THE DIFFUSION OF INSTABILITY IN AUTHORITARIAN REGIMES

A Dissertation in
Political Science
by
Thomas Brawner

© 2017 Thomas Brawner

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2017
The dissertation of Thomas Brawner was reviewed and approved* by the following:

Joseph G. Wright  
Associate Professor of Political Science  
Dissertation Adviser  
Committee Co-Chair

Donna Bahry  
Professor of Political Science  
Committee Co-Chair

Bumba Mukherjee  
Professor of Political Science

Johannes W. Fedderke  
Professor of International Affairs

James Piazza  
Liberal Arts Research Professor of Political Science  
Director of Graduate Studies

*Signatures are on file in the Graduate School.
Abstract

Spatial models of political instability typically constrain the trigger events of interest to lagged values of the dependent variable, and such a lag is often intended to capture positive diffusion of the event. This dissertation explores a wider range of spatiotemporal lags in descriptive and predictive models of authoritarian regime failure, democratic transition, and coups d’état. Expanding the set of spatiotemporal lags allows for more precise directional hypotheses of transnational learning. The results indicate that lags of autocratic transitions, for example, may inhibit the realization democratic transition, while lags of democratic transition may facilitate it. The dissertation is also inspired by a methodological concern for model generalizability, and therefore the importance of model validation and regularization is given extensive treatment. These principles are applied to descriptive models, in which the primary objective is to explain, and to predictive models, in which the primary objective is to predict unseen data. Moreover, these principles are applied to linear models commonly employed in political science and to more flexible machine learning models that are less commonly used in the field. In the end, a template for improving the generalizability of model estimates is provided, and straightforward tools for interpreting those estimates are explored.
# Contents

List of Figures vi

List of Tables vii

Chapter 1  Diffusion & Autocratic Instability 1

1.1 Core concepts 3
1.2 Learning and Emulation 7
1.3 Variation in trigger events & outcomes 10
1.4 Pathways of diffusion 13

Chapter 2  Empirical Model of Regime Failure 15

2.1 Data 15
2.1.1 Dependent variables 15
2.1.2 Spatiotemporal lags 17
2.1.3 Domestic attributes 21
2.1.4 Missing data 27
2.2 Statistical model 28
2.2.1 Elastic net regularization 29
2.2.2 Feature standardization 32
2.2.3 Cross-validation 33
2.2.4 Parameter uncertainty 34
2.2.5 Melding quantities of interest 35
2.2.6 Predicted probability plots 36
2.2.7 Results 36
2.3 Remarks 46

Chapter 3  Predictive Models of Regime Failure 50

3.1 Validation strategy 52
3.1.1 Random forest 53
3.1.2 Support vector machine 55
3.1.3 Ensemble Bayesian model averaging 57
3.2 Model performance metrics 58
3.3 Model interpretation 60
3.4 Prediction results 61
3.4.1 Regime failure 61
3.4.2 Democratic transition 69
3.5 Remarks 77
List of Figures

1.1 Autocratic regimes and population under autocratic rule ............... 6
2.1 Regime failure and democratic transition, yearly counts 1946–2010 .... 16
2.2 Geographic connectivity .................................................. 19
2.3 Economic connectivity ...................................................... 21
2.4 Authoritarian regime type counts ....................................... 22
2.5 Authoritarian regime type duration ...................................... 23
2.6 Authoritarian regime type institutionalization ............................ 24
2.7 Economic attributes ........................................................ 26
2.8 Proportions of missing data ................................................ 27
2.9 Regime failure estimates, bootstrap intervals .............................. 39
2.10 Regime failure estimates, permutation significance ..................... 40
2.11 Predicted probabilities of regime failure ................................ 42
2.12 Democratic transition estimates, bootstrap intervals ................... 44
2.13 Democratic transition estimates, permutation significance ............ 45
2.14 Predicted probabilities of democratic transition ........................ 47
3.1 Distributions of test predictions, regime failure ........................... 63
3.2 Probability of failure by lagged autocratic transitions .................... 65
3.3 Probability of failure by lagged democratic transitions ................. 66
3.4 Predicted and observed regime failures .................................. 68
3.5 Distributions of test predictions, democratic transition .................. 71
3.6 Probability of democratic transition by lagged autocratic transitions .. 73
3.7 Probability of democratic transition by lagged democratic transitions .. 74
3.8 Probability of democratic transition by regime type ..................... 76
4.1 Coups d’état, yearly counts 1950–2010 .................................. 86
4.2 Coup estimates, bootstrap intervals ....................................... 88
4.3 Coup estimates, permutation significance ................................ 89
4.4 Predicted probabilities of coups ......................................... 91
4.5 Distributions of test predictions, democratic transition ................. 94
4.6 Probability of coup attempts by lags ................................... 96
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Elastic net parameters, regime failure &amp; democratic transition models</td>
<td>49</td>
</tr>
<tr>
<td>3.1</td>
<td>Prediction performance for models authoritarian regime failure across classifiers for the test period</td>
<td>62</td>
</tr>
<tr>
<td>3.2</td>
<td>Prediction performance for models of authoritarian regime failure for the ensemble BMA and component models for the calibration period</td>
<td>69</td>
</tr>
<tr>
<td>3.3</td>
<td>Prediction performance for models of authoritarian regime failure for the ensemble BMA and component models for the test period</td>
<td>70</td>
</tr>
<tr>
<td>3.4</td>
<td>Prediction performance for models democratic transition across classifiers for the test period</td>
<td>70</td>
</tr>
<tr>
<td>3.5</td>
<td>Prediction performance for models of democratic transition for the ensemble BMA and component models for the calibration period</td>
<td>77</td>
</tr>
<tr>
<td>3.6</td>
<td>Prediction performance for models of democratic transition for the ensemble BMA and component models for the test period</td>
<td>77</td>
</tr>
<tr>
<td>4.1</td>
<td>Elastic net parameters, coup model</td>
<td>92</td>
</tr>
<tr>
<td>4.2</td>
<td>Prediction performance for models coup attempts across classifiers for the test period</td>
<td>93</td>
</tr>
<tr>
<td>4.3</td>
<td>Prediction performance for models of coup attempts for the ensemble BMA and component models for the calibration period</td>
<td>94</td>
</tr>
<tr>
<td>4.4</td>
<td>Prediction performance for models of coup attempts for the ensemble BMA and component models for the test period</td>
<td>95</td>
</tr>
</tbody>
</table>
Chapter 1

Diffusion & Autocratic Instability

Scholars have long recognized the importance of international context for the stability of governments, and accordingly there is a rich subset of the comparative and international politics literature that has treated spatiotemporal clustering of political unrest, including of civil war (Gleditsch, 2007; Weidmann and Ward, 2010), coups d’état (Huff and Lutz, 1974; Li and Thompson, 1975; Lutz, 1989; Miller, Joseph, and Ohl, 2016), democratization (Beissinger, 2007; Brinks and Coppendge, 2006; Houle, Kayser, and Xiang, 2012), and protest (Bunce and Wolchik, 2006; Gause, 2011; Hill and Rothchild, 1986). It is apparent, therefore, that mechanisms of diffusion are common to a range of political objects, and distinguishing among domestic and transborder determinants of those objects is a key consideration in building and evaluating empirical models of political instability.

The prevailing literature on the diffusion of political instability has been defined in part by three characteristics that are addressed in the current work. First, from a conceptual perspective, the direction of diffusion is often assumed to be positive, implying that political objects of a given type tend to lead to produce more political objects of that type. In other words, civil wars spill over state boundaries, protests occur in waves (as in the so-called Arab Spring), and democratization events beget democratization attempts (as in the fall of the Soviet Union). This characteristic makes two important assumptions for models of spatial diffusion: 1) lags of the the dependent variable contain the only information relevant for the diffusion of regime instability; 2) the direction of the diffusion is positive, suggesting spatial clustering of the event captured by the dependent variable.

In contrast, the current work does not put such a strict constraint on the events that matter for the diffusion of instability. Rather, it departs under the assumption
that the relevant actors engaged in the spread of political objects across state boundaries observe and learn from a wide array of events, not merely the event captured by the dependent variable, and we further expect that some of these events encourage while others discourage the diffusion of political instability. For example, the observation of democratization in one authoritarian regime may prompt other authoritarian regime failures, some of which may result in democratic regimes while others do not. On the other hand, a regime failure that leads to the emergence of a separate authoritarian regime may prevent authoritarian regimes elsewhere from failing in the first place, perhaps because opposition actors are not willing to risk a challenge to the incumbent regime only to have a new authoritarian regime emerge in its place.

Second, from a methodological perspective, empirical studies of diffusion often use data that are aggregated temporally at the yearly level, a symptom of which is difficulty in the construction of spatiotemporal lags of events. By contrast, the analyses below leverage specific dates for events of political instability, allowing for the construction of temporally precise spatial lags for those events, thus reducing the measurement error for these variables and, in turn, permitting the use of straightforward classification methods.

Last, in place of purely descriptive models, the analyses below are driven first and foremost by the objective of generalizability, and description is performed only after generalizable insights are uncovered through model validation. Indeed, a core value in this work is located in its exploration of those events that matter for the diffusion of political instability (i.e., do the findings generalize?), how those events matter (i.e., what are the suggested effects?), and the extent to which the findings help researchers and policymakers understand what will happen moving forward under counterfactual scenarios. At the end of the day, if model validation procedures suggest that the findings generalize, and the direction and magnitude of the effects are substantively interesting, then the findings should help us make better predictions about where political instability is likely to occur moving forward in time.

The work proceeds in three sections. The first chapter outlines the core concepts
in this work and details the theories underlying transnational learning and diffusion of political instability. The second chapter presents an initial set of models aimed at inferring the associations between authoritarian regime failures and their spatiotemporal lags. The third chapter extends the initial analysis by including more flexible machine learning algorithms, evaluating their predictive performance, and then exploring the findings of the best performing models to understand the spatiotemporal relationships uncovered by the models. The final chapter repeats this analysis on separate target event, coups d’état.

1.1 Core concepts

The fundamental concepts in this work are inspired by Huntington (1991), who outlines a process by which the observation of political event $X$ in country $i$ at time $t = 1$ is a function to some degree of the observation of $X$ in another country $j$ at time $t = 0$. In this process, $i$ is the target state, and the initial event in $j$, the sending state, is known as a trigger event. In simple terms, political events in one country may subsequently alter the political trajectory of other countries.

The implication is that two temporally ordered events across political units may not be independent of one another, and this in turn prompts an interesting counterfactual scenarios. In particular, would we have observed $X_i$ in the absence of having first observed $X_j$? The assumption in this formulation of the dependence between the political trajectories of the two states is that the effect of the trigger event in the sending state is to increase the likelihood of observing a similar event in the target state. An alternative question asks if we might have observed $X$ in a separate target country at $t = 1$, where it was in fact not observed, had we not observed $X_j$. In this case, the dependence among the two countries implies that the trigger event in $j$ subsequently reduces the likelihood of observing the event in would-be target countries.

These are the primary questions under examination. Specifically, is there a systematic temporal and spatial dependence among politically destabilizing events that occur in the
international system? If there is a systematic dependence among these events, is the
direction of that association positive? In other words, does the observation of political
instability in one country subsequently increase the probability of observing a similar
event in other countries? Or is the direction of the association negative, thus suggesting
that the observation of political instability in one country tends to reduce the probability
of observing it in other countries?

To address these questions, quantitative models of regime failure, democratic transi-
tion, and coups d’etat in authoritarian regimes are specified and evaluated. The choices
made here with respect to dependent variables and the subset of countries (i.e., authorita-
tarian countries) under evaluation are intentional. Regarding the subset of events under
evaluation, regime transitions and coups are representative of the relevant destabilizing
events from the perspective of authoritarian countries. To see this, we can assume (as a
simplification) that the principal objective of a political leader, authoritarian or otherwise,
is to withstand challenges to its existence and remain in power (e.g., Geddes, 2003). In
the case of authoritarian regimes, we typically think of these threats as originating from
domestic sources (e.g., Geddes, 1999b), but we should also consider the possibility that
domestic actors are influenced by the international environment. Therefore, reducing the
domestic impact of a trigger event, taking the form of an authoritarian regime failure, a
democratic transition, or a coup, is of central importance. For this reason, democratic
transitions and coups have been regular targets in scholarly treatments of the diffusion of
political instability (e.g., Gleditsch and Ward, 2006; Li and Thompson, 1975; Weyland,
2010).

In addition, limiting the domain to authoritarian regimes makes for a more coherent
analysis for several reasons. First, the political dynamics at play between opposition
groups and regime elites are fundamentally distinct from the political contestation in
democratic regimes (Geddes, 2003). In this way, the decision calculus on the part of
opposition groups in an authoritarian regime, which lack political channels to power,
to attempt to emulate a regime transition observed abroad entails different risks and
rewards than that of opposition groups in democracies, which have well-defined institutional mechanisms for the acquisition of power. Moreover, the diffusion of instability across state borders presupposes a receptive audience in the target state(s) based on a prevailing set of grievances. For example, the wave of reforms in African authoritarian regimes in the years after the end of the Cold War reflected long-standing preferences for reform among the masses, though the timing of the reforms was clearly related to the collapse of the World War international system (Decalo, 1992). In this way, regarding the mobilizations in Africa at this time, “external factors serve[d] as precipitating conditions, rather than causal ones” (Bratton and van de Walle, 1992, 420). A similar argument can be made with regard to the spread of protests during the Arab Spring, where common socioeconomic grievances made the authoritarian regimes of the Middle East and North Africa ripe for emulation (Campante and Chor, 2012). In sum, limiting the analysis to authoritarian countries allows for us to hold structural political factors constant and focus on how trigger events alter expectations among opposition groups whose grievances are defined by these precise political factors.

Second, democratic transition is obviously not a potential outcome for existing democracies, and underlying this fact is a subjective preference for the analysis of why authoritarian regimes fail in the first place (e.g., Ulfelder, 2005). That is, because a non-negligible proportion of governments are non-democratic and because the quality of life for subjects of non-democratic governments is assumed to be inferior to that of citizens of democratic governments, understanding the determinants, whether domestic or foreign, of regime failures, and in particular those regime failures that result in a democratic transition, is a worthy endeavor. Figure 1.1 plots the proportion of autocratic countries alongside the proportion of the total world population living under autocracy from 1946 to 2010.

Finally, the work below is primarily quantitative and is motivated in large part by an interest in learning models of political instability that generalize well. To some degree, the emphasis is prediction over inference or description, and the flexible machine learning models featured in the analyses are chosen because of their superior performance
in prediction. However, a purely predictive exercise is at odds with the larger aim of understanding and qualifying the direction and magnitude of the transmission of political instability across state boundaries, as outlined above. As such, care is taken at each step of the analysis to understand the effects suggested by the model. In the end, a key and deliberate feature of the methodological workflow employed here is to first learn the model that best generalizes, then to open the black box and explore the relationships in the data. The hope is both to (1) present a realistic view on the importance of transnational diffusion in the larger picture of determinants of instability for authoritarian regimes and (2) describe the nature of the diffusion learned by the models.

In part because statistical inference is often not prioritized in machine learning, and in part because the nature of political instability as operationalized here implies analysis of a rare event, custom statistical programs that extend the functionality of built-in statistical and machine learning libraries are used throughout the analysis. Indeed, some of the value added in this contribution is likely in the development of a set of programmatic tools that facilitate the joint pursuit of generalizability and inference in statistical models. Because this work presents a series of binary classification problems, these programs have a particular slant toward classification, and they are further tailored to address
two widespread issues for quantitative models in the social sciences, class imbalance and missing data. In the interest of satisfying calls for replicability and transparency (e.g., Eubank, 2016), a subset of the relevant Python programs that comprise the workhorse tools for data wrangling and statistical modeling are presented and described in Appendix 5.

1.2 Learning and Emulation

To begin, we need to outline a theory of dependence in time and space for contentious political events across state borders. As implied above, dependence among such events might take one of two broad forms if it is present at all. In one case, a trigger event in a sending state increases the likelihood of a subsequent event in a target state, and in the other case a trigger event decreases the likelihood of a subsequent event in target states. Franzese and Hays (2008) concisely categorize interdependence in the domestic politics of two states into two categories, strategic complements and strategic substitutes. If the best response function for some country \( i \), given a trigger event in country \( j \), makes emulation more likely, the events in \( i \) and \( j \) are strategic complements. On the other hand, if the trigger event in \( j \) dissuades emulation in \( i \), the events are strategic substitutes.

While the notions of strategic complements and substitutes in Franzese and Hays (2008) are conceived with respect to the diffusion of domestic policies relevant for economic globalization, they are nonetheless helpful in framing the present discussion. The key distinction for contentious politics within states is that it is necessarily a two-player interaction at the domestic level. In other words, the regime and an opposition group must be contending for power in order to manifest contentious politics. For Franzese and Hays (2008), states are unitary actors at the domestic level, though engaged in strategic interactions at the international level. That said, states in both cases are subject to the international context. Namely, one can imagine that the best response function in a target state, given a trigger event in a sending state, is simply the aggregate response of all relevant groups in the target state. In other words, if the sum of the respective best
responses for the opposition group, which perhaps is motivated to undertake emulation of a trigger event, and the regime, which is motivated to inhibit such an emulation, is positive, we might label the dependence among the sending and target states a strategic complement. On the other hand, if the aggregate responses of the opposition and regime has the effect of a lowering the probability of an event such as the trigger event beyond what would have been in the absence of the trigger event, we could usefully classify this dependence as strategic substitution.

At this point, we should further flesh out the expectations for the diffusion of contentious politics across state borders. With regard to challenges to a political regime, a manifestation of contentious politics by *early risers* in sending state $j$ may generate perceptions of political opportunity, which otherwise would not have been realized, among the political opposition in target state $i$, and which may in turn lead to a contentious interaction in $i$ (Tarrow, 1998). A necessary condition implied in this formulation is that political actors in structurally similar situations in states $i$ and $j$. In this way, there are “actor-specific” mechanisms of diffusion, and these are contingent on perceived commonalities in group identities across state borders (Gleditsch, 2007; Hill, Rothchild, and Cameron, 1998).

For example, opposition groups in authoritarian regimes, which likely share many structural challenges in their efforts to achieve political change, may engage in transborder learning. Specifically, the introduction of new information as a consequence of $X_j$ interacts with the prevailing preference for regime change in $i$, thus increasing the probability of *emulation*, represented by $X_i$ (e.g., Hill, Rothchild, and Cameron, 1998; Kuran, 1991; Weyland, 2010). In other words, opposition actors observe contentious politics abroad, note what is successful, and apply the observed model to their own situation, updating their strategies to account for the new information introduced by $X_j$. Beissinger (2007) and Tarrow (1998) refer to this as *modular action*. In this light, and under the assumption that the opposition in $i$ undertakes an emulation in response to $X_j$, the events $X_j$ and $X_i$ are understood to be strategic complements. This yields the expectation that challenges
to incumbent authoritarian regimes should cluster in time and space in response to a trigger event.

However, authoritarian regimes and pro-regime actors likewise observe and learn from the international environment, suggesting that the effect of a trigger event may be to reduce the likelihood of spatiotemporal clustering of regime challenges. That is, even where the target state is ripe for receipt of diffusion, due to structural similarities with the sending state and opposition learning, regimes can still prevent emulation. This point is frequently made with reference to democratic transitions and the diffusion of anti-regime protests across state borders. Beissinger (2007) and Weyland (2010), for example, note that elites in favor of the incumbent regime learn from a trigger event and impose constraints on the desire and capabilities of the opposition to emulate. Similar arguments are put forth by Houle, Kayser, and Xiang (2012), among others, who emphasize that the responses of elites to changes in the perceived threat levels to the regime is central to understanding variation in outcomes following attempts at emulation of a successful model. In the end, an effective response by the regime to guard itself from a perceived increase in domestic threats may indeed preemptively prevent the diffusion of instability.

Several cases demonstrate the tactics employed by authoritarian regimes in response to a trigger event. Koesel and Bunce (2013) detail Russian and Chinese responses to the Colour Revolutions and the Arab uprisings, emphasizing that both regimes feared emulation by their citizens. To reduce the probability of an emulation, these regimes framed the democratization events in a negative fashion, implying that conditions had gotten worse in the aftermath of the regime failure, and placed restrictions on the free flow of information to their citizens. In addition, the regimes increased their constraints on opposition political parties and civil society organizations. In this way, Koesel and Bunce argue that the Russian and Chinese regimes impeded diffusion by both attacking the willingness and ability of domestic opposition to emulate.

Vines and Weimer (2011) discuss a similar response by the Angolan regime to the Arab uprisings and to protests in Mozambique, outlining a range of tools from state control of
the flow of information about protests to repression and co-optation of opposition groups to impede diffusion. In Gabon, domestic opposition groups were fueled by successful protests in Tunisia, but the regime, concerned about the spread of protests throughout the region, responded by deploying security forces and closing television stations sympathetic to the opposition (Ndong, 2011).

In sum, regimes account for instability abroad when estimating and addressing threats to stability at home. To the extent that these steps lower the probability of opposition challenges to the regime, an event such as $X_j$ may actually supplant the realization of $X_i$, and in this way they are strategic substitutes. The expectation, contrary to opposition learning, is that challenges to incumbent authoritarian regimes tend not to cluster in time and space in response to a trigger event.

### 1.3 Variation in trigger events & outcomes

To restate the expectations thus far, both opposition groups and regimes observe and learn from their international context (e.g., trigger events), and the respective effects in response to that learning move in opposite directions. Therefore, following a trigger event in the form a regime challenge, the probability of emulation in potential target states (i.e., authoritarian regimes having structural similarities) should both increase and decrease as a function of the diverging best response functions for the opposition and pro-regime actors. Yet, we should note that there is important variation among trigger events, and this should correspond to important variation in how opposition groups respond to those events.

To begin, the outcome under study in this section is authoritarian regime failure, which is a potential outcome given contentious politics in a non-democratic regime. From this broad category, we can further split off a subset of regime failures, democratic transitions, that are likewise of particular interest. So the two specific questions under consideration are: (1) *What are the conditions, domestic and international, that are associated*
with authoritarian regime failure? and (2) What are the conditions, domestic and international, that are associated with a transition from an authoritarian to a democratic regime? For the analyses below, four international determinants, in the form of trigger events, are considered: regime failures of any type, regime failures that lead to a democratic transition, regime failures that lead to an autocratic transition, and coerced regime failures.

On the part of regimes, we can make the assumption that transnational learning should take basically the same form in response to the observation of a trigger event of any type. Again, this is traced to the fundamental objective of authoritarian regimes to stay in power (Geddes, 2003). Even military regimes, which are notable in their softer preferences for holding onto political power (Geddes, 2003; Ulfelder, 2005), are likely to respond to perceived threats to the regime with a repressive crackdown for the maintenance of order (Ulfelder, 2005) and more generally desire to control the process by which it passes power off to a subsequent group (e.g., Geddes, 1999a; Huntington, 1991). As a result, the best response of an authoritarian regime should always be to take action to prevent domestic instability, and doing so should include efforts to prevent emulation of a trigger event.

On the part of opposition groups who seek a transition to democracy, the update in preferences for challenging the incumbent regime should vary with the nature of the trigger event. For example, in response to a democratic transition, the expectation is that opposition groups abroad should learn from the trigger and try to emulate it. That is, a successful trigger should bolster anti-regime sentiment and increase the likelihood that a regime challenge will take place, irrespective of whether the challenge is unsuccessful or whether it is successful and results in either a democratic or an autocratic transition. This is the model of diffusion commonly encountered in the literature, emphasizing positive diffusion of a particular form of regime transition (e.g., Kuran, 1991; Weyland, 2010), but in this case we acknowledge that emulation may result in an outcome other than democratic transition. We formally hypothesize, therefore, that democratic transitions in
one state should be associated with subsequent regime failures in other states including, but not limited to, democratic transitions.

However, in response to a regime failure that results in a transition to a separate authoritarian regime (an autocratic transition), the lesson learned by opposition groups may be entirely different. Namely, despite a prevailing preference for unseating the incumbent regime, observing a transition abroad that results only in a new state of authoritarianism might dissuade the opposition from pursuing a regime change. That is, the costs of contentious political activity may not be worth the benefits of achieving a regime change if there is a perception that the regime change will not result in political improvements (Meirowitz and Tucker, 2013). As such, contrary to democratic transitions, autocratic transitions should reduce the likelihood of subsequently observing a regime failure in other states.

The broad category of regime failures, which nests both democratic and autocratic transitions, is therefore less theoretically specific. Rather, the objective in specifying a spatial lag for all authoritarian failures is to test the argument that failures of any type cluster in time and space. That is, if the failure of the incumbent regime is the predominant preference of opposition groups, irrespective of the outcome, then we should expect that a failure in one state should increase the probability of subsequent failures in other states.

Similarly, coerced regime failures, which fall into the subset of regime failures caused by popular uprising, military coup, or civil war, are not specific about the outcome of the regime failure, but are highlighted here because of the costs to the opposition in challenging the incumbent regime. In particular, any transnational lessons learned by opposition groups necessarily include knowledge of the costs to be paid in order to bring about political change. By itself, it is not clear that coerced failures should be associated with lower or higher probabilities of emulation. Rather, it may crucially depend on

---

1Svolik (2013) outlines a similar logic for explaining political instability in new democracies. Namely, if voters are dissatisfied with elected officials, their support for democratic institutions suffers, and an authoritarian reversal is more likely.
whether such failures result in democratic or autocratic transitions. No explicit hypothesis is therefore provided at this point.

1.4 Pathways of diffusion

This work features three notions of connectivity, or channels through which political objects are transmitted from a sending to a target state, including geographic, economic, and cultural ties. To begin, geography has been the primary metric of distance employed in the diffusion literature on instability (e.g., Buhaug and Gleditsch, 2008; Gleditsch, 2007; Weidmann and Ward, 2010). In part, the influence of geography in these works is related to the direct and material flow of people and goods across geographic space, and the connectivity among states is enhanced where the potential for such flows is greater. On the other hand, geographic proximity has also been used as a proxy for many other types of connectivity, including cultural similarities and economic activity (Huntington, 1991; Simmons and Elkins, 2004). Franzese and Hays (2008, 764) note, for example, that economic activity tends to concentrate at borders to reduce transportation costs and that multinationals choose their location to maximize benefits through exporting to geographically proximate markets. In this work, because the political objects diffused across borders are assumed to be “air-borne” (Iqbal and Starr, 2008), in that learning and emulation is not bound by physical distance, geography is used to implicitly capture country attributes that tend to cluster geographically but are not explicitly specified as a measure of connectivity.

While metrics of connectivity other than geography are less commonly encountered in spatial models of political phenomena (Beck, Gleditsch, and Beardsley, 2006), this work explores the extent to which economic and linguistic ties specifically act as a mechanism for diffusion. Following Beck, Gleditsch, and Beardsley (2006) and Simmons and Elkins (2004), the level of trade between two states is posited to be an indication of the strength of interaction among actors in those countries and thus the potential for the flow of infor-
information among those actors. Likewise, common spoken language is assumed to facilitate the flow of information across state borders by reducing transaction costs in interactions (Hutchinson, 2002, 2005; Isphording and Otten, 2013).

By using bilateral trade flows and linguistic commonalities among states, this work builds on prevailing work that extends the notion of *distance* among countries for conceptualizing the potential for interaction and, ultimately, transnational learning. That said, there remain several indicators of connectivity, whether economic or cultural, that are not captured in these indicators, including religious similarities, common colonial heritage, and so on. As indicated above, geographic distance is intended as a proxy for these omitted metrics of connectivity.
Chapter 2

Empirical Model of Regime Failure

This chapter details the specification of the statistical models of authoritarian regime and democratic transition. In Section 2.1, the data sources and transformations performed on those data to yield the features and target variables in the analysis are described. Subsequently, in Section 2.2, the statistical modeling techniques to yield parameter estimates are described in detail, including the specification of regularization parameters for logistic regression and methods for combining and interpreting the model results across multiple imputations of the data.

2.1 Data

This section outlines the data sources, transformations, and feature engineering conducted in support of the analysis of authoritarian regime failure and transitions from an authoritarian to a democratic regime. The country-year observations included in the study are determined by the set of authoritarian countries identified in Geddes, Wright, and Frantz (2014), hereinafter denoted GWF, which codes authoritarian regimes from 1946 to 2010. The data identify 4591 authoritarian country-year observations for 118 countries.

2.1.1 Dependent variables

There are two binary dependent variables in the analysis, and both are obtained or derived from GWF. The first of these is the simple indicator of regime failure in GWF, which aggregates events irrespective of the nature of the failure or the subsequent regime type, and these events account for approximately 4.9% of authoritarian regime-year observations. Models of this outcome test the generalized fragility of regimes to episodes of
political instability in other regimes. The rationale for looking at the full set of regime failures as an outcome follows from the argument that we should seek to understand the circumstances under which the repressive institutions of dictatorships fail (Ulfelder, 2005). The second outcome variable, *democratic transition*, indicates autocratic regime failures that are subsequently replaced by a democratic regime, accounting for 2.2% of the total observations in the data. This outcome is used to specifically test models of democratic diffusion and to explore how the probability of democratic transition is associated with different classes of political instability abroad. Counts of each of these event types over time are shown in Figure 2.1.


2.1.2 Spatiotemporal lags

There are two key objectives in generating the spatiotemporal lags of politically destabilizing events, the features used to test hypotheses related to the diffusion of instability in the models below.

1. Avoid simultaneity bias in estimation by including in values of the lag only those spatially lagged events that temporally precede the realization of the dependent variable.

2. Leverage multiple metrics of connectivity between countries to determine if they differentially associated with the likelihood of a regime failure event.

To address the first point, because failure events in GWF are given on the day on which they occurred, a precise date can be applied to the observation of the dependent variable. In turn, when constructing a spatial lag of the relevant events for a country-year unit of observation, events occurring within that year in countries external to country $i$ (e.g., in country $j$) but following the date on which the event in $i$ is observed are appropriately omitted. The inclusion of such events in the construction of spatial lags, a consequence of temporal aggregation, is the source of the issue of simultaneity bias in the spatial statistics literature (e.g., Franzese and Hays, 2007). Instead of relying on statistical estimators that account for the improper temporal ordering, or contemporaneous observation, of events in a country-year format, this study proposes a more appropriate measure of spatial lags that includes both spatial and temporal dimensions for lag construction.

Specifically, regarding the temporal criterion for inclusion of an event in the spatiotemporal lag, this approach takes as a starting point the date attached to the dependent variable, then sets up a one year window in advance of that date, to the day.\footnote{More precisely, a window of 365 days is used, so leap years are incorrectly, if only slightly, measured by this definition of a year.} Any event occurring external to $i$ within this window is included in the spatiotemporal lag of
failure events for that country-year observation for \( i \). For example, if country \( i \) experiences a regime failure event on June 11 of a given year, the lag incorporates information on regime failures from June 11 of the previous year until June 10 of the current calendar year. For years in which there is not an event in \( i \), the date attached to the dependent variable is the final day of the current year and the window maps onto the calendar year, capturing all events external to \( i \) in the prevailing calendar year, the strategy followed by Miller, Joseph, and Ohl (2016). The clear benefits of disaggregating the dependent variables temporally are twofold. First, more precise measurement in the spatiotemporal lags yields the ability to explicitly test the association of only those events that precede the observation of the dependent variable in time. Second, this process allows the straightforward application of a range of standard classification methods for binary dependent variables.

To address the issue of multiple metrics capturing the distance between countries, three different metrics of connectivity are used, including geographic, economic, and linguistic distances. Each of these is described in detail below.

**Geographic distance**

Geographic distance is operationalized here as the distance in kilometers between capital cities (Gleditsch and Ward, 2001).\(^2\) I choose a continuous measure of geographic connectivity, as opposed to a binary contiguity weights matrix, reflecting the idea that events beyond an arbitrary minimum distance are relevant for spatial interdependence. Notably, scholars have acknowledged that material interaction or relative geographic proximity is not necessary for interdependence, especially where the mechanism of diffusion is one of learning (Beissinger, 2007; Gleditsch, 2007; Iqbal and Starr, 2008; Ross and Homer, 1976). In turn, by limiting the information in the lag to a constrained geographic neighborhood, there is a risk that substantively important events (i.e., external events expected to be

\(^2\)The capital cities distance data are made available on Kristian Skrede Gleditsch’s website at [http://privatewww.essex.ac.uk/~ksg/data/capdist.csv](http://privatewww.essex.ac.uk/~ksg/data/capdist.csv).
associated with the probability of a regime failure) are omitted from the analysis.

Because geographic connectivity decreases in distance (i.e., countries farther apart are less connected), the weights are constructed by first reversing the distances assigned to dyads of countries, then constraining the values between 0 and 1 by dividing each distance by the maximum observed distance. This weighting scheme yields nicer distributions than inverse distance or inverse log distance schemes, which result in high right skew in the lags because a small number of geographically proximate countries are given very high influence and the overwhelming majority of countries are effectively ignored. Moreover, this normalization scheme yields a non-zero weight for all country dyads, which is appropriate in light of the argument that actors within a country see beyond their immediate neighborhood for information on potential threats and opportunities. In the end, values closer to 0 imply less connectivity (greater geographic distance) and values closer to 1 imply greater connectivity (less geographic distance). The distribution of the raw distances and the transformed geographic weights are shown, respectively, in the left and right panels of Figure 2.2.
Economic distance

Economic distance is captured by dyadic trade flows using Correlates of War Bilateral Trade data v3.0 (Barbieri, Keshk, and Pollins, 2009; Barbieri and Keshk, 2012). Before converting the trade flows to weights, two steps are taken to prepare these data. First, because the data are in current US Dollars, GDP deflator data are generated and employed to convert the trade figures to constant 2009 US Dollars. Second, to deal with the degree of missingness in the data, missing values are filled via a multiple imputation model using the available monadic and dyadic trade figures, yielding estimates for a large number of the missing values. The specific value employed for the distance metric is the aggregate flow of imports and exports between the two countries in the dyad. Once this aggregate measure is obtained, the natural log of the measure is used to reduce the extreme right skew, then the result is constrained to the 0 to 1 scale in the same manner as the geographic distance metric. In Figure 2.3, the distributions of the raw trade bilateral trade values and these values after transformation are provided.

Linguistic distance

Linguistic distance is operationalized as the intersection of a major language between two countries, implying a binary connectivity scheme. In other words, countries sharing a major language take a value of 1 and countries not sharing a major language take a value of 0. To construct this measure, the Ethnologue online database of national languages is programmatically leveraged to obtain dyadic relationships. The site maintains a comprehensive list of languages spoken in every country in the world, but critically it

---

3The dyadic and monadic trade data are made available on the Correlates of War Project website at http://correlatesofwar.org/data-sets/bilateral-trade.
4Missing values prior to imputation make up 25.75% of the dyadic trade data. After imputation, the percent missing is reduced to 0.38%. The multiple imputation model is run in the Amelia II library in R (Honaker and King, 2010; Honaker, King, and Blackwell, 2012; R Core Team, 2016), which permits specification of the cross-sectional and temporal structure of the data. The practice of using a single imputed value is admittedly crude, but certainly an improvement on list-wise deletion or simple mean imputation.
5The website is http://www.ethnologue.com/browse/countries.
Figure 2.3: Economic connectivity: (a) natural log of aggregate trade flows between dyads; (b) normalized trade weights. The values in (b) are used to weight events in the construction of spatiotemporal lags.

also specifies if a language is either a *statutory national language* or a *de facto language of national identity*, together considered to be the principal languages spoken in the country. The custom program first generates a list of all principal languages for each country in the world, then evaluates if there is intersection in the sets of principal languages for each country in the dyad. If there is an intersection, countries are considered to be connected linguistically, otherwise they are considered not connected. In the resulting connectivity data, there are 37,828 country dyads, of which 17.83% are connected by principal languages.

### 2.1.3 Domestic attributes

In addition to specifying spatiotemporal lags of key events occurring beyond state borders, the analyses control for several attributes accounting for the domestic context of regimes. Key among these is the type of authoritarian regime as coded in GWF. The authors classify regimes into one of four categories, including single-party regimes, personalist regimes, military regimes, and monarchies. As noted in (Geddes, 2003), the institutional
structures of each of these regime types informs the way in which they withstand or succumb to challenges. Counts of authoritarian regime types over time in GWF are shown in Figure 2.4.

To model temporal dependence for regime failure, the natural log of regime duration in years is specified in the models. Controlling for duration dependence with a third order polynomial of duration (Carter and Signorino, 2010) or with a spline of duration (Beck, Katz, and Tucker, 1998; Keele, 2008) does not substantively alter any of the findings, and the spline estimate on the natural log of duration in a generalized additive model provides no evidence of non-linearities in the relationship between duration and the regime failure events. Figure 2.5 shows the distribution of regime duration in years by regime type classification in GWF. We see in these data that military and personalist regimes tend to be more short-lived than monarchies and single-party regimes.

Beyond regime type, the models specify the degree of institutionalization for regimes. Institutions are posited to increase the life of authoritarian regimes by providing a means of co-optation, supplying a controlled environment for reducing transaction costs between leaders and the opposition, and increasing the credibility of commitments made by leaders, thus increasing prospects for growth through investment (Gandhi, 2008; Wright, 2008).
On the other hand, institutions are thought to increase the likelihood of a democratic transition since the interests of incumbent regimes are better protected in the transition to a democratic regime (Wright and Escribà-Folch, 2012). In this light, there is an expectation that institutionalized autocratic regimes are more likely to survive, but there is also the expectation that these regimes are more likely to transition to democracy. The institutions data in the analyses here capture legislatures as coded in Cheibub, Gandhi, and Vreeland (2010). In Figure 2.6, the distribution of this variable is displayed across the regime types in GWF. A value of 0 reflects the absence of a legislature, 1 indicates an appointed legislature, and 2 indicates an elected legislature.

The models also control for population size. As noted in Wright (2008), the tradeoff between public and private goods provision is due in part to the size of the population, and decisions about the allocation of goods should be associated with the durability of authoritarian regimes. The models include the natural log of population, and population data are obtained from World Bank (2014).

Several economic indicators are included in the models as well. The distributions
of these variables and their bivariate associations are shown in Figure 2.7. To start, GDP per capita is included to capture the effect of level of development on authoritarian breakdown (e.g., Boix and Stokes, 2003). While the association between development and regime failure has yielded mixed findings in the empirical literature (e.g., Przeworski et al., 2000), increases in the level of development are generally associated with the emergence of pro-democratic forces in non-democratic countries (Campante and Chor, 2012; Miller, 2012). The measure of GDP per capita is taken from Heston, Summers, and Aten (2011), and the natural log of the variable is used to address right skew.

Economic growth is likewise included to capture general economic performance and changes in the level of development. Strong economic growth is commonly associated with political stability, and likewise poor performance is associated with higher probabilities of regime failure (Haggard and Kaufman, 1995). The measure used is from World Bank (2014), and the measure in the specification is the lagged two-year moving average of growth, and the resulting figures are divided by ten to reduce the spread caused by extreme negative and positive growth spurts.
The degree of economic inequality is additionally a common factor present in models of regime breakdown. The basic expectation is that high inequality increases the demand for redistribution of wealth, which we can think of as a regime transition, and therefore increases the costs to elites under democracy, which implies a reduction in the likelihood a regime transition outside of some change in the balance of power (Acemoglu and Robinson, 2006; Boix, 2003). The inequality measure is taken from the University of Texas Inequality Project (2014).

The degree of economic integration is also considered in these models. Acemoglu and Robinson (2006), Boix (2003), and Freeman and Quinn (2012) have argued, for example, that the mobile capital decreases the ability of democratic governments to tax, thereby reducing the costs to elites in democratization. Moreover, trade is also theorized to reduce the degree of inequality by raising incomes for the poor, again reducing demands for redistribution and, in turn, the costs of democratization for elites (Acemoglu and Robinson, 2006). To capture these concepts, measures of capital account openness (Chinn and Ito, 2008) and economic openness (World Bank, 2014) are included in the specification.

The last domestic economic indicator included in the model is natural resource revenue. High levels of resource wealth have been argued to inhibit authoritarian regime failure via several mechanisms, including repression, co-optation, and the rentier state (e.g., Madhavy, 1970; Ross, 2001; Ulfelder, 2007; Wright, Frantz, and Geddes, 2012). The measure used in the models is the natural log of total resource revenue per capita, derived from the Haber and Menaldo (2011) data.

Finally, the models below specify region and two-year period indicators to control for unmodeled cross-sectional and temporal unit heterogeneity. Controls for temporal heterogeneity are particularly useful in these models to help distinguish between systemic shocks that are common to all countries and the influence of non-global spatiotemporal lags (Franzese and Hays, 2008; Plümper and Neumayer, 2010; Miller, Joseph, and Ohl, 2016). The choice to aggregate at the regional and two-year period levels reflects a tradeoff between the desire to capture unmodeled similarities among countries and time.
Figure 2.7: Distributions and bivariate associations for economic control variables.
periods, while also dealing with separation issues, which is a consequence of modeling rare events such as authoritarian regime transitions (e.g., Ulfelder, 2005). It should be noted, however, that the substantive findings change little by substituting country- and year-specific fixed effects, and the statistical associations are stronger in those models.

### 2.1.4 Missing data

Merging data from the sources above results in non-negligible missingness, a common problem in political science research. A histogram of the missingness across the variables in the data, prior to constructing any binary indicators for levels of categorical variables, is shown in Figure 2.8. In place of listwise deletion, which is known to introduce bias into statistical models (Honaker and King, 2010), each of the models below is constructed on ten imputed versions of the original data. Multiple imputation is performed with the Amelia II library for R (Honaker, King, and Blackwell, 2012; R Core Team, 2016).

Because the methods employed throughout the study do not produce analytical standard errors, prevailing tools for melding parameter estimates and variances thereof are
not readily available. The strategy for addressing uncertainty across the multiple imputations is to numerically simulate the sampling distributions for parameter estimates or predicted values for models built on individual imputations, then aggregate across the imputations to produce an aggregate measure of the variance for quantities of interest.

2.2 Statistical model

As an initial step toward understanding the associations between the probability of a regime transition and the domestic and international context of regimes, a series of logistic regression models are specified. To start, we can specify a basic regression model with the following form:

$$
Pr (y_i = 1) = \logit^{-1} \left[ \beta_0 + \sum_p \beta_p x_{pi} + e_i \right]
$$

(2.1)

Here, $y_i$ is the binary dependent variable for country $i$ in year $t$, $\beta_0$ is the intercept term, and the $x_p$ and corresponding coefficients $\beta_p$ capture domestic attributes for country $i$. The $e_i$ is the error term. At this point, we have a model of regime failure that accounts only for domestic context, and thus we can extend this form to account for the international context.

$$
Pr (y_i = 1) = \logit^{-1} \left[ \beta_0 + \sum_p \beta_p x_{pi} + \sum_{w,m} \rho_{wm} \phi_{wmi} + e_i \right]
$$

(2.2)

In this specification, the coefficients $\rho_{wm}$ capture the influence of the spatiotemporal lag terms $\phi_{wmi}$ using distance weighing scheme $w$ and event $m$. The lags are defined as follows:
\[ \phi_{wmi} = \sum_{j \neq i} d_{wij}z_{mj}. \]

Broadly, the spatiotemporal lag is a weighted sum of events taking place in all countries \( j \neq i \). The weights in the lag, the \( d_w \), reflect the three different weighting schemes described in Section 2.1.2, so the \( w \) index the possible distance weighting schemes including geographic, economic, and linguistic distances. Specifically, \( d_{wij} \) is the distance between countries \( i \) and \( j \) using the distance metric indexed by \( w \). In turn, the \( z_m \) refer to the temporally lagged events in all countries \( j \). For the models of authoritarian regime failure and democratic transition, the events indexed by \( m \) include regime failures, democratic transitions, autocratic transitions, and coerced transitions. Again, each of these events falls within a specified window of 365 days prior to the observation of the dependent variable \( y_i \).

### 2.2.1 Elastic net regularization

Once the weighted lags are constructed, (2.2) is a simple logistic regression model, and for the purpose of simplicity, the notation in (2.1) is used moving forward.\(^6\) Scholars have increasingly expressed reservation regarding the suitability of linear models to uncover generalizable relationships in social science data. In part, this is attributed to a practice of statistical inference by which \( p \)-values are used as a benchmark for understanding a given independent variable to be systematically related to the dependent variable. Yet, statistically significant coefficient estimates from linear models are often a poor indicator of improved predictive performance (Schrodt, 2014; Ward, Greenhill, and Bakke, 2010). This is traced to a key concern in models fit with unconstrained maximum likelihood methods, which is the potential of such models to overfit the training data (e.g., Hastie, \[^6\text{To be explicit, the spatiotemporal lags } \phi_{wmi} \text{ are treated in the same fashion as the } x_i, \text{ such that the estimates } \rho_{wm} \text{ can accordingly be expressed as } \beta.\]
Tibshirani, and Friedman, 2009). In particular, maximum likelihood estimates, while perhaps statistically significant, are potentially idiosyncratic to the sample in which they are learned and therefore of limited use in uncovering generalizable relationships. In the extreme, if we were to adopt a more flexible classifier, either in terms of functional form or an increasingly complex specification, it would be relatively straightforward to learn a decision boundary to correctly identify a high proportion of the samples in the training data, yet this model would almost certainly perform worse than a less complex model in classifying new samples. In response, we need a mechanism for constraining the complexity of the model fit to the training data to reduce the variance of predictions made on new data.

With these concerns in mind, the estimates presented in this section are subjected to elastic net penalization (Zou and Hastie, 2005). In general, regularization methods for linear models introduce bias by putting imposing an external constraint on maximum likelihood estimates. To see this, we can start by expressing the objective function for our logistic regression model using the specification in (2.1) as follows:

$$\argmin_{\beta_0, \beta_1, \ldots, \beta_p} \frac{1}{N} \sum_{i} l \left[ y_i, g \left( \beta_0 + \sum_{p} \beta_p x_{pi} \right) \right].$$

In (2.3), the loss function $l(y_i, g(x_i))$ is the negative log-likelihood for observation $i$, $g(x)$ is the cumulative distribution function of the logistic distribution, or $\frac{1}{1+e^{-x}}$, and the vector of coefficients minimizing the average loss over the $N$ observations is the maximum likelihood solution. To constrain this fit, we could introduce regularization in any number of common forms, such as $L_2$ (ridge) or $L_1$ (lasso) penalties. The elastic net penalty combines the benefits of both ridge and lasso penalization, and indeed the ridge and lasso solutions, as well as the unconstrained maximum likelihood solution, are nested in a hyperparameter search for the elastic net. The revised objective function with the elastic net penalty is below.
\[
\arg\min_{\beta_0, \beta_1, \ldots, \beta_p} \frac{1}{N} \sum_{i} l \left[ y_i, g \left( \beta_0 + \sum_{p} \beta_p x_{pi} \right) \right] + \lambda \left[ \frac{1}{2} (1 - \alpha) \sum_{p} \beta_p^2 + \alpha \sum_{p} |\beta_p| \right] \tag{2.4}
\]

There are two hyperparameters for the elastic net. The overall level of regularization is controlled by \( \lambda \), and as \( \lambda \) approaches 0, the solution to (2.4) approaches the solution to (2.3). The other hyperparameter is \( \alpha \), which controls the tradeoff between \( L_2 \) and \( L_1 \) penalization. As \( \alpha \) approaches 1, (2.4) reduces to the lasso solution (the sum of the absolute values of the estimates), and as \( \alpha \) approaches 0, the ridge solution (the sum of the squared values of the estimates). In general, the balance between these two penalties in the elastic net has been shown to have nice properties in terms of a sparse solution, which is associated with the \( L_1 \) penalty, and reducing the magnitude of parameter estimates in the presence of high multicollinearity, which is associated with the \( L_2 \) penalty (Zou and Hastie, 2005; Kuhn and Johnson, 2013). Estimation of (2.4) is done using the generic stochastic gradient descent algorithm for classification in scikit-learn in Python (Pedregosa et al., 2011), which permits specification of the correct loss function for logistic regression with elastic net penalization.

The models in this section are relatively complex in the sense that we are fitting a large number of parameters, including country and year fixed effects as well as a long list of standard “control” variables common in quantitative comparative politics, and the elastic net is therefore a natural and elegant solution. Specifically, \( L_1 \) penalization offers sparsity or variable selection in high dimension problems. In addition, because the spatiotemporal lags and many of the domestic indicators are highly correlated, thus producing issues with multicollinearity, the \( L_2 \) penalty is crucial for fitting a model of regime failure as a linear function of these variables.

Alternative solutions for dealing with multicollinearity and/or large numbers of parameters include selectively removing the offending features or preprocessing in the form of unsupervised dimension reduction (e.g., principal components analysis). The former
of these methods seems unacceptable because of the potential for omitted variable bias. The latter decreases the interpretability of the estimates from the model since the resulting components are linear combinations of the original features, and it would further prevent parsing the different mechanisms through which political events are transmitted between states by clustering the information in the spatiotemporal lags. Moreover, since matrix decomposition is unsupervised, there is no guarantee that the components with the highest eigenvalues demonstrate any systematic association with the dependent variable.

Among the general benefits of regularization are (a) to allow for complex specifications, thus avoiding omitted variable bias, and (b) to constrain the complexity of the solution by reducing the magnitude of the parameter estimates, thus avoiding the high variance in estimates common when fitting linear models in the context of multicollinearity. As Schrodt (2014, 288) contends, “The specification of a linear model must always steer between the rock of collinearity and the hard place of omitted variable bias”, and elastic net regularization is intended here as a concise manner of doing just that.

2.2.2 Feature standardization

A necessary prerequisite for fitting linear models with regularization is standardization of the independent variables. This need is a direct function of the penalty term in (2.4). Because the magnitudes of the estimates in linear models are traced both to the strength of the association and to feature scaling, the elastic net penalty unfairly punishes features with a smaller scale.\(^7\) To ensure that the features are equally subjected to the regularization penalty, they are first standardized by subtracting each element by the mean value and dividing by the standard deviation (Hastie, Tibshirani, and Friedman, 2009, 63-4). This ensures that the resulting distribution of each feature has a mean value of zero and unit variance.

\(^7\)Precisely, a one-unit change in a binary feature is associated with a much larger marginal effect in the dependent variable than a one-unit change in a continuous variable that takes values between 0 and 10 million, also being equal.
There are two main consequences of standardization. First, the interpretation of the estimates is made more difficult by changing the scale of the features. For example, interpretation of the odds of an event given a one-unit change in a given feature is no longer as substantively meaningful.\textsuperscript{8} However, direct interpretation of the estimates is largely avoided in the analyses below. Rather, the preferred method of conveying the substantive significance is to plot the predicted probability of an event over values of the feature in question, and where relatively large changes in the predicted probability are observed in high-density ranges of the feature, those changes are assumed to be of substantive importance.

Second, a benefit of standardization is that the magnitudes of the estimates are contingent only on the strength of the association, and therefore they can be compared directly across the features (Gelman, 2008). This aspect directly motivates the visual mode of presentation for the parameter estimates. Namely, because the distributions of parameter estimates are shown visually for both the permutation and bootstrapping procedures, described in Section 2.2.4, it is easy to compare the strength of the associations and locate features that are of greater relative importance.

### 2.2.3 Cross-validation

In order to minimize (2.4), the hyperparameters $\lambda$ and $\alpha$ must be held constant. To choose the combination of these values that best produces a generalizable model, a grid search over a specified range of $\lambda$ and $\alpha$ is evaluated in terms of validation error. Cross-validation is helpful for this purpose in that it permits averaging out the variance associated with the models trained on different subsets of the original training data. In this case, stratified 10-fold cross-validation is used, and the metric employed for validation error is the negative log-likelihood for logistic regression. This metric is chosen over alternative metrics, such as classification accuracy, because of the low baseline probability of observing an authoritarian regime failure. Binary classification problems are known to be difficult

\textsuperscript{8}In this case, a one-unit change corresponds to a change of one standard deviation.
in the presence of class imbalance (e.g., King and Zeng, 2001), and the choice of accuracy as an evaluation metric tends to bias models in favor of the majority class.\footnote{Note, for example, that we can achieve 95\% accuracy by simply and naively classifying all cases as a non-event when the rate of events is 0.05.} Using the log-likelihood as the validation error metric ensures that we get a more representative picture of performance on each of the samples in the validation set. In the end, one combination of $\lambda$ and $\alpha$ yields the minimum loss in cross-validation, and this hyperparameter set is used for estimating the parameters in (2.4).

\subsection*{2.2.4 Parameter uncertainty}

Once optimal values of the hyperparameters are chosen, parameter estimates are obtained from a model fit on the full data, so the focus turns to statistical inference. While regularization provides benefits in terms of the generalizability of the model, it does so at the expense of standard analytical methods of quantifying uncertainty. This section details the two numeric methods employed in place of standard errors.

\textbf{Permutation significance}

The first method is to numerically simulate the sampling distribution of the null hypothesis for each parameter estimate, which is here referred to as permutation significance. The procedure to achieve this is to randomly permute the values of the dependent variable 100 times, refitting the model with each permutation. In expectation, random permutation of the dependent variable $y$ should nullify any systematic association between the vector of features $X$ and the outcome $y$, except by chance. With 100 iterations, we therefore obtain a null distribution for the estimates, centered at zero and with tails representing the upper and lower bounds, that we would observe by chance. Comparison of the actual statistic to this distribution therefore yields a quantity similar in motivation to a $p$-value, which captures the probability of observing a statistic of equal or greater magnitude given that the null hypothesis is true. In this case, the percent overlap in the estimate and the
simulated null distribution yields a quantity that is treated effectively as a $p$-value for a two-tailed hypothesis test.

**Bootstrap confidence intervals**

The second method for gauging the uncertainty in the estimates is to generate bootstrap confidence intervals. In this case, 100 bootstrap samples are drawn from the full set of training samples and the model is fit on each, yielding a numeric simulation of the sampling distribution for the test statistics. The result is treated as a confidence interval for the statistic. Where the distribution is distinguishable from zero, in the sense that it does not overlap (significantly) with zero, there is greater confidence of a statistically significant association between the feature and the dependent variable. Last, we should note that the median estimates from this bootstrapping procedure are used as the observed statistics for evaluating permutation significance.

### 2.2.5 Melding quantities of interest

One additional source of complexity in obtaining parameter estimates is the use of multiple imputed versions of the data, each of which are slightly different given the inherent uncertainty in the imputation process. In the analyses below, the procedures outlined in Sections 2.2.2, 2.2.3, and 2.2.4 are performed independently on each imputed data set, and the estimates from each analysis are then combined, or melded, as suggested by Honaker and King (2010, 564). As 100 estimates are learned on each data set for both the permutation significance procedure and the bootstrap confidence intervals, the melding across the ten imputed data sets yields distributions of 1000 estimates.

It should be noted that combining results across data sets implies that the set of hyperparameter values selected at the validation stage is allowed to vary over the imputations. In sum, the variance of the parameter estimates in a single data set is obtained via the permutation or the bootstrap procedure, and the variance in the estimates attributed to the imputation process is accounted for by combining those procedures over
the ten imputed data sets.

### 2.2.6 Predicted probability plots

Having the distribution of coefficient estimates aggregated across imputations, it is then a simple extension to produce simulated marginal effects from the models for each estimate as a means for understanding the substantive significance of the estimates. Such a procedure is similar in spirit to the algorithm described in King, Tomz, and Wittenberg (2000), though the coefficients in this case are not drawn from a multivariate normal distribution specified with parameters learned by the model, but rather they are generated through nonparametric resampling (i.e., the estimates learned from the 1,000 bootstrap samples drawn from the data). Specifically, using each bootstrap estimate, the expected values of the predicted probability of an event are evaluated for the full range of relevant values of some variable of interest, $x_k$. In this case, the variables $x$ identified for evaluation are a subset of the variables found to have associations statistically different from zero, using the criteria in Sections 2.2.4. Following the recommendations of Hanmer and Kalkan (2013), the observed-value approach for simulating the expected values is used in lieu of holding the other independent variables in the model at their average values. This procedure amounts to holding $x_k$ constant at each value in its range, then obtaining predicted probabilities using each bootstrap estimate, and finally plotting the distribution of expected predicted probabilities for each value of $x_k$. In the plots, the median predicted probability and a 90% confidence band is superimposed on a histogram of $x_k$ to visually convey the areas of greatest density, which speaks to the substantive relevance of the predicted value at that value of the variable.

### 2.2.7 Results

There are two models presented in this section, one for authoritarian regime failure and the other for regime failures resulting in a democratic transition. The results are presented
visually in three formats. For the bootstrap confidence intervals, the distributions of the coefficient estimates are provided in boxplots and are directly comparable to each other as well as to a line representing the null hypothesis. In presenting permutation significance, the null hypothesis is represented by a boxplot centered at zero, while the median bootstrapped point estimate is shown to represent the observed coefficient estimate. Last, a selection of the features found to have a statistically significant and relatively large association with the respective failure indicators are chosen for presentation in predicted probability plots.

**Authoritarian Failure**

The results for the model of authoritarian regime failure are shown in Figures 2.9 and 2.10. Many of the findings are consistent with prevailing models of regime failure. In particular, military regimes are more likely to fail, all else equal, than regimes of other types, which is consistent with the arguments laid out in Geddes (2003). Single-party regimes, on the other hand, are less likely to fail than regimes of other types. In addition, the results suggest that regimes with higher GDP per capita, regimes higher economic growth rates, and regimes with higher levels of natural resource income are all less likely to fail, all else equal.

To start with the spatiotemporal lags, the lagged variables weighted by geographic distance are generally the largest in magnitude. Most notably, the lag of democratic transitions is statistically distinguishable from the line at \( x = 0 \), and the geographic lag of regime failures of all types just barely crosses this line. Both of these findings are consistent with expectation, in that regime failures and, more specifically, democratic transitions lead to learning on the part of opposition groups. While overlapping the line at \( x = 0 \) more heavily, the geographic lag of autocratic transitions is the also the largest in magnitude among the negative associations. This provides suggestive evidence that regime failures that result merely in the establishment of a new authoritarian regime tend to dissuade regime failures in other locations.
A key insight in these findings is that the lag of democratic transitions, and not merely the lag of authoritarian failures of any kind, provide statistical leverage in the model of failures, which in turn speaks to the idea that an array of events beyond the dependent variable may need to be considered in spatial models of political instability. Second, that the findings speak primarily to transnational learning on the part of opposition groups is preliminary evidence that regimes are not able to withstand pressure when opposition groups are motivated by their international context. In Chapter 3, this notion is explored further to see if regimes of certain types are better able to survive in response to trigger events than others.

To explore the substantive importance of the spatiotemporal lags highlighted here, the predicted probability of regime failure is compared across the full range of these variables as well as the ranges of variables that are prominent in the literature and that are found here to have statistical associations with failure, including economic growth, natural resource income, and GDP per capita. A subset of these are shown in Figure 2.11. Consistent with the findings in Figures 2.9 and 2.10, the predicted probability of a failure is shown to be increasing in both the geographic lag of failures and the geographic lag of democratic transition (panels a and b, respectively), while it is decreasing in GDP per capita and natural resource income (panels c and d). In the plots, the 90% confidence band is shaded gray, while the median is represented by a dark line.

What is apparent from these plots is that the change in the probability of failure with respect to changes in any of these attributes is quite small in magnitude. To illustrate, given a change from the 25th to the 75th percentile of the geographic lag of failures, the increase in the probability of failure is 0.0010 (from 0.0407 to 0.0417). For the lag of democratic transitions, the change in the probability of failure corresponding to the same change in the lag is 0.0013 (from 0.404 to 0.0417). Again, these effects are very small, but they are similar to the changes due to domestic attributes that are well established in the literature. Notably, the same query for the effect of GDP per capita yields a change in the probability of failure of -0.0019 (0.0423 to 0.0404). For natural resource income
Figure 2.9: Bootstrap confidence intervals for the parameter estimates in the model of authoritarian regime failure. Estimates for regional and two-year period indicators are omitted for clarity.
Figure 2.10: Permutation significance for the parameter estimates in the model of authoritarian regime failure. Estimates for regional and two-year period indicators are omitted for clarity.
per capita, the effect is -0.0034 (0.0433 to 0.0399), and for economic growth, -0.0028 (not shown, 0.0427 to 0.0399).

In sum, the effects associated with substantively important changes in the spatiotemporal lags are small, but the effects associated with the most important continuous domestic attributes are likewise small, if larger comparatively. This is to be expected in light of the fact that the baseline rate of failure is so low in the data. Therefore, in spite of the small effect sizes, we can suggest with reasonable uncertainty that international context, captured here by geographic proximity to regime instability, is relatively important for explaining instances of regime failure.

**Democratic Transition**

The results for the model of democratic transition are shown in Figures 2.12 and 2.13. Similar to the model of authoritarian failure, many attributes common to models of regime instability are found to have the expected associations here. Namely, we see that monarchies and single-party regimes are relatively less likely to end in a democratic transition, and military regimes are relatively more likely to end in a democratic transition. Personalist regimes are found to have no distinguishable association in this model. In addition, economic growth is again found to be negatively associated with the outcome, suggesting that periods of growth are associated with regime stability. Resource income, is likewise found to be negatively associated with democratic transitions, though with slightly less certainty than the model of authoritarian failure. In contrast to the model above, however, GDP per capita is found to have no statistical association with democratic transition.

The spatiotemporal lags in this model suggest some interesting associations. First, economic ties tend to be more important in general for explaining democratic transitions than for explaining authoritarian failure more broadly. Second, we see clearer separation here between the lags of democratic and autocratic transitions. Namely, the economic and geographic lags of democratic transitions appear positively associated with democratic
Figure 2.11: Predicted probability of authoritarian regime failure over select variables: (a) the geographic weighted lag of regime failures; (b) the geographic weighted lag of democratic transitions; (c) the lagged two-year moving average of economic growth; and (d) natural resource income. Probability estimates are derived from estimates shown in Figures 2.9 and 2.10.
transition, while the geographic and, with lesser certainty, economic lags of autocratic transition appear to reduce the probability of a democratic transition. These findings provide evidence in support of the idea that opposition groups update their expectations about the success of challenging the regime based on the experiences of opposition groups in similar positions abroad.

Moreover, the findings once again seemingly suggest that regimes learn less well than opposition groups. Notably, if we assume that regimes wish to stay in power irrespective of the outcome of a transition event, it is not readily clear why they would learn from observing autocratic transitions abroad while not learning from democratic transitions. On the other hand, if regimes do better in a transition to democracy than in a transition to autocracy, there is reason to expect a more concerted effort to prevent a transition (of any kind) in response to observing autocratic transitions. As noted in Section 1.3, regimes desire to control the process of a transition to democracy in certain contexts, thus ensuring a favorable position in the new democratic regime. In this case, they may be more willing to cede to international pressures for democratization in light of wave of democratic transition events. Again, we explore this idea more carefully across regime type classifications and regions in Chapter 3.

In Figure 2.14, the change in the probability of a democratic transition as a function of select features are visualized. As above, the effects suggested by the model amount to only very small changes in the probability of an event. As an example, given a change from the 25th to the 75th percentile of the geographic lag of democratic transitions, the resulting increase in the probability of a democratic transition is 0.0005 (from 0.0181 to 0.0186). For the same change in the economic lag of democratic transitions, the change in probability is slightly greater at 0.0008 (0.0180 to 0.0188). The drop in the probability of a democratic transition given a shift from the 25th to the 75th percentile of the geographic lag of autocratic transitions is here estimated to be 0.0006 (0.0188 to 0.0182). By comparison, the drops associated with similar shifts in economic growth (not shown) and natural resource income are estimated to be 0.0009 (0.0189 to 0.0180) and
Figure 2.12: Bootstrap confidence intervals for the parameter estimates in the model of democratic transition. Estimates for regional and two-year period indicators are omitted for clarity.
Figure 2.13: Permutation significance for the parameter estimates in the model of democratic transition. Estimates for regional and two-year period indicators are omitted for clarity.
0.0008 (0.0189 to 0.0181), respectively.

The model of democratic transition, despite suggesting several statistical significant associations in Figures 2.12 and 2.13, nonetheless also suggests that these associations translate into only slight adjustments in the predicted probability of an event in Figure 2.14. While small, these changes may nevertheless be substantively important for explaining variation in the realization of democratization events.

2.3 Remarks

To this point, we have seen preliminary evidence that spatiotemporal lags of authoritarian regime failure events are statistically associated with both regime failure and democratic transitions. Particularly, we have seen evidence that lags of autocratic transitions tend to reduce the probability of the events of interest, whereas lags of democratic transitions tend to increase the probability of such events. Further, we have seen evidence that two specific channels of diffusion, geography and trade, tend to better explain the transmission of these events than a third, linguistic ties. Last, while the estimated substantive importance of these lagged variables is small in magnitude, the results have likewise shown that they are not categorically less important than domestic attributes commonly associated with regime stability and democratization, including economic growth, GDP per capita, and natural resource income per capita.

Moreover, because penalized logistic regression is presented in Section 2.2.1 as an appropriate method for addressing common specification problems in quantitative models in comparative politics, it is natural at this juncture to question whether grid search over hyperparameters for the elastic net is worth the trouble. To reiterate the problem, the concern for fitting maximum likelihood models in high dimensional feature spaces is that the model picks up idiosyncracies in the training data that do not generalize. In the context of the models of regime failure and democratic transition above, this is a valid concern since we are specifying two-year period and region dummies on top of a long
Figure 2.14: Predicted probability of democratic transition over select variables: (a) the geographic weighted lag of democratic transitions; (b) the trade weighted lag of democratic transitions; (c) the geographic weighted lag of autocratic transitions; and (d) the lagged two-year moving average of economic growth. Probability estimates are derived from estimates shown in Figures 2.12 and 2.13.
list of “control” variables. Accounting for the spatiotemporal lags and the intercept, this sums to 68 parameters to be estimated by the model.

On top of the simple count of terms in the model, many of those terms are highly correlated, which is a known violation of the assumptions underpinning statistical inference in linear models. This is evident in Figure 2.7, where we see several non-negligible bivariate correlations between the economic features. It is also evident in the relationships among the spatiotemporal lags, which are correlated by construction. Namely, for any given unit of observation (i.e., regime \( a \) in year \( b \)), the events falling within the prespecified temporal window of 365 days are invariant. The variation in the lags is therefore attributable only to the distance weighting scheme. In addition, these weighting schemes are related. As described in Section 1.4, geographic distance has historically been used as a proxy for other notions of connectivity among states. This is intuitive, since countries closer in geographic space are more likely to engage in trade and because languages tend to cluster in geographic space. Given this level of correlation among the inputs, \( L_2 \) regularization is a necessity to constrain the magnitude of the estimates learned by the model.

In turn, we can see the hyperparameters for the elastic net chosen in cross-validation in Table 2.1 for the models of authoritarian regime failure and democratic transition. Again, as we see in (2.4), \( \lambda \) controls the overall amount of regularization imposed on the parameter estimates and \( \alpha \), otherwise known as the \( L_1 \) ratio, controls the tradeoff between \( L_1 \) and \( L_2 \) penalization, which are the lasso and ridge penalties, respectively. The results in Table 2.1 show that the models with the best performance in 10-fold cross-validation have a non-zero amount of regularization, which indicates that a logistic regression with no regularization is less appropriate for this specification. Because the elastic net nests the unpenalized logistic regression (i.e., the case where \( \lambda = 0 \)), these results would reveal if penalization were unnecessary. Given that we expect collinearity to be a problem, verification of a positive value for \( \lambda \) makes sense. Second, the optimal \( L_1 \) ratio is found to be zero for both models across all imputations, in which case the parameterization
Table 2.1: Elastic net parameters chosen in cross-validation for each imputation.

<table>
<thead>
<tr>
<th></th>
<th>Regime Failure</th>
<th>Democratic Transition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\lambda$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Imputation 1</td>
<td>0.2512</td>
<td>0.0</td>
</tr>
<tr>
<td>Imputation 2</td>
<td>0.2512</td>
<td>0.0</td>
</tr>
<tr>
<td>Imputation 3</td>
<td>0.1995</td>
<td>0.0</td>
</tr>
<tr>
<td>Imputation 4</td>
<td>0.2512</td>
<td>0.0</td>
</tr>
<tr>
<td>Imputation 5</td>
<td>0.1995</td>
<td>0.0</td>
</tr>
<tr>
<td>Imputation 6</td>
<td>0.1995</td>
<td>0.0</td>
</tr>
<tr>
<td>Imputation 7</td>
<td>0.2512</td>
<td>0.0</td>
</tr>
<tr>
<td>Imputation 8</td>
<td>0.2512</td>
<td>0.0</td>
</tr>
<tr>
<td>Imputation 9</td>
<td>0.3162</td>
<td>0.0</td>
</tr>
<tr>
<td>Imputation 10</td>
<td>0.2512</td>
<td>0.0</td>
</tr>
</tbody>
</table>

in (2.4) simplifies to a ridge penalty on the parameter estimates. Again, this is intuitive because ridge penalization is found to perform better than the lasso in the presence of multicollinearity (e.g., Kuhn and Johnson, 2013), and the consequence for the results of removing $L_1$ penalization is that all parameter estimates are non-zero.

In the end, based on the results in this section, it could be argued that spatiotemporal lags of an array of events that speak to regime instability are found to be statistically and substantively important for explaining authoritarian regime failures and democratic transition. However, the analysis has thus far assumed that these outcome variables are a linear and additive function of the inputs. In addition, beyond the cross-validation procedure employed for selecting hyperparameters, the models have not been subjected to careful evaluation in terms of out-of-sample predictive performance. In the next section, the analysis is extended to more flexible machine learning algorithms to allow for less biased functional forms, and then model performance is compared in a series of prediction exercises. These tests allow for a more robust exploration of counterfactual scenarios under which authoritarian regime failure and democratization are more or less likely to occur.
Chapter 3

Predictive Models of Regime Failure

Linear models of political phenomena are widely employed in the social sciences, and they offer many attractive benefits. Chief among these is the fact that we obtain a single parameter estimate describing the learned association between each input variable and the output variable. As such, the results from linear models are relatively easy to interpret. Even in cases where linear models do not produce analytical standard errors, such as the model described in Section 2.2.1, it is still relatively straightforward to numerically engineer estimates of the uncertainty in the estimates and, in turn, make inferences about the statistical and substantive associations in the data.

Yet, linear models rely on the assumption that a linear functional form is appropriate for the data. Moreover, an important shortcoming of these models is that any suspected (hypothesized) multiplicative interactions between input variables must be specified by the analyst. In the models above, the number of potentially meaningful interactions among the inputs is quite high. For example, consider whether the association between spatiotemporal lags of authoritarian failure events is variable across regime type classifications. With twelve spatiotemporal lags and four regime types, an additional 48 terms would need to be specified and interpreted. When interactions among the other inputs are considered, not to mention triple or even higher order interactions, understanding which among the total set of possible interactions are meaningful from a statistical or substantive point of view quickly becomes an intractable exercise for both the analyst and the reader. For these reasons, the linear model is typically characterized by high bias, in the sense that its specification does not often accurately reflect the true data generating process. In other words, the linear model is potentially underfitting the data.

To better accommodate the possibility of more flexible functional forms in the rela-
tionships between input features and outputs, there exist a number of machine learning algorithms capable of learning functions of the data without attempting to specify an unknown number of higher order terms. As such, they are more resistant to biases traced to specification and the assumption of linearity. In this section, two additional classification algorithms, the random forest and support vector machine classifiers, are used to address these shortcomings in the linear model. In contrast to the linear model, the random forest and support vector machine can be very low bias, in the sense that they can learn functions of the data that very accurately predict the samples on which the model is trained. On the other hand, these classifiers tend to be high variance, given that the function learned by the model may be idiosyncratic to the training data (i.e., overfitting). As a result, the models below are subjected to rounds of model validation to select tuning parameters with the objective of learning the most generalizable model from the training data.

The primary objective of the models presented in this section is prediction. This is motivated primarily by the observation that statistically significant findings in comparative politics and international relations research have not typically translated into accurate, generalizable predictions that are perhaps more useful from a policymaking perspective (Montgomery, Hollenbach, and Ward, 2012; Schrodt, 2014; Ward, Greenhill, and Bakke, 2010). Therefore, the models below are evaluated only in terms of their ability to predict outcomes in withheld future observations. That being said, consistent with the linear models presented above, the goal of understanding associations in the data and historical explanation is not discarded. Rather, given the models with the best predictive performance, this section also extracts and presents substantive effects to better understand the generalizable conditions under which regime failure and democratization events have occurred historically and are more likely to be observed moving forward.

This section presents and compares the predictive performance of three classification algorithms, including logistic regression with elastic net penalization, random forests, and support vector machines. In addition, the performance of each of these individual
classifiers is compared to their cumulative performance in an ensemble Bayesian model averaging (EBMA) framework. The component classification algorithms are all available in scikit-learn (Pedregosa et al., 2011), and the programs used to validate the models in the analyses are presented in Appendix 5. The EBMA functionality is available in the EBMAforecast library in R (Montgomery, Hollenbach, and Ward, 2016).

3.1 Validation strategy

As is the case with elastic net regularization for the linear model, the random forest and support vector machine classifiers have hyperparameters that must be tuned when fitting the models on the training data. Again, the objective in choosing hyperparameters through model validation is to ensure that the function learned on [subsets of] the training data generalize to withheld samples. In the case of the random forest, models are validated on out-of-bag samples, and stratified 10-fold cross-validation is used to tune the support vector machine. The validation metric is again the negative log-likelihood function used in logistic regression, so the objective is to minimize

$$
-\frac{1}{N} \sum_{i}^{N} y_i \log (\hat{\pi}_i) + (1 - y_i) \log (1 - \hat{\pi}_i).
$$

(3.1)

Here, $y_i$ is the observed outcome and $\hat{\pi}_i$ is the predicted probability assigned by the model. This metric is preferable to common alternatives such as classification accuracy in light of the class imbalance. Model validation is again performed on each of the ten imputed data sets, and the hyperparameters selected in validation are allowed to vary over the imputations.

For each of the classifiers in this section, including the regularized logistic regression model described in Section 2.2.1, the model is trained and validated on observations for the period from 1946 to 2000. The observed proportions of authoritarian regime failures
and democratic transitions in the training data are 0.051 and 0.022, respectively. Once the best performing models are learned on the training data, the predictive performance for each of the classifiers is compared in the test period, having observations for the years 2001 to 2010. The observed proportions of regime failures and democratic transitions in the test data are, respectively, 0.030 and 0.020, noticeably lower than the rates in the training data.

The underlying motivation in this procedure is to engineer a scenario in which future unobserved samples are predictive given only the information in historical observed samples. Similar procedures are used in Montgomery, Hollenbach, and Ward (2012) and Beger, Dorff, and Ward (2014). In this way, we can observe the utility of historical models for uncovering patterns in the data that generalize to the future. In the end, the results reveal a useful picture of model performance as relates to the primary objective in this research agenda. Because the central question in this work is whether transnational learning is occurring among actors in authoritarian regimes, predictive performance is compared between models that contain the spatiotemporal lags of regime failure events and restricted models that suppress those variables. Therefore, the results indicate how generally useful spatiotemporal lags of regime instability are for correctly predicting authoritarian regime failures and democratic transitions.

3.1.1 Random forest

The random forest classifier is an ensemble of decision tree classifiers. Whereas trees tend toward high variance, random forests aggregate the output of many independent and decorrelated trees, which reduces variance without sacrificing the low bias of a single tree (Breiman, 2001; Hastie, Tibshirani, and Friedman, 2009, 588, 600). The decorrelation is traced to two sources of randomness in the construction of the trees. First, individual trees are grown on bootstrapped samples of the data, and thus random forests are a form of bootstrap aggregation, or bagging. Second, at each stage of the tree construction,
only a random subset of the total number of variables are used to generate node splits (Breiman, 2001).

A convenient consequence of bagging in the random forest is that each tree is trained on a subset of the total observations. On average, approximately two-thirds of the training samples are selected for each tree estimator, and the remaining one-third are referred to as out-of-bag, or OOB, samples. The benefit of OOB samples is that the model can be validated on withheld observations without the computational expense of performing cross-validation. Specifically, consider that k-fold cross-validation requires the construction of k random forest models, whereas validation on the OOB samples only requires fitting a single random forest, so the latter option is much more desirable from a computational standpoint. Validation is achieved in this framework by passing each of the training samples down each of the tree estimators in which it was not sampled, averaging the predicted probability generated across estimators for those samples, and then evaluating the loss (Hastie, Tibshirani, and Friedman, 2009, 592–3). The aggregate loss is the validation error for the model.

A complicating factor in using OOB samples to validate the model is that the only built-in validation metric for ensemble classifiers is accuracy. As such, a custom Python program was coded to allow specific of any valid classification loss metric for binary classification problems using ensembles of trees, shown in Appendix 5.\textsuperscript{1} Using this tool, a grid search over hyperparameters is conducted to find the model that best generalizes. The grid search is conducted independently on the ten imputations, and each trained model is used to make predictions on the future withheld observations. The mean predicted probability is obtained across the imputations for each of the test samples, and this value is able to be evaluated by a number of classification performance metrics, described below in Section 3.3.

The random forests used in the analysis are made up of 2,000 classification trees.

\textsuperscript{1}In addition to the random forest classifier, this class would be valid for a pure bagging classifier as well as for extremely randomized trees classification. In these cases, as with the random forest classifier, the trees in the ensemble are constructed independently of one another.
Because random forests do not overfit in the number of individual tree estimators, this number is chosen to simply have a sufficient number of trees to average out the variance of complex functional forms learned by the trees in the ensemble. Hyperparameters considered in the grid search include the interaction depth of the trees and the number of features randomly selected at each node to choose the optimal split. Each of these tuning parameters controls, in part, the bias-variance tradeoff for the random forest. Specifically, as classification trees grow deeper, they are able to make more splits and partition the input space to classify the training samples, which results in low bias but high variance estimates. In the extreme, a classification tree with no constraints on depth could perfectly fit the training data, but such an estimator is unlikely to generalize well to test data. The random forest is able to average out the results of these high variance trees, but constraints on the depth of its constituent trees are considered in the grid. The specific set of values of the interaction depth included in the grid is \{3, 5, 9, 15, None\}.

The objective in varying the number of terms \(m\) considered at each split of the trees constituting the random forest is to control the correlation between those trees. As \(m\) increases, the trees in the forest behave more like unconstrained classification trees, in which all terms are considered in determining each node split and are therefore deterministic on the same data. Because trees are deterministic, increasing \(m\) in a random forest results in higher correlation among the trees, which inhibits the ability of the forest to reduce the variance of any single tree classifier fit on the data. With \(p\) total terms in the model, a typical value for \(m\) is \(\sqrt{p}\). The set of values included in the grid is \(\{\sqrt{p}, \ln(p), 4\}\).

### 3.1.2 Support vector machine

The support vector machine classifier seeks a decision boundary in the feature space to best separate the classes, but it allows for boundaries that may not be linear in the inputs by using a kernel function to transform the inputs (Hastie, Tibshirani, and Friedman, 2009, 423–4). For example, a simple kernel function might be a polynomial transformation.
of the original inputs, yielding a specification that includes the original inputs and all possible interactions of those inputs up to the desired order. The models below use the radial basis function (RBF), or Gaussian kernel, defined as

$$K(x, x') = \exp\left(-\gamma ||x - x'||^2\right),$$

where $\gamma$ is defined as $\frac{1}{2\sigma^2}$ and $||x - x'||^2$ is the Euclidean norm of the distance between two samples (or vectors) $x$ and $x'$. In practical terms, because the norm in this case is a measure of similarity between vectors, $\gamma$ controls the effect a sample $x'$ has in determining the predicted value of $x$. Namely, if $\gamma$ is large, the implication is small $\sigma$, and therefore the radius around $x$ is quite small and the information from very few samples $x'$ is used in classifying $x$. In this way, as $\gamma$ increases, the potential for overfitting increases. By contrast, a small value for $\gamma$ implies a large $\sigma$ and a wider radius that captures information from a larger proportion of the total training samples in making a decision about the predicted value of $x$, which constrains overfitting. Values for $\gamma$ that provide the best balance in the tradeoff between bias and variance are chosen in cross-validation and allowed to vary over imputations.

The decision boundary learned by the support vector machine is also affected by a cost parameter, $C$. Left unconstrained, the objective of the support vector machine is to learn a boundary that perfectly separates the classes and maximizes the distance between samples of different classes. Intuitively, we could imagine that a complicated boundary that perfectly separates the classes in the training data may in fact be picking up idiosyncrasies in those data, which in turn limits the generalizability of the model. To constrain the complexity, we can reduce the cost paid for incorrect classification by reducing $C$, which reduces the overall loss in the training data and generates a smoother decision boundary. High values of $C$ impose higher loss for misclassification, producing a more complicated decision boundary. Again, $C$ is chosen in cross-validation and is
variable in the imputed data.

3.1.3 Ensemble Bayesian model averaging

Prevailing work on predicting political events has found that aggregating the information in component classification models to form an ensemble classification produces better performance than the component models individually. In particular, Montgomery, Hollenbach, and Ward (2012) use ensemble Bayesian model averaging (EBMA) to weight the contribution of the predictions from component models, which in this case are the predictions from the penalized logistic regression, random forest, and support vector machine classifiers that perform best in the training period. The weights are determined as a function of the performance of each of component classifier and the uniqueness of its contribution. That is, the weight of any one classifier in the ensemble is increasing in the extent to which its prediction deviates from the other component classifiers and the performance of that prediction.

The weighting scheme described here is referred to as model calibration, which serves as additional form of model validation to what has been described thus far. In practical terms, the predictions of the component models are calibrated on a subset of the observations from the test period defined above (2001 to 2010). Specifically, to compare the EBMA predictions to those of its component models, the predictions from the latter are calibrated on the years 2001 to 2005, hereinafter called the calibration period, and the resulting weighted ensemble (i.e., the calibrated EBMA predictions) can then be compared to the predictions of the other models in the calibration period and a subsequent test period, defined here as the remaining years from 2006 to 2010. As a result, because the calibration period uses up some of the test observations used for comparing the component models individually, the comparison of EBMA to the component models is done on a shorter test period with fewer observations. The proportion of authoritarian regime failure events in the calibration and test periods are 0.033 and 0.027, respectively, while
the proportions of democratic transitions across these periods are 0.020 and 0.021. Refer to Montgomery, Hollenbach, and Ward (2012) and Beger, Dorff, and Ward (2014) for other examples of this validation design.

### 3.2 Model performance metrics

Once model training and validation is complete, the best performing models are used to make predictions on the withheld samples for future years. Having obtained predicted probabilities for the withheld observations in each imputed data set, the predictions are aggregated and evaluated with a number of classification metrics, including negative log-likelihood in (3.1), Brier score loss, the area under the Receiver Operating Characteristic (ROC) curve, and the area under the Precision-Recall (PR) curve. The use of Brier scores and ROC curves for classification problems in political science are described in Greenhill, Ward, and Sacks (2011) and Montgomery, Hollenbach, and Ward (2012), among other places. Brier score loss captures the average squared difference between the observed discrete values and the predicted probabilities.

\[
Brier = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{\pi}_i)^2
\]

As a loss metric, Brier scores closer to zero are indicative of better fit, and in this way it is similar to the loss defined in (3.1). The ROC curve captures the tradeoff between the true positive rate and false positive rate while varying the classification threshold to take all observed values of predicted probabilities in the data. Areas under the curve closer to one are preferable, and this occurs when the observed classes are well separated by their predicted probabilities. Similarly, the area under a precision-recall curve gauges the tradeoff between recall, which is the true positive rate, and precision. These metrics are typically evaluated to determine how well the model predicts the positive class, which
refers to regime failure events in this case.

\[
\text{Recall} = \frac{TP}{TP + FN} \quad \text{Precision} = \frac{TP}{TP + FP}
\]

In essence, for a given classification threshold, recall describes how well the model correctly predicts the observed events in the data, and precision describes how well the model performed on the predicted events in the data. For the former, the primary concern is false negatives (FN), whereas false positives (FP) are the greater concern for the latter. Both precision and recall are increasing in the number of true positives (TP). The precision-recall curve is constructed by again allowing the classification threshold to vary along each observed predicted probability in the test data and evaluating each of these metrics at each threshold. As with the ROC curve, the objective is to maximize the area under the curve.

Uncertainty in the Brier score loss and the area under the curve metrics are obtained by bootstrapping the predicted probabilities, stratifying on observed class to ensure roughly proportional representation of events and non-events, and then evaluating the statistic. The plots below compare the distribution of these statistics for restricted models that omit the spatiotemporal lags and full models that include them. This speaks to the central objective of this exercise, which is to understand the degree of value added in predictive models by specifying the spatiotemporal lags.

It is notable here that classification metrics that rely on discrete predictions are absent. The focus on predicted probabilities in place of discrete classifications reflects the continuing concern with the low baseline probability of observing a rare event. In other words, because each of the models assigns probabilities well below the default classification threshold of 0.50 for predicting an event, metrics such as recall go to zero and global accuracy simply yields the proportion of non-events, which are non-informative results. In spite of the low predicted probabilities produced by the models for the events
in question, the chosen metrics allow us to see how well the models separate events from non-events in terms of their predicted probabilities. That is, ideally the probabilities assigned to observed events should be higher than the probabilities to non-events, and in turn judgment could be made about the cases in which a regime failure event is most likely to occur.

3.3 Model interpretation

Once predictive performance is compared, we can proceed with interpreting the associations between the features in the model and the probability of a regime failure in the training data that produced the finding. The procedure for displaying the change in the predicted probability over substantively interesting values of the features is similar to the procedure described in Sections 2.2.6 and shown in the results above. Namely, probabilities are evaluated for a sequence of relevant values of the feature in question, and the change in probability is visible over the distribution of that variable. The key difference in this section is that confidence bands to show the uncertainty in the probability estimate are omitted. This is attributed to the computational burden of drawing bootstrap samples from the training data and, for example, passing those samples down each of the 2,000 tree estimators in the fitted random forest, especially in light of the fact that this has to be done independently for ten imputed training data sets. As such, only the mean predicted probability across the ten imputations is plotted across the range of the features.

Despite paying a price in interpretability in terms of not being able to express uncertainty, the probability plots in this section benefit from the complexity of the algorithms in that interactions can be tested by holding the values of more than one of the features constant. For example, in the results below, we are interested in how the association between the spatiotemporal lags is variable over authoritarian regime types. The use of the random forest and support vector machine classifiers allows the analyst to test such
hypotheses without varying the model specification, only the counterfactual for which the predicted values are evaluated.

The random forest and support vector machine do not provide concise information about the features that are most important to the model.\(^2\) As such, there is no simple way to decide which of the features to explore via predicted probability plots, and therefore a brute force method of producing the predicted probability plots and exploring the results is employed to uncover the associations that are of most substantive interest. These are the plots chosen for presentation in the next section.

### 3.4 Prediction results

This section summarizes the prediction results for each of the learning algorithms for the two outcomes under examination, authoritarian regime failure and democratic transition. For both outcomes, the restricted specification that omits the spatiotemporal lags is compared to the full specification. The associations between select independent variables, as well as interactions therebetween, and the outcomes are also explored in predicted probability plots.

#### 3.4.1 Regime failure

The performance for the three algorithms on the test period 2001–2010 in predicting authoritarian regime failure is displayed in Table 3.1. There are two notable takeaways in these results. First, in terms of the algorithms, the SVM classifier produces the lowest average loss across the test observations, while the penalized logistic regression produces the greatest area under the ROC and PR curves. The random forest classifier is dominated

\(^2\)Feature importances are available for random forests, but these tend not to give precise information about the importance of the features in the presence of correlated inputs, such as the spatiotemporal lags in the present problem. For example, for two correlated inputs, the change in OOB performance by permuting values of the feature in question will be roughly equivalent for both inputs, given that they are randomly sampled into the set of features considered for determining optimal node splits for all tree classifiers in the ensemble.
Table 3.1: Prediction performance for models authoritarian regime failure across classifiers for the test period.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Specification</th>
<th>Log Loss</th>
<th>Brier Loss</th>
<th>AUC ROC</th>
<th>AUC PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penalized Logit</td>
<td>Full</td>
<td>0.1300</td>
<td>0.0288</td>
<td>0.8342</td>
<td>0.1185</td>
</tr>
<tr>
<td>Penalized Logit</td>
<td>Restricted</td>
<td>0.1275</td>
<td>0.0287</td>
<td>0.8397</td>
<td>0.1200</td>
</tr>
<tr>
<td>SVM</td>
<td>Full</td>
<td>0.1286</td>
<td>0.0286</td>
<td>0.8186</td>
<td>0.0991</td>
</tr>
<tr>
<td>SVM</td>
<td>Restricted</td>
<td>0.1233</td>
<td>0.0284</td>
<td>0.8298</td>
<td>0.1055</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Full</td>
<td>0.1360</td>
<td>0.0298</td>
<td>0.7470</td>
<td>0.0729</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Restricted</td>
<td>0.1275</td>
<td>0.0286</td>
<td>0.7609</td>
<td>0.0851</td>
</tr>
</tbody>
</table>

both in terms of the loss and area under the curve metrics. This is suggestive evidence that the relatively simpler findings in the linear model and support vector machine generalize best to the test period and, by extension, that the more complex nonlinear findings learned by the random forest classifier are picking up on associations that are idiosyncratic to the training period. Direct comparisons of some of these findings are presented below in the predicted probability plots.

Second, the restricted models dominate the models that include the spatiotemporal lags, which is evidence against the specification of the latter for the purpose of better predicting instances of authoritarian regime failure. It likewise indicates that the lags are explaining variation in the training data that does not generalize to the test period, since otherwise we would expect their influence on the test predictions to be minimal. Indeed, we can see from the distribution of the predicted probabilities for the test observations in Figure 3.1 that the predictions tend to be higher for the specifications containing the spatiotemporal lags, and this pattern is consistent across the classifiers. In this way, we expect that the lags are helping to explain failures in the training period, but failing to help predict failures in the test data.

There are many potential reasons for this, and principal among them is the possibility that the observations in the test samples are systematically different from the observations in the training data. For one thing, as noted above, the base rate of authoritarian regime failure in the test period is substantially lower than in the training period, dropping
from approximately 5% to 3% of the total number of observations. As such, the test performance in this sample of observations may not be decisive evidence against the value of the spatiotemporal lags. Moreover, in addition to fewer failures, the mechanisms underlying those failures may have less to do with transnational learning in the test period than in the training period. Again, the extent to which the period 2001–2010 marks a shift in the value of the lags for explaining or predicting regime failure events is an open question.

To get a sense of the effects learned by each of the classifiers, we can explore the change in the predicted probability of failure given changes in spatiotemporal lags. In Figure 3.2, the probability of failure is shown across values of the spatiotemporal lags of autocratic transitions weighted by geography and trade. The findings for the penalized logistic regression model in this case are muted, while the support vector machine suggests a negative association with failure. A negative associate was suggested without great
confidence in the results in Section 2.2.7, but here they are found to contribute little to the predictive performance of the model. The random forest model, in contrast, reveals a more complex U-shaped pattern wherein the probability of failure falls rapidly at lower values of the lag, but rebounds slightly at higher values of the lag. Returning to the question of comparative model performance, what these findings show is that the more flexible findings from the random forest and, to a lesser degree, the support vector machine are picking up on non-negligible relationships between the lags and authoritarian regime failure in the training data, but these associations are not generalizing to the test data. By contrast, the logistic regression does not suggest an association between the lag of autocratic transitions and regime failure, and we in turn observe only a minor difference in the performance of the restricted and full specifications for this model. In sum, although the support vector machine uncovers the negative association expected by theories of transnational learning, these findings ultimately support the null findings in the linear model in Section 2.2.7 in light of the comparative predictive performance.

A similar story can be told for the models in Figure 3.3, where the predicted probability of regime failure is plotted across lags of democratic transitions for all three classifiers. Namely, the logistic regression reveals a slight positive association between the geographic lag of democratic transitions, but no clear association between the lag weighted by trade. These findings echo the earlier results in Figures 2.9, 2.10, and 2.11. The support vector machine suggests stronger positive associations for these variables, while the random forest classifier again uncovers considerably more complex functional forms. The strength of these findings can likely be associated with the weaker performance of these methods.

We can look at specific cases to explore the comparative performance of the models. In Figure 3.4, the predicted probability of regime failure is plotted against observed regime failure events, marked by solid vertical lines, across the training and test periods. Several patterns stand out in these plots. First, the predictions from the random forest

---

3This mode of presentation is inspired by the separation plot described in Greenhill, Ward, and Sacks (2011).
Figure 3.2: Estimated change in the predicted probability of authoritarian regime failure across values of (a) the trade weighted spatiotemporal lag of autocratic transitions and (b) the geographic weighted spatiotemporal lag of autocratic transitions.
Figure 3.3: Estimated change in the predicted probability of authoritarian regime failure across values of (a) the trade weighted spatiotemporal lag of democratic transitions and (b) the geographic weighted spatiotemporal lag of democratic transitions.
classifier tend to be more extreme than those of the support vector machine and the logistic regression. One important consequence of this tendency is that the random forest often assigns the highest probability of regime failure to the observed regime failures. In Figure 3.4, this is true of the autocratic transition in Chad in 1990, the autocratic transition in Zaire in 1997, the string of failures in Haiti from the mid-1980s to mid-1990s, and the democratic transition in Haiti during the test period in 2004. On the other hand, the bold predictions of the random forest also produce a high number of false positives in the test period relative to the other two algorithms. For example, the random forest is generally paying a higher price for its predictions for the Democratic Republic of the Congo and Togo post-2001. This prediction behavior can be contrasted with the predictions of the logistic regression model, which tracks the baseline of regime failures, at roughly 0.05, quite closely through the training and test periods. The variability in the predictions of the support vector machine is greater than the logistic regression, but its test predictions are considerably more conservative on average than those of the random forest. These patterns mesh with the patterns presented in the distribution of predicted probabilities by classification algorithm (Figure 3.1) as well as the predicted probability plots directly above.

Because there is important variation in the predictive behavior of the algorithms, it may well be the case that weighting the contributions of each algorithm may produce better predictions than any single model would otherwise produce. To this end, we turn to ensemble Bayesian model averaging. Table 3.2 displays the performance results in the calibration period 2001–2005 and Table 3.3 shows the results for the test period 2006–2010. The results diverge in important ways. To start with the calibration period, the loss metrics indeed show that the EBMA model outperforms any of the constituent models, and this is true of both the restricted and full specifications. The area under the curve metrics show no improvement of the EBMA over the best performing classifier, which in this case is the logistic regression model. Importantly, regarding both the loss and AUC metrics, the full specification is found to outperform the restricted specification.
Figure 3.4: Predicted probabilities and observed authoritarian regime failures for four countries across the training and test periods. Observed failure events are marked by solid vertical lines, and the test period is shaded gray.
The performance on the test data are also different in important ways. In this period, the EBMA is again able to reduce the test loss, and again this is true of both the restricted and full specifications. The AUC metrics are less conclusive, however. The highest AUC for the ROC curve is produced by the random forest classifier using the full specification, but the highest AUC for the precision-recall curve is produced by the EBMA and logistic regression model using the restricted specification. It is also notable that the performance is much better in the test period than in the calibration period. This is likely due in part to the lower observed rate of failure from 2006 to 2010, but it could also be the case that the data generating process in this period tracks that of the training data more closely than does the period 2001 to 2005. In the end, the importance of including the lags in the specification is still somewhat of an open question for predicting instances of regime failure. As new data are obtained, additional experiments are needed to shed more light on the predictive power of spatiotemporal lags.

### 3.4.2 Democratic transition

The performance for each of the classifiers in predicting democratic transitions for the test period is shown in Table 3.4. The classifier with the lowest test loss metrics in this case is the random forest classifier with the restricted specification. In terms of the area
<table>
<thead>
<tr>
<th>Classifier</th>
<th>Specification</th>
<th>Log Loss</th>
<th>Brier Loss</th>
<th>AUC ROC</th>
<th>AUC PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBMA</td>
<td>Restricted</td>
<td>0.1011</td>
<td>0.0249</td>
<td>0.9344</td>
<td>0.3947</td>
</tr>
<tr>
<td>Penalized Logit</td>
<td>Restricted</td>
<td>0.1166</td>
<td>0.0260</td>
<td>0.9344</td>
<td>0.3947</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Restricted</td>
<td>0.1089</td>
<td>0.0250</td>
<td>0.9239</td>
<td>0.2810</td>
</tr>
<tr>
<td>SVM</td>
<td>Restricted</td>
<td>0.1162</td>
<td>0.0256</td>
<td>0.8499</td>
<td>0.1301</td>
</tr>
<tr>
<td>EBMA</td>
<td>Full</td>
<td>0.1024</td>
<td>0.0250</td>
<td>0.9124</td>
<td>0.3126</td>
</tr>
<tr>
<td>Penalized Logit</td>
<td>Full</td>
<td>0.1202</td>
<td>0.0262</td>
<td>0.9124</td>
<td>0.3126</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Full</td>
<td>0.1158</td>
<td>0.0255</td>
<td>0.9423</td>
<td>0.2941</td>
</tr>
<tr>
<td>SVM</td>
<td>Full</td>
<td>0.1216</td>
<td>0.0260</td>
<td>0.8442</td>
<td>0.3340</td>
</tr>
</tbody>
</table>

Table 3.3: Prediction performance for models of authoritarian regime failure for the ensemble BMA and component models for the test period.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Specification</th>
<th>Log Loss</th>
<th>Brier Loss</th>
<th>AUC ROC</th>
<th>AUC PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penalized Logit</td>
<td>Full</td>
<td>0.0947</td>
<td>0.0196</td>
<td>0.8241</td>
<td>0.0929</td>
</tr>
<tr>
<td>Penalized Logit</td>
<td>Restricted</td>
<td>0.0945</td>
<td>0.0195</td>
<td>0.8242</td>
<td>0.0918</td>
</tr>
<tr>
<td>SVM</td>
<td>Full</td>
<td>0.0915</td>
<td>0.0196</td>
<td>0.7751</td>
<td>0.0587</td>
</tr>
<tr>
<td>SVM</td>
<td>Restricted</td>
<td>0.0936</td>
<td>0.0198</td>
<td>0.7645</td>
<td>0.0508</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Full</td>
<td>0.0959</td>
<td>0.0200</td>
<td>0.7181</td>
<td>0.0404</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Restricted</td>
<td>0.0894</td>
<td>0.0191</td>
<td>0.7221</td>
<td>0.0717</td>
</tr>
</tbody>
</table>

Table 3.4: Prediction performance for models democratic transition across classifiers for the test period.

under the curve metrics, the logistic regression model performs best, and performance is comparable in terms of ROC, but the full specification scores slightly higher in the tradeoff between precision and recall. The support vector machine does not outperform the other classifiers in terms of any metric, but the full specification outperforms the restricted specification for each metric for this model.

In looking at the distribution of predicted probabilities from each of these classifiers and specifications in Figure 3.5, we see that the full specifications produce slightly higher predicted probabilities on average than the restricted specifications, indicating that the spatiotemporal lags are helping to explain the observed democratic transitions in the training data. In addition to the highest variability, the random forest again produces
event probabilities that are greater on average than those of the other classifiers, though
the support vector machine is responsible for the boldest predictions of democratic transi-
tions. The predictions from the logistic regression model demonstrate very little variance
relative to the more flexible methods, and its distributions are tightly centered around
the baseline rate of democratic transitions in the data.

While the results in Table 3.4 provide no decisive evidence that the spatiotempo-
ral lags improve predictive performance over the restricted specification, it is clear from
Figure 3.5 that the lags are contributing importantly to the associations learned by the
algorithms through validation on the training data. We can again explore these associa-
tions in predicted probability plots. To begin, in Figure 3.6, the result from Section 2.2.7
indicating a negative association between the lag of autocratic transitions and the obser-
vation of democratic transitions is reiterated (see Figures 2.12, 2.13, and 2.14). As was
the case in predicting the broader category of regime failures, the associations suggested
here by the logistic regression and support vector machine classifiers are relatively muted, whereas the random forest classifier indicates a much sharper drop in the probability of a democratic transition in the areas of greatest density for the lags of autocratic transition. These findings are evidence in favor of the respective arguments offered in, for example, Beissinger (2007), Weyland (2010), and Meirowitz and Tucker (2013) that prevailing negative examples of regime failure either contribute to adaptation on the part of regimes to prevent diffusion of a regime failure or otherwise dissuade regime opposition from attempting to undertake a challenge to the regime in the first place.

The positive association between lagged democratic transitions and observed democratic transitions is likewise reiterated in this set of models. In Figure 3.7, the probability of a democratic transition is found to be increasing in the values of the lag of democratic transitions weighted by both geography and trade for the logistic regression and support vector machine. As in Figures 2.12 and 2.13, when we compare the results of the logistic regression and support vector machine in panels (a) and (b) of Figure 3.7, trade appears to be a slightly more important mechanism of diffusion in these models in the sense that the probability of a democratic transition increases more apparently than it does for geographic proximity. In this way, we might suggest that bilateral trade acts a relatively important mechanism of political instability, or at least of successful democratic transitions, among authoritarian regimes and that we should observe authoritarian regimes transition with higher likelihood when their authoritarian trading partners do so in advance.

It is worth noting that the random forest classifier again produces the most complex function of each of the lags in Figure 3.7, with the probability of a democratic transition decreasing to a certain point, then shooting higher at higher values of the lag, suggesting a U-shaped relationship. This finding is interesting in the sense that it might suggest interesting dynamics in the two-player game between regimes and regime opposition within authoritarian regimes in response to trigger events. For example, is the opposition hesitant in the absence of overwhelming modular examples of democratic
Figure 3.6: Estimated change in the predicted probability of democratic transition across values of (a) the trade weighted spatiotemporal lag of autocratic transitions and (b) the geographic weighted spatiotemporal lag of autocratic transitions.
Figure 3.7: Estimated change in the predicted probability of democratic transition across values of (a) the trade weighted spatiotemporal lag of democratic transitions and (b) the geographic weighted spatiotemporal lag of democratic transitions.
transition? Does the ability of the regime to suppress emulation fare better when there is not overwhelming pressure from several lagged democratic transitions, then diminish when the pressure increases? One might rightly argue that the comparative performance of the full and restricted specifications in Table 3.4 indicates that the lags offer little help in out-of-sample prediction, and thus the findings suggested in Figure 3.7 should not be interpreted too seriously. Nonetheless, it is important to remember that these findings were validated on the training data, and so they have been subjected to a tougher test than simple in-sample prediction. As more data are obtained moving forward, repeated experiments with new test observations may provide support for this finding, or otherwise the extended training period may temper the result.

As mentioned in Chapter 2, an open question to this point is whether regimes of a certain type are more resistant to diffusion of political instability than others. In Figure 3.8, the probability of democratic transition is plotted over the lag of democratic transitions weighted by bilateral trade and binary regime type classifications in GWF. Among these plots, military regimes clearly stick out, primarily because the base rate of democratic transitions among military regimes is higher. Namely, for most observations, the predicted probability given the classification of a military regime is 0.07 and 0.09, while for other regimes the probability is located between 0.02 and 0.04. This finding is likewise evident in Figures 2.12 and 2.13, where the coefficient on the military indicator is quite large in magnitude, both in absolute terms and relative to other regime indicators. In addition, however, the increased probability of a democratic transition in the value of the lag is sharper for military regimes, especially when looking at the findings from the random forest classifier, indicating that military regimes are potentially more vulnerable to trigger events.

Beyond the potential modifying role of regime types, other interactions have been explored as well, including combinations of indicators of wealth and economic growth. No substantive change in the shape of the relationships of the lags were found across wealthy-high growth, wealthy-low growth, poor-high growth, and poor-low growth sce-
narios, where wealth accounts for both GDP per capita and natural resource income per capita. These findings are omitted from the text but available upon request.

In Table 3.5, the results from the ensemble Bayesian model average of the three individual classifiers fit on the training data is provided on the calibration data from 2001–2005. In these results, the EBMA performs favorably relative to the other classifiers in terms of loss, but does only as well as the best performing model, the logistic regression, in terms of the area under the curve metrics. Notably, however, the full specification performs better on the calibration data than the restricted specification, evidence that the models fit with the spatiotemporal lags on the training data in fact generalize to some extent beyond those data.

However, the comparative performance of the calibrated EBMA and the component classifiers in Table 3.6 reveals no clear dominant model. The random forest classifier with the full specification performs best in terms of log loss, the support vector machine fit with the restricted specification performs best in terms of Brier loss, and the logistic
<table>
<thead>
<tr>
<th>Classifier</th>
<th>Specification</th>
<th>Log Loss</th>
<th>Brier Loss</th>
<th>AUC ROC</th>
<th>AUC PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBMA</td>
<td>Restricted</td>
<td>0.0928</td>
<td>0.0192</td>
<td>0.7096</td>
<td>0.0413</td>
</tr>
<tr>
<td>Penalized Logit</td>
<td>Restricted</td>
<td>0.0947</td>
<td>0.0192</td>
<td>0.7096</td>
<td>0.0413</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Restricted</td>
<td>0.1042</td>
<td>0.0206</td>
<td>0.6444</td>
<td>0.0267</td>
</tr>
<tr>
<td>SVM</td>
<td>Restricted</td>
<td>0.0968</td>
<td>0.0197</td>
<td>0.6555</td>
<td>0.0309</td>
</tr>
<tr>
<td>EBMA</td>
<td>Full</td>
<td>0.0909</td>
<td>0.0192</td>
<td>0.7547</td>
<td>0.0500</td>
</tr>
<tr>
<td>Penalized Logit</td>
<td>Full</td>
<td>0.0944</td>
<td>0.0192</td>
<td>0.7547</td>
<td>0.0500</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Full</td>
<td>0.1007</td>
<td>0.0202</td>
<td>0.6644</td>
<td>0.0287</td>
</tr>
<tr>
<td>SVM</td>
<td>Full</td>
<td>0.1063</td>
<td>0.0210</td>
<td>0.6243</td>
<td>0.0243</td>
</tr>
</tbody>
</table>

Table 3.5: Prediction performance for models of democratic transition for the ensemble BMA and component models for the calibration period.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Specification</th>
<th>Log Loss</th>
<th>Brier Loss</th>
<th>AUC ROC</th>
<th>AUC PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBMA</td>
<td>Restricted</td>
<td>0.0836</td>
<td>0.0192</td>
<td>0.9295</td>
<td>0.3244</td>
</tr>
<tr>
<td>Penalized Logit</td>
<td>Restricted</td>
<td>0.0943</td>
<td>0.0199</td>
<td>0.9295</td>
<td>0.3244</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Restricted</td>
<td>0.0825</td>
<td>0.0191</td>
<td>0.8817</td>
<td>0.2320</td>
</tr>
<tr>
<td>SVM</td>
<td>Restricted</td>
<td>0.0818</td>
<td>0.0185</td>
<td>0.7861</td>
<td>0.2046</td>
</tr>
<tr>
<td>EBMA</td>
<td>Full</td>
<td>0.0826</td>
<td>0.0190</td>
<td>0.8969</td>
<td>0.2411</td>
</tr>
<tr>
<td>Penalized Logit</td>
<td>Full</td>
<td>0.0951</td>
<td>0.0199</td>
<td>0.8969</td>
<td>0.2411</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Full</td>
<td>0.0817</td>
<td>0.0190</td>
<td>0.8800</td>
<td>0.2133</td>
</tr>
<tr>
<td>SVM</td>
<td>Full</td>
<td>0.0851</td>
<td>0.0189</td>
<td>0.8071</td>
<td>0.1780</td>
</tr>
</tbody>
</table>

Table 3.6: Prediction performance for models of democratic transition for the ensemble BMA and component models for the test period.

regression and EBMA models perform best in terms of the area under the curve metrics. As a result, it is again unclear if the spatiotemporal lags are contributing meaningfully to the predictive capability of the models, and predictive experiments on future observations are likely required to shed further light on the question.

### 3.5 Remarks

The out-of-sample performance in this chapter has not conclusively offered evidence in support of including spatiotemporal lags for the purposes of prediction, but neither has
it rejected their specification completely. Namely, the full specification outperforms the restricted specification for a subset of the performance metrics and for a subset of the test periods under examination. Equally important, however, is that the findings first uncovered in the linear model in Chapter 2 are upheld in the training data in the present chapter using a wider range of classification algorithms and appropriate model validation techniques. For example, the lag of democratic transitions is found to be positively associated with both observed regime failures and democratic transitions, while the lag of autocratic transitions is found to be negatively associated with observed democratic transitions. The key question that remains is whether the underwhelming out-of-sample performance is due to a fundamental change in the mechanisms that lead to authoritarian regime failure post-2000 or to overfitting the training data as a function of including the spatiotemporal lags in the specification.

In addition, this chapter has generally shown that a simpler functional form produces better out-of-sample than more flexible methods. Even a post-hoc search for interactions between key terms in the model proved relatively fruitless in yielding insight into the conditions most associated with regime failure and democratic transition. In part, this is likely due to the fairly small number of observations in the data, and the performance of the more flexible learning algorithms is likely to improve as they pick up on meaningful interactions in a larger training sample.

That said, there is ample room for improvement on the methods outlined above. One obvious improvement would entail exploring a wider range of alternative specifications, including the dyadic weights used to construct the connectivity matrices, the lagged events chosen for specification, and the time and unit effects elected in this study. It might be better, for example, to use year indicators instead two-year period effects, country instead of region indicators, or perhaps it is better to omit these indicators altogether. In addition, as noted by Franzese and Hays (2008, 774), if the connectivity matrices are misspecified, and we are not accurately capturing the channels of diffusion, then the effect of diffusion is understated. In this light, any non-finding in the limited number of specifications
expressed here should not be interpreted as definitive evidence against diffusion. Rather, it could simply be the case that a more robust set of specifications of these connections is required to elicit the mechanism through which states in the international system learn from each other.

The extreme bounds approach of Miller, Joseph, and Ohl (2016) offers a potential way forward in this discussion. In that work, the authors cycle through all combinations of an array of possible specifications, including connectivity matrices and fixed effects. While their emphasis is on in-sample explanation, fitting a larger number of models with a diverse set of specifications might offer predictive gains through the ensemble Bayesian model averaging approach outlined above. The major limiting factor is computational resources, but with sufficient resources, it is feasible to produce predictive models with a number of different algorithms and variable sets of features, allowing them to pick up on different aspects of the data generating process, then combining those different perspectives with EBMA on calibration data. Moreover, this approach is more computationally feasible than a test for all possible restrictions for the lagged variables outlined in this study as would be required to understand which among them are important for improving the predictive performance for models of regime failure and democratic transition.
Chapter 4

Coups

To this point, the relevance of diffusion as a factor in explaining and predicting instances of authoritarian regime failure and democratic transition has been explored, and some evidence has been presented that diffusion of such instability is present through both geographic and economic linkages. In this section, we move beyond regime failures to another source of manifestation of regime instability that has received attention in the scholarly literature on diffusion, coups d’état. Coups as an outcome is an appropriate complement to the previous study of regime failures. Particularly, as pointed out in Miller, Joseph, and Ohl (2016), regime failures and democratic transitions are often “mass-based” events while coups are often orchestrated by a small group of military elites, suggesting that the pathways for transnational learning and updating expectations about the success of emulation may be distinct. In addition, while the literature on regime transitions demonstrates fairly well-defined expectations regarding diffusion, the majority of theories regarding coups treat domestic attributes exclusively (e.g., Powell, 2012). Systematic analyses of the diffusion of coups, commonly referred to as coup “contagion”, are fewer in number, especially in recent years. Moreover, with the notable exception of the recent contribution by Miller, Joseph, and Ohl (2016), these analyses have not leveraged the modern cross-national datasets on coups available to researchers today and, additionally, have not explored alternative pathways through which diffusion is posited to occur, such as those outlined in Secion 1.4. Following the procedures outlined in the preceding sections, this section builds on the work of Miller, Joseph, and Ohl in presenting a comprehensive analysis of diffusion with respect to coups.

As the label *contagion* might indicate, the bulk of theories of diffusion in the case of coups consider how coups cluster spatiotemporally. The crux of the contagion theory,
as argued by Li and Thompson (1975) and Lutz (1989), explains that military officers with a desire to oust the incumbent regime update their expectations concerning the success of a coup attempt in response to viewing a successful model abroad. Central to this idea is that the triggering coup event is interpreted by would-be emulators as relevant to their own situation, implying a degree of structural connectivity between the countries. Notably, “Contagious recipiency... is determined almost wholly by the degree of readiness of the recipient to be influenced. The initiator of contagion need not establish the goals or the techniques of successive coups. It suffices merely to reinforce the aspiring coup-maker’s desire to achieve his own ends” (Li and Thompson, 1975, 81).

As an example, Saine (2008) explains that the 1994 coup in The Gambia was partly motivated by the earlier coup in Sierra Leone led by Valentine Strasser. Saine argues that the widespread instability and rent-seeking opportunities in weakly institutionalized countries generate a situation in which elements of the military can draw on successful models of intervention to their own advantage. This argument is consistent with prevailing theories of military intervention in politics, thought to be a consequence of the disposition on the part of the security apparatus for ousting the incumbent and the opportunity to do so (e.g., Finer, 2002; Powell, 2012). The likelihood of receipt by an observing group is a function of a prevailing disposition for intervention, and opportunity is influenced by a shift in expectations on the part of this group.

There are two important assumptions underlying the above. In the first place, anti-regime actors are assumed to update their expectations based on the observation of successful coups. However, as with the case of regime failures more broadly, military officers with a prevailing disposition for regime change are perhaps just as likely to learn from unsuccessful coups as they are successful coups. That is, the learning that takes place on the part of regime opposition should either increase or decrease the probability of attempting a coup, and this direction depends crucially on whether the trigger events were successful or unsuccessful.

Second, the direction of the diffusion is positive, suggesting that coups clustering
in time and space are strategic complements. By this logic, the trigger event is relevant only to actors, in this case military officers, who oppose the incumbent regime and possess the desire and capacity to initiate a regime change via a coup. Notably, the potential for regimes to likewise observe and learn from trigger events is not considered, thus dismissing the possibility that a trigger event may serve as a strategic complement for subsequent events in other countries. Yet, as with regime learning in the case of authoritarian regime failures and democratic transitions, we expect incumbent regimes to take steps to preclude coup emulation. In the context of coup emulation, authoritarian regimes should focus on addressing the perceived threat posed by dissident actors within the military or security apparatus who are most likely to undertake an emulation.

How do authoritarian regimes typically insulate from the threat posed by security forces? That regimes manipulate the security forces, including their composition, organization, and leadership, is widely recognized in the cross-national and case study literature on coups. Decalo (1989), for example, outlines several commonalities among resilient authoritarian regimes, including placement of co-ethnics and family members in key security posts and the creation of paramilitary security organizations to counterbalance the conventional armed forces. In fact, Miller, Joseph, and Ohl (2016) use counterbalancing forces as an indication of coup-proofing in their analysis of negative coup diffusion. Notable cases of this type of coup-proofing include Zaire under Mobutu and Kenya under Kenyatta and Moi. In each case, the regime used elite and loyal security units as well as ethnic congruence in the security forces to guard against challenges from the military (Emizet, 2000; N'Diaye, 2002).

As these coup-proofing strategies speak to structural vulnerability to coups, they are largely invariant with respect to time within the duration of a regime. Once the regime has created parallel security organizations or placed loyal figures at the head of important security units, the expected baseline exposure to domestic threats from the military are lower and essentially constant throughout the course of the regime, all else equal. This is potentially problematic for two reasons. First, structural coup-proofing mechanisms are
perhaps just as likely to increase the disposition of the military for intervention. N’Diaye (2002), for example, notes that ethnic bias and politically-inspired hierarchies generated discontent in the Kenyan security forces and, accordingly, increased the risk of regime challenges by these organization. This suggests that political manipulation of the security forces does not generate a clean expectation with regard to regime stability, and moreover it does not offer a great deal of variance to be explained over time for regimes.

Second, and drawing on the notion that coup-proofing does not necessarily eliminate the threat of political intervention from the security forces, the probability of a coup attempt responds in part to changes in the opportunity structure. Here, the dynamic opportunity structure for coup attempts is a function of both domestic and international instability. This implies that regimes need short-term tools, even in the presence of structural coup-proofing measures, to respond to fluctuations in the threat level. In other words, in a dynamic setting where new information is revealed by way of triggering events and beliefs are subsequently updated by anti- and pro-regime actors, the latter have tools at their disposal beyond, say, the ethnicization or reorganization of the armed forces, which again only shifts the baseline susceptibility to intervention. Rather, regimes may turn to allegations of coup conspiracies, selective purges of the military, or even executions to remove or deter threats to survival. In Zimbabwe, for example, Chitiyo (2009, 12) notes:

Over the past five years there has been a significant exodus of lower- and middle-ranking officers from both the [Zimbabwe National Army] and the [Air Force of Zimbabwe]. Many of the personnel were unhappy with the increasing politicisation, low pay, indierent accommodation and the decreasing professionalism of the security sector. Nor was the disaffection confined to the lower ranks; in June 2007, 400 low- to mid-ranking ZNA personnel were arrested in connection with an alleged coup attempt. Many were court-martialled and there were numerous ‘disappearances’. There were also a number of mysterious deaths including those of three senior officers in the ZNA and the AFZ.
Ndlovu-Gatsheni (2006) also notes that the Mugabe regime has routinely linked political opponents with coup plots in Zimbabwe. Similarly, the Biya regime in Cameroon and the Jammeh regime in Gambia use purges from the security forces, among other repressive measures, in response to coup allegations (AfricaPresse.com, 2010; Smith and Noyes, 2015). While these examples do not speak directly to the connection between regime actions and an external triggering event, they do indicate that regimes respond to a dynamic environment with dynamic coup-proofing strategies, not simply aligning key security forces along ethnic lines and expanding the number of security forces as a means of counterbalancing potential coup-plotters. In this light, in response to perceived changes in the threat level, and irrespective of whether the source of the change is located in domestic or international instability, we can expect regimes to try to deter a coup attempt with preemptive action. As such, because regimes are more conscious of threats to survival in the aftermath of challenges to survival beyond their borders, they are likely to take steps to prevent the diffusion of the challenge, thus lowering the probability of emulation.

In sum, as was the case with regime failures and following the analyses in Miller, Joseph, and Ohl (2016), there are competing expectations to be examined for the diffusion of coups. First, because military officers with a preference for unseating the incumbent regime are likely to have their expectations of success reinforced by successful trigger events, we may expect coups clustered in time and space to be strategic complements (positive diffusion). On the other hand, because pro-regime actors likewise learn from trigger events, a trigger coup in one regime may serve as a strategic substitute for coups in other regimes by undermining the capacity of anti-regime actors in the security forces through dynamic coup-proofing measures (negative diffusion).
4.1 Dependent variable, spatiotemporal lags, & controls

The dependent variable for the analysis is a binary variable capturing attempted coup events, both successful and unsuccessful, and I use the Powell and Thyne (2011) data to capture all coups from 1950. The analysis is limited to attempted coups, as opposed to a separate model for successful coups, since the primary interest is to uncover the set of attributes and contextual circumstances that lead military officers to undertake a coup, while the set of attributes that determine the success of that attempt might be slightly different.\footnote{Because the unit of analysis is the country-year, a small number of coup events are omitted from the data because some countries experience more than one event in a given year. In these cases, the first coup occurring in the calendar is adopted as the operative event for that unit of observation. Again, following the previous analysis as well as the analysis in (Miller, Joseph, and Ohl, 2016), the date used for observations for which there is no observed coup event is the final day of the calendar year. The timeline of observed events in the data are displayed in Figure 4.1.}

In turn, just as above, the spatiotemporal lags evaluated for observations are constructed with a 365-day window set up from the date of the event. In this section, the lagged events of interest are the number of coups attempted (i.e., aggregate successful and failed coups, as in Figure 4.1 and, additionally, successful coups. As argued above, the total number of attempted coups is likely to send a signal to incumbent regimes, who are likely to adjust their expectations about threats from the security forces, and they are likely to learn from both successes and failures. The lag of successful coups is expected to send a separate signal to officers who wish to initiate a coup against the incumbent regime.

\footnote{Of course, the same may be true of authoritarian regime failures, but democratic transition as a subject of study is much more prominent that successful coups in comparative politics historically. There is also no guarantee that coups set in motion a transition to democracy, though there is some evidence that this may be so (Thyne and Powell, 2014).}
The remaining features entering the statistical models are equivalent to those used in the models of regime failure above, including authoritarian regime attributes, economic indicators, and two-year and region fixed effects. As pointed out by Plümper and Neumayer (2010) and echoed in Miller, Joseph, and Ohl (2016), the period indicators help to separate the influence of spatiatemporal terms from shocks felt commonly by all observations, and the use of two-year periods again helps deal with the issue of separation in the model. It is also worth mentioning that instead of modeling duration as a function of the number of years since the last coup attempt, the natural log of regime duration from GWF is again used. This choice offers several advantages. First, successful coups are nested within regime turnovers in the GWF data. Second, regime turnovers not associated with a successful coup are still likely to affect temporal dependence in the observation of coups. Namely, a change in regime, entailing a shift in ruling coalition or the political institutions that shape policy-making, is likely to alter the capacity or willingness of actors within the military or security forces to attempt a coup. Third, because the Powell and Thyne data only date to 1950, a duration feature derived from
these data alone may underestimate the stability of long-standing regimes with respect to military intervention in politics.

4.2 Linear model of coups

The melded results for the logistic regression model with elastic net penalization are shown in Figures 4.2 and 4.3. Looking first at the domestic attributes in the model, we see several findings that are consistent with expectation. First, military regimes are found to be the most vulnerable regime type with regard to coups, a finding that is regularly highlighted in the coup literature (Belkin and Schofer, 2003; Powell, 2012). In contrast, single-party regimes appear the most resistant to coups, which matches the regime failure models in the previous sections. In addition to regime type classifications, regimes with higher levels of institutionalization are found to be negatively associated with coups. This finding is consistent with arguments that coups are a mechanism for political change when political outlets are largely constrained (Thyne and Powell, 2014). That is, in the absence of institutions providing a means for opposition actors to be engaged in the political process, coups may emerge as a more viable option for unseating an authoritarian regime.

We further see preliminary evidence that key economic attributes for regimes are negatively associated with coups. For example, GDP per capita and economic growth are both found to reduce the likelihood of a coup, which is consistent with previous work (e.g., Londregan and Poole, 1990; Powell, 2012). The model also finds that natural resource income is negatively associated with coups, which may offer support for the argument that resource income increases state capacity through an increase in spending on internal security (Ross, 2001).

As for the spatiotemporal lags, the results show the geographic and trade weighted lags of coup attempts and successful coups demonstrate no systematic association with coups. These findings match the results in Miller, Joseph, and Ohl (2016), who find
Figure 4.2: Bootstrap confidence intervals for the parameter estimates in the model of coups in authoritarian regimes. Estimates for regional and two-year period indicators are omitted for clarity.
Figure 4.3: Permutation significance for the parameter estimates in the model of coups in authoritarian regimes. Estimates for regional and two-year period indicators are omitted for clarity.
no evidence of coup “contagion”. The linguistic weighted lags of these outcomes, however, are both found to be positively associated with coups. That successful coups are associated with higher subsequent coup risk falls in line with the arguments above, but the association between all coups, whether successful or unsuccessful, with coup risk is a surprising result and offers evidence that regimes do not learn from trigger events, at least when they are transmitted through the linguistic mechanism. In any case, these results support the coup contagion hypothesis, though the subset of target states for a given trigger coup seems to be defined by linguistic commonality.

To better understand the substantive importance of the associations suggested in the model, the predicted probability of a coup attempt over the range of a subset of the features is shown in Figure 4.4. As expected from the estimates, the probability of a coup attempt is increasing in the values of the linguistic lags of all attempted coups and successful coups. Specifically, in panel (a), the increase in the probability of a coup attempt is approximately 0.0047 (from 0.0450 to 0.0497) when the linguistic lag of all coup attempts increases from its 25th to its 75th percentile. In panel (b), when the linguistic lag of successes increases from its 25th to its 75th percentile, the increase in probability is approximately 0.0027 (from 0.0459 to 0.0486).

While these increases are seemingly small, we can see that they are comparable in magnitude to variables routinely associated with coup attempts in the academic literature. For example, in panel (c), the drop in the probability of a coup attempt when GDP per capita moves from its 25th to its 75th percentile is approximately 0.0034 (from 0.0490 to 0.0456). And the drop in probability given a drop in natural resource income per capita, panel (d), is approximately 0.0030 (from 0.0489 to 0.0459).

Thus, given that we are dealing with fairly small nudges in the probability of a coup attempt over the most important features in the model, the linguistic lag carries relatively important information about the likelihood of a coup attempt. Yet, it is surprising that the lag of coups that aggregates both failures and successes is substantively more important than the lag of successes alone and indicative of positive diffusion. No evidence
Figure 4.4: Predicted probability of coup attempts over select variables: (a) the linguistic weighted lag of coup attempts; (b) the linguistic weighted lag of successful coups; (c) log GDP per capita; and (d) log resource income per capita. Probability estimates are derived from estimates shown in Figures 4.2 and 4.3.
is found in the model for negative diffusion that would indicate learning on the part of regimes.

Finally, we can see in Table 4.1 the values of the elastic net hyperparameters chosen in cross-validation across the ten imputed datasets. As with the models of regime failure, non-negligible amounts of regularization, controlled by the parameter $\lambda$, are found to be perform best in validation for each of the imputations. In this way, logistic regression without any regularization is not the preferred model. In addition, also similar to the models of regime failure, the regularization is tilted in favor of the $L_2$ penalty, although the $L_1$ ratio, represented by the parameter $\alpha$ in Table 4.1, is greater than zero for two of the imputations, implying that at least some of the estimates in these models are driven to zero. In the end, these results suggest that the ridge penalty is the most useful for learning a generalizable solution in the linear model of coups.

### Table 4.1: Elastic net parameters chosen in cross-validation for each imputation in the model of attempted coups.

<table>
<thead>
<tr>
<th>Imputation</th>
<th>$\lambda$</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2512</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>0.2512</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>0.2512</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>0.2512</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>0.1259</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>0.1995</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>0.1259</td>
<td>0.05</td>
</tr>
<tr>
<td>8</td>
<td>0.1585</td>
<td>0.00</td>
</tr>
<tr>
<td>9</td>
<td>0.1585</td>
<td>0.00</td>
</tr>
<tr>
<td>10</td>
<td>0.1585</td>
<td>0.05</td>
</tr>
</tbody>
</table>

4.3 Predictive models of coups

Following the same procedure as in Chapter 3, in this section the predictive performance of more flexible learning algorithms is compared to the logistic regression model for the test period 2001–2010, which are displayed in Table 4.2. The penalized logistic regression
Table 4.2: Prediction performance for models coup attempts across classifiers for the test period.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Specification</th>
<th>Log Loss</th>
<th>Brier Loss</th>
<th>AUC ROC</th>
<th>AUC PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penalized Logit</td>
<td>Full</td>
<td>0.0861</td>
<td>0.0177</td>
<td>0.8613</td>
<td>0.0965</td>
</tr>
<tr>
<td>Penalized Logit</td>
<td>Restricted</td>
<td>0.0891</td>
<td>0.0178</td>
<td>0.8433</td>
<td>0.0945</td>
</tr>
<tr>
<td>SVM</td>
<td>Full</td>
<td>0.0918</td>
<td>0.0180</td>
<td>0.8844</td>
<td>0.1158</td>
</tr>
<tr>
<td>SVM</td>
<td>Restricted</td>
<td>0.0932</td>
<td>0.0181</td>
<td>0.8791</td>
<td>0.1165</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Full</td>
<td>0.0988</td>
<td>0.0186</td>
<td>0.8388</td>
<td>0.1167</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Restricted</td>
<td>0.0928</td>
<td>0.0183</td>
<td>0.8377</td>
<td>0.0848</td>
</tr>
</tbody>
</table>

Model fit with the full specification performs best in terms of the loss metrics, which is not surprising. As in Sections 3.4.1 and 3.4.2, the predicted probabilities of an event in the logistic regression model tend toward the baseline rate of the event in question, which implies lower loss in a rare events problem. The distribution of predicted probabilities by classifier and specification in the test data is shown in Figure 4.5.

By contrast, the random forest and support vector machine classifiers, and specifically those trained with the full specification, perform better in terms of the area under the curve metrics. These models, therefore, are found to perform better with respect to the observed events in the test data. Moreover, that these models are fit using the full specification is evidence that the spatiotemporal lags of coup events offer some leverage in predicting coup attempts. Therefore, building on the effects suggested above in Figures 4.2, 4.3, and 4.4, these outcomes support the notion that coup events diffuse across authoritarian regimes.

In turn, we split the period 2001–2010 into separate calibration and test periods to gauge weight the contributions of the individual classifiers in an ensemble. The results on the calibration period are shown in Table 4.3. The EBMA model is able to make improvements on the loss metrics in the calibration period, but not for the area under the curve metrics. In addition, for this period, the restricted specification is found to perform marginally better, though only slightly so.

Moving to the test performance for years 2006–2010 in Table 4.4, the EBMA model
Figure 4.5: The distribution of predicted probabilities of coup attempts for each of the classifiers, comparing the restricted to the full specification.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Specification</th>
<th>Log Loss</th>
<th>Brier Loss</th>
<th>AUC ROC</th>
<th>AUC PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBMA</td>
<td>Restricted</td>
<td>0.0904</td>
<td>0.0237</td>
<td>0.8998</td>
<td>0.1884</td>
</tr>
<tr>
<td>Penalized Logit</td>
<td>Restricted</td>
<td>0.1107</td>
<td>0.0246</td>
<td>0.8561</td>
<td>0.1522</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Restricted</td>
<td>0.1103</td>
<td>0.0243</td>
<td>0.8998</td>
<td>0.1884</td>
</tr>
<tr>
<td>SVM</td>
<td>Restricted</td>
<td>0.1089</td>
<td>0.0245</td>
<td>0.8502</td>
<td>0.1320</td>
</tr>
<tr>
<td>EBMA</td>
<td>Full</td>
<td>0.0905</td>
<td>0.0235</td>
<td>0.8902</td>
<td>0.1863</td>
</tr>
<tr>
<td>Penalized Logit</td>
<td>Full</td>
<td>0.1091</td>
<td>0.0247</td>
<td>0.8691</td>
<td>0.1591</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Full</td>
<td>0.1088</td>
<td>0.0242</td>
<td>0.8986</td>
<td>0.1890</td>
</tr>
<tr>
<td>SVM</td>
<td>Full</td>
<td>0.1147</td>
<td>0.0246</td>
<td>0.8388</td>
<td>0.2079</td>
</tr>
</tbody>
</table>

Table 4.3: Prediction performance for models of coup attempts for the ensemble BMA and component models for the calibration period.
is able to reduce the log loss by combining the predictions of the individual classifiers, but the logistic regression model still performs better in terms of Brier loss. Second, the support vector machine produces the best performance in terms of the area of the curve metrics, with the full specification doing better for the ROC curve and the restricted specification maximizing area under the precision-recall curve.

Because the importance of the spatiotemporal lags with respect to predictive performance appears sensitive to the definition of the test period, we can look at the predicted probability plots learned by the models in the training data to get a better sense of the substantive effects for these variables. The predicted probabilities of coup attempts across values of the geographic and linguistic weighted lags of both attempted and successful coups are displayed in Figure 4.6, and the findings echo the results in Figures 4.2 and 4.3. Specifically, the models all suggest positive diffusion of attempted and successful through the channel of shared primary language. In addition, neither geographic proximity or trade ties as a transmitter of lagged coup events are found to have much of an effect on the probability of a coup attempt.
Figure 4.6: Predicted probability of coup attempts over the linguistic, geographic, and trade weighted lags of attempted and successful coups.
4.4 Remarks

The findings in the present chapter cast some doubt on the claim in Miller, Joseph, and Ohl (2016) that coups are not contagious. While the findings here support their conclusion that diffusion does not occur through economic ties or geographic proximity, evidence is presented that positive diffusion does in fact occur through a linguistic channel. Admittedly, improvements in predictive performance do not speak conclusively to such a claim, but the substantive finding of positive diffusion is robust across logistic regression, random forest, and support vector machine classifiers that have all been subjected to rigorous validation procedures. Moreover, in spite of the seemingly small effects presented for the spatiotemporal lags, in the sense that the predicted probability of a coup attempt changes only slightly in the value of the lags, the magnitude of these substantive effects are on par with domestic indicators, such as GDP per capita and natural resource income per capita, that are well established in the prevailing literature on coups.

That said, the models presented here are subject to the same criticisms as those in Chapter 3. Miller, Joseph, and Ohl (2016) are to be commended for raising the important question of uncertainty regarding proper model specification in studies of spatial effects, and the models here are not without their biases as a result of presenting only a small subset of a much larger population of potential specifications. In future work, more model specifications, including splitting the spatiotemporal lags of distinct events across different models, considering alternative connectivity matrices that speak to the mechanisms through which political events diffuse, and different indicators for time and unit effects, are needed to ensure that at least some of the models are picking up on the true mechanisms that underlie coup attempts. In turn, ensemble Bayesian model averaging offers great promise for meaningfully combining the unique contributions of each model and producing better predictions.
Chapter 5

Conclusion

The analysis in this dissertation is motivated by two primary concerns, as outlined in Chapter 1, one conceptual and one methodological. Conceptually, previous literature treating the diffusion of political events in time and space, particularly those events that speak to political instability, are constrained in the events that are thought to influence the probability of an event. In contrast, the analysis above extends the set of destabilizing events that are thought to influence the likelihood of observing instability in authoritarian regimes. Notably, we consider the spatiotemporal lags of total authoritarian regime failures, democratic transitions, autocratic transitions, and coerced failures in spatial models of authoritarian regime failure and democratic transition. In the spatial models of coup attempts, we consider the lags of the full set of coup attempts and further look specifically at the lag of successful coups. In each of these cases, disaggregating events such as regime failures and coups into subsets of more precise events allows for the exploration of new and specific hypotheses. Indeed, in the above, it allows for testing the effects of lagged variables that are suggested to point in opposite directions, such as the spatial lags of autocratic and democratic transitions.

As it turns out, parsing regime failures in the construction of spatiotemporal lags reveals statistically and substantively interesting results, particularly in the models of democratic transition. In Chapter 2, evidence is presented that lagged democratic transitions are positively associated with regime failures and with democratic transitions, while autocratic transitions are found to be negatively associated with democratic transitions. The substantive effects corresponding to these findings, while small, are comparable in magnitude to the effects corresponding to well-established indicators in the comparative politics literature on regime failure. The spatiotemporal lags of coerced failures, on the
other hand, do not appear to be associated with the two outcomes under investigation in Chapter 2, which may indicate that the outcomes of prevailing regime failures, as opposed to the means by which they occur, is more informative to actors who learn from those events in other authoritarian countries. In future work, these indicators could perhaps be dropped from the specification. In addition, the findings for the lags of the full set of regime failures offer less concrete evidence of diffusion. In this light, these lags could most likely be dropped from future models, especially since the disaggregated indicators (i.e., autocratic vs. democratic transitions) offer clearer directional hypotheses.

Additionally, in the descriptive model of coups, the lag of total coup attempts and the lag of successful coups are found to be positively associated with coup attempts. In contrast to prevailing findings that suggest no clustering of coups in time and space, the models in this analysis find positive diffusion through two lagged variables, at least in the subset of authoritarian countries under examination. Future work should focus on the robustness of linguistic linkage as a mechanism of diffusion, and further why this specific mechanism best informs the transmission of coups across state borders among military elites.

On the methodological side, the key concern in the prevailing analysis is generalizability, which is manifested in two ways. In the first place, validation plays a central role in model training. For the descriptive models of regime failure and democratic transition in Chapter 2 as well as for the descriptive model of coup attempts in Chapter 4, cross-validation is employed to choose elastic net hyperparameters and model parameter estimates that best generalize. For these models, validation is crucial to prevent overfitting as well as to constrain the estimates on correlated inputs, such as domestic economic indicators and the spatiotemporal lags. In Chapter 3, cross-validation is likewise used to validate the logistic regression and support vector machine classifier, while out-of-bag validation is used for the random forest classifier. These procedures ensure that the correct hyperparameters are chosen during model training to better generalize to new test data.
What is clear from these models is that model validation procedures that introduce regularization, whether for logistic regression or for flexible machine learning algorithms, still permit analysts to explore the statistical and substantive significance of their results. For example, although analytical standard errors are not provided for regularized linear models, simple numeric procedures such as the bootstrap and permutation significance provide straightforward estimates of parameter uncertainty with a negligible computational burden. Though presented graphically here, the results of these procedures could very well be presented in tabular form, similar to the regression summary output common to statistical software packages. With greater computational complexity, the substantive effects of machine learning models can likewise be explored in predicted probability plots, which are commonly employed in political science research for interpreting linear models of discrete outcomes.

Beyond model validation in the training phase, the second step to ensure generalizability is to subject trained models to new, unseen data. This is the key objective of the predictive models of regime failure and democratic transition in Chapter 3, and predictive models for coup attempts are also shown in Chapter 4. The results are variable regarding the importance of spatiotemporal lags to the prediction of events during the 2001–2010 period. Particularly, there is some evidence that spatiotemporal lags offer little in terms of predicting regime failures and democratic transitions during the test period, and there is some slightly stronger evidence that the lags help in predicting coups during the test period, but the evidence is not overwhelming in either case. As indicated in the text, the substantive importance of the lags is consistent through each of the outcomes under investigation, and we see this in regression summary figures and in the predicted probability plots. The limited generalizability of the models learned in the training period may be attributed to idiosyncrasies in the test period, in the sense that the dynamics underlying regime failure events and democratic transitions in the test period are distinct from those in the training period and because the events occur at lower rates. In the end, one of the key takeaway points from the analyses is that the spatiotemporal lags are
important for *explaining* the outcomes under investigation although they are perhaps less
important for *predicting* those outcomes in the single test period delineated in this study.

In sum, the conceptual advances and methodological procedures outlined above yield
novel results, including the negative association between lagged autocratic transitions
and democratic transitions and the positive diffusion of coups through linguistic linkages.
Further, the methods here provide a template for political scientists to construct and
interpret generalizable models as a relatively simple extension of models that are already
in widespread use. Appendix 5 provides a Python implementation of the workhorse
programs used in the dissertation, and anyone can use or extend these programs to
validate and present models with very little coding. Future work on generalizable models
of regime failure, democratic transition, and coups should build on these models and
methods to explore a greater number of potential specifications, both in terms of omitting
those spatiotemporal lags that offered little information in the models here and, following
the example of Miller, Joseph, and Ohl (2016), in terms of considering alternative spatial
weights and unit and time indicators.
Appendix

This section details the workhorse programs for training the statistical models in the dissertation. As outlined in the main text, two flavors of model validation are employed. For the logistic regression models with elastic net penalization and the support vector machine classifier, k-fold cross-validation is used. Out-of-bag validation is used for the random forest classifier. These programs are intended to provide functionality for model training, validation, and prediction for classification problems in very few lines of code.

**k-fold cross-validation**

The Python class defined in `KFoldsValidationPrediction` performs a grid search over all possible combinations of hyperparameters, which are supplied as a dictionary, for a classifier instance to which those hyperparameters correspond. For example, to use logistic regression with elastic net regularization, we first import the generic stochastic gradient descent classifier from scikit-learn, then instantiate it with the loss function for logistic regression in (3.1) and specify the elastic net penalty.

```python
from sklearn.linear_model import SGDClassifier
logit = SGDClassifier(loss='log', penalty='elasticnet')
```

Next, we set up a parameter grid in a dictionary, which will specify the overall degree of regularization, which is denoted by $\lambda$ in (2.4), and the $L_1$ ratio. These values are given by the arguments `alpha` and `l1_ratio`, respectively.

```python
params = {'alpha': np.logspace(-4, 1, num=51),
          'l1_ratio': np.linspace(0, 1, num=11)}
```

With an instance of the classifier and its tuning parameters, we can specify an instance of the `KFoldsValidationPrediction` class. To perform ten-fold cross-validation, set the argument `k` to take a value of 10. The training data, including features $X$ and target $y$, are passed to the `X_train` and `y_train` arguments.

```python
kfold = KFoldsValidationPrediction(model=logit, params=params,
                                    k=10, X_train=X, y_train=y)
```

The model is trained on each combination of parameters and the set of parameters that generates the best validation performance is stored as an attribute. This procedure is executed in the `evaluate_model` method, which allows the user to specify the validation metric. For example, to use the loss function for logistic regression, specify `log_loss`, which will be passed to the scikit-learn’s grid search functionality.

```python
kfold.evaluate_model(metric='log_loss')
```

A range of metrics are available for evaluating validation performance. For example, we could use area under the ROC curve.

```python
from sklearn.metrics import roc_auc_score
kfold.evaluate_model(metric=roc_auc_score)
```

Once the optimal parameters are learned, the model is fit and can be used for making predictions for the training samples or for test samples. These predictions are returned in
the form of predicted probabilities, which allow for a wider range of evaluation metrics, and are especially useful in the case of rare events such as authoritarian regime failure and coups.

```python
kfold.predict_train()
kfold.predict_test(X_test)
```

If we fit a logistic regression, we can also use the class to obtain bootstrap confidence intervals for the parameter estimates as well as null distributions for evaluating permutation significance.

```python
kfold.bootstrap_estimates()
kfold.permutation_estimates()
```

In the end, we can access the predicted probabilities, the bootstrap estimates, and the estimates for the null by accessing the relevant attributes on the class instance.

```python
kfold.probabilities
kfold.boot_estimates
kfold.perm_estimates
```

These results can be used to generate plots like those in the main text. The full class definition for KFoldsValidationPrediction is shown below.

```python
class KFoldsValidationPrediction(object):
    def __init__(self, model, params, k, X_train, y_train):
        '''k-fold cross-validation for prediction'''
        self.model = model
        self.params = params
        self.k = k
        self.X_train = X_train
        self.y_train = y_train

    def evaluate_model(self, metric, random_search=False):
        '''grid search to find optimal parameters using given metric'''
        if random_search:
            combos = [x for x in apply(product, self.params.values())]
            iters = int(len(combos) * 0.5)
            search = RandomizedSearchCV(self.model, self.params,
                                         n_iter=iters, scoring=metric, n_jobs=-1, cv=self.k)
        else:
            search = GridSearchCV(self.model, self.params,
                                   scoring=metric, n_jobs=-1, cv=self.k)
            search.fit(self.X_train, self.y_train)
            self.optimal_params = search.best_params_
            self.cv_score = search.grid_scores_[1][1]

    def predict_test(self, X_test):
        '''predicted probabilities for given test data'''
        self.model.set_params(**self.optimal_params)
        self.model.fit(self.X_train, self.y_train)
        self.probas = self.model.predict_proba(X_test)[:, 1]

    def predict_train(self):
        '''predicted probabilities for the training data'''
```

103
self.model.set_params(**self.optimal_params)
folds = StratifiedKFold(self.y_train, n_folds=self.k)
test_indices, test_data, test_probs = [], [], []
for train_idx, test_idx in folds:
    Xtr, Xte = self.X_train[train_idx], self.X_train[test_idx]
    ytr, yte = self.y_train[train_idx], self.y_train[test_idx]
    train = self.model.fit(Xtr, ytr)
    test_probs.append(train.predict_proba(Xte)[:, 1])
    test_indices.append(test_idx)
    test_data.append(yte)
    test_indices = np.argsort(np.concatenate(test_indices))
    self.y_test = np.concatenate(test_data)[test_indices]
    self.probabilities = np.concatenate(test_probs)[test_indices]

def bootstrap_estimates(self, n_boot=100):
    '''bootstrap confidence intervals for linear model'''
    self.model.set_params(**self.optimal_params)
est = [np.hstack([self.model.fit(iX, iy).coef_.ravel(),
                     self.model.fit(iX, iy).intercept_])
          for iX, iy in (resample(self.X_train, self.y_train)
                         for _ in xrange(n_boot))]
    self.boot_estimates = np.vstack(est)

def permutation_estimates(self, n_perm=100):
    '''permutation significance for linear model'''
    self.model.set_params(**self.optimal_params)
yc = self.y_train.copy()
est = []
    for i in xrange(n_perm):
        np.random.shuffle(yc)
        self.model.fit(self.X_train, yc)
est.append(np.hstack([self.model.coef_.ravel(),
                         self.model.intercept_]))
    self.perm_estimates = np.vstack(est)

**Out-of-bag validation**

To reiterate the argument in the main text, cross-validation with random forests are (a) computationally expensive and (b) unnecessary, since the bootstrapping component to tree fitting leaves us with validation data. Taking advantage of these out-of-bag samples leads to much faster validation and the results are known to be similar to cross-validation.

The Python programs for conducting out-of-bag validation are split across two classes. The first, `OOBPerformance`, evaluates the predictive performance on the validation data (i.e., the OOB samples) for each tree in the ensemble and aggregates the performance across trees. The key contribution in this program is the ability for the user to specify the validation metric. For classification problems, the OOB validation functionality in scikit-learn is limited in that it only permits the use of accuracy as the validation metric. Further, the program offers a method for evaluating permutation importance, one of the variable importance measures available for random forest and other tree ensemble models, functionality which is also not available in scikit-learn.
The second, `OOBValidation`, calls `OOBPerformance` and executes its functionality in a grid search of the user-provided hyperparameter options, selecting the hyperparameter combination that yields the best validation performance. In this way, the user only interacts with an instance of `OOBValidation`. To demonstrate, we can first set up a random forest classifier with 1,000 tree estimators along with the parameter grid we want to search over.

```python
generate code here
```

Having the model instance and parameter grid, we can set up an instance of `OOBValidation` and execute the training with the desired validation metric.

```python
generate code here
```

The best performing model is selected in the fitting stage, so a trained random forest is now available for making predictions.

```python
generate code here
```

The full definitions for these classes are provided below.

```python
generate code here
```
def _oob_permute_predict(self, estimator_idx, metric):
    '''Evaluate change in metric randomly permuting each feature'''
    estimator = self.classifier.estimators_[estimator_idx]
    oob_samples_mask = self.oob_index[:, estimator_idx]
    oob_samples = self.X[oob_samples_mask]
    y_true = self.y[oob_samples_mask]
    oob_y_true = metric(y_true, estimator.predict_proba(oob_samples)[:, 1])
    error_baseline = metric(y_true, estimator.predict_proba(oob_samples)[:, 1])
    errors_perm = []
    for col in range(self.X.shape[1]):
        oob_samples_c = oob_samples.copy()
        np.random.shuffle(oob_samples_c[:, col])
        probs_perm = metric(y_true, estimator.predict_proba(oob_samples_c)[:, 1])
        errors_perm.append(probs_perm)
    return np.array(errors_perm) - error_baseline

def permutation_importance(self, metric):
    '''Average change in metric permuting features'''
    n_trees = len(self.classifier.estimators_)
    return np.vstack([self._oob_permute_predict(i, metric)
                      for i in range(n_trees)]).mean(axis=0)

class OOBValidation(object):
    '''Grid search using OOB validation for tree ensembles'''
    def __init__(self, model, param_dict):
        self.model = model
        self.params = param_dict
        self.param_grid = self.make_param_grid()

    def make_param_grid(self):
        '''Converts dict of hyperparams and values to grid'''
        combos = [x for x in apply(product, self.params.values())]
        return [dict(zip(self.params.keys(), p)) for p in combos]

    def fit(self, X, y, metric, minimize=True):
        '''Grid search over params, evaluate with given metric'''
        self.X = X
        self.y = y
        self.oob_scores = []
        for params in self.param_grid:
            self.model.set_params(**params)
            self.model.fit(X, y)
            oob = OOBPerformance(self.model, X, y)
            self.oob_scores.append(oob.oob_score(metric))
        best_idx = np.argmin(np.array(self.oob_scores)) if minimize
        else np.argmax(np.array(self.oob_scores))
        self.best_params = self.param_grid[best_idx]
        self.best_model = self.model.set_params(**self.best_params)
        self.best_model.fit(X, y)
        self.oob_score_ = self.oob_scores[best_idx]

    def predict_test(self, X):
        '''Predictions for test samples X'''
        return self.best_model.predict_proba(X)[:, 1]
def predict_train(self):
    '''Predictions for train samples using OOB predictions'''
    oob = OOBPerformance(self.best_model, self.X, self.y)
    oob.oob_score(score=False)
    return oob.probabilities
Bibliography


Thomas Brawner  
Curriculum vitae

Employment

- Postdoctoral Researcher, Institute for Quantitative Social Science, Harvard University (2016–)
- Research Statistician, ChildTrauma Academy (2015–)
- Senior Data Scientist, Galvanize, Inc. (2015–2016)

Education

- The Pennsylvania State University, Doctor of Philosophy, Political Science (expected May 2017)
- The Pennsylvania State University, Master of Arts, Political Science (2012)
- American University, School of International Service, Master of Arts (2010)
- Georgetown College, Bachelor of Arts *Summa cum laude*, Political Science (2005)

Research & Conference Participation


Technical Skills

- **Software**: Python, R, Stata, Spark, Bash, C++, PostgreSQL, MongoDB, tmux, Vim, Git[Hub], \LaTeX