Predictive Analytics: Analyzing Complex Threats for Operations and Readiness (ACTOR)

College of Charleston

Courtney Beckham

Honors College
College of Charleston
Spring 2018

Contents

1 Abstract

This paper explores the accuracy of the Analyzing Complex Threats and Operations Readiness (ACTOR) model employed by the United States government and theoretical approaches to surmount the challenges quantitative predictive models pose when used to predict qualitative information. The ACTOR model, also known as a "second image" model, views conflicts as arising from "factors and failings internal to states."[1] In accordance with this belief, the ACTOR model aggregates data on 12 macro-structural factors from 159 countries over the period 1975-1999 to predict the level of intensity a country will experience instability. This paper will explore the accuracy of such findings as well as the future directions and developing technologies that can be applied in accordance with later projects employed by the United States such as Integrated Crisis Early Warning Systems (ICEWS).

2 Introduction

Using technological advances in conjunction with mathematical processes gives way to new and exciting methods used to ensure a country's security. One such way this partnering has manifested is in the creation of the ACTOR model used by the United States government. The goal of ACTOR is to identify key macro-structural factors that contribute to country instability. Through the use of predictive analytics and technological capabilities on attaining and storing this macro-structural data, predictions are able to be made with up to 80% accuracy in predicting country instability. However, as the intellectual field of predictive analytics grows, this accuracy can be improved. This paper discusses the current mathematics and computational procedures behind the current ACTOR model as well as theoretical developments both in mathematics and computer science that could lead to greater accuracy for these findings, as well as methods to present the ACTOR predictions to analysts.

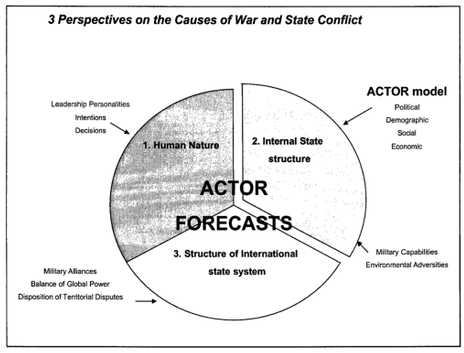


Figure 1: This figure illustrates the three different models for country instability, with ACTOR classified as a second image model.[1]

This paper discusses the accuracy of the ACTOR model findings based on the same data sources used to aggregate data for the creation of the ACTOR model. The ACTOR model gathered data on 12 macro-structural factors for 159 over 1975-1999. These macro-structural factors are outlined in Table 1.

Table 1: The table below lists all of the macro-structural factors considered in the ACTOR model.[1]

Macro-Structural Factors	
Percent of history spent in conflict	Infant Mortality Rate
Trade Openness	Youth Bulge
Civil Liberties Index	Life Expectancy
Political Rights Index	Democracy
Religious Diversity	Caloric Intake
GDP per Capita	Ethnic Diversity

It also discusses theoretical approaches to solve various issues with the ACTOR problem, that will later be tested on the expansion of the ACTOR model known as the Integrated Crisis Early Warning System (ICEWS) that is employed by the United States government.

3 Country Instability: The Predictions

First, here are a few examples of what predictions from ACTOR look like and an examination of the accuracy of the results. This offers a glimpse of what factors the ACTOR model takes into consideration, and the type of information that analysts may receive. The section entitled "Predictions for Improvement" offers a look at three different countries, Bangladesh, Iran, and Israel, that present their own unique backgrounds and current situations to the

Table 2: The following table provides insight into what classifies as low, moderate, and high instability for a country.

3 Levels of Instability Intensity			
Instability Level	**Conflict Type**	**Examples**	**Definition**
High intensity	War/violent crisis	WWII	Systematic use of force
Moderate intensity	Violent crisis	Ethnic conflict in Bosnia	'War-in-sight' crises
None/low intensity	None/crisis	Possession of strategic weapons	Mostly non-violent

ACTOR model. Bangladesh has one of the highest populations in the world, Iran has a unique history of conflict, and Israel experienced a high amount of conflict at the end of the initial ACTOR forecasting period consisting of 2000-2015. The section entitled "Predictions for Decline" takes a look at three different false predictions from ACTOR and what variables contributed the most to these false alarms.

3.1 Predictions for Improvement

Case Study: Bangladesh As a country with one of the highest populations in the world, Bangladesh presents its own unique situation to the ACTOR model. ACTOR predicted that over the period of 2001-2015, Bangladesh would have an improved infant mortality rate (from 67 to 37), as well as an improved GDP per capita, and an improved average life expectancy (from 61 to 68). The ACTOR model also predicted that the youth bulge in Bangladesh would decline and that trade openness would increase by 50 percent. Here is a breakdown of how accurate each of these predictions were:

- The infant mortality rate in Bangladesh in 2015 was 34 per 1,000 live births. Not only did the ACTOR model correctly predict that the infant mortality rate would improve, they were pretty close to the exact figure. The Bangladesh infant mortality rate hit 37 a few years earlier than expected, reaching 37 deaths per 1,000 births by the year 2013.[2]

- The GDP per capita in Bangladesh in 2015 was 1,210.2, as compared to 402.6 in 2001. This was a correct prediction for ACTOR.[3]

- The average life expectancy in Bangladesh in 2015 was 73 years and ACTOR predicted that the average life expectancy would be approximately 68 years in 2015, a conservative estimate[2].

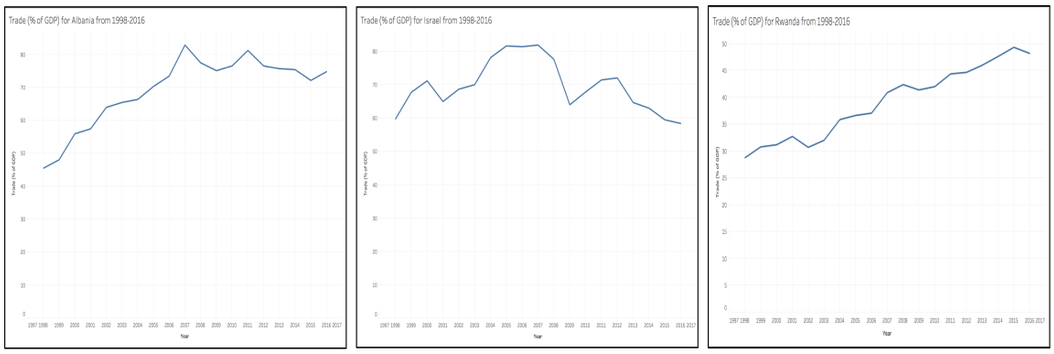
Case Study: Iran Iran presents its own unique history of conflict to the ACTOR model.

- ACTOR predicted that the infant mortality rate in Iran would decline from 28 deaths per 1,000 births to 14 deaths per 1,000 births in 2015. However, the rate leveled off at 17 deaths per 1,000 births in the year 2013, and stayed constant through the year 2015, never reaching the projected 14 deaths.

- Youth bulge declines (1.29 to .78). ACTOR measured the youth bulge factor by taking the ratio of the population aged 15-29 to those aged 30-54. The youth bulge in Iran was expected to decline from 2001 to 2015. The population of those aged 15-29 was 22,020,020, and the population of those aged 30-54 was 29,204,969. Thus the youth bulge was 0.754, or slightly better than the ACTOR model predicted. [2]

Case Study: Israel Israel presents it's own unique situation to the ACTOR model because it experienced a high level of conflict near the end of the ACTOR forecasting period. This case study illustrates the accuracy of the ACTOR model. The model accurately predicted two-thirds of the forecast for Israel, and the incorrect prediction may be due to sudden conflict experienced near the end of the ACTOR forecasting period.

- GDP per capita improves. This prediction from ACTOR was correct, with the GDP per capita improving from 1,891 in 2001 to 4,862 in 2015, measured in current US dollars. [3]



Scheme 1: These figures illustrate the trade openness measure over the years from 1998 to 2015 for (left to right) Albania, Israel, and Rwanda.[3]

- Youth bulge declines slightly. The population of people aged 15-29 in Israel in 2015 was 1,824,604, whereas the population of those aged 30-54 was 2,411,794. Thus the ratio is 0.757 in 2015, compared to .866 in 1997. Thus ACTOR was yet again correct with their prediction. [2]
- Trade openness improves. In 1999, the trade openness measure of Israel was 67.8, whereas in 2015 the trade openness had declined to 59.5. This was a drastically incorrect prediction for ACTOR. However, as noted before, Israel did undergo higher levels of conflict near the end of the forecasting period, and for half of the forecasting period, the trade openness measure was improving.[3]

3.2 Predictions for Decline

Case Study: Rwanda Rwanda is an example of a country that was a complete false positive prediction from the ACTOR model.

- The life expectancy in Rwanda was projected to decrease by the ACTOR model, but the life expectancy actually increased an exceptional amount throughout the period from 2001-2015. In 1999 the life expectancy was 50 years, whereas in 2015 the life expectancy had increased to 64 years.[2]
- Trade openness was expected to decline over the period from 2001-2015 for the country of Rwanda, but it has actually steadily increased over the years, as illustrated in Scheme 1.[3]

Case Study: Albania

- ACTOR predicted that the GDP per capita would decline for Albania, but it has actually improved from 1,098 in 1999 to 3,935 in 2015. This was yet another false positive for the ACTOR model.
- Trade openness in Albania was predicted to decline throughout the predicted period, but increased as well. [3]

Case Study: Djibouti

- Youth bulge was projected to increase, from 1.19198467 in 1999, but in 2015 was 1.13909113. [2]
- The trade openness of Djibouti was predicted to decline, but the last recorded data the World Bank Database has available is from 2007, which had a much higher value compared to years prior. [3]

4 FASE: Fuzzy Analysis of Statistical Evidence

General Overview

Fuzzy Analysis of Statistical Evidence, also known as FASE, uses fuzzy set and statistical theory for solving problems of pattern recognition and classification. The aim of FASE is to

mimic how human judgment operates. ACTOR claims that this is the advantage of using the FASE technique over a similar method like Bayesian classifiers.

Methodology

As illustrated in the paper by Yuan Yan Chen on the FASE method, FASE operates with the current model definition

$$Pos(C|A_1, ..., A_n) = Pr(A_1, ..., A_n|C)/sup_c Pr((A_1, ..., A_n|C)$$

where C is the class variable and $A_1, ..., A_n$ are the attributes variable. Pos is the possibility measure. Thus, this equation uses the "fuzzy membership that an instance belongs to class C" and the belief measure that an instance belongs to class C. This is only slightly different than the traditional Bayes formula. This formula simply has a difference normalization constant. FASE defines this as being where "in possibility the sup norm is 1, while in probability measure the additive norm is 1."[4]

Advantages and Disadvantages

This method offers several advantages, but also has its own drawbacks. This method doesn't highlight the relations between specific attributes, which would be valuable information for analysts to receive because it would provide a way to get a more targeted look at specific variables and their relation to country instability. On the positive side, this method weighs the attributes, which is a crucial aspect of being able to increase predictive model accuracies. FASE is also noise tolerant, but it isn't noise tolerant to the point that it learns the noise and details of the dataset and accidentally skews the predictive findings. This would occur if the FASE model was overfit to the training data set.

FASE eliminates less plausible hypotheses based on evidences. Thus, the model is good at learning certain rules of the data. FASE also adequately deals with the sparseness of the ACTOR model dataset, because classification is based on the lesser number of attributes. However, there are specific methods that could be used to present this information in a clear manner to analysts, as well as give analysts a more honed-in view on specific attributes and countries.

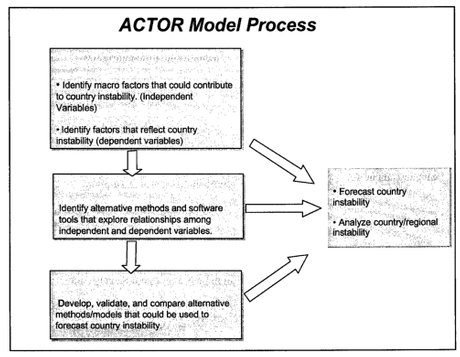


Figure 2: This figure illustrates the general process of the ACTOR model.[1]

5 Limitations of the ACTOR Model

ACTOR does not claim to be an all-encompassing model, but rather is a second image model that is meant to provide specific insight into the structural elements within a country that effects its instability. Thus, there is very little risk in using ACTOR in conjunction with other models such as first and third image models, as well as with the discretion of expert analysts.

However, the question of what the Army wants to find with this technology is crucial. The accuracy of the model could certainly be improved if the data was preprocessed by removing some variables. But, this is problematic if the Army wants an all-encompassing picture of these 12 macro-structural factors for all 159 countries. There is a trade-off between accuracy of predictions and the scope of these predictions. Due to this being an extremely large dataset with 12 variables serving as predictors, it creates a more complex issue in relation to prediction accuracy. Below are three theoretical methods that could add to the ACTOR model capabilities. The first model is a different approach to modeling this problem, the second provides a concise look at which countries are predicted to experience the most instability, and the third allows user manipulation to calculate how much a specific shifting macro-structural element will affect a unique country.

6 Theoretical Extensions and Updates to ACTOR

6.1 Theory 1: Decision Tree Approach

General Methodology

A decision tree is "a machine learning algorithm that partitions the data into subsets." When approaching a predictive modeling problem, decision trees can serve to "discover features and extract patterns in large databases that are important for...predictive modeling." In approaching the large dataset that the ACTOR model is working with, finding patterns in the data is imperative to proceeding with the model. More particularly, finding a set of decision rules within the data that serve to "provide an informative and robust hierarchal classification model" is one of the first steps. [5] To begin, the subsets of the data must be determined. Specific variables will be split at a location and two determinations will be formed: the predictor variable used for the split, and the set of values for the predictor variable. To aid in possible partitioning values, an equation for information gain is used. The higher the value of the information gain variable, the better a split. The information gain is determined by the following equation

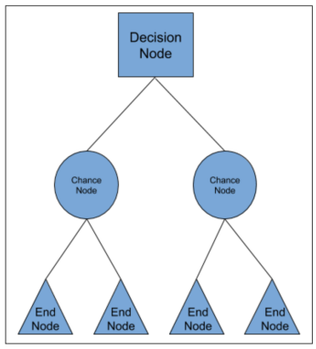


Figure 3: This figure illustrates a typical decision tree, where the square is the decision node, the circles are the chance nodes, and the triangles are the end nodes.

$$Info = -\sum \left(\frac{N_j(t)}{N(t)}\right) log_2 \left(\frac{N_j(t)}{N(t)}\right)$$

where N_j is the number of samples in class j, $N(t)$ is the number of samples in node t, and $N_j(t)$ is the number of class j samples in node t. Due to this model being a classification model, using the Gini Index value (also known as the Gini impurity value) will help determine where to split variables. In order to calculate the Gini Index, we must first calculate the impurity of a class via the following equation:

$$impurity = 1 - \sum \|p(j)N_j(t)/N_j\|^2$$

This equation is then used to calculate the Gini Index

$$Gini = impurity(Parent) - \sum (p_k) impurity(Child_k)$$

where $p(j)$ is the probability that a sample belongs to class j, and $\|g\|$ is the normalization of the vector g to the unit length. [6]

Advantages and Disadvantages

Decision trees in predictive modeling afford several advantages. Decision trees implicitly perform feature selection, which determines which variables offer the maximum amount of information gain. It also doesn't require an overwhelming amount of preprocessing of the data, which means that a decision tree scales extremely well, but still doesn't have extremely high computation costs. Also, nonlinear relationships do not affect the model performance, which is not true when using regression models. Lastly, the best feature of decision trees is their easy interpretability. This quality may be the most advantageous to the specific uses of the ACTOR model, because it creates an easy way to clearly communicate ideas and predictions to employees in less technical fields. Decision trees also create a range of possible outcomes, which allows the information to be used in conjunction with other processes to provide analysts with several likely outcomes. On the flip side, decision trees are "based on expectations" [7] which can lead to errors in the tree. Also, small changes in the input data can cause large changes in the decision tree. [8]

6.2 Theory 2: Nearest Neighbor Outlier Detection

General Overview

The main goal of the ACTOR project is to predict which countries are going to experience the highest amount of instability in the future, because these countries pose the biggest threat

to U.S. prosperity and security. Considering that this is the main goal, it may be superfluous to need information on the state of all 159 countries. Instead, it may suffice for the Army to focus on the major outliers in the dataset: those who are highly unstable. Instead of using this to replace the ACTOR model, this may present a more easily interpretable and communicable output that includes the countries experiencing the extremes of instability. This technique would be used best after the initial predictions for the ACTOR model are made.

This method centers on the assumption that "instances of normal data occur in dense neighborhoods, while outliers occur far away from their closest neighbors." [9] This assumption is consistent with the dataset used by the ACTOR model, and thus this model can provide concise information on outliers, that will aid users of the ACTOR model in clearly conveying information to expert analysts.

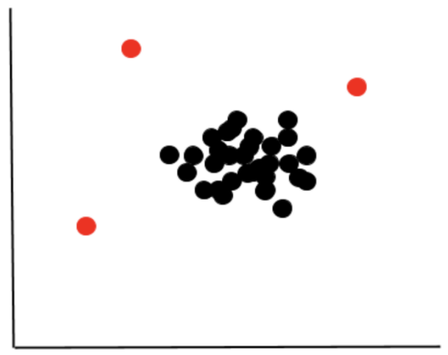


Figure 4: This figure is a simple illustration of what outliers in this data would look like. When comparing two variables, one on the x axis, another on the y axis, the red dots represent countries that vary the most from others in relation to x and y variables.

Methodology

The nearest neighbor based outlier detection technique requires finding which instances are most dissimilar. For the continuous variables that are used in the ACTOR model, Euclidean distance would be used in measuring dissimiliarity. The Euclidean distance is calculated using the following formula

$$d = \sqrt{\sum_{i=1}^{v} (p_{1i} - p_{2i})^2}$$

where p_{1i} and p_{2i} are two data instances that are summed for v variables. For the categorical attributes used in the ACTOR model, a simple matching coefficient would be used to measure dissimilarity.

To find an outlier score for a data instance (in this case a country) is "its distance to its k^{th} nearest neighbor in a given data set." [9] Then, the overall outlier score for a country would be it's sum of distances from its k nearest neighbors. This approach operates under the assumption that most countries are at a state of stability, and those that deviate most from stable countries are experiencing the most instability.

Advantages and Disadvantages

This method allows clear communication of information to expert analysts, which is advantageous. In communicating highly technical processes to less technical experts, nearest neighbor outliers gives each country a score that reflects how much they deviate from other countries in terms of stability. This allows expert analysts to receive a concise report with only the most important information they need. Instead of parsing through information for 159 countries, this method allows the most concise and polished presentation of the information that the ACTOR model calculates. However, this approach does operate under the idea that most countries are at a state of stability, so there could be issues with the model when there is a high amount of countries at what would be a seemingly unstable state for a certain variable.

6.3 Theory 3: Multinomial Logistic Regression

General Overview

Logistic regression is a new way of looking at this problem of categorizing and predicting country instability. Logistic regression is a "statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome."[10] In the case of traditional logistic regression, there must only be two possible outcomes. However, the extension to logistic regression known as multinomial logistic regression allows for there to be two or more outcomes. The goal of this modeling technique is to find the best model that

accurately describes the relationship between the dependent variable (i.e. level of stability) and the independent variables (i.e. the 12 macro-structural factors).

Methodology

Multinomial logistic regression uses the following regression equation

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_kX_k$$

where $b_0, b_1, b_2, ..., b_k$ are the regression coefficients of the independent variables. This presents several advantages, because it allows unimportant variables to be identified. Out of the 12 macro-structural factors, there may be some that are less important than others and multinomial logistic regression identifies those through the coefficients. Also, as an extension, discriminant analysis could be used to evaluate the significance of the independent variables. However, simply using the regression coefficients in this case will suffice. If the regression coefficient is not significantly different from 0 ($P > 0.05$), then the variable can be removed from the regression model. However, if the $P < 0.05$, then the variable is significant in the prediction of the outcome variable. The logistic regression coefficients illustrate the change in the predicted log odds of having the outcome of interest for a one-unit change in the independent variables. If b_k is greater than 0, then the odds are higher, and similarly if b_k is less than 0, then the odds are lower. [10]

This allows expert analysts to clearly see how different variables contribute to the instability of the country. The log odds are calculated by using the following equation for odds

$$odds = \frac{1}{1-p}$$

where p is the probability of the presence of the variable of interest. This equation is then plugged into this equation to yield the log odds: $logit(p) = ln(odds)$. [10]

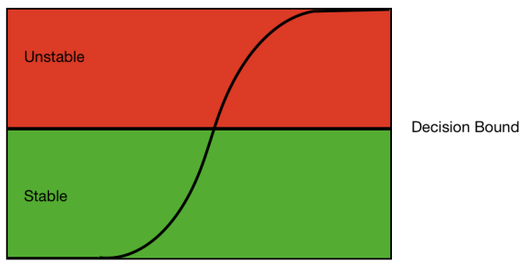


Figure 5: This figure is an example of what the simplest case of a logistic regression model would look like for the ACTOR model, with only two categories used.

Advantages and Disadvantages

Logistic regression can give analysts a more targeted look how specific change in one variable impacts a country's chances of being unstable. For example, say we know that the GDP per capita in Bangladesh is steadily dropping. Using logistic regression, we can predict, based on the history of instability in Bangladesh used in conjunction with the history of instability of other countries, at what point the GDP per capita will push Bangladesh into a state of instability. This allows analysts to analyze ongoing situations. The biggest drawback of the logistic regression approach is the data preprocessing. The data that the ACTOR model uses up until year 2000, would need already be classified as stable, moderately unstable, or highly unstable for each year. This information would be used in conjunction with the data from that year to allow the logistic regression model to learn what classifies and stable, moderately unstable, and highly unstable. This can take quite a bit of data preprocessing. However, through this process it would provide an easily interpretable way to present the conclusions to expert analysts. This method would give analysts clear looks at how shifts in the 12 macro-structural factors affect individual countries.

7 Future Work

This paper serves largely theoretical approaches to later be hard-coded and used in conjunction with the updated ACTOR model known as the Integrated Crisis Early Warning System (ICEWS). The dual approaches include:

- Creating a decision tree model that uses the 159-country dataset to compile reports that would be useful to analysts.
- Designing a web applet that allows the user to select a variable and a country and outputs how a shift in the variable affects the country's instability.

8 Conclusions

As mentioned in previous sections, the ACTOR model is not meant to be an all-encompassing, stand alone approach to predicting country instability. This methodology is to be used in conjunction with expert analysts as well as a myriad of other predictive modeling techniques and information retrieval. The changes that could be made to the ACTOR model, and to its descendant, the ICEWS (Integrated Crisis Early Warning System) rely first and foremost on the question that the government would like for ACTOR to answer. If the government needs accuracy on just a few important factors, the accuracy of the model could increase drastically. However, if the government desires a more holistic view of the individual country, the ACTOR model provides an excellent picture of 159 countries that allow policymakers and expert analysts to help inform them in important decisions for the protection and furtherance of prosperity for the United States government.

Acknowledgements

I would like to express my gratitude to Professor Lancie Affonso for his continued guidance on this project.

References

[1] e. a. O'Brien, Sean P., "Analyzing complex threats for operations and readiness," 2001.

[2] U. C. Bureau, "International programs, international data base," www.census.gov/data-tools/demo/idb/informationGateway.php., 2011.

[3] "World devlopment indicators — databank," databank.worldbank.org/data/reports.aspx?source=world-development-indicators.

[4] Y. Y. Chen, "Fuzzy analysis of statistical evidence," pdfs.semanticscholar.org/8f1b/b1a27bd35637e7efbf108882808b87616827.pdf.

[5] "Explanation of the decision tree model," webfocusinfocenter.informationbuilders.com/wfappent/TLs/TL_rstat/source/DecisionTree47.htm.

[6] e. a. Myles, Anthony J., "An introduction to decision tree modeling," onlinelibrary.wiley.com/doi/pdf/10.1002/cem.873., 2004.

[7] B. P. Management, "A review of decision tree disadvantages," www.brighthubpm.com/project-planning/106005-disadvantages-to-using-decision-trees/., 2011.

[8] B. Deshpande, "4 key advantages of using decision trees for predictive analytics," www.simafore.com/blog/bid/62333/4-key-advantages-of-using-decision-trees-for-predictive-analytics.

[9] D. S. Upadhyaya and K. Singh, "Nearest neighbour based outlier detection techniques," pdfs.semanticscholar.org/b4c9/3848b0808566b06e6527901fd07a3aa2a2f4.pdf., 2012.

[10] F. Schoonjans, "Logistic regression," www.medcalc.org/manual/logistic_regression.php., 2017.