Empirically Predicting International Diplomatic Decisions Regarding the Libyan Civil War

A Thesis

Presented to

the Faculty of the School of Engineering and Applied Science

University of Virginia

In Partial Fulfillment

of the requirements for the Degree

Master of Science (Systems and Information Engineering)

by

Alex Pape

December 2015

Approval Sheet

This Thesis is submitted in partial fulfillment of the requirements for the degree of

Master of Science (Systems and Information Engineering)

Alex Pape

This Thesis has been read and approved by the Examining Committee:

Matthew Gerber

Prof. Matthew Gerber, Advisor

Laura Barnes

Prof. Laura Barnes, Committee Chair

Jonah Schulhofer–Wohl

Prof. Jonah Schulhofer–Wohl, Committee Member

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, Dean, School of Engineering and Applied Science

December 2015

Abstract

The evolution of international political systems is heavily influenced by the choices made by diplomatic decision makers. Improved predictive modeling of these decisions could significantly contribute to improved prediction of political as well as military processes. Much of the conventional work on political prediction frames political data in the form of events sorted into a taxonomy of types. In this work the type of event I focus on recognition of the Libyan National Transitional Council (NTC) by members of the international community. The NTC was the main opposition authority in the 2011 Libyan Civil War. The recognition of authorities like the NTC can have both political and economic impact on the course of the conflict and its aftermath. The specific predictive task I focus on is that of estimating the time to recognition for a selection of countries. In contrast, existing work has mainly focused on predicting the occurrence events within a fixed time horizon. I experimented with both parametric and non–parametric approaches using both country–specific data (like indicators of economic development), general temporal data (like text features from international news), and dyadic relationships between countries (like trade relationships). Overall, I found predictive power through the influence of dyadic relationships between countries. Specifically, certain models based on dyadic relationships predict recognition with about the same accuracy as some baseline models. The models with dyadic relationships can ultimately be considered superior because they do not presuppose knowledge of the underlying distribution of event times as the baseline models do. However, once reframed as a classification problem rather than a regression, the effect of the dyadic relationships becomes even more salient. The results strongly suggest that while events like recognition are difficult to predict because they depend on human agents, each diplomatic decision is heavily influenced by pre–existing diplomatic relationships. The results also point to the ontological heterogeneity of the data that describes sociopolitical processes and the resulting difficulties of handling it in a classically inductive manner. Thus, the present work provides strong motivation for exploring alternative methods, such as those of Statistical Relational Learning, to make inferences on such structurally complex datasets.

Contents

Contents iii

1 Introduction 1

2 Related Work 3
- 2.1 Causal and Empirical Approaches . . . 3
- 2.2 Data . . . 4
- 2.3 Contributions and Gaps . . . 5
- 2.4 Survival Analysis . . . 5

3 Data 6
- 3.1 Recognition Subjects and Data . . . 6
- 3.2 Data Collection . . . 8
  - 3.2.1 Affinity Metrics . . . 8
  - 3.2.2 Text Features . . . 11

4 Methods 12
- 4.1 Survival Modeling . . . 12
- 4.2 Survival Modeling for NTC Recognition . . . 14
  - 4.2.1 Feature Extraction . . . 15

5 Experimental Setup 19
- 5.1 Baseline Methods . . . 19
- 5.2 Censoring . . . 19
- 5.3 Experiment Execution . . . 20

6 Results 21
- 6.1 Univariate Results . . . 21
- 6.2 Parametric Models . . . 22
- 6.3 Non–Parametric Model Results . . . 22
- 6.4 Ensemble Models . . . 22

7 Discussion 24
- 7.1 Conversion to Static Classification . . . 26

8 Conclusion 28
- 8.1 Summary of Research Problem and Results . . . 28
- 8.2 Impacts . . . 30
- 8.3 Limitations . . . 31
  - 8.3.1 Data Quality . . . 31
  - 8.3.2 Problem Formulation . . . 31
- 8.4 Future Work . . . 32

A Recognition Dates and Sources 33

B Variables From WDI 36

C Markov Logic Networks 37

Bibliography 38

1 Introduction

The diplomatic decisions made by the international community are important but uncertain factors in conflict prediction. In general, improved prediction of the conflict process would aid decision makers in anticipating and proactively preparing for and responding to future actions of their opponents. At a strategic level, conflict prediction can inform decisions about scale and timing of intervention. At an operational level, prediction aids decision–makers in more efficiently allocating limited resources (such as troops and munitions) and balancing the needs of the combat mission with other objectives, like protecting civilians and providing humanitarian aid.

However, prediction can be seen not only as a means of preparing for conflict, but also as a means of preventing the occurrence of conflict. One of the key explanations for war between rational actors outlined by Fearon [1] is mutual misperception and mistrust (because neither side is willing to disclose military secrets) that leads to overconfidence. Prediction using open–source data, which conflict actors have neither the ability nor motivation to conceal, would provide an independent basis for resolution through negotiation rather than military confrontation.

The specific case considered in this thesis is recognition of the Libyan National Transitional Council (NTC), which was the de facto governing body of the rebels fighting the forces of Muammar Gaddafi in the Libyan Civil War. Major protests began around 17 February 2011, but the actual military conflict started at the beginning of March and continued until October of that year. Beginning on 19 March 2011, a UN–authorized no–fly zone was imposed by several NATO states, including the US. The first country to recognize the NTC was France, which did so on 8 March 2011. The United Nations General Assembly (UNGA) voted to allow the NTC to fill Libya's seat on 16 September 2011.

International recognition of opposition authorities like the NTC can have both political and economic impact on the course of the conflict and its aftermath. In addition to the obvious external political credibility, recognition can have major impacts on the internal structure of the opposition. On one hand, it can help the opposition follow a more unified and coordinated strategy both in pursuing its military objectives and in the civic administration of areas under its control. On the other hand, recognition can fracture the opposition by aligning the interests of the recognized authority with the interests of the recognizing states. For instance, it is easy to imagine how, under slightly different conditions, extremists could have perceived the NTC as a puppet of the West to such an extent that they would form an opposing militant group, especially considering the factionalization that has subsequently taken place in Libya.

In addition to the political role, recognition plays an important economic role. For example, it can facilitate trade: Qatar recognized the NTC in connection with their oil deal in late March 2011. This is particularly significant given the general reluctance of other Arab states to do so. Further, in many countries, it is a necessary condition for granting the frozen funds of the incumbent administration to the opposition. However, it is far from a sufficient condition: as of 30 Nov 2011, only $18 billion of Gaddafi's originally–frozen $150 billion had been released, even though virtually all countries as well as the UNGA had recognized the NTC by that point [2]. Overall, though its precise effect is complex (and is itself an important topic of study), recognition is nonetheless potentially very significant to conflict processes.

Conceptually, there are four basic levels of recognition, though in practice, recognition is not as simple [3–9]. For example, there is no universal vocabulary for each level, but rather types of statements and terms that are generally associated with each level. Further, countries may intentionally and unofficially push the boundaries of these levels. However, major players in international security like the US, UK, and France draw a great deal of attention to their recognition announcements.

The levels are as follows (we/us refers to the country that is or is not doing the recognizing):

1. Non–recognition: We take the formal stance that nothing is happening and that the incumbent government is the only representative of the country.

2. Recognition of Belligerency: We recognize that there is a conflict and that some specific authority has the right to speak for the opposition (for instance, in negotiations). As a historical example, Britain and France recognized the US Civil War in this manner, to which the Lincoln administration protested, since it was an internal matter (as many civil wars are).

3. Recognition as representative of the country's people: We recognize that some specific authority is working in the interest of and has the support of some large proportion of the people. Because this is not legally binding, only unofficial diplomatic relations can be established, and diplomats of the incumbent government cannot be expelled or forced to hand over the embassy to their opposition counterparts. For example, Gaddafi diplomats were still in the UK until late July 2011 (though one wonders what exactly they had been doing for the period of over four months since the NATO bombing began). Interestingly, some of Gaddafi's diplomats sympathized with the opposition and simply switched allegiances while maintaining their diplomatic roles.

4. Recognition as representative of the state: We recognize some specific authority as the full–fledged government of the country, with all the rights and responsibilities thereof. This involves breaking off formal diplomatic relations with the incumbent government (often symbolically by changing the flag at the embassy). Another reason that this is significant is that it absolves the former government from its

responsibilities – for example, it no longer needs to guarantee immunity for our diplomats (though in practice they would have long since been evacuated).

The predictive task is to estimate how long each of a set of countries will wait to grant recognition to a new government (specifically the NTC) in one particular country (specifically Libya). Thus, in modeling terms, the predictor variables are features related to each of the would–be recognizing countries, and the response variable is the corresponding time to the recognition date.

The remainder of this thesis proceeds as follows. In Section 2 I review previous work in the area of political prediction. In Section 3 I outline what data I collected and how I did so. In Section 4 I discuss our specific modeling approaches. In Section 5 I describe how the models are tested. In Section 6 I present the results obtained. In Sections 7 and 8 I discuss the implications of the work for Libya specifically and for political prediction in general.

2 Related Work

Predicting conflicts or conflict progression is a special case of the more general problem of predicting political events. Data–driven conflict prediction has been studied since the Cold War [10] and has since then taken many forms.

2.1 Causal and Empirical Approaches

One recent review [11] points out how the pursuit of political prediction displays the classic dichotomy between causal (i.e., mechanistic) and empirical approaches. For instance, researchers following the causal approach would seek to understand the process of escalation or de-escalation between an opposition group (anything from peaceful protesters to militants) and an incumbent government (not necessarily repressive), in which each side adopts some level of aggression but faces highly uncertain payoffs [12–15]. This sort of research examines theories like that of the "murder in the middle" hypothesis, which holds that the most harmful political systems are not authoritarian states, but those that reside between authoritarianism and democracy, because such systems have both a somewhat repressive government and an opposition active enough to incur that repression. The advantage of taking such an approach is that it can very directly inform and be informed by political science theory. However, a disadvantage is that such models only rely on assumptions and inputs that are not necessarily available or reliable. For example, game theory models rely heavily on well–defined payoff functions that have to be estimated by a human expert. Meanwhile, they ignore the many readily available sources of information that lack a clear place in the model.

In contrast, the empirical approach places a higher priority on accurate prediction, with less concern about the causal connections between the predictors and response. A good example of the empirical approach is presented in [16], which uses textual indicators from news on a given day to predict a high or low count of events in the Conflict and Mediation Event Observations (CAMEO) Material Conflict class the next day. One of the key advantages of this approach is that it is open to any available source of data that could be germane to the process under consideration. Primarily motivated by this advantage, the present work follows the empirical approach.

2.2 Data

Much of the previous work on political prediction has used techniques such as point processes [17], Autoregressive Integrated Moving Average (ARIMA) models [18, 19], or Hidden Markov Models (HMMs) [20, 21] that attempt to model a process with few or no covariates. When used, covariates for empirical models usually fall into three general classes: long–term structural data, event data, or textual features. Structural variables, such as economic, demographic, and environmental factors are used in [22–25]. Structural factors can also be expressed as qualitative characteristics describing political and economic context, as used in [26, 27]. Alternatively, data about events distinct from the response can be used as predictors. For instance, Shellman et al. use event data about state repression as a predictor for violence of insurgent groups [22]. Finally, a less–explored type of predictor is that of textual data. For example, Leetaru [16] uses textual features of daily news reports as predictors of high or low counts of CAMEO Material Conflict events. The present work applies several different data types, which are discussed in Section 3.

One of the key characteristics of any temporal prediction is time scale, i.e., the unit of time over which each observation of predictors and response is defined. Many political prediction studies are defined on a scale of months or years (e.g., [22]). One notable exception is [19], which predicts the daily status of dyadic interactions in the Balkans conflict in the late 1990s. As discussed above, [16] also uses news text to predict a high or low Material Conflict event count on a daily basis. It is important to note that Vector Autoregression (VAR) models of political interactions change when the time scale is varied, so predictive power found at fine–grained time scales may not exist at coarse–grained time scales [28]. The present work adds to the small body of research targeting day–resolution predictions of conflict progress. Another important characteristic of temporal prediction is response formulation. Previous efforts have formulated responses as counts (e.g., air strikes per day) and scales (e.g., the cooperation-opposition continuum) [29]. A third alternative is to focus on events of one particular type and predict the time to the occurrence of those events, which is formally referred to as survival analysis, and is further discussed in Section 2.4.

2.3 Contributions and Gaps

An important gap in the current work is commensurability between human and machine-based analysis. As thoroughly documented in Philip Tetlock's *Expert political judgment: how good is it? How can we know?*, sociopolitical prediction problems are very difficult to pose in a manner that is both understandable and relevant to human experts and decision makers but also empirically tractable [30]. An empirical model may often make predictions that are technically more accurate but far less insightful than those of a domain expert. However, though an expert may offer a wealth of factually true information that is relevant and insightful, he may still have inaccurate (and yet overconfident) answers to a few particularly crucial questions. A good example of such a shortcoming was the planning of the US military operation that resulted in the death of Osama Bin Laden, in which President Obama expressed consternation at being presented by his advisors with estimates of the probability of success ranging from 30% to 95% [31, 32].

A further gap is generalization: existing work in empirical political prediction has mostly been very specific, focusing on particular event types in particular sociopolitical contexts. For example, [29] exclusively considers military events in Afghanistan, and thus is only useful within that very limited context. Further, since the data used in [29] comes from a time period of specific conditions, its predictive power is uncertain once those conditions are no longer in place (e.g., once coalition troops leave Afghanistan).

Another key gap is that much existing work relies on predictors that are too proximate to the actual political process to be useful. For example, [33] discusses how DARPA's Integrated Conflict Early Warning System relied on some near–trivial correlations, such as one between expressions of ethnic hostility and acts of ethnic hostility. While examining the details of such a relationship may be interesting from the mechanistic perspective discussed in Section 2.1, its predictiveness is a rather uninteresting hypothesis from an empirical perspective.

Since I am focusing specifically on the Libyan Civil War as a case and time to NTC recognition as a response, and doing so with conventional empirical methods, the present work does not overcome the limitations of commensurability or generalization in a direct manner. However, the methodology and results do point to some possible improvements to these gaps, which are discussed in Section 8. Further, the present work avoids the problem of trivial predictors by focusing on a problem for which no such predictors exist.

2.4 Survival Analysis

Predicting the time to events (NTC recognition, in this case) is formally referred to as survival analysis. It is distinguished from generic statistical modeling by two characteristics. First, since the response variable represents elapsed time, it is strictly non–negative. Second, survival models must handle events whose time of occurrence is not strictly known, but is known to have occurred before or after a certain point. Such observations are called "censored." For example, if a cancer patient dies in a traffic accident, his or her time of death due to cancer is not known exactly, but would have occurred after the time of the accident. This is not to be confused with truncation, a special type of selection bias in which subjects whose event times occur outside a certain range are never observed at all. For example, if a zoologist first observes a litter of wolf pups at age x, she has no way of knowing whether or not the litter also included pups that died before age x. Such an observation would be considered truncated at x. There are many innovative variations within the general class of survival analysis, such as modeling recurrent (as opposed to one–time) events, modeling time to the first of several events (known as "competing risks"), and modeling exogenous prevention (i.e., when the event is prevented from occurring in some proportion of the subjects, known as the "cure fraction") [34]. However, the models in the present work are formulated in the conventional single–type, single–occurrence event manner and thus these more complicated forms of survival analysis have little bearing.

As the name suggests, survival analysis has traditionally been and continues to be applied in medical and actuarial domains, in which a single person, rather than a social group, is the unit of analysis. Survival analysis in the political domain is rare. Exceptions include [35] and [36], but these were relatively isolated cases that focused very specifically on modeling the longevity of parliamentary governments and bear little relation to the present work in terms of either methodology or application.

3 Data

3.1 Recognition Subjects and Data

One of the advantages of predicting time to recognition is that one can arbitrarily select a pool of governments (that is, countries) for this analysis. I considered countries that (1) are in Europe, Africa, and the Middle East, and (2) had limited or no involvement in the NATO operation in Libya. I chose the geographic restriction because those were the countries for which the Libyan Civil War was most politically and economically relevant. I imposed the NATO non–involvement restriction because involvement would constitute material support for the NTC and thus make the act of verbal support (i.e., recognition) meaningless.

Unfortunately, there is no centralized source for recognition dates, so I collected dates from open–source news articles and press releases from foreign ministries (or equivalent institutions) of the various governments under consideration. Wikipedia[1] provided a useful starting point for links to such sources. Overall, I was able to identify clear dates for either Level 3 (also known as *de facto*) recognition or Level 4 (also known as

[1] https://en.wikipedia.org/wiki/International_recognition_of_the_National_Transitional_Council

de jure) recognition for 54 countries. A list of dates and sources is shown in Appendix A.

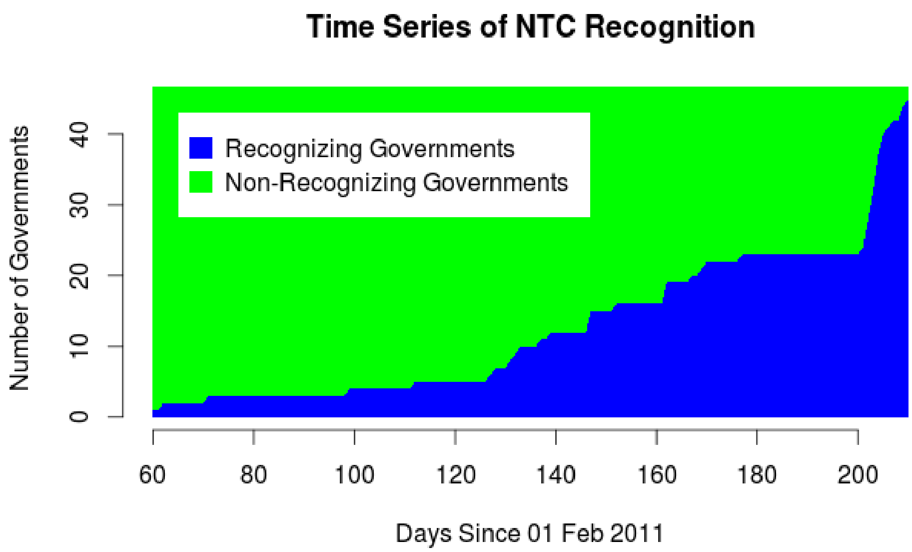


Figure 3.1: Cumulative Count of NTC Recognitions

For the response variable, I pooled all the countries and used whichever recognition level (Level 3 or Level 4) for which a date of recognition was known. The only country for which both dates were known was Denmark, so I dropped that observation. That left a pool of 53 countries. I arbitrarily set $t = 0$ at 01 February 2011. The time series of cumulative recognitions is shown in Figure 3.1.

A major problem that arises is heterogeneity among countries in the pool. For instance, consider a large and influential government like Germany contrasted with a much less influential government like Bosnia. The timing of Germany's decision was very likely chosen with a great deal of deliberation and thought regarding the current situation and public sentiment, especially since Germany initially had reservations about the NATO mission. In other words, there were clear decision factors for and against such action. Other states with such clear motivations include states like Algeria, Saudi Arabia, and South Africa (because of their complicated history with the Gaddafi regime) as well as Russia and China (because of their general opposition to foreign intervention in such cases). In contrast, it is difficult to find any sort of motivations behind the decision of a small country like Bosnia; the timing of its recognition may simply have been a matter of executive or diplomatic convenience. Needless to say, this depends on the situation; for instance, great significance would be attached to Bosnia's recognition of Kosovo because of the many social, cultural, and historical ties they share. In the present work, this heterogeneity must simply be accepted as a source of error.

It is important to note that the pool of countries represents a selection bias: I am using only cases for which it is known that a recognition date can be identified. In a real application, an analyst would have no such hindsight to distinguish between governments that would grant recognition at some point and those that would indefinitely remain silent on the matter. I am not aware of any particular property that countries in the pool possess and that outsiders lack (other than an obvious interest in Libyan relations). It is also important to consider that many countries likely treated their vote in the UNGA (on 16 September 2011) in support of allowing the NTC to occupy the Libyan seat as a statement of recognition.

Besides this selection bias, there are several sources of error in this process, the most obvious of which is the aggregation of the two fundamentally distinct event types (*de facto* and *de jure* recognition) into a single class. The reporting process from which I obtained the recognition dates is another source of error: agreements between the NTC and various governments could have taken place behind closed doors before or after the date that was officially announced. Further, the governments may have chosen the timing of their recognition for convenience or for other more or less random reasons.

3.2 Data Collection

3.2.1 Affinity Metrics

Affinity metrics are variables that characterize the relationship between two countries, including UNGA voting records, formal alliances, diplomatic exchange, membership in inter–governmental organizations, and bilateral trade.

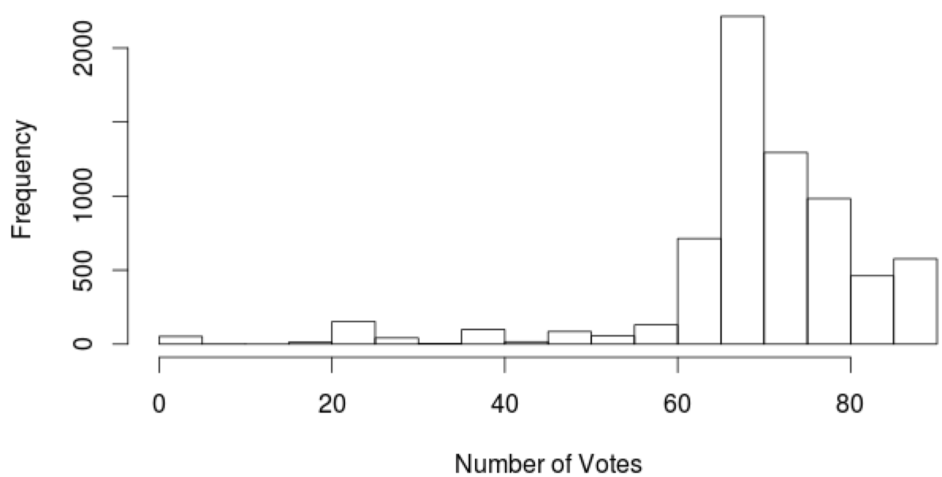


Figure 3.2: Histogram of Pairwise Vote Counts

UNGA Voting Records

UN voting records are available (from [37]) in two forms: as shown in Table 3.1 and as pre–computed affinity measures. The calculation of those measures (called S indicators [38]) uses the votes (called resolutions) for which the two countries in question were both present (i.e., both entered a vote of yes or no or abstained). Specifically, if v_A and v_B represent the vector of n votes for countries A and B, respectively,

$$S = \frac{1}{n(v_{max} - v_{min})} \sum |v_A - v_B|, \tag{3.1}$$

where v_{max} and v_{min} are the highest and lowest possible vote levels (in this case, $v_{max} = 3$ and $v_{min} = 1$). In other words, S is just the mean absolute difference between the vote vectors normalized by the range of possible levels. Resolutions for which either was absent are omitted. Because only two countries at a time are considered, this calculation is fairly robust to sparse attendance. As shown in Figure 3.2, between 60 and 90 resolutions are on average shared between any two countries. Because v in Equation 3.1 can only take on values for votes of ”yes,” ”no,” and ”abstain,” I treated absences as abstentions (only about 4% of all possible votes are listed as absent).

Table 3.1: Example of UNGA Voting Records

resolution	govt 1	govt 2	$\cdots$	govt N
R/<session #>/1	Yes	Abstain	$\cdots$	No
R/<session #>/2	No	No	$\cdots$	Abstain
R/<session #>/3	Yes	Absent	$\cdots$	Yes
R/<session #>/4	Absent	Yes	$\cdots$	Yes
R/<session #>/5	Abstain	No	$\cdots$	Yes
R/<session #>/6	Yes	No	$\cdots$	Absent
R/<session #>/7	No	Abstain	$\cdots$	No
R/<session #>/8	No	No	$\cdots$	Yes
R/<session #>/9	No	Absent	$\cdots$	Yes
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
R/<session #>/n	Yes	No	$\cdots$	No

Formal Alliances

Data on formal alliances between countries was obtained from the Correlates of War (COW) project, version 4.1 [39]. Types of alliances range from mutual defense (e.g., that of NATO members) to simple consultation agreements (technically called “entente alliances”). The correlation between these was almost perfect, so I kept only the variable representing the presence of a mutual defense alliance.

Diplomatic Exchange

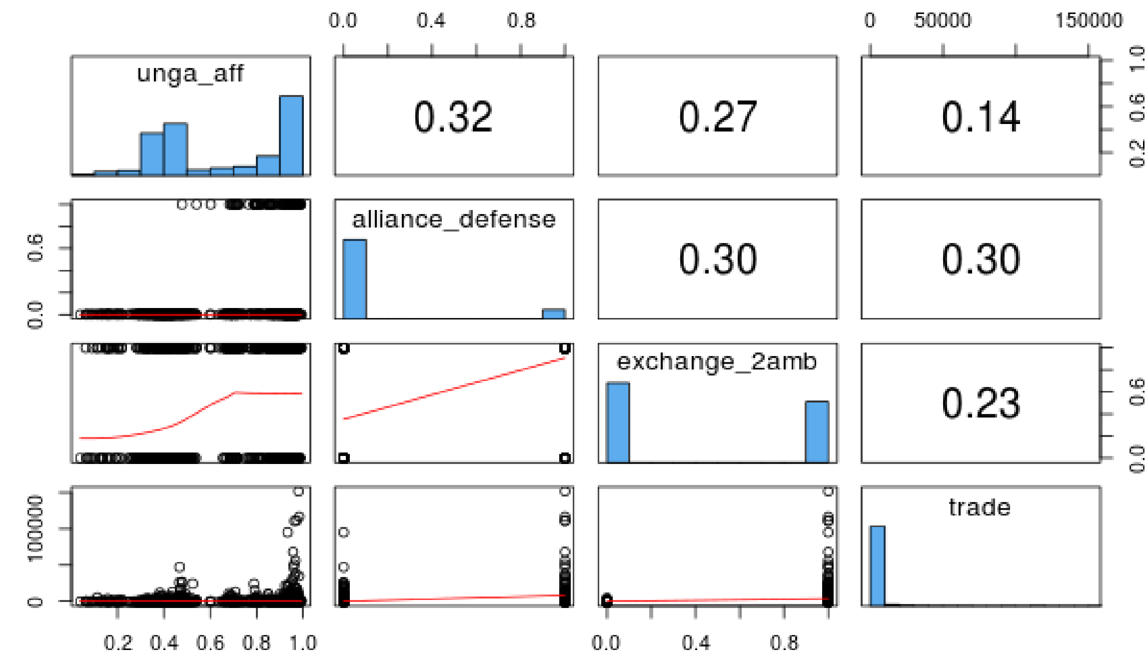


Figure 3.3: Scatter Plot Matrix for General Pairwise Metrics

Another way to characterize the relationship between two countries is the type of diplomatic representation of each at the capital of the other. I used the COW dataset (v2006.1) [40]. Unlike almost all the other dyadic measures, this is technically directed (i.e., it is possible for government A to have an ambassador to government B without government B having an ambassador to government A). Also, there are many different levels of diplomatic relations recorded in the dataset (in order of increasing strength: chargé d'affaires, minister, and ambassador, plus a miscellaneous category). As shown in Table 3.2, there are few countries in the pool that have some diplomatic exchange but do not maintain mutual ambassadors. Thus, for simplicity, I set the exchange metric to the boolean indicator of whether or not the two countries have mutual ambassadors (shown as `exchange_2amb` in Figure 3.3).

Table 3.2: Diplomatic Exchange Measures

		Mutual Ambassadors?	
		True	False
Any Diplomatic Exchange?	True	598	213
	False	0	567

Bilateral Trade

The COW project also has data on trade (version 3.0) [41, 42]. Like diplomatic exchange, this is directed (flow from A to B and from B to A). For simplicity, I added the two together to get a trade volume. There is a great deal of significance in the relative size of the trade flows, but such structure is too complicated for the model.

3.2.2 Text Features

There are several options for data sources. First, the Global Database of Events, Language, and Tone (GDELT) could be used as a surrogate for text. This would be very easy to disaggregate by country, since each event has a well–defined (though often inaccurate) location. However, the lack of a literally textual baseline could be a disadvantage if comparison to other textual sources was needed. A second option is to use the textual sources upon which GDELT is built. These are accessible via Lexis Nexis (GDELT conveniently provides the exact query). While offering the real text, they would involve a great deal of processing (especially to disaggregate): the AFP alone provides over 250 articles per day. A third option is to use news agency twitter feeds. This approach is the more "elegant" one because it is essentially just restricting analysis to a few specific twitter users rather than the general population. However, it would be the hardest to disaggregate, because one would have to identify commensurate news agencies in each country (especially difficult in the developing world). I eventually chose to compromise between the second and third positions. Specifically, I chose to use headlines from Lexis Nexis articles, primarily on grounds of convenience.

The challenge of using headlines is that they have to be downloaded manually with a maximum batch size of 3 000. So as to limit the number of batches, I tried querying for Libya–related articles (i.e., containing the word "libya") from the Associated Press, AFP, UPI, and Xinhua. This ended up giving only about 24 000 headlines (after dropping meaningless headlines, like "News Summary ..."), leaving a few days without any headlines at all. To supplement this, I downloaded all headlines (not just Libya–related) from BBC Monitoring, which gave approximately 149 000 headlines (again, after dropping meaningless headlines), for a total of about 173 000 between February 1, 2011 and December 31, 2011. In addition to the removing meaningless headlines, I removed some headline–specific stopwords, like "Feature ..." and "Special Report."

For textual analysis, the normal preprocessing steps of tokenization (by spaces), stemming, and stop word removal were carried out. Specifically, English stemming was done with the standard Snowball stemmer from the Natural Language Toolkit [43], and English stopword removal was done with the standard English Stopword list, also from [43]. In addition, I used the Stanford POS tagger [44] to filter out all except following parts of speech: JJ (adjective or numeral, ordinal), JJR (adjective, comparative), JJS (adjective, superlative), NN (noun, common, singular or mass), NNP (noun, proper, singular), NNPS (noun, proper, plural), NNS (noun, common, plural), RB (adverb), RBR (adverb, comparative), RBS (adverb, superlative), VB (verb, base form), VBD (verb, past tense), VBG (verb, present participle or gerund), VBN (verb, past participle), VBP (verb, present tense, not 3rd person singular), and VBZ (verb, present tense, 3rd person singular). Overall, the corpus was composed of approximately 1.3 million tokens before stopword removal and 1.2 afterwards.

After preprocessing I extracted topic distribution features. Specifically, I used Mallet [45] to build a 100−topic Latent Dirichlet Allocation (LDA) model, with both training and inference on the full dataset. The most popular words in each topic had little semantic connection to each other, besides a few that were obviously related to specific geopolitical themes (e.g., the Israeli–Palestinian conflict). In general, the topics were composed of the generic terms one would expect to find frequently in international news headlines, such as "leader," "military," and "kill."

4 Methods

4.1 Survival Modeling

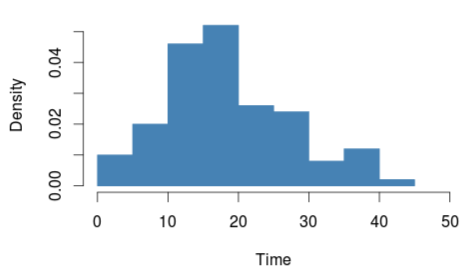
As discussed in Section 2.4, survival analysis is distinguished from conventional statistical modeling by the non–negative time response and the effects of censoring and truncation. In general, a survival model can be expressed in the form:

$$t_i = g(\boldsymbol{Z}_i) + \epsilon \tag{4.1}$$

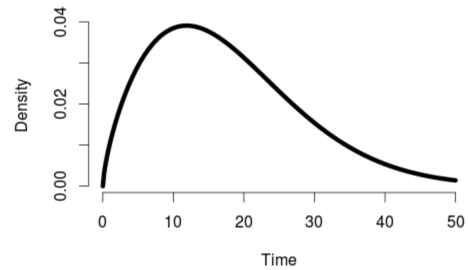
where t_i and $\boldsymbol{Z}_i$ are the time to event and feature vector for the ith subject, respectively, and ϵ is a generic error term. In other words, $g(\cdot)$ is estimating a density (or mass) $f(t)$ on the time T to event. Practically, however, the survival function $S(t)$ and the hazard rate $\lambda(t)$ are often of more direct interest than $f(t)$. The survival function is the complement of the cumulative distribution function, or $S(t) = \Pr(T > t)$. The hazard rate is the probability of event occurrence in the next unit of time given that it has not already occurred. In instantaneous form, the hazard rate is $\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$. Figure 4.1 shows an example of the above quantities calculated for a Weibull distribution with shape parameter 1.7 and scale parameter 20.0.

Survival models fall into two general classes: parametric and non–parametric. Parametric models are those that assume an underlying parametric distribution for the time to event. The simplest parametric model possible is the Accelerated Failure Time (AFT) model, which assumes that covariates scale the argument of a baseline survival function [46]:

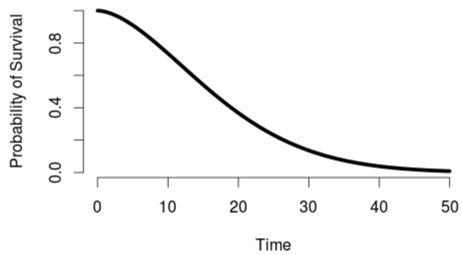
$$S(t|\boldsymbol{Z}) = S_0(t \exp(\boldsymbol{\beta}^T \boldsymbol{Z})) \tag{4.2}$$



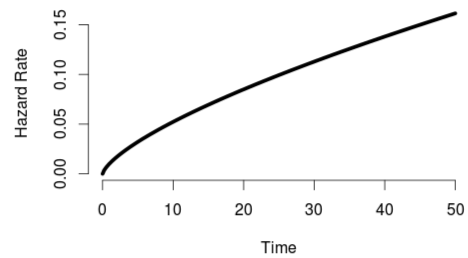
(a) Histogram of Sampled Points



(b) True Density Function



(c) True Survival Function



(d) True Hazard Function

Figure 4.1: Survival–Related Quantities Calculated from Simulated Data

The AFT model has a convenient interpretation in direct linear form as $\ln(T) = \mu - \boldsymbol{\beta}^T \boldsymbol{Z} + \sigma W$, where μ and σ are constants and W is a random variable that follows an arbitrarily–selected parametric distribution. Non–parametric models are those that assume no underlying parametric distribution for T. Thus, they do not necessarily conform to any formulation besides Equation 4.1.

The effects of censoring and truncation, which are really forms of missingness, do not directly impinge on Equation 4.1 itself but rather play a role in the model fitting process. In principle, as with a conventional ordinary least squares regression, it is assumed that all the observations are independent and identically distributed, though in practice this is often violated. However, the assumption is of primary importance for a model built for mechanistic rather than predictive purposes. For example, in many biomedical applications, obtaining predictive accuracy is unimportant compared to conducting a statistically rigorous test of the efficacy of certain therapies. Further implications of violations of the *iid* assumption are explored in Section 8.3.

Figure 4.2 illustrates the time variables as defined for a single censored subject. An uncensored subject would be the same but without censoring time, so t_r' and t_c would be undefined.

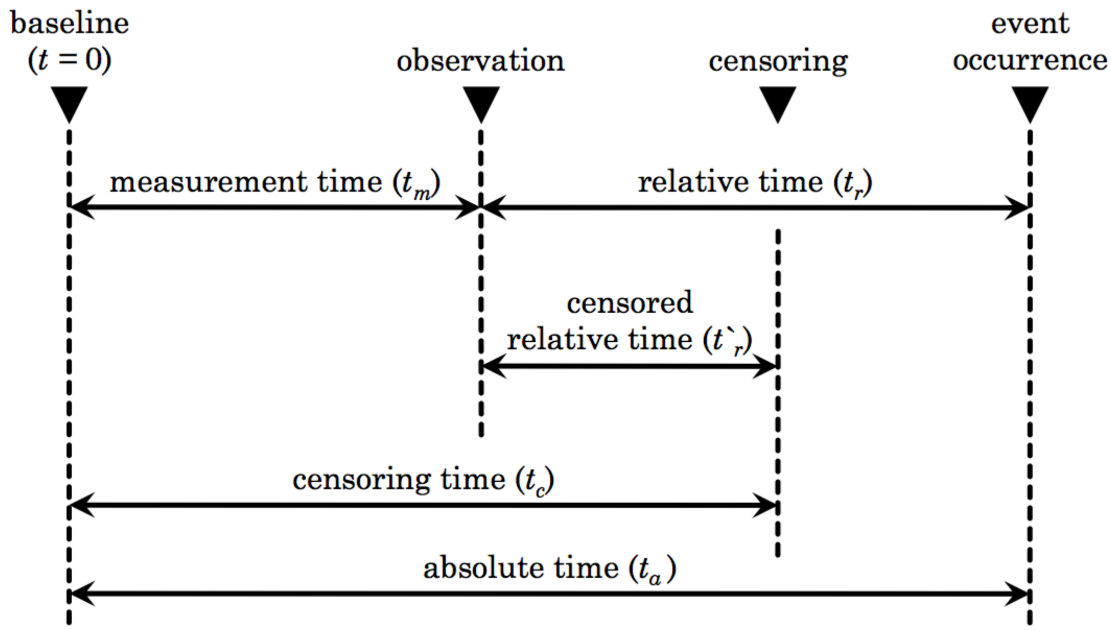


Figure 4.2: Diagram of Time Variables for Generic Censored Observation

4.2 Survival Modeling for NTC Recognition

As discussed in Section 4.1, the overall model is a survival analysis of time to recognition. There are two main challenges in choosing a unit of analysis: spatial and temporal. Spatially, there are features that are aggregated across all countries under consideration (such as global economic indices like the price of oil), while there are other features that are disaggregated (such as country–level features like GDP). Temporally, there are static features (like previous alignment with the Gaddafi regime) as well as dynamic features (like text features from news). As a result, there are three types of features, as illustrated in Table 4.1. Static aggregated features, by definition, are constant over space and time, so they offer no utility for discriminating between recognition times.

Table 4.1: Features Types

		Temporal Dimension: Static	Temporal Dimension: Dynamic
Spatial Dimension	Aggregated	N/A	Not specific to any country, varying over time (e.g., price of oil), symbolized by T
Spatial Dimension	Disaggregated	Specific to each country, but constant over time (e.g., GDP), symbolized by S	Specific to each country and varying over time (e.g., currency exchange rate), symbolized by U

For modeling the present problem, let the ith country be represented by $S_i = (x_i(t_1), x_i(t_2), \ldots, x_i(t_c))$, where $x_i(t)$ is a sample of features at time t. Specifically, $x_i(t) = (S_i, T(t), U_i(t))$, where $S_i = (s_{i,1}, s_{i,2}, \ldots, s_{i,n_S})$ is a vector of n_S static disaggregated features for the ith country, $T(t) = (t_1(t), t_2(t), \ldots, t_{n_T}(t))$ is a vector of n_T dynamic aggregated features sampled at time t, and $U_i(t) = (u_{i,1}(t), u_{i,2}(t), \ldots, u_{i,n_U}(t))$ is a vector of n_U dynamic disaggregated features for country i sampled at time t.

I use these features to estimate remaining time (t_r) at measurement time t_m as:

$$\widehat{t_r} = f_m(f_s(x_i(t_1)), f_s(x_i(t_2)), \ldots, f_s(x_i(t_m))) + \epsilon \tag{4.3}$$

where $f_s(x_i(t))$ is a model for estimating time to event based on $x_i(t)$, that is, based on data observed at time t, and $f_m(\cdot)$ is a function for estimating time to event given a history of estimates from the initial time t_1 to current time t_m. Conventional survival models like AFT were used for $f_s(x_i(t))$, while I define $f(\cdot)$ to simply be the median of the $f_s(x_i(t))$ for last 90 time units (days) prior to t_m. Figure 4.2 illustrates the various time variables relevant to the model.

Overall, the modeling approach used in the present work is just one solution to the problem of integrating the three different units of analysis in Figure 4.1. Specifically, it is the approach that takes the most specific unit of analysis possible (that is, the country–day). Alternatively, I could have used each subject (each row in Figure 4.3) as a single observation, incorporating dynamic features as time–dependent covariates. That is, rather than having two functions $f_m(\cdot)$ and $f_s(\cdot)$, I could have used a single function, say $f_0(\cdot)$, to estimate $\widehat{t_r}$ from $(x_i(t_1), x_i(t_2), \ldots, x_i(t_m))$. While this is useful in developing explanatory models, it is not conducive to a predictive application because it would require extrapolation of the predictors, as noted in [47] (pg. 304).

Figure 4.3 provides a graphical illustration of all three types of features; in this case, each subject is a country, whereas in traditional medical applications of survival analysis each subject is a patient. Each x varies across time and across subjects (that is, across space). Elements of x that are shared across subjects (vertical axis) are spatially aggregated, elements that are constant over time (horizontal axis) are temporally static, and elements that change over both dimensions are dynamic disaggregated.

4.2.1 Feature Extraction

Both the affinity metrics discussed in Section 3.2.1 and the text features discussed in Section 3.2.2 can form the basis of the three basic types of features shown in Figure 4.1. The following explains how each type of feature was extracted from the relevant data.

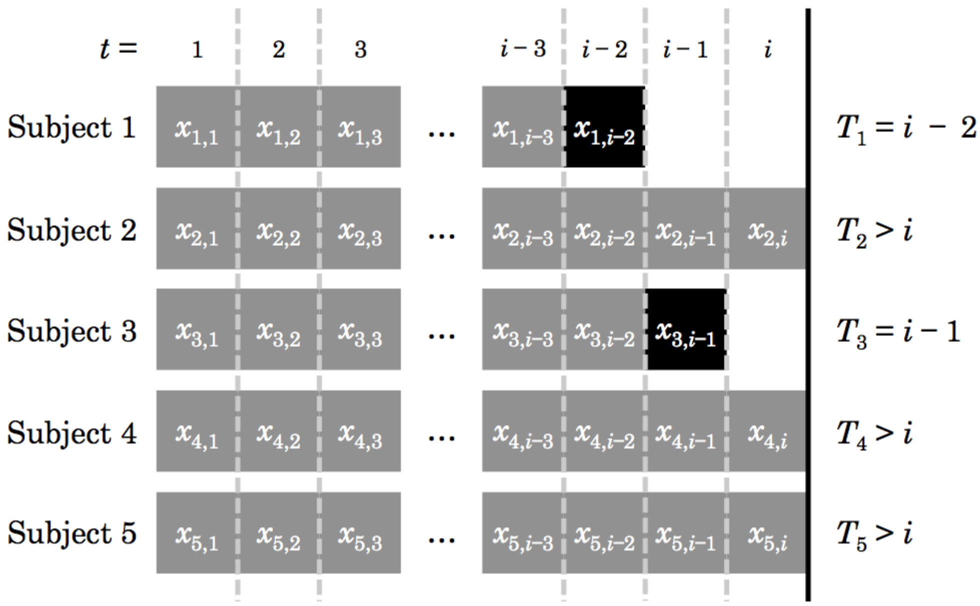


Figure 4.3: Time–Dependent Covariates Observed Until Event Occurrence (Black Squares) or Time $t = i$ for a Set of Hypothetical Subjects

Dynamic Aggregated Features

Features of this type could be utilized directly with no further processing. Specifically, dynamic aggregated features consisted of a single economic variable (the WTI oil price) and text features extracted from news headlines, as discussed in Section 3.2.2.

Static Disaggregated Features

Features of this type were extracted from the World Bank's World Development Indicators (WDI), as well as from features characterizing the dyadic relationship between Libya and each country in the pool, features from the Polity IV project, and the Composite Index of National Military Capability (CINC). The WDI reflect development–related data like environmental, industrial, and educational factors. The Libya–related dyadic features, Polity IV features, and CINC complement this by reflecting diplomatic and political variation among the countries in the dataset. To avoid collinearity, I filtered the static disaggregated variables such that no two had a correlation higher than 0.75.

The WDI dataset [48] covers approximately 1300 unique features, but only 243 were available for all of the countries in the pool for 2010 (which would be the appropriate year from which to get data for 2011 predictions). For conciseness, I manually selected 38 variables that represented the general categories of information in the dataset. The variables are listed in Appendix B.

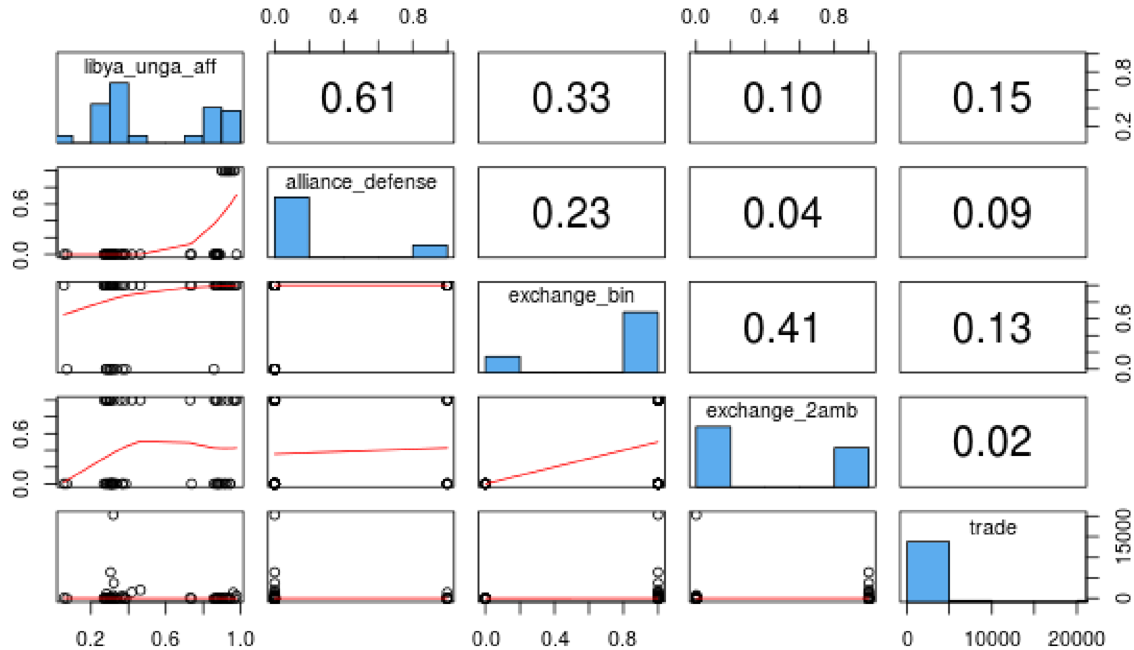


Figure 4.4: Scatter Plot Matrix Showing Low Correlations Between Metrics of Alliance With Gaddafi Regime

The dyadic features, which reflect the alignment of each country with the Gaddafi regime, were constructed with the affinity metrics in Section 3.2.1. Allies of the regime may be more reluctant to recognize the NTC, or perhaps, if disgruntled, eager to see a change of power. Figure 4.4 shows that these metrics are relatively uncorrelated with each other, indicating that each one conveys some distinct information about the relationships between a given pair of countries.

Lastly, two features from government type were obtained from the Polity IV Project dataset [50]. Specifically, I used the features "POLITY2" (revised combined polity score) and "DURABLE" (regime durability). The "POLITY2" feature expresses where a particular country lies on the continuum between autocracy and democracy, and the "DURABLE" variable expresses how long it has remained on a particular location on that continuum. CINC data came from the Correlates of War National Military Capabilities dataset version 4.0 [49].

Dynamic Disaggregated Features

Features of this type consist of currency exchange rates and features extracted from dyadic affinity metrics (which remain constant over time) between countries in the pool, which are then combined with the recognition status vector (which can change at each time step). The currency variables were relatively straightforward; I

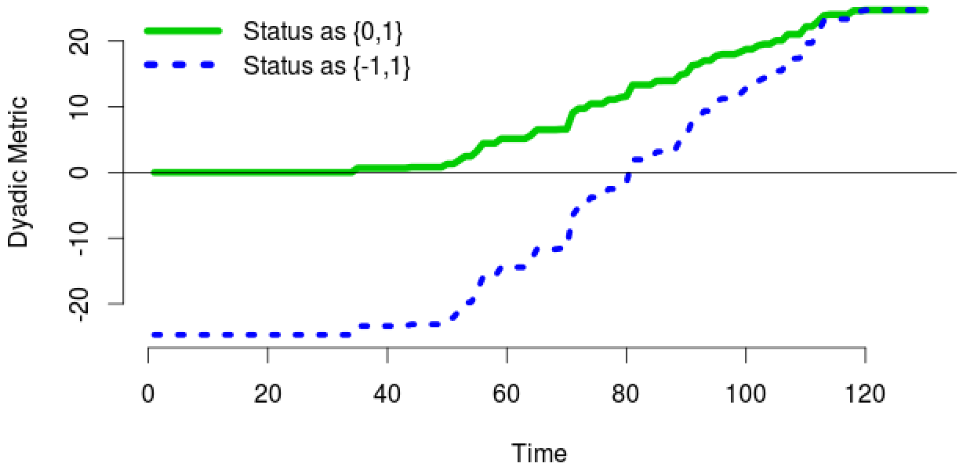
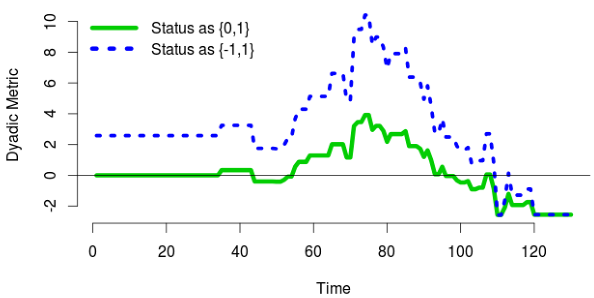
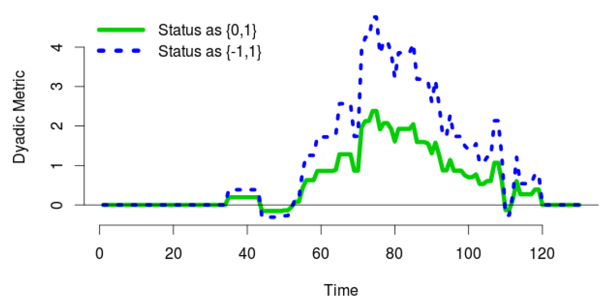


Figure 4.5: Dyadic Metric Computation with Non–negative Dyads



(a) Dyadic Metric Computation With Real–Valued Dyads



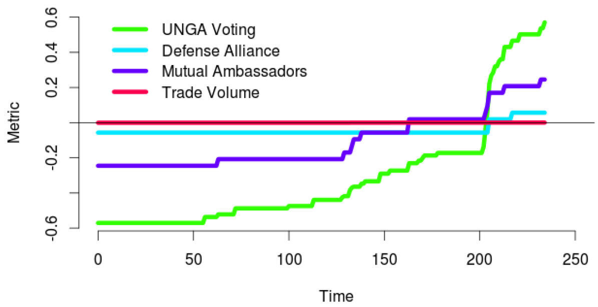
(b) Dyadic Metric Computation WIth Real–Valued Zero–Mean Dyads

Figure 4.6

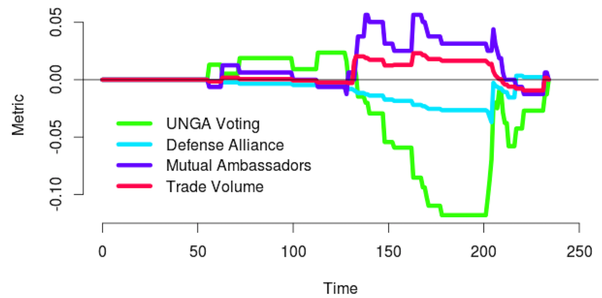
downloaded the daily exchange rates (midpoint) from [51].

The approach I have been using for combining affinity metrics and the recognition status vector is that of taking the dot product of the dyadic affinity vector (which is strictly positive) and the recognition status vector, either as–is ($\{0, 1\}$) or adjusted to $\{-1, 1\}$. An example with dummy data corresponding to that of a single country over the entire time period is shown in Figure 4.5. The alternative is to allow the dyadic metrics to be negative, as shown in Figure 4.6a (with the same dummy data), as well as possibly shifting to zero–mean as shown in Figure 4.6b (again, same dummy data). Shifting to zero–mean has the effect of forcing the initial and final values to both be zero.

I chose to experiment with the positive–valued, zero–mean dyads, as in Figure 4.7b as opposed to Figure 4.7a. It should be noted that Kenya is a special case because it was the final country in the pool to recognize the NTC. For all other countries, the metrics are truncated before they reach zero, but only because the



(a) Alliance Metrics for Kenya Using Positive–Valued Dyads



(b) Alliance Metrics for Kenya Using Real–Valued Zero–Mean Dyads

Figure 4.7

metrics cease to be predictively meaningful once the country has recognized the NTC; if the computation were extended, they too would finally reach zero.

5 Experimental Setup

5.1 Baseline Methods

There are two baselines with which to compare predictive models for this problem. One is a univariate estimate assuming a parametric distribution for the time to event. In formulaic terms, a univariate model is an AFT model with a zero–length $\boldsymbol{Z}$. As illustrated in Figure 5.1, I did a censored distribution fit on the event times observed up to a given threshold, then used the median of the upper tail (shaded area in Figure 5.1) as the the predicted time for all other governments in the dataset.

The other baseline is a null model that predicts that all events that have not yet occurred by t_c will occur on $t_c + 1$. Since the null model underestimates as much as possible, any model that performs worse than it does must be doing so by overestimation.

5.2 Censoring

As discussed in Section 2.4, censoring refers to an observation in which the event is known to occur before or after, rather than precisely at, a certain point in time (for example, subjects 2, 4, and 5 in Figure 4.3 are censored). A basic assumption of survival analysis is that censoring is random (also referred to as non–informative). In technical terms, this means that there exists some density (or mass) $f_C(t)$ for time to censoring and a density (or mass) $f_E(t)$ for time to event, and the recorded time is the minimum of a single draw from each of the distributions. As long $f_E(t)$ and $f_C(t)$ as they are independent, then censored

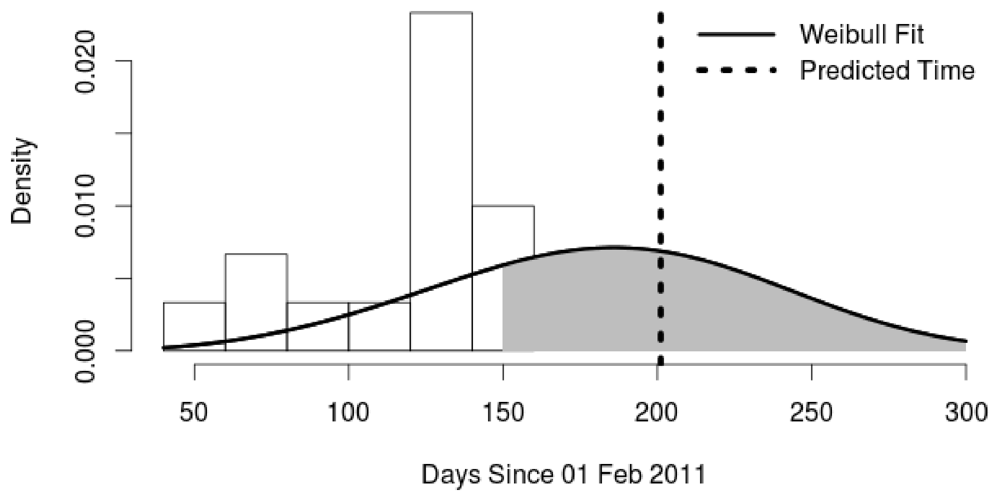


Figure 5.1: Histogram with Censored Univariate Fit and Prediction with Thresholding at 150 Days

observations will not be distributed differently than uncensored ones, and so knowing that an observation in censored does not provide any information about how long after the censoring point the event would have occurred.

Is this assumption justified in my case? (Thresholding all subjects at the same time is technically called Type 2 censoring.) Intuitively, it seems like it is not, since subjects with long event times will be censored but those with short event times will not be. However, this is simply because of the relative shape and scale of the $f_C(t)$ and $f_E(t)$. Specifically, the effect is achieved by assigning a lower variance to $f_C(t)$ than to $f_E(t)$. My scenario (Type 2 censoring) is a extreme case of this paradigm because censoring time is deterministic (i.e., could be thought of as having infinitesimally small variance). Overall, it is safe to say that the assumption about non–informativeness is satisfied because the two distributions are still independent.

5.3 Experiment Execution

I tested the predictive models as follows. Let t_c be a particular thresholding time. I arbitrarily selected thresholding times at five–day intervals. Next, all observations made after t_c were dropped since, by definition, they could not have been observed at that time. Observations corresponding to countries that had experienced the event were marked as uncensored ("dead"). Observations corresponding to countries that had yet to experience the event were marked as censored ("alive"). The resulting dataset is analogous to that shown in Figure 4.3. Finally, the time to event for countries that are "alive" was set to $t_c - t_m$, where t_m is the measurement time. Overall, this process restricted the dataset to information known at t_c. In other words, each threshold level makes a distinct train/test split. By sweeping through a range of threshold values, the performance of the model over a range of conditions could be observed. The resulting accuracy curves were plotted on accuracy–vs–time axes, such as those on Figure 6.1. The accuracy score at each time step was computed using the prediction for all future events, so the curve itself can only be computed at the final time step.

For the parametric models, I used the AFT implementation in the `survreg` package in R [52] (with `maxiter=200`, `rel.tolerance=1e-12`, `outer.max=50`). I did forward variable selection with the `step` function. For non–parametric modeling I used a variant on random forests called conditional forests, which I implemented with the `cforest` function (with $n = 100$ trees) in the R package `party` [53–55].

6 Results

6.1 Univariate Results

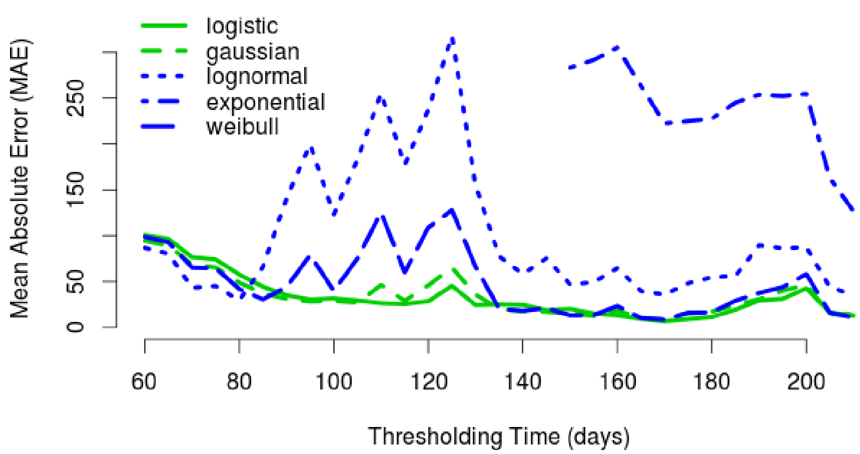


Figure 6.1: Accuracy Comparison for Univariate Prediction by Distribution

The results of univariate prediction, computed as shown in Figure 5.1, are shown in Figure 6.1 for a selection of distributions. All the models have roughly equal performance until approximately $t_c = 85$, when the asymmetric distributions (weibull and lognormal) diverge from those that are symmetric (gaussian and logistic) until approximately $t_c = 135$. Overall, the symmetric distributions are clearly superior to the asymmetric ones. For comparison, Figure 6.2 shows the histogram of all event times (that is, uncensored,

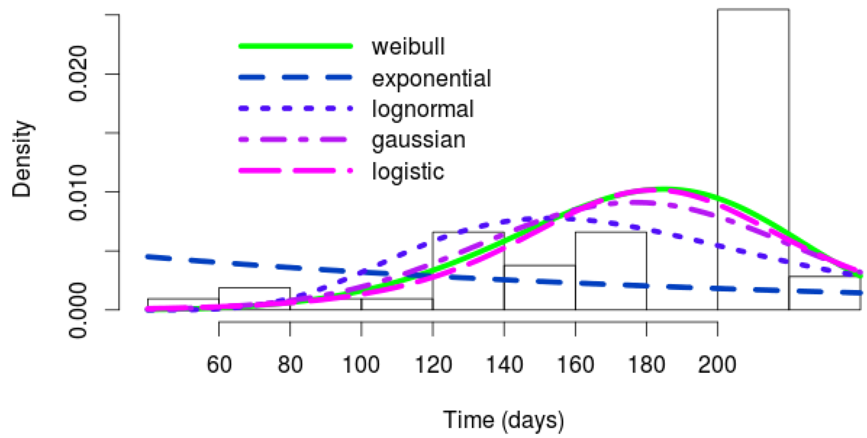


Figure 6.2: Histogram of Event Times With Parametric Distributions

in contrast to Figure 5.1) along with the corresponding fitted parametric distributions. Each time step in Figure 6.1 corresponds to distributions fitted on the data up to that time step. The erratic behavior of the predictions prior to $t_c = 135$ can be explained by the fact they have few observations from which to extrapolate, as shown in Figure 6.2.

6.2 Parametric Models

Results from AFT models using a logistic distribution (based on the univariate results) are shown in Figure 6.3. There is an obvious similarity between them and the univariate prediction results, which is to be expected since they both rely on the same parametric distribution. However, there are many points at which the AFT error rate is higher than that of the univariate, indicating that the addition of covariates is somehow impairing predictive performance.

6.3 Non–Parametric Model Results

As shown in Figure 6.4, the conditional forests almost always perform better than the null model, and usually do about as well as the best univariate model. Except for the interval from approximately $t_c = 105$ to $t_c = 130$, the parametric (AFT) results are better than those of the trees. After approximately $t_c = 185$, the tree models become worse than the parametric models as well both baselines, which is surprising since, as t_c grows, there is more training data available.

6.4 Ensemble Models

All of the parametric models discussed above (including those that are univariate) presuppose a certain distribution. Though the logistic and gaussian turn out to be the best when considered in retrospect, there

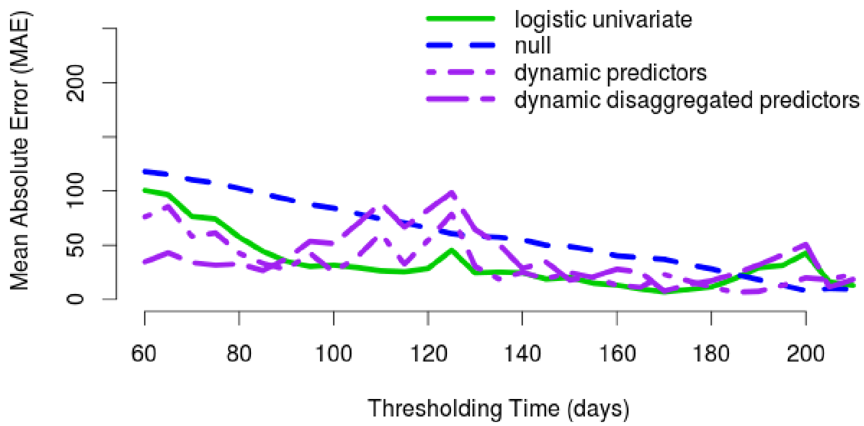


Figure 6.3: Accuracy of AFT Model on Dynamic Predictors

would be no way of knowing this fact prospectively. In other words, the models above do not take into account the full constraints of limited information. Suppose I take an alternative approach in which I fully constrain even the choice of distribution to information known at each t_c. That is, at each time, the distribution is selected automatically rather than being specified beforehand. I refer to this as the "ensemble" approach. An algorithmic description of this approach is shown in Algorithm 1.

Algorithm 1: Ensemble–like parametric modeling

Data: vector of t_c values $\{t_c^{(1)}, t_c^{(2)}, \ldots, t_c^{(n)}\}$; predictor matrix x; vector of true event times $\mathbf{t}$; vector of possible parametric distributions $\{f_1(\cdot), f_2(\cdot), \ldots, f_m(\cdot)\}$

Result: MAE measurement for each time step $\{e_{abs}^{(1)}, e_{abs}^{(2)}, \ldots, e_{abs}^{(n)}\}$

1 **for** $i \leftarrow 1:n$ **do**

2 $\quad(\tilde{x}, \tilde{\mathbf{t}}) \leftarrow doCensoring(x, \mathbf{t}, t_c^{(i)})$; `// remove observations after` t_c

3 $\quad$**for** $j \leftarrow 1:m$ **do**

4 $\quad\quad\tilde{f}_j(\cdot) \leftarrow fit(f_j(\cdot), \tilde{\mathbf{t}})$; `// fit each to previously-observed times`

5 $\quad$**end**

6 $\quad f^*(\cdot) \leftarrow argmax_j\{\mathscr{L}(\tilde{f}_j(\cdot) \,|\, \tilde{\mathbf{t}})\}$; `// get` $f()$ `with highest likelihood`

7 $\quad\hat{\mathbf{t}} \leftarrow extrapolate(\tilde{x}, \tilde{\mathbf{t}}, f^*(\cdot))$; `// train and predict`

8 $\quad e_{abs}^{(i)} \leftarrow MAE(\hat{\mathbf{t}}, \mathbf{t})$; `// calculate error`

9 **end**

In textual form, Algorithm 1 is simply saying the following: at each time step, select the distribution with the best fit (that is, the highest log likelihood) and make predictions with a model using that distribution. The model (line 7 of Algorithm 1) could be univariate or multivariate.

Results of the ensemble approach are shown in Figure 6.6. The univariate ensemble model starts off

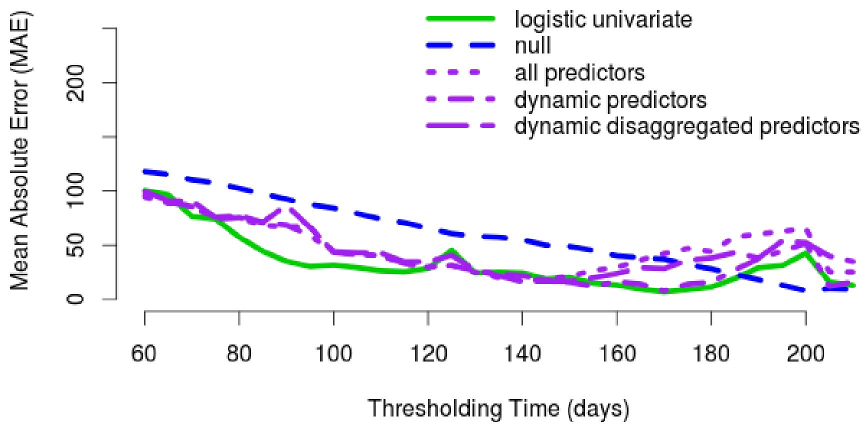


Figure 6.4: Accuracy of Conditional Forests

using the lognormal distribution around $t_c = 85$ and switches to the logistic or gaussian distribution around $t_c = 135$. Because of the switch to lognormal, the univariate ensemble model is clearly much worse than the logistic univariate model, and the multivariate ensemble model is even worse than that. Because the ensemble approach reflects realistic uncertainty (i.e., cannot know what the most accurate distribution will be), the shortcoming in the univariate ensemble models allows us to "beat" it with non–parametric models like conditional forests, as shown in Figure 6.6. In other words, once the logistic univariate is demoted as a baseline in favor of the ensemble univariate, models like conditional forests emerge as clearly superior.

7 Discussion

The fundamental hypothesis is that of predictive power of one or more of the given predictors for the time to recognition. There exists no previous work on predicting time to recognition, so I can only consider this hypothesis with the results at hand. I can reject the null hypothesis for conditional forests, as shown in Figure 6.4, depending on how I generalize beyond this particular process (recognition of the Libyan NTC). Do the non–parametric models just "get lucky" between $t_c = 85$ and $t_c = 135$ or did the parametric models just "get unlucky" on that interval? In other words, my rejection of the null hypothesis depends on how much uncertainty exists in the superiority of the logistic distribution. This uncertainty is very difficult to express because there exists no clear underlying mechanism to the process. One approach is to simply identify some distributions for which the ensemble–selected distribution shows relatively poor performance. However, only response data can be randomly generated, so I can only compare univariate models to each other and to a

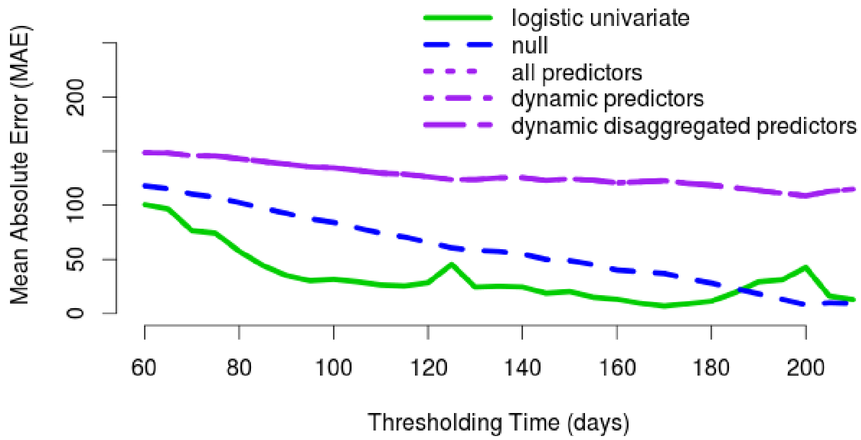


Figure 6.5: Accuracy of Random Forests

null model.

To address this problem I randomly generated event times that were distributed in a manner similar to that of the actual event times. The specific process was as follows. First, I drew five points from a uniform distribution between the minimum and maximum actual event times. Next, I used those to generate a kernel density estimate (gaussian kernel with $sd = 30$). Next, I ran univariate and ensemble prediction on 53 points (the number of countries in the pool) drawn from the kernel density estimate. To compare the results of these predictions I needed to compute a scalar performance metric, since I could not manually compare all the curves. I defined the overall performance score as the area under the curve on the maximum interval over which all univariate models were defined (some were undefined at early times because of insufficient simulated observations prior to that time).

Differences in these performance metrics between different models are shown in Figure 7.2. The horizontal axis of each histogram shows the difference between the performance of two model types. The areas of the histogram below zero correspond to simulation runs in which the first model performed better than the second, and areas of the histograms above zero correspond to simulation runs in which the second performed better than the first. The histogram of gaussian performance minus null performance (middle right–hand side) shows that the gaussian sometimes performs worse than the null (tail above zero). Further, the histogram of ensemble–selected model performance minus null performance (bottom right–hand corner) shows that the ensemble–selected model can also do worse than the null. Overall, this indicates that I must have at least some uncertainty about whether or not the univariate models will do better than the null, which means I have more uncertainty about whether or not they will do better than the conditional forest model.

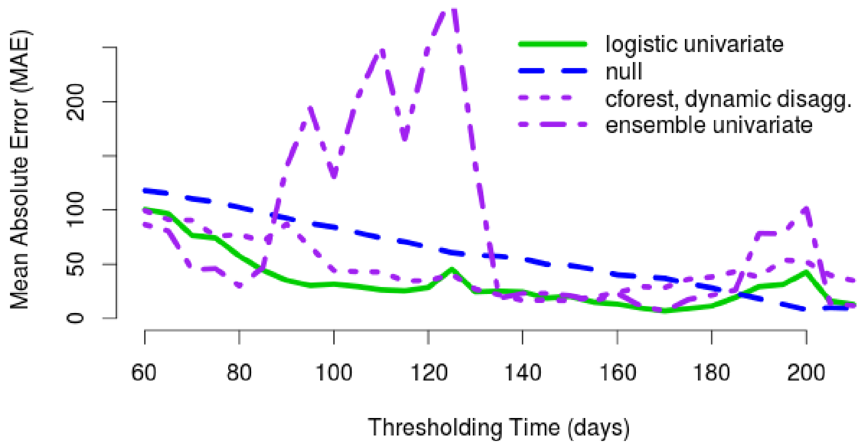


Figure 6.6: Accuracy for Ensemble Models

As shown in Figure 6.4, essentially all the predictive power lies with the dynamic disaggregated features, which are those based on the combination of affinity metrics with the recognition status vector. Because the status vector is responsible for the dynamic nature of these features, they are essentially autoregressive. That is, they represent the history of the process itself rather than any dynamic exogenous features.

7.1 Conversion to Static Classification

As discussed above, the predictive power resides in the features representing the links between countries. Specifically, the dataset includes links that reflect voting affinity, diplomatic exchange, trade relationships, and formal alliances. As discussed in Section 4.2.1, these links (which are, by definition, static and disaggregated) were combined with the recognition status vector (which is, by definition, dynamic and aggregated) to create the dynamic disaggregated features. To further test the effect of the these static features, one could convert the problem from a time series regression to a static classification. That is, rather than predicting time to event, I could predict the occurrence of the event before or after a certain time threshold. This conversion would render the model useless in a practical sense, since one would have to wait until the threshold passed to collect the training data on both sides of the threshold. At that point, by definition, predicting which countries lay on which side would be trivial. Thus, classifying event times is purely of theoretical interest in this case. Specifically, it would allow us to compare a classical *iid* approach with a collective classification. However, it is worth noting that such a classification approach is perfectly suited to predicting diplomatic decisions in contexts for which all decisions are made simultaneously, such as that of a vote in an international body. Such applications are further explored in Section 8.2.

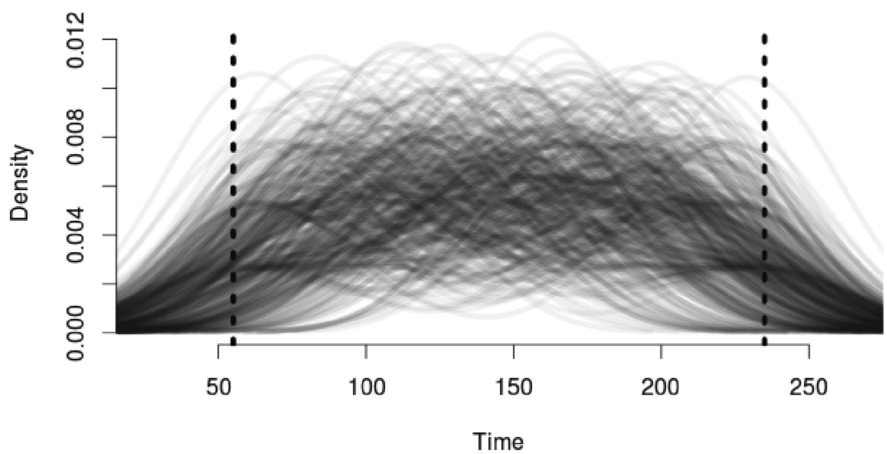


Figure 7.1: Automatically Generated Distributions

I set up a simple static classification problem as follows. First, I converted all the static disaggregated variables to binary values. Where necessary, I used the mean of the variables across all countries as the threshold. Next, I converted the dyadic relationships to binary values, using the mean across all dyads as the threshold. Finally, I set the response variable for each country to be whether or not it recognized the NTC before or after August 20, 2011. This was a convenient temporal threshold because shortly thereafter the rebel army took control of the city of Tripoli, which prompted a large number of countries to grant recognition. Next, I set up three types of models, each to be evaluated with $n-$fold cross–validation. First, I set up models with no dyadic data (that is, purely static disaggregated features). I used logistic regression and neural networks with this type of model. (Specifically, I used R's `glm()` function with `family="binomial"` for the logistic regression and the `nnet()` function with 2 hidden layers for the neural networks.) Second, I set up models with the dynamic disaggregated features as well as dyadic features as explicit binary variables. Specifically, I added a column for each country and each dyadic link type (of which there were four, as discussed in Section 3.2.1). That is, for every country i (represented by the ith row), there was a column indicating whether or not country i had the kth type of link with the jth country. The addition of dyadic link information in this manner extended the length of the feature vector to several time the number of countries. As above, I used logistic regression and neural networks with this type of model. Third, I set up a simple markov logic network (MLN) model with static disaggregated features and dyadic links represented by predicate logic. MLNs are a type of statistical relational learning (SRL), which is an approach to handling data that is not composed of independent and identically distributed (*iid*) observations. In this specific case, MLNs allow us to take into account the non–independence represented by the dyadic links between countries. Technical details can be found in Appendix C.

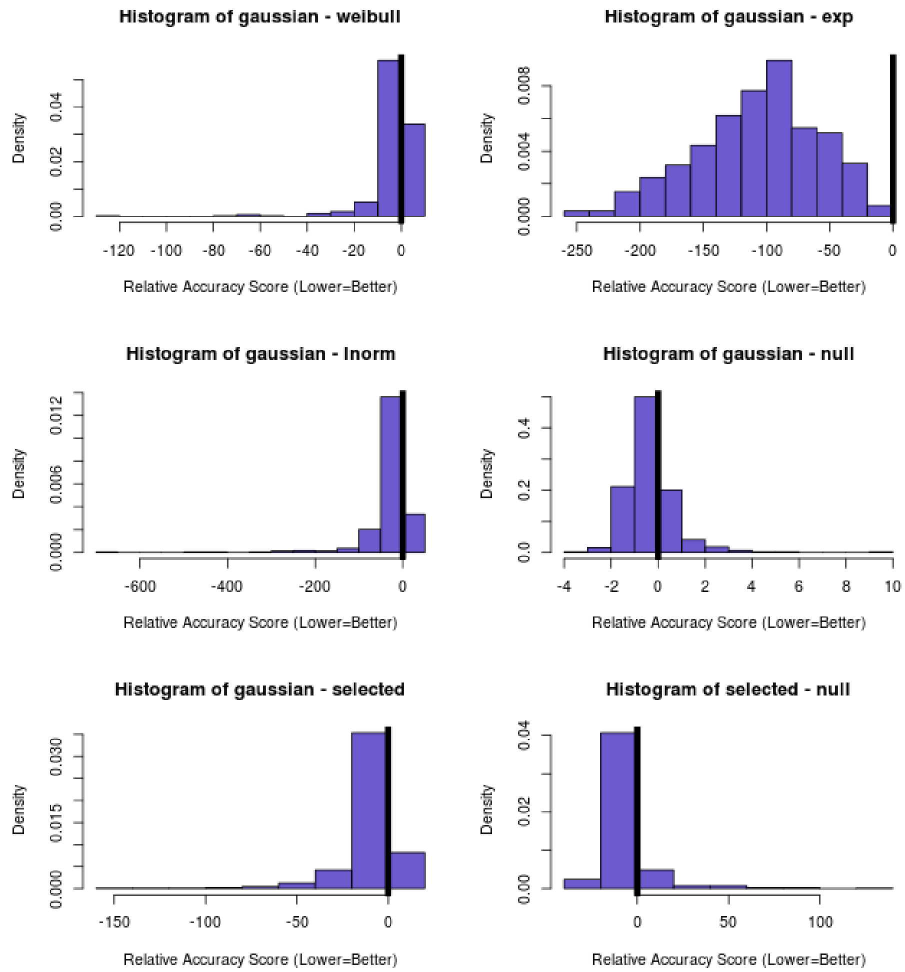


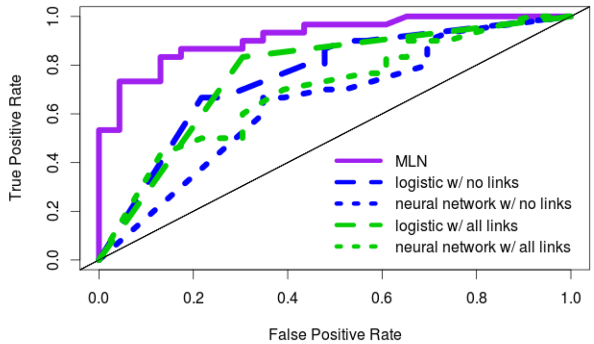
Figure 7.2: Histograms of Accuracy Score Differences

Figure 7.3 shows the results of the five different models outlined above. Generally, these results show conclusively that the static disaggregated features and the link data do have some predictive power. More specifically, the results indicate that MLNs perform at least as well and usually better than the conventional alternatives. This is what one would expect given that MLNs are designed to handle non–independent observations.

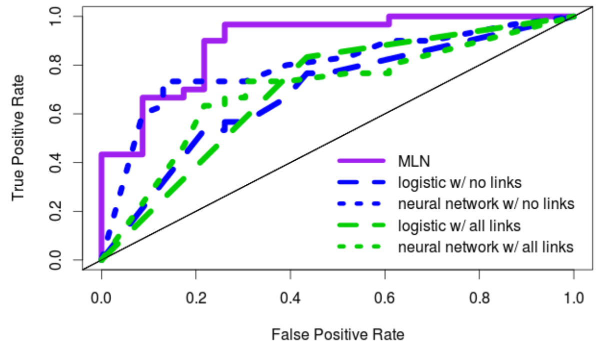
8 Conclusion

8.1 Summary of Research Problem and Results

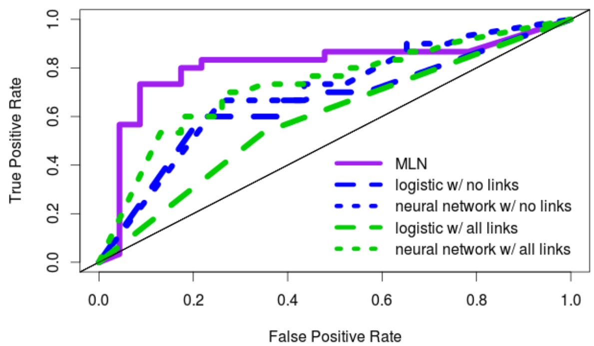
The goal of the present work was to formulate spatially and temporally–varying features and test how well they predict time to recognition of the Libyan NTC. Specifically, I identified and formulated features that



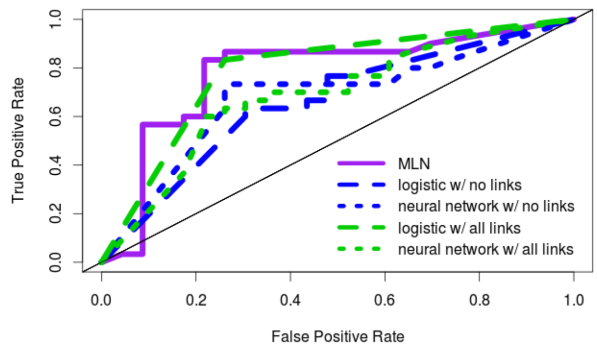
(a) 5 folds



(b) 10 folds



(c) 25 folds



(d) 53 folds

Figure 7.3: ROC Plot for Recognition Classification With $n-$fold Cross–Validation

were temporally dynamic but spatially constant (dynamic aggregated features), others that were temporally constant but spatially varying (static disaggregated features), and still others that varied in both space and time (dynamic disaggregated features). I tested the predictiveness of these features with several different models, including parametric and non–parametric formulations. I compared those results to a null baseline (predicting that all events will occur on the next day) as well as several univariate baselines using different parametric distributions. The best results are those of conditional forests (non–parametric multi–variate model) and are comparable to those of the univariate baselines with gaussian or logistic distributions. The conditional forest models can be considered superior because, being non–parametric, they do not rely on the assumption of any particular distribution as the univariate baselines do. However, we should refrain from attributing any significance to the structure of the conditional forest models since they were beaten by univariate models, which are, by definition, spatially non–discriminative (meaning that they make the same prediction for all countries at each time step). In other words, the fact that there exists at least one spatially non–discriminative model that does better than the conditional forest model (which does discriminate between countries) casts very serious doubt on the idea that the structure of the conditional forest models could tell us anything about the mechanism underlying the recognition process. Overall, however, evidence for the predictive power of the dyadic features can be found in the relatively strong performance of the MLNs in the static classification.

8.2 Impacts

As discussed in Section 1, prediction in the political domain in general and prediction of conflict progression in particular is important for two reasons. First, it enables decision makers to better plan and prepare for future events. Second, prediction can, at least in some cases, prevent the escalation of conflict in the first place by reducing the uncertainty associated with the outcome. This work in particular provides one case of such prediction specifically for NTC recognition.

Because little to no predictive power was found in the features examined in this work, it is difficult to identify any policy implications. However, three limited conclusions may be drawn. First, the very lack of predictive power found may suggest that some countries would be open to incentives to recognize sooner or later. In other words, since there appears to be little pattern to the recognition decisions, it is possible that they are being made at the discretion of a few key decision–makers, who could be influenced through the usual diplomatic channels. Second, since the predictiveness that was found resided in the dyadic features, these decision–makers are to some extent taking into account the decisions of their allies and enemies, which means that the decisions of many members of the international community could be influenced by incentivizing a few key countries. Third, it might be advisable to offer incentives early enough in the course of the conflict to affect its overall trajectory, but not so early as to waste resources incentivizing countries that would recognize early anyway. As shown in Figure 3.1, the recognition process stagnated between $t = 180$ and 200 days, before spiking around 200, corresponding to a stalemate followed by the capture of Tripoli by the rebels. Incentives, if offered early enough, could potentially have prevented such stagnation and accelerated rebel victory.

In the more general sense, recognition could be thought of as just one example of a voting process. Thus, by extension, these results are germane to any process by which international actors make individual decisions on a shared issue, both in formal context, like the UN General Assembly or official summits on issues like climate or human rights, as well as informal contexts, like making voluntary offers of humanitarian or military aid. In fact, the present work is a more challenging form of this general class of problems, since the survival response is essentially a latent measure of the underlying proclivity of each country's government to recognize to the NTC; the countries that waited longer needed to be convinced further before making a commitment. Other cases would provide indicators that are less ambiguous, such as simple yes/no votes or commitments to control carbon emissions to a certain extent.

8.3 Limitations

Overall, the results give some evidence that dynamic disaggregated factors are predictive of diplomatic decisions such as recognition. However, the weakness of the evidence suggests at least two areas for improvement.

8.3.1 Data Quality

The data I used was available from open sources and was readily accessible (notwithstanding the difficulties of processing). Data that is more theoretically well–informed and more specific to the political context would likely improve the results. For example, I included only very simple factors to represent economic and diplomatic relationships. More complex factors (such as those more specific to the oil industry) may yield better results. Further, more detailed static disaggregated, such as those representing differences in political structure rather than simply location on the autocracy/democracy scale as with the polity score. One especially salient shortcoming is the uncertainty in the recognition times, which turned out to be a more difficult dataset to gather than initially anticipated. Better sources of information about diplomatic decisions could go far in improving their prediction.

8.3.2 Problem Formulation

The major shortcoming of these results is that the data that best describes the process being modeled fundamentally violates the classical assumption of independent and identically distributed (*iid*) observations. In general, the task of building an *iid* dataset consists of identifying a feature space of which a sample of instantiations can be observed. The necessity for repetition is inherent to inductive reasoning itself, which by definition relies on having a multiplicity of specific cases from which to generalize. This task of identifying a feature space has been called the "reference class" problem [56], since every observation is measured with reference to the abstract class defined by the feature space. For example, the performance of an cyclist is modeled with reference to other cyclists, or the cyclist's own performance in the past. A cyclist's performance is not modeled with reference to gymnasts, because gymnasts belong to a distinct reference class. The reference class problem is particularly difficult because virtually all of machine learning and data mining operates on the assumption that it has already been solved. To put it into philosophical terms, one must answer ontological questions about what entities exist in the world before one can answer epistemological questions regarding what is true of those entities. The reference class problem is the specific ontological question that precedes conventional statistical modeling, machine learning, and data mining. Though not practically relevant to the present work, it is helpful to note that the reference class problem also plays an important role in epistemological thought experiments like the Sleeping Beauty problem and the doomsday argument [57], as well as in the measure problem in contemporary cosmology [58].

In some machine learning application areas like medical decision support or economic forecasting the reference class problem is trivial to solve because of well–defined ontology. For example, economists have well–defined indices that can be measured at time t to predict their value at some time $t' = t + h$ in the future, where h is a time horizon, but the variables that represent the state of a political or military conflict at time t may not even be meaningful or measurable at $t' = t + h$.

In addition to the general reasons why sociopolitical data is non–*iid*, the data for this specific problem is not *iid* even when simplified to a static format, as explored in Section 7.1, and even though I selected it for the purpose of *iid*–style modeling. It is important to note that this problem of data quality is concomitant with a problem of quantity. Simply obtaining the amount of data discussed in Section 3 has been the primary difficulty in developing this type of model. This was not due to a lack of information in a general sense (as discussed in Section 3.2.2, a handful of media sources generated around 24 000 articles related to Libya during the relevant time frame). Rather, the problem is specifically a lack of data organized according to the *iid* assumption. Thus, the most crucial improvement needed is a means of capitalizing on non–*iid* data. In metaphorical terms, modeling sociopolitical phenomena such as the Libyan Civil War is challenging not because the territory is unexplored but because it is unmapped. First–order logic, of which Section 7.1 provided only a very limited example, has promising flexibility for such problems.

8.4 Future Work

In terms of application, the present work explores predictive modeling of only one among many types of decisions in the political domain, as discussed in Section 8.2. In terms of methodology, Section 8.3.2 describes how a predicate logic approach can take advantage of non–*iid* data. Crucially, however, it does so not by eliminating the reference class problem (or more generally, the ontological problem), but by providing a more flexible way to solve it. In other words, even if we do not organize the data into single reference class, we still have to organize it into some ontological structure, and then find a way to make inference within that structure. For example, if we were to organize the countries into several different reference classes, each with its own feature vector, we would also need corresponding multi–class predictive models. Developing such models is the primary challenge of future work, since by definition they lie almost entirely outside the paradigm of conventional machine learning and data mining.

Acknowledgements

Many thanks to Dr. John Owen in the Department of Politics for his advice on recognition and Peter Wu for assistance on various programming and modeling problems.

A Recognition Dates and Sources

Table A.1: Recognition Date and Sources

Government	Level	Date	URL
Qatar	3	2011-03-28	http://www.wsj.com/articles/SB10001424052748704559904576228341860730126
Italy	4	2011-04-04	http://www.esteri.it/MAE/EN/Sala_Stampa/ArchivioNotizie/Approfondimenti/2011/04/20110404_FocusLibia_frattini_Cnt.htm
Kuwait	3	2011-04-13	http://sputniknews.com/voiceofrussia/2011/04/22/49293259/
Jordan	3	2011-05-24	http://www.cnn.com/2011/WORLD/africa/05/24/libya.jordan.council/
Spain	3	2011-06-08	http://af.reuters.com/article/egyptNews/idAFLDE7571E720110608
Australia	3	2011-06-09	http://www.jpost.com/Breaking-News/Australia-says-recognizes-Libya-rebel-council
United Arab Emirates	3	2011-06-12	http://af.reuters.com/article/libyaNews/idAFLDE75B0DW20110612
Germany	3	2011-06-13	http://www.bbc.co.uk/news/world-europe-13753422
Canada	3	2011-06-14	http://www.bbc.co.uk/news/world-us-canada-13771946
Austria	3	2011-06-18	http://www.shabablibya.org/news/austria-recognises-libyan-rebels
Latvia	3	2011-08-29	http://www.latvija.kz/en/news/press-releases/2011/august/29-3/
Lithuania	3	2011-06-20	http://www.urm.lt/default/en/news/in-vilnius-lithuanian-foreign-vice-minister-meets-with-representatives-of-libyas-transitional-council-
Bulgaria	3	2011-06-28	http://www.sofiaecho.com/2011/06/28/1113888_bulgaria-and-croatia-recognise-libyas-transitional-national-council
Croatia	3	2011-06-28	http://www.sofiaecho.com/2011/06/28/1113888_bulgaria-and-croatia-recognise-libyas-transitional-national-council
Turkey	3	2011-07-03	http://news.xinhuanet.com/english2010/world/2011-07/04/c_13963431.htm
Poland	3	2011-05-11	http://www.eubusiness.com/news-eu/libya-conflict.b64
Belgium	3	2011-07-13	http://www.reuters.com/article/2011/07/13/us-libya-rebels-recognition-idUSTRE76C1P120110713
Luxembourg	3	2011-07-13	http://www.reuters.com/article/2011/07/13/us-libya-rebels-recognition-idUSTRE76C1P120110713
Netherlands	3	2011-07-13	http://www.reuters.com/article/2011/07/13/us-libya-rebels-recognition-idUSTRE76C1P120110713

Table A.2: Recognition Date and Sources (cont.)

Government	Level	Date	URL
Albania	3	2011-07-18	http://www.balkaninsight.com/en/article/albania-recognizes-libya-s-rebel-government
Slovenia	3	2011-07-20	http://www.vlada.si/en/media_room/government_press_releases/press_release/article/the_government_of_slovenia_recognises_the_libyan_national_transitional_council_18693/
Montenegro	3	2011-07-21	http://www.gov.me/en/News/107518/Government-of-Montenegro-recognises-Libyan-National-Transitional-Council.html
Portugal	4	2011-07-28	http://www.pcp.pt/en/decision-portuguese-government-recognize-%E2%80%9Cnational-transitional-council-libya%E2%80%9D
Tunisia	3	2011-08-21	http://english.alarabiya.net/articles/2011/08/21/163403.html
Egypt	4	2011-08-22	http://www.dailynewsegypt.com/2011/08/22/egypt-recognizes-libyan-ntc-arab-league-offers-support/
Ireland	4	2011-08-22	https://www.dfa.ie/news-and-media/press-releases/press-release-archive/2011/august/developments-in-libya/
Morocco	3	2011-08-22	http://allafrica.com/stories/201108231326.html
New Zealand	3	2011-08-22	http://www.radionz.co.nz/news/political/83167/nz-to-recognise-libya%27s-transitional-national-council
Bahrain	3	2011-08-23	http://www.todayszaman.com/latest-news_bahrain-recognizes-libyan-rebels_254642.html
Greece	3	2011-08-23	http://news.xinhuanet.com/english2010/world/2011-08/24/c_131069716.htm
Nigeria	3	2011-08-23	http://www.voanews.com/content/nigeria-recognizes-libya-rebels-128255143/158719.html
Norway	4	2011-08-23	http://www.newsinenglish.no/2011/08/24/norway-backs-libya-amidst-scolding/
Burkina Faso	3	2011-08-24	http://www.mae.gov.bf/miseajour/Declaration-BF-sur-la-situation-en-Libye.htm
Chad	3	2011-08-24	http://blogs.aljazeera.com/topic/libya/libya-aug-24-2011-2025
Ethiopia	4	2011-08-24	https://now.mmedia.me/lb/en/archive/ethiopia_recognizes_libyan_rebels
Hungary	4	2011-08-24	http://www.politics.hu/20110824/hungary-recognises-rebel-council-as-legitimate-representative-of-libya/
Sudan	3	2011-08-24	http://news.xinhuanet.com/english2010/world/2011-08/24/c_131071287.htm

Table A.3: Recognition Date and Sources (cont.)

Government	Level	Date	URL
Bosnia and Herzegovina	3	2011-08-25	http://www.klix.ba/vijesti/bih/predsjednistvo-bih-priznalo-novu-libijsku-vlast/110825093
Ivory Coast	3	2011-08-25	https://now.mmedia.me/lb/en/archive/ivory_coast_recognizes_libyan_rebels
Cyprus	3	2011-08-26	http://www.khaleejtimes.com/DisplayArticle09.asp?xfile=data/international/2011/August/international_August1246.xml&section=international
Estonia	3	2011-06-28	http://vm.ee/en/news/estonian-and-italian-foreign-ministers-agree-violence-libya-must-end-soon
Togo	3	2011-08-27	http://www.republicoftogo.com/Toutes-les-rubriques/Diplomatie/Le-Togo-reconnait-la-victoire-des-rebelles
Czech Republic	3	2011-08-29	http://www.radio.cz/en/section/news/czech-republic-recognizes-libyan-national-transitional-council
Slovakia	4	2011-08-30	http://www.mzv.sk/servlet/tripoliszu?MT=/App/WCM/ZU/TripolisZU/main.nsf/vw_ByID/ID_621F5291AE4A5FD4C125715B004FFE51_EN&OpenDocument=Y&LANG=EN&OB=1001&HM=0-XXX&NCH=Y&%20DS=Y&TG=BlankMaster&URL=/App/WCM/Aktualit.nsf/vw_ByID/ID_A1955CD90279252CC12578FC004974FA
Finland	4	2011-09-01	http://formin.finland.fi/public/default.aspx?contentid=227728&nodeid=15146&contentlan=2&culture=en-US
Romania	4	2011-09-01	https://now.mmedia.me/lb/en/archive/romania_recognizes_libyan_rebels
Russia	4	2011-09-01	http://www.mid.ru/brp_4.nsf/0/95F1415C6C130007C32578FE0028500D
Ukraine	3	2011-09-01	http://en.interfax.com.ua/news/general/78179.html
Central African Republic	3	2011-09-05	http://www.radiondekeluka.org/politique/item/5204-la-rca-reconnait-la-l%C3%A9gitimit%C3%A9-du-cnt-en-libye
Ghana	4	2011-09-09	http://www.ghananewsagency.org/politics/ghana-recognizes-libyan-national-transitional-council-33335
South Africa	3	2011-09-20	http://www.dfa.gov.za/docs/2011/liby0921.html
Algeria	3	2011-09-22	http://uk.reuters.com/article/2011/09/22/libya-algeria-report-idUKLDE78L04M20110922
Kenya	3	2011-09-24	http://www.nation.co.ke/News/Kenya+embraces+new+leadership+in+Libya+/-/1056/1242284/-/11766tjz/-/index.html

B Variables From WDI

Table B.1: Variables Selected from WDI Dataset

Abbreviation	Description
AG.LND.AGRI.ZS	Agricultural land (% of land area)
AG.PRD.CROP.XD	Crop production index (2004-2006 = 100)
AG.SRF.TOTL.K2	Surface area (sq. km)
BX.KLT.DINV.CD.WD	Foreign direct investment, net inflows (BoP, current US$)
EG.ELC.ACCS.ZS	Access to electricity (% of population)
EG.NSF.ACCS.ZS	Access to non-solid fuel (% of households)
EN.ATM.CO2E.KT	CO2 emissions (kt)
EN.POP.DNST	Population density (people per sq. km of land area)
EP.PMP.DESL.CD	Pump price for diesel fuel (US$ per liter)
ER.LND.PTLD.ZS	Terrestrial protected areas (% of total land area)
ER.PTD.TOTL.ZS	Terrestrial and marine protected areas (% of total territorial area)
FP.CPI.TOTL	Consumer price index (2010 = 100)
IT.CEL.SETS	Mobile cellular subscriptions
IT.MLT.MAIN	Telephone lines
IT.NET.USER.P2	Internet users (per 100 people)
PA.NUS.ATLS	DEC alternative conversion factor (LCU per US$)
SE.PRM.DURS	Primary education, duration (years)
SE.SEC.DURS	Secondary education, duration (years)
SG.GEN.PARL.ZS	Proportion of seats held by women in national parliaments (%)
SH.ANM.CHLD.ZS	Prevalence of anemia among children (% of children under 5)
SH.DTH.IMRT	Number of infant deaths
SH.DYN.NMRT	Mortality rate, neonatal (per 1,000 live births)
SH.H2O.SAFE.UR.ZS	Improved water source, urban (% of urban population with access)
SH.IMM.IDPT	Immunization, DPT (% of children ages 12-23 months)
SH.MMR.RISK	Lifetime risk of maternal death (1 in: rate varies by country)
SH.PRG.ANEM	Prevalence of anemia among pregnant women (%)
SH.STA.MMRT	Maternal mortality ratio (modeled estimate, per 100,000 live births)
SH.TBS.INCD	Incidence of tuberculosis (per 100,000 people)
SH.XPD.PCAP	Health expenditure per capita (current US$)
SL.EMP.TOTL.SP.ZS	Employment to population ratio, 15+, total (%) (modeled ILO estimate)
SL.TLF.CACT.ZS	Labor force participation rate, total (% of total population ages 15+) (modeled ILO estimate)
SL.UEM.TOTL.ZS	Unemployment, total (% of total labor force) (modeled ILO estimate)
SM.POP.TOTL	International migrant stock, total
SP.ADO.TFRT	Adolescent fertility rate (births per 1,000 women ages 15-19)
SP.DYN.CBRT.IN	Birth rate, crude (per 1,000 people)
SP.POP.DPND	Age dependency ratio (% of working-age population)
SP.RUR.TOTL	Rural population
SP.URB.TOTL	Urban population

C Markov Logic Networks

Markov logic networks are (MLNs) are a popular type of statistical relational learning (SRL) model introduced in [59]. I used the Alchemy implementation [60]. In this case I are using them to do the same task as conventional machine learning models. The analog of a predictor–response dataset is a file of instantiations of predicates with literal arguments (called "ground atoms"), and the analog of a model structure is a set of rules, expressed in first–order logic, to which Alchemy assigns weights according to how well they fit the training data. For prediction (that is, inference), Alchemy uses the weighted rules to estimate the truth value of a query predicate, which in this case is the response variable.

For the ground atoms, static disaggregated features represented with binary predicates of the form $\texttt{hasProperty}(< \textit{country id} >, < \textit{feature name} >)$ and dyadic links represented with ternary predicates of the form $\texttt{areLinkedBy}(< \textit{country id} >, < \textit{country id} >, < \textit{link type} >)$. The response variable was coded as a unary predicate of the form $\texttt{beforeThreshold}(< \textit{country id} >)$.

This was implemented in Alchemy with the following two rules: First, a simple rule for correlation between static disaggregated properties and the response class:

$$\texttt{hasProperty}(c, +f) \rightarrow \texttt{beforeThreshold}(c)$$

Second, a rule indicating that linked countries are likely to share recognition decisions:

$$\texttt{areLinkedBy}(c_1, c_2, +t) \rightarrow (\texttt{beforeThreshold}(c_1) \longleftrightarrow \texttt{beforeThreshold}(c_2))$$

Bibliography

[1] J. D. Fearon, "Rationalist explanations for war," *International organization*, vol. 49, no. 03, pp. 379–414, 1995.

[2] P. Worsnip, "UN council wants to free more Libyan assets-envoy," *Reuters*, Nov. 2011. [Online]. Available: http://af.reuters.com/article/energyOilNews/idAFN1E7AT16U20111130

[3] S. R. Power, "The role of the National Transitional Council in the Economic Reconstruction of LibyaSome Legal Challenges," *Socio-Legal Studies Review*, vol. 1, 2012.

[4] J. Van Essen, "De Facto Regimes in International Law," *Merkourios*, vol. 28, no. 74, pp. 31–49, 2012. [Online]. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2017736

[5] S. Talmon, "Recognition of the Libyan National Transitional Council," *ASIL Insights*, vol. 15, no. 16, p. 13, 2011. [Online]. Available: http://www.asil.org/sites/default/files/insight110616.pdf

[6] S. F. Halabi, "Traditions of Belligerent Recognition: The Libyan Intervention in Historical and Theoretical Context," *Am. U. Int'l L. Rev.*, vol. 27, pp. 321–991, 2012.

[7] A. ubuku, "The Responsibility to Protect: Libya and the Problem of Transnational Solidarity," *Journal of Human Rights*, vol. 12, no. 1, pp. 40–58, Jan. 2013. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/14754835.2013.754291

[8] B. S. Minteh, "Recognition of Libya's Revolutionary National Transitional Council - Diplomatic Policy Variations of Nigeria and South Africa," *SSRN Electronic Journal*, 2012. [Online]. Available: http://www.ssrn.com/abstract=2177642

[9] M. Dickson, "Nigeria's Recognition of The Transitional National Council of Libya: Exploration and Critical Assessment." *British Journal of Arts & Social Sciences*, vol. 14, no. 1, 2013.

[10] P. A. Schrodt and D. J. Gerner, "Empirical Indicators of Crisis Phase in the Middle East, 1979-1995," *Journal of Conflict Resolution*, vol. 41, no. 4, pp. 529–552, Aug. 1997. [Online]. Available: http://jcr.sagepub.com/cgi/doi/10.1177/0022002797041004003

[11] P. A. Schrodt, J. Yonamine, and B. E. Bagozzi, "Data-based Computational Approaches to Forecasting Political Violence," in *Handbook of Computational Approaches to Counterterrorism*, V. Subrahmanian, Ed. New York, NY: Springer New York, 2013, pp. 129–162. [Online]. Available: http://link.springer.com/10.1007/978-1-4614-5311-6_7

[12] Y. M. Zhukov, "An Epidemic Model of Violence and Public Support in Civil War," *Conflict Management and Peace Science*, vol. 30, no. 1, pp. 24–52, Feb. 2013. [Online]. Available: http://cmp.sagepub.com/cgi/doi/10.1177/0738894212456936

[13] E. H. Ritter, "Policy Disputes, Political Survival, and the Onset and Severity of State Repression," *Journal of Conflict Resolution*, vol. 58, no. 1, pp. 143–168, Jan. 2013. [Online]. Available: http://jcr.sagepub.com/cgi/doi/10.1177/0022002712468724

[14] S. Shellman, "Leaders' Motivations and Actions: Explaining Government-Dissident Conflict-Cooperation Processes," *Conflict Management and Peace Science*, vol. 23, no. 1, pp. 73–90, Feb. 2006. [Online]. Available: http://taylorandfrancis.metapress.com/Index/10.1080/07388940500503820

[15] J. H. Pierskalla, "Protest, Deterrence, and Escalation: The Strategic Calculus of Government Repression," *Journal of Conflict Resolution*, vol. 54, no. 1, pp. 117–145, Dec. 2009. [Online]. Available: http://jcr.sagepub.com/cgi/doi/10.1177/0022002709352462

[16] K. H. Leetaru, "Can We Forecast Conflict? A Framework for Forecasting Global Human Societal Behavior Using Latent Narrative Indicators," 2013. [Online]. Available: http://www.kalevleetaru.com/Publish/Leetaru_Dissertation_Can_We_Forecast_Conflict-Dissertation.pdf

[17] A. Zammit-Mangion, M. Dewar, V. Kadirkamanathan, and G. Sanguinetti, "Point process modelling of the Afghan War Diary," *Proceedings of the National Academy of Sciences*, vol. 109, no. 31, pp. 12 414–12 419, Jul. 2012. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1203177109

[18] P. T. Brandt, J. R. Freeman, T.-m. Lin, and P. A. Schrodt, "Forecasting Conflict in the Cross-Straits: Long Term and Short Term Predictions," 2013. [Online]. Available: http://www.utdallas.edu/~pbrandt/Patrick_Brandts_Website/Research_files/ForecastWindows.pdf

[19] J. C. Pevehouse and J. S. Goldstein, "Serbian Compliance or Defiance in Kosovo? Statistical Analysis and Real-Time Predictions," *The Journal of Conflict Resolution*, vol. 43, no. 4, pp. 538–546, Aug. 1999. [Online]. Available: http://www.jstor.org/stable/174616

[20] V. B. Petroff, J. H. Bond, and D. H. Bond, "Using hidden markov models to predict terror before it hits (again)," in *Handbook of Computational Approaches to Counterterrorism*. Springer, 2013, pp. 163–180. [Online]. Available: http://link.springer.com/chapter/10.1007/978-1-4614-5311-6_8

[21] P. Schrodt, "Forecasting Conflict in the Balkans using Hidden Markov Models," in *Programming for Peace*, ser. Advances in Group Decision and Negotiation, R. Trappl, Ed. Springer Netherlands, 2006, vol. 2, pp. 161–184. [Online]. Available: http://dx.doi.org/10.1007/1-4020-4390-2_8

[22] S. M. Shellman, B. P. Levey, and J. K. Young, "Shifting sands: Explaining and predicting phase shifts by dissident organizations," *Journal of Peace Research*, vol. 50, no. 3, pp. 319–336, May 2013. [Online]. Available: http://jpr.sagepub.com/cgi/doi/10.1177/0022343312474013

[23] M. M. Hussain and P. N. Howard, "What Best Explains Successful Protest Cascades? ICTs and the Fuzzy Causes of the Arab Spring," *International Studies Review*, vol. 15, no. 1, pp. 48–66, Mar. 2013. [Online]. Available: http://doi.wiley.com/10.1111/misr.12020

[24] J. A. Goldstone, R. H. Bates, D. L. Epstein, T. R. Gurr, M. B. Lustik, M. G. Marshall, J. Ulfelder, and M. Woodward, "A global model for forecasting political instability," *American Journal of Political Science*, vol. 54, no. 1, pp. 190–208, 2010. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1540-5907.2009.00426.x/full

[25] E. Miguel, S. Satyanath, and E. Sergenti, "Economic Shocks and Civil Conflict: An Instrumental Variables Approach," *Journal of Political Economy*, vol. 112, no. 4, pp. 725–753, Aug. 2004. [Online]. Available: http://www.jstor.org/stable/10.1086/421174

[26] A. Sliva, G. Simari, V. Martinez, and V. S. Subrahmanian, "SOMA: Stochastic opponent modeling agents for forecasting violent behavior," in *Handbook of Computational Approaches to Counterterrorism*. Springer, 2013, pp. 99–127. [Online]. Available: http://link.springer.com/chapter/10.1007/978-1-4614-5311-6_6

[27] G. I. Simari, D. Earp, M. V. Martinez, A. Sliva, and V. S. Subrahmanian, "Forecasting group-level actions using similarity measures," in *Handbook of Computational Approaches to Counterterrorism*. Springer, 2013, pp. 181–199. [Online]. Available: http://link.springer.com/chapter/10.1007/978-1-4614-5311-6_9

[28] S. Shellman, "Measuring the Intensity of Intranational Political Events Data: Two Interval-Like Scales," *International Interactions*, vol. 30, no. 2, pp. 109–141, Apr. 2004. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/03050620490462603

[29] J. Yonamine, *Working with Event Data: A Guide to Aggregation Choices1*. MS, 2011. [Online]. Available: http://jayyonamine.com/wp-content/uploads/2013/04/Working-with-Event-Data-A-Guide-to-Aggregation-Choices.pdf

[30] P. Tetlock, *Expert political judgment: How good is it? How can we know?* Princeton University Press, 2005.

[31] M. Bowden, *The finish: the killing of Osama Bin Laden*. New York: Atlantic Monthly Press : Distributed by Publishers Group West, 2012.

[32] J. A. Friedman and R. Zeckhauser, "Handling and Mishandling Estimative Probability: Likelihood, Confidence, and the Search for Bin Laden," *Intelligence and National Security*, vol. 30, no. 1, pp. 77–99, Jan. 2015. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/02684527.2014.885202

[33] S. P. OBrien, "A multi-method approach for near real time conflict and crisis early warning," in *Handbook of Computational Approaches to Counterterrorism*. Springer, 2013, pp. 401–418. [Online]. Available: http://link.springer.com/chapter/10.1007/978-1-4614-5311-6_18

[34] P. C. Lambert, J. R. Thompson, C. L. Weston, and P. W. Dickman, "Estimating and modeling the cure fraction in population-based cancer survival analysis," *Biostatistics*, vol. 8, no. 3, pp. 576–594, Jul. 2007. [Online]. Available: http://biostatistics.oxfordjournals.org/cgi/doi/10.1093/biostatistics/kxl030

[35] G. King, J. E. Alt, N. E. Burns, and M. Laver, "A Unified Model of Cabinet Dissolution in Parliamentary Democracies," *American Journal of Political Science*, vol. 34, no. 3, p. 846, Aug. 1990. [Online]. Available: http://www.jstor.org/stable/2111401?origin=crossref

[36] C. Cioffi-Revilla, "The Political Reliability of Italian Governments: An Exponential Survival Model," *The American Political Science Review*, vol. 78, no. 2, p. 318, Jun. 1984. [Online]. Available: http://www.jstor.org/stable/1963367?origin=crossref

[37] E. Voeten, A. Strezhnev, and M. Bailey, "United Nations General Assembly Voting Data," 2015. [Online]. Available: http://hdl.handle.net/1902.1/12379

[38] C. S. Signorino and J. M. Ritter, "Tau-b or Not Tau-b: Measuring the Similarity of Foreign Policy Positions," *International Studies Quarterly*, vol. 43, no. 1, pp. 115–144, 1999.

[39] D. M. Gibler, *International military alliances, 1648-2008*. CQ Press, 2009.

[40] R. Bayer, "Diplomatic exchange data set, v2006. 1," *Online: http://correlatesofwar. org*, 2006.

[41] K. Barbieri, O. M. Keshk, and B. M. Pollins, "Trading data evaluating our assumptions and coding rules," *Conflict Management and Peace Science*, vol. 26, no. 5, pp. 471–491, 2009.

[42] K. Barbieri and O. Keshk, *Correlates of War Project Trade Data Set Codebook, Version 3.0*, 2012, published: Online: http://correlatesofwar.org.

[43] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. " O'Reilly Media, Inc.", 2009.

[44] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 2003, pp. 173–180. [Online]. Available: http://dl.acm.org/citation.cfm?id=1073478

[45] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," 2002, http://mallet.cs.umass.edu.

[46] J. P. Klein, *Survival analysis: techniques for censored and truncated data*, 2nd ed., ser. Statistics for biology and health. New York: Springer, 2003.

[47] R. J. Cook, *The statistical analysis of recurrent events*, ser. Statistics for biology and health. New York: Springer, 2007.

[48] *World Development Indicators 2014*. The World Bank, 2014. [Online]. Available: http://elibrary.worldbank.org/doi/abs/10.1596/978-1-4648-0163-1

[49] J. D. Singer, "Reconstructing the correlates of war dataset on material capabilities of states, 1816–1985," *International Interactions*, vol. 14, no. 2, pp. 115–132, 1988.

[50] M. G. Marshall, K. Jaggers, and T. Gurr, "Data user's manual. polity iv project." [Online]. Available: http://www.systemicpeace.org/inscr/p4manualv2013.pdf

[51] "Historical Exchange Rates." [Online]. Available: http://www.oanda.com/currency/historical-rates/

[52] T. M. Therneau, *A Package for Survival Analysis in S*, 2015, version 2.38. [Online]. Available: http://CRAN.R-project.org/package=survival

[53] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional Variable Importance for Random Forests," *BMC Bioinformatics*, vol. 9, no. 307, 2008. [Online]. Available: http://www.biomedcentral.com/1471-2105/9/307

[54] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution," *BMC Bioinformatics*, vol. 8, no. 25, 2007. [Online]. Available: http://www.biomedcentral.com/1471-2105/8/25

[55] T. Hothorn, P. Buehlmann, S. Dudoit, A. Molinaro, and M. V. D. Laan, "Survival Ensembles," *Biostatistics*, vol. 7, no. 3, pp. 355–373, 2006.

[56] A. Hájek, "The reference class problem is your problem too," *Synthese*, vol. 156, no. 3, pp. 563–585, May 2007. [Online]. Available: http://link.springer.com/10.1007/s11229-006-9138-5

[57] Nick Bostrom, *Anthropic bias: observation selection effects in science and philosophy.* Routledge, 2010.

[58] M. Tegmark, *Our mathematical universe: my quest for the ultimate nature of reality.* Penguin UK, 2014.

[59] P. Domingos and M. Richardson, "Markov Logic: A Unifying Framework for Statistical Relational Learning," in *Statistical Relational Learning*, 2007, p. 339.

[60] S. Kok, M. Sumner, M. Richardson, P. Singla, H. Poon, and P. Domingos, *The Alchemy system for statistical relational AI (Technical Report). Department of Computer Science and Engineering, University of Washington, Seattle, WA*, 2006.