exercise in accurately determining the unknown true value, especially when using multiple imputation techniques [43]. Despite his angst for root mean square error, he concedes that it is a good metric to evaluate the compromise between bias and variance if the desire is to assess accuracy and precision [43].

$$NRMSE = \sum_{p=1}^{P} \frac{\sqrt{\sum_{i=1}^{N_{1p}} (\hat{x}_{1ip} - x_{1ip})^2 / N_{1p}}}{\max(x_{2p}) - \min(x_{2p})} \tag{5}$$

3.5 Model Results

The majority of data elements converged after only two iterations for the LR model and four iterations for NL. In other words, the difference between the regression point estimates in most vectors were less than three standard deviations of the first iteration residuals. The LR model converged more vectors faster than the other two as seen in Figure 7; and with the fastest time to compute a data element, remained the fastest model type to reach the stopping condition.

The convergence rate appears counter intuitive when considering the average adjust-ed-R^2 of the models seen in Table 8. It was hypothesized that better model fit would increase convergence, however, it was observed that the correlation between the convergence iteration and the data vector adjusted-R^2 was weak (<0.3). Despite this finding, all models produced a high average adjusted-R^2. With NL models producing a higher adjusted-R^2 than LR, the assumption remains plausible that many of the data elements should be characterized in curvilinear form. And supporting Rubin's claim, imputation models benefit further in adjusted-R^2 when modeling independent variables up to at least first-order interactions.

As far as the accuracy of the models, the median NRMSE for the data elements

Table 8: Model average adjusted-R^2, N=10

	Iteration 1		Iteration 20	
Model	Avg	Std Dev	Avg	Std Dev
LR	0.8732	0.0001	0.8541	0.0012
NL	0.8939	0.0000	0.8892	0.0037
NFI	0.9257	0.0003	0.9357	0.0035

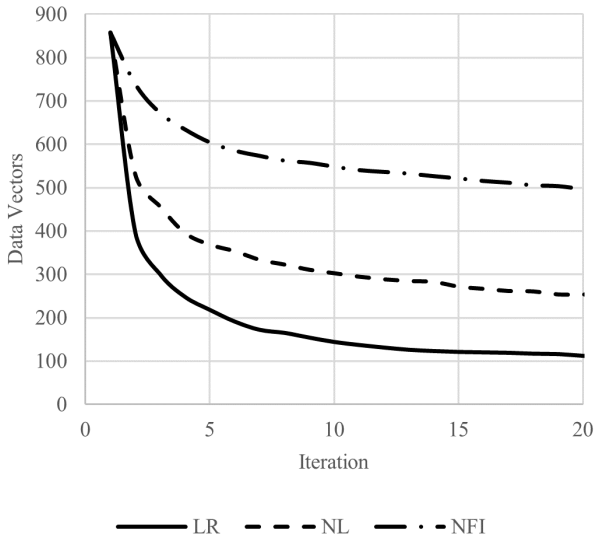

Figure 7: Model convergence rate of data vectors

demonstrated low values after 20 iterations with 0.019 (LR), 0.0216 (NL), and 0.645 (NFI). However, the sum NRMSE was less optimistic with 1,903 (NL) and magnitude higher for NL and NFI. For the LR model, 4 of the 858 vectors had extremely high NRMSE values ranging from 11 to 1,154 inflating the overall NRMSE. All 4 vectors had very high adjusted-R^2 and no connection to percent missingness could be established. It was observed that some data vectors may have imputed values outside the plausible distribution. For example, known values in positive-only vectors had imputed observations with negative values. This remains an obstacle for regression methodologies that do not add limiting bounds like predictive mean matching. The "out-of-bounds" imputations exacerbate the issue for squared terms in the NL and NFI models when selected as independent variables, which lead to a larger number of outliers concerning vector NRMSE.

3.6 Conclusion

This paper presents a methodology to impute large datasets based on convergence of iterations within confidence bounds set by initial regression model residuals and using the information contained within the data correlation matrix. Large datasets increase the presence of numerical issues causing other imputation methods to fail. The regression methodology presented, demonstrated through the country conflict dataset, appears to overcome numerical issues without failed or stalled iterations. The methodology processes data elements quickly and generates high adjusted-R^2 models. Through developing the methodology, a stopping criterion to dynamically define convergence was presented offering a more tailorable condition for when data elements are of different scales. The exploitation of the initial regression model residuals overcomes any guesswork that may be present when submitting a static stopping tolerance offered in other imputation packages. The algorithm balances computation time, computational power, and accuracy to achieve a traceable, defensible approach to imputing large data sets where many preferred commercial packages fail. Despite the mentioned advantages, the methodology could benefit from further refinement. Although the methodology produces useable and defendable results, further work is needed to assure the user of plausible values. Notably, the issue of "out-of-bounds" imputations should be addressed to take further advantage of the improvements from NL and NFI type modeling. Other aspects of research could include investigating multicollinearities within the independent variables, while the dependent variable capitalizes on high correlation selection.

IV. Multicollinearity Applied Stepwise Stochastic Imputation: A Large Dataset Imputation through Correlation-based Regression

4.1 Abstract

This paper presents a stochastic imputation approach for large datasets using a correlation selection methodology when preferred commercial packages struggle to iterate due to numerical problems. A variable range-based guard rail modification is proposed that benefits the convergence rate of data elements while simultaneously providing increased confidence in the plausibility of the imputations. A large country conflict dataset motivates the search to impute missing values well over a common threshold of 20% missingness. The Multicollinearity Applied Stepwise Stochastic imputation methodology (MASS-impute) capitalizes on correlation between variables within the dataset and uses model residuals to estimate unknown values. Examination of the methodology provides insight toward choosing linear or nonlinear modeling terms. Tailorable tolerances exploit residual information to fit each data element. The methodology evaluation includes observing computation time, model fit, and the comparison of known values to replaced values created through imputation. Overall, the methodology provides useable and defendable results in imputing missing elements of a country conflict dataset.

4.2 Introduction

Many popular multiple imputation methods rely on a regression framework to develop plausible missing values [57]. Although no one imputation methods succeeds at being the best in all imputation applications [42], some studies demonstrate k-nearest neighbors as the best single imputation methods and predictive mean match-

ing (pmm) as the best multiple imputation method for the datasets considered [58]. Prior country conflict dataset imputations by Ahner & Brantley [19] and Kane [49] also contend that pmm, a regression approach multiple imputation bounded to only known values for estimates, exhibited superior performance toward their country conflict datasets compared to other tested approaches. However, these prior studies were limited to small datasets of 32 variables. When expanding the 32-variable country conflict dataset into a very large dataset, the preferred pmm approach broke down due to numerical problems [59]. In [59], a new regression approach investigated capitalizing on dependent variable correlation in a stochastic regression framework to overcome numerical problems. Although the approach provided promise with favorable results, the algorithm also suffered from some "out-of-bounds" imputations and concerns over multicollinearities within the independent variables [59]. This research extends the Large Dataset Imputation through Correlation-based Regression approach found in [59] to develop robust imputations by including variable range-based guard rails and exploring correlation selection discounts.

The large dataset considered consists of 932 continuous data proxies or data elements from the Internal Conflict Database [59] allowing direct comparisons between the initial method's results and the extension presented in this paper. The scope of observations involves annual data over 10 years from 173 United Nations (UN) member countries that possess a total population of over 250K. The observations are recorded as country-year pairs for a total of 1,730 observations. This dataset supplies a diverse selection of multiple data elements spanning all three country conflict aspects of political, economic, and social influences. Completing the dataset with plausible imputations assists peace researchers in developing solutions through increasing sample size power, especially when employing analytical modeling.

Within the dataset, 74 of the 932 data elements were complete cases with all

country-year observations. Considering all three patterns of missingness, the missingness of an observation in a data element averaged 17.5% while the missingness of a data element for a country-year pair observation averaged 14.0%. The diversity of missingness in this large dataset presents a good opportunity for multiple imputation which has demonstrated to be robust even when datasets depart from the normality assumption or when the proportion of missingness may be high [60].

The overwhelming majority of data analysis techniques require complete data as mathematical operations cannot be applied to non-values. An easy solution to overcoming this problem is using listwise deletion on the observations with missing values, however, such methods increase biasness, underpower sample sizes, or insert unreliable estimates [61]. For example, when considering this dataset, listwise deletion would reduce the desired 1,730 observations down to an unacceptable 3. Imputation, a mathematical process of inferring a value to an undocumented attribute of an observation, is then necessary to create a more useable dataset for analysis.

With Rubin's proposal of multiple imputation, the identification of three distinct patterns of missingness became standard practice, which include missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [62]. Within the large dataset, all three categorizations can be observed, which accounts for some of the numerical problems encountered when applying pmm. Country conflict data often carries the complexity of missing data through multiple lenses: seeing missingness through unique country-year pairs as observations, missingness through unique countries over a time-series of years, and missingness as individual occurrences across multiple variables. Some of the missingness could be identified as missing at random, while some are obviously worse case as missing not at random.

MCAR is data missing as a random effect in the sample, or in more colloquial terms, due to just bad luck. The missingness is not correlated or dependent to

any observed or unobserved independent or dependent measurement. MCAR data is rarely found in practice, however, it can be perceived that very low missingness in a dataset could be identified as such. Each complete data vector would consist of 1,730 observations, where each represented country may have 10 time-series data points. A worst-case scenario would imply that all missingness came from one country. Therefore, keeping missingness less than half the country observations and claiming MCAR would place a data element with at most 4 missing observations and still be considered MCAR. This happens in 73 of the 932 data elements.

More commonly, MAR ties the missingness to an observed measurement, however, the missing data does not depend on the value of the missing data. In the dataset, such missingness may manifest in areas such as the *Corruption Perception Index* score not being recorded for a country that is in conflict or data not being measured due to a country having an autocratic government and controlling what information is available to the public. However, the missing values may be validly imputed by considering other observed variables in a model. This assumption would fit the majority of missingness in the dataset.

MNAR ties the value of the missingness to the missing value itself or when the missingness may not be understood by any other observed value. This could manifest at the intersection where both high missingness rates are seen across the time-series and within a variable column. Such examples include a country having no time-series data for a variable and the variable across countries also having high missingness; for instance, observational data for the Democratic Republic of Korea having no time-series data for *Battle-Related Deaths* along with the data element also having a cumulative 84% missingness. This applies to at least three data elements which are observed with scrutiny.

Two main issues surfaced in [59] while developing the concept for the Large

Dataset Imputation through Correlation-based Regression approach. First, there are concerns about multicollinearity of independent variables effecting the stability of regression coefficients. Second, regression results may produce imputation estimates that are outliers to the distribution of known values undermining the confidence in the plausibility of the imputed values. The combination of these two issues are assumed responsible for the imputed data vectors that experienced extremely high root mean square error values [59]. These issues are addressed in this new imputation process Multicollinearity Applied Stepwise Stochastic Imputation (MASS-impute).

4.3 Model Implementation

Reviewing the original proposed algorithm, the steps can be categorized into three main segments: pre-processing, regression modeling, and imputation development. Pre-processing consists of two parts: developing a correlation matrix used for nominating variables and a ranking of variables by missingness. The correlation matrix consists of the absolute value Pearson correlation coefficients, $\boldsymbol{r}$, used for variable selection, and designated as matrix $\mathbf{Q}$. The rank ordering of data elements establishes that order for imputation with the least missing elements undergoing the imputation process before data vectors with more missingness. This is consistent with other imputation methods using multiple imputation by chain equations (MICE) [51]. No changes to the pre-processing segment were made from the original method in [59].

The modeling segment selects up to 10 variables for inclusion into a regression model to estimate the missing values of a single data element, of which 96% of data elements typically select the maximum, varying slightly from iteration to iteration. Producing candidate regression coefficients uses a stepwise process of evaluating candidate variables with the goal of increasing the adjusted-R^2 statistic. The original method selected candidate independent variables that had high r linear correlation

scores with the dependent variable. The method structures itself by leveraging variables that provide as much useful information as possible to estimate missing data points. Theoretically, if one independent variable were perfectly correlated with the dependent data element having missing values, then it would be expected that perfect prediction could be obtained. Therefore, when selecting individual candidate variables for inclusion into the model, the main criteria focuses on increasing the adjusted-R^2 without regard to multicollinearity with other independent variables. Often, analysts highlight multicollinearity as a concern when building models; modeling with highly correlated independent variables produce unstable estimates, inflated variances, and confounding effects, although coefficient instability may be a consequence of multicollinearity rather than a product of it [50, 63]. To clarify, the perceived multicollinearity problem consists between only the univariate independent data elements themselves but not with modeling constructs such as interaction product terms. Modeling square terms or product terms often highly correlate with the individual independent variables, yet do not create multicollinearity problems as "multicollinearity neither affects the value of the coefficient of the product term nor inflates its standard error" [64]. The multicollinearity problem typically concerns model analysis rather than modeling for imputation purposes. Still, imputation practitioners pause for concern when reading van Buuren's statement that using several hundred variables in multiple imputation cannot be feasible due to multicollinearity and computational problems [45]. Solutions to the multicollinearity problem often include removing variables to increase parsimony. Some suggest removing variables in imputation models should they have large amounts of missing data due to incomplete cases, failure to have adequate association with the dependent variable (absolute correlation value greater than 0.5), or high correlation with other independent variables resulting in not adding additional value to the model [50]. Yet excluding variables

with high partial correlation simultaneously increases the risk of omitted variable bias [63]. Despite the hazards of multicollinearity, the implications may be better described as a problem of degree rather than kind [65], therefore this research presents variable selection conditioned on degrees through correlation discounting. In other words, multicollinearity of independent variables may be addressed through variable selection with a discount.

Before applying a discount to the variable selection criteria, it is necessary to establish when to apply a discount. If correlation is too high, multicollinearity concerns exist and discounting is deemed necessary. If discounting is applied too heavily, the algorithm may omit valuable variables resulting in a less than optimal imputation. To aid in proper variable selection, five categories of correlation are defined: very high (1.0-0.9), high (0.9-0.7), moderate (0.7-0.5), low (0.5-0.3), and negligible (0.3-0.0) as illustrated in Table 9. Some data elements have correlation values in each correlation category while others data elements may only be represented in a few categories. It was noted that 5 of the 932 data elements consisted of all correlation values below 0.5, suggesting they would not be strong candidates for inclusion in the model.

Table 9: Correlation categories with no discounting

Correlation Category	Number of Data Elements	Percent of Elements Including Category
Very High (1.0-0.9)	597	64%
High (0.9-0.7)	709	76%
Moderate (0.7-0.5)	829	89%
Low (0.5-0.3)	919	99%
Negligible (0.3-0.0)	932	100%

The method uses a forward stepwise linear regression approach, in the form of $\hat{y} = \boldsymbol{\beta} * \boldsymbol{X}$ where $\hat{y}$ are imputed results from $\mathbf{X}$ data elements with associated $\boldsymbol{\beta}$ coefficients, which economizes on computational effort. Through this method of stepwise addition by correlation value, the method nominates variables with the highest r

absolute value. Limiting the multivariate regression equation to only 10 variables, it is highly unlikely that correlation values below 0.5 are included in any models. However, addressing multicollinearity, the high correlation between independent variables adding little value is addressed through exploring discounting of a variable's r value based on its correlation with variables already in the model. This provides the first deviation from [59] as variables are now nominated through a discount matrix rather than matrix $\mathbf{Q}$ in order to mitigate multicollinearity between independent variables. Data elements with at least one very high correlated variable account for 64% of the dataset, having a median number of just one variable, as illustrated in Figure 8. If no discounting is present when selecting variables, 415 data elements have at least the first two candidate variables with an absolute value collinearity above 0.9 and 139 data elements potentially consisting of all 10 variables within that very high category. The discount process alleviates this situation.

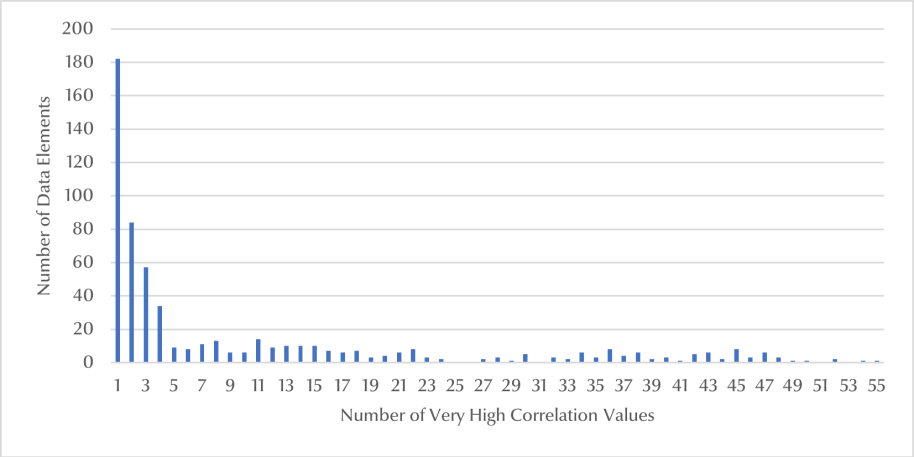

Figure 8: "Very high" correlated values in data elements

Four discount strategies were examined on a degree scale and are described as follows: None, Cube, Square, and Max. The None discount is the base case where no discount is applied. The algorithm chooses the next best variable based on correlation with the dependent variable. This baseline case illustrates the effects of

multicollinearity among independent variables in the imputation model and whether multicollinearity should be a concern. Although almost-linear related predictors are frequently a source of problems for imputation [61], this baseline assists in quantifying how much a problem may be present within the large country conflict dataset considered [45]. At the other extreme is the Max discount. The Max discount chooses the next best variable based on adjusted correlation with the dependent variable by comparing each candidate variable's correlation with the dependent variable after subtracting the maximum correlation between the candidate variable and the variables already included in the model. The Max discount, along with the Square and Cube variant can be seen in Equations 6-8 (Max, Square, Cube respectively), where $\mathbf{A}_{i,j}$ is the discounted absolute value correlation score, $\mathbf{Q}_{i,0j}$ are the original absolute value Pearson correlation coefficients for dependent variable i and nominated independent variable j, and $\mathbf{Q}_{i,nj}$ are the original absolute correlation values of variables currently added to the model associated with the dependent variable. Matrix $\mathbf{A}$ then, in all cases, is the transformed correlation matrix after the appropriate discount from with to choose the next data element j with the maximum discount value.

$$Max : \mathbf{A}_{i,j} = \mathbf{Q}_{i,0j} - max(\mathbf{Q}_{i,1j}, \mathbf{Q}_{i,2j}, ..., \mathbf{Q}_{i,nj}) \tag{6}$$

$$Square : \mathbf{A}_{i,j} = \mathbf{Q}_{i,0j} - [max(\mathbf{Q}_{i,1j}, \mathbf{Q}_{i,2j}, ..., \mathbf{Q}_{i,nj})]^2 \tag{7}$$

$$Cube : \mathbf{A}_{i,j} = \mathbf{Q}_{i,0j} - [max(\mathbf{Q}_{i,1j}, \mathbf{Q}_{i,2j}, ..., \mathbf{Q}_{i,nj})]^3 \tag{8}$$

For example, using Max discount, consider the data *a-f* as shown in Table 10. Data element a is set as the dependent variable and data elements b and c as independent variables already in the model. For every unmodelled data element, elements *d-f*, subtract from the associated correlation value of a, the maximum value between the correlation values associated to the already modeled data elements b and c. Data

element e would be selected for the next modeled independent variable having the highest adjusted correlation value after discount. The Square and Cube discounts choose the next best variable based on adjusted correlation with the dependent variable after subtracting the respective squared maximum or cubed maximum absolute value correlation value of each currently modeled term from the dependent variable's correlation value, thus reducing the effect of the discount. Additionally, the adjusted correlation value of the candidate data element must have a positive value, $\mathbf{A}_{i,j} \geq 0$, or the algorithm stops adding variables to the model. All values in $\mathbf{Q}$ are absolute values, so a negative discount value would be an imaginary number and infeasible for consideration.

Table 10: Adjusted correlation using max discount

Unmodeled Variable	Correlation with a	Correlation with b	Correlation with c	Discount (Max)	Adjusted Correlation
d	0.8	0.1	0.7	0.7	0.1
e	0.5	0.2	0.1	0.2	0.3
f	0.4	0.1	0.2	0.2	0.2

Through the discounting, all data elements eliminate any second candidate variables having very high collinearity as seen in Table 11. The quantity of data elements potentially choosing a second candidate variable with high collinearity is noted under the column quantity of data elements. However, by the third selection of a candidate variable, all variables would be in the moderate category thus satiating any concerns about multicollinearity, but potentially increasing the risk of omitting key variables. Comparing the validation statistics between the degrees of discounting should identify where the balance may lie between too much collinearity and key variable omittance.

Now that the degree of multicollinearity is addressed, the implementation of preserving the stochastic element of imputation is modeled. Within the modeling segment, the stochastic noise values are saved. For the first iteration of the algorithm,

Table 11: Second variable correlation categories after discounts

	Number of Data Elements			Percent Including Category		
Correlation Category	Cube	Square	Max	Cube	Square	Max
High (0.9-0.7)	25	8	3	3%	1%	0%
Moderate (0.7-0.6)	82	41	7	9%	4%	1%
Moderate (0.6-0.5)	230	98	28	25%	11%	3%
Low (0.5-0.3)	902	244	153	97%	26%	16%

only known values (non-imputed) are used to select variables and produce coefficients. The residuals from this first iteration are saved and set aside to be used for all subsequent iterations as stochastic variation as well as in determining convergence.

The final segment, imputation, takes the regression coefficients from modeling, applies them to the related independent missing data values, and produces a point estimate. Randomly choosing a value from the normal distribution of residuals saved on the first iteration provides the uncertainty added to the estimate. The second modification to the original algorithm concerns setting limiting bounds for estimating imputations. The original regression models were unbounded and therefore could produce unreasonable estimates unlike a contrasting methodology such as pmm. Taking the *Battle-Related Deaths* data element as an example, the model regression coefficients could estimate some imputed observations with negative numbers. A negative death has no clear or rational interpretation, which implies that the data element should not allow for such values. Additionally, there are no known negative values in the original data distribution, which would cause further plausibility concerns if left unchecked. Therefore, the imputed estimates are assessed with consideration toward the known values within the data vector. This assessment acts like guard rails. There are three types of variable range-based guard rails implemented in the algorithm. First, if the minimum and maximum of the known data points are 0 and 100, it is assumed the data vector is a percentage and therefore all imputations are bounded

between 0 and 100. Second, if the known data points present no negative values, it is assumed the data must be positive only and bounded as such. Third, if the known data vector contains both positive and negative values, then the bound set is 1.5x the maximum and minimum known values. The wider range accounts for potential unobserved nonresponses outside the observed values in the model without allowing extreme extrapolation. Any imputed point estimates that are outside the bounds are set to the bound and then applied with applicable noise to stay within the bounds. However, some point estimates that are already within the bounds still may produce imputations outside the bounds when the stochastic element of noise is applied, therefore, the data vector is assessed a second time after the noise application to ensure all imputed values remain inside the bounds.

The final step in the imputation segment considers the stopping condition. Due to the importance to the process, the idea of convergence is expounded. One of the largest issues plaguing multiple imputation techniques manifests in knowing when enough iterations are complete. Defining convergence becomes even more of a nebulous term because of the stochastic nature of the algorithm accounting for the uncertainty of the imputed value. Stochastic convergence has four main definitions: observing a convergence in distribution, a convergence in probability, a convergence almost surely, and convergence in r-mean. Van Buuren notes that there is no clear-cut method for determining convergence in multiple imputation, however, the MICE package in R defines convergence as "when the variance between the different sequences is no larger than the variance with each individual sequence" [45]. A Python implementation of MICE in Iterative Imputer notes that their experimental algorithm could warrant more investigation into their convergence criteria (#14338) where certain datasets fail to converge and debate continues on what criteria to use against the tolerance parameter [66]. The Autoimpute documentation does not expound upon

stopping conditions and settles with simply stating that increasing the posterior sampling chains may improve the chance of convergence [67]. Nevertheless, an algorithm benefits from a stopping condition to assess the completion of the imputation outside a user defined value for iterations, which typically is convergence within a tolerance.

The noise aspect in stochastic regression adds uncertainty to the regression point estimate by exploiting the residuals in the known data points. Leveraging van Buren and the sentiments expressed in Iterative Imputer and Autoimpute, consecutive iterations of dependent variables within the distribution range of the residuals should satisfy a classification of convergence. The difference between the regression point estimate and its prior iteration estimate becomes the assessment for convergence. These estimates are prior to the addition of the stochastic noise. If every observation in the data element for the iteration has an absolute value difference less than the stopping criteria, then the data element is converged and no longer assessed for imputation. This leads toward the question of a good stopping criteria. In the three previously mentioned commercial programs, the stopping criteria is a user inputted tolerance. However, a user inputted tolerance does not account for the different scales that may be present in the large set of data elements. Capitalizing on the residuals used for the stochastic nature of the algorithm can assist in formulating tailored stopping tolerances for each data element. The return on a tailored tolerance manifests in observing the distribution of the first iteration residuals for each data vector. Observing the adjusted-R^2, experimenting with various standard deviation tolerances of the residuals, little improvement manifests in selecting a tighter than three standard deviation parameter for the stopping condition tolerance.

Acknowledging the initial presentation of the algorithm presented in [59], the modifications to the pseudocode are presented in Figure 9. The inclusion of the discount strategy is seen in step 3b with additional effects in 3c and 3h. Instead of

eliciting candidate variables from the **Q** matrix, candidate variables are selected from the **A** matrix, which is updated with every variable selection. The variable range-based guard rails are introduced in step 4c with a second variable range-based guard rail check in 4e.

1. Create **Q**, a matrix of absolute value Pearson correlation coefficients **r** of all **p** data vectors.
2. Rank all **p** data vectors in the dataset from least proportion of missingness to greatest proportion of missingness to identify the order in which imputation is processed. Data vectors with few missing elements are imputed first.
3. Create the stepwise regression models.
    a. Using the order from (2), select a data vector as the dependent variable requiring imputation.
    b. Create matrix **A** using the discount strategy.
    c. Add candidate data vector as independent variable based upon the maximum value in matrix **A**.
    d. Listwise delete all observations from the model that incorporate a missing value across all variables.
    e. If the number of observations is below the threshold, go to (3c) and choose the next best candidate data vector up to 10 possible attempts. Otherwise solve model and go to (4).
    f. Solve model.
    g. If the adjusted-R^2 fails to improve, go to (3c) and choose the next best candidate data vector up to 10 possible attempts. Otherwise solve model and go to (4).
    h. If there are less than 10 variables in the model, go to (3b) to update matrix **A** and select another candidate data vector.
    i. Save the model regression coefficients.
    j. If this is the first iteration model with no imputed values, save residuals to be used as noise.
4. Impute missing values in the dependent data vector.
    a. Restore all observations removed during (3d).
    b. Using the model coefficients from (3i) produce point estimate $\hat{y}$ for missing values in the dependent data vector.
    c. Using variable range-based guard rails, ensure all point estimates are plausible.
    d. Add model residual noise to the estimated $\hat{y}$, using a randomly selected residual from the first iteration model developed in (3j).
    e. Recheck guard rails to ensure all imputed estimates are plausible.
5. Assess the stopping rule for iterations against the convergence factor. If data vector has not converged, continue back to (3).

Figure 9: Multicollinearity Applied Stepwise Stochastic Imputation (MASS-impute)

Methodology Evaluation

Validation of the methodology continues with the same three metrics conducted in [59]: time evaluation illustrated by number of data element convergences, model fit calculated by adjusted-R^2, and prediction accuracy through the proxy of recreating known values through imputations and assessed under a normalized root mean square error (NRMSE). Due to different scales between the data elements, normalization of the error is necessary to make comparisons between data elements. The normalization used in this study leverages the original range of the data vector as illustrated in Equation 9, where x_{1ip} are the known values in the test set, $\hat{x}_{1ip}$ are the imputed values corresponding to x_{1ip} with N_{1p} test set observations, and x_{2p} are the known values in the original set, all for the p^{th} data element of P total elements. The test set randomly selected 8% of known observations to be recreated through imputation. Finally, since an instance of an imputed value is not unique, 30 imputed complete datasets are used for analysis.

$$NRMSE = \sum_{p=1}^{P} \frac{\sqrt{\sum_{i=1}^{N_{1p}} (\hat{x}_{1ip} - x_{1ip})^2 / N_{1p}}}{\max(x_{2p}) - \min(x_{2p})} \tag{9}$$

Randomly removing known values and checking the accuracy of the imputations against the known values provides an evaluation akin to MCAR. When data is MAR, other validation measures may be more appropriate. Van Buuren observed distributions and scatterplot values to observe if the estimates overlayed with known results appeared as if nothing had ever been missing when checking the plausibility of multiple imputation results [43]. For a set of imputation results, visual inspection via a scatterplot should present further evidence about the plausibility of the imputation, both in distribution and in position.

There are also a few statistical tests to evaluate the plausibility of results. Should

the known values follow a normal distribution, or the quantity of imputed values be sufficiently small, a parametric two-sample t-test would highlight inconsistencies in the means. The null hypothesis being that the known data and the imputed data are drawn from populations that share the same mean. If the p-value of the test is greater than some confidence level, then the difference in means appears insignificant and the perception is that the sample means are the same and assumed to come from similar distributions. However, if the distribution is unknown, the non-parametric Wilcoxon-Mann-Whitney (WMW) test also highlights inconsistencies between two independent groups, but with relation to medians. WMW test asserts that if the data values of two quantities x_n and y_m are ordered, the arrangement when counting how many times y precedes x, designated as U, is significant if $P(U \leq \overline{U})$ is under some confidence interval [68]. The null hypothesis states that the known data and the imputed data are drawn from populations that share the same median. These two inferential tests examine descriptive metrics of the imputations; therefore, a goodness-of-fit test is also examined. Two well-known goodness-of-fit tests are the Kolmogorov-Smirnov and the Anderson-Darling. A simple understanding of the two tests see the one-sample Kolmogorov-Smirnov test as a supremum proximity analysis of the empirical distribution function, and the one-sample Anderson-Darling test as an evaluation of how close the points are to a straight line estimated in a probability graphic [69]. The two-sample Anderson-Darling (AD) test is similar to the Kolmogorov-Smirnov in that it is a goodness-of-fit test, but is said to dominate Kolmogorov-Smirnov in observing smaller moments in the distribution [19] due to its sensitivities in the extreme ends of distributions [70]. For this research, the AD is used, consistent with other country conflict imputation research [19, 49]. The null hypothesis proposes that the known and imputed values are drawn from the same population without having to specify the distribution function of that population.

Python SciPy packages [71] provided ease of use implementations to generate p-values. The ttest_ind package set the assumed variance between the vectors as not equal. The mannwhitneyu and the anderson_ksamp packages used default values. This study used a significance level of 95%. Each imputed dataset was assessed against the known values in the data element vector to quantify how many imputation sets satisfied the test.

4.4 Model Results

The model results highlight the benefits of the methodology in three aspects: micro, macro, and comparative. The micro aspect looks at the application of variable range-based guard rails, a change in controlling the aperture of the results, with a focus on improving the methodology to the previous evolution. The macro aspect evaluates the application of discounting, a change in nominating variables for inclusion, with a focus on identifying the degree of multicollinearity hindrances and objectively selecting the optimal discount. The comparative aspect dives into the imputations themselves when the method is optimally configured to defend the plausibility of the method's results.

When researching categories of correlation, Nguyen highlighted independent variables with inadequate association with the dependent variable should be removed from the model [50]. Five data elements had a maximum correlation value below 0.5, which, using Nguyen's advice, would recommend no modeling variables for imputation. Of the five data elements containing only correlation values below 0.5, their percent missingness were 0.3%, 2.1%, 5.2%, 7.9% and 36.2%. Despite their correlation limitation, the three lowest missingness met the convergence criteria in all model-runs by at least iteration 8 and therefore should not be a cause for concern for instability. The two with higher missingness would often converge by iteration 5, although

3 of 10 exploratory model-runs saw non-convergence when allowed to run out to 100 iterations. Still, the data elements below the minimum threshold by Nguyen do not appear to unduly suffer regarding the validation metrics within this methodology and therefore it is likely that Nguyen's bottom threshold of 0.5 may be set too high.

4.4.1 Micro Aspect

The application of variable range-based guard rails provided many benefits to the models. As in [59], three regression model constructs were considered, linear (LR), nonlinear (NL), and nonlinear with first-order interactions (NFI). The LR model retained the fastest convergence rates compared against the NL and NFI models, and the 'with variable ranged-based guard rails (WGR) continued to improve all models compared to the original (Orig) models from [59] as illustrated in Figure 10. The overall time comparisons between the Orig models compared to the WGR models is less pronounced, although the NL-WGR model converged faster than the LR-Orig model. In practice, the LR-Orig model completed 20 iterations after 52.3 ± 0.4 minutes using an Intel i7-9700K with 64GB of RAM in Python 3.8.8, however, even with the inclusion of the variable range-based guard rails increasing the checks within the algorithm, increasing the complexity of the models with squared terms still enjoys similar completion times due to converging on earlier iterations. This is significant where each additional unconverged data element adds compounding time to the completion of an iteration where the NFI-Orig model finished after 3 hours 50 minutes for an average of 461 unconverged data elements. The algorithm with variable range-based guard rails included more statistical outputs, so a direct comparison is not practical, but even with the additional workload, the NFI-WGR finished significantly faster with an average of 2 hours 49 minutes. In fact, the NFI-WGR averaged only 33 unconverged data elements more than the LR-Orig model, converging sooner, and

therefore theoretically outputting faster than NL-Orig or NFI-Orig models.

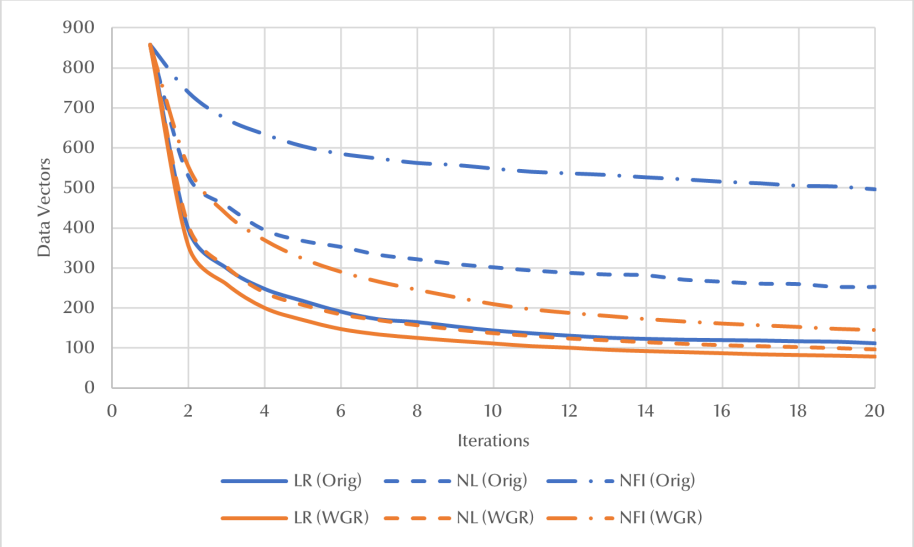

Figure 10: Model convergence rate of data vectors, N=10

The model fit continued to retain similar features with or without variable range-based guard rails. As seen in Figure 11, the second iteration saw a decrease in adjusted-R^2, which is an artifact of both a preliminary mean imputation for missing values in the independent variables for only iteration 1 as well as the iteration 2 models being constructed with more observations from the first round of imputations. As with the Orig models, this one-time mean imputation bias decreases as each round of imputations develops more plausible results and converges on a value within the range of noise. Although the WGR models do not rebound to the level of the Orig models, the measurement retains average values above 80% and demonstrate more stable results, especially when implementing the NFI model.

The main benefit of the variable range-based guard rails surfaces when cross-checking known values against imputed values. Although the median NRMSE values of the Orig models demonstrated low values, the LR-Orig model sum value was quite high due to four outliers and the NL-Orig and NFI-Orig sums were excessive due to compounding artifacts of outliers. When variable range-based guard rails are im-

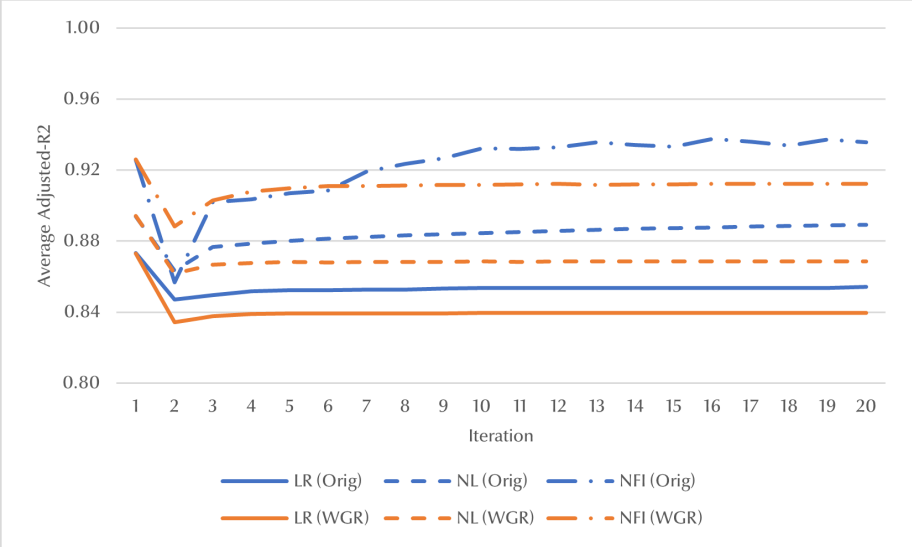

Figure 11: Model average adjusted-R^2, N=10

plemented, these outliers are severely reduced. The maximum data element NRMSE were 0.471 (LR-WGR), 0.477 (NL-WGR), and 0.417 (NFI-WGR) with variable range-based guard rails as opposed to values from the Orig models without variable range-based guard rails that ranged into the thousands. As with results from the Orig models, the distributions are still not normal as shown by averages of 0.054 (LR-WGR), 0.049 (NL-WGR) and 0.042 (NFI-WGR), and median values lower at 0.022, 0.019 and 0.013 respectively. This brings the NRMSE sums into reporting range: 50.120 ± 0.059 (LR-WGR), 48.797 ± 0.054 (NL-WGR), and 38.716 ± 0.040 (NFI-WGR). The immediate change from the implementations of the original algorithm without guard rails is that now the LR model has the worst NRMSE with the models incorporating increased complexity subsequently improving, as expected. This agrees with the hypothesis that many data elements contain curvilinear relationships within the variable as observed with some economic indicators as well as first-order interactions. Checking to ensure that bias is not a factor with either missingness or the rate of convergences, the indications appeared weak at best. The NRMSE of the data elements were contrasted against the number of missingness within the data

element producing an average correlation coefficient of 0.15. Administering a similar test against the iteration of convergence, the correlation coefficient was -0.25. If the data element did not converge, the iteration was designated as N=21, which is not necessarily true and may underestimate the correlation strength. With a negative correlation coefficient, it appears that data elements that converge later may benefit from a lower NRMSE. The maximum NRMSE always came from an iteration 2 data element and many outliers disappeared after iteration 7. A future modification to the algorithm may include pausing the stopping condition check until at least seven iterations have concluded to benefit from a closer threshold in reproducing known values with the imputations.

4.4.2 Macro Aspect

The benefit of variable range-based guard rails brought the imputations into a more plausible and defendable range of values. However, concerns of multicollinearity between independent variables used within the imputation models are still present despite the variable range-based guard rails. To dispel the concerns or minimize collinearity, collinearity discounts were applied to the variable selection process. Depending on the model, the discount had varying effects concerning convergence as illustrated in Figure 12. The LR model benefited from increased convergence rates with each degree of discounting. The standard error between runs was small regardless of model or discount combination averaging just over 1 data element. The NL models experienced an initial improvement toward convergence with the cubed degree of discounting, but subsequent degrees of discounting were statistically the same. NFI models saw the opposite effect. No discounting and the squared degree of discounting were statistically the same, while max discounting saw appreciable benefit in convergence. Although the discounting saw gains in convergence, it remains unclear if faster

convergence produces more plausible imputations as the other metrics would indicate more defensible results from the slower NFI model rather than the faster LR model.

The average adjusted-R^2 values tells a different story. The increasing degree of discounting for all models reduced the adjusted-R^2 as seen in Figure 13. The standard error for N=20 was extremely tight with a maximum of 0.0003, meaning each model-discount pair were statistically different. As with the comparison between with variable range-based guard rails and without guard rails, the difference in adjusted-R^2 is small. However, with all models showing similar trends and small standard error, multicollinearity does not appear to be as big of an influence as first feared. This isn't to say that high multicollinearity does not exist, but that it does not hinder the development of plausible imputations. When the dependent variable is highly correlated to at least one dependent variable, then the R^2 value should be high. The problem with multicollinearity surfaces when multiple dependent variables are highly correlated so that the coefficients cannot differentiate stable relationships to the dependent variable. In other words, there may be multiple solutions to the coefficients to exact the same value to the dependent variable. As mentioned previously, this is a problem of analysis, not necessarily a problem with result. Seeing how adjusted-R^2 penalizes adding independent variables of little value, high adjusted-R^2 compared between model discounts infers more defensible imputations.

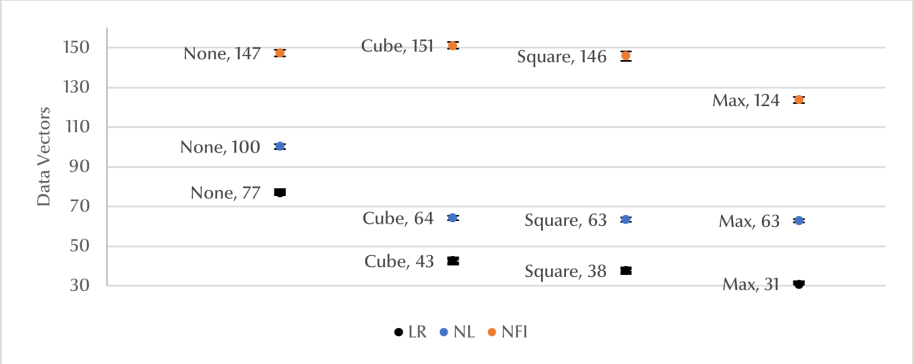

Figure 12: Remaining unconverged data elements, Iteration 20, N=20

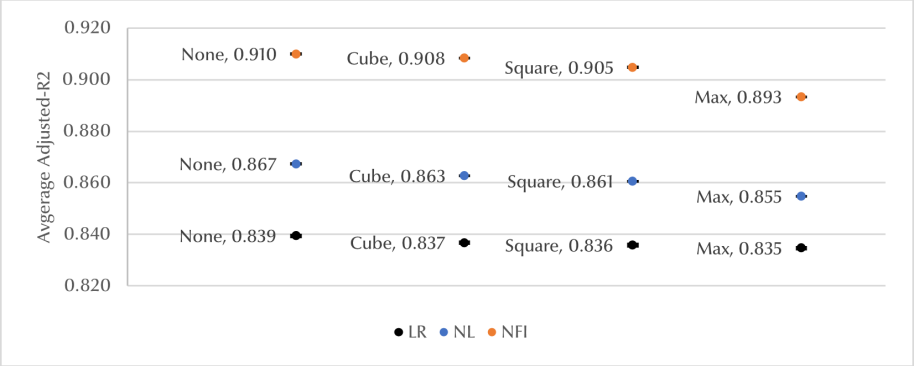

Figure 13: Discount model average adjusted-R^2, Iteration 20, N=20

Verifying the assumption that higher adjusted-R^2 leads to more defensible imputations is supported by the NRMSE metric. Lower deviation from the known value is better and the NRMSE results shown in Figure 14 confirm that the high adjusted-R^2 of NFI produces lower NRMSE than the other models. The translation for the LR and NL models is consistent with the adjusted-R^2 results, however, the NFI is less clear. Again, the standard error is tight signifying that all model-discount pairs are statistically different. The cube and square degree of discount for the NFI model produces imputations closer to their known values over using no discount. But discounting too heavily nominates independent variables that are too far removed from alternate variables that have higher correlation values with the dependent variable. Looking at each of the validation metrics supports a different model-discount approach. However, the NRMSE defense could be weighted the heaviest by explicitly connecting imputations to known values. With the NFI model in agreement between NRMSE and adjusted-R^2 concerning the best modeling approach, it can be concluded that multicollinearity does cause a degree of problem for generating the best imputations and that a cube discount for correlation selection is warranted.

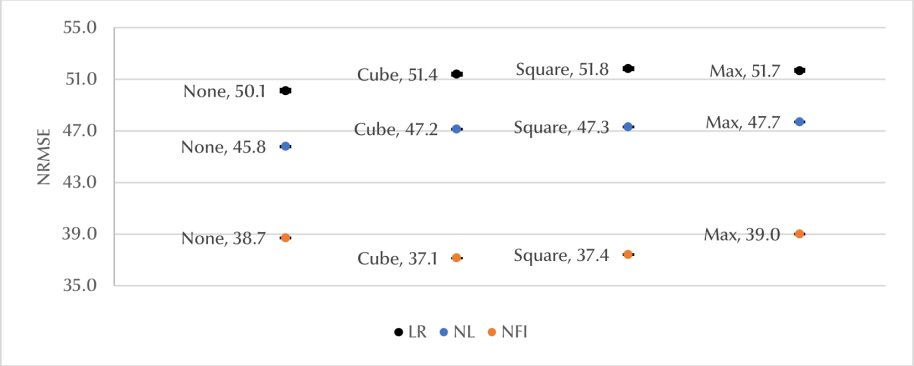

Figure 14: Discount model NRMSE, Iteration 20, N=20

4.4.3 Comparative Aspect

Thirty imputed complete datasets were generated for comparative testing by configuring the methodology to allow for first-order interaction while nominating variables with a cubed correlation discount following with variable range-based guard rails. Since it is computationally challenging to quickly test all data elements, this report compares only three data elements, one each from three different categories: low (<5%) missingness with quick convergence, high (>50%) missingness with quick convergence, and significant missingness (20%-50%) without convergence. The low missingness converged on iteration 2 requiring only 1 value for imputation. The high missingness converged on iteration 6 requiring 1437 imputed values, or 83.1% of the data element. The significant missingness required 526 imputed values, or 30.4%.

The scatterplots for the three datasets are provided in Figures 15-17. The blue points indicate the known data points, while the orange points indicate the imputed data for the selected variable across all 30 imputed datasets. In Figure 15, the one missing value in low missingness had an imputed value varying between 100 and 95.36. The missing value was in the year 2006, with the other nine years showing 100. One might assume that 2006 would also be 100, but the stochastic nature of the unknown allows for a chance of deviation. In Figure 16, the high missingness

tells a different story. Each column of orange data points shows up to 30 alternative values. At first sight, there may be questions about the plausibility, however, the descriptive statistic of standard deviation places the plausibility into perspective. For further analysis, the standard deviation across years was assessed for each country using the known data. When the same was accomplished for the imputed data, no country exceeded the maximum standard deviation of the known data, allowing the variability shown in the scatterplot. As for the non-converged data in Figure 17, the standard deviation analysis was an as straightforward, where all 30 datasets had a high maximum standard deviation. The maximum standard deviation of the known data was 2.33E14 whereas the imputed data ranged between 2.78E14 and 6.90E14. As a positive, it appears that the known data may see trends of increasing values over time as observations 1-173 are in year 2006 and subsequent ranges proceeding by year. The imputed values also demonstrate that potential movement. The takeaway from all three figures is that the imputed values appear to be within a reasonable distribution of the known data.

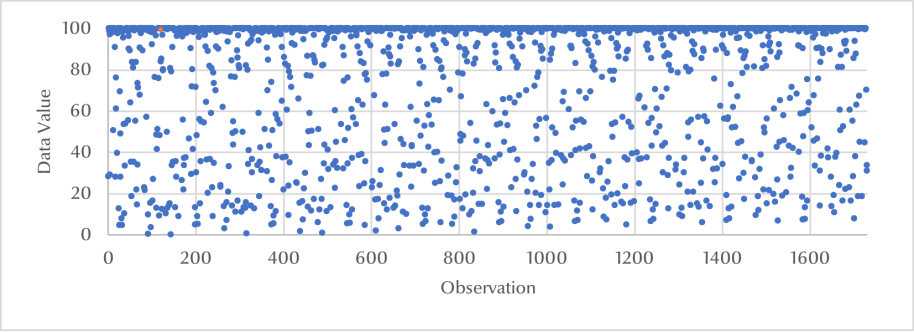

Figure 15: Converged data element, 0.1% missingness

Concerning the inferential tests, it was no surprise that all three tests showed no statistical significance when comparing the distribution of the 30 generated datasets to each other in the low missingness scenario. The imputed values were all within the range of known values and it was unlikely that one data point would signifi-

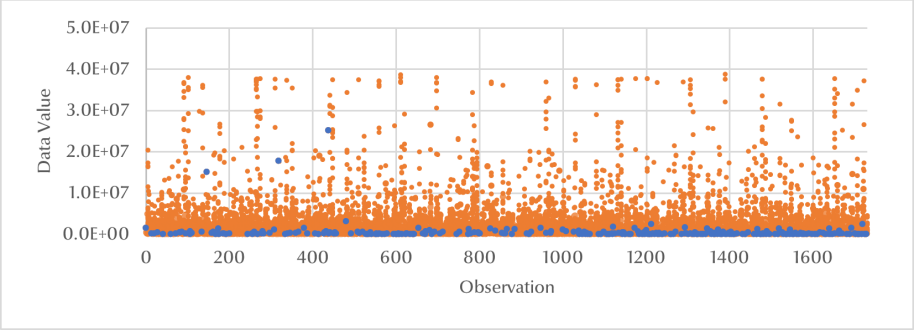

Figure 16: Converged data element, 83% missingness

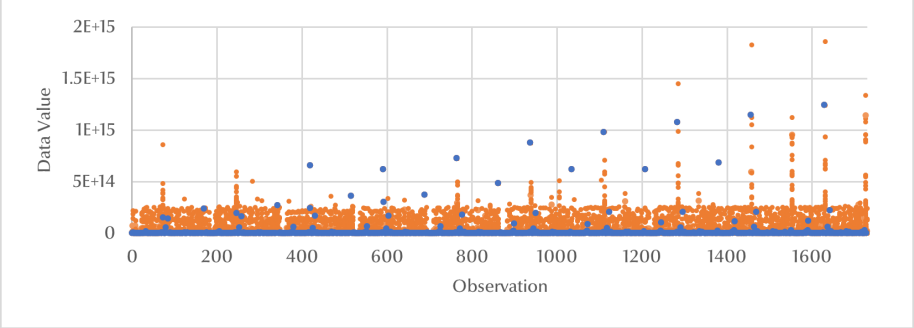

Figure 17: Non-converged data element, 30% missingness

cantly skew the mean, median, or distribution shape. The high missingness example saw a significant difference in mean for 16 of the 30 datasets. Furthermore, all 30 datasets saw p-values below 0.05 signifying statistical differences in the median and distribution shape. Despite these results, the consideration of high missingness and the MAR assumption could still find the results plausible. The imputed values could be categorically from samples that either are adverse from measuring or are difficult to measure, as expressed in the earlier examples of the *Corruption Perception Index* or the Democratic Republic of Korea. Similar findings were observed in the non-converged example; 6 datasets demonstrating statistical differences in mean and all 30 datasets demonstrating statistical differences with the WMW and AD tests.

4.5 Summary

The original Large Dataset Imputation through Correlation-based Regression method [59] demonstrated much promise through a multiple imputation stepwise correlation approach. It provided a balance between the analyst's trade-off of time, computational power, and accuracy. Two main concerns of this original approach revolved around multicollinearity and the potential for extreme outlier values. This paper alleviates both of those concerns through exploring a full range of discounts to the variable nomination process and bounding imputation estimates within a variable ranged-based guard rail process. Both processes strengthened the plausibility and defensibility of the imputed results.

Multicollinearity is a problem of analysis in determining coefficients for cause and effect, rather than a bias in output. The None discount demonstrated superior results in the LR and NL models. Only when a small degree of discounting was applied to the NFI model did any perceived effect of collinearity surface resulting in the Cube discount being superior for the dataset considered. However, specifying the appropriate model type, from LR to NL to NFI, demonstrated greater gains than the effects of discounting collinearity.

To further enhance the prior approach, variable range-based guard rails were developed that bounded the imputations into a plausible range and deterred subsequent iterations within the algorithm to exacerbate outliers. In hindsight, it aligns with the superiority of ppm on small datasets where imputations are likewise bounded to values already seen in the dataset. Unlike ppm, the variable ranged-based guard rails allow values that are probable in the distribution yet not observed, widening the aperture for plausible values.

Providing three aspects of analysis assisted in quantifying progress while increasing the defensibility of the method. The micro aspect analysis highlighted the im-

provements in convergence rates of individual data elements compared to [59] while maintaining strong goodness of fit. The macro aspect analysis quantified how little collinearity effects hinder the imputation through the adjusted-R^2 results demonstrating decreasing values with discounting and all but the interactions modeling showing lower NRMSE without discounting, dispelling concerns over using a correlation-based selection process. The comparative aspect analysis visualized the imputations to the known value distributions for a qualitative approach to plausibility. The inferential tests conducted alongside the visual assessment and descriptive statistics demonstrated opposing theories on plausibility, which cautions analysts from relying on a single metric when evaluating imputations. When working with MAR and NMAR data, an expert in the data is necessary for more conclusive analysis.

Outside of overcoming numerical problems in generating imputations, the improved approach also provided insight into the rate of convergence. Rather than providing a user-specified static tolerance for a stopping condition, the approach relied on the data itself to generate tailored data element tolerances by exploiting the residuals in modeling the known data. The concept leans on the definition of stochastic convergence of the r-th order mean where the difference of successive iterations is statistically zero. Using the distribution of the residuals captured in the first iteration of only known values, which were also used for noise, the algorithm conducts a check between iteration N and N+1 to measure the difference between estimates. Should the difference be within 3 standard deviations of the distribution of residuals convergence is assumed and the stopping condition applied.

Although the MASS-impute algorithm improved the original correlation-based approach, there are still areas that require further refinement. It was noted that the worst NRMSE values were captured during the first iterations, so further modifications to the algorithm may investigate not allowing stopping conditions until after

a set number of iterations. Such changes would increase the processing time of the algorithm, but at the potential benefit of improved accuracy. The investigation would illuminate the trade space between these two analytical trade-offs for balancing out the algorithm's parameters. Additionally, as seen in the comparative analysis, some of the high standard deviations in the imputations continue to be a concern. Known outliers in some datasets may be allowing too much variability in the noise element of the algorithm. The high missingness scatterplot showed three known values that would pull at the regression line used to generate the noise residuals. These outliers could potentially be adding too much variability to the stochastic nature of the estimates, especially when the outliers are more prevalent as in the non-converged example. Future modifications may investigate better accounting for these outliers when producing the pool of noise.

Using MASS-impute, the multiple imputations appear plausible while dispelling concerns about variable selection based on correlation. As with the finding in [59], the evolution of the methodology continues to balance computation time, power and accuracy in achieving traceable, defensible imputations for large datasets, including those that may exhibit over 20% missingness for some variables.

V. A Hierarchical Cluster Approach Toward Understanding the Regional Variable in Country Conflict Modeling

5.1 Abstract

This paper examines the regional variable common in country conflict modeling, specifically how to group countries into regions, by considering the country's regional proximity and data similarity effect on conflict prediction. Two key components contribute toward identifying a 7-region model that demonstrates high training accuracy with competitive validation accuracy using logistic regression. First, the application of feature extraction, which produces a manageable number of independent variables from the 932 political, economic, and social indicators. Second, the utilization of hierarchical clustering to develop insights into constructing contiguous regions using logistic stepwise regression as a goodness metric. While the study proposes one primary solution to the region construction question, it also identifies issues that will require further research and refinement.

5.2 Introduction

War is a messy business. Not only does war pay a cost in current lives, but it impacts future lives, fortunes, and honor (prestige). Even though the 1940s event in Germany occurred over seventy years ago, people continue to have mental anguish concerning the religious genocide of the Holocaust. In Japan, survivors of Nagasaki continue to face increased cases of cancer, especially leukemia, well past the initial loss of homes and family. The Iraq and Kuwait conflict saw oil resources razed lest the enemy control them, regardless of the economic impact to the world. Today, political conflict in Yemen stunts development as factions vie for official government legitimacy. Yes, war claims more than lives; it seeps into every aspect of living.

It is no wonder that from the highest levels of power to the lowest trenches of poverty, researchers seek and strive to understand the constructs that perpetuate the flames of war – much time, resources, and research drive modeling country conflict and peace. The irony, however, is that research often takes a narrow view of conflict to assume it is about the distribution of economic resources and the game theory of information [72]. Yet, country conflict has always been more complex than that – it is a product that incorporates both political, economic, and social aspects. While investigating significant variables toward predicting country conflict, five proxies continue to surface: Polity through regime types, gross domestic product (GDP) per capita, conflict history, population size and regions. However, many non-government organizations expend significant time and funding resources in developing data on specific datasets. All the variables except regional groupings trace to an open-source database. Regions, however, are often qualitative in their construct while at the same time showing integral toward increasing prediction accuracy [16, 18, 20]. Although prior research categorizes countries into regions, there remains a gap to uncover what drives this region proxy and why it is so important. One hypothesis states that regions represent a complex mixture of variables that produce a common culture, driving how other variables influence country instability. In other words, the region proxy sets the level of coefficients for all other proxies in a robust country conflict prediction model. The task then is to develop these regions to maximize the predictive influence of other independent variables.

This research considers far more variables than previously considered in the literature to develop a whole of culture concept while also forming regions to better model country conflict, cultural boundaries. Most notably, it investigates the optimal number of regions to consider within modeling and where to delineate the geographic boundaries for each region, while also considering data similarity.

5.3 Literature Review

Multiple country conflict researchers demonstrate the benefits of a region component toward modeling predictions. Over a decade ago, Goldstone noted that different regions facilitate different propensities for instability and therefore used regions as a control for building the modeling dataset [9]. His research explicitly noted five regions with different propensities for instability and made efforts to account for similar "regional and temporal distributions" in both the control and problem datasets [9]. Although the modeling approach was global, a single model to predict "all of the onsets of instability that occurred worldwide" for a given time period, the results concluded regional differences with striking results showing the Africa and East Asia region having higher risk of instability onset within a five year prediction [9]. An interesting contribution from the research focused on modeling conflict in a single region, their specific case study being sub-Saharan Africa. It was noted that by modeling by region rather than globally, model accuracy increased. However, regions for every country were not addressed.

Shortly thereafter, researcher Hegre demonstrated a modeling approach that included regions as predictor variables [16]. Instead of the five regions annotated by Goldstone, Hegre defined nine regions revised from the United Nation's regional definitions. He posited that the region variable improves the quality of predictions by maximizing the explained variance in the dataset, but questioned the duration of this assistance for distant forecasts [16]. The basis for the claim revolves around how long the heterogeneity of the regions may remain and surmises that prediction benefits may degrade after a decade [16].

A third example of regional modeling surfaced with the Boekestein logistic regression study, where his study investigated five different categories of a regional variable [18]. The study concluded that a six-region categorization presented the best mod-

eling accuracy for the modeling employed, specifically a categorization inspired by a 2006 talk presented by statistician Hans Rosling. Rosling's presentation dissected a six-region categorization asserting that semi-geographical aggregation of data hides the diversity of country-level and even within-country level data [73]. His examples, such as population versus fertility rates, or child survival versus GDP, foster conclusions that social changes precede economic changes while economies trend toward homogeneity. Despite the theme that inter-national culture may be too diverse to conclude national culture (discriminant properties ranging between societal and economic variables), other studies using hierarchical clustering techniques refute any claim that national culture cannot be a worthwhile analysis unit [74]. Notably, missing from Rosling's presentation was rationale for the categorization of the regions. Despite the lack of rationale for the categorizations, Boekestein's use of region as a variable assisted in reducing both false negatives and false positives within a global model. Furthermore, when treating each region as its own model with tailored classification cut-off parameters at 0.28, model accuracies increase by at least 5% [18].

Other works have improved upon Boekestein's research while maintaining the consistency of using the same 6 distinct regions for modeling [20, 22]. Shallcross incorporated a dependent variable, dividing the modeling dataset into in-conflict and not-in-conflict Markov states, focused on the transitional state of conflict rather than the current year's static state, further improving prediction results [22]. Later, Neumann sought to find further improvements by reevaluating region categories using both the transitional dependent variable from Shallcross and her new modified k-means approach for clustering countries [21]. This capitalized on Hegre's idea that heterogeneity of the regions may change over time. Using a modified k-means algorithm, Neumann improved prediction accuracies by as much as 2.5% through redefining 6 United States Combatant Command regions using a combination of political,

military, economic, and social variables [21]. Her combination of 30 diverse variables transformed into 9 principal components (PCs) alludes to the idea of a cultural association between countries. Previous studies have shown support for cultural clusters as a combination of religion, language, geography, ethnicity, and economics, among other factors [74, 75]. Gupta's study classified 10 distinct clusters through discriminant analysis indicating shared societal goals or values between countries, culminating toward the conclusion that regions are a relevant unit of analysis and a reliable study indicator [75].

Unresolved is a consensus on the number of cultural clusters, or regions, and how they should be formed. Neumann assumed 6 clusters using a mathematical approach based on k-means clustering that maintains consistency with the current number of U.S.-defined geographic commands. However, concerning the Gupta study, his mathematical approach using discriminant analysis concluded that more distinct clusters may exist. Another study recognized the inconsistency of published reports toward identifying the number of distinct cultural clusters, which varied from as little as 6 towards as many as 18 clusters, and applied a hierarchical mathematical approach settling on 11 global clusters [76]. Although these studies apply mathematical approaches to defend their conclusions, they were limited in how many culture-defining variables they considered. Neumann's study presented the most culture-defining variables, considering up to 30 variables. This study greatly increases the culture-defining variables considered, and thus the complexity, by considering 932 possible variables.

Capitalizing on the increased availability of possible variables, this study seeks to address assumptions feeding prior work. First, does the increase in variables considered assist in producing better country conflict prediction regions? Second, what number of regions produce the best country conflict prediction models? And third, what country regional groupings produce superior country conflict prediction fore-

casts?

5.4 Methodology

The dataset contains variables on 173 United Nations' member countries whose population total exceeds 250K as of 2016. The Political, Military, Economic, Social, and Information (PMESI) Database, which is the Air Force Institute of Technology's repository of several open-source databases, provided the 932 independent variables. Any variables missing values from their open-source databases were imputed using multiple imputation.

Feature extraction and clustering techniques require complete-case data, so observations with missing values must be discarded or estimated. List deletion of missing values seriously degrades the ability to detect effects of interest as various statistical estimates would be severely biased [43]. The alternative, considered to be the method of choice for addressing country conflict missing values, multiple imputation, estimates a plausible value that is statistically valid for the missing data [44]. This study used MASS-impute, a type of multiple imputation, to complete the dataset [40]. Originally, 30 datasets were created to account for the stochastic nature of imputation but due to the computational complexity of the method's algorithm, only one dataset was explored. However, the preliminary exploration of parameters for the number of PCs used all 30 datasets.

Consistent with Neumann, the methodology follows a process transforming the variables into PCs before running a clustering algorithm. Also, as with Neumann, the last period of observation generates the clusters, in this case, the year 2015. Once the countries are clustered into new regions, each region is modeled independently through logistic regression to predict each country's conflict state. Within the methodology, there are four types of control parameters: 2 types of dependent

variables, 7 quantities of PCs, and up to 10 possible clusters with and without geographic connections. This totals 1,925 different regional logistic regression models. Furthermore, this study applies an automated stepwise logistic regression approach to develop a goodness metric based upon the accuracy of the best resultant found. This approach also expedites the modeling building process in comparison to the 7-step purposeful selection of covariates approach found in [36], as used by both Shallcross and Neumann, whom built only 24 models.

The observation period consists of 10 years, employing 2006-2012 as a training set and 2013-2015 as a three-year validation set. The logistic regression modeling assesses two variants of dependent variables, both of which are derived from the Heidelberg Institute for International Conflict Research (HIIK). HIIK maps a highest level of conflict intensity score to each country according to a conflict means and conflict consequences approach [23]. One of the dependent variable variants, static-state, borrows from Boekestein [18] where HIIK intensity levels 0-2 are coded as not-in-conflict and levels 3-5 are coded as in-conflict. The other variant, transition-state, borrows from Shallcross [22] and Neumann [21] where the Boekestein static-states transition its conflict state given the nation's previous year conflict status. Nations that transition into or remain not-in-conflict are coded as not-in-conflict, while nations that transition into or remain in-conflict are coded as in-conflict [22]. An overview of the new methodology is in Figure 18.

5.4.1 Dimension Reduction

By increasing the number of variables, challenges arise concerning applying clustering techniques. Kriegel investigated clustering high-dimensional data and imparted four key considerations [77]. The four key issues when employing clustering techniques are typically referred to as the curse of dimensionality. The first issue revolves around

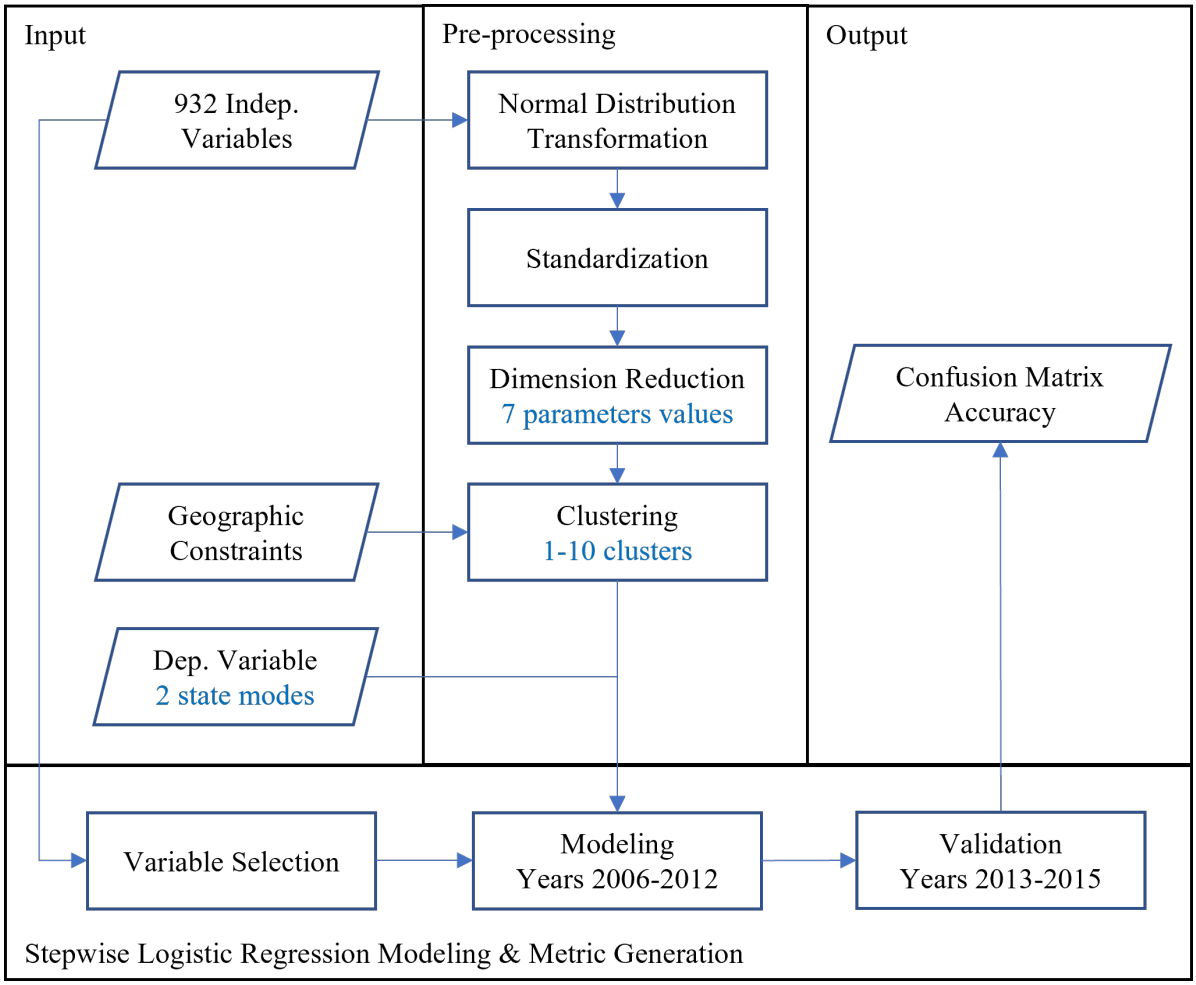

Figure 18: Overview of methodology

the ratio of data elements (p) to observations (n), the general principle that when $p > n$, there aren't enough simultaneous equations to solve for a solution. Kriegel noted that clustering enables "users to identify the functional dependencies resulting in the dataset", but as more variables are added, the complexity of the relationships increase making it difficult to visualize interesting insights [77]. The second issue states that as more variables are considered, the idea of proximity or distance becomes less meaningful because of increasing dimensionality; "the distance of the farthest point and the nearest point converge to 0" [77]. The third issue considers the difference between global and local subspaces, where variables are more likely to be

irrelevant in certain subspaces, in turn, increasing the amount of noise at the global level [77]. The fourth issue dives into the redundancy of variables, thus artificially weighting distances, from a correlation perspective [77]. The advice to overcome all four issues remains the same though: narrow variable selection below 10-15 variables. Beyer demonstrated that using more than 15 dimensions produces meaningless results [78]. The Beyer study focused on distance measures within clustering algorithms, showing that this multi-dimensional upper bound is agnostic to distance type if the clustering method used employs distance as a metric. The premise is "that the minimum and maximum distances from the query point to points in the dataset become closer and closer as dimensionality increases" [78]. Through simulation, the dataset size and the data distribution remained consistent showing that the primary restrictor is dimensionality, and that the inflection point is between 10 and 20 dimensions [78].

There are two overarching mechanisms toward reducing dimensions in a dataset: feature selection and feature extraction. Feature selection selects and only uses the most relevant variables in the dataset. However, this study dramatically increases the number of variables for consideration, therefore, using feature selection would disregard a core study motivation with ignoring the influences of over 900 additional variables. On the other hand, feature extraction reduces the number of dimensions by considering all 932 variables, creating a small subset of new variables as linear combinations of the original variables. With the aim to retain as much of the original information captured while reducing the overall dimensions of the dataset, feature extraction is preferred and used for this study.

For the clustering portion of the study, there is no dependent variable or current meaningful label, so unsupervised approaches as opposed to supervised approaches, like discriminant analysis, facilitate feature extraction. Principal component analysis (PCA) and factor analysis (FA) cover the two primary unsupervised approaches.

PCA seeks to solve the optimization problem of developing linear combinations of all variables subject to loading scalars that sum to one, while accounting for variance [79]. Meanwhile, FA models the correlation structure of all variables to illuminate rotatable latent variables with associated factor loadings [80].

PCA assumes that the dataset is multivariate normal and has been standardized so scaling is not a factor. FA assumes the dataset has no outliers, multicollinearity is manageable, and there is no homoscedasticity between variables. Management of the assumptions were dealt with through various measures such as Box-Cox normal distribution transformations, Min-Max standardization scaling, and exploring the removal of variables with high pair-wise correlation. Feature extractions seeks to reduce the number of variables to some $m < p$, where p would be the full 932 variables and m being the number of newly created variables that explain most of the information. Due to FA having multiple solutions because of its rotatability, PCA is preferred for this study. For PCA, there are p number of PCs, but m number of PCs explaining the interesting information (information with limited amounts of white noise) through representing much of the variation in the data [79]. There is no ideal solution to identify the optimal number of PCs, but there are a battery of methods from which to form a consensus, or at least a plausible range [81].

For this study, the following tests influenced the number of PCs retained: the combined assessment of the percent variance explained, the broken-stick model, the Jolliffe modification to the Guttman-Kaiser rule, and the log-eigenvalue diagram. For PCA, the ratio of each eigenvalue to the sum of all eigenvalues captures the variance explained in the model. The goal contends to use as few PCs as possible to explain the variance in the dataset. Typically, a predetermined ratio of 90% total explained variance is sought after, but for data with more white noise, the threshold can be lower. Cangelosi notes that in practice, common thresholds are between 70% to 95%

[81]. The broken-stick model, presented by MacArthur during a bird study, compares eigenvalues against an apportioned resource distribution [81]. The broken-stick distribution follows Equation 10, where p is the number of partitions and j is subinterval for the corresponding k-th element component. The element components are compared to the eigenvalue loadings, retaining the number of components that have a greater value than the broken-stick elements. The Guttman-Kaiser rule simply states that interesting components have eigenvalues obtained from the correlation matrix exceeding unity. In practice, the rule may be too conservative, so Jolliffe's modifications lowers the threshold to 0.7. Finally, the log-eigenvalue diagram, which is a modification of the scree plot, plots the log of eigenvalues against the number of components. This modified way of looking at eigenvalues can clarify some of the subjectivity inherent in the scree plot. The log-eigenvalue diagram displays the eigenvalue such that the smaller values will eventually form a geometric line, identifying those components that are conjectured to be noise [81].

$$E_k = \frac{1}{p} \sum_{j=k}^{p} \frac{1}{j} \tag{10}$$

5.4.2 Clustering and Geography

Two objectives motivate developing regions. The first objective seeks to apply mathematical rigor to the prediction models where studies [18, 20] demonstrate that grouping countries provide higher prediction results over just one global model. The second objective seeks to apply practical rigor to the models where political, economic, or military application may only be useful for countries that are contiguous.

Neumann studied the dichotomy of the objectives through her modified k-means approach. Her algorithm weighted the distance formula in k-means clustering between the Euclidean distance of the first two PCs and the Euclidean distance of each

country's center of power (the capital city) [21]. K-means finds a local optima influenced by the initial assignment of countries to clusters. Which also infers that there is no consistency between observing countries within a 6-cluster solution and a 7-cluster solution as the within cluster variation versus the without cluster variation is influenced by by first assuming k. There are two factors in the Neumann study that this research challenges.

The first factor is that the modified k-means approach does not always produce contiguous regions. In her final groupings, Morocco and Libya are attached to Combatant Command (COCOM) 1 with Algeria from COCOM 2 separating their contiguousness. Additionally, Tunisia and Albania are attached to COCOM 2 with Italy from COCOM 3 separating their contiguousness. These anomalies in contiguous regions arise from, practically speaking, developing two separate models, and finding a compromise between them. K-means develops clusters only by observing the dimensional likeness within the dataset. The modifed approach presents a solution to combine a geographic constraint, but it is still a compromise between the data solution and the geography solution.

The second factor addresses the contiguousness from a different aspect – the geographic constraint has not been defined and therefore left to the modeler to approach a solution. Neumann used a Great Circle distance between country capitals [21]. Where this may be a valid approach, distance biases may occur when the capitals are not centrally located within the country. For example, Russia borders 14 countries, but Moscow is 3,200 miles closer to Minsk, Belarus than Beijing, China, where both countries border Russia. It is uncertain if an assumption of centralized centers of power factored into the weights between the mathematical rigor and the practical rigor of the Neumann study. This research proposes that using country borders overcomes center of power assumptions when considering contiguous regions. To assist in

capturing many of the island nations, a country is considered bordering if the country pair's borders are within 100km of each other. For island nations further than 100km from any other country, the next closest country is considered bordering. One exception is made to the border matrix; the border connection between Russia and the United States is severed to assist in keeping North America and Asia as separate geographic regions. This exception assists leaders in setting policy and strategy as the Atlantic and Pacific Oceans present natural lines of demarcation.

Hierarchical clustering accommodates these two new factors innately, making it preferable over developing another modified k-means approach. Unlike k-means, where observations are randomly assigned one of predefined k-number of clusters with the algorithm reassigning observations to clusters by minimizing the within cluster variable (or PCs) Euclidean distance variation, hierarchical clustering starts with each observation as its own cluster and then combines 'like clusters' or 'two least dissimilar pairs' together until only one cluster exists. An output of this process is a tree-like diagram called a dendrogram. A k-number of clusters can be obtained from hierarchical clustering by stopping the algorithm prematurely. 'Like cluster' observations are defined as the two cluster observations that share the least distance when calculating the Euclidean distance difference of their associated variables (or PCs). To accommodate the geographic constraint, the algorithm considers a connection parameter, which only assesses the Euclidean distance difference for observations that have valid connection points.

5.4.3 Model Building and Comparison

Referencing Figure 1, independent variables may undergo transformations to meet assumptions for PCA. A Box-Cox transformation assists in transforming the variables to appear as close to a normal distribution as the data allows. The data is then

standardized using a min-max approach placing all values between the range of 0 to 1. Once the data meets the assumptions of standardized, multivariate normal, PCA is applied to create the specified number of PCs that are used for the dimensions establishing clusters. Agglomerative hierarchical clustering, using a ward linkage, builds a tree to identify which countries belong in which regions. The clustering is completed using both no additional connectivity constraints as well as using a country border matrix connectivity constraint.

Once the countries are identified by region, individualized regional models are created through a stepwise logistic regression method. For the transition-state dependent variable, two models are developed for each region: given in-conflict static-state and given not-in-conflict static-state. Selection of variables come from the pre-transformed datasets. Unlike linear regression, logistic regression does not have a model-fit measure such as adjusted-R^2 to assess variable selection. One pseudo-R^2 method that doesn't use maximizing the likelihood function, which coincidentally is also what logistic regression uses to develop model coefficients, is the Tjur statistic [82]. Tjur saw similarities between graphically comparing differences in two "parallel histograms" and the graphical check of the Hosmer-Lemeshow test [83]. This led to Tjur developing the coefficient of discrimination, D, which characterizes "a good model" of high explanatory power that predicts a high percentage of true positives and true negatives [83]. The Tjur statistic, as seen in Equation 11, identifies statistically significant variables for the models, where $\hat{\pi}_{i1}$ and $\hat{\pi}_{j0}$ denote the fitted values for successes and failures, respectively, of N true successes and M true failures, for the binary outcomes of logistic regression.

$$D = \frac{\sum_{i=1}^{N} \hat{\pi}_{i1}}{N} - \frac{\sum_{j=1}^{M} \hat{\pi}_{j0}}{M} \tag{11}$$

Accuracy from the confusion matrix quantifies the predictive power of the mod-

els. A weighted and unweighted (average) accuracy score provides insight into the analysis. The weighted score uses the number of observations per region to provide perspective into how many country-year pair observations predict accurately, whereas the unweighted score averages the accuracy of all regional models for the specified modeling parameters.

5.5 Results

Predictive accuracy remains the core focus in assessing models for country conflict. Focusing on just the dependent variable, the naïve approach assumes that transitions into or out of conflict are rare occurrences ("black swans") presenting an assumption that countries will remain in their current state for the next three years. Therefore, anchoring on the last year in the training set (year 2012), the following three years of naïve predictions would be accurate 87.3%, 85.0%, and 86.1% for a cumulative average of 86.1%. Considering the 932 independent variables through the stepwise logistic regression modeling approach, some global predictions using either 6 or 7 clusters achieved similar results. A global prediction averages all regional predictions given the number of worldwide clusters and dependent variable states. One global prediction may incorporate a single cluster; therefore, the global prediction and a 1-cluster regional prediction would be the same. However, another global prediction may incorporate 6 clusters; therefore, the global prediction would be the average of 6 regional predictions. If the global model uses the transition-state dependent variable, 12 regional predictions aggregate for the global prediction, as each region would have a prediction given a not-in-conflict static-state model and given an in-conflict static-state model. At the regional level of modeling, 296 of the 1,925 regional models surpassed the naïve global baseline. However, it is noted that prior research had lower goal thresholds – a goal to be above 80% [9, 18, 20, 22], which this research

achieved in the majority of models.

5.5.1 Pre-processing Results

A Box-Cox transformation was applied to each variable to optimize normality of the data. The lambdas of the transformation ranged between 16 and -18, where the mean and median lambda were 0.45 and 0.18, respectively. Although some of the transformations required large lambdas, over 20% of the variables were within 0.5 of a linear transformation, or basically no transformation required at all to assume normal.

PCA demonstrated superiority over FA for the dataset. After optimizing the normality of the data through Box-Cox transformations and standardizing the data, the explained variance after 15 variables for PCA was 71.3%, whereas FA was only 54.8%. The first principal component explained 36.6% of the variance, whereas the first latent variable of FA only explained 18.7% of the variance. Due to more information being retained in the reduced dimensions of PCA, the study used the PCA technique for the remainder of the study.

Observing the tests to determine the number of PCs to keep, the range varied between 6 and 32 components. The two statistical methods producing the maximum and minimum range of PCs for consideration were the broken-sticks model and Jolliffe's method, retaining 32 (range between 30-32) and 6 components, respectively. Cangelosi noted that his research observed that the broken-stick method consistently retained the fewest number of components compared to other techniques [81], yet in this research, the broken-stick method retained the most PCs. This is most likely due to a much larger number of variables in the original dataset compared to Cangelosi, where examples by Cangelosi were much smaller on the scale of 10s rather than 100s considered here. Still, 32 out of 932 components is a 96.6% reduction in

dimensions, which is a better reduction than Cangelosi demonstrated in his study. Figure 19 illustrates that although 32 components (black lines) statistically quantify the threshold (red line, broken-stick distribution), graphically, it could be argued that little is gained by retaining more than 16 components, with how close the distribution lines are to each other. The Jolliffe method result of 6 PCs remained consistent across all 30 datasets and presented the minimum number of components to retain. Ironically, this is also contrary to reports that the Jolliffe method in practice errs on retaining too many components [81]. Again, the recommendations were made on much smaller dimension sizes with the example examining only 9 variables [81] compared to our over 900 variables.

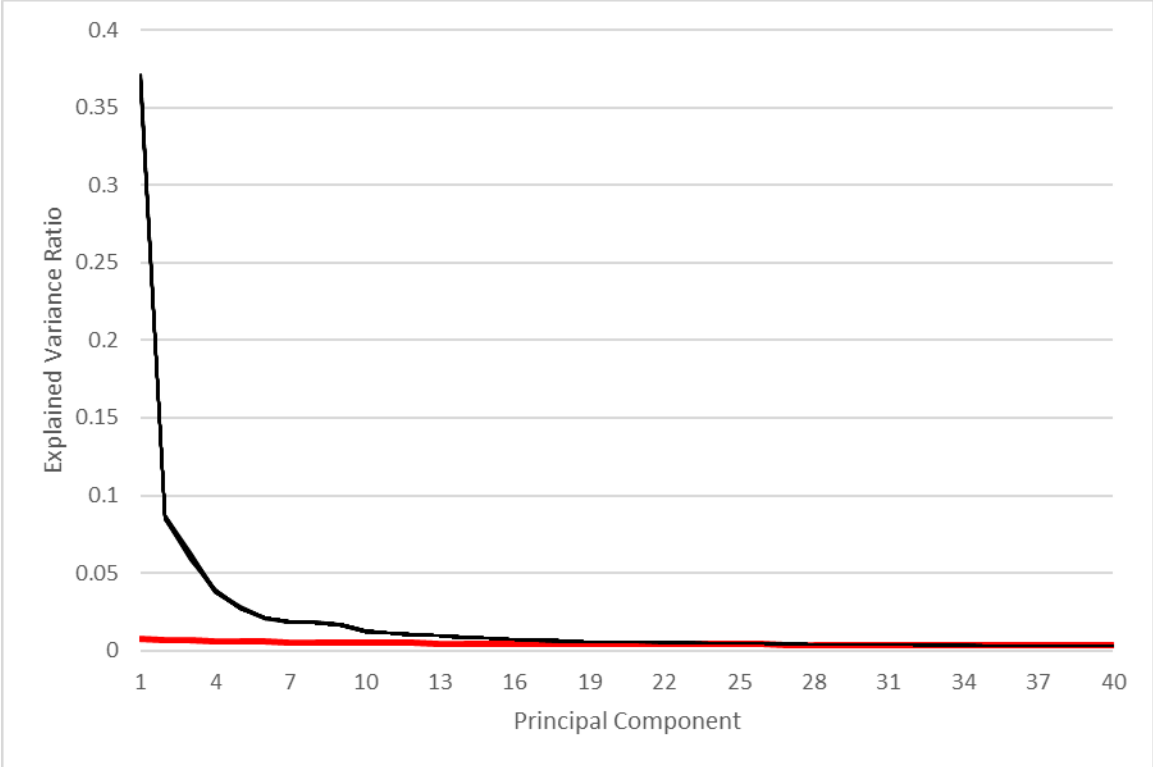

Figure 19: Broken-stick model

The log-eigenvalue diagram, as illustrated in Figure 20, presents a subjective interpretation of how many components should be retained. The log theory conjectures

that noise decays geometrically, meaning the graphical representation of noise in the data should manifest as a straight line as shown in red. Taken strictly, the graph demonstrates a maximum of 18, but taken less strict, a minimum of 10 components could possibly suffice.

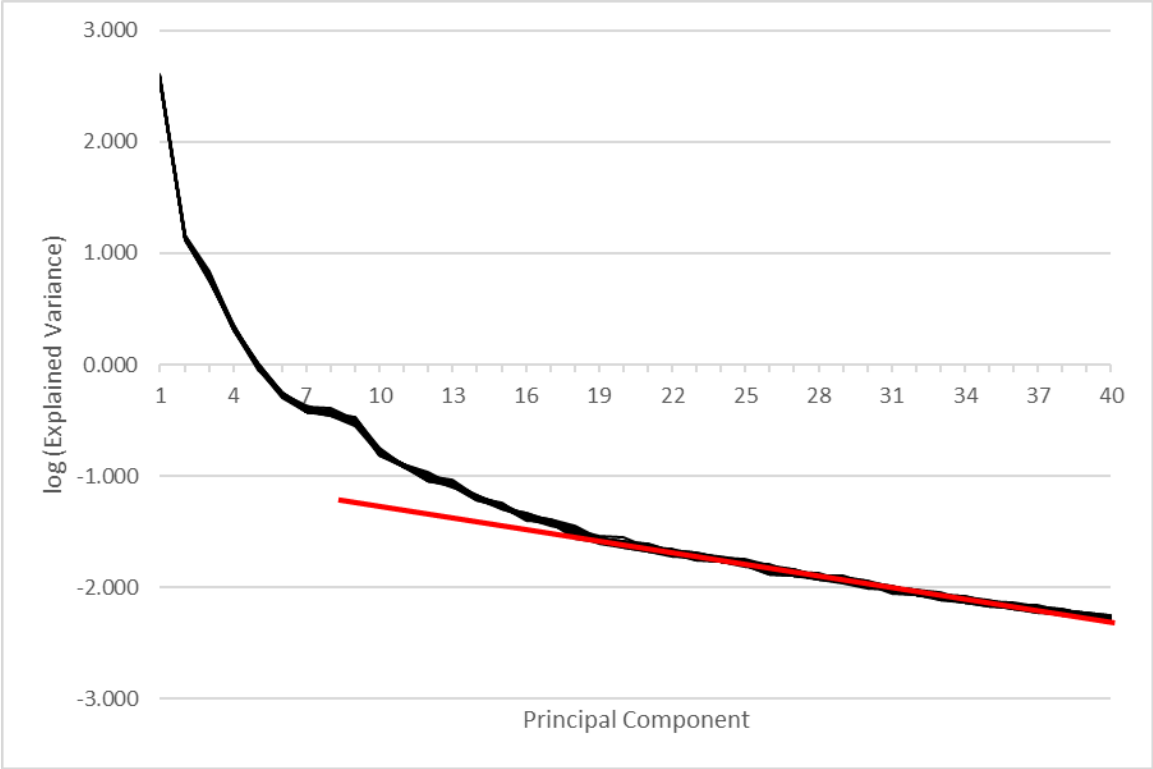

Figure 20: Log-eigenvalue diagram

Considering the mentioned three tests, there was no consensus between them, which suggested the need to explore multiple values: 6, 10, 16, 18, and 32. Retaining too few PCs results in a loss of information, while retaining too many attaches meaning to noise, or as Franklin refers to it, underextraction and overextraction [84]. The percentage of variance explained after 6 components is only 60.02%, as seen in Figure 21, which does not meet the window of explained variance desired – between 70% and 95%. It's not until 14 components are included that the lower threshold is achieved at 70.50%. The disparity of results from the preliminary tests does not

come to a consensus, therefore, all suggestions for the number of PCs are tested in the modeling phase for further examination. An additional point was added for testing on the higher end of the scale making the PCs quantities tested 6, 10, 14, 16, 18, 21, and 32.

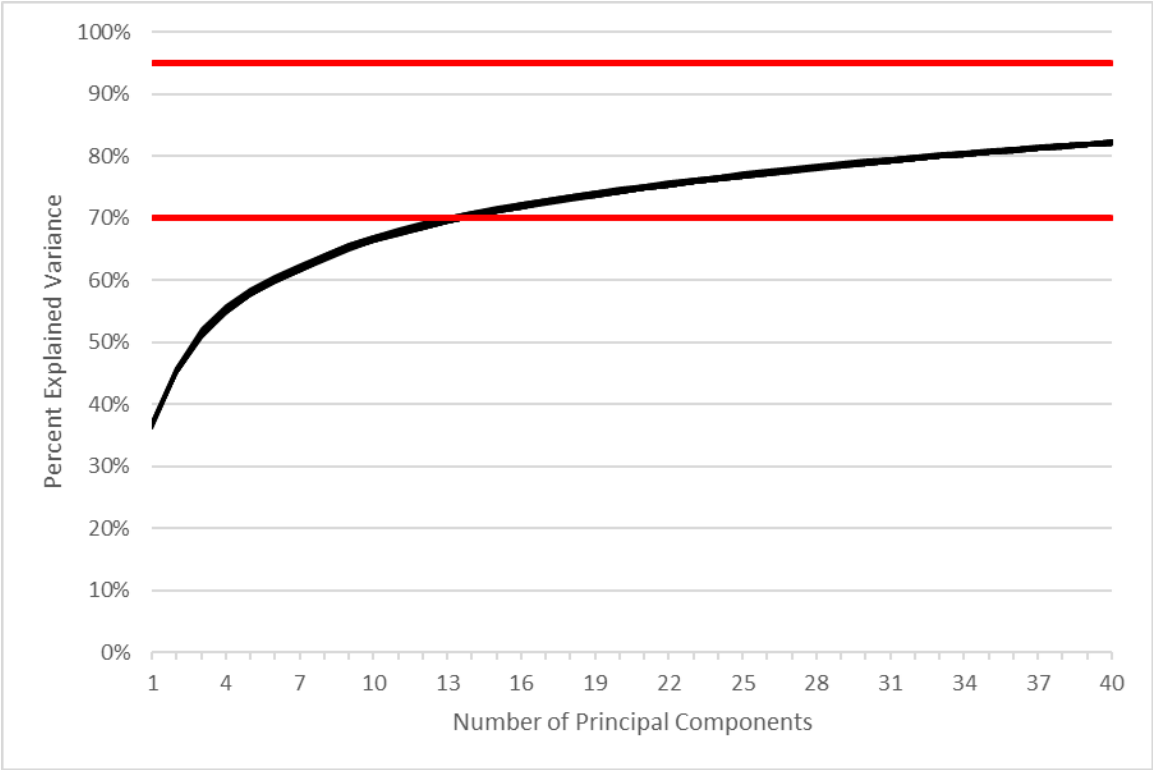

Figure 21: Percent explained variance

Later in the study, it is recommended that 10 PCs are optimal for certain models. Using 10 PCs would be a 98.9% reduction in dimensions while explaining 66.4% of the total variance, as seen in Table 12. Only the unemployment description explicitly states similarities in the findings compared to Neumann. However, Neumann's top principal component quantified as 'Quality of Life' comes from multiple variables: birth rate, fertility rate, infant mortality rate, youth bulge, and population growth [38]. These variables are similar to the description of our 'Population Sizes', which quantifies percentages across population generations affected by birth rates, fertility

rates, and so forth. It is also noted that Neumann presented conflict intensity, the data for the proxy dependent logistic regression variable, in the clustering data, where this study chose to keep that influence apart from the clustering segment. Overall, this study observed more economic influences explaining data variation than what Neumann observed, suggesting that modeling regions may be more economic based rather than a hypothesized holistic culture. This may be in part to the dataset containing 558 economic indicators, whereas Neumann's dataset contained only 4. This may also explain why Rosling's regions worked well when combining countries together, like the Organizations for Economic Co-operation and Development.

Table 12: Principal components descriptions and variance

Principal Component	Leiby Descriptions	% Variation	Neumann Descriptions	% Variation
PC1	Private Non-Guaranteed Debt	36.6%	Quality of Life	24.0%
PC2	Population Sizes	8.5%	Military and Government	11.0%
PC3	Amortization	6.0%	Freedom	7.8%
PC4	Consumption Spending	3.8%	Unemployment	5.6%
PC5	Imports	2.7%	Trade and Religious Diversity	5.1%
PC6	Unemployments	2.1%	Anarchy Government	4.9%
PC7	Natural Resource Values	1.9%	Arable Land	4.3%
PC8	Interest-Free Loans	1.8%	Fresh Water	3.8%
PC9	Purchasing Power Parity	1.7%	Conflict Intensity	3.3%
PC10	Publicly-Guaranteed Debt	1.3%		
Total Variation		*66.4%*		*69.8%*
Dimension Reduction		*98.9%*		*70.0%*

5.5.2 Modeling & Validation Results

Three model types demonstrated the selected combinations of PCs and cluster configurations: static-state with no connection (SSNC), transition-state with no connection (TSNC), and transition-state with geographic connection (TSGC). Confusion matrix accuracy results for all combinations are in Appendix B. For all model types given the available data, the clustering parameter had more influence on predictive outcome than the PCA parameter – meaning varying the number of clusters changed the accuracy more than varying the quantity of PCs used to develop the clusters.

The best training accuracy results for the no connection model demonstrated a preference toward few PCs with static-state demonstrating an average training accuracy of 97.8% with 10 clusters (95.6% weighted) and the transition-state demonstrating an average training accuracy of 98.5% with 8 clusters (96.9% weighted) for 6 PCs. The geographic connection model demonstrated a preference for more PCs, where 18 PCs demonstrated both 100% average and weighted training accuracy for both 9 and 10 clusters. As far as predictive power to assess the number of PCs to anchor analysis on, the average weighted test accuracy of all cluster parameters was examined; results are in Appendix C. Choosing between different numbers of PCs resulted in a maximum difference of only 2.7% predictive accuracy, suggesting that adding more PCs, for the regression models explored and the available variables in the dataset, may add little value. To restate, the 15% explained variance gains between using 6 PCs (59.8% explained variance) or 21 PCs (74.7% explained variance) garnered only a 2.7% modeling confusion matrix accuracy change for the TSGC model type. Furthermore, adding 21 or more PCs saw decreases in predictive accuracy confirming the curse of dimensionality with clustering. Referencing the charts in Appendix B, all the validation results share similar patterns except for using 6 PCs in the SSNC model type. All models demonstrated severe diminishing return for average validation accuracy when increasing the number of clusters, whereas the SSNC model type with 6 PCs did not demonstrate this trend of diminishing returns. It may be assumed that 59.8% explained variance for the 6 PCs model may not be enough information to provide discriminating models.

The highest overall accuracy models were compared between the three types as seen in Figure 22: 16 PCs for SSNC, 14 PCs for TSNC, and 10 PCs for TSGC. In all three cases, there is a point where the average accuracy (blue line) diverges from the weighted accuracy (orange line). These divergences, to no surprise, are due to small

sample sizes within a region. For example, SSNC developed regions with over 150 observations up through 3 clusters. At 4 clusters, a divergence is detected from which a fourth cluster contained only 21 training observations and 9 validation observations. Despite the small number of observations, the models continue to increase in training accuracy while only predicting at naïve levels. The dramatic decrease in accuracy at 9 clusters is due to a region becoming small enough to not have observations containing both states. One of the regions contained only one state from which a model cannot be generated (default accuracy = 0). This is consistent with drops in accuracy for the transition-state models as well, except the occurrence happened with less clusters due to the splitting of models given their static-state. The geographic constraint minimized this occurrence through maintaining larger observation sizes per region cluster.

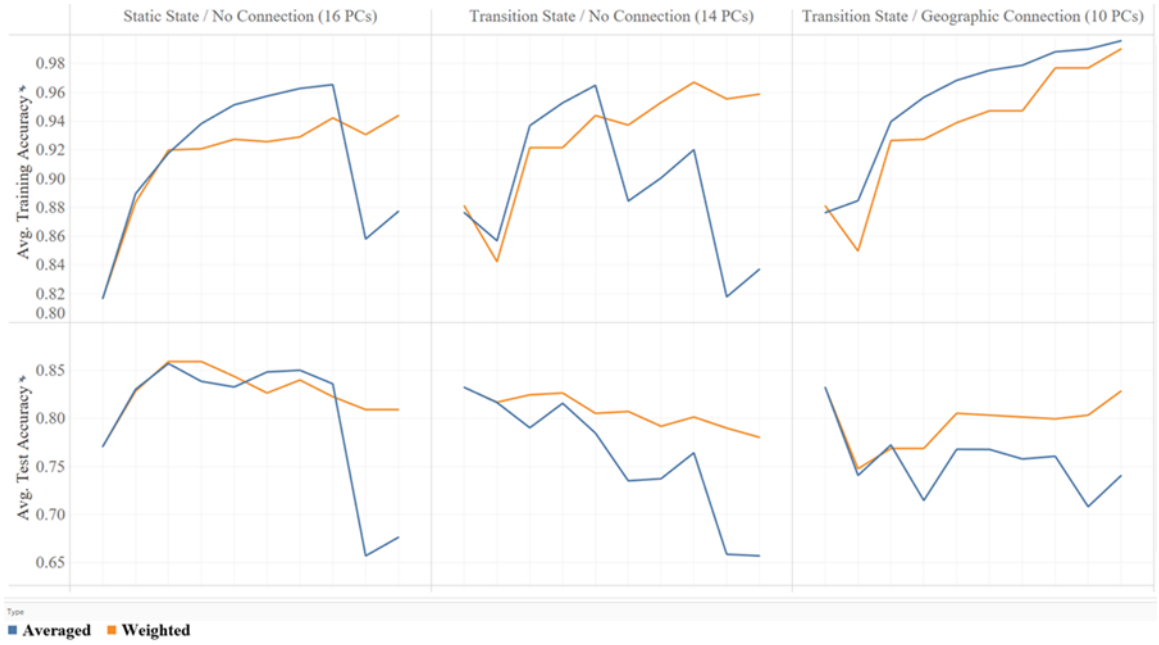

Figure 22: Model type's accuracy across clusters for best PCA parameter

5.5.3 Discussion and a Heuristic Model

Although the basic validation results did not surpass the naïve three-year prediction, most of the models for training accuracy demonstrated potential for good forecasting. However, there are some insights observed within this exploratory study in concert with the refined Shallcross [22] and Neumann [21] studies.

Shallcross proposed that using the dependent variable transition-state would increase the accuracy of the models [22], with Neumann demonstrating a comparison between the Shallcross transition-state study and the Boekestein static-state study increasing by 6% [21]. Although the gains in this study are not as pronounced, the weighted training accuracy as observed in Figure 22 demonstrated the potential for better models using the transition-state dependent variable, especially when employing over 5 regional clusters. Shallcross tailored study years for training and validation sets, meaning not all regions were consistent for every model. Neumann included an interpolation year for validation rather than only extrapolating validation years. This study was not able to tailor years to each region to fine tune each model, as the objective was a wide exploration of multiple quantities of PCs representing the explained variance in the dataset and adjusting the number of clusters to gain insight into quantifying the number of appropriate cluster regions. The data, however, did demonstrate that using a 6-region world model may be too conservative, and that more regions may produce better models.

Another insight that may explain the less pronounced confusion matrix accuracy gains considers the non-stationarity of data. As countries transition into conflict, the quality and accuracy of the data may become suspect, which also may explain why in-conflict predictions are typically lower than their not-in-conflict counterparts [22]. Recalling the method setup, the validation of the data considered a 3-year period. However, as seen in Table 13, years trained has an impact on the prediction of

subsequent years. Years 2013 and 2014 increase the variation to the dataset leading to lower training accuracies, however, their inclusion increase the validation prediction. Unfortunately, this can only be assessed for past data and identifying factors to help assist in selecting appropriate training data periods for future data is outside the scope of this study.

Table 13: Global accuracy for different validation periods

		Averaged		Weighted	
T.Years	V.Years	T.Accuracy	V.Accuracy	T.Accuracy	V.Accuracy
2006-2012	2013-2015	97.5%	76.8%	94.7%	80.4%
2006-2012	2013	97.5%	76.8%	94.1%	80.3%
2006-2014	2015	96.4%	78.8%	92.6%	82.1%

T. - Training, V. - Validation

** TSGC Model with 6 Regions*

One of the issues pointed out when using Neumann's modified k-means approach was the non-contiguousness that could occur. Using the hierarchical clustering method with connectivity should solve this problem. However, a constraint was to force a disconnect between North America and Asia. Relying on scikit-learn's structured agglomerative clustering requires the connectivity matrix to be complete [66]. When the connection matrix is disjointed, the algorithm overrides any connection point constraint and uses dimensional Euclidean space to pair observations. It was assumed that the algorithm would override the connectivity matrix when all possible connections were made, which for the supplied matrix would be the last connection. However, for the TSGC 6-cluster model, a connection between Asia and South America was made on the 10th to last pairing resulting in a noncontiguous region, as seen in Figure 23.

A gem of hierarchical clustering is that the dendrogram product provides an insightful benefit to the construction of the regions. Pairs that are connected early portray closer dimensional Euclidean distance than pairs made later. This assisted

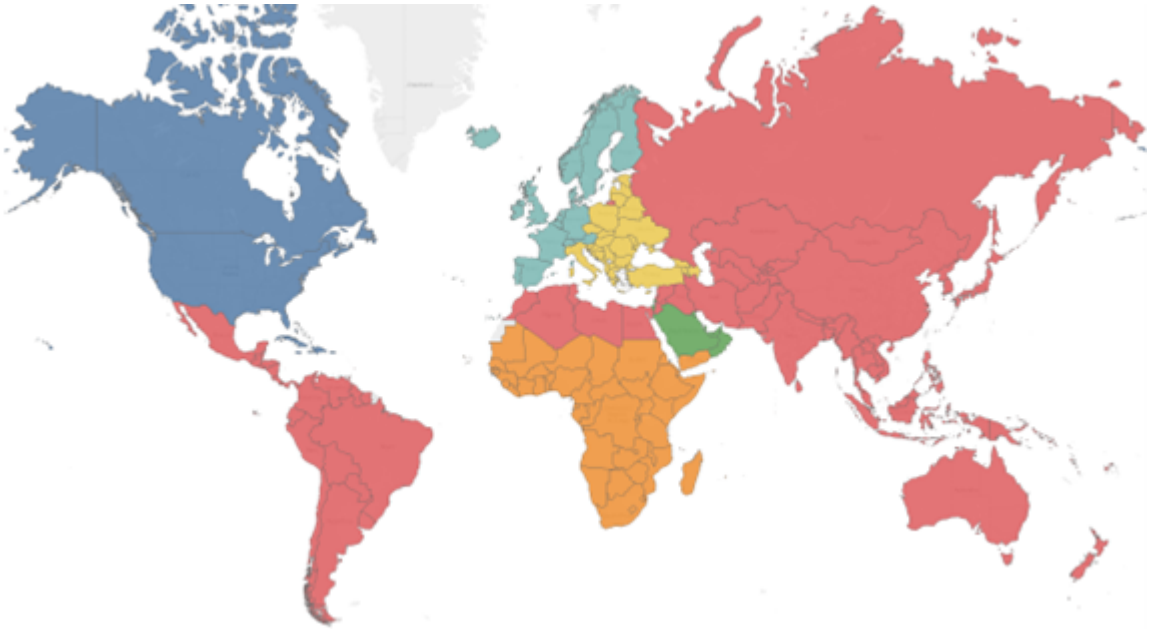

Figure 23: 6-cluster TSGC regional map

in developing a heuristic approach model to observe increasing the number of regions above six. The heuristic approach observed three rules. First, the regions would adhere to the strict connection constraint provided through the geographic connection matrix. Second, each region would retain at least six training observations. Third, the regions are created using the dendrogram by moving the "least likely" trees until the first two constraints are satisfied. These "least likely" trees refer to the fusion of observations through dimensional Euclidean differences rather than the geographic connection constraint. Normally when viewing a hierarchical clustering dendrogram, all tree branches would spread upward in the same direction, but using a connection constraint, some branches become inverted to satisfy the connection constraint as well as the dimensional likeness. Observing these inverted branches highlight potential countries to move to other clusters as their dimensional likeness is weak and heavily constrained by the geographic connection.

Although TSGC results demonstrated increased accuracy up to 10 clusters, the heuristic map resulted in only 7 regions. Clusters 8-10 contained small amounts of

observations when broken down between state-country pairs resulting in infeasible regions. For example, cluster 8 included Cuba, Haiti, the Dominican Republic, Jamaica, and the Bahamas. However, logistic regression requires observations of both categories of the binary dependent variable. For TSGC, the rare observation needed is the *change in transition state* given the prior year's *static state*. Globally, this occurs 15% of the time, but the distribution is not distributed equally across the globe. Therefore, cluster 8, along with clusters 9 and 10, did not contain enough observations to meet the second heuristic rule. The branch was also inverted, suggesting a defense for potentially reassigning its subsequent cluster connection.

The new heuristic constructed regional map is presented in Figure 24. The model incorporated the three gained insights: transition-state dependent variable combined with more than 6 regions, a 9-year training set (2006-2014) with a 1-year validation set (2015), and ensuring all regions are contiguous and well represented with observations. The results demonstrated a high training accuracy of 96.1% with an 85.4% validation accuracy, as seen in Table 14. It is worth highlighting that the in-conflict accuracy is greater than the not-in-conflict accuracy, overcoming quality and accuracy issues innate to in-conflict data.

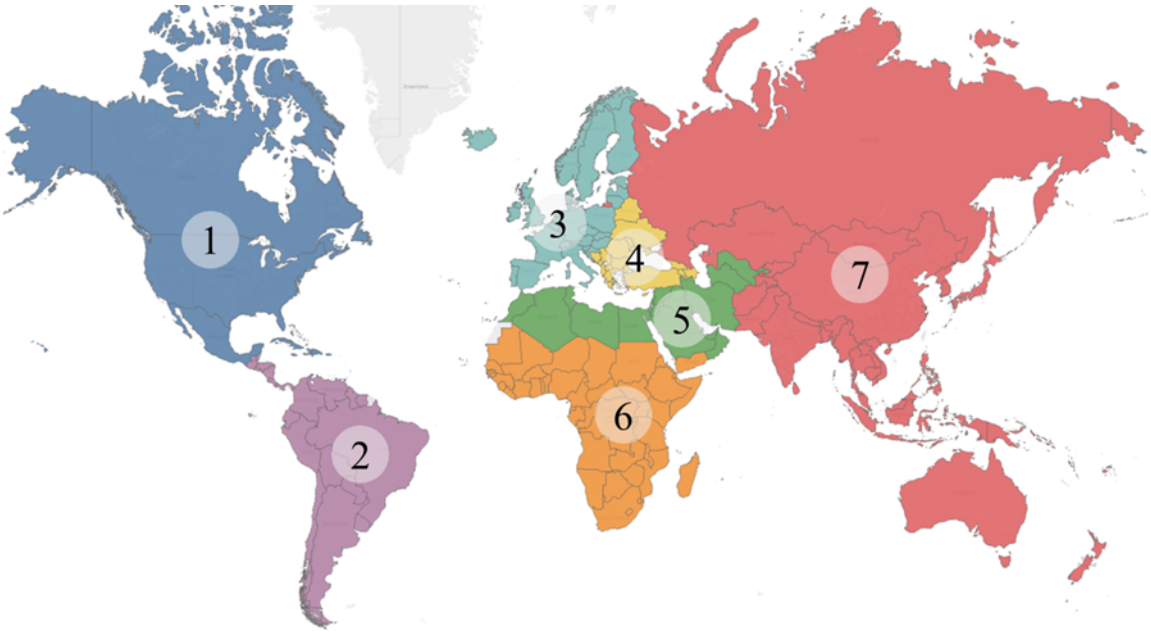

Figure 24: Modified 7-cluster transition-state regional map

Table 14: Modified 7-cluster TSGC regional results

		Training		Validation	
Region	Transition-State	Obs	Accuracy	Obs	Accuracy
1	Not-In-Conflict	45	100.0%	4	100.0%
2	Not-In-Conflict	98	96.9%	9	100.0%
3	Not-In-Conflict	212	100.0%	23	73.9%
4	Not-In-Conflict	84	100.0%	9	77.8%
5	Not-In-Conflict	73	82.2%	6	83.3%
6	Not-In-Conflict	214	91.1%	18	72.2%
7	Not-In-Conflict	148	96.6%	13	69.2%
1	In-Conflict	27	100.0%	4	75.0%
2	In-Conflict	91	97.8%	12	100.0%
3	In-Conflict	22	100.0%	3	100.0%
4	In-Conflict	69	100.0%	8	75.0%
5	In-Conflict	98	100.0%	13	92.3%
6	In-Conflict	218	90.8%	30	86.7%
7	In-Conflict	158	90.5%	21	90.5%
Total	*Not-In-Conflict*		*95.3%*		*82.4%*
Total	*In-Conflict*		*97.0%*		*88.5%*
Total	**Global**		**96.1%**		**85.4%**

Training Years (2006-2014), Validation Year (2015)

** TSGC Model with 7 Modified Regions*

5.6 Summary

The goal of the research sought to identify an optimal number of clustering regions and delineate regional boundaries for conflict modeling. The additional constraint of contiguousness assumes that geographic proximity is as or more important than country indicators alone. Furthermore, maintaining contiguous modeling regions assists decision makers with distributing resources and aide.

This study challenged two assumptions from Neumann producing insights otherwise left unknown in prior research. The first challenged the k-means approach, which assumes a pre-defined number of regions. The second challenged the method to provide contiguous regions. The use of hierarchical clustering allows researchers to observe the pairing of countries based on political, economic, and social aspects. Of the three aspects, this research demonstrated that economic indicators provide a large bulk of the influence for establishing dimensions that feed the country clustering method. Demonstrating an economic heavy influence for partitioning the world into regions supports other successful country conflict region studies relying on Rosling's partitions. This became more apparent only when increasing the number of independent variables from 30 to 932. Although increasing the number of variables also increases the number of dimensions clustering methods need to contend with, feature extraction assists in reducing over 96% of the dimensions, solving the curse of dimensionality.

Many parameters are involved with constructing country conflict models. This research explored a framework to increase predictive accuracy. Although other metrics quantify the statistical viability of a model, predictive accuracy provides the practical usefulness for decision makers. Given the available variables in the dataset, this research provides insight into the desirable number of PCs to use for clustering countries into regions. The methodological setup further provides insight into segmenting

the world into regions for modeling. Using hierarchical clustering highlights not only which countries should define a region, but also how those regions formed. The formation aspect adds value over other clustering methods, such as k-means clustering, which suffers from local optima based upon the initial random state. The dendrogram facilitates observing which countries have the strongest cultural connection to one another, adding yet further information toward constructing regions constrained outside dimensional Euclidean distance.

This explorational study highlighted classifying countries to regions through balancing cultural boundaries with geographical boundaries. Russia geographically borders both Kazakhstan and Belarus, but the cultural boundary between Russia and Kazakhstan is much greater than between Russia and Belarus. Given the available dataset, Russia's first connection to form regions always culturally links to Kazakhstan. However, the discriminating link for Belarus between region 4 and region 7 for the modified 7-cluster transition-state global model is weaker yet places it in region 4. Similarly, Australia remains the last country to link to region 7, leading toward a hypothesis that geographic boundary heavily influences the link rather than cultural factors. These insights are easily seen through hierarchical clustering's dendrogram, balancing geographic and cultural boundaries. As regions play a significant role in developing accurate prediction models, the methodology of using hierarchical clustering becomes valuable.

There do remain several obstacles when implementing hierarchical clustering to produce regional maps. Practically speaking, the Pacific Ocean creates a natural delineation between regions, but algorithms do not always handle forced connections (or disconnections) as expected. An adequate distribution of observations, in addition to number of country observations, plays a vital role for adequate statistical modeling when constructing the regions. The severe drop in global accuracy after a

sufficient number of clusters clearly demonstrates this influence as clusters increase hindering the distribution of observations. Solving both maintaining a strict adherence to the geographic connection constraint and maintaining adequate observations for robust modeling may require a modified hierarchical clustering algorithm for conflict modeling. Once solved, more emphasis on selection of variables for the logistic regression models, possibly through purposeful selection, should further increase the global predictive output of the model.

Finally, the research exposed the assumption that emphasizing a 6-cluster regional map for conflict modeling may be a limiting factor. This hierarchical approach methodology demonstrates that regional model accuracy increases when exploring a greater number of regions. Specifically, the modified 7-region map garnered high training accuracy with competitive validation accuracy. These insights will propel advances in conflict modeling and assessments, ultimately assisting leaders to have a greater understanding of threats and vulnerabilities within their regions so that they may more effectively plan, prepare, and palliate possible threats.

VI. Conclusions

6.1 Summary

The quality of the decision is only as good as the data informing it. Combatant Commands provide a demand signal to other organizations for limited resources, but justification can be more qualitative in nature rather than quantitative. Country conflict research provides the qualitative rigor toward defending the demand signal. This research reinforced and added to bolstering country conflict modeling. Through codifying the taxonomy and ontology of country conflict research, we've cemented the foundation for all future work concerning the country conflict modeling. Like what Ward did in the area of debunking the reliance on p-value [30], this research advocates for a whole of concept approach to modeling country conflict. This includes gathering as much data as possible on countries to ensure that culture is accounted for along with the geographic, political, economic, and societal influences. It is not lost that as more data is accumulated, new problems arise in how to handle and process it all. New methods were developed to ensure the integrity and completeness of data is retained for use in statistical modeling methods. In fact, the new MASS-impute method may expand outside the realm of country conflict "data repairing" and become a vital tool for other research areas when other commercial imputation methods fail to iterate and converge. Finally, to address specific questions about the appropriate number of Combatant Commands or worldwide regions, hierarchical cluster was explored to include parameters that are otherwise left as assumptions. The research community is left with a 7-region division of countries that will improve the predictive accuracy of future work.

Finally, this research addressed three questions contributing to the body of country conflict and peace modeling knowledge, providing decision makers with the informa-

tion to lead their organizations.

Research Question 1 What data sources are available and what data elements provide statistical insight to country conflict modeling? A variety of data sources were identified along with variables used in prior research, consolidated in Table 6. Furthermore, core variables were identified through investigating trends in prior research, suggesting that future models should consider proxies listed in Table 5. However, the major contribution highlighted through this exploration of literature is the functional ontology that demonstrates that accurate country conflict predictions are a function of political, economic, and social aspects, all of which should be present in future models. This was only apparent after mapping proxy variables into a concise taxonomy.

Research Question 2 How can incomplete country data be addressed through imputation methods? For small datasets, commercial imputation packages are suffice for developing plausible estimates for missing values. However, for larger datasets, numerical problems necessitated an alternative method to develop these estimates. Through a multicollinear applied stepwise stochastic process, plausible estimates are possible with defensible observations and metrics. Additional contributions include addressing the multiple imputation tolerance problem and providing an alternative solution through the leverage of comparing known residuals by variable to the variance of the estimates. The community is also left with the MASS-impute algorithm that can be applied to more than just country conflict data.

Research Question 3 Are there defensible, analytical arguments for partitioning the world into management sectors? Although prior research had suggested that country conflict modeling predictions improve when modeling worldwide regions rather than globally, this research demonstrated the improvements while

overcoming some of the assumptions. A contribution of an alternative methodology using hierarchical clustering was presented strengthening the mathematical rationale to assign countries to regions rather than compromising a clustering method's output with a weighted geographic constraint. The assessment of multiple combinations of partitions illuminated that predictive accuracies for identifying country instability increase when using 7 geographic regions as opposed to 6.

6.2 Future Work

Two areas are suggested to expand on the work provided: one in the area of imputation and one in the area of clustering. The MASS-impute methodology introduced a new ways of thinking toward developing plausible estimates. Innate to the approach facilitated an answer to an otherwise ambiguous answer toward solving convergence tolerance: the variability in estimate for the unknown true value. Multiple imputation sought to account for the variability, but in return creates skepticism concerning convergence. MASS-impute addresses this. However, the current state of MASS-impute may be too accommodating allowing too much variability in the noise element to affect estimates. Future work should examine different numbers of initial iterations to balance processing time with reduced variability of estimates upon convergence. As it stands, the parameter allows convergence after iteration 1, while there is a hypothesis that setting a parameter to not allow convergence until later iterations will statistically improve estimate behavior. Furthermore, all residuals are currently used for supplying the stochastic nature to the methodology. Whereas all known data points should be observed in research, the methodology does not address the distribution of all known data points. With extreme outliers present in some of the country conflict variables, allowing all residuals to form the stochastic nature of

the estimates may be too generous. Future research should investigate the solving the balance between the plausible variability in estimates and the distribution of known data points.

As far as clustering countries into regions for country conflict modeling, hierarchical clustering provides benefits not explored in prior research. This exploration into the clustering method highlighted gaps inherent in commercial applications of the methodology. This calls for a modified algorithm to accommodate specific constraints required in developing practical regions for decision makers. Accurate predictions of instability are needed, but there also needs to be balance between the practical use of the model, which includes contiguousness in regions. The modified approach should incorporate the ability to handle force disconnections along with a mechanism to distribute observations equitably for adequate modeling. Furthermore, due to the complexity of the research problem, this study stopped short of presenting a case for variable selection and model coefficients within each region. Developing statistical models remains an open research area to continue improving the accuracy of predictions.

Appendix A. Author Reference for Paper 1

Table 15: Author reference for paper 1

Index	Study	Year	Reference
1	Gartzke	2001	(Gartzke et al., 2001)
2	Fearon & Laitin	2003	(Fearon & Laitin, 2003)
3	Collier & Hoeffler	2004	(Collier & Hoeffler, 2004)
4	Gates	2006	(Gates et al., 2006)
5	Hegre & Sambanis	2006	(Hegre & Sambanis, 2006)
6	Ostby	2008	(Østby, 2008)
7	Goldstone	2010	(Goldstone et al., 2010)
8	Hegre	2013	(Hegre et al., 2013)
9	Buhaug	2014	(Buhaug et al., 2014)
10	Boekestein	2015	(Ahner et al., 2015)
11	Muchlinski	2016	(Muchlinkski et al., 2016)
12	Shallcross	2016	(Shallcross, 2016)
13	Celiku & Kraay	2017	(Celiku & Kraay, 2017)
14	Leiby	2017	(Leiby, 2017)
15	Brantley	2018	(Ahner & Brantley, 2018)
16	Neumann	2018	(Neumann, 2018)

Appendix B. Model dataset accuracy across clusters

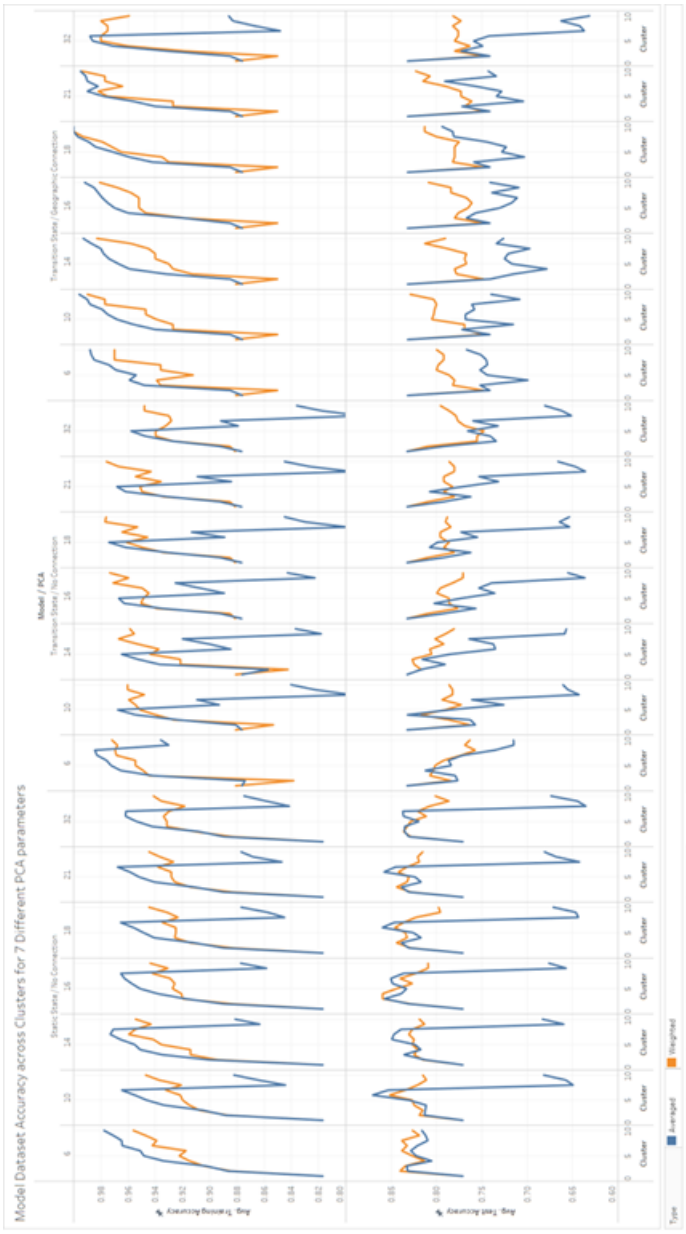

Figure 25: Model dataset accuracy across clusters

Appendix C. Model dataset accuracy across principal components

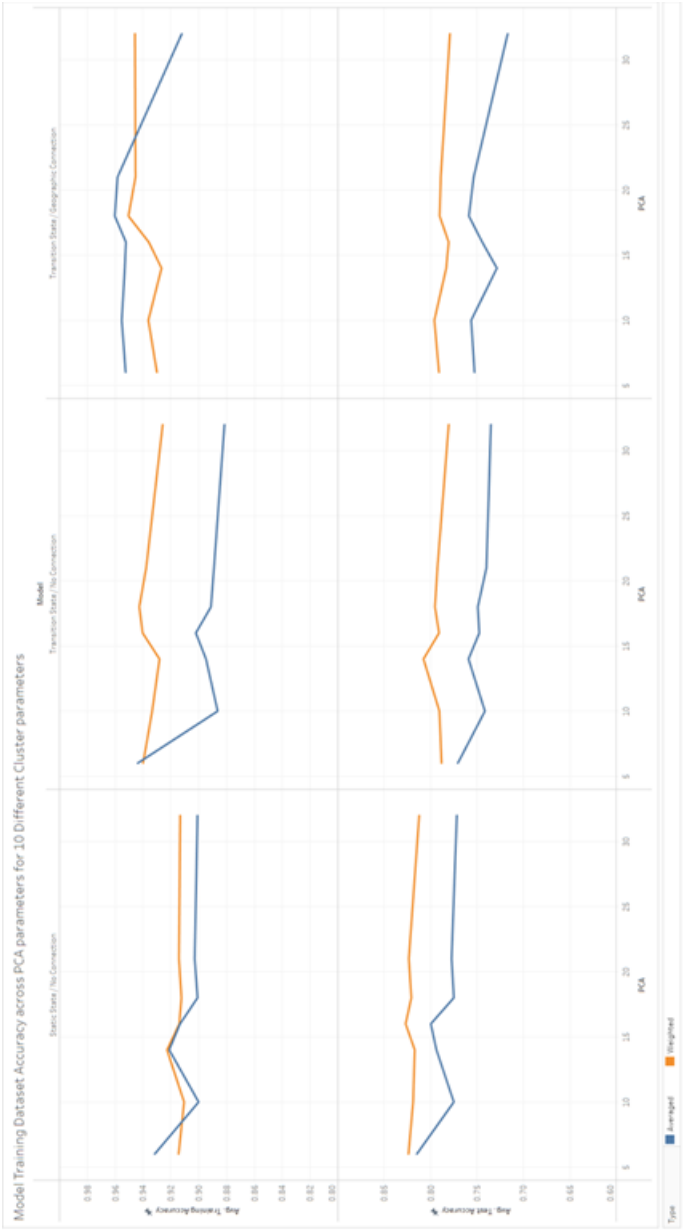

Figure 26: Model dataset accuracy across principal components

Bibliography

1. Mackenzie Eaglen. Putting combatant commanders on a demand signal diet. *War on the Rocks*, 9, 2020.

2. Steve Ferenzi and Keith Weber. We need to reorganize more than the military. *Defense One*, 16, 2020.

3. Albert Einstein and George Bernard Shaw. *Einstein on cosmic religion and other opinions and aphorisms*. Courier Corporation, March 2012.

4. Håvard Hegre, Nils W Metternich, Håvard Mokleiv Nygård, and Julian Wucherpfennig. Introduction: Forecasting in peace research. *Journal of Peace Research*, 54(2):113–124, March 2017.

5. Michael D. Ward. Can We Predict Politics? Toward What End? *Journal of Global Security Studies*, 1(1):80–91, February 2016.

6. Thomas Wencker, Christoph Trinn, and Aurel Croissant. Data Bases and Statistical Systems: Security and Conflict. In James Wright, editor, *International Encyclopedia of the Social and Behavioral Sciences*, pages 836–843. Elsevier Ltd, Amsterdam, March 2015.

7. James D. Fearon and David D. Laitin. Ethnicity, Insurgency, and Civil War. *The American Political Science Review*, 97(1):75–90, February 2003.

8. Scott Gates, Håvard Hegre, Mark P Jones, and Håvard Strand. Institutional Inconsistency and Political Instability: Polity Duration, 1800–2000. *American Journal of Political Science*, 50(4):893–908, October 2006.

9. 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, January 2010.

10. Erik Gartzke, Quan Li, and Charles Boehmer. Investing in the peace: Economic interdependence and international conflict. *International Organization*, 55(2):391–438, April 2001.

11. Halvard Buhaug, Lars Erik Cederman, and Kristian Skrede Gleditsch. Square Pegs in Round Holes: Inequalities, Grievances, and Civil War. *International Studies Quarterly*, 58(2):418–431, June 2014.

12. Gudrun Østby. Inequalities, the Political Environment and Civil Conflict: Evidence from 55 Developing Countries. In Frances Stewart, editor, *Horizontal Inequalities and Conflict: Understanding Group Violence in Multiethnic Societies*, pages 136–159. Palgrave Macmillan UK, London, April 2008.

13. Bledi Celiku and Aart Kraay. Predicting Conflict. World Bank Group Policy Research Working Paper, May 2017.

14. Håvard Hegre and Nicholas Sambanis. Sensitivity Analysis of Empirical Results on Civil War Onset. *Journal of Conflict Resolution*, 50(4):508–535, August 2006.

15. David Muchlinski, David Siroky, Jingrui He, and Matthew Kocher. Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis*, 24(1):87–103, January 2016.

16. Håvard Hegre, Joakim Karlsen, Håvard Mokleiv Nygård, Håvard Strand, and Henrik Urdal. Predicting Armed Conflict, 2010-2050. *International Studies Quarterly*, 57(2):250–270, June 2013.

17. Peter Wallensteen and Margareta Sollenberg. Armed Conflict, 1989–2000. *Journal of Peace Research*, 38(5):629–644, September 2001.

18. Darryl Ahner, Benjamin Boekestein, and Richard Deckro. A Predictive Model of World Conflict using Open Source Data, Mar 2015.

19. Darryl Ahner and Luke Brantley. Finding the Fuel of the Arab Spring Fire: a Historical Data Analysis. *Journal of Defense Analytics and Logistics*, 2(2):58–68, Jan 2018.

20. Benjamin D. Leiby. A Conditional Logistic Regression Predictive Model of World Conflict Considering Neighboring Conflict and Environmental Security. Master's thesis, Air Force Institute of Technology, WPAFB, Ohio, March 2017.

21. Sarah Neumann, Darryl Ahner, and Raymond Hill. Forecasting Country Conflict Using Statistical Learning Methods. *Journal of Defense Analytics and Logistics*, 6(1):59–72, Apr 2022.

22. Nicholas Shallcross and Darryl Ahner. Predictive Models of World Conflict: Accounting for Regional and Conflict-State Differences. *The Journal of Defense Modeling and Simulation*, 17(3):243–267, jul 2019.

23. Heidelberg Institute for International Conflict Research (HIIK). Conflict Barometer 2019. Technical report, HIIK, Heidelberg, 2020.

24. Therése Pettersson and Magnus Öberg. Organized violence, 1989–2019. *Journal of Peace Research*, 57(4):597–613, June 2020.

25. Heidelberg Institute for International Conflict Research (HIIK). Conflict Barometer 2010. Technical report, HIIK, Heidelberg, 2010.

26. Mikael Eriksson and Peter Wallensteen. Armed Conflict, 1989–2003. *Journal of Peace Research*, 41(5):625–636, September 2004.

27. Nicholas Shallcross. A Logistic Regression and Markov Chain Model for the Prediction of Nation-state Violent Conflicts and Transitions. Master's thesis, Air Force Institute of Technology, WPAFB, Ohio, March 2016.

28. Thomas F. Homer-Dixon. On the Threshold: Environmental Changes as Causes of Acute Conflict. *International Security*, 16(2):76–116, August 1991.

29. Paul Collier and Anke Hoeffler. Greed and Grievance in Civil War. *Oxford Economic Papers*, 56(4):563–595, oct 2004.

30. Michael D. Ward, Brian D. Greenhill, and Kristin M. Bakke. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):363–375, July 2010.

31. Kristian Skrede Gleditsch, Nils W Metternich, and Andrea Ruggeri. Data and progress in peace and conflict research. *Journal of Peace Research*, 51(2):301–314, March 2014.

32. Lotta Themnér and Peter Wallensteen. Armed Conflict, 1946–2010. *Journal of Peace Research*, 48(4):625–636, July 2011.

33. Minorities at Risk Project. *Minorities at Risk Dataset*. Center for International Development and Conflict Management, College Park, MD, 2009. Retrieved from http://www.mar.umd.edu/.

34. Gunter Bolch, Stefan Greiner, Hermann de Meer, and Kishor S. Trivedi. *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. John Wiley & Sons, Hoboken, NJ, 2nd edition, mar 2006.

35. Frederick S. Hillier and Gerald J. Lieberman. *Introduction to Operations Research*. McGraw Hill, New York, 10th edition, jan 2014.

36. David W. Hosmer Jr, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, Hoboken, New Jersey, 3rd edition, April 2013.

37. Zoran Bursac, Clinton Heath Gauss, David Keith Williams, and David W. Hosmer. Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*, 17(3):1–8, December 2008.

38. Sarah Neumann. Forecasting Country Conflict Within Modified Combatant Command Regions Using Statistical Learning Methods. Master's thesis, Air Force Institute of Technology, WPAFB, Ohio, March 2018.

39. Hannes Mueller and Christopher Rauh. Reading Between the Lines: Prediction of Political Violence Using Newspaper Text. *American Political Science Review*, 112(2):358–375, May 2018.

40. Benjamin D. Leiby and Darryl K. Ahner. Multicollinearity applied stepwise stochastic imputation: A large dataset imputation through correlation-based regression. preprint on webpage at `https://doi.org/10.21203/rs.3.rs-1894388/v1`, 2022.

41. Håvard Hegre, Håvard Mokleiv Nygård, and Peder Landsverk. Can We Predict Armed Conflict? How the First 9 Years of Published Forecasts Stand Up to Reality. *International Studies Quarterly*, jan 2021.

42. Julián Luengo, Salvador García, and Francisco Herrera. On the Choice of the Best Imputation Methods for Missing Values Considering Three Groups of Classification Methods. *Knowledge and Information Systems*, 32(1):77–108, Jul 2012.

43. Stef van Buuren. *Flexible Imputation of Missing Data*. CRC Press, Boca Raton, 2nd edition, Jul 2018.

44. Donald B. Rubin. Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, 91(434):473–489, jun 1996.

45. Stef van Buuren and Karin Groothuis-Oudshoorn. Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67, Dec 2011.

46. Yajuan Si, Steve Heeringa, David Johnson, Roderick Little, Wenshuo Liu, Fabian Pfeffer, and Trivellore Raghunathan. Multiple imputation with massive data: An application to the panel study of income dynamics. *arXiv Preprint arXiv:2007.03016*, Jul 2020.

47. Yi Deng, Changgee Chang, Moges Seyoum Ido, and Qi Long. Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Scientific reports*, 6(1):1–10, Feb 2016.

48. Roderick J Little. On algorithmic and modeling approaches to imputation in large data sets. *Statistica Sinica*, 30(4):1685–1696, Jan 2020.

49. Zachary J Kane. An Imputation Approach to Developing Alternative Futures of Country Conflict. Master's thesis, Air Force Institute of Technology, Mar 2019.

50. Cattram D. Nguyen, John B. Carlin, and Katherine J. Lee. Practical Strategies for Handling Breakdown of Multiple Imputation Procedures. *Emerging Themes in Epidemiology*, 18(1):1–8, Dec 2021.

51. Catrin O. Plumpton, Tim Morris, Dyfrig A. Hughes, and Ian R. White. Multiple Imputation Of Multiple Multi-Item Scales When A Full Imputation Model Is Infeasible. *BMC Research Notes*, 9(1):1–16, Dec 2016.

52. Eduardo Nunez, Ewout W Steyerberg, and Julio Nunez. Regression modeling strategies. *Revista Española de Cardiología (English Edition)*, 64(6):501–507, Jun 2011.

53. John A Nelder. The selection of terms in response-surface models—how strong is the weak-heredity principle? *The American Statistician*, 52(4):315–318, May 1998.

54. John R Oneal and Bruce Russett. Rule of three, let it be? when more really is better. *Conflict Management and Peace Science*, 22(4):293–310, Sep 2005.

55. George Smith Patton and Paul Donal Harkins. *War As I Knew It*. Houghton Mifflin Harcourt, 1995.

56. Yuan Luo. Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*, 23(1), Jan 2022.

57. Geeta Chhabra, Vasudha Vashisht, and Jayanthi Ranjan. A Comparison of Multiple Imputation Methods for Data with Missing Values. *Indian Journal of Science and Technology*, 10(19):1–7, May 2017.

58. Anil Jadhav, Dhanya Pramod, and Krishnan Ramanathan. Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, 33(10):913–933, Jul 2019.

59. Benjamin D. Leiby and Darryl K. Ahner. A Large Dataset Imputation Approach Applied to Country Conflict Prediction Data. *International Journal of Mathematical and Computational Sciences*, 16(3):11–17, Mar 2022.

60. Jeffrey C. Wayman. Multiple Imputation For Missing Data: What Is It And How Can I Use It? In *Annual meeting of the American Educational Research Association, Chicago, IL.*, volume 2, page 16, Apr 2003.

61. Paul Lodder. To Impute or Not Impute : That's the Question. In Gideon J. Mellenbergh and Herman J. Adér, editors, *Advising on research methods: Selected*

topics (2013), pages 1–7. Johannes van Kessel Publishing, The Netherlands, Jul 2013.

62. Vincent Arel-Bundock and Krzysztof J. Pelc. When Can Multiple Imputation Improve Regression Estimates? *Political Analysis*, 26(2):240–245, Mar 2018.

63. Thomas Lindner, Jonas Puck, and Alain Verbeke. Misconceptions About Multicollinearity in International Business Research: Identification, Consequences, and Remedies. *Journal of International Business Studies*, 51(3):283–298, Apr 2020.

64. David Disatnik and Liron Sivan. The Multicollinearity Illusion in Moderated Regression Analysis. *Marketing Letters*, 27(2):403–408, Jun 2016.

65. Andrew C. Harvey. Miscellanea: Some Comments on Multicollinearity in Regression. *Applied Statistics*, 26(2):188–191, Jun 1977.

66. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, Jul 2011.

67. Joseph Kearney and Shahid Barkat. Autoimpute Documentation, Jan 2021.

68. Henry B. Mann and Donald R. Whitney. On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, Mar 1947.

69. Lorentz Jäntschi and Sorana D. Bolboacă. Computation of Probability Associated with Anderson-Darling Statistic. *Mathematics*, 6(88):1–16, May 2018.

70. Sonja Engmann and Denis Cousineau. Comparing Distributions: the Two-Sample Anderson-Darling Test as an Alternative to the Kolmogorov-Smirnoff Test. *Journal of Applied Quantitative Methods*, 6(3):1–17, Sep 2011.

71. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

72. Dagobert L Brito and Michael D Intriligator. Conflict , War , and Redistribution. *The American Political Science Review*, 79(4):943–957, dec 1985.

73. Hans Rosling. The Best Stats You've Ever Seen. In *TED Conferences*, Monterey, California, feb 2006.

74. Michael Minkov and Geert Hofstede. Is National Culture a Meaningful Concept? Cultural Values Delineate Homogeneous National Clusters of In-Country Regions. *Cross-Cultural Research*, 46(2):133–159, may 2012.

75. Vipin Gupta, Paul J. Hanges, and Peter Dorfman. Cultural Clusters: Methodology and Findings. *Journal of World Business*, 37(1):11–15, mar 2002.

76. Simcha Ronen and Oded Shenkar. Mapping World Cultures: Cluster Formation, Sources and Implications. *Journal of International Business Studies*, 44(9):867–897, dec 2013.

77. Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering. *ACM Transactions on Knowledge Discovery from Data*, 3(1), mar 2009.

78. Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When Is "Nearest Neighbor" Meaningful? In *International Conference on Database Theory*, pages 217–235, Springer, Berlin, Heidelberg, jan 1999.

79. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer, New York, jun 2013.

80. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, 2nd edition, aug 2009.

81. Richard Cangelosi and Alain Goriely. Component Retention in Principal Component Analysis with Application to cDNA Microarray Data. *Biology Direct*, 2(2):1–21, jan 2007.

82. Paul Allison. What's the Best R-squared for Logistic Regression. *Statistical Horizons*, 13, feb 2013.

83. Tue Tjur. Coefficients of Determination in Logistic Regression Models - A New Proposal: The Coefficient of Discrimination. *American Statistician*, 63(4):366–372, nov 2009.

84. Scott B. Franklin, David J. Gibson, Philip A. Robertson, John T. Pohlmann, and James S. Fralish. Parallel Analysis: A Method for Determining Significant Principal Components. *Journal of Vegetation Science*, 6(1):99–106, feb 1995.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
15-09-2022	Doctoral Dissertation	Sept 2019	Sept 2022

4. TITLE AND SUBTITLE

Improving Country Conflict and Peace Modeling: Datasets, Imputations, and Hierarchical Clustering

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)

Leiby, Benjamin, Major, USAF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way, Building 640
WPAFB OH 45433-8865

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-22-S-065

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Intentionally Left Blank		

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRUBTION 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

Many disparate datasets exist that provide country attributes covering political, economic, and social aspects. Unfortunately, this data often does not include all countries nor is the data complete for those countries included, as measured by the dataset's missingness. This research addresses these dataset shortfalls in predicting country instability by considering country attributes in all aspects as well as in greater thresholds of missingness. First, a structured summary of past research is presented framed by a developed casual taxonomy and functional ontology. Additionally, a novel imputation technique for very large datasets is presented to account for moderate missingness in the expanded dataset. This method is further extended to establish the MASS-impute algorithm, a multicollinearity applied stepwise stochastic imputation method that overcomes numerical problems present in preferred commercial packages. Finally, the imputed datasets with 932 variables is used to develop a hierarchical clustering approach that accounts for geographic and cultural influences that are desired in the practical use of modeling country conflict.

15. SUBJECT TERMS

conflict models, conflict datasets, country conflict prediction, correlation, hierarchical clustering, imputation, stochastic regression, taxonomy

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	UU	154

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Dr. Darryl Ahner, AFIT/ENR	(937)255-6565, ext 4708 Darryl.Ahner@afit.edu