Received 2 November 2022
Revised 8 April 2023
16 April 2023
Accepted 18 April 2023

A hierarchical cluster approach toward understanding the regional variable in country conflict modeling

Benjamin Leiby
Department of Operational Sciences, Air Force Institute of Technology, Wright-Patterson AFB, Ohio, USA, and
Darryl Ahner
Air Force Institute of Technology, Wright-Patterson AFB, Ohio, USA

Abstract

Purpose – This paper aims to examine how the regional variable in country conflict modeling affects forecast accuracy and identifies a methodology to further improve the predictions.
Design/methodology/approach – This paper uses statistical learning methods to both evaluate the quantity of data for clustering countries along with quantifying accuracy according to the number of clusters used.
Findings – This study demonstrates that increasing the number of clusters for modeling improves the ability to predict conflict as long as the models are robust.
Originality/value – This study investigates the quantity of clusters used in conflict modeling, while previous research assumes a specific quantity before modeling.

Keywords Cluster analysis, Country conflict, Forecasting, Principal components evaluation
Paper type Research paper

1. 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 70 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 on 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 (Brito and Intriligator, 1985). For example, Gartzke focused mainly on economic contributions of capital interdependence (Gartzke *et al.*, 2001), while Goldstone stressed political factors such as regime type and state-led discrimination (Goldstone *et al.*, 2010), and Østby narrowed in on social inequalities (Østby, 2008). Yet, country conflict has always been more complex than that – it is a product that incorporates both political, economic and social aspects.

Journal of Defense Analytics and Logistics
Vol. 7 No. 1, 2023
pp. 48-68
Emerald Publishing Limited
2399-6439
DOI 10.1108/JDAL-11-2022-0011

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 (Hegre *et al.*, 2013; Ahner *et al.*, 2015; Leiby, 2017). 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 culture concept while also forming regions to better model country conflict through 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.

2. 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 (Goldstone *et al.*, 2010). 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 (Goldstone *et al.*, 2010). 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 (Goldstone *et al.*, 2010). 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 (Hegre *et al.*, 2013). 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 (Hegre *et al.*, 2013). 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 (Hegre *et al.*, 2013).

A third example of regional modeling surfaced with the Boekestein logistic regression study, where his study investigated five different categories of a regional variable (Ahner *et al.*, 2015). The study concluded that a six-region categorization presented the best modeling 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 (Rosling, 2006). His examples, such as population vs fertility rates, or child survival vs 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 (Minkov and Hofstede, 2012). 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 the 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% (Ahner *et al.*, 2015).

Other works have improved upon Boekestein's research while maintaining the consistency of using the same six distinct regions for modeling (Leiby, 2017; Shallcross and Ahner, 2019). 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 (Shallcross and Ahner, 2019). 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 (Neumann *et al.*, 2022). This capitalized on Hegre's idea that the heterogeneity of the regions may change over time. Using a modified k-means algorithm, Neumann improved prediction accuracies by as much as 2.5% by redefining six United States Combatant Command regions using a combination of political, military, economic and social variables (Neumann *et al.*, 2022). 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 (Gupta *et al.*, 2002; Minkov and Hofstede, 2012). 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 (Gupta *et al.*, 2002).

Unresolved is a consensus on the number of cultural clusters, or regions, and how they should be formed. Neumann assumed six clusters using a mathematical approach based on k-means clustering that maintains consistency with the current number of US-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 six toward as many as 18 clusters, and applied a hierarchical mathematical approach settling on 11 global clusters (Ronen and Shenkar, 2013). 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 forecasts?

3. Methodology

The dataset contains variables on 173 United Nations' member countries whose population total exceeds 250 K 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.

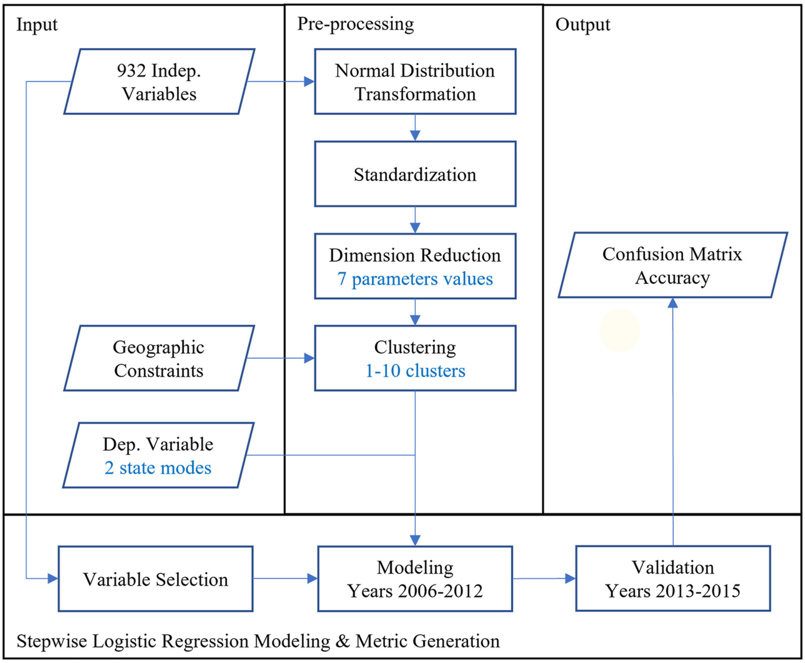
Feature extraction and clustering techniques require complete-case data, so observations with missing values must be discarded or estimated. List deletion of missing values seriously degrades the ability to detect effects of interest as various statistical estimates would be severely biased (van Buuren, 2018). 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 (Rubin, 1996). This study used MASS-impute, a type of multiple imputation, to complete the dataset (Leiby and Ahner, 2023). 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 for this investigational study employing a new approach to the region generation problem. However, the preliminary exploration of parameters for the number of PCs used all 30 datasets.

Consistent with Neumann, the methodology follows a process of 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 using the Tjur coefficient of determination to select appropriate independent variables for the regional models. This approach also expedites the modeling-building process in comparison to the seven-step purposeful selection of covariates approach found in (Hosmer *et al.*, 2013), as used by both Shallcross and Neumann, who 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 the highest level of conflict intensity score to each country according to a conflict means and conflict consequences approach (Heidelberg Institute for International Conflict Research (HIIK), 2020). HIIK is an alternative to other possible conflict variable databases such as the Correlates of War Project or the Uppsala Conflict Data Program, which focus on casualties per event year and delineations between inter- and intra-state conflict. One of the dependent variable variants, static-state, borrows from Boekestein (Ahner *et al.*, 2015), 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 (Shallcross and Ahner, 2019) and Neumann (Neumann *et al.*, 2022), 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 (Shallcross and Ahner, 2019). An overview of the new methodology is illustrated in Figure 1.

3.1 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 (Kriegel *et al.*, 2009). The four key issues when employing clustering techniques are typically referred to as the curse of dimensionality. The first issue revolves around the ratio of data elements (p) to observations (n), the general principle that when $p > n$, there are not enough simultaneous equations to solve for a solution. Kriegel noted that clustering enables "users to identify the functional dependencies resulting in the dataset," but



Source(s): Figure by authors

Figure 1.
Overview of methodology

as more variables are added, the complexity of the relationships increases making it difficult to visualize interesting insights (Kriegel *et al.*, 2009). 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” (Kriegel *et al.*, 2009). 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 (Kriegel *et al.*, 2009). The fourth issue dives into the redundancy of variables, thus artificially weighting distances, from a correlation perspective (Kriegel *et al.*, 2009). 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 (Beyer *et al.*, 1999). 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” (Beyer *et al.*, 1999). 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 (Beyer *et al.*, 1999).

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 by 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 (James *et al.*, 2013). Meanwhile, FA models the correlation structure of all variables to illuminate rotatable latent variables with associated factor loadings (Hastie *et al.*, 2009).

PCA assumes that the dataset is multivariate normal and has been standardized, so scaling is not a factor. Therefore, variables are first assessed for normal distribution and transformed as appropriate in the pre-processing stage. Standardization is also applied during pre-processing to alleviate any scaling biases between independent variable measures. FA assumes the dataset has no outliers, multicollinearity is manageable, and there is no homoscedasticity between variables. Management of the assumptions was 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 seek 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 (James *et al.*, 2013). 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 (Cangelosi and Goriely, 2007).

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% and 95% (Cangelosi and Goriely, 2007). The broken-stick model, presented by MacArthur during a bird study, compares eigenvalues against an apportioned resource distribution (Cangelosi and Goriely, 2007). The distribution follows (Equation 1), 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 lower 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 (Cangelosi and Goriely, 2007).

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

Equation (1): Broken-stick distribution

3.2 Clustering and geography

Two objectives motivate developing regions. The first objective seeks to apply mathematical rigor to the prediction models where studies (Ahner *et al.*, 2015; Leiby, 2017) 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 applications 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) (Neumann *et al.*, 2022). K-means finds a local optimum influenced by a specified initial assignment of countries to clusters. This constrains comparisons between different numbers of clusters and infers that there is no consistency between observing countries within, for example, a 6-cluster solution to a 7-cluster solution. First, because local and global optima are not the same, initial assignment matters. Second, changes in the number of regions influence associations between countries due to the mechanism of the within-cluster variation vs the without cluster variation. Any major shifts between country associations then are hard to explain when comparing different k solutions. 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 modified 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 (Neumann *et al.*, 2022). 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 100 km of each other. For island nations further than 100 km 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 USA 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 yet another modified k-means approach. Unlike k-means, where observations are initially (sometimes randomly) assigned one of (predefined k) k-number of clusters, 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.

3.3 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 and 1. Once the data meet 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. Independent variables are assessed for increasing the model accuracy as evaluated through the Tjur coefficient of determination. 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. The selection of independent variables comes from the pre-transformed datasets, where the model is limited to a maximum of 10 variables to curb overfit. 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 does not use maximizing the likelihood function, which coincidentally is also what logistic regression uses to develop model coefficients, is the Tjur statistic (Allison, 2013). Tjur saw similarities between graphically comparing differences in two "parallel histograms" and the graphical check of the Hosmer–Lemeshow test (Tjur, 2009). 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 (Tjur, 2009). In previous research, an AUC-ROC was considered among goodness measures which value both positive and negative cases equally but suffer when applied to unbalanced data. For imbalanced data, AUPRC is heralded as a superior metric, but only when one case class is deemed important (Baillie *et al.*, 2021). The Tjur statistic provides an alternative to the area-under-the-curve debate. The Tjur statistic, as seen in Equation (2), identifies statistically significant variables for the models, where $\widehat{\pi}_{i1}$ and $\widehat{\pi}_{j0}$ denote the fitted values for successes and failures, respectively, of N true successes and M true failures, for the binary outcomes of logistic regression.

$$D = \frac{\sum_{i=1}^{N} \widehat{\pi}_{i1}}{N} - \frac{\sum_{j=1}^{M} \widehat{\pi}_{j0}}{M} \tag{2}$$

Equation (2): Tjur coefficient of discrimination

Accuracy from the confusion matrix quantifies the predictive power of the models. 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.

4. 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 (the year 2012), the following three years of naïve predictions would be accurate at 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% (Goldstone *et al.*, 2010; Ahner *et al.*, 2015; Leiby, 2017; Shallcross and Ahner, 2019), which this research achieved in the majority of models.

4.1 Pre-processing results

A Box–Cox transformation was applied to each variable to optimize the 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 assessed for less than 0.5 of a linear transformation, which means basically no transformation is required at all to assume normal.

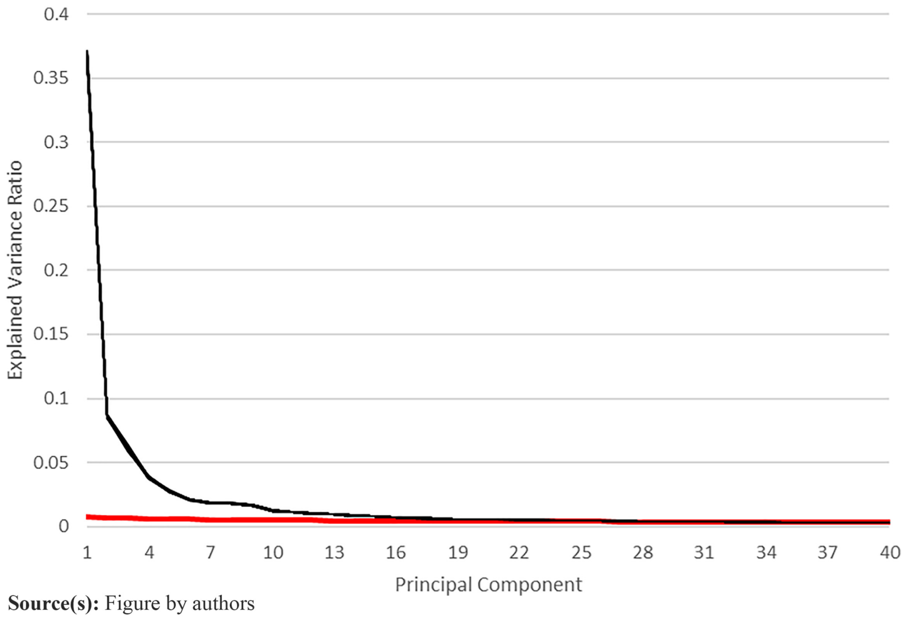
PCA demonstrated superiority over FA for the dataset. After optimizing the normality of the data through Box–Cox transformations and standardizing the data, the explained variance after 15 variables for PCA was 71.3%, whereas FA was only 54.8%. The first principal component explained 36.6% of the variance, 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 and 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 (Cangelosi and Goriely, 2007), 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 a 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 2 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 (Cangelosi and Goriely, 2007). Again, the recommendations were made on much smaller dimension sizes with the example examining only nine variables (Cangelosi and Goriely, 2007) compared to our over 900 variables.

The log-eigenvalue diagram, as illustrated in Figure 3, 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 strictly, a minimum of 10 components could possibly suffice.

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 (Franklin *et al.*, 1995). The percentage of variance explained after six components is only 60.02%, as seen in Figure 4, which does not meet the window of explained variance desired – between 70% and 95%. It is 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.



Source(s): Figure by authors

Figure 2.
Broken-stick model

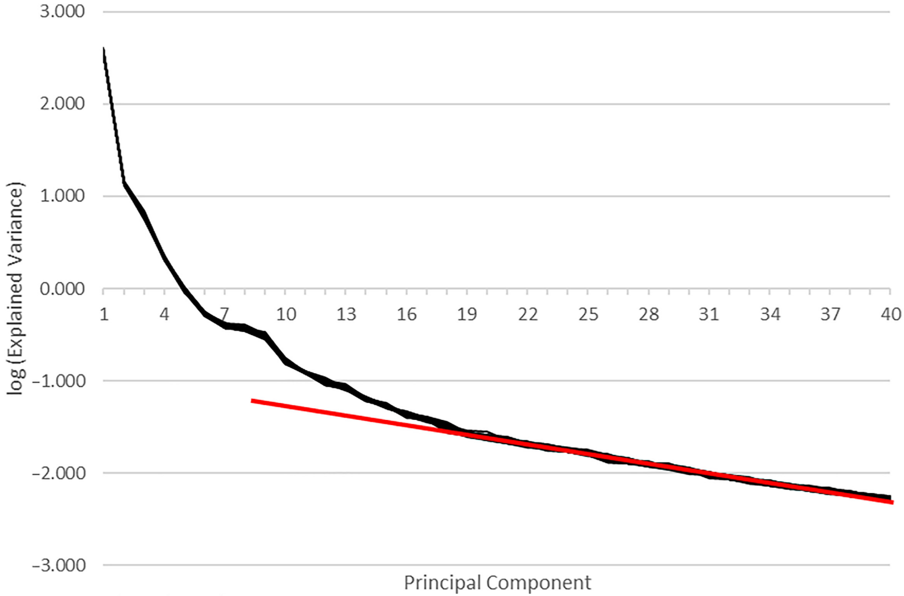


Figure 3.
Log-eigenvalue diagram

Source(s): Figure by authors

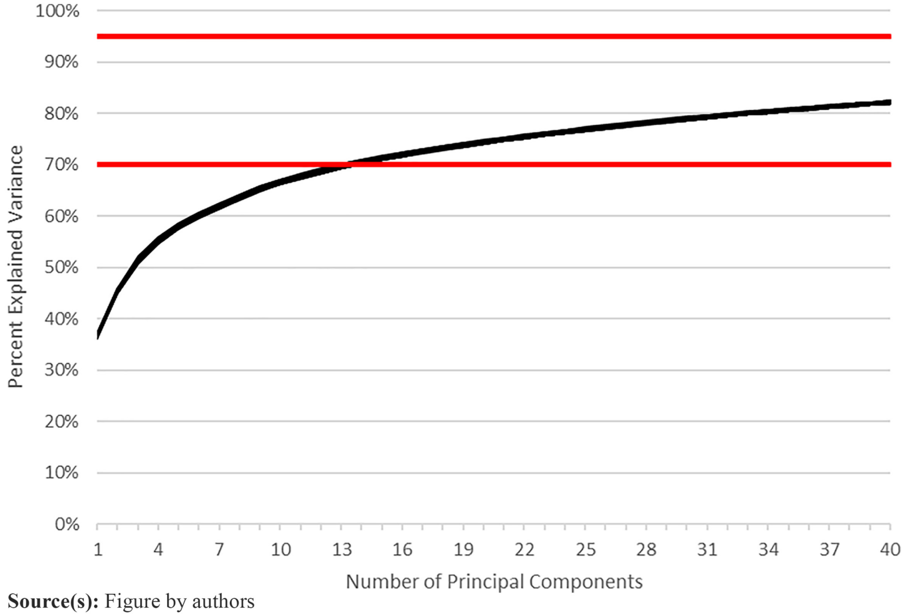


Figure 4.
Percent explained variance

Source(s): Figure by authors

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 1. Comparing PC description names between this study and Neumann, only the

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*

Source(s): Table by authors

Table 1. Principal components descriptions and variance

"unemployment" PC explicitly stands out in the comparison. However, other PCs were implicitly similar to 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 (Neumann, 2018). 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 (60%), whereas Neumann's dataset contained only 4 (13%). This may also explain why Rosling's regions worked well when combining countries together, like the Organizations for Economic Co-operation and Development.

4.2 Modeling and 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). Accuracy results for all combinations are in Figure A1. For all model types given the available data, the clustering parameter had more influence on the predictive outcome than the PCA parameter. 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 Figure A2. Choosing between different numbers of PCs resulted in a maximum difference of only 2.7% predictive accuracy, suggesting that more PCs, for the regression models, explored and the available variables in the dataset, may add little value. In fact, adding 18 or more PCs saw decreases in predictive accuracy confirming the curse of dimensionality with clustering. Referencing the charts in Figure A1, all the validation results share similar patterns except for using six PCs in the SSNC model type. All models demonstrated severe diminishing returns for average validation accuracy when increasing the number of clusters, whereas the SSNC model type with six 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 5: 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 four 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 nine 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 by maintaining larger observation sizes per region cluster.

4.3 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 (Shallcross and Ahner, 2019) and Neumann (Neumann *et al.*, 2022) studies.

Shallcross proposed that using the dependent variable transition-state would increase the accuracy of the models (Shallcross and Ahner, 2019), with Neumann demonstrating a comparison between the Shallcross transition-state study and the Boekestein static-state study increasing by 6% (Neumann *et al.*, 2022). Although the gains in this study are not as pronounced, the weighted training accuracy as observed in Figure 5 demonstrated the

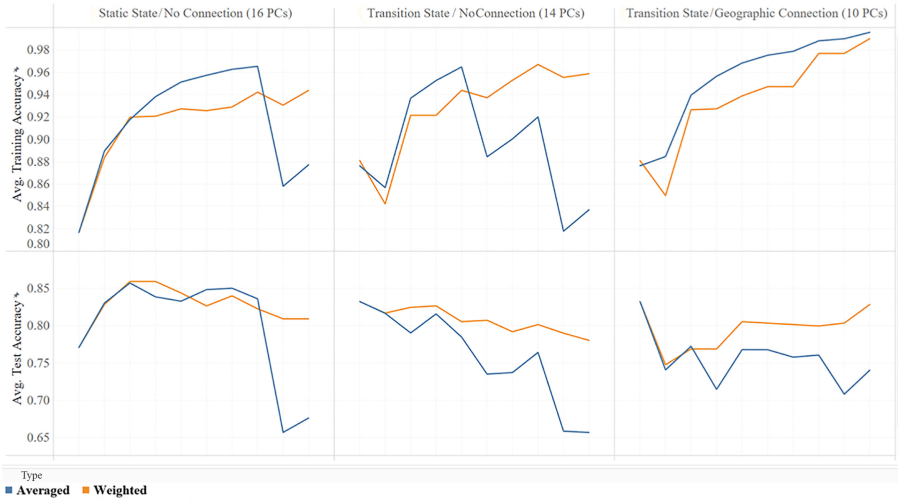


Figure 5.
Model type's accuracy across clusters for best PCA parameter

Source(s): Figure by authors

potential for better models using the transition-state dependent variable, especially when employing over five 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 (Shallcross and Ahner, 2019). Recalling the method setup, the validation of the data is considered a three-year period. However, as seen in Table 2, years trained has an impact on the prediction of subsequent years. Using 9 years of training data (2006–2014) increases the variation to the dataset leading to lower training accuracies, however, the inclusion of two additional years (2006–2012 vs 2006–2014) increases the validation prediction by 2%. 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.

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 (Pedregosa *et al.*, 2011). 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 6.

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 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

		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

Note(s): T. – Training, V. – Validation
* TSGC Model with 6 Regions
Source(s): Table by authors

Table 2.
Global accuracy for different validation periods

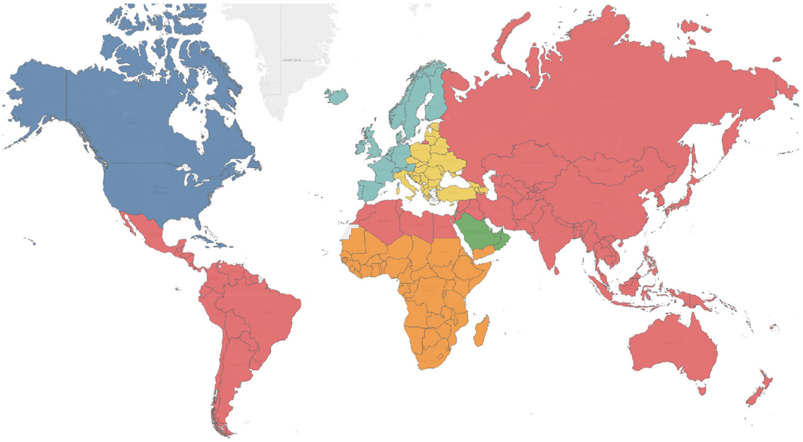


Figure 6.
6-Cluster TSGC regional map

Source(s): Figure by authors

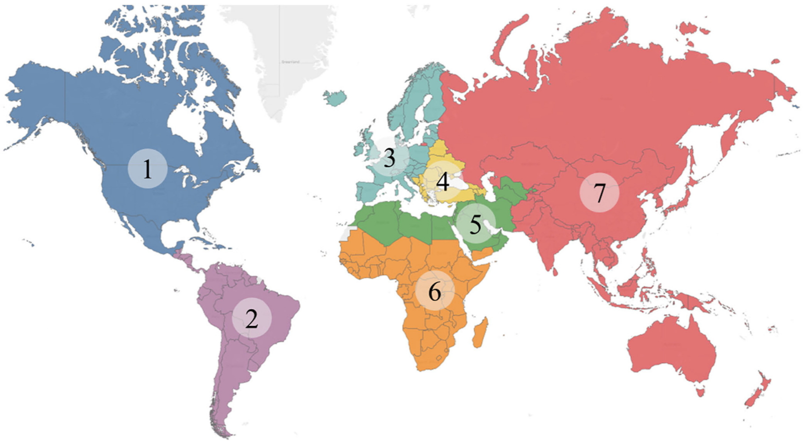
likely" trees refer to the fusion of observations through dimensional Euclidean differences rather than the geographic connection constraint. Normally when viewing a hierarchical clustering dendrogram, all tree branches would spread upward in the same direction, but using a connection constraint, some 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. As a 1-region global model, this occurs 15% of the time, but the distribution of transitions is not observed equally as the world is divided into regions. 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 7. The model incorporated the three gained insights: transition-state dependent variable combined with more than six 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 3. 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.

5. 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



Source(s): Figure by authors

Figure 7. Modified 7-cluster transition-state regional map

			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

Note(s): Training Years (2006–2014), Validation Year (2015)
* TSGC Model with seven Modified Regions
Source(s): Table by authors

Table 3. Modified 7-cluster TSGC regional results

indicators alone. Furthermore, maintaining contiguous modeling regions assists decision-makers with distributing resources and aid.

This study challenged two assumptions from Neumann producing insights otherwise left unknown in prior research. The first challenge is the k-means approach, which assumes a pre-defined number of regions. The second challenge is 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 in constructing country conflict models. This research explored an automated, data-driven framework to increase country conflict-state predictive accuracy one to three years into the future. Although other metrics quantify the statistical viability of a model, predictive accuracy provides 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 into regions by 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 that 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.

Several obstacles still remain 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 in 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 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 the 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 seven-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.

References

Ahner, D., Boekestein, B. and Deckro, R. (2015), *A Predictive Model of World Conflict Using Open Source Data*, Air Force Institute of Technology, Wright-Patterson AFB, Ohio.

Allison, P. (2013), "What's the best R-squared for logistic regression", *Statistical Horizons*, Vol. 13, available at: https://statisticalhorizons.com/r2logistic/

Baillie, E., Howe, P.D.L., Perfors, A., Miller, T., Kashima, Y. and Beger, A. (2021), "Explainable models for forecasting the emergence of political instability", *PLOS ONE*, Vol. 16 No. 7, pp. 1-18, doi: 10.1371/journal.pone.0254350.

Beyer, K., Goldstein, J., Ramakrishnan, R. and Shaft, U. (1999), "When is 'nearest neighbor' meaningful?", *International Conference on Database Theory*, Springer, Berlin, Heidelberg, pp. 217-235.

Brito, D.L. and Intriligator, M.D. (1985), "Conflict, war, and redistribution", *The American Political Science Review*, Vol. 79 No. 4, pp. 943-957.

Cangelosi, R. and Goriely, A. (2007), "Component retention in principal component analysis with application to cDNA microarray data", *Biology Direct*, Vol. 2 No. 2, pp. 1-21, doi: 10.1186/1745-6150-2-2.

Franklin, S.B., Gibson, D.J., Robertson, P.A., Pohlmann, J.T. and Fralish, J.S. (1995), "Parallel analysis: a method for determining significant principal components", *Journal of Vegetation Science*, Vol. 6 No. 1, pp. 99-106, doi: 10.2307/3236261.

Gartzke, E., Li, Q. and Boehmer, C. (2001), "Investing in the peace: economic interdependence and international conflict", *International Organization*, Vol. 55 No. 2, pp. 391-438, doi: 10.1162/00208180151140612.

Goldstone, J.A., Bates, R.H., Epstein, D.L., Gurr, T.R., Lustik, M.B., Marshall, M.G., Ulfelder, J. and Woodward, M. (2010), "A global model for forecasting political instability", *American Journal of Political Science*, Vol. 54 No. 1, pp. 190-208, doi: 10.1111/j.1540-5907.2009.00426.x.

Gupta, V., Hanges, P.J. and Dorfman, P. (2002), "Cultural clusters: methodology and findings", *Journal of World Business*, Vol. 37 No. 1, pp. 11-15, doi: 10.1016/S1090-9516(01)00070-0.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York.

Hegre, H., Karlsen, J., Nygård, H.M., Strand, H. and Urdal, H. (2013), "Predicting armed conflict, 2010-2050", *International Studies Quarterly*, Vol. 57 No. 2, pp. 250-270, doi: 10.1111/isqu.12007.

Heidelberg Institute for International Conflict Research (HIIK) (2020), "Conflict barometer 2019", Heidelberg.

Hosmer, D.W., Jr, Lemeshow, S. and Sturdivant, R.X. (2013), *Applied Logistic Regression*, 3rd ed., John Wiley & Sons, Hoboken, NJ.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An Introduction to Statistical Learning with Applications in R*, Springer, New York.

Kriegel, H.-P., Kröger, P. and Zimek, A. (2009), "Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering", *ACM Transactions on Knowledge Discovery from Data*, Vol. 3 No. 1, doi: 10.1145/1497577.1497578.

Leiby, B.D. (2017), *A Conditional Logistic Regression Predictive Model of World Conflict Considering Neighboring Conflict and Environmental Security*, Air Force Institute of Technology, available at: https://apps.dtic.mil/sti/citations/AD1055147

Leiby, B.D. and Ahner, D.K. (2023), "Multicollinearity applied stepwise stochastic imputation: a large dataset imputation through correlation-based regression", *Journal of Big Data*, Vol. 10 No. 1, p. 23, doi: 10.1186/s40537-023-00698-4.

Minkov, M. and Hofstede, G. (2012), "Is national culture a meaningful concept? Cultural values delineate homogeneous national clusters of in-country regions", *Cross-Cultural Research*, Vol. 46 No. 2, pp. 133-159, doi: 10.1177/1069397111427262.

Neumann, S. (2018), *Forecasting Country Conflict within Modified Combatant Command Regions Using Statistical Learning Methods*, Air Force Institute of Technology, available at: https://scholar.afit.edu/etd/1854/

Neumann, S., Ahner, D. and Hill, R. (2022), “Forecasting country conflict using statistical learning methods”, *Journal of Defense Analytics and Logistics*, Vol. 6 No. 1, pp. 59-72, doi: 10.1108/jdal-10-2021-0014.

Østby, G. (2008), “Inequalities, the political environment and civil conflict: evidence from 55 developing countries”, in Stewart, F. (Ed.), *Horizontal Inequalities and Conflict: Understanding Group Violence in Multiethnic Societies*, Palgrave Macmillan, London, pp. 136-159, doi: 10.1057/9780230582729_7.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É. (2011), “Scikit-learn: machine learning in python”, *Journal of Machine Learning Research*, Vol. 12 No. 85, pp. 2825-2830.

Ronen, S. and Shenkar, O. (2013), “Mapping world cultures: cluster formation, sources and implications”, *Journal of International Business Studies*, Vol. 44 No. 9, pp. 867-897, doi: 10.1057/jibs.2013.42.

Rosling, H. (2006), “The best stats you’ve ever seen”, *TED Conferences*, Monterey, California.

Rubin, D.B. (1996), “Multiple imputation after 18+ years”, *Journal of the American Statistical Association*, Vol. 91 No. 434, pp. 473-489, doi: 10.1080/01621459.1996.10476908.

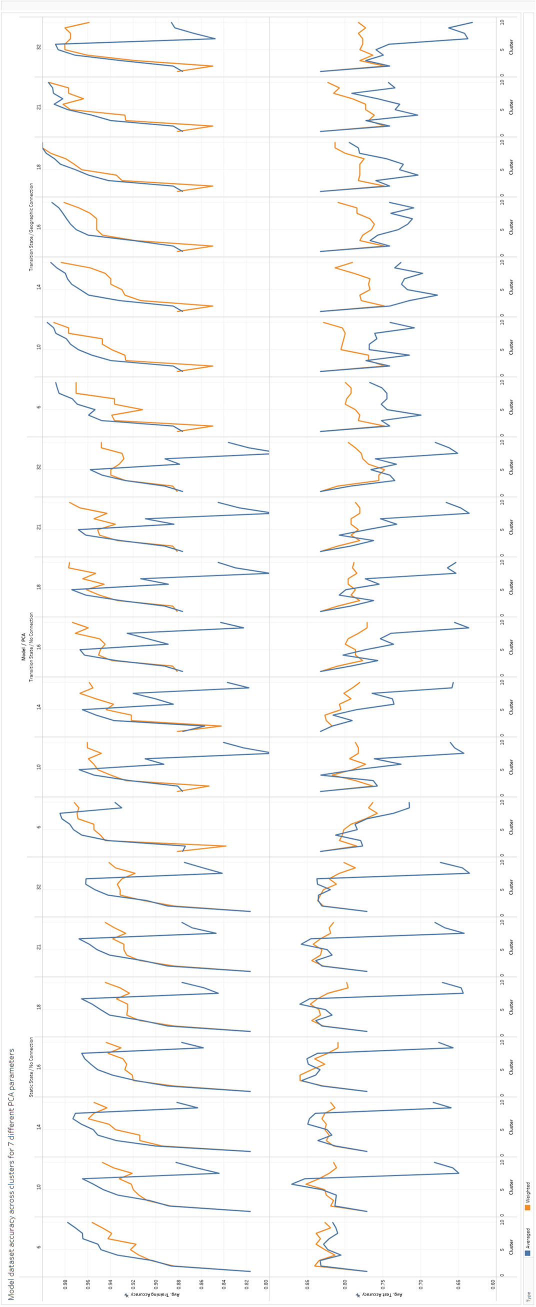
Shallcross, N. and Ahner, D. (2019), “Predictive models of world conflict: accounting for regional and conflict-state differences”, *The Journal of Defense Modeling and Simulation*, Vol. 17 No. 3, pp. 243-267, doi: 10.1177/1548512919847532.

Tjur, T. (2009), “Coefficients of determination in logistic regression models – a new proposal: the coefficient of discrimination”, *American Statistician*, Vol. 63 No. 4, pp. 366-372, doi: 10.1198/tast.2009.08210.

van Buuren, S. (2018), *Flexible Imputation of Missing Data*, 2nd ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, available at: https://stefvanbuuren.name/fimd/

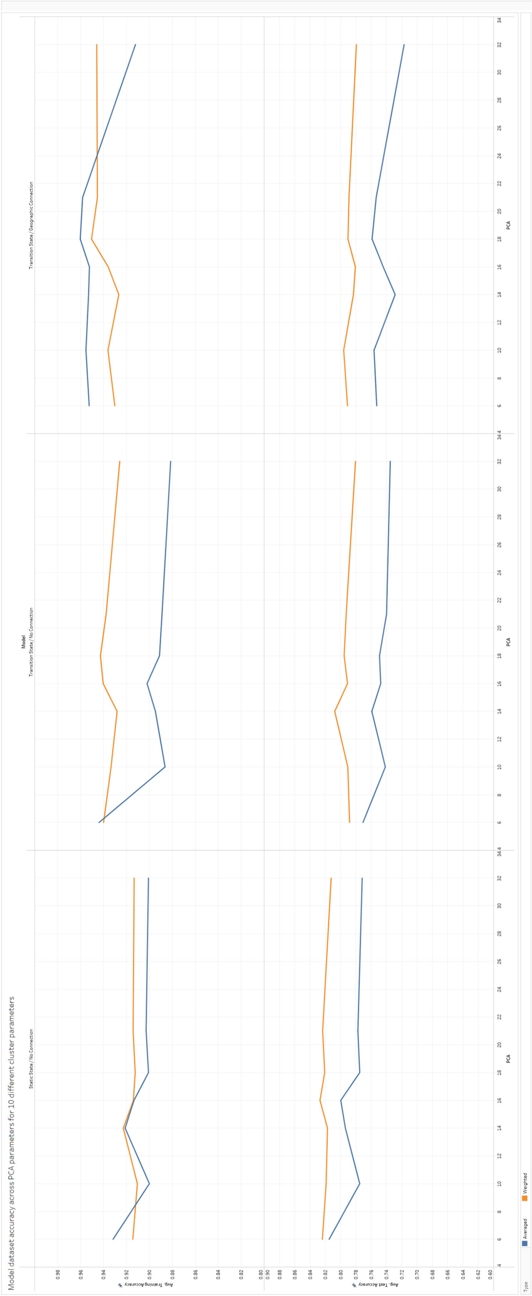
Corresponding author
Benjamin Leiby can be contacted at: benjamin.leiby@us.af.mi

Appendix



Source(s): Figure by authors

Figure A1.
Training and test model accuracies given cluster and PCA pairings



Source(s): Figure by authors

Figure A2.
Training and test model accuracies given PCA parameter averaging across clusters