An Empirical Evaluation of Theories in Government Formation and Duration

Lars Sutterud

Department of Political Science

Faculty of Social Sciences

University of Oslo

May 2015

An Empirical Evaluation of Theories in Government Formation and Government Duration

Lars Sutterud

May 18, 2015

2015

An Empirical Evaluation of Theories in Government Formation and Duration

Lars Sutterud

http://www.duo.uio.no/

Print: Reprosentralen, University of Oslo

Words: 22 120

Abstract

Theories of government formation and duration have been subjected to rigorous empirical testing. Numerous articles each argue for their own approach, pointing to an empirical model with significant explanation power. Being in a position to gather data not used in government formation and duration classics, this thesis has the benefit of testing their predictive power out-of-sample. The main question is: Can the models predict what they say they can explain when facing unseen observations? Another question at the forefront of explaining the life cycle of governments is the trade-offs regarding a cornerstone in political science methodology - decreasing or increasing complexity of the models which are tested. Recent literature on the life cycle of governments have pulled in the respective directions. This thesis sets out to empirically evaluate the trade-off between parsimony and complexity.

The empirical evaluation can be summed up in two main points: Only one out of the four original models predict new observations better than how they predict the original sample. Complex modeling of government formation yields more predictive power compared to the more parsimonious approach. Parsimonious modeling of government duration has more predictive power compared to complex modeling.

Acknowledgements

This thesis is a function of contributions from many great people.

Thank you, Bjørn Høyland, for research idea, wisdom and kindness. Thank you, Cristina Bucur, for thorough and important corrections along the way.

To all the people from the 9th floor, who have benefited from my coffee making skills, I am sincerely grateful to have met each and every one of you. To Martin, thank you for sharing your intellect and for pulling me through these last years at Blindern.

Thanks to Magnus, Malin and Peter for vital comments on language and content. Thank you, Haakon, for reading the final draft and for being an inspiration.

To my family, thank you for teaching me the value of persistence and endurance. Thank you also for unconditional support, without which this thesis would probably not have happened.

The rest is attributed to Kristina for bringing happiness to my life.

All flaws are mine alone.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

The life cycle of governments is a cornerstone in legislative studies. The making and breaking of governments in parliamentary democracies are opaque processes. The situation is the following; a change in government is demanded, whether on the basis of an election, a resigned incumbent government, a change in the parties represented in government, or a change in the prime minister post. As is often the case in parliamentary democracies, single party majority governments are rare (Gallagher et al. 2011, 413). The individual parties, consequently, must bargain with other parties to find a viable cabinet. This process often occurs in private, smoke-filled rooms. In the end the result is a new government. How can one explain the process of forming a government, going from the smoke-filled rooms to a viable cabinet? The unobservable features of the bargaining lay the ground for competing explanations of the government formation phase.

Immediately after the cabinet has formed, the termination clock starts ticking. Some governments last longer than others. Is the duration dependent on unpredictable critical events, such as personal scandals or financial crises? Or can attributes of the political system and the political actors involved explain government duration? This thesis seeks to evaluate the predictive power of theories in both the formation and the duration phase of the life cycle of governments in parliamentary democracies.

1.1 The Life Cycle of Governments

The life cycle of governments has been subjected to deep theoretical coverage, and the theories have subsequently been exposed to heavy empirical testing. Contributions range from formal models and case studies to cross-national, quantitative studies. Theories

have most often been directed towards explaining the formation and the termination of governments. My study will evaluate renowned studies from both fields. I will test their out-of-sample predictive power as a mean for evaluating how well the models perform. Earlier works have resorted to evaluate their models in terms of how well the model explains and fits the data, mostly by looking at statistical significance, direction and strength of the effects and residual statistics. Bäck and Dumont (2007) takes this a step further and calculates the in-sample predicted probability of a coalition formation. The problem is that the predictions stem from the same sample the model is fitted on. Taking empirical evaluations a step further, I will utilize the out-of-sample method to evaluate the predictive power of the models.

The end of the 1980s and the beginning of the 1990s brought on significant theoretical, methodological, and empirical advances in both fields. This guides the selection of articles evaluated in this thesis. I have chosen models that uses real-world data up until the end of the 1980s. This is mainly because the most cited empirical models in both fields are either published around the 1990s, or they use data only covering the period up until 1990. The next task is to gather updated data. I will stretch this time series to include the most recent governments, using both existing and self-coded data. The final step is to analyse the predictions of the original articles using data from 1945 up until today.

Recent contributions have stressed the need for improvement in the study of the life cycle of governments. Golder et al. (2012b) approaches the field of government formation in a new way. Their main argument is simple - models of government formation does not predict the phenomenons it sets out to because the models are too complex. A model which only considers two institutional constraints is better at predicting different phases of the formation process than models making many more constraints and using more indicators (Golder et al. 2012b, 443). Chiba et al. (2015), looking at government duration, suggest going towards higher complexity. The argument in the article is that government duration is dependent on the government formations process. In effect, they estimate both the formation and duration processes simultaneously.

The more general argument from this debate is the discussions of how to model the life cycle of governments in terms of more parsimony or more complexity. This serves as the rationale behind the selection of original articles to evaluate. I will use one less complex and one more complex model in both the formation and duration literature. The

predictive power of each will be used as evidence for going in one direction rather than the other regarding modeling the life cycle of governments.

1.2 Government Formation

The number of different study subjects within the government formation research reflect the complexity of the formation process. An early aim in the literature was explaining the variation in types of government (see for example Crombez (1996)) - would the government formation bargaining result in a minority, a minimal winning, or a surplus majority government? Advances in the explanations lead to new areas of research on the formation process. A more recent project has been directed towards explaining cabinet composition (see for example Martin and Stevenson (2001, 2010)). This field seeks to explain which of all the potential cabinets are chosen. This thesis will focus on a third large government formation field, namely formation delays in parliamentary democracies. This field was deduced from the observed variation in the duration of formation bargaining processes. Some countries experienced longer bargaining periods than others. This variation breached with the theoretical prediction of no formation delays from renowned bargaining models, such as the bargaining model proposed by Baron and Ferejohn (1989).

Three empirical contributions stand out as important for the field of explaining bargaining delays. Diermeier and van Roozendaal (1998) represents the first published cross-national quantitative study on bargaining delays, and also the first of the formation models that will be evaluated in this thesis. Models from non-cooperative game theory does not predict any formation delay (Baron and Ferejohn 1989). In light of observing that formation delays in reality often stretches over a certain period of time, Diermeier and van Roozendaal (1998) sought an explanation using an information uncertainty approach. They point out that understanding the bargaining process - what happens between an election or other government termination and formation of a new government - is a vital part of understanding the formation process as a whole. They ground their theoretical argument in non-cooperative bargaining theory, where actors have incomplete information. Their main argument is that information uncertainty among the government actors about the government formation process has substantial explanation power on the duration of the government formation process. Whereas Diermeier and van Roozendaal (1998) argued for the information uncertainty approach, Martin and Vanberg (2003)

took an alternative approach. Their main claim is that complexity in the bargaining environment explains variation in the observed bargaining duration rather than the uncertainty approach.

Golder (2010) unifies the information uncertainty and bargaining complexity approaches. This is the second article I will evaluate in the government formation field. In the article, the effects of the bargaining complexity indicators contingent on the level of uncertainty in the bargaining situation is estimated. When combining two theoretical approaches into one explanation of bargaining delays, the article takes a step towards increased theoretical complexity. Therefore, the Golder (2010) article will be used as the complex approach of modeling bargaining delays, while the information uncertainty approach by Diermeier and van Roozendaal (1998) will be used as the more parsimonious approach.

1.2.1 Back to basics?

Models which are unsuitable to the real world must be revised. Golder et al. (2012b), motivated by the failure of theoretical models to predict the observed variation in the government formation process, seeks to replace earlier contributions with a 'zero-intelligence' model of government formation. Their main claim is that formation models, to be able to theoretically predict the formation process, must go back to basics. Their theoretical model puts only two constraints on the formation process - that there always exists an incumbent government and that the government must have majority support in the legislature. The 'zero-intelligence' concept comes from economics. The zero-intelligent agent is "one who acts randomly subject to minimal constraints" (Golder et al. 2012b, 429). This means that institutional constraints guide individual behavior. But it does not mean that other contextual factors are not important. The main motivation behind the article is to empirically predict different stages of the government formation process, not to explain "the bargaining behavior of actors included in the government formation process" (Golder et al. 2012b, 429). To prove that their model performs better than the existing ones, the authors test their model on three government formation research areas - government type, bargaining duration and portfolio allocation. They show that their models predict real world observations better than any of the other approaches in the fields.

In a reply, Martin and Vanberg (2014) criticized the model for being to reliant on the random proposal mechanism. The random proposal mechanism in the Golder et al. (2012b) approach suggests that there are no rules which guide when and how parties propose coalition alternatives. This random proposal mechanism rejects the importance of the bargaining model introduced by Baron and Ferejohn (1989). The bargaining model gives great weight to the actor with the power of giving the first proposal, the formateur. The selection of the formateur follows an exogenous selection rule. This assumption was not supposed to be realistic in the first place, but Golder et al. (2012b) decide to remove all of the formateur assumptions. Combined with the rest of their approach, the 'zero-intelligence' model could be interpreted as an attempt of pushing coalition theory in the direction of weighting country-specific institutional structures the most (Diermeier 2014, 35).

Having established the complexity of the government formation process, an overview of the life cycle of governments is incomplete without including the process of duration and termination of the governments. The following section introduces the field of government duration.

1.3 Government Duration

Government stability is vital to the functioning of parliamentary democracies and has been a study subject since the 1970s (Laver 2003, 23). The duration of a government is an observable phenomenon. A government takes office one day and exits the same offices some months or years later. But what is the probability of a present or future government to break down? This is a question which many have tried to model.

Early research focused on how exogenous events controlled the termination of governments. In particular, Browne et al. (1984, 1986) did systematic research on government duration, the introduction of the critical events approach as the most important contribution. They claimed that the termination of governments could be explained mostly by critical events such as economic shocks, scandals or deaths within the cabinet. Their claim was backed by the observation that the actual distribution of terminations was close to the random poisson distribution (Laver 2003, 28).

A different approach wanted to incorporate both cabinet- and institution specific attributes into a regression framework on explaining terminations. Strøm (1985) was

one of the main proponents of the attributes approach. The main claim from this camp was that an explanation of government termination could not only focus on critical events. Cabinet characteristics and institutional design of the political systems had to have explanatory power on how long governments lasted (Laver 2003, 28).

In this thesis, I will evaluate two government duration classics. The first is the only book-length coverage of government survival, written by Paul Warwick (1994). The second is an article by Diermeier and Stevenson (1999). These two works further developed the usage of the event history models to empirically model government duration. Event history analysis, often also called survival analysis, helped researchers to unify the the government duration field by estimating both critical events and attributes at the same time - that is, the simultaneous estimation of the underlying hazard of government breakdown and how different covariates influenced the risk of termination.

The rationale for choosing these two studies as replication for my study is straightforward - Warwick (1994) represents the parsimonious theoretical account of government duration. The model I evaluate from Warwick (1994) represents one of his most prominent theoretical contributions - the introduction of cabinet ideological diversity as an indicator of government duration. The approach from Diermeier and Stevenson (1999) introduces the strategic actor assumption to government duration. Operationalized, this meant that cabinets would seize the possibility of maximising utility by choosing to terminate in order of gaining power in the following step. This meant that Diermeier and Stevenson tested an assumption that governments ended in two different modes, one in replacement and the other in legislative dissolution. Hence, the Diermeier and Stevenson (1999) approach represents the more theoretically complex approach, while the Warwick approach is framed as the more parsimonious model.

1.4 Evaluating Theories Empirically

A heavy focus on one direction of philosophy of science needs to be justified. I will base my conclusions on the predictive power of models of formation and duration. The failure of a model to predict an outcome will be used as supporting evidence for improving or renewing theoretical approaches. My approach leans on an argument which says that theories able to explain but not to predict represent less theoretical improvement than theories which are shown to both explain *and* predict (Shmueli 2010, 292).

1.4. Evaluating Theories Empirically

How have formation and duration theories been evaluated empirically? Both literatures have found important empirical evidence. However, the theoretical validity of has been drawn on the basis of statistical significant effects. This way of empirically testing the explanatory power of theories is a necessary step, but does not represent an evaluation broad enough for making far-reaching inferences regarding the performance of the underlying theories.

In this thesis, I evaluate the predictive power of four empirical models from different theoretical perspectives in the government formation and duration literature. My contribution stems from a growing sense in the field of political science that significance testing is not the only way to test theoretical predictions. A result table in a scholarly article is a result of one draw from a distribution of coefficients that could possibly describe the relationship between two social or political variables. As Ward et al. (2010) has pointed out, this is not the way of advancing empirical testing of theories, and hence no way of advancing theoretical work in light of empirical evidence in the formation and duration literature.

Ward et al. (2010) is one of the first articles in political science which implements the out-of-sample method in order of estimating the predictive power of theoretical models. Hill Jr. and Jones (2014) is a more recent article which uses the field of state repression to show how and why quantitative political science articles should include out-of-sample evaluations of statistical models. In addition to the substantive claims I make in this thesis, I also aim to contribute to this increasing trend towards arguing that p-values do not represent a sufficient tool for evaluating the performance and utility of theoretical models.

The danger of over-fitting is imminent in research fields with limited data. Over-fitting is a generic term. In this context it is used to mean that empirical results do not generalize to new data. In effect, this can mean that empirical testing of formation and duration theories have been results of specific circumstances in the specific data sets used and not general trends which support the causality which is modeled. Hence, the empirical results that have been published in the fields are results of model that have been over-fitted to the specific data set and therefore do not inhibit predictive capability.

The out-of-sample framework enables researchers to evaluate the generalizability of their empirical results with greater confidence. The original results are used to test

predictions in a new data set. The deviances between the observations in the new data set and how the original model predicts the new observations can be used as evidence for claiming that original articles does not generalize well. Low generalization can be interpreted in the direction of the model being over-fitted to the data, the consequences of which is described in chapter 3.

The approaches used in Ward et al. (2010), and in this thesis, are made possible due to computational developments. The theoretical accounts of government formation and duration was heavily tested around 1990[1]. The period since has been represented by enormous advances in technology. Therefore, this thesis is not a critique of previous empirical testing. It represents instead a significant improvement of how to evaluate theories in government formation and duration.

An overarching issue regarding the evaluation of statistical models is choosing between parsimonious and complex models. Parsimony is often fronted as the panacea of the social sciences. This trade-off is important to my thesis because of the theme underlying the ongoing debate between Golder et al. (2012b), Martin and Vanberg (2014) and Golder et al. (2014). As mentioned in section 1.2.1, Golder et al. (2012b) wants to go back to basics in the formal modeling of government formation. On the opposite side is the Martin and Vanberg article where the authors more or less want to keep the status quo regarding the theoretical modeling of government formation. The underlying question of the debate is: what is the cost of complexity? Golder et al. (2012b) show how they successfully predict different stages of the formation process by simulating values based on the two institutional constraints. However, how does one interpret the results from a model which is only two constitutional constraints away from being a model of pure randomness?

This trade-off guides the selection of replications I have chosen. Regarding government formation literature, the Diermeier and van Roozendaal (1998) article tests one theoretical approach towards explaining bargaining duration, while the Golder (2010) article combines both the Diermeier and van Roozendaal approach and the Martin and Vanberg (2003) approach as described in section 1.2. The assumptions behind the empirical test is more complex than the approach from Diermeier and van Roozendaal (1998). The predictive power test will give a basis for claiming something about which of the models

[1]Except of the articles and book under evaluation here comes for example Laver and Schofield (1990), Strøm (1990) and Laver and Shepsle (1996) which all contributed heavily to theorize the life cycle of governments.

that predicts the outcome the best.

Regarding the government duration literature, the Warwick (1994) approach represents the basic version of a government duration model. The second replication represents a further development of the Warwick approach. The Diermeier and Stevenson (1999) article pursues, following newly published theoretical accounts, that actors involved in the life cycle of governments are not only cooperative. The non-cooperative theoretical account suggest that dissolving a cabinet yield further gains than remaining in office until the general election. Hence, Diermeier and Stevenson (1999) estimates the duration of governments as two different processes - one where the cabinet ends in replacement and one where it strategically dissolves and calls a new election. This competing risk framework serves as a theoretically more complex approach than Warwick (1994).

The guiding research question in this thesis is motivated by the most recent approaches to the fields of government formation and duration. Golder et al. (2012b) wants formation models to be built from the ground up using a parsimonious theoretical model. Chiba et al. (2015), on the other hand, wants the research agenda to incorporate formation and duration to one continuous process. This is a clear argument for increasing the complexity in how researches model the life cycle of governments. Therefore, my research question is the following: How well does parsimonious and complex theoretical models in the fields of government formation and duration generalize when faced with new data?

1.5 Outline

The following section describes how I will proceed in evaluating theories in government formation and duration, contextualized by the research question at the end of the previous section.

Chapter 2 introduces the state of the art of both government formation and duration literature. The chapter is directed towards explaining the theoretical development in both fields in general by pointing to the explicit theoretical implications which are being tested. Secondly, I will show how the articles evaluated chose their indicators and the theoretical rationale behind the selection of predictors.

Chapter 3 goes through the research design. Firstly, I will display the data underlying this thesis. Secondly, I will describe the variables used, and the sources used in situations where I needed to code new data. Thirdly, I will focus on how to test predictive power

empirically. Finally, a section will be used to discuss theoretical and practical problems of using the chosen statistical models.

Chapter 4 presents the evaluation of the government formation models. Here, I will first show that my data gives the same substantial conclusions as the original models. I start the evaluation by describing the in-sample predictive performance of the models. This will be used as a motivator for part two of the evaluation which is the out-of-sample predictions. Here, I describe the results both graphically and substantially. The third section test the models using the cross validation method. The last section sums up the chapter.

Chapter 5 evaluates the government duration models. The same procedure will be applied as in 4. I will show through substantial effects how my data is comparable to the original data. Then the in-sample, out-of-sample and cross validation methods will be applied. This chapter will also end with a section summarizing the results.

Chapter 6 nuances the in-sample, out-of-sample and cross validation results from chapters 4 and 5. Here, I show how the models are able to predict the 25th quantile, median and the 75th quantile. This section will also show the large deviances between the predicted and observed duration. This will be used to motivate the final section which investigates an alternative way of calculating prediction error.

Chapter 7 deals with the summary and the implications of the analysis. Firstly, I will summarize the main findings. Secondly, I will discuss challenges and possible solutions to the empirical and theoretical modeling of government formation and duration.

CHAPTER 2

Making and Breaking Governments

Two questions stand out in the literature on the life cycle of governments: which governments form, and what explains the duration of governments (Diermeier 2014, 41). This chapter introduces literature that have evolved around these two questions, that is - around the subjects of making and breaking governments.

2.1 Formation

"The government formation literature is one of the largest literatures in all of political science" (Golder et al. 2012b, 443). Significant attention has been focused on many different aspects of how governments are forming in parliamentary democracies. The theoretical development on the bargaining process surrounding the formation point stems from game theorists. Country experts have also contributed with their detailed knowledge of the individual political system, and the country-specific characteristics of the formation process (Martin and Stevenson 2001, 33). The area is therefore rich, especially regarding theoretical accounts of the formation process.

First, this section introduces the theoretical development surrounding the specific feature of government formation this thesis looks into - namely the formation duration. The main theoretical accounts are the information uncertainty approach and the combined information uncertainty and bargaining complexity approach. This section will also present the empirical implications derived from these two different theoretical explanations of formation duration.

2.1.1 Information uncertainty

The actuality of the study of government formation duration was revitalized in light of the Belgian formation process that lasted from early 2010 to late 2011, making it a total of 541 days since the last election (Devos and Sinardet 2012, 167). Empirical models of government formation have used delays in the formation process as indicating political crises. This is so because of the prominent bargaining model proposed by Baron and Ferejohn (1989). The bargaining model yields no room, theoretically, for any formation delay. Hence, observed delays such as the recent example from Belgium caused legislative study researchers to investigate the empirical clear deviance from the bargaining model.

The bargaining process as laid out by Baron and Ferejohn (1989) is shown in figure 2.1. The bargaining model assumes rational actors with complete information, and decisions under majority rule. The guiding assumption is the existence of a political party with the agenda power of making the first proposal, often called the formateur (Druckman et al. 2014, 202). The proposal from the chosen formateur consist of a distribution of pay-offs regarding cabinet positions. A clear implication is that the proposal from the formateur will immediately be accepted by all actors - because of the complete information assumption.

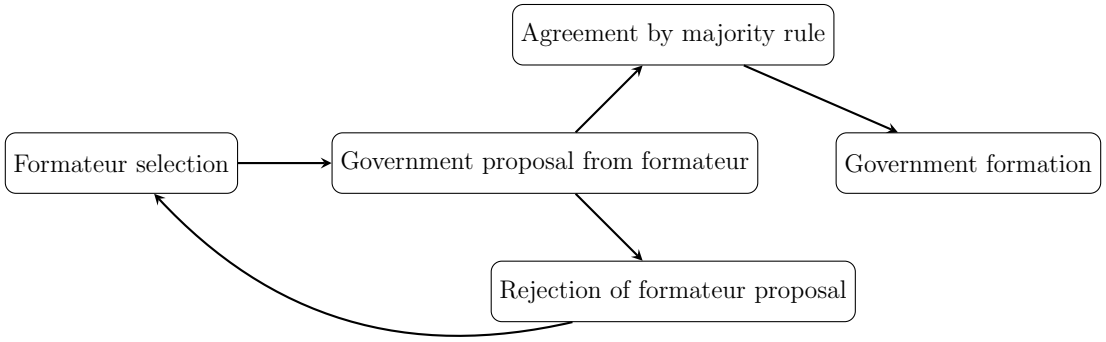

Figure 2.1: The Baron-Ferejohn bargaining model. Based on description in Golder et al. (2012, 428).

However, as with Belgium and other parliamentary democracies, the empirical backing for the formal bargaining model is slim. Only when the information assumption is relaxed is it possible to theoretically predict formation delays (Golder 2010, 19). Therefore, Diermeier and van Roozendaal (1998) suggests testing an implication of the incomplete information bargaining approach - "the degree of uncertainty over relevant bargaining parameters is positively correlated with the expected formation time" (Diermeier and van Roozendaal 1998, 612). Formation delay, then, becomes a function of political actors'

information uncertainty regarding preferences and strategies of the other political actors in the formation bargaining.

The parties, knowing that they have uncertain information of the preferences and strategies of the other actors, can benefit from using different institutional opportunities, such as inner-cabinet negotiations or policy-specific agreements (De Winter and Dumont 2008, 134-135). However, the consequence of parties searching for information through these kinds of institutional characteristics is formation duration.

Diermeier and van Roozendaal (1998) uses two indicators to investigate the empirical validity of the information uncertainty approach. The first indicator is the timing of the government formation bargaining - whether the bargaining happens immediately after or in between elections. Theoretically, governmental actors have more information about the preferences and strategies of the other actors when they have interacted over some time in the parliament (Diermeier and van Roozendaal 1998, 620). A government bargaining which happens in an inter-election period has this characteristic. The opposite happens in an post-election context. The game has changed as a consequence of an election. Following, there are new party platforms, new policy goals and in some cases also changes to the legislative party composition. The actors then have yet to experience the real preferences of the other actors in the legislature. Since formation bargaining is often guided by parties having to reach a comprise on a viable governing policy agreement (Gallagher et al. 2011, 419), government formations are delayed because of higher levels of information uncertainty among the bargaining actors.

The second indicator of information uncertainty is the mode of termination for the previous cabinet. The argument is that a previously defeated governments, indicated by a cabinet loss to a confidence-vote or other oppositional pressure leading to termination, leads to leadership battles within parties. One thing is to lose after an election, another thing is to be politically beaten in the parliament. Therefore, party leaders can be challenged by ambitious party members inside their own party. This conflictual environment, caused by the previous defeat, decreases the degree of trust in the information which is available to the other parties. Hence, this also decreases the level of information on the real preferences of the other actors. The information uncertainty indicators and their expected effects on bargaining duration are summed up in table 2.1 below.

Table 2.1: Information uncertainty indicators, Diermeier and van Roozendaal (1998)

Variable	*Expectation*
Post-Election	Government formation duration increases in light of a post-election context relative to an inter-election context.
Previous Defeat	Government formation duration is longest in situations where there exist leadership battles as a result of a previous parliamentary defeat.

2.1.2 Combining information uncertainty and bargaining complexity

Martin and Vanberg (2003) argue for a bargaining complexity approach rather than the information uncertainty approach by Diermeier and van Roozendaal. The bargaining complexity approach is characterized by the assumption that a complex bargaining situation yields increased bargaining duration, compared to a simple bargaining situation. Martin and Vanberg (2003) uses indicators such as the effective number of parties in the legislature and ideological diversity inside the cabinet to operationalize bargaining complexity. The rationale behind the inclusion of the effective number of parties is that it leads to an increased number of possible government options, and hence, whoever is guiding the bargaining must consider more options than in a bargaining with fewer government options. This argument and the subsequent empirical implication comes from the government formation classic by Laver and Schofield (1990), and have since been the most used indicator of bargaining complexity (De Winter and Dumont 2008, 136).

Golder (2010) combines the two theoretical approaches. The main argument in the article is that an explanation of bargaining duration is not either uncertainty or complexity. Information uncertainty always affects delays in government formation, and the effect of complexity is contingent on uncertainty among the actors (Golder 2010, 9). The reasoning is simple - complexity indicators are known to the actors involved in the bargaining. This is the main reason for Golder to expect that the effects of the complexity indicators either gives constant or decreasing formation delays. In effect, theoretically, this means that complexity should not yield any effect in inter-election contexts. When there exist uncertainty among the actors about the preferences and strategies of the other actors involved in the bargaining, the effect of bargaining complexity should indicate a further increase in formation delays. For Golder, the uncertainty approach has therefore causal priority compared to the complexity approach (Golder 2010, 12). To measure information uncertainty Golder uses only the post-election indicator.

The bargaining complexity indicators used by Golder include the effective number of

parties mentioned above, Ideological polarization and positive parliamentarism. Higher levels of ideological polarization in the legislature will lead to bargaining delays. The underlying assumption is simple - ideologically tighter connected legislatures can find a viable government solution quicker than ideologically diverse legislatures. Positive parliamentarism is defined as an institutional rule which says that the new cabinet must pass a majority investiture vote in the parliament to officially be invested (Rasch 2004, 115). This indicator is expected to increase complexity, and hence increase formation delays, because of the explicit demand that the final cabinet must have majority backing in the legislature. The theoretically relevant indicators for complexity are summarized in table 2.2.

Table 2.2: Bargaining complexity indicators, Golder (2010)

Variable	*Expectation*
Effective number of parties in legislature	A higher number of effective parties increases bargaining duration contingent on an uncertain context. In inter-election contexts the expectation is constant or negative effect.
Polarization in legislature	Higher levels of polarization within the legislature is expected to increase bargaining duration in an uncertain context.
Positive parliamentarism	Countries having the investiture rule are expected to have longer bargaining duration contingent on an uncertain context.

2.2 Duration

Whereas the government formation literature has been characterized by a heavy theoretical focus, the government duration literature has been skewed towards the empirics (Laver 2003, 38). The result has been that theoretical and empirical accounts of duration has taken separate ways. The theoretical strand of the duration literature suffers from the definitional weakness of a priori modeling - the models are static, and not capable of predicting terminations happening during the lifetime of a cabinet. The duration of a government is not a static process. This is also why a typical *a priori* model as the Baron-Ferejohn model can not be used in duration studies (Diermeier 2014, 43).

This section will introduce developments in the government duration literature. First, I will describe the early developments in the 1980s, which was based mainly around a debate between the critical event and the attributes approaches. Second, I will explain how the field was unified and highlight the ideological approach by Warwick (1994). Third, the strategic actor assumption is presented and connected to the competing risk approach

of Diermeier and Stevenson (1999).

2.2.1 Attributes and critical events

Before a unified attempt by King et al. (1990), there was a heated debate between critical events and attributes approaches on the determinants of government duration. Browne et al. (1984) represented the critical events approach. This approach lifted the critical events hypotheses up and into the light. The argument was that the attributes approach essentially assumed that "governments more likely to form are also likely to endure longer" (Browne et al. 1988, 931). Browne et al., instead, claimed that the attributes approach had to be amended by taking critical events into a more complete account of government duration. The critical events are never explicitly defined in the literature, but implicitly understood as outside events like economic shocks, minister scandals and other non-structural events which causes a government termination (Strøm 1988, 926).

Against this, Strøm (1985, 1988) argued for the attributes approach. He claimed that cabinet duration had causal structures which could be best revealed by using institutional and cabinet specific predictors. Opposite to the critical events approach, where cabinet termination only could be explained by randomness, the turn towards the attributes approach was a significant progress towards giving a systematic of the observed variation in government duration.

After King et al. (1990) the debate between the two competing approaches slowed down. By introducing the event history model framework, King et al. (1990) became a milestone for the government duration literature. The real benefit of using the event history model compared to the flawed OLS-framework used before them was that both attributes and critical events could be estimated simultaneously. The concept of the baseline hazard replaced critical events, and the attributes could contribute to explain the systematic variation in government duration. Hence, the field of government duration was unified.

King et al. (1990) also introduced the concept of censoring governments that sat the whole period, but were terminated due to a constitutionally mandated election. These governments could be assumed to have survived longer if there had not been an election. King et al. (1990, 853) chose to censor every government that sat within 12 months of the forthcoming constitutionally mandated election. The specifications King et al. chose had

several shortcomings. This was corrected in a book-length coverage by Warwick (1994), and will be discussed below. However, more importantly here, Warwick retained some of the attributes used by King et al. (1990).

These attributes are summed up in table 2.3, along with their expected effect on government duration. A majority cabinet is expected to last longer than cabinets without majority support in the legislature. The logic is straightforward - majority cabinets can not be beaten by the legislature, unless one of the governing parties turns on them. The post-election indicator measures the effect that a cabinet invested in the beginning of a constitutionally mandated period will have the opportunity to sit longer than inter-election cabinets. Cabinets that have passed the investiture vote have been showed trust by the legislature, and hence these cabinets are expected to last longer than cabinets in countries that do not practice investiture.

Table 2.3: Attributes of government duration, Warwick (1994)

Variable	*Expectation*
Majority status	Majority cabinets have longer duration than minorities.
Post-election status	Post-election cabinets have longer duration compared to inter-election cabinets.
Investiture	Cabinets passing the investiture rule are expected to have the longest government duration.

One of the main surges in Warwick (1994) was to replace the exponentially distributed hazard of termination used in King et al. (1990). The exponential parametrization of the baseline hazard means that the King et. al. (1990) article expected the hazard rate to be flat, i.e. the risk of cabinet termination remains constant, independent of how long the cabinet had been in power (Box-Steffensmeier and Jones 2004, 22). Warwick's point was that the Cox proportional hazard model gave more theoretically sense. In the Cox proportional hazard model the baseline hazard of termination is left unspecified, and hence the hazard does not need to follow an assumed distribution as the parametric survival models must[1]. Therefore, Warwick argued that this approach was a better way of testing attributes and critical events without having to expect a certain distribution of government duration.

Another key development brought on by Warwick (1994) was to introduce indicators of the policy-seeking government actor to the study of cabinet duration. Essentially,

[1]For more on the cox proportional hazard model and other survival models, see Box-Steffensmeier and Jones (2004, 47-68).

the policy-seeking[2] government actor is not only concerned with getting into office and stay there. The actor is also concerned with introducing and implementing the favoured policies of the actor. Warwick claimed that earlier studies neglected the fact that ideological distances within the cabinet had large consequences for how long the cabinet lasted.

Table 2.4 summarizes the ideological indicators used by Warwick. Warwick uses three dimensions to evaluate the policy-seeking assumption, the general left/right dimension, clerical versus secular dimension and the regime support dimension. These are the ideological indicators, amongst other measures and dimensions which is reported to have a significant impact on government duration (Warwick 1994, 59). The expected effect of ideological distance is that larger policy deviances between cabinet parties can lead to severe conflicts, and higher levels of conflict increases the probability of termination. Hence, higher levels of policy diversity within the cabinet should cause shorter government durations.

Table 2.4: Policy-seeking indicators, Warwick (1994)

Variable	*Expectation*
Left-Right diversity	Increased left-right cabinet diversity decreases expected government duration.
Clerical-Secular diversity	Increased clerical-secular cabinet diversity decreases expected government duration.
Regime Support diversity	Increased regime support cabinet diversity decreases expected government duration

2.2.2 Strategic dissolution

King et al. (1990) and Warwick (1994) assume implicitly that all cabinet actors want to stay in government as long as they can. This means that strategic decisions by parties is left out of the early explanations of government duration. Lupia and Strøm (1995) introduced a formal model with the strategic actor assumption. Cabinet parties, in countries where the constitution mandates the executive the opportunity of dissolving the legislature and calling a new election, have the possibility of strategically choose to opt out of government. Lupia and Strøm (1995), therefore, formalized a game which could explain the benefits from strategically bringing down the cabinet (Laver 2003, 35). Implications of this strategic approach has been under some empirical scrutiny, for example by Strøm and

[2]The policy-seeking assumption is often attributed to Riker (1962).

Swindle (2002). They validate empirically that dissolutions happens, and they explain the presence of the strategic choice of ending a cabinet with indicators such as the role of the actor responsible for enacting the dissolution (Strøm and Swindle 2002, 575). This supports the argument that government terminations are not random events - they can be explained by systematic indicators[3].

The inclusion of the strategic actor assumption represented great progress in the government duration field. The Diermeier and Stevenson (1999) article attempts to validate the strategic actor assumption by testing empirically implications of Lupia and Strøm (1995) model. Lupia and Strøm (1995) claims that there exist two separate processes of government termination which strategic actors can take advantage of. One mode of termination is the choice of dissolving the parliament and calling new elections. This is done because the governmental actors calculate more benefits from dissolving the parliament and calling new elections than staying in office until the end of the constitutionally mandated period. The other mode of termination is situations in which cabinet changes do not directly follow from elections, and hence it does not involve dissolution and the calling of new elections. The decision to be replaced is also a strategic choice, where the benefits of being replaced beats the cost of staying in office until the next election (Diermeier and Stevenson 1999, 1052). The termination indicators are summed up in table 2.5.

Table 2.5: Mode of government termination, Diermeier and Stevenson (1999)

Variable	*Expectation*
Dissolution termination	Cabinets ending in dissolutions are expected to have an increasing hazard of failure.
Replacement termination	Cabinets ending in replacements are expected to have constant or decreasing hazard of failure.

[3]And, often times, these indicators are institutional rules, (Strøm and Swindle 2002, 589).

CHAPTER 3

Research Design

In the following section, I will describe the data underlying this thesis. Additionally, I will explain the sources and the methods behind the data I have gathered. The second section outlines how the empirical models will be evaluated - that is, in-sample, out-of-sample and 5-fold cross validation. The main surge here is directed towards explaining the benefits of doing out-of-sample prediction. The section also presents the cross validation method as an additional method which as well adapted to research areas with restricted data. The last section gives a walk-through regarding the statistical models which will be used in the analyses. Additionally, I will discuss problems and a proposed solution regarding the statistical modeling of government formation and duration.

3.1 Data

I have based my data collection on the cabinet counting regime in the European Representative Democracy Data base (ERD) (Andersson et al. 2014). The origin of this data set is the data as presented in Müller and Strøm (2000). The most recently updated version of the ERD data set contains cabinet-level data on 29 European countries from 1945 to 2012. It has also updated errors and other misspecifications from the original Müller and Strøm data.

Table 3.1 below shows descriptives of the countries which will be used in the replication of the original models of government formation and duration. Table 3.2 shows cabinets after 1989 up until the most recent data point. This is the data which will be used in the out-of-sample prediction approach. Included are also statistics over how the cabinets are coded due to their mode of termination, and how they are censored. This data is

elaborated on in the variable section, below[1].

Table 3.1: Cabinets 1945 - 1989

Country	Period	N cabinets	Duration	Barg. duration	Replacements	Dissolutions	Pooled	Censored
Austria	1945-1987	17	921	35	7	8	15	2
Belgium	1946-1988	30	514	42	18	10	28	2
Canada	1945-1988	18	964		4	8	12	6
Denmark	1945-1988	26	625	9	5	16	21	5
Finland	1945-1987	39	397	28	27	3	30	7
France	1947-1988	40	393	11	9	5	14	28
Germany	1949-1987	22	666	19	20	2	22	0
Greece	1977-1989	7	639	7	1	2	3	4
Iceland	1944-1989	23	716	22	11	8	19	4
Ireland	1944-1989	18	954	13	2	14	16	2
Italy	1946-1989	42	346	44	35	7	42	0
Luxembourg	1945-1989	14	1240	29	11	2	13	1
Norway	1945-1989	22	738	9	18	0	18	4
Portugal	1976-1987	12	416	54	5	5	10	1
Spain	1977-1989	6	942	30	1	4	5	1
Sweden	1945-1988	22	761	5	17	1	18	4
Netherlands	1946-1989	20	798	78	14	5	19	1
UK	1945-1987	17	960	16	3	10	13	4

Table 3.2: Cabinets 1990 - 2015

Country	Period	N cabinets	Duration	Barg. duration	Replacements	Dissolutions	Pooled	Censored
Austria	1990-2013	8	975	74	1	3	4	4
Belgium	1991-2010	10	636	42	6	1	7	3
Canada	1993-2015	9	806		1	6	7	2
Denmark	1990-2011	9	839	3	2	5	7	2
Finland	1990-2011	10	736	18	6	0	6	4
France	1991-2012	11	696	1	2	1	3	7
Germany	1990-2013	7	1162	33	2	1	3	4
Greece	1990-2011	8	982	3	0	4	4	4
Iceland	1991-2013	9	884	10	5	0	5	4
Ireland	1992-2011	7	969	20	2	3	5	2
Italy	1991-2011	13	535	44	7	1	8	4
Luxembourg	1994-2013	5	1377	37	0	1	1	4
Norway	1990-2013	8	1029	15	3	0	3	5
Portugal	1991-2011	7	996	21	1	3	4	3
Spain	1993-2011	5	1306	42	0	2	2	3
Sweden	1991-2014	7	1192	8	2	0	2	5
Netherlands	1994-2012	8	748	90	2	3	5	1
UK	1990-2015	7	1271	7	2	2	4	3

One of the main goals of this thesis is to evaluate the generalizability of empirical models on government formation and duration when tested on new data. For three of the four original articles replication material was available. However, my reasoning on collecting all data on my own is that the validity of the data has improved, all the while there has been much improvement in measurements and data coverage today as opposed to the end of the 1980's.

A challenge using newly collected data to replicate data collected in the late 1980's

[1]Notes to tables 3.1 and 3.2: The *Period* column shows the first and the last formation year for each country in the old data. For the new data, the first year is the formation year and the last is the termination year. Canada has no value indicating bargaining duration. This is because Canada is not used in the bargaining duration studies used in this thesis. The censoring and termination data is only relevant for the government duration studies.

is the differences in counting cabinets. Differing numbers of governments also gives differences in the data on duration. The ERD data has in general higher recorded duration, both regarding bargaining and cabinet duration than the data I am trying to replicate. The deviances are biggest in countries with the most complex party systems, such as Italy and Finland. My choice of the ERD stems from the fact that it represents the most recent approach of collecting and updating cabinet data. Given that newer data is more valid compared to old data, I will continue forward relying on the ERD data.

3.1.1 Dependent variables

The dependent variables are bargaining duration and government duration. Bargaining duration measures the length in days between the date of an election or a cabinet resignation and the following date of the official formation of a new government. Government duration is measured as the number of days between the official formation of a cabinet until the official end of a cabinet. The main rule, as has been recognized throughout the literature (see for example Müller and Strøm (2000) and Strøm et al. (2008)) is that a change in cabinet is counted when facing i) an election, ii) a change in the prime minister position, iii) a successful vote of confidence or iv) technical termination such as death of prime ministers or changes in party composition.

In the following sections I will describe the sources of the variables, and how they have been constructed.

3.1.2 ERD data

Post-election: Cabinets forming as a result of an election are expected to have longer bargaining duration than cabinets formed between elections. This indicator is used in both of the bargaining duration articles used in this thesis.

Investiture: This structural attribute is included in three of the four models replicated. Countries that demand an explicit legislative majority for investing the cabinet are coded as having investiture.

Effective number of parties in the legislature: This variable counts the number of parties in the legislature and weights the number on the seat share of each party.

Single party majority and majority status: These data are based on the seat share of each cabinet recorded in ERD. Single party majority cabinets are identified different from

cabinets with majority status.

Cabinet ideological diversity: Cabinet ideological left-right diversity is calculated using the rile score, which is calculated from the coding of party manifestos, see Volkens et al. (2014). The general left-right measure is identified for each party, and the value for each cabinet is the absolute distance between the most extreme parties according to their left-right placement.

3.1.3 Other sources

Previous defeat: This covariate is the second information uncertainty indicator, used only by Diermeier and van Roozendaal (1998). The measure consist of termination mode of the previous cabinet. The theoretical rationale is that a previous defeat leads to longer duration of government formation because these situations inhibit more conflict between and within the bargaining parties (Diermeier and van Roozendaal 1998, 620).

I have used data from the "Party Government Dataset" (PGD) by Woldendorp et al. (2000) to measure this variable. The data set is recently updated, and covers governments up until 2012, as discussed in Seki and Williams (2014). The variable from PGD codes different government termination according to 7 categories. The termination category I use to find previous defeat is the "lack of of parliamentary support" category (Seki and Williams 2015, 9). The PGD data code this variable in cases of a successful vote-of-confidence[2] or where parties withdrew support from the government.

The cabinets are identified and matched by using the date of formation. Where the formation dates in PGD differed from the cabinets in ERD I recoded the formation dates in the PGD data to ERD dates. Next, I lag the variable one position, meaning that the original variable gives the termination mode for the original cabinet, while the lagged variable gives the termination mode of the previous government.

Caretaker status: Caretaker governments are included in Diermeier and van Roozendaal (1998) and excluded from the sample in Golder (2010). Diermeier and van Roozendaal (1998) controls for caretaker governments with a dummy variable. Diermeier and Stevenson (1999) and Warwick (1994) also use caretaker governments, but they do not control for their specific effect. Caretaker governments are expected to have shorter bargaining durations than regular cabinets because they are considered as a short-term

[2]Successful from the perspective of the instigators of the investiture vote.

solution that does not really influence the policies implemented (Diermeier and van Roozendaal 1998, 620). I used the "Parliament and Government composition database" (ParlGov) Holger and Manow (2012) to identify cabinets with caretaker status.

Identifiability: This variable is only used by Diermeier and van Roozendaal (1998). The variable comes from Strøm (1990) who coded the decade-basis recognition of pre-electoral coalition alternatives by using a 0, 0.5 to 1 scale, depending on the level of identifiability. The original data is unavailable. Instead, I utilize data on pre-electoral coalitions from Golder (2005). She identifies pre-electoral coalitions following two indicators - the pre-electoral coalition must be officially announced and the parties involved in the pre-electoral coalition cannot compete in the election as truly independent parties (Golder 2005, 652). Again, these data are connected to the cabinets in ERD by using the date of formation. And, dates that do not fit in the data from Golder (2005) are recoded to fit with ERD data.

Polarization: Ideological polarization in the legislature is measured using the Esteban and Ray polarization index (1994). Golder (2010) uses this index, but her replication data ranges only up until 1998. Therefore, I coded the rest of the legislatures from 1999 and up until today. The measure consist of data on policy distances and party seat shares. I used the general left-right measure from the Chapel Hill expert survey (Bakker et al. 2015) as the baseline. In case of missing information on parties, I filled in with values from the combined left-right expert survey measure in ParlGov, which combines data from multiple sources[3]. I used the seat share data from the same data set. The data for the period I needed came mainly from the expert survey of Benoit and Laver (2006). The formula for calculating the polarization index is

$$Polarization = \sum_{i=1}^{n} \sum_{j=1}^{n} \Pi_i^{\alpha+1} \Pi_j |y_i - y_j|, \tag{3.1}$$

where i is a party in the legislature and j is a second party in the legislature. Π is the seat share of party i and j. α is a fixed parameter, which is set at 1.3. The maximum value for α is 1.6. I chose 1.3, following Indridason (2011, 715), who created the polarization data used in Golder (2005). A higher α means increased sensitivity to polarization (Indridason 2011, 694). Party i is compared to all the party j's in the

[3]The measure combines expert surveys from Castles and Mair (1984), Huber and Inglehart (1995), Benoit and Laver (2006) and Bakker et al. (2015).

legislature. Then the next party in the legislature becomes i, which again is compared to the rest of the party j's in the legislature. Finally, the pairwise comparisons are summed up to the legislative polarization index.

Continuation rule: This variable is used in the articles investigating bargaining duration. It controls for the presence of a continuation rule. The continuation rule states that the incumbent government always makes the first cabinet proposal for the next bargaining round (Diermeier and van Roozendaal 1998, 620). Diermeier and van Roozendaal identify the continuation rule to exist in Norway, Sweden and Denmark (Diermeier and van Roozendaal 1998, 621). Sona Golder (2010, 30) uses the same coding of the confounder, except that she also adds Great Britain to the list of countries having the continuation rule. The variable is coded 1 for the countries identified to follow the rule and zero otherwise.

Clerical-secular and regime support dimensions: To have comparable policy estimates for parties ranging from 1945 to 2015 I found the manifesto coding from the Comparative Manifesto Group (CMP) most beneficial (Volkens et al. 2014). The clerical-secular and the regime support measure used in Warwick (1994) and Diermeier and Stevenson (1999) comes mainly from Dodd (1976). Those data are not available. Therefore, I coded every cabinet in my data set according to the absolute distance between the most extreme cabinet parties. The cabinet parties were identified using information on cabinet party composition in the ERD data set and then matched with party names in the CMP data. Clericalism and secularism is coded according to the traditional morality category in the manifesto data[4]. Regime support is covered by the anti versus pro constitution dimension[5].

Returnability: This indicator comes from Warwick (1994). He argues that there exists an underlying probability of returning to power for each political system. This variable is operationalized as the mean of the proportion of parties in the previous cabinet which re-enter the following cabinet. Returnability is fixed for each political system. The replication data set from Chiba et al. (2015) contains the proportion of parties which also were present in the previous cabinet. This data was merged with the ERD data using the formation date as merge key. There were some missing information. I calculated the missing returnability values in the Chiba et al. (2015) data on the basis of the identification

[4]Category 603 and 604 in Volkens et al. (2014)
[5]Category 203 and 204 in Volkens et al. (2014)

of cabinet parties in ERD. Firstly, I estimated a single probability for each country up until 1990 (or 1998 in the case of Golder (2010)). Secondly, I re-estimated the probability for the data from 1990 (1999) up until today using the full data set. This means that the returnability score for the old data was calculated with the mean returnability from 1945 to 1989(1998) for each political system, while returnability for the new data was calculated on the mean returnability from 1945 to 2015, for each political system.

Censoring government duration: I have relied on the most recently updated data on the mode of termination. The date originates from the original Martin and Stevenson (2001) data. This data was updated through 2012 for use in Chiba et al. (2015). Some cabinets had missing values. I coded these cabinets manually, using information from the election database in ParlGov (Holger and Manow 2012). Chiba et al. (2015) cites Diermeier and Stevenson (1999) regarding their competing risk approach, and codes cabinets in relation to if they ended in dissolution or replacement (Chiba et al. 2015, 52). The two modes are used separately for the competing risk approach, while the two modes are pooled to make the censoring variable for the Warwick (1994) approach to government duration.

All variables and data sources mentioned in this section are summarized in table 3.3 and 3.4, according to the two respective fields.

Table 3.3: Government Formation

Variable	**Source**
Majority status	ERD (2014)
Post-election	ERD (2014)
Investiture	ERD (2014)
Eff. number of parties	ERD (2014)
Previous defeat	PGD (2014)
Identifiability	Golder (2005)
Polarization	Golder (2010)
Caretaker	ParlGov (2012)
Continuation	Self-coded

Table 3.4: Government Duration

Variable	**Source**
Majority status	ERD (2014)
Post-election	ERD (2014)
Investiture	ERD (2014)
Ideological diversity	ERD (2014)
Clerical-Secular diversity	CMP (2014)
Regime support diversity	CMP (2014)
Returnability	Chiba (2015)
Censoring	Chiba (2015)

3.2 Evaluating Predictive Power

3.2.1 In-sample prediction

One way of testing the quality of a model is to calculate the error rate between the predicted and observed values of the dependent variable on the original sample. The

empirical literature on government formation used in-sample predictions as a model evaluation tool. Strøm (1990) shows that his model for explaining the presence of minority governments predicts 70% of all formed minority governments in-sample. Another example is Bäck and Dumont (2007), who evaluate the in-sample predictive performance of office- and policy seeking theories on the selection of the government coalition.

A weakness when doing in-sample predictions is that the estimated predictions are based on the particular sample of data used to fit the model. The catch is that the researcher can never know whether the estimates are just results of over-fitting or that they describe real causal effects (Hill Jr. and Jones 2014, 662). Over-fitting means that the estimated model is merely describing the data, or worse, only describing noises in the data. The real interest is to catch real, underlying effects of a causal theory in the available data.

A more specific point to the government formation and duration literature is that the fields use data which are naturally restricted. There are only *that* many cabinets, no matter how one chooses to count them. As a consequence, researchers using cabinets as their units "[...]have been picking over the entrails of essentially the same dataset" (Laver 2003, 27). One consequence of the restricted nature of cabinet data is that the probability of over-fitting increases (Jones and Linder 2014, 21). Researches can not afford to hold sets of data outside their analysis, and hence the models are fitted using all data available. This does not present a problem with the analysis in itself, but it present a problem of how to evaluate the ability of the model to describe real effects.

In all, the in-sample prediction method is an alternative way of evaluating a model's performance rather than looking only at the signs of coefficients and statistical significance. However, because the method offer slim possibilities for knowing whether the resulting predictions are results of underlying trends in the data or just noises, the in-sample prediction method is not bulletproof. This weakness can be amended by evaluating the performance of the model on a holdout data set, i.e. the out-of-sample method.

3.2.2 Out-of-sample prediction

Out-of-sample prediction has to do with an estimation of a model on the original sample and then measure the error rate between the predicted and observed values in a new

data set. For example, Paul Warwick (1994), estimates government duration models on observations from 1945-1989. My task of evaluating the theoretical approach from Warwick is to gather data ranging from 1990 and up until today. The original model will then be used to predict the duration in the new data, contingent on the values of the predictors used. The errors between the actual and the predicted cabinet durations are measured and used to calculate a unified error measure. This is the out-of-sample prediction method of holding out a test set to validate the original model (James et al. 2013, 176). Testing the original model on a holdout data set, assumed to be based on the same data generating process as the original sample, enables researches to put their models up to a real test - is the model really able to predict unseen observations?

Scholars doing cross-national quantitative studies often want to infer something about a general trend of the phenomenon in interest. Out-of-sample prediction is a way of testing how a theoretical model is able to predict unseen events, and hence a very relevant tool for evaluating theories and their generalizability. The method is illustrated in figure 3.1 below.

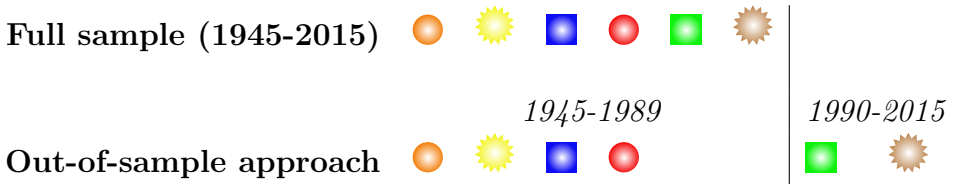

Figure 3.1: Illustration of out-of-sample validation approach. The full sample is split in two. One part ranging from 1945 to 1989. 1990 to 2015 are held out and used to validate the original model using data from the first part.

The predictive power of the old model estimated on new data will tell something about how the model is capable of finding the real causal effects of the phenomenon in interest. A model is expected to yield poorer predictions out-of-sample compared to in-sample. This is because many statistical methods are attuned to minimizing the in-sample prediction error (James et al. 2013, 32). Therefore, the consideration of traces of over-fitting in the original model, as mentioned in section 1.4, does not automatically follow from poorer out-of-sample predictions. The trends toward over-fitting must be judged independently for each model, based on a combination of quantitative and qualitative judgements.

The out-of-sample method has two drawbacks. One is that the resulting error rate from splitting data into two folds, as with the Warwick example above, is highly dependent on where and how the split is applied to the data. The data in this thesis has been clearly divided, one old data set from 1945 to 1989, 1998 for Golder (2010), and one new

data set from 1990 (1999) and up until today. In effect, this means that the comparison of predictions in- and out-of-sample is guided by time specific traits. The method will therefore possibly yield variable results if a different split had been made, for example by letting the computer randomly divide the data randomly 90% and a holdout set of the remaining 10%.

The second drawback is that the model is fitted on the test data set, which has fewer observations. Statistical methods do usually perform worse when applied to a limited set of observations. This can lead to overestimated error rates (James et al. 2013, 178). One solution to these two problematic areas of the out-of-sample approach is the k-fold cross validation method.

3.2.3 K-fold cross validation

Instead of splitting the data into an original set and a holdout set, the k-fold cross validation method randomly splits the observations into k equally large random folds. To increase the randomness of the method, I estimate 100 models from each fold. The mean results for each fold are used in the analyses chapters. The number of folds can be as high as equal to the number of observations. However, when tested empirically, the 5- or 10 fold cross validation analysis yields error rates that are neither highly biased nor have high variance (James et al. 2013, 184). In this thesis the 5-fold cross validation procedure is chosen. After choosing the amount of folds, the next step is to fit the model on *k-1* folds and use the fitted model to predict the data in the k fold, which is the fold left out. This process is repeated k-times (James et al. 2013, 181). Figure 3.2 illustrates the 5-fold cross validation method.

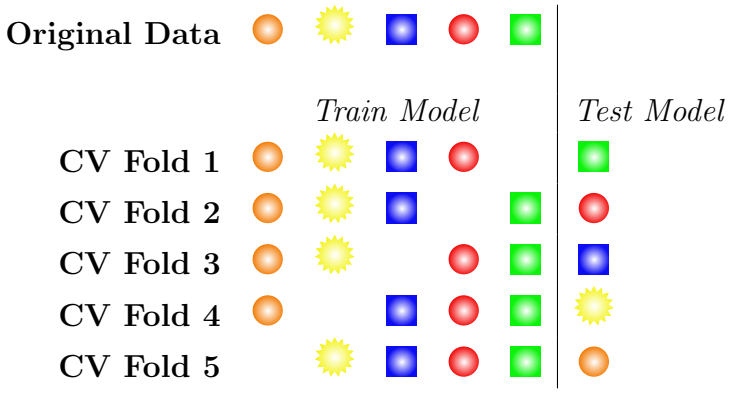

Figure 3.2: Illustration of the 5-fold cross validation approach. The 5 different figures separately illustrate the full sample, while the 5 fold individually illustrates the 5 equally large folds the full sample is split into.

The final result is a mean error estimate calculated from errors between predictions and observations in the five different folds. The benefits of the method is the minimal dependency it puts on how the observations in the data is split. Hence, the resulting predictive power is less reliant on time-specific traits in the data and more reliant on the generalizability of the model to all parts of the sample.

3.2.4 Metric for measuring predictive power

To evaluate predictive power it is common to choose a metric that summarizes the deviance between observations and predictions. I will use the *root mean squared error* (RMSE). RMSE is often juxtaposed to mean absolute error, MAE. The real difference between the measures is that RMSE penalizes larger deviances more than MAE (Chai and Draxler 2014, 1247). This makes RMSE the most conservative measure of the error rate and is therefore chosen over MAE in the main analyses. The formal calculation of the RMSE is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2} \tag{3.2}$$

where y is the observed duration of either the bargaining or the government for the individual cabinet i, whilst the estimated $\hat{f}(x)$ is the predicted duration for the corresponding cabinet i. The mean deviance between y and $\hat{f}(x)$ is then squared to get one measure of deviance between the predicted and observed outcomes. The model with RMSE closest to zero has the comparatively best predictive power (James et al. 2013, 30). For interpretation, the value of the RMSE is dependent on the scale of the dependent variable. Bargaining duration has a max value of 272, while government duration has 1956 days as maximum. The RMSE values are limited by these maximum values.

3.3 Statistical Model

All the original articles use a duration dependent variable. The property of duration variables is that they measure the time from a start point to an end point, observed or unobserved. Therefore, negative duration is impossible. Nor are the duration variables normally distributed. Choosing an OLS-framework would hence give biased coefficients and also the possibility of meaningless predictions. Additionally, in the government

duration context, cabinets fall because of a constitutionally mandated election. Which means that "some observations may have left the sample before their termination mode can be observed" (Diermeier and Stevenson 1999, 1059)[6]. The OLS-framework is not adept to tackle these censored events. The solution is survival analysis.

One type of survival model is the Cox proportional hazard model. It has been used by all four original articles used in this thesis. The benefits of using the Cox proportional model compared to other approaches is the restricted amount of assumption one has to make. Underlying a distribution of duration times is a failure probability. This is called the baseline hazard. In parametric survival models one has to assume a distribution of the failure times, i.e. if the risk of failure increases, decreases or is the same along the duration period. The Cox model has the baseline hazard incorporated, but does not explicitly model it. The drawback is that a substantial interpretation of the effects of predictors require that the different levels of the predictors gives proportional hazards, or have the same effect across the duration distribution. If the effect varies with time it will be difficult to interpret the failure rate between the mean observations between groups of observations.

Another drawback from using the Cox model is the implicit baseline hazard in the Cox model. This implies that the model does not yield an intercept, and hence that making predictions are somewhat unavailable. The benefit of using a parametric model over the semi-parametric Cox model is that the parametrization of the baseline hazard makes it possible to make substantially meaningful predictions.

To be able to use parametric models rather than the semi-parametric approach, both approaches must be shown to give substantially the same conclusions. I have used the Weibull parametrization of the baseline hazard. The Weibull model assumes that the baseline hazard can both be increasing and decreasing over time (Box-Steffensmeier and Jones 2004, 25). The comparisons is made in the tables in the appendix, where I show both the Cox results from my data versus the Weibull results using my data. To show how my data corresponds with the data used in the original articles, I will therefore use the parametric Weibull model to show the estimated coefficients and the substantial effects.

As mentioned above, there are difficulties interpreting the substantial effects of a Cox model. The accelerated failure time estimates from parametric models is more intuitive

[6]In the bargaining duration literature censoring is not an issue. There is always an end result, i.e. a cabinet always forms, eventually.

because the specification uses log of time as the response variable (Box-Steffensmeier and Jones 2004, 26-27). It means that a negative coefficients is interpreted as decreasing survival time, while a positive coefficient is interpreted as increasing survival. Therefore, I will use the accelerated failure time specification of the Weibull model instead of the proportional hazard rate given by the Cox model.

CHAPTER 4

Predicting Government Formation

In this chapter I will assess the predictive power of the government formation duration models presented in section 1.3. The first step will be to replicate the theoretically most relevant results. This will serve to show that the data I have gathered are comparable to the data used in the original articles. A weakness in the government formation literature is the lack of illustrating the full range of substantial effects of the independent variables (Golder et al. 2012a, 249). I will make up for this by showing the substantial effects of the theoretically most interesting variables from the original models. The substantial effects are calculated as predicted bargaining durations in a fixed context, where only the value of the covariate in interest is varied.

Secondly, the in-sample predictive power of the original results will be calculated and illustrated. The in-sample results are interpreted as an independent evaluation. In addition the results are used as the baseline against which the out-of-sample and the cross validation predictions will be compared regarding better or worse predictive power.

Thirdly, I will perform the out-of-sample analysis. Here, I will attempt to predict the new data using the original models. The guiding question is to what extent the models can generalize their claims to new cabinets formed after 1990, or after 1998 for Sona Golder.

Finally, I will further validate the original models by using the 5-fold cross validation method. This chapter ends with a summary of the results presented in this section. The main message will be that both of the formation duration models evaluated experience a drop in their predictive performance when faced with new data, compared to when tested on the original sample.

4.1 The Information Uncertainty Approach: Diermeier and van Roozendaal (1998)

The Duration of Cabinet Formation Processes in Western Multi-Party Democracies by Daniel Diermeier and Peter van Roozendaal (1998) was the first article to investigate bargaining delays in parliamentary democracies by using a cross-national quantitative analysis. In legislative studies, bargaining duration was used as a proxy for political crisis - the longer the duration of the government formation process, the higher the level of crisis in a political system (see for ex. Strøm (1990)). Diermeier and van Roozendaal saw this gap and moved bargaining duration from the right to the left side of the equation. The main concern in the article was to find evidence for the theoretical expectation that higher levels of information uncertainty among the political actors involved in the bargaining lay the ground for longer formation delays. The chosen indicators for information uncertainty was post-election status and the mode of termination for the previous cabinet (Diermeier and van Roozendaal 1998, 620).

4.1.1 Original results

Diermeier and van Roozendaal (1998) uses bargaining formation level data from 13 Western European countries in the period between 1945 to 1989. This amounts to a total of 304 individual formation processes. The authors use the Cox proportional hazard model to estimate four different specifications. In this thesis, the reduced model, model 3, in Diermeier and van Roozendaal (1998, 625) is chosen for evaluation. This model is both the model which is singled out by the authors, and it is also the most parsimonious model of the different specifications.

The replication of the original using my data, and using the Weibull model as discussed in chapter 3, is presented in figure 4.1[1]. Negative coefficients are interpreted as leading to shorter survival, while a change in value in a positive coefficient is interpreted as contributing to longer survival. The coefficients are quite close to the original effects, except for the identifiability variable, the nature of which is discussed in section 3.1.3. The substantial effects of the information uncertainty indicators are identical.

[1]The comparison between Cox and Weibull, as well as the original results, are shown in table A.1 in the appendix.

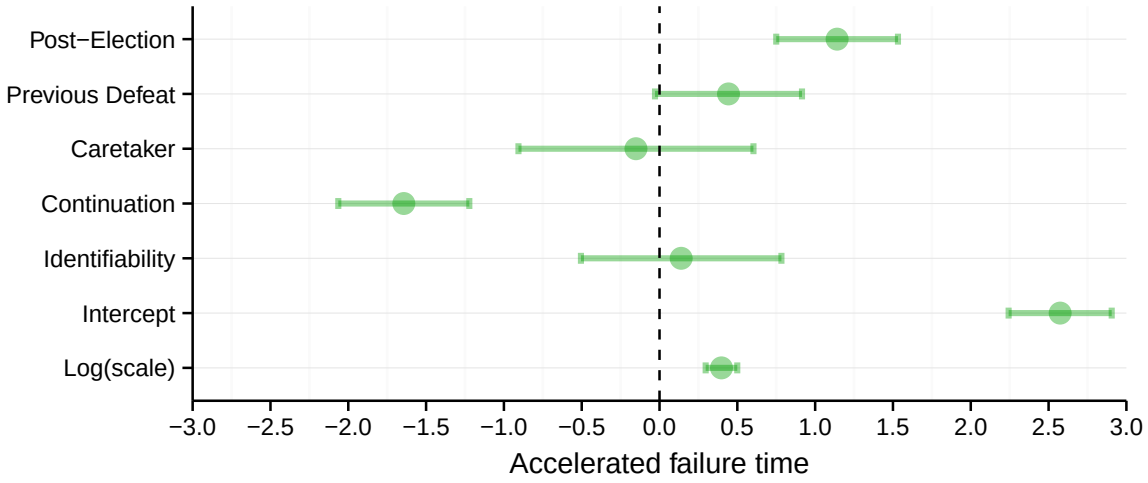

Figure 4.1: Coefficient plot showing the coefficients from the replication of model 3, Diermeier and Roozendaal (1998, 625). The dashed line represents zero effect.

The differences can be due to the loss of observations from using the identifiability and the previous defeat variables. The full sample from my data, using countries and periods as described in Diermeier and Stevenson (1998, 619), consists of 318 cabinets. The identifiability variable drops 35 of them, whilst the previous defeat variable drops 23 observations. Combined, the two indicators drop 48 observations, leaving the total original sample to 270 non-missing observations. Additionally, differences can also be traced by how they have counted cabinets. However, the authors do not specify their sample, and it is therefore possible that there exists some definitional differences between how the cabinets are counted in ERD and in the data used in the original article. I move on by showing the substantial effects.

Formation bargaining in a post-election context yields longer bargaining duration than formation bargaining in an inter-election context. A formation bargaining preceded by a parliamentary defeat have longer formation delays than a formation bargaining where there were other reasons for cabinet termination. These two findings are replicated in the data I have collected. This gives support to the main argument in Diermeier and van Roozendaal (1998) that more uncertainty leads to longer bargaining situations. Additionally, by making the replication as close as shown in table A.1 in the appendix, the reliability of the findings in the article is strengthened.

The effects of the theoretically most important predictors are illustrated in figure 4.2. It shows the predicted bargaining duration on the two levels of the two uncertainty

indicators. Post-election context, coded as 1, increases government formation duration by around 25 days [13-42][2], compared to an inter-election context, coded 0. The difference between the effect of inter-election and post-election context is highly significant.

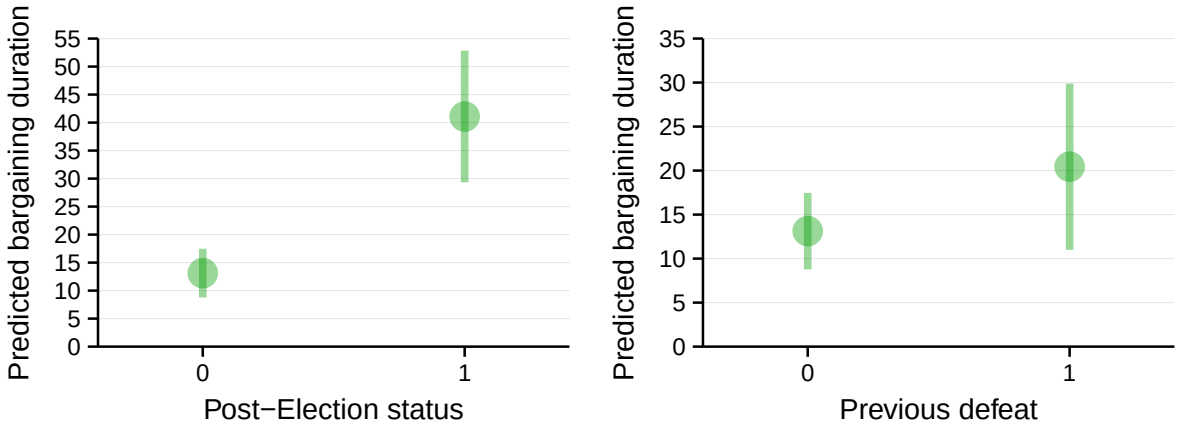

Figure 4.2: Predicted bargaining duration from different values of the uncertainty indicators from Diermeier and van Roozendaal (1998).

A bargaining context with a previous defeat gives predicted point estimates indicating an increase in bargaining duration. However, the effects are not statistically separable, as indicated by the crossing confidence intervals in the right panel of figure 4.2. The substantial effects have been shown to be identical with the original results.

The predictive ability of the bargaining duration model by Diermeier and van Roozendaal (1998) is estimated without the identifiability variable. Firstly, the identifiability variable I have used introduces a substantial amount of missing values in the data ranging from 1989 up until today. Secondly, there is some confusion as to how the original identifiability variable has been coded, and hence, there is also some confusion of what the variable actually measure. The data comes from Strøm (1990), which codes the identifiability of pre-electoral coalition alternatives "impressionistically as low(0), medium(.5) and high(1) on a decade-to-decade basis" (1990, 73). It would be hard to code the new data following this guidance. Table A.1 in the appendix shows exactly how the effect of the indicator varies. When the pre-electoral agreement variable is used, the effect goes from being a small positive coefficient to a small negative coefficient. After checking the importance of the identifiability predictor, both in terms of estimates and in-sample predictive ability, I find that the differences are minimal. Hence, identifiability is excluded from the evaluation of the predictive power of the uncertainty approach.

[2]Highest and lowest estimate.

The next step is now to test how well the information uncertainty model predicts the duration of the original sample.

4.1.2 In-sample prediction

Figure 4.3 shows the density of the observed and predicted values from the in-sample evaluation of the reduced model in Diermeier and van Roozendaal.

One trend is evident. The predicted values are mainly represented by three spikes. The largest spike comes at around 10 days. This corresponds to the largest density for the observed values. This is positive - the model is able to predict the largest amount of observed duration. However, the density areal is smaller for the observed than the predicted duration. The second largest bulks up around 40 days, and the smallest spike comes after around 65 days of bargaining.

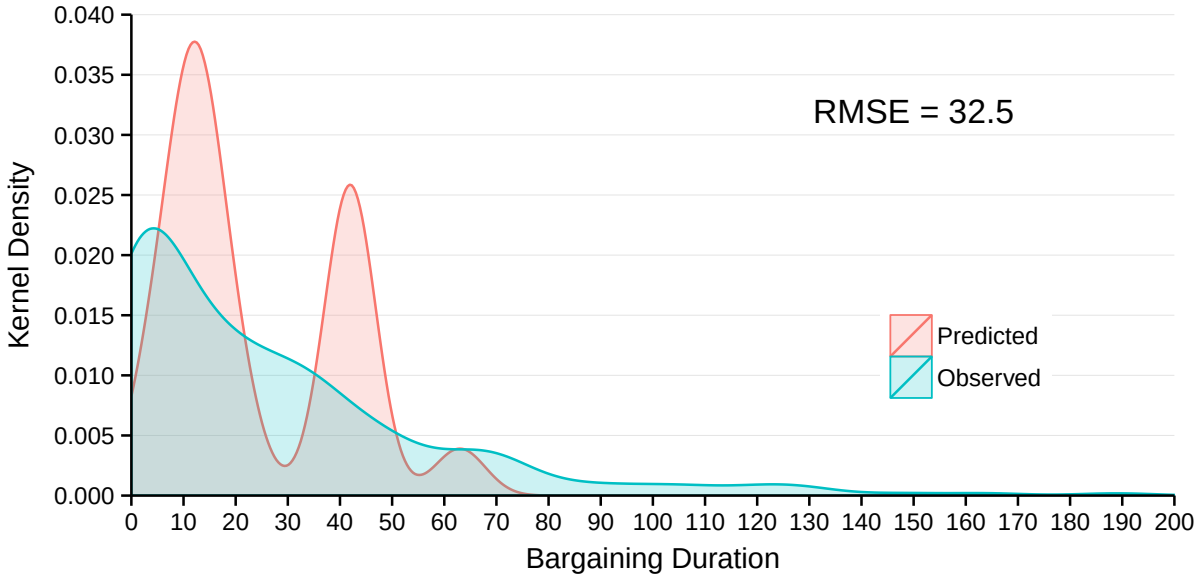

Figure 4.3: In-sample predictive accuracy from the reduced model in Diermeier and van Roozendaal (1998, 625). The y-axis indicates the kernel density. Kernel density is an improved alternative to histograms for illustrating the distribution of continuous variables. The x-axis shows bargaining duration in days.

The main trends are over-estimation of the minimum values, over-predicting mean values and not predicting durations longer than 70 days. This is further shown by the RMSE of 32.5 days average prediction error, weighted on individual deviances between prediction and observation. A statistical explanation for the spike-trend and the high RMSE can be found in the binary nature of the predictors used by Diermeier and van Roozendaal (1998) to explain variation in bargaining days. Every predictor is a dummy,

as shown in table 4.1. Hence there is a limit as to the range of values the model can predict. The binary nature of the model prohibits the predictions to follow the more natural distribution of the observations, and the consequence is large deviances. These large deviances, then, lead to the inflated RMSE score discussed above.

Table 4.1: Descriptive statistics - Diermeier and van Roozendaal (1998)

Variable	**No**	**Yes**
Post-election	144	151
Previous Defeat	239	56
Caretaker	277	18
Continuation	231	64
N		295

Therefore, the in-sample predictions show evidence of a model not able to follow a natural distribution due to the binary nature of the predictors chosen. This trend is likely to be shown in the next section. However, will the model be able to show signs of generalizability to the new data? The next step will be to evaluate the information uncertainty model out-of-sample.

4.1.3 Out-of-sample prediction

The in-sample evidence pointed to poor predictive ability of the information uncertainty approach. Figure 4.4 shows how the model performs out-of-sample. The figure is based on predictive accuracy on 96 western European cabinets from 1999 to 2015. The trends described in the previous section are nearly replicated when the model is tested on new cabinets. However, in figure 4.4, there are two spikes, compared to the three spikes that was shown for the in-sample predictions. As opposed to the in-sample evidence, the largest predicted density does not correspond with the largest observed density.

It is fair to say that the ability of the Diermeier and van Roozendaal (1998) model to predict new data is worse than the ability to predict the in-sample observations. However, the difference between the RMSE values indicates that the in- and out-of-sample differences should not be exaggerated. In-sample predictions miss by 32.5 days whilst the model misses by 36.5 days out-of sample, on average, and where the RMSE is weighted by large deviances. Given that the target is zero, both the in- and out-of-sample predictions miss significantly, but the differences between the two are not that large.

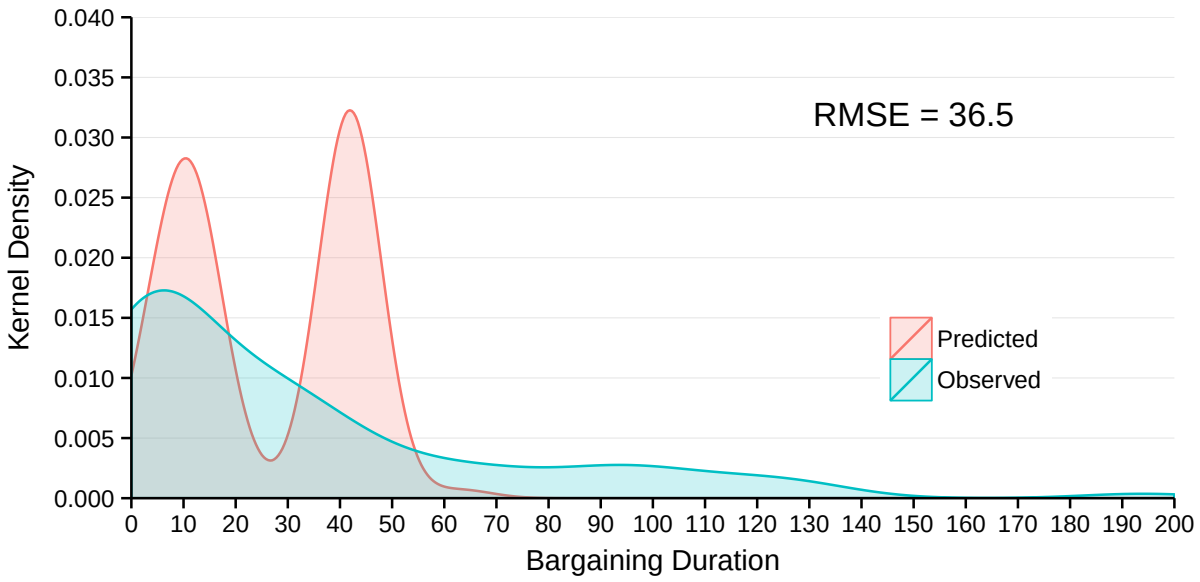

Figure 4.4: Out-of-sample predictive accuracy, Diermeier and van Roozendaal (1998). The density plot is based on observed and predicted values from the test set.

Summed up, the information uncertainty model has explanatory power on data ranging up until 1990, but it does not show an impressive ability of predicting the same observations. The trend continues out-of-sample with an increase in RMSE relative to the in-sample RMSE. An explanation is that the model has pushed modeling simplicity too far to be able to pick up the real underlying trends in the data.

As mentioned in chapter 3, the out-of-sample method utilizing only a training set and a test set is highly sensitive to how the data is split. Since the split has been chosen from a possible confounder, time, it makes sense to further evaluate how the model performs when the full sample is split to 5 random folds. The next step uses the 5-fold cross validation to search for a more broad answer to the generalizability of the model.

4.1.4 Cross validation

Having established that there exist a difference in predictive power in-sample and out-of-sample, the last step is an evaluation of the generalizability of the Diermeier and van Roozendaal model which is less dependent on time, and hence, more concerned with how well the model picks up what it tries to model. Figure 4.5 presents the results of the cross validation analysis. The first thing to note is that the RMSE value for the 5-fold cross validation analysis is closer to the RMSE value in-sample than the RMSE out-of-sample. This indicates that the model predicts better when re-fitted on random

samples than when the data is split to before and after 1990.

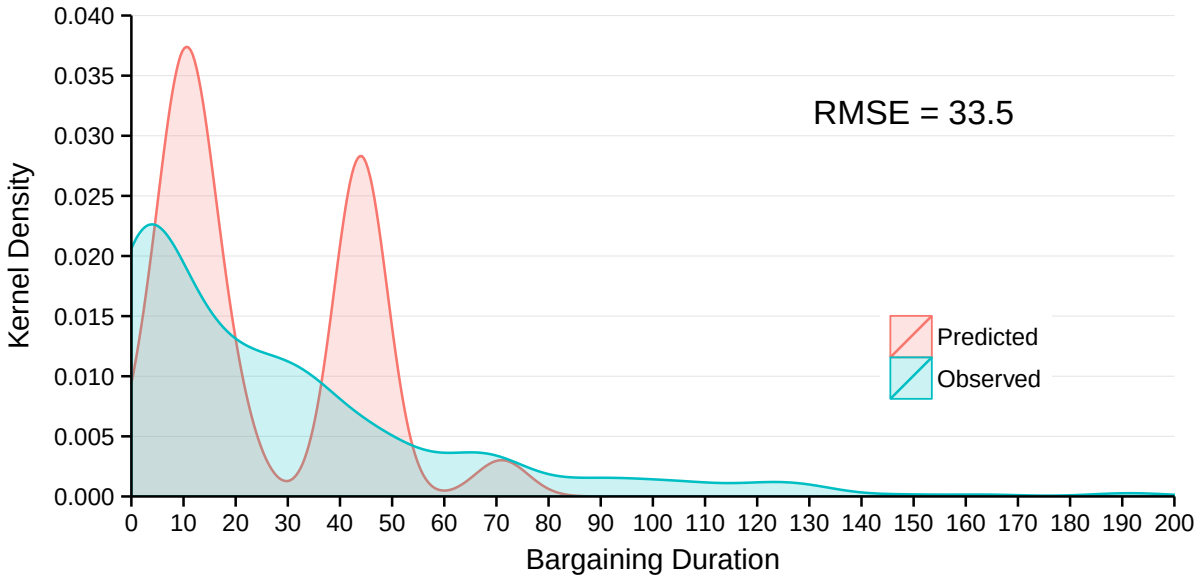

Figure 4.5: Predictive accuracy from 5-fold cross validation, Diermeier and van Roozendaal (1998). The density plot is based on observed and predicted values from splitting and predicting on 5 different folds of the full data set.

A second point is that the density plot shows an almost identical distribution of predicted bargaining durations compared to the in-sample results. This gives another indication that the predictive power of the model is reduced out-of-sample, but when all observations are randomly folded and predicted, the results indicate the same trend as the in-sample predictions. This also means that over-fitting is not a prominent problem of the model when tested on random samples. The results are not only dependent on the original sample when the cross validation RMSE is close to the in-sample predictions [3].

This concludes the evaluation of the more parsimonious approach of modeling one aspect of the government formation process. Next follows the more complex approach, represented by Golder (2010).

4.2 The Combined Uncertainty and Complexity Approach: Golder (2010)

Bargaining Delays in the Government Formation Process by Sona Golder (2010) is a unification of two competing theoretical explanations of the observed deviance from the Baron-Ferejohn bargaining model (1989). Following the argument in Golder (2010),

[3]There are some variance in the RMSE for the 5 different folds, meaning that some folds can consist of harder predicted observations than other folds. Hence, folds with harder observations will also have higher RMSE. The RMSE given here is the mean of the 5 different RMSE's. The five different RMSE measures are shown in figure A.1 in the appendix.

the two approaches should be empirically validated by intertwining the theories. The information uncertainty and the bargaining complexity approaches have been thoroughly covered in chapter 2. The argument in Golder (2010) is that both approaches are valuable. However, the effect of complexity is contingent on the degree of information uncertainty among actors involved in the government formation bargaining. This means that bargaining complexity only have an effect on bargaining duration when the actors involved in the bargaining are uncertain about the preferences and strategies of the other actors. Therefore, the main empirical, and theoretical, focus for Golder (2010) is to interact the information uncertainty indicator, post-election status, with the bargaining complexity indicators, such as the effective number of parties, the ideological polarization in the legislature and the presence or absence of the investiture rule.

4.2.1 Original results

Sona Golder (2010) uses data from 16 Western European countries from 1944 to 1998. This gives a total of 383 observations of government formation bargaining. The dependent variable is bargaining duration, measured in days between the date of previous government's termination and the date when the new government is invested. I have replicated model 4 from table 2 in Golder (2010, 20-21). This model is the only one which incorporates the contingent theoretical account, interacting the uncertainty variables with complexity indicators.

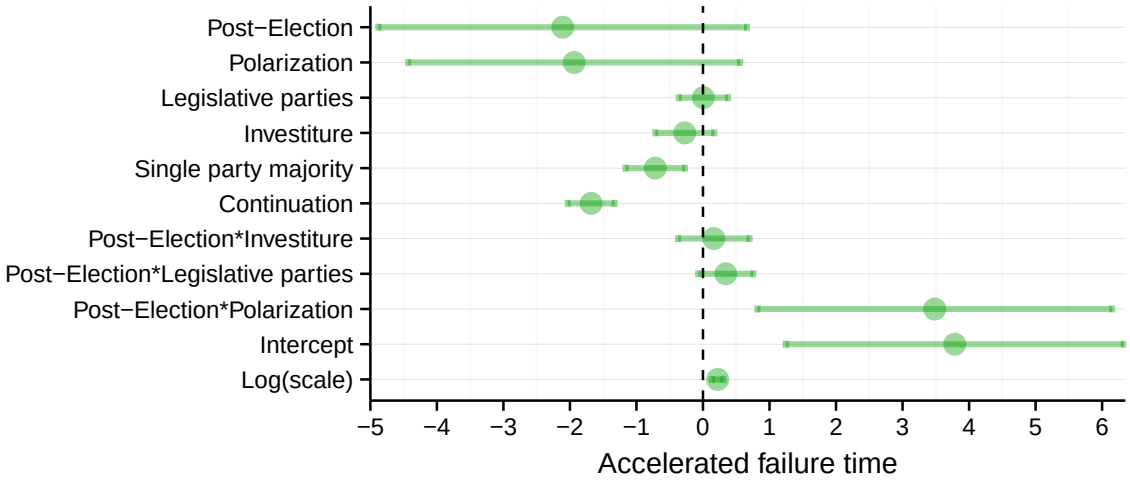

Figure 4.6: Coefficient plot showing the coefficients from Golder (2010). The dashed line indicates no effect.

The replication is shown in figure 4.6. The results illustrated are accelerated failure

time estimates from the Weibull model, estimated on the original sample. As in the previous section, negative estimates yield shorter survival times and positive estimates indicate longer survival. The replicated results are almost identical to the results in Golder (2010, 20-21)[4]. The effects in figure 4.6 are almost impossible to interpret from the coefficients alone, due to the three interaction terms included in the model. This means that an insignificant coefficient in figure 4.6, i.e. a effect with confidence intervals crossing zero, does not necessarily imply no explanation power. It only means that the effects are dependent on the value of the other predictors in the model.

Figure 4.7 shows the effects of the variables that are interacted in the combined model. The effects of the uncertainty indicator, post-election status, and the three bargaining complexity indicators are illustrated. The substantial effects are illustrated by using predicted bargaining duration, calculated by holding the remaining dummy variables at their zero value and the continuous predictors at their mean.

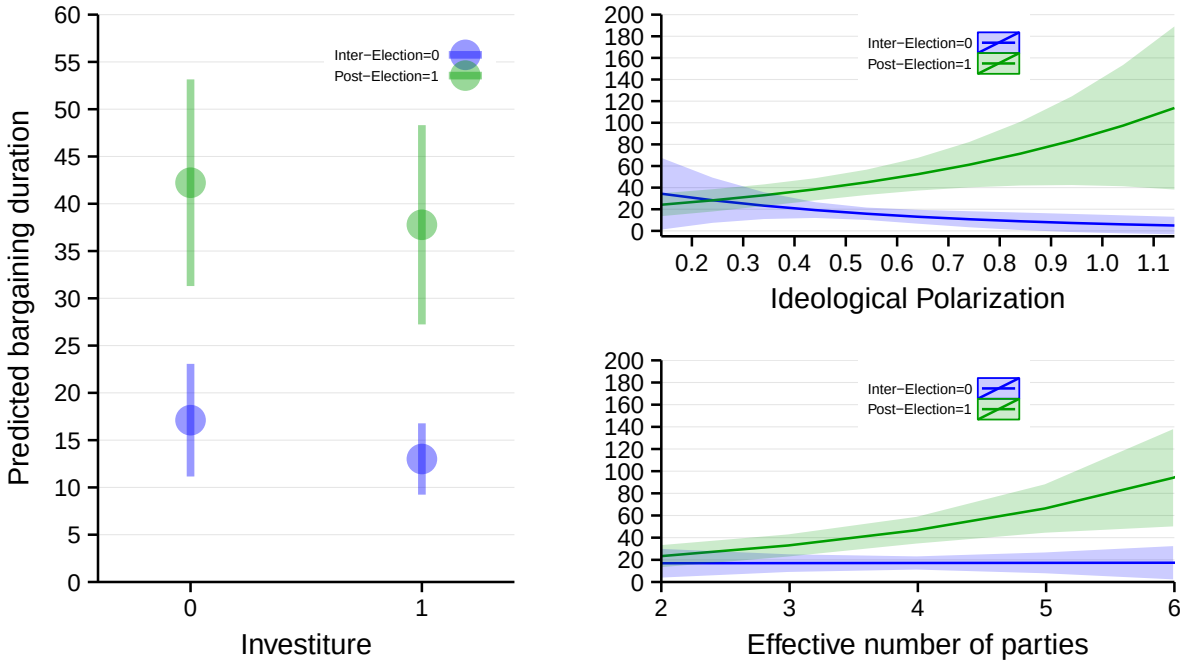

Figure 4.7: Illustrating the theoretically most interesting interaction effects from Golder (2010). All y-axes indicates predicted bargaining duration, measured in days.

The influence of positive parliamentarism is shown in the left panel of figure 4.7. This replicates the finding from Golder (2010, 23). Having the investiture rule does actually reduce the predicted length of bargaining. That is, countries practising positive

[4]See table A.4 in the appendix for the comparison between the Cox and Weibull models, as well as the comparison between the original results and the replication using my data.

parliamentarism are estimated to reduce the length of their formation duration. However, the effects are not statistically different across the two levels of uncertainty. Even though having positive parliamentarism is theoretically expected to increase complexity, and hence bargaining duration, Golder points out that only in minority situations are the effect realistic (2010, 24).

A post-election context interacted with higher levels of polarization in the legislature yields significant differences in predicted bargaining duration. And, as argued in Golder (2010, 15), inter-election periods have constant or decreasing effect. This is in line with the estimated effects of ideological polarization, shown in the top figure in the right panel of figure 4.7.

The same significant difference is found when post-election context is interacted with the number of effective parties in the legislature. The more parties in the legislature the longer the predicted bargaining duration. For inter-election contexts, the effect of the effective number of parties is basically constant. This indicates that effective number of parties has significantly different effects depending on the degree of information uncertainty.

I have shown that the substantial results using my data are the same as the findings in Golder (2010). The next step of the evaluation is in-sample predictions, where the model as a whole is tested for how well it predicts bargaining duration.

4.2.2 In-sample prediction

The in-sample predictions plotted in figure 4.8 show a model which is comparatively better at predicting bargaining duration than the pure uncertainty approach shown in section 4.1. Similarly to the in-sample evidence from the information uncertainty model, the highest density of predicted bargaining durations comes at around 15 days. While the in-sample predictions from the information uncertainty model bulked up to two and three spikes, the in-sample predictions for the combined approach follow the distribution of the observed duration more closely. In addition, the combined approach is able to predict the largest bulk of durations. The largest deviance comes around 40 days, where the observed distribution monotonically decreases while the predictions shows a sudden increase.

Another indicator of better predictive ability in-sample is that the RMSE of 9 days indicates a decrease in average error relative to the in-sample evidence from the

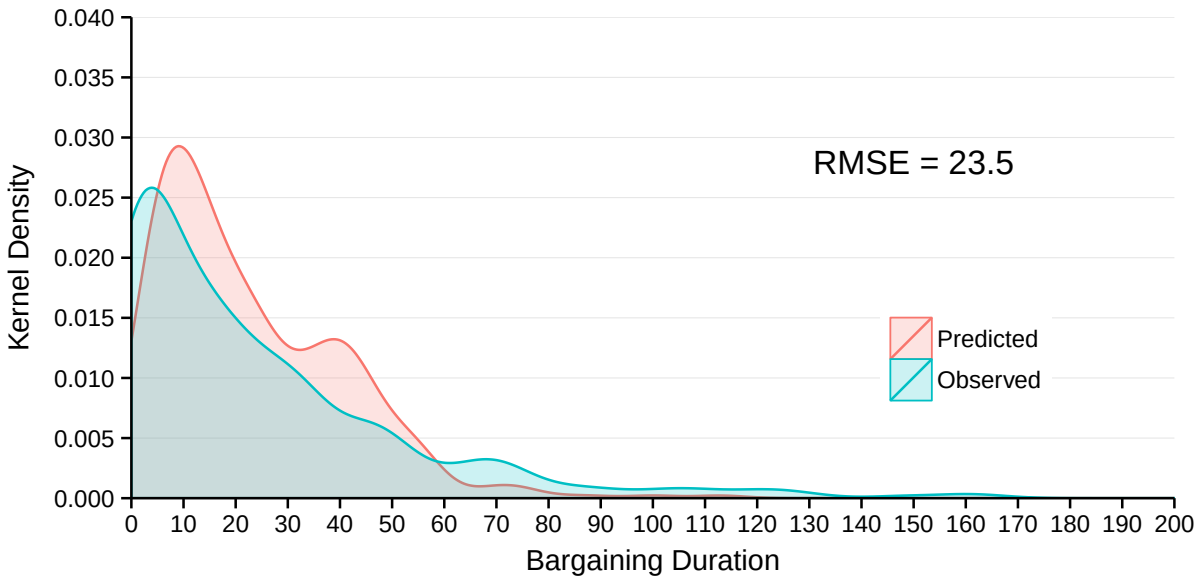

Figure 4.8: In-sample predictive accuracy, for the combined model in Golder (2010, 20-21). The x-axis gives bargaining duration in days.

information uncertainty approach. This means that the combined model is a better model to the original sample than the uncertainty model, measured by in-sample predictive power.

4.2.3 Out-of-sample prediction

The in-sample evidence indicated that the model could be able to both predict and explain. The out-of-sample evidence shown in figure 4.9 is made from the predictions and observations of 70 Western European cabinets from 1999 to 2015. The figure illustrates a model less adapted to predict new cabinets compared to how well it predicted the observations in-sample. Here, the model predicts the highest density to be around 25 days, while the highest observed density comes at around 10 days.

However, the model is still able to follow the observed durations from 45 days and onward. This clearly separates the combined approach from the uncertainty approach, which only predicted durations up to around 50 bargaining days. The out-of-sample predictions show a small bump at 140 days, which indicates that the combined model is better able to follow the bargaining duration distribution.

Still, the out-of-sample evidence shows a model yielding predictions worse than the in-sample evidence. The RMSE of 38 days mean deviance as compared to the in-sample RMSE of 23.5 days, illustrate the poorer ability of the combined model to predict new

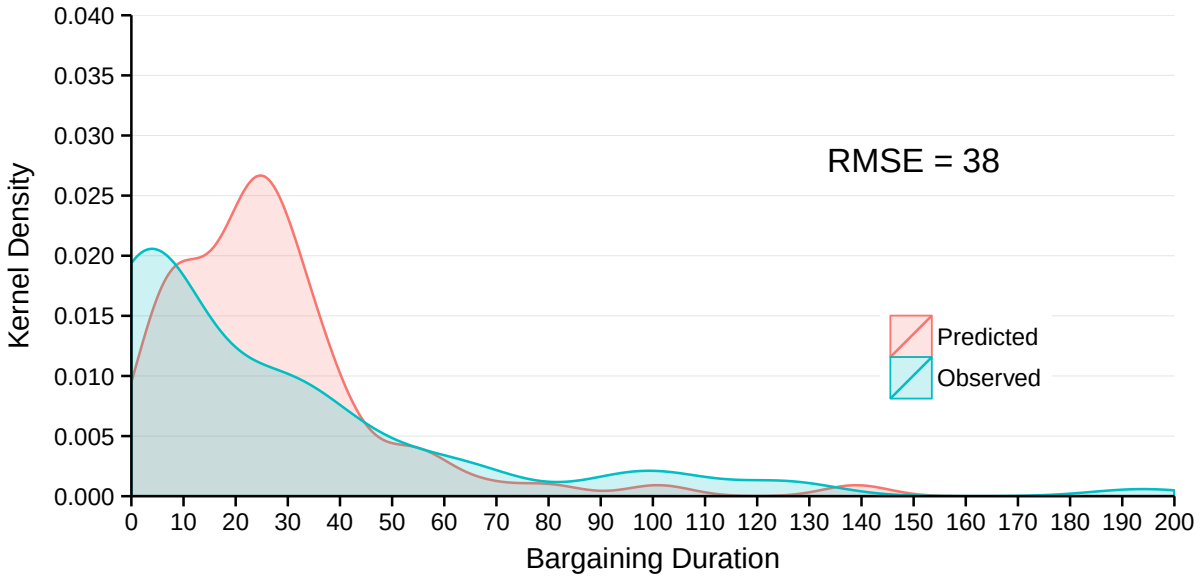

Figure 4.9: Out-of-sample predictive accuracy, Golder (2010). The density plot is based on observed and predicted values from the test set.

observations. One reason for the increase in RMSE by 14.5 days out-of-sample is the choice of RMSE as metric for evaluating predictive power. The metric punishes large deviances comparatively harder than other measures of prediction error. In effect, this means that the out-of-sample evidence from the combined approach consists of larger individual errors than in the information uncertainty approach.

4.2.4 Cross validation

After splitting the data to 5 folds and predicting the outcome in the left-out fold, the RMSE shows a decrease from 38 days out-of-sample to around 25.5 days. The trend equals that of the uncertainty approach from Diermeier and van Roozendaal (1998) - higher RMSE out-of-sample than in-sample, while cross validation yields a RMSE comparatively closer to the in-sample predictions.

The main interpretation of this section is that when the prediction error rate is tested by randomly dividing the observations into 5 folds, the results show similarities to the in-sample results. Hence, the combined information uncertainty and bargaining complexity model is better adapted to explain general trends in the data compared to testing the model on specific time periods[5].

[5]Similarly as in the information uncertainty approach, the RMSE varies between the 5 folds. The RMSE's are shown in figure A.1 in appendix A.

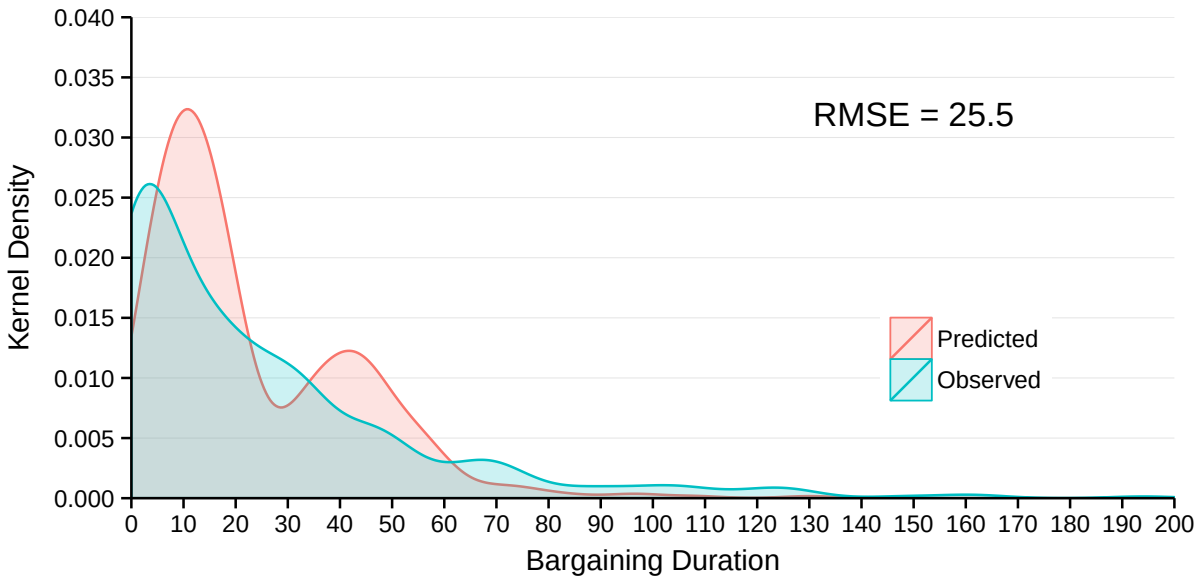

Figure 4.10: Predictive accuracy from 5-fold cross validation, Golder (2010). The density plot is based on observed and predicted values from splitting and predicting on 5 different folds of the full data set.

Therefore, as with the information uncertainty approach, the cross validation of the combined model does rather well when compared to the in-sample approach and measured with RMSE. Because the model is able to generalize as well on a random sample as to the original sample, this evidence demonstrate that the combined model show fewer sings of being over-fitted to the original sample.

4.3 Summarizing Predictive Power in Government Formation

In concluding this chapter it would firstly seem evident that the theoretically, and empirically, most complex approach to the prediction of bargaining duration gives the best predictive power. The combined approach from Golder (2010) predicts best in-sample. The out-of-sample evidence shows that the information uncertainty approach has the lowest RMSE of 37.5 as compared to 38.5 for the combined model. However, the graphical illustration shows that the combined approach is better able to predict a more natural distribution of bargaining duration than the uncertainty approach which consisted of large bulks.

Secondly, both models perform worse out-of-sample than in-sample. This means that none of the models fitted on data from 1945 to 1989 generalizes well to new data. This indicates that the empirical models might be over-fitted. However, this is nuanced by the cross validations. These results are close to the in-sample predictions. This can be

interpreted as an ability of the model to generalize its results not only to the specific original sample but also when tested on random folds.

Thirdly, the substantially relevant results of the original articles have been replicated. Information uncertainty has a significant impact on bargaining duration, but the combined approach which incorporates the information uncertainty approach, has also done so. This means that the information uncertainty approach might suffer from omitted variable bias.

CHAPTER 5

Predicting Government Duration

In this chapter, I will evaluate the predictive performance of two government duration models, Warwick (1994) and Diermeier and Stevenson (1999). I will apply the identical procedure as in the previous chapter - firstly, showing how the theoretically most important substantial effects, using my data, are similar to the corresponding effects from the original results. Secondly, estimate the in-sample predictive performance. Thirdly, calculating prediction error for the original model out-of-sample, using the data from 1990 to 2015. And, finally, applying the cross validation method to the full sample.

5.1 The Importance of Ideology: Warwick (1994)

Government Survival in Parliamentary Democracies by Paul Warwick (1994) introduced ideological indicators to the study of government duration. The main theoretical argument was that increased level of policy disagreement within the cabinet would causally correlate with lower government duration. The rationale is simple - the bigger the distances on policy-dimensions between cabinet parties lead to more conflict, and higher levels of conflict is a mechanism which contributes to earlier termination.

5.1.1 Original results

Warwick (1994) uses data from 15 Western European countries in the period between 1945 to 1989. The original data set consist of 374 cabinets. However, Warwick lacks ideological data on Finland, Spain, Portugal and Iceland (1994, 55). These countries are therefore dropped in the analysis, resulting in 284 non-missing observations[1]. The

[1] I have done the same with my data. The drawback is reducing observations in the new data. However, it makes the results more comparable to the original empirical model.

dependent variable is the duration in days between the date of formation and the date of termination. Cabinets are censored in situations of a constitutionally mandated election, technical failures such as the death of a prime minister or cabinets that have yet to experience termination[2]. Warwick (1994), because of the nature of the dependent variable, applies the Cox proportional hazard model. I will evaluate model 7 (Warwick 1994, 59) because it includes three ideological indicators, which are found to have the most explanatory power. The model estimated using my data is shown in figure 5.1. The coefficients are accelerated failure time estimates from the Weibull specification[3].

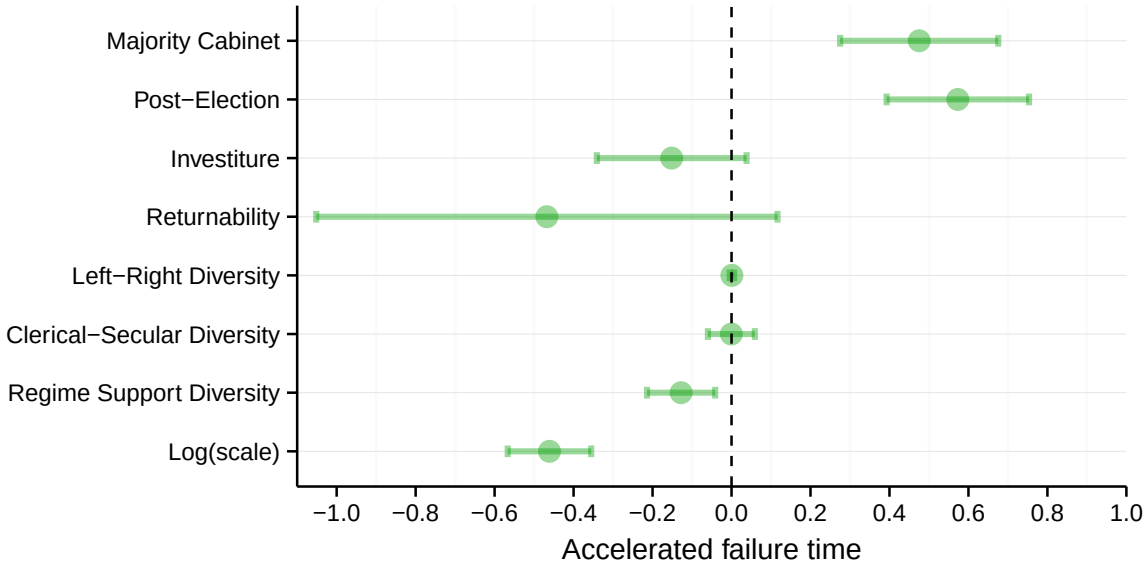

Figure 5.1: Intercept: 6.53[6.10-6.96]. Coefficients from original article, Warwick (1994). The dashed line represents no effect.

A note to the replication of the original results is the remarkable uncertainty estimates for the returnability variable. As explained in the data section, chapter 3, I have calculated this measure using data from Chiba et al. (2015). One possible explanation is that returnability is a variable which have been calculated in different ways. Hence, it is hard to find out how the original variable is coded, and therefore the replication is made harder.

The main theoretical point in Warwick (1994) is the importance of the policy-seeking assumption for the explanation of government duration. In figure 5.2 I show how the

[2]For Warwick (1994) this was the cabinets that had not terminated as of 1989. For my data, cabinets are censored if the have not been terminated by May 2015.

[3]The comparisons between Cox and Weibull, as well as the comparison between the original and the replication, are shown in table B.1 in the appendix.

ideological indicators affect the predicted government duration. The only statistically significant ideological predictor is that higher degrees of diversity on the regime support dimension decreases the duration of cabinets, which is consistent with the argument in Warwick (1994). As seen in the left panel, changes in the regime support variables leads to significant decreases in predicted government duration. However, the right panel illustrates the zero-effects both from the left-right diversity and clerical-secular predictors.

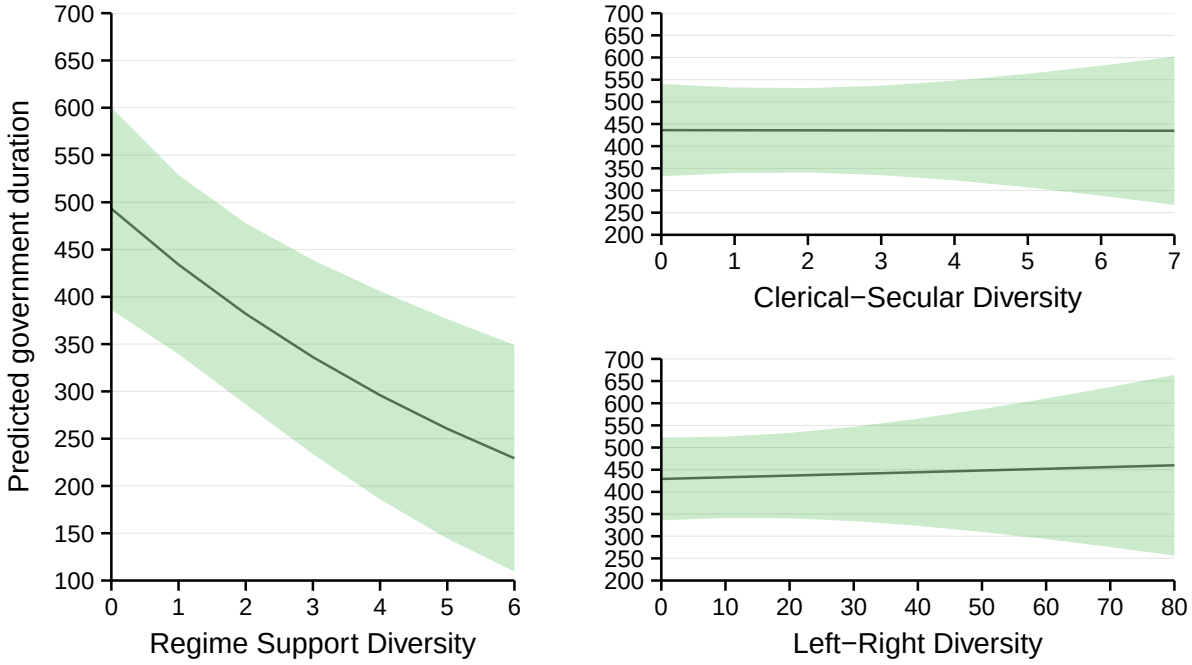

Figure 5.2: Effect of ideological indicators on predicted government duration, Warwick (1994). Y-axes indicate predicted government duration, in days.

Warwick (1994, 57) reports a small positive proportional hazard estimate of both the left-right and the clerical-secular dimension. This indicates that higher levels of left-right diversity lead to shorter expected government survival. The results in figure 5.1 show that the left-right dimension has as close to zero-effect as possible. The same goes for the clerical-secular dimension. In Warwick (1994, 57), the effect of the clerical-secular dimension is positive, indicating decreasing survival for cabinets with higher levels of clerical-secular diversity.

One possible explanation of this finding is that I have coded the clerical-secular and regime-support indicators with CMP data (Volkens et al. 2014). Warwick's ideological data comes from several expert surveys and using scales from Dodd (1976). The left-right cabinet ideological diversity measure from the ERD data comes also from CMP. This measure has been shown in other studies as well to give the same zero effect for the

general left-right measure - in Chiba et. al. (2015, 54) and in Saalfeld (2008, 340) regarding government duration, and De Winter and Dumont (2008, 149-150) regarding bargaining duration.

A second explanation can be that the regime support dimension steal effect from in particular the left-right measure. The general left-right measure is supposed to inhibit different sub-dimensions. When regime support is the strongest predictor, one explanation could be that ideology only has effect on a sub-dimensional level. Another explanation is that regime support takes effect from the general measure. However, when I exclude the regime support predictor, the result stays the same - no effect from the general left-right ideological, and the effect of the clerical-secular dimension is constant. It is clear, therefore, that the effect of the left-right and clerical-secular dimensions found in Warwick (1994) is not replicable when using manifesto data.

5.1.2 In-sample prediction

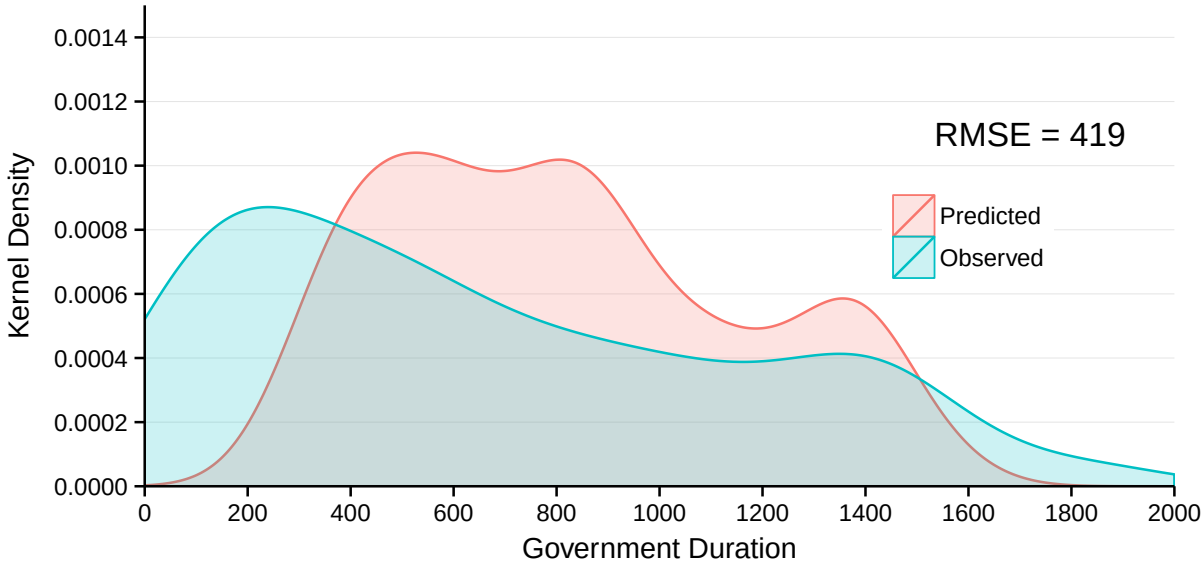

Figure 5.3: In-sample predictive accuracy from model 7 in Warwick (1994, 59). The x-axis represents government duration measured in days.

The in-sample predictions in figure 5.3 show that the ideological model delays the predicted peak of terminations to 500 days, whereas the observed durations peak around 200 days. The observed risk of termination increases as the number of days increases from around 1200 days. The model predicts the same slight bump at around 1400 days. These two observations indicate that the ideological model is able to roughly follow the main

trends in the observed cabinet duration data.

However, between 400 and 900 days the prediction errors are quite substantial. This is also shown in the RMSE of 419, which means that on average the ideological approach misses the observed durations by 419 days, weighted by large individual deviances. In effect, the predictions are one year and almost three months off, on average, for each cabinet in the data from 1945 to 1989. In all, the Warwick model does not predict cabinet duration very well in-sample. This gives support to the claim that explanation power does not need to mean predictive power. The next step is to test the predictive performance of the original model on new data.

5.1.3 Out-of-sample prediction

The out-of-sample evaluation is performed on 92 cabinets from 1990 to 2015. Summed up, the ideology approach does not generalize well to the new sample. Figure 5.4 shows mean deviance between observed and predicted observation of 507 days, which is an 83 day increase from the performance in-sample.

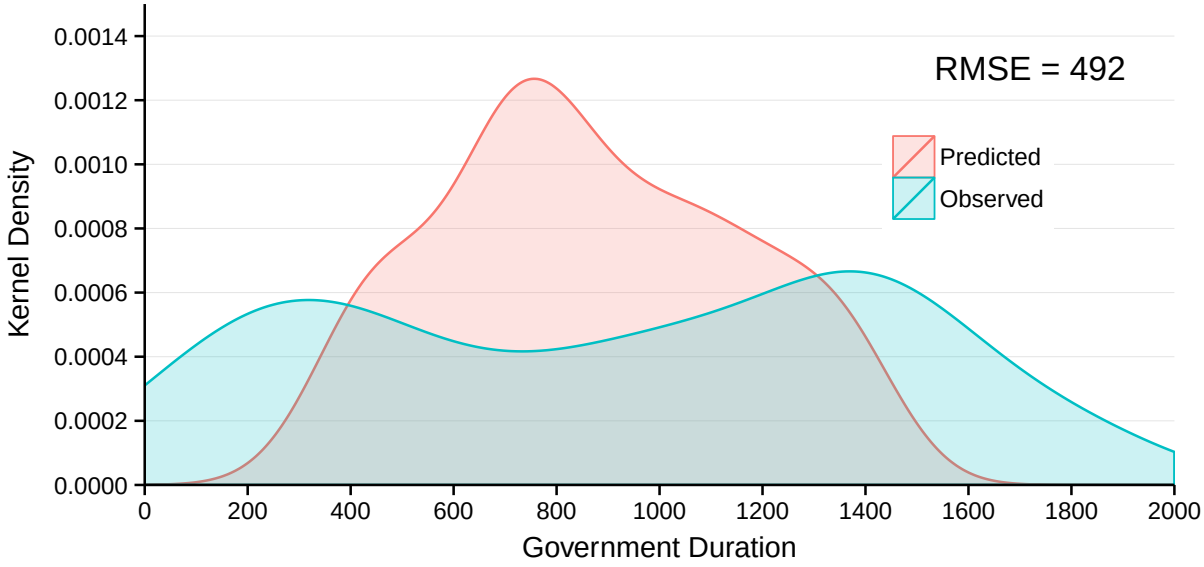

Figure 5.4: Out-of-sample predictive accuracy, Warwick (1994).

The majority of the predictions are concentrated between 400 and 1200 days with a peak after 800 days, whereas the distribution of the observed duration peaks at 1400 days. A noticeable change from the in-sample results is the distribution of the actual duration days in the data from 1990 and up until today. It shows that cabinets formed after 1990 last longer, compared to cabinets formed before 1990. This change represents

a possible explanation for the poor out-of-sample predictive power. Because of changes in government duration before and after 1990, the ideological model predict cabinets more poorly out-of-sample.

This section has shown that when the model is applied on data from 1990 and onwards it shows decreased predictive power compared to the in-sample predictions. This can be a sign of over-fitting, because the model is more adept at predicting the original sample compared to the new sample. If different splits are applied to the data, will it then show better predictive performance?

5.1.4 Cross validation

The density plot in figure 5.5 shows a model less able of predicting low government durations. However, the predictions follow roughly the same patterns as the in-sample results. First, the prediction density shows slight over-predictions around the middle of the distribution. Second, the density peaks at the beginning of the distribution for both observed and predicted durations. However, for the predictions, the peak is delayed by around 400 days compared to the observations.

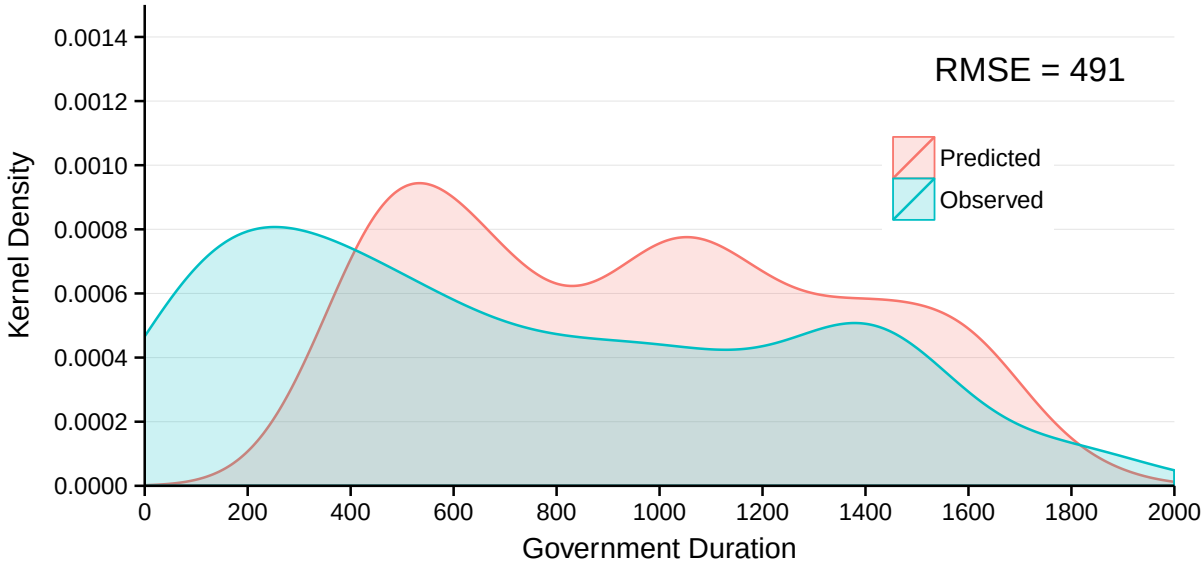

Figure 5.5: Predictive accuracy from 5-fold cross validation, Warwick (1994). The density plot is based on observed and predicted values from splitting and predicting on 5 different folds of the full data set.

Interpreted, the cross validation shows signs of an original model which is dependent on the original sample to be able to predict government duration. When the cabinets are split randomly into 5 folds, the RMSE is lower than for the out-of-sample approach,

but higher than the in-sample approach. Combined, this means that the model is not as generalizable to samples other than the original[4].

5.2 Strategic Dissolution: Diermeier and Stevenson (1999)

Cabinet Survival and Competing Risk from Diermeier and Stevenson (1999) introduces the strategic assumption of government actors in the modeling of government duration. Their main point is that earlier approaches, Warwick (1994) and King et al. (1990) among others, assumes that government actors does not possess the ability making of strategic decisions. After showing formally that a member of the government can benefit from leaving the cabinet, Diermeier and Stevenson (1999) demonstrate empirically that there exist two ways for a government to terminate - one in dissolution and one in replacement. This differs from previous attempts that have been pooling different modes of terminations, as in Warwick (1994). Pooling terminations can therefore be a potential source of biased results.

5.2.1 Original results

The Diermeier and Stevenson (1999) data set covers 15 western European countries as well as Canada and Israel[5]. Diermeier and Stevenson (1999) uses the ideological data from Warwick (1994), mentioned in the previous section. Therefore, the same countries are dropped from their analysis - that is, Finland, Iceland, Spain and Portugal.

The replication of the original results is presented in figure 5.6 [6]. The results point to a close to zero effect on the general left-right measure, both for the dissolution and the replacement models. The clerical-secular diversity indicator shows a negative coefficient in the replacement model, meaning that increased clerical-secular diversity predicts a decrease in government duration. This effect is the opposite in the original model. This can be explained by differences in the sample. Or, as mentioned in the Warwick (1994) evaluation, it can be because my ideological data stems from party manifesto data,

[4]There are some variance in the RMSE for the 5 different folds, depending on if the folds have more or less predictable observations. The RMSE given here is the mean of the 5 different RMSE's. The five different RMSE measures are shown in figure B.1 in appendix B.

[5]Canada and France, 4th republic, are not included in the ERD data set. I have collected data for both countries. I have not found appropriate data for Israel, and Israel is therefore excluded from the analysis.

[6]The comparison between Cox and Weibull and the comparison between the original results and my replication are shown in table B.4 in the appendix.

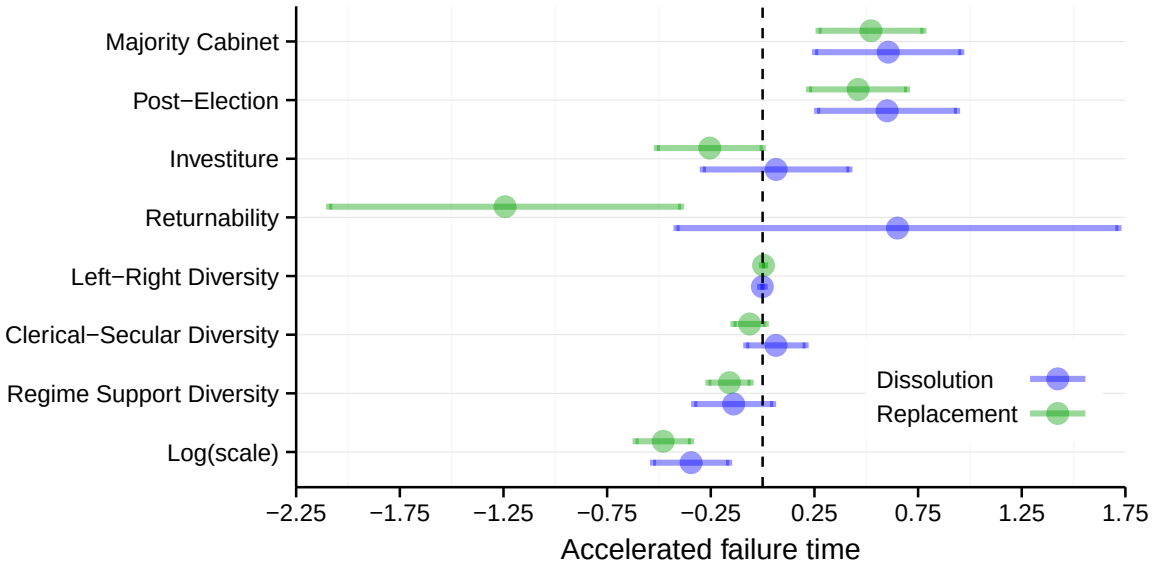

Figure 5.6: Intercept Dissolution: 6.19[5.44-6.94]. Intercept Replacement: 7.50[6.86-8.14]. Coefficient plot from the competing risk approach, Diermeier and Stevenson (1999). The dashed line equals no effect.

while the original data are coded from different expert surveys. Again, the returnability indicator shows unstable and unreliable results. This was shown in the Warwick model and is still the case for the competing risk approach. The inference one can draw is that the effect of returnability is highly sensitive to calculation and coding.

Figure 5.7 shows the effect of the full range of left-right diversity for both modes of termination. The predicted government durations are calculated holding dummy variables at their zero value and the continuous predictors at their mean, while the value of the left-right diversity indicator is allowed to vary. The main message here is the uncertainty in the results. The estimates show a cabinet diversity indicator which decreases and increases the number of predicted government duration days, respectfully for the dissolution and replacement model. However, the effects cannot be trusted on conventional levels of significance.

A potential source of error, which can explain the somewhat different results when using my data, is the coding of dissolution and replacement terminations. Diermeier and Stevenson (1999) used an automatic coding procedure for how the cabinet ended. They calculated the remaing time until the constitutionally mandated election for each cabinet, and recorded a dissolution if an election happened before the period had ended. Replacements were coded for cabinets which did not directly follow an election. Technical failures, cabinets lasting until 6 month before an end to the constitutionally mandated,

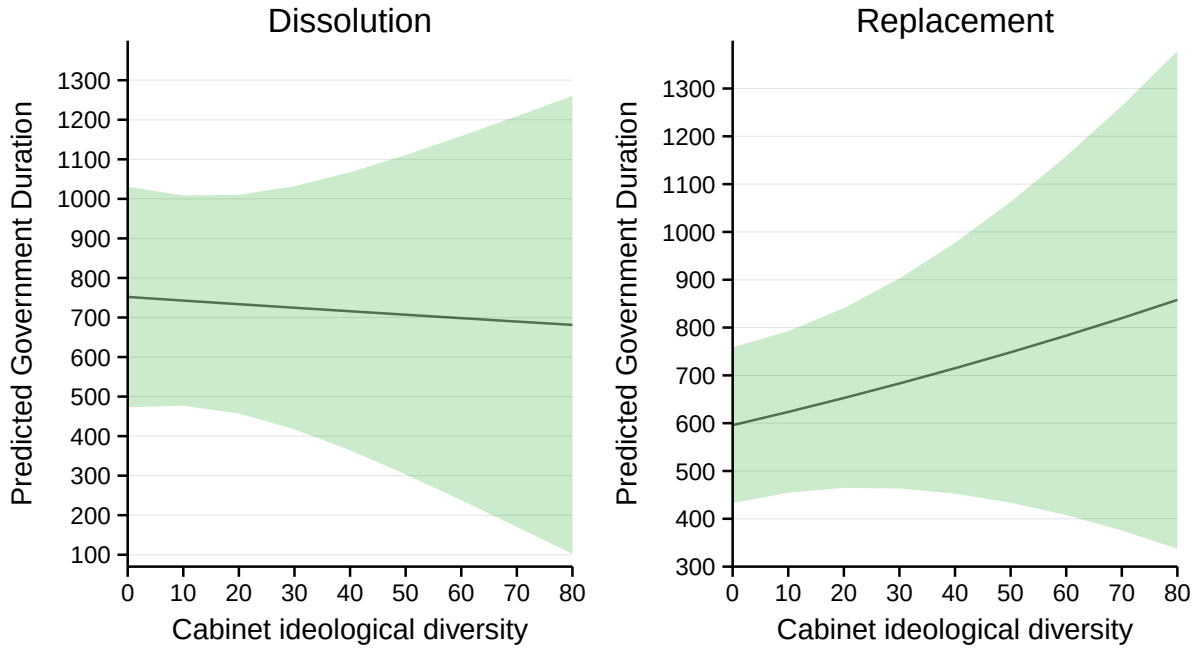

Figure 5.7: Predicted effect of Left-Right cabinet ideological diversity, Diermeier and Stevenson (1999). Predicted government duration is measured in days.

and cabinets which had not experienced termination within 1989 were all censored[7]. As mentioned in section 3.1, I have used a newly updated data set from Chiba et al. (2015), which explicitly used dissolution and replacement censoring in the study of government duration. Diermeier and Stevenson (1999, 1063) reports 124 dissolutions and 117 replacements, out of 268 observations between 1945 to 1989. Chiba et al. (2015, 52) reports 112 dissolutions and 231 replacements out of a total 432 cabinets from 1945-2012. This means that over half the observations are coded as replacements while only a quarter are dissolutions as opposed to Diermeier and Stevenson where number of dissolutions and replacements are nearly identical.

5.2.2 In-sample prediction

The in-sample predictions in figure 5.8 show a clear weakness in predicting cabinet durations of cabinets ending in dissolution. The RMSE-value indicates that on average, the in-sample predictions miss the observed durations by 1202 days. It also shows that the dissolution model predicts almost up to 3500 days government duration. The predictive power is improved when using the replacement terminations as the event in question. The

[7]Information on termination and censoring coding retrieved from e-mail correspondence with Randolph Stevenson.

RMSE is more than halved from the dissolution approach, and as shown in figure 5.8, the model has fewer predictions ranging outside the observed duration distribution compared to the dissolution model.

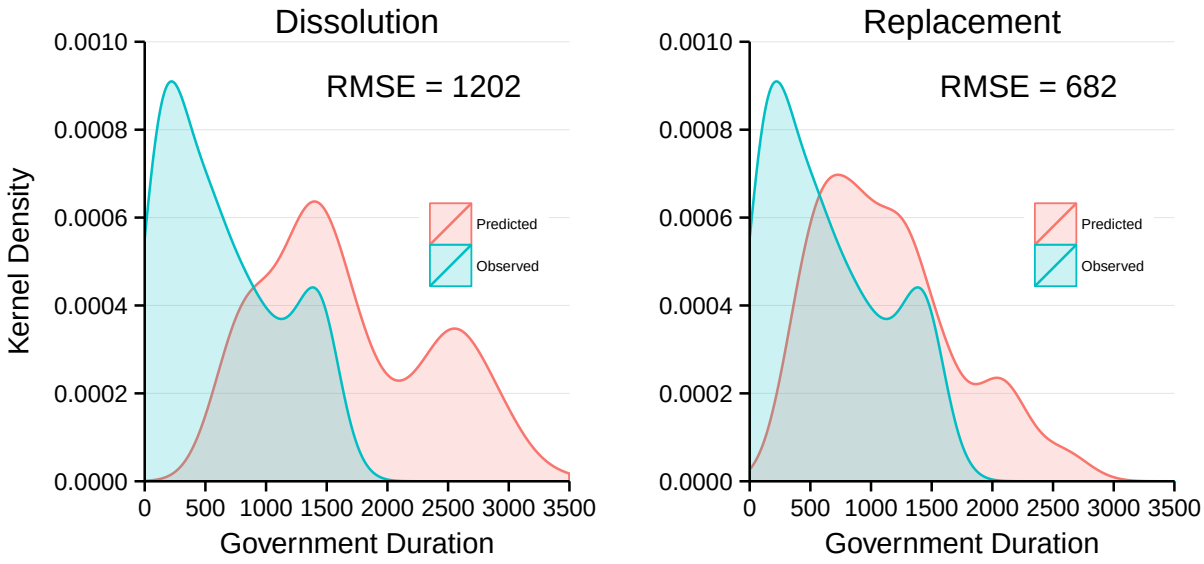

Figure 5.8: In-sample predictive accuracy of governments ending in dissolution and replacements, Diermeier and Stevenson (1999).

A statistical explanation for the poor performance of the dissolution model is that there are only 79 non-missing observations which were dissolution terminations. The remaining observations are censored, and therefore do only contribute to the survival function. This leads to a high intercept which over-estimate the survival rate, which again makes the model unable to predict any observations under the value of the intercept [8]. When the largest bulk of the observed durations are under 500, this means that the model has explanatory power, but not much predictive power, especially for low durations.

The replacement model is somewhat better compared to the dissolution model. This could be explained statistically by the fact that there are 132 observations replacement terminations. The number of observations that contribute information to the failure rate is over half of the total number (259), and this makes the model better able to predict lower durations, because fewer observations are censored. Still, compared to the Warwick (1994) approach in-sample, the RMSE is 263 days higher. Since both Warwick (1994) and Diermeier and Stevenson (1999) use effectively the same sample, the differences must be traced mostly to competing risk framework and the restricted empirics which restrains

[8] For the dissolution model, the intercept is at exp(6.19)= 488 days.

the empirical evaluation of the approach.

5.2.3 Out-of-sample prediction

The results in this section are based on predicted and observed values of 102 cabinets from 1990 until today. The expectations formed from testing the competing risk approach in-sample are more or less continued when tested on the new data. The out-of-sample evidence in figure 5.9 shows that both models, but especially the dissolution model, are incapable of predicting cabinet duration. Dissolution predictions miss by 1213 days on average, making it even worse than the in-sample dissolution predictions.

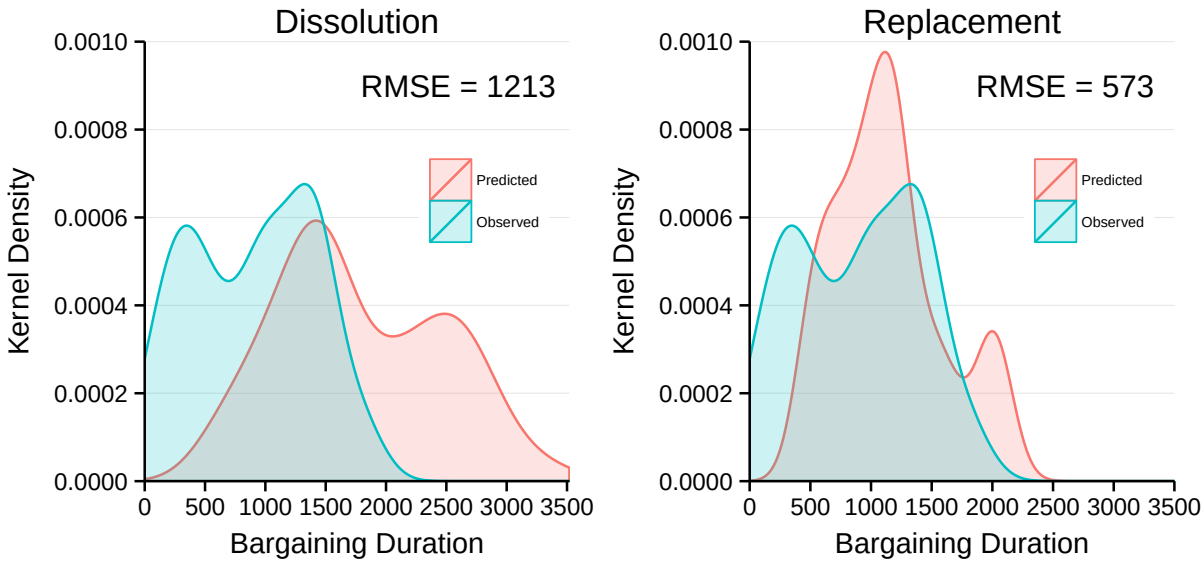

Figure 5.9: Out-of-sample predictive accuracy of governments ending in dissolution and replacements, Diermeier and Stevenson (1999).

The replacement model tells a different story. The RMSE has decreased by 109 days compared to the in-sample results. The data, hence, shows that the replacement model is better adept to predict cabinets from 1990 up until today rather than predicting the durations in the original sample. Given that the new data have a higher mean cabinet duration and that the high intercept leads to predictions biased towards finding high observed durations, a tentative explanation could be that the replacement model has higher probability to predict higher durations than shorter durations.

5.2.4 Cross validation

The 5-fold cross validation gives the worst RMSE's, both for the dissolution and the replacement termination models. Here, the full dataset consisting of 363 observations, is split randomly to 5 folds. The main interpretation of figure 5.10 follows in line with the in-sample and out-of-sample results - both models are incapable of predicting actual government durations, and the dissolution model is the worst. The cross validation evidence helps illustrate the general inability of the competing risk approach to predict the government duration.

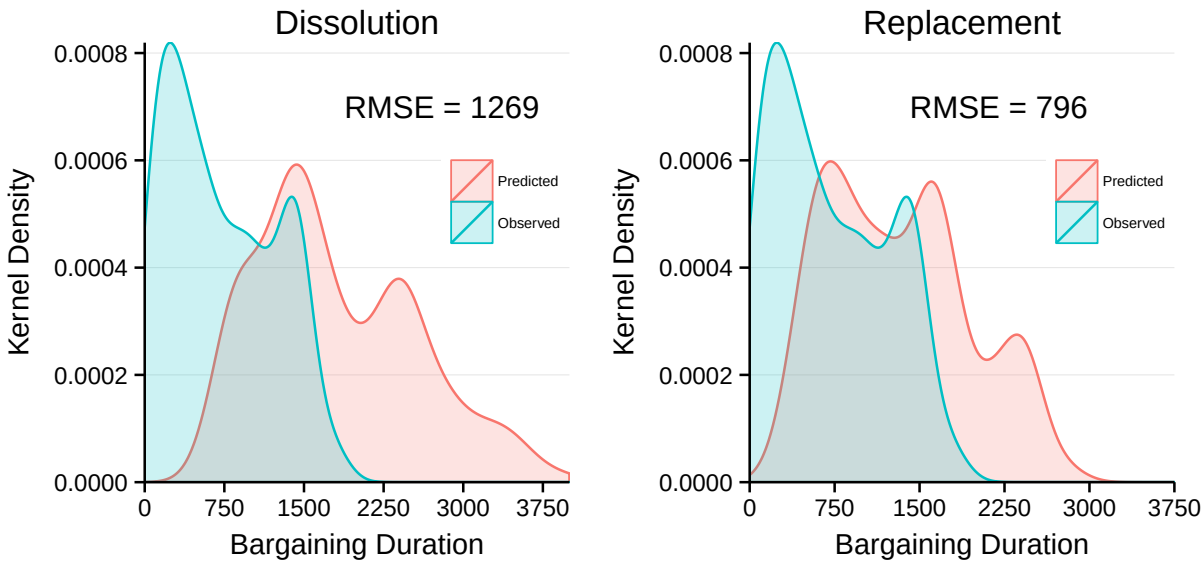

Figure 5.10: Predictive accuracy from 5-fold cross validation, governments ending in dissolution and replacements, Diermeier and Stevenson (1999).

One nuance is due. While the dissolution model is fairly stable across the evaluations, the replacement model show some signs of a better ability of predicting duration. However, when the replacement model is evaluated using 5-fold cross validation the RMSE is the worst compared to in- and out-of-sample. This can indicate over-fitting of the in-sample results. When the Diermeier and Stevenson (1999) competing risk approach is not estimated on the original sample, but rather on a random sample from the full sample, the results show that the competing risk approach has low generalizability[9].

[9]There is also some variance in the RMSE for the 5 different folds. The RMSE for the 5 folds are shown in figure B.1 in appendix B.

5.3 Summarizing Predictive Power in Government Duration

Firstly, this chapter has shown how the more parsimonious approach from Warwick (1994) yields far better predictive power compared to the more complex approach of Diermeier and Stevenson (1999). However, it must be stressed that none of the models are particularly good at predicting government duration. This means that the significant explanatory effects, shown in sections 5.1.2 and 5.2.1, do not contribute much predictive power, either for the original cabinets, the new cabinets or the random split sample. This weakness can be checked and tested on other government duration models, and it is possible that there exist government duration models which predicts better than the models I have chosen for this thesis. Nevertheless, the evidence shown in this chapter points to the usage of more parsimonious models when dealing with the subject of government duration.

Secondly, the out-of-sample evidence show conflicting stories. On one side, the ideological approach and the dissolution model have decreased predictive power on new observations. On the other side, however, the replacement model shows a decreased out-of-sample RMSE compared to the in-sample RMSE. This can be explained by the observed increase in government duration in general after 1990. Tentatively, this has been explained by the high intercept in the replacement model, which leads to high predictions. This coincides with the descriptive fact that mean government duration has increased after 1990.

CHAPTER 6

Evaluation Robustness

This chapter aims to broaden the scope of the evaluations. In chapters 4 and 5 I have evaluated government formation and duration models using tough criteria. Firstly, I have measured predictive power of how the models were able to predict days, and not how they predicted the right amount of weeks or even months. The logic here is that a larger target is easier to hit than a smaller target. Secondly, it has been the expected that the models was able to predict the full scope of both the bargaining and government duration distributions. A more relaxed demand would be to analyze predictive power of models regarding how they for example predicted median observations. Theories aim, broadly speaking, at predicting the general cases. Hence, one would expect a decent model to at least manage to predict median observations.

Therefore, this chapter has the following two parts. Firstly, I will group the in- and out-of-sample predictions to three levels of the duration distribution. Additionally, this section will also group the three levels further into the individual countries. This makes it possible to evaluate how easy or hard to predict the individual countries are. And also the potential consequences this has for the conclusions from the main evaluations. Secondly, as a consequence of the results in the first section, I will use an alternative measure for prediction error which weights large errors less.

6.1 Nuancing Prediction Errors

In this section I will show how the models are able to predict cabinets in the different countries on different levels of duration. I will go through the models, one by one, and show their nuanced in- and out-of-sample predictions. I will group the observations to

the 25th quantile, the median and the 75th quantile of the duration distribution. These durations will be compared to the ability of the model to predict the corresponding low, median and high duration values. Following, I will discuss some deviances and give possible explanation as to why the deviances are observed.

6.1.1 Information uncertainty: Diermeier and van Roozendaal (1998)

Chapter 4 demonstrated that the information uncertainty model have restricted predictive power due to the binary nature of the indicators used to explain formation delays. Figure 6.1 nuances these predictions, by showing how the information uncertainty model predicts the 25th quantile, the median and the 75th quantile observations in-sample, grouped by the countries used to estimate the model.

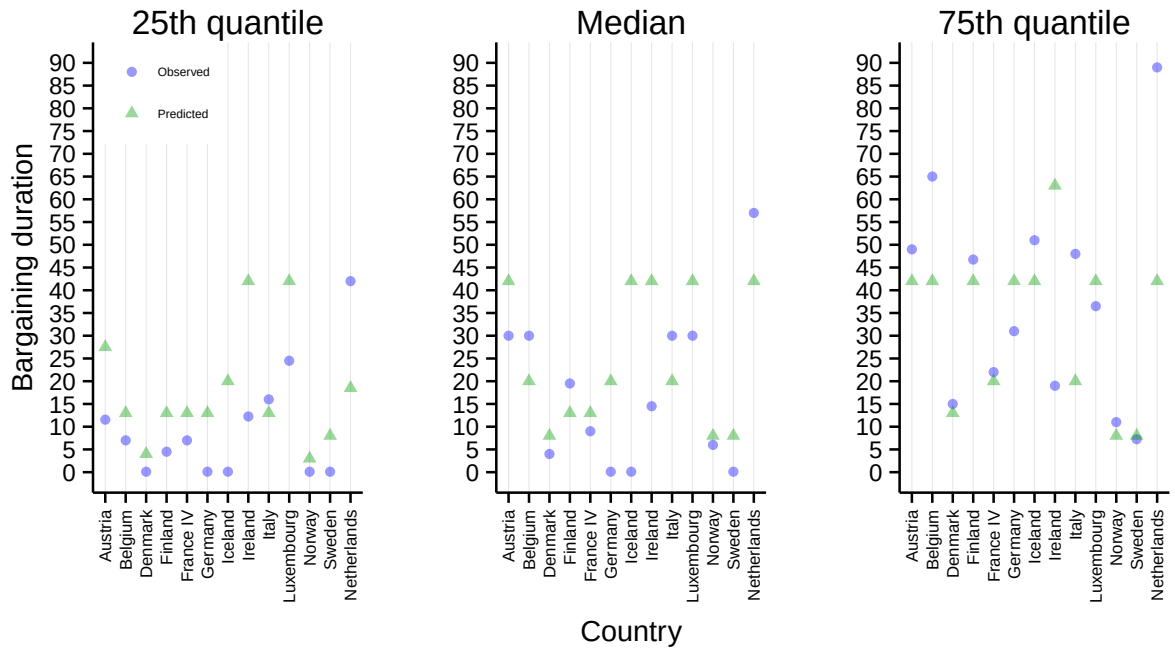

Figure 6.1: In-sample predictions, Diermeier and van Roozendaal (1998). The y-axes indicate predicted bargaining duration. The x-axes indicate the individual countries.

The model does not predict median observations in Iceland and Ireland well. The rest of the Nordic countries, however, are almost perfectly predicted. The Scandinavian countries have been characterized with high government stability, discussed for example in Bergman and Strøm (2011, 51). The in-sample predictions shown in 6.1, hence, do support the stability claim for Scandinavian countries.

Another trend is evident. With the exception of the Netherlands and Italy, all low bargaining delays are over-predicted by the model. The opposite trend is shown in the 75th

quantile plot, under-prediction is here the most common trait. This backs the expectation that the model should be able to predict the median observations better than lower and higher durations. Additionally, there are clear differences between the countries that are well predicted and the countries that the information uncertainty model struggles more with. The Scandinavian countries are at one end of the spectrum, while countries like the Netherlands and Italy, countries with traditionally very complex political systems are at the opposite side, gives the highest prediction errors.

Figure 6.2 shows interesting out-of-sample predictions. As the in-sample predictions, the Nordic countries are the easiest for the information uncertainty model to predict regarding the median values. But this is also true for countries such as Italy and Belgium - two countries are often classified differently from the Nordic countries regarding the political system.

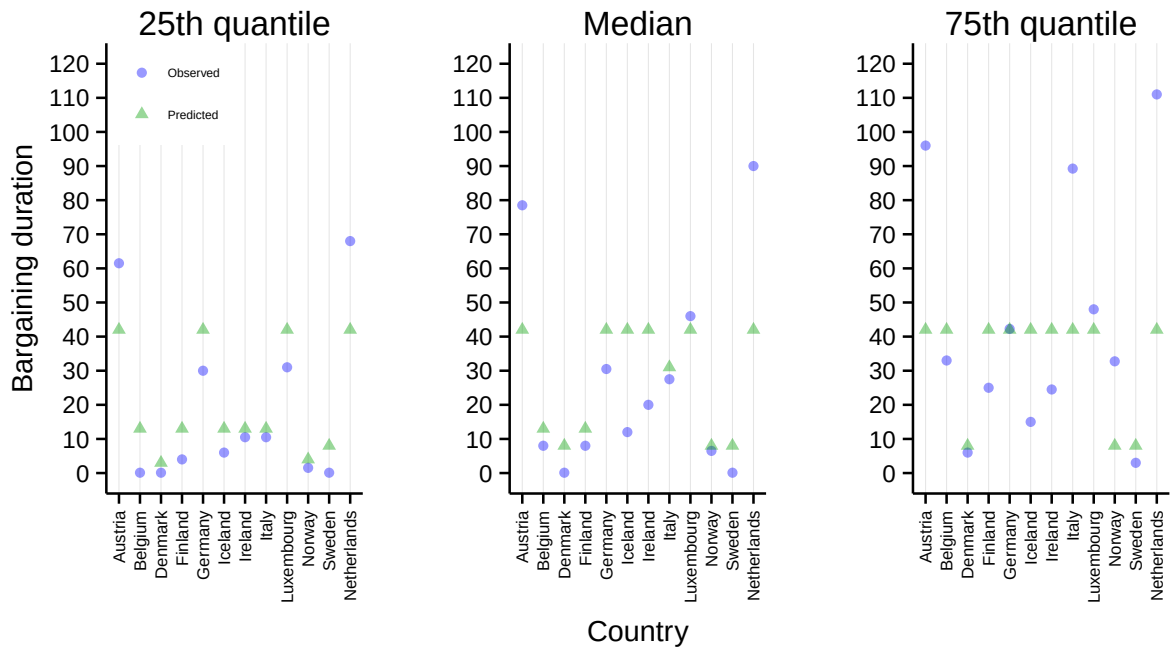

Figure 6.2: Out-of-sample predictions, Diermeier and van Roozendaal (1998).

The main causes for the increase in RMSE in the main analysis[1] are the Netherlands and Austria. The countries are the hardest to predict, no matter the level of duration. One main cause is the change in observed bargaining duration. Table 3.1 in chapter 3 shows that the Netherlands have a mean of 90 bargaining days from 1990 to 2012, which is a 12 day increase for the mean from 1945 to 1989. This point is also raised by De

[1]From 32.5 days in-sample to 36.5 days out-of-sample.

Winter and Dumont (2008, 1521), which labels it as "Dutch exceptionalism", regarding bargaining duration. 16 days behind the Netherlands comes Austria. The 39 day increase from 1945 to 1989 in Austria is large. Evidently, this is a clear sign that the information uncertainty model does not generalize well to new data because of changing political circumstances, illustrated here by the heavy increase in bargaining duration after 1990.

Another sign which supports the conclusion that the model has poor predictive power is the trend most clearly shown in the 75th quantile out-of-sample prediction plot, but also for the two other levels. Here, the inability, because of the binary nature of the model, to predict duration over 50 days is evident. And, because of the binary nature, the RMSE is guided by the following large deviances.

For the main conclusions regarding the information uncertainty approach, it does well at predicting median formation delays in-sample. The lower and higher durations are where the large individual prediction error happens - and hence these formation delays are most responsible for the in-sample RMSE of 32.5 days. The out-of-sample RMSE is heavily guided by prediction errors from the Netherlands and Austria, on all levels of duration.

6.1.2 Combining uncertainty and complexity: Sona Golder (2010)

The second part of chapter 4 demonstrated the decent capability of the combined information uncertainty and bargaining complexity approach to predict in-sample. Figure 6.3 shows the same patterns. The combined model has limited prediction errors for the median observations. Additionally, the model is also quite good at predicting 25th and 75th quantile durations. This is a sign of quality, and also what separates the combined approach most clearly from the pure information uncertainty approach regarding predictive power in-sample.

A clear exception is the Netherlands, which, as also shown for the information uncertainty approach, is a hard country to predict. As mentioned above, these outliers are hard for any model to predict. And, as pointed out before, the large errors from Netherland inflates the value of the RMSE. The point is that Netherland contributes much more to the high RMSE than for example the Scandinavian countries. The Scandinavian countries are nearly perfectly predicted through all three stages.

An observation regarding the comparison between the two theoretical approaches is

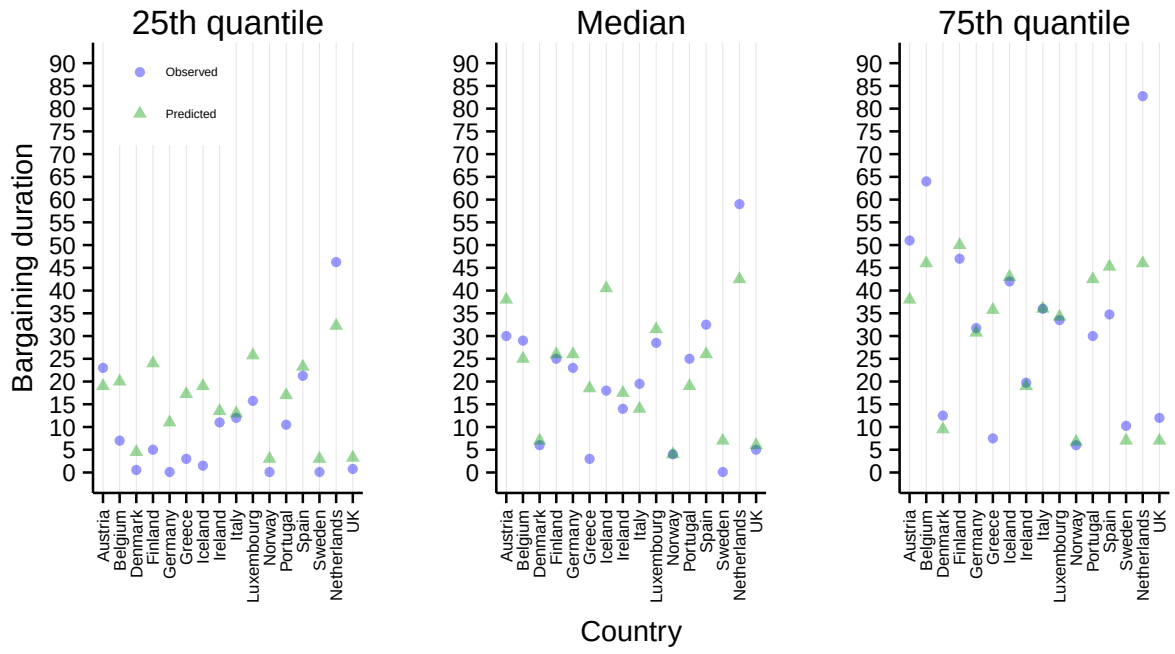

Figure 6.3: In-sample predictions, Sona Golder (2010). The y-axes indicate predicted bargaining duration. The x-axes indicate the individual countries.

that the combined model predicts the Netherlands better than the information uncertainty approach. One explanation is that the combined model has demonstrated its ability to predict the natural distribution of bargaining duration better than the information uncertainty approach. This has been traced to the binary nature of the information uncertainty model compared to the combined model which includes continuous predictors and interaction terms. Substantially, it means that the combined approach is the best at predicting outliers such as the Netherlands. This, again, strengthens the conclusion from chapter 4.

As reported for the in-sample predictions, the out-of-sample predictions, in figure 6.4, has some countries with large prediction errors, Austria in particular. Austria was well accounted for in-sample, as shown in figure 6.3. In light of the in- versus out-of-sample discussion in this thesis, this case is illustrative. The combined model is well adept to predict formation delays in Austria in-sample, showing prediction errors at maximum 13 days[2]. However, when the model is tested on the new sample, the prediction error for Austria increases to a maximum of 66 days[3]. This, again, points towards there being a change in political circumstances after 1999.

[2] 4 days for the 25th quantile, 8 days for median formation delays.
[3] 56.5 days for the 25th quantile and 64 days for median formation delays.

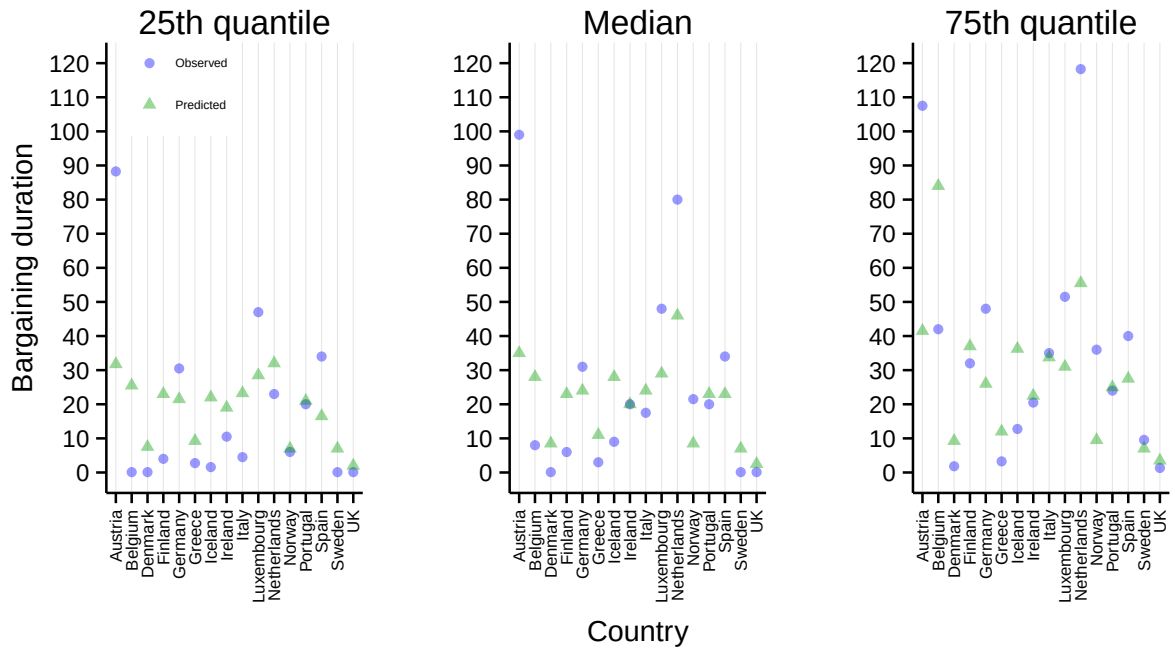

Figure 6.4: Out-of-sample predictions, Sona Golder (2010).

Summing up the nuancing of the government formation models, the combined model shows generally better predictive power than the information uncertainty approach, and it also shows better ability to predict observations outside the median. This gives further support for the conclusions reached in chapter 4 - that the more complex modeling of bargaining duration has better predictive power than the more parsimonious modeling.

6.1.3 Importance of ideology: Warwick (1994)

The nuanced in-sample evidence from Warwick shows an interesting pattern, not reflected in the full analysis in the first part of chapter 5. The ideological approach was shown in chapter 5 to predict poorly in-sample. When broken down, as in figure 6.5, the ideological approach seems to be well adept to predict government duration in the 75th quantile. However, this comes as no surprise considering the high value of the intercept in the original Weibull model, as discussed in chapter 5. A high intercept means that the predictions will be somewhat biased towards the high durations. Hence, the large prediction errors are mostly found for low durations.

In general, none of the countries are demonstrating significantly larger prediction errors than others. This can be interpreted as the strong side of the model's performance in-sample. The model is able to more or less give identical predictions to cabinets across

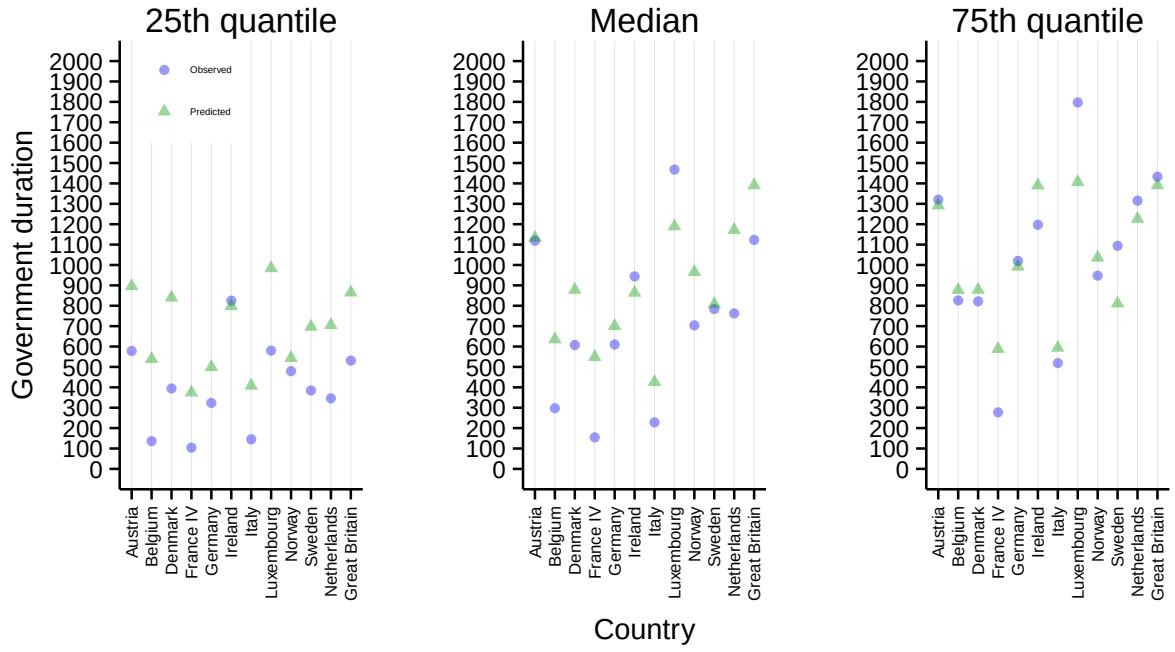

Figure 6.5: In-sample predictions, Paul Warwick (1994). The y-axes indicate predicted government duration. The x-axes indicate the individual countries.

the 12 countries in the sample - this kind of consistency is a sign of quality regarding the possibilities for generalizing the model to the whole sample.

Figure 6.6 shows substantial deviances for all three chosen levels of duration out-of-sample. The in-sample predictions were shown to perform best on high durations. From the out-of-sample results no such conclusions can be drawn. None of the three stages can effectively be separated from the others regarding how close they are to the observed government duration. The conclusion one can make is that the inference of poor out-of-sample predictive power for the ideological approach is supported, even when tested for different stages of the full duration distribution.

The country hardest to predict using the original model on new data is Luxembourg. Luxembourg has prediction errors ranging from 600 to 800 days in the three stages. One explanation for this can be that Luxembourg has the highest observed mean government duration in the sample after 1989, see table 3.2. However, this does not represent a change from before to after 1990, seeing as Luxembourg also had the highest mean duration from 1945 to 1989. A second explanation could be that there are a total of 5 cabinets recorded, and 4 of them are right censored. It means that the cabinets from 1990 and onwards have mostly ended due to a general election. Compared to the period between 1945 to 1989,

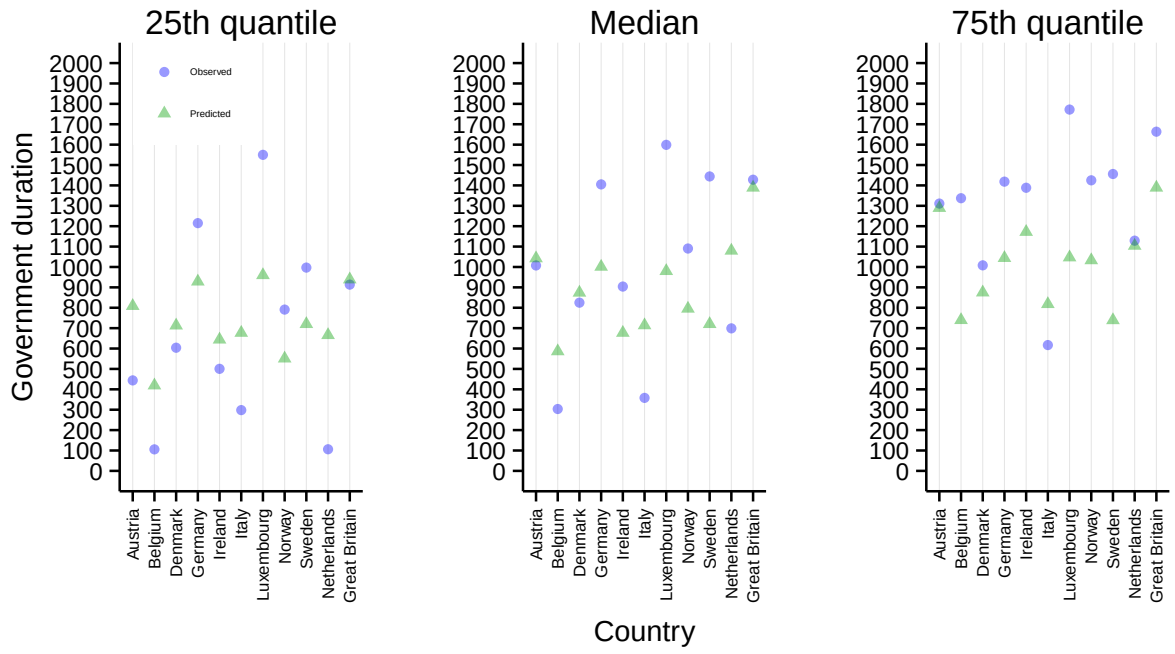

Figure 6.6: Out-of-sample predictions, Paul Warwick (1994).

Luxembourg only had 1 out of 14 cabinets being right censored. This can support an explanation that political realities are changing after 1990, a change which the ideological approach does not generalize well to.

Sweden is almost as hard to predict as Luxembourg. The difference between the median prediction error in-sample compared to the out-of-sample predictions is large, from almost perfect predicted in-sample to over 700 days prediction error out-of-sample. Again, a possible explanation can be found in the descriptive table in chapter 3. In the original data, Sweden are only recorded to have 4 censored cabinets out of 22 in total, i.e. the 4 cabinets are those which terminate due to a general election or some other technical reasons. In the new data, 5 out of 7 cabinets are censored. Hence, relatively speaking, this represents a large deviance from the observed duration before 1990. As the explanation used before, this overweight of censored cabinets after 1990 means that it is harder for the model to estimate the timing of the failure.

Summed up, the in-sample evidence shows a model capable of predicting every country in the sample with constant quality. The out-of-sample evidence demonstrate that the main conclusions from chapter 5 stands, with the exception of Luxembourg and Sweden.

6.1.4 Strategic dissolution: Diermeier and Stevenson (1999)

As shown in the second part of chapter 5, the competing risk model from Diermeier and Stevenson (1999) does not predict government duration well. However, as shown in the previous section, traces of quality can still be found by nuancing the predictions.

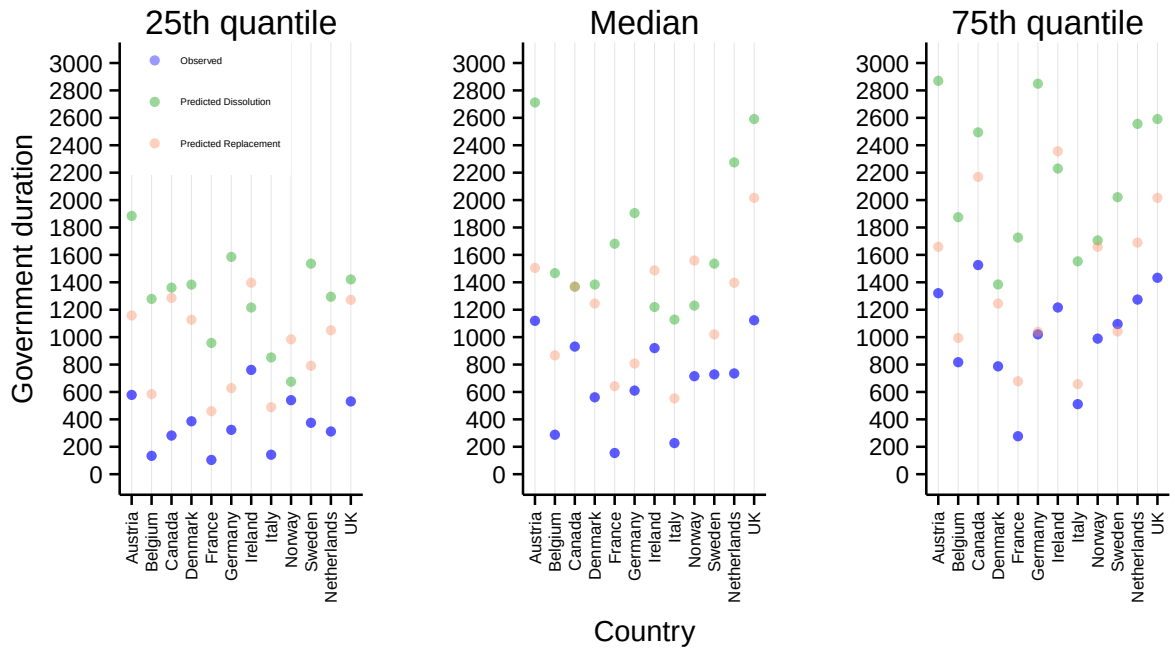

Figure 6.7: In-sample predictions for both dissolution and replacement termination, Diermeier and Stevenson (1999). The y-axes indicate predicted government duration. The x-axes indicate the individual countries. The blue points indicate the observed government duration, both for the dissolution and the replacement model.

The in-sample evidence for both the dissolution and replacement models is shown in figure 6.7. The obvious trend that the replacement model is better able to predict government duration compared to the dissolution model. This is mostly true both across countries and across the levels of government duration, except for Norway and Ireland. The two countries are consistently better predicted by the dissolution model compared to the replacement model. Why does the dissolution model suddenly give better predictions compared to the replacement model? An explanation of Ireland can be found in table 3.1. 14 out of 18 Irish cabinets from 1945 to 1989 are coded as dissolutions. Hence, Ireland is an easier country to estimate the failure rate from, and therefore also easier to predict. However, this explanation can be rejected with the Norwegian case - the country does not practice parliamentary dissolution (Rasch 2004, 27). The next step is to nuance the out-of-sample predictions, looking for traces of the same trend regarding Ireland and

Norway.

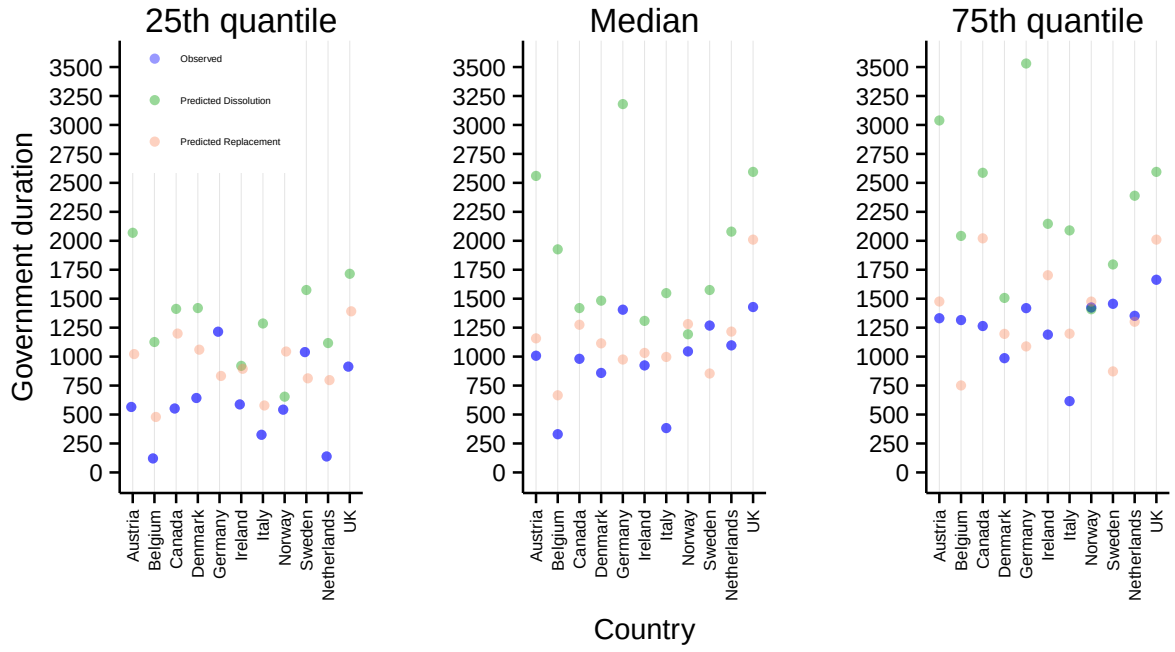

Figure 6.8: Out-of-sample predictions for both dissolution and replacement termination, Diermeier and Stevenson (1999).

Nuancing the out-of-sample evidence, in figure 6.8, gives no news compared to the description of the in-sample predictions above. The replacement model does generally predict better than the dissolution model. Here, the trend of over-predicting is even more imminent. The largest over-prediction error recorded is a 2112 day error for Germany, for high levels of duration. Errors this large does not come alone, Austria does also contribute to the large out-of-sample errors for the dissolution model. Consequently, the RMSE values have been inflated.

As a final note, the inclusion of Norway in the sample of the dissolution model does not make sense. The original article defends the inclusion by arguing that the article would be more comparable to other studies (Diermeier and Stevenson 1999, 1060). However, as mentioned above, Norway does not practice dissolution. Hence, when using Norway in the dissolution approach, the model fails to account for its unique effect. A possible consequence of which is that it introduces some bias regarding the results for the other countries actually having the dissolution rule. This is a possible explanation of the results shown for Norway. In all levels[4], both in- and out-of-sample, the dissolution model predict duration better than the replacement model.

[4]Except for the 75th percentile in-sample, where the replacement model is closest, minimally.

Summing up government formation, the ideology approach from Warwick (1994) gave a slightly more nuanced picture of the predictive ability of the model. It showed that it could predict high durations quite well, and also that it predicted all countries in-sample equally. The opposite can be said for the competing risk approach. The results shown did not change the picture painted in chapter 5 - it only strengthened it, finding that the replacement model performs better than the dissolution model, but that both models have large individual prediction errors.

6.2 Alternative Prediction Error Metric

In the previous section I have pointed to specific countries with large individual prediction errors. They were shown to guide the RMSE's. This section will demonstrate a prediction error calculation less vulnerable to large individual errors. Hence, the prediction error estimate will show the prediction error without the large error weight found for countries such as the Netherlands, Austria and Luxembourg. The mean absolute error (MAE) metric, discussed in 3.2.4, will be used for this.

The tables in this section sum up the RMSE and MAE scores from chapter 4 and 5. The RMSE scores will always be higher or equal to the MAE, never lower, precisely because it does not give particular weight to large deviances. The evidence shown below will be used as potentially further support to the conclusions reached in chapter 4 and 5.

6.2.1 Diermeier and van Roozendaal (1998)

Table 6.1 gives the prediction error measures for the information uncertainty approach. The previous evaluation of the approach have shown that the Netherlands and Austria in particular guide the RMSE value. The MAE scores illustrates this point. The in-sample MAE has around 4.5 days of average error, while the out-of-sample MAE is around 8 days. It means a 28 days average prediction error decrease, respectively. This is a significant reduction from the RMSE values. Additionally, the trend from the RMSE values are replicated by the MAE. The out-of-sample predictions gives higher prediction errors compared to the in-sample predictions.

Table 6.1: RMSE and MAE, from evaluation of Diermeier and van Roozendaal (1998)

	RMSE	MAE
In-sample	32.4	4.5
Out-of-sample	36.5	8.0

6.2.2 Golder (2010)

Table 6.2 shows the RMSE and MAE for the Golder (2010) evaluation. Above, the Netherlands was identified as the country with the largest prediction errors, and hence the greatest contributor to the RMSE value shown below. The MAE's are shown to significantly reduce the prediction errors, as expected. More interesting is that the differences between in- and out-of-sample are reduced from 14 days between the RMSE values, to around 2 days between the MAE values. One explanation is that the deviances shown for Netherlands is not weighted for. Even if the large errors are controlled for, the main results from the evaluation of the combined model hold.

Table 6.2: RMSE and MAE, from evaluation of Golder (2010)

	RMSE	MAE
In-sample	23.5	1.8
Out-of-sample	37.9	3.6

Comparing the information uncertainty to the combined approach, the MAE results give further support to the main conclusion regarding the comparison in chapter 4 - the combined approach by Golder (2010) has more predictive power than the more parsimonious information uncertainty approach by Diermeier and van Roozendaal (1998).

6.2.3 Warwick (1994)

Table 6.3 shows the RMSE and MAE for the ideological approach in Warwick (1994). The ideological approach has previously been described as the model which most consistently predicts all countries in the sample. I singled out Luxembourg, especially for the out-of-sample results, as the country most heavily guiding the RMSE. From a difference of 73 days error on average in RMSE, the MAE show 60 days differences between in- and out-of-sample predictions. This relative reduction demonstrate the effect of removing the weight on Luxembourg out-of-sample. When this is controlled for, the MAE show exactly the same pattern as the RMSE reported in chapter 5. This illustrates that the findings

for the ideological approach is not dependent on the prediction error metric used.

Table 6.3: RMSE and MAE, from evaluation of Warwick (1994)

	RMSE	MAE
In-sample	419	342
Out-of-sample	492	402

6.2.4 Diermeier and Stevenson (1999)

Table 6.4 shows the RMSE and MAE for the competing risk approach of Diermeier and Stevenson (1999). Considering the large errors reported in chapter 5, and in section 6.1.4 the MAE show the expected reduction. For the dissolution model the MAE shows better out-of-sample than in-sample predictive power. However, arguably, when the prediction errors are as large as reported here, the sequence does not matter that much. The main message for the dissolution model is that it does not predict well.

Table 6.4: RMSE and MAE, from evaluation of Diermeier and Stevenson (1999)

	RMSE	MAE
Dissolution: In-sample	1202	1047
Dissolution: Out-of-sample	1213	1005
Replacement: In-sample	682	558
Replacement: Out-of-sample	573	471

For the replacement termination model, the RMSE show better out-of-sample predictions compared to predictions in-sample. This is also found when using the MAE. Basically, it means that the replacement model is better adapted to predict unseen observations than the original observations. One explanation could be that the model showed the best ability at predicting high durations. When the mean duration of the cabinets in the sample has been shown to increase in the new data, the model hence has a higher probability of making correct predictions contra the somewhat lower mean cabinet duration before 1990. Another possible explanation is that, simply spoken, it is a sign of the whole competing risk framework being hard to evaluate, and that the findings presented from it are artefacts of a poorly fitted statistical model to the data,.

This section has in general shown support for the conclusions reached in chapter 4 and 5. The alternative prediction error metric has shown how large individual errors have inflated the RMSE's. However, the MAE's demonstrated the same story as the RMSE's, except for the dissolution model in the competing risk framework.

CHAPTER 7

Concluding Remarks

I have shown, through using classics in the fields of government formation and duration, that none of the models are particularly good at predicting new observations. What can explain this evidence? One explanation could be that the poor out-of-sample predictions are caused by changing political realities after 1990. This argument has, firstly, been supported descriptively by the mean observed increases in both bargaining and government duration after 1990, compared to before 1990. Secondly, it has been supported by showing that prediction errors increase when the original models are tested on data from 1990 and onwards. This is hard evidence which must be considered both in the development of the life cycle of governments and whether it needs a new approach more adept to changing political circumstances.

Regarding parsimony versus complexity for the government formation literature, this thesis does not support the call for going back to basics. The results from chapter 4 show a clear trend. The more complex theoretical approach, represented by the combined information uncertainty and bargaining complexity approach of Golder (2010) has good in-sample predictive power, but the out-of-sample predictions are somewhat weaker. The more parsimonious approach by Diermeier and van Roozendaal (1998) had lower predictive power over all compared to the combined model - which is also supported by the robustness checks in chapter 6. Consequently, the more complex theoretical approach has more predictive power than the more parsimonious approach.

For the government duration studies, claims have been directed towards even more theoretical and empirical complexity. The demand for more complexity is not backed up empirically, as shown in chapter 5. The more parsimonious approach by Warwick (1994) is

significantly better at predicting the duration of new governments than the more complex approach from Diermeier and Stevenson (1999), both in- and out-of-sample. Furthermore, caution must be exercised regarding the competing risk model from Diermeier and Stevenson. The statistical, and hence also theoretical, problem of predicting from the competing risk framework is the heavy amount of censored cabinets, which mean less information goes to estimating the failure rate, and hence the model becomes less fit for making predictions.

Theoretically, this thesis have pointed towards using more complexity in the modeling of the formation process and more parsimony in the modeling of the government duration process. Methodologically, this thesis has argued for developments in the way government formation and duration models should be evaluated empirically. One solution proposed has been the out-of-sample prediction method. This thesis has, therefore, served as an example of how researchers could improve their evaluations by utilizing the out-of-sample method. Empirically, this thesis have contributed to the investigations of the life cycle of governments by updating and expanding cabinet data[1].

7.1 Implications and Future Research

The cabinet level data used to model the life cycle of governments is to a certain degree too far removed from where the actual decisions are made. Here, I will make the case for party-elite level data. Having behavioral data on the actual players involved in the bargaining or, when the government has formed, or in life of the government duration, yields a more nuanced and realistic picture of everyday politics. This is opposed to the restrictions the data on the larger political context sets. Behavioral indicators can yield better predictive power of models on the life cycle of governments, than the approach chosen at the moment. In particular, gathering this data can help open up the literature; behavioral indicators such as ambitions or other kinds of incentives, could help to explain the observed variation of formation delays and government duration.

How are institutional indicators contributing to predicting the life cycle of governments? All articles evaluated in this thesis have tested the effect of the investiture rule. These kinds of institutional variables are often coded as binary predictors - a country

[1]For example by including Canada and the 4th republic of France to the ERD data, and coding cabinet and legislative ideological data.

either has the institutional arrangement or not. Consequently, all cabinets in a country are coded the same way, which, in effect, means that an institutional indicator essentially introduces fixed country effects on the statistical model. One consequence is that the dummy variable constrains the variation in statistical model. This leads to inefficient use of data. Another problem, more difficult to resolve, is that it becomes difficult to entangle and interpret the effect of the institutional indicator. The dummy can represent many things which separates one country from another, and not just the effect of the constitutional constraint.

A possible solution for the investiture predictor is given in Cheibub et al. (2013). Here, the authors argue for a more nuanced operationalization of the investiture rule. This solution contributes to the possibility of observing more variation within countries, and hence the effect of the constitutional constraint can more efficiently be estimated. Furthermore, this approach makes it easier to entangle the effects from the institutional indicators different form only being pure country fixed effects.

The nuancing of the investiture rule gives actuality to an implication demonstrated in this thesis. As shown in chapter 4, for the information uncertainty model by Diermeier and van Roozendaal (1998), binary predictors put natural limitations on the range of possible predictions. The substantial effect of using only binary predictors is that it becomes difficult to have naturally distributed predictions - meaning that there are difficulties in evaluating how well the model actually performs. This points towards using more nuanced operationalization, which enable the use of categorical predictors with more categories, or also the use of continuous predictors.

The competing risk approach, theorized by Lupia and Strøm (1995) and empirically validated by Diermeier and Stevenson (1999), represented a breakthrough in the government duration literature. This thesis, however, has shown a weakness in the approach. As a consequence of the poor ability of the Cox model to make predictions, I chose the parametric Weibull model. Shown in chapter 5, this way of modeling the competing risk approach does not give good predictions. I have traced the failure of predictive ability to the restricted amount of cabinets which experience the two separate modes of termination. The main problem follows a logic of theoretical richness and empirical restrictions; there are only that many cabinets to test the theories.

Because of the restricted amount of cabinets coded as dissolutions or replacements,

the consequence of implementing the competing risk approach is that there remains very few cabinets actually contributing to the failure rate. As argued before, this makes the intercept in the Weibull model biased towards finding long durations. This happens because the model expects durations to last longer than the maximal observed duration, due to the heavy amount of cabinets without an observed end point. Following Clarke and Primo (2012), the quality of a theoretical model can be judged, amongst other alternatives, by its ability to generate empirically testable hypothesis. This, arguably, points in the disfavor of the competing risk approach because of the difficulty in evaluating the predictions from the model.

APPENDIX A

Formation

Replication materials are fully available, send request to lars_sutterud@hotmail.com.

The following models are Cox models if not indicated otherwise. The coefficients can be interpreted as follows: A negative coefficient in the Cox models indicates that an increase in the predictor means an increase in survival time. A positive coefficient in the Cox models indicates decreasing survival time with an increase in the predictor. A negative Weibull coefficient means that an increase in the predictor leads to decreased duration. A positive coefficient indicates longer survival time. Zero indicates no effect both for Cox and Weibull coefficients. In sum, this means that a negative Cox coefficient is interpreted identical to a positive Weibull coefficient.

A.1 Diermeier and van Roozendaal (1998)

Table A.1: Reduced model - Diermeier and van Roozendaal (1998, 625). Standard errors are not given in the original article. Replication(2) and Weibull(2) are without the identifiability indicator

Predictor	Original	Replication(1)	Replication(2)	Weibull(1)	Weibull(2)
Post-Election	-0.85	-1.02 (0.14)	-1.06 (0.13)	1.14 (0.20)	1.16 (0.17)
Previous defeat	-0.31	-0.33 (0.17)	-0.30 (0.16)	0.44 (0.24)	0.41 (0.22)
Caretaker	0.54	0.25 (0.26)	0.22 (0.25)	-0.15 (0.38)	-0.17 (0.36)
Continuation	1.03	1.46 (0.17)	1.43 (0.16)	-1.64 (0.22)	-1.60 (0.21)
Identifiability	0.32	-0.04 (0.22)		0.14 (0.33)	
Intercept				2.57 (0.17)	2.57 (0.15)
N	304	270	295	270	295

Table A.2: Descriptive statistics - Diermeier and van Roozendaal (1998)

Variable	**No**	**Yes**
Post-election	123	147
Previous Defeat	221	49
Caretaker	253	17
Continuation	206	64
Identifiability	245	25

Table A.3: Descriptive statistics for continuous variable, Diermeier and van Roozendaal (1998)

Statistic	N	Mean	St. Dev.	Min	Max
Formation duration	270	29.64	38.42	0	272

A.2 Golder (2010)

Table A.4: Model 4 - Sona Golder (2010, 20-21)

	Cox prop. hazards		*Weibull*
	Original	Replication	
Post-Election	1.03 (0.86)	1.07 (1.01)	−2.11 (1.41)
Polarization	0.13 (0.71)	−0.05 (0.83)	−1.94 (1.27)
Effective No. parties	−0.09 (0.10)	−0.13 (0.14)	0.01 (0.19)
Investiture	0.37 (0.19)	0.45 (0.19)	−0.27 (0.23)
Continuation	0.73 (0.17)	0.72 (0.19)	−0.72 (0.23)
Single party majority	1.75 (0.17)	1.81 (0.17)	−1.68 (0.18)
Post-El*Polarization	−1.62 (0.81)	−1.65 (0.92)	3.48 (1.36)
Post-El*Effective No. parties	−0.29 (0.13)	−0.29 (0.16)	0.34 (0.21)
Post-El*Investiture	−0.20 (0.22)	−0.22 (0.22)	0.16 (0.27)
Constant			3.78 (1.30)
Observations	383	383	383

Note: St.error in parantheses.

Table A.5: Descriptive statistics for binary variables - Golder (2010)

Variable	**No**	**Yes**
Post-election	155	228
Investiture	221	162
Single party majority	331	52
Continuation	280	103

Table A.6: Descriptive statistics for continuous variables, Golder (2010)

Statistic	N	Mean	St. Dev.	Min	Max
Formation duration	383	24.59	29.95	0	208
Polarization	383	0.50	0.22	0.14	1.24
Legislative parties	383	3.70	1.22	1.99	8.41

A.3 5-Fold Cross Validation Variability

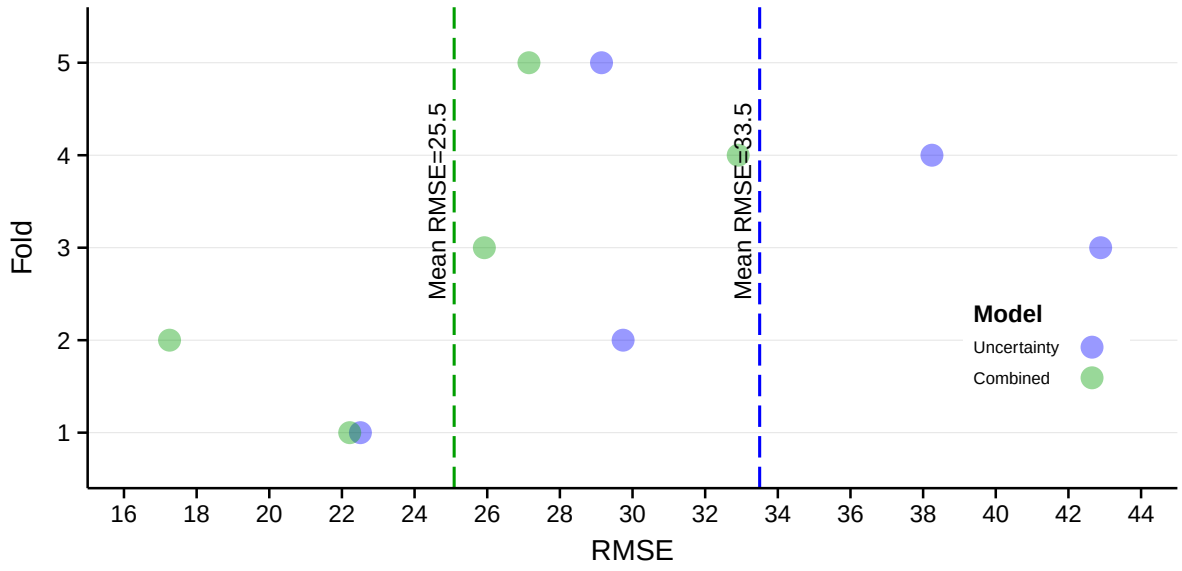

Figure A.1: RMSE values from 5-fold cross validation, Diermeier and van Roozendaal (1998) and Golder (2010). The y-axis shows the 5 different folds. The dashed lines represent the mean RMSE for the two models, indicated by the respective color. The points indicates the RMSE value from the individual folds, interpreted as the mean deviance in days between predictions and observations.

APPENDIX B

Duration

B.1 Warwick (1994)

Table B.1: Original, replication and weibull estimates from Warwick (1994, 59)

Predictor	Original	Replication	Weibull
Majority status	-1.76 (0.21)	-0.91 (0.17)	0.47 (0.10)
Postelection status	-0.60 (0.17)	-1.06 (0.16)	0.58 (0.09)
Investiture	0.68 (0.18)	0.32 (0.16)	-0.15 (0.10)
Returnability	1.63 (0.49)	0.71 (0.48)	-0.49 (0.30)
Left-Right diversity	0.25 (0.08)	0.002 (0.004)	0.001 (0.003)
Clerical-Secular diversity	0.15 (0.06)	-0.08 (0.05)	0.002 (0.03)
Regime Support diversity	0.17 (0.06)	0.22 (0.07)	-0.13 (0.04)
Intercept			6.53 (0.22)
N	284	264	264

Table B.2: Descriptive statistics for binary variables - my data, Warwick (1994)

Variable	**No**	**Yes**
Majority status	100	164
Post-election	117	147
Investiture	126	138

Table B.3: Descriptive statistics for continuous variables, Warwick (1994)

Statistic	N	Mean	St. Dev.	Min	Max
Formation duration	264	666.27	497.78	2	1,935
Returnability	264	0.71	0.14	0.38	0.89
Left-Right diversity	264	17.73	20.08	0.00	81.43
Clerical-Secular diversity	264	1.65	1.80	0.00	7.43
Regime Support diversity	264	0.97	1.47	0.00	6.64

B.2 Diermeier and Stevenson (1999)

Table B.4: Comparison between original results from Diermeier and Stevenson and replication using my data

	Dissolution		Replacement	
	Original	Replication	Original	Replication
Majority status	−1.07 (0.24)	−1.11 (0.27)	−1.36 (0.26)	−1.00 (0.22)
Post-Election	−2.00 (0.25)	−1.05 (0.26)	−0.16 (0.22)	−0.77 (0.21)
Investiture	−0.06 (0.21)	−0.12 (0.26)	1.03 (0.24)	0.45 (0.22)
Returnability	1.12 (0.51)	−0.88 (0.76)	1.52 (0.65)	2.03 (0.71)
Left-Right Range	0.20 (0.11)	0.005 (0.01)	0.24 (0.11)	−0.01 (0.01)
Clerical-Secular Range	0.02 (0.08)	−0.14 (0.11)	0.14 (0.08)	0.06 (0.07)
Regime Support Range	−0.19 (0.16)	0.23 (0.14)	0.30 (0.07)	0.28 (0.09)
Observations	268	259	268	259

Note: Partial likelihood estimates.
St.errors in parantheses.

Table B.5: Comparison between Cox and Weibull using my data, Diermeier and Stevenson model 3 and 4 (1999:1063)

	Dissolutions		Replacements	
	Cox prop. hazards	*Weibull*	*Cox prop. hazards*	*Weibull*
Majority status	−1.11 (0.27)	0.61 (0.18)	−1.00 (0.22)	0.52 (0.13)
Post-Election	−1.05 (0.26)	0.60 (0.17)	−0.77 (0.21)	0.46 (0.12)
Investiture	−0.12 (0.26)	0.07 (0.18)	0.45 (0.22)	−0.25 (0.13)
Returnability	−0.88 (0.76)	0.65 (0.54)	2.03 (0.71)	−1.24 (0.43)
Left-Right Diversity	0.005 (0.01)	−0.001 (0.01)	−0.01 (0.01)	0.005 (0.004)
Clerical-Secular Diversity	−0.14 (0.11)	0.06 (0.07)	0.06 (0.07)	−0.06 (0.04)
Regime Support Diversity	0.23 (0.14)	−0.14 (0.10)	0.28 (0.09)	−0.16 (0.05)
Constant		6.19 (0.38)		7.50 (0.33)
Observations	259	259	259	259

Note: Partial likelihood estimates.
St.errors in parantheses.

Table B.6: Descriptive statistics for continuous variables, Diermeier and Stevenson (1999)

Statistic	N	Mean	St. Dev.	Min	Max
Formation duration	259	644.78	483.69	2	1,748
Returnability	259	0.71	0.14	0.38	0.89
Left-Right diversity	259	15.96	19.77	0.00	81.43
Clerical-Secular diversity	259	1.39	1.71	0.00	7.43
Regime Support diversity	259	0.89	1.46	0.00	6.64

Table B.7: Descriptive statistics for binary variables, Diermeier and Stevenson (1999)

Variable	**No**	**Yes**
Majority status	103	156
Post-election	115	144
Investiture	126	133

B.3 5-Fold Cross Validation Variability

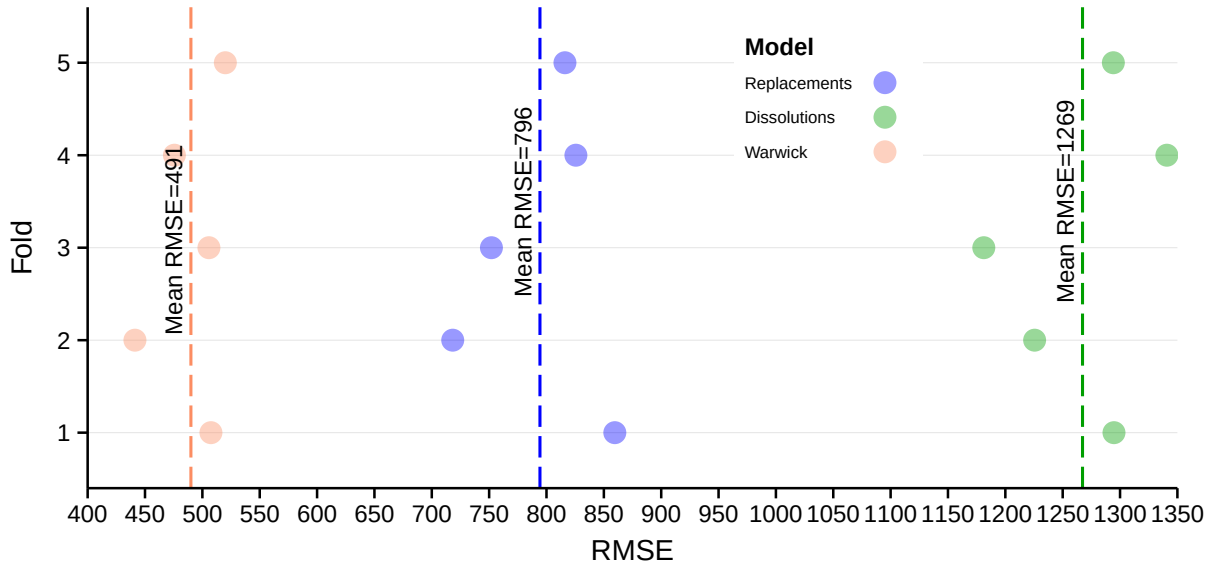

Figure B.1: Variability of CV-estimates, Warwick (1994) and Diermeier and Stevenson (1999). The x-axis shows the RMSE, the dotted lines show the mean RMSE for each model according to their color. The y-axis show the results of the 5 different folds. The points indicates the RMSE value from the individual folds, interpreted as the mean deviance in days between predictions and observations.

Bibliography

Andersson, S., Bergman, T. and Ersson, S.: 2014, *The European Representative Democracy Data Archive, Release 3.* Main sponsor: Riksbankens Jubileumsfond (ln2007-0149:1-E); `http://www.erdda.se/index.php/projects/erd`.

Bäck, H. and Dumont, P.: 2007, Combining Large-N and Small-N Strategies: The Way Forward in Coalition Research, *West European Politics* **30**(3), 467–501.

Bakker, R., De Vries, C., Edwards, E., Hooghe, L., Jolly, S., Marks, G., Polk, J., Rovny, J., Steenbergen, M. and Vachudova, M. A.: 2015, Measuring Party Positions in Europe The Chapel Hill Expert Survey Trend File, 1999–2010, *Party Politics* **21**(1), 143–152.

Baron, D. P. and Ferejohn, J. A.: 1989, Bargaining in Legislatures, *The American Political Science Review* **83**(4), 1181–1206.

Benoit, K. and Laver, M.: 2006, *Party Policy in Modern Democracies*, New York: Routledge.

Bergman, T. and Strøm, K.: 2011, Nordic Europe in Comparative Perspective, *in* T. Bergman and K. Strøm (eds), *The Madisonian Turn: Political Parties and Parliamentary Democracy in Nordic Europe*, Ann Arbor: University of Michigan Press, chapter 2.

Box-Steffensmeier, J. M. and Jones, B. S.: 2004, *Event History Modeling: A Guide for Social Scientists*, New York: Cambridge University Press.

Browne, E. C., Frendreis, J. P. and Gleiber, D. W.: 1984, An Events Approach to the Problem of Cabinet Stability, *Comparative Political Studies* **17**(2), 167–197.

Browne, E. C., Frendreis, J. P. and Gleiber, D. W.: 1986, The Process of Cabinet Dissolution: An Exponential Model of Duration and Stability in Western Democracies, *American Journal of Political Science* pp. 628–650.

Browne, E. C., Frendreis, J. P. and Glieber, D. W.: 1988, Contending Models of Cabinet Stability: A rejoinder, *American Political Science Review* **82**(3), 930–941.

Castles, F. G. and Mair, P.: 1984, Left–Right Political Scales: Some Expert Judgments, *European Journal of Political Research* **12**(1), 73–88.

Chai, T. and Draxler, R. R.: 2014, Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? Arguments Against Avoiding RMSE in the Literature, *Geoscientific Model Development* **7**(3), 1247–1250.

Cheibub, J. A., Martin, S. and Rasch, B. E.: 2013, The Investiture Vote and the Formation of Minority Parliamentary Governments. Workshop on 'The Importance of Constitutions: Parliamentarism, Representation, and Voting Rights'.

Chiba, D., Martin, L. and Stevenson, R.: 2015, A Copula Approach to the Problem of Selection Bias in Models of Government Survival, *Political Analysis* **23**(1), 42–58.

Clarke, K. A. and Primo, D. M.: 2012, *A Model Discipline: Political Science and the Logic of Representations*, Oxford: Oxford University Press.

Crombez, C.: 1996, Minority Governments, Minimal Winning Coalitions and Surplus Majorities in Parliamentary Systems, *European Journal of Political Research* **29**(1), 1–29.

De Winter, L. and Dumont, P.: 2008, Uncertainty and Complexity in Cabinet Formation, *in* K. Strøm, W. C. Müller and T. Bergman (eds), *Cabinets and Coalition Bargaining: The Democratic Life-Cycle in Western Europe*, Oxford: Oxford University Press, pp. 123–157.

Devos, C. and Sinardet, D.: 2012, Governing Without a Government: The Belgian Experiment, *Governance* **25**(2), 167–171.

Diermeier, D.: 2014, Formal Models of Legislatures, *in* S. Martin, T. Saalfeld and K. W. Strøm (eds), *The Oxford Handbook of Legislative Studies*, Oxford: Oxford University Press, chapter 2.

Diermeier, D. and Stevenson, R. T.: 1999, Cabinet Survival and Competing Risks, *American Journal of Political Science* **43**(4), 1051–1068.

Diermeier, D. and van Roozendaal, P.: 1998, The Duration of Cabinet Formation Processes in Western Multi-Party Democracies, *British Journal of Political Science* **28**(4), 609–626.

Dodd, L. C.: 1976, *Coalitions in Parliamentary Government*, Vol. 176, Princeton: Princeton University Press.

Druckman, J. N., Leeper, T. J. and Mullinix, K. J.: 2014, The Experimental Study of Legislative Behaviour, *in* S. Martin, T. Saalfeld and K. W. Strøm (eds), *The Oxford Handbook of Legislative Studies*, Oxford: Oxford University Press, chapter 2.

Esteban, J.-M. and Ray, D.: 1994, On the measurement of polarization, *Econometrica: Journal of the Econometric Society* pp. 819–851.

Gallagher, M., Laver, M. and Mair, P.: 2011, *Representative Government in Modern Europe*, New York: McGraw-Hill.

Golder, M., Golder, S. N. and Glasgow, G.: 2012a, New Empirical Strategies for the Study of Parliamentary Government Formation, *Political Analysis* **20**(2), 248–270.

Golder, M., Golder, S. N. and Siegel, D. A.: 2012b, Modeling the Institutional Foundation of Parliamentary Government Formation, *The Journal of Politics* **74**(2), 427–445.

Golder, M., Golder, S. N. and Siegel, D. A.: 2014, Evaluating a Stochastic Model of Government Formation, *Journal of Politics* **76**(4), 880–886.

Golder, S. N.: 2005, Pre-Electoral Coalitions in Comparative Perspective: A Test of Existing Hypotheses, *Electoral Studies* **24**(4), 643–663.

Golder, S. N.: 2010, Bargaining Delays in the Government Formation Process, *Comparative Political Studies* **43**(1), 3–32.

Hill Jr., D. W. and Jones, Z. M.: 2014, An Empirical Evaluation of Explanations for State Repression, *American Political Science Review* **108**(3), 661—687.

Holger, D. and Manow, P.: 2012, *Parliament and government composition database (ParlGov): An infrastructure for empirical information on parties, elections and governments in modern democracies.* `www.parlgov.org`.

Huber, J. and Inglehart, R.: 1995, Expert Interpretations of Party Space and Party Locations in 42 Societies, *Party Politics* **1**(1), 73–111.

Indridason, I. H.: 2011, Coalition Formation and Polarisation, *European Journal of Political Research* **50**(5), 689–718.

James, G., Witten, D., Hastie, T. and Tibshirani, R.: 2013, *An Introduction to Statistical Learning*, New York: Springer.

Jones, Z. M. and Linder, F.: 2014, Data Mining as Exploratory Data Analysis. Paper presented at the 2014 Society for Political Methodology summer meeting.

King, G., Alt, J. E., Burns, N. E. and Laver, M.: 1990, A Unified Model of Cabinet Dissolution in Parliamentary Democracies, *American Journal of Political Science* **34**(3), 846–871.

Laver, M.: 2003, Government Termination, *Annual Review of Political Science* **6**(1), 23–40.

Laver, M. and Schofield, N.: 1990, *Multiparty Government: The Politics of Coalition in Europe*, Oxford: Oxford University Press.

Laver, M. and Shepsle, K. A.: 1996, *Making and Breaking Governments: Cabinets and Legislatures in Parliamentary Democracies*, Cambridge: Cambridge University Press.

Lupia, A. and Strøm, K.: 1995, Coalition Termination and the Strategic Timing of Parliamentary Elections, *American Political Science Review* **89**(3), 648–665.

Martin, L. W. and Stevenson, R. T.: 2001, Government Formation in Parliamentary Democracies, *American Journal of Political Science* **45**(1), 33–50.

Martin, L. W. and Stevenson, R. T.: 2010, The Conditional Impact of Incumbency on Government Formation, *American Political Science Review* **104**(3), 503–518.

Martin, L. W. and Vanberg, G.: 2003, Wasting time? The Impact of Ideology and Size on Delay in Coalition Formation, *British Journal of Political Science* **33**(2), 323–332.

Martin, L. W. and Vanberg, G.: 2014, A Step in the Wrong Direction: An Appraisal of the Zero-Intelligence Model of Government Formation, *Journal of Politics* **76**(4), 873–879.

Müller, W. C. and Strøm, K.: 2000, Coalition Governance in Western Europe: An Introduction, *in* W. C. Müller and K. Strøm (eds), *Coalition Govenrments in Western Europe*, Oxford: Oxford University Press, chapter 1.

Rasch, B. E.: 2004, *Kampen om Regjeringsmakten: Norsk Parlamentarisme i Europeisk Perspektiv*, Oslo: Fagbokforlaget.

Riker, W. H.: 1962, *The theory of political coalitions*, Vol. 578, New Haven: Yale University Press.

Saalfeld, T.: 2008, Institutions, Chance and Choices: The Dynamics of Cabinet Survival, *in* K. Strøm, W. C. Müller and T. Bergman (eds), *Cabinets and Coalition Bargaining: The Democratic Life Cycle in Western Europe*, Oxford: Oxford University Press, chapter 10, pp. 327–368.

Seki, K. and Williams, L. K.: 2014, Updating the *Party Government* Data Set, *Electoral Studies* **34**, 270–279.

Seki, K. and Williams, L. K.: 2015, *Updating the Party Government data set, Codebook for Data Set 1: Governments*. `http://web.missouri.edu/~williamslaro/SW%202014%20Codebook--Governments.pdf`.

Shmueli, G.: 2010, To Explain or to Predict?, *Statistical Science* pp. 289–310.

Strøm, K.: 1985, Party goals and government performance in parliamentary democracies, *The American Political Science Review* **79**(3), 738–754.

Strøm, K.: 1988, Contending models of cabinet stability, *American Political Science Review* **82**(3), 923–930.

Strøm, K.: 1990, *Minority Government and Majority Rule*, Cambridge: Cambridge University Press.

Strøm, K., Müller, W. C. and Bergman, T.: 2008, *Cabinets and Coalition Bargaining: The Democractic Life Cycle in Western Europe*, Oxford: Oxford University Press.

Strøm, K. and Swindle, S. M.: 2002, Strategic Parliamentary Dissolution, *American Political Science Review* **96**(3), 575–591.

Volkens, A., Lehmann, P., Merz, N., Regel, S. and Werner, A.: 2014, *The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2014b*, Wissenschaftszentrum Berlin fur Sozialforschung, Berlin. with Schultze, Henrike.

Ward, M. D., Greenhill, B. D. and Bakke, K. M.: 2010, The Perils of Policy by P-Value: Predicting Civil Conflicts, *Journal of Peace Research* **47**(4), 363–375.

Warwick, P.: 1994, *Government Survival in Parliamentary Democracies*, Cambridge: Cambridge Univ Press.

Woldendorp, J., Keman, H. and Budge, I.: 2000, *Party Government in 48 Democracies (1945-1998): Composition-Duration-Personnel*, Dordrecht: Kluwer Academic Publishers.