The Statistical Analysis of Judicial Decisions and Legal Rules with Classification Trees

Jonathan P. Kastellec*

A key question in the quantitative study of legal rules and judicial decision making is the structure of the relationship between case facts and case outcomes. Legal doctrine and legal rules are general attempts to define this relationship. This article summarizes and utilizes a statistical method relatively unexplored in political science and legal scholarship—classification trees—that offers a flexible way to study legal doctrine. I argue that this method, while not replacing traditional statistical tools for studying judicial decisions, can better capture many aspects of the relationship between case facts and case outcomes. To illustrate the method's advantages, I conduct classification tree analyses of search and seizure cases decided by the U.S. Supreme Court and confession cases decided by the courts of appeals. These analyses illustrate the ability of classification trees to increase our understanding of legal rules and legal doctrine.

I. INTRODUCTION

A necessary condition for a legal system to be considered as operating under a rule of law is a high degree of consistency between similar fact situations and similar judicial outcomes: if two courts hear cases with the same facts, the likelihood that both come to the same decision should be high. This principle guides the norm of stare decisis in the U.S. common-law system, under which courts are supposed to decide cases by looking to past precedents in a given area of the law. Accordingly, legal texts advise law students that to "predict how future cases are likely to be resolved, you study how similar cases have been resolved in the past" (Berch et al. 2002:11).

To enhance the ability to predict, appellate courts in common-law systems frequently adopt legal rules that attempt to articulate how future cases should be decided. Lower court judges then attempt to apply these rules to cases at hand. Although legal

———

*Assistant Professor of Politics, Princeton University, Department of Politics, 39 Corwin Hall, Princeton, NJ 08544; email: jkastell@princeton.edu.

rules range greatly in their degree of specificity (as well as their effectiveness), in general they help improve the consistency of the law by coordinating expectations for both judges and litigants. Indeed, if judges did simply resolve the dispute before them without articulating how the case at hand fits into the framework of either a legal rule or a series of precedents, it is easy to imagine a legal system devolving into chaos due to lack of predictability.

Recognizing that legal rules are important, however, is a much easier task than defining what exactly a rule is or studying rules (either qualitatively or quantitatively). Over the past half-century, political scientists and legal scholars have sought to document the extent to which a relationship between case facts and judicial decision making holds—mainly with respect to the U.S. Supreme Court, but also for the lower federal courts and state supreme courts. More precisely, they have sought to explain judicial decisions by coding for the presence or absence of certain case facts in various areas of the law. These studies have contributed greatly to the field of judicial politics, often showing high consistency between fact patterns and judicial outcomes and suggesting that judicial decision making, especially at the Supreme Court level, is not as capricious as some would argue.[1] They have also allowed scholars to assess competing theories of judicial decision making (George & Epstein 1992); test whether lower courts tend to be faithful agents of the Supreme Court (Songer et al. 1994; Benesh 2002); and explore how the Supreme Court chooses which cases to review (Cameron et al. 2000).

This article summarizes and utilizes a statistical method relatively unexplored in political science and legal scholarship—classification trees—that offers a flexible way to study legal doctrine through the use of fact-pattern analysis. I argue that this method, while not replacing traditional statistical tools for studying judicial decisions, can better capture the mapping from fact space to outcome space. In doing so, it can also better connect empirical analyses of judicial decision making to their theoretical underpinnings (Cameron & Kornhauser 2005). To illustrate the method's advantages, I conduct classification tree analyses of search and seizure cases decided by the U.S. Supreme Court and confession cases decided by the courts of appeals. These analyses illustrate the ability of classification trees to increase our understanding of legal rules and legal doctrine.

The article proceeds as follows. Section II discusses the notion of legal rules and their importance for drawing inferences from fact-pattern analysis. Section III discusses the methodological approach utilized by judicial politics scholars in fact-pattern analyses and its implicit assumptions about the nature of judicial decision making. It also illustrates why the standard methodological approach may be too restrictive to capture the structure of legal reasoning. Section IV introduces the method of classification trees, and discusses their advantages in examining judicial decision making and presenting a more flexible approach to legal doctrine. Section V applies classification trees to search and seizure cases and confession cases. Section VI concludes.

––––––

[1]Compare, for example, Segal (1984) and Kritzer and Richards (2005) with Amar (1994) on the consistency of the Supreme Court's Fourth Amendment jurisprudence.

II. The Content of a Legal Rule

More than a century ago, Oliver Wendell Holmes (1897:461) argued that the "prophecies of what the courts will do in fact, and nothing more pretentious, are what I mean by the law." In many areas of the law, such prophesying is enabled by legal rules established by courts—frequently, in the United States, by the U.S. Supreme Court—that structure how cases concerning a certain set of facts in a particular area of the law will ultimately be decided. Although a court must decide who wins and who loses in a given case, the structure given by legal rules is important because the significance of a Supreme Court decision— and, to a lesser, though still significant extent, appellate courts and state supreme courts— extends well beyond the immediate parties involved in a case. Under the doctrine of vertical stare decisis, a legal rule established by a higher court's reasoning in prior cases both binds lower courts and creates expectations for potential litigants in a given area of the law. "Put another way, the rules articulated by the Court direct subsequent behavior by providing information about mutual expectations and sanctions for compliance" (Wahlbeck 1997:780). To use a famous example, the procedures outlined in *Miranda v. Arizona* (384 U.S. 386) compelled police officers to question suspects in a certain manner and lower courts to uphold the rules set forth by the Supreme Court. If they did not, they faced the sanction of having the confessions they elicited thrown out in court.

What, then, is the structure of a legal rule? While rules can take many forms, and range from the very specific to the very abstract, all legal rules must in some form establish a meaningful relationship between certain sets of facts and certain case outcomes. A legal rule is, at its core, a sorting device: certain facts can be classified as producing one legal outcome, while other facts can be classified as producing a different legal outcome. "A legal doctrine indicates which fact situations are to be grouped together and treated similarly. In other words, it creates a set of equivalence classes in a fact or case space" (Cameron et al. 2000:102).[2]

More formally, using the notation advanced by Kornhauser (1992), we can describe a case c as a vector in an N-dimensional space. A legal rule, then, is a function r that maps cases, c, to outcomes, {0,1}, where the latter can be thought of as a generic dichotomous choice faced by a court (e.g., admit or exclude evidence from a police search).[3] Thus, the rule partitions the N-dimensional fact space into equivalence classes, and dictates which cases will receive one judicial treatment, and which cases will receive the other. It is this mapping from fact space to outcome space that provides the content of a legal rule. Whether a rule is successful depends on the degree to which lower courts, litigants, and

---

[2]This formulation, of course, abstracts from the multitude of forms that a rule might take (Twining & Miers 1999), as well as the difference between legal rules and legal standards (Sunstein 1996:27–28), the latter of which provide for much greater discretion in creating equivalence classes. My purpose in following this abstraction is to advance a conception of rules that facilitates the translation between the theory of judicial decision making and the way it is studied quantitatively.

[3]Some legal rules may involve continuous outcomes—for example, sentencing decisions—but Kornhauser (1992:171) notes that "[l]egal questions almost always have yes or no answers." Indeed, almost all the fact-pattern analyses noted below involve a dichotomous outcome.

police officers understand how the Court intended to classify certain types of fact patterns involving the law of confessions.

Of course, judges with different ideological views in an area of the law will differ on how they believe different case facts should be treated (Segal & Spaeth 2002). At the same time, it is crucial that ideological battles on courts are battles over legal rules, or what form the mapping between case facts and outcomes should take. As Lax (2007:591) notes: "Even the most ideological and policy-oriented appellate judge must make policy by telling lower court judges what facts to consider and what those facts mean for case outcomes." Thus, even granting that decision making on appellate courts, particularly the Supreme Court, is influenced by ideology does not reduce the need to analyze judicial decision making in a rule-based framework.

A. The Role of Legal Reasoning

In considering how best to study the mapping from case facts to case outcomes, it is important to consider how judges come to their decisions. This process usually consists of legal reasoning, which in turn is often marked by reasoning by analogy. Sunstein (1996:65) describes this process as consisting of the following five steps.

> (1) Some fact pattern A—the "source" case—has certain characteristics, call them X, Y, and Z. (2) Fact pattern B—the "target" case—has characteristics X, Y and A, or characteristics X, Y, Z, and A. (3) A is treated a certain way in law. (4) Some principle, created or discovered in the process of thinking through A, B and their interrelations, explains why A is treated the way that it is. (5) Because of what it shares in common with A, B should be treated the same way.

According to this process, cases consist of clusters of facts; legal reasoning is then used to determine whether the addition or presence of a certain case fact, *in combination with the absence or presence of clusters of other case facts*, results in a case being deemed in the same class or an altogether different class.[4] "Indeed, it is *the aspects in which their facts are similar* which [gives one the] first guidance to what *classes* will be found legally relevant, that is, will be found *to operate alike*, or to operate *at all*, upon the court" (Llewellyn 1951:49, emphasis in original).

The implications for assessing the influence of legal reasoning on the mapping between case facts and case outcomes are at least two-fold. First, the absence or presence of a single fact may not simply move the case toward one classification or the other; instead, the absence or presence of a single fact may be enough to shift case *c* from classification A to classification B, even if *c* has many other things in common with other cases that fall under classification A. This decision-making process can frequently be seen in judicial opinions, in which judges move sequentially, or hierarchically, through the facts of the case, discussing how the presence of a case fact or cluster of facts, perhaps in combination with

―――――

[4]Legal rules, of course, are not static, and may evolve or even be altogether revoked as new cases arise or as judicial preferences change (Wahlbeck 1997, 1998). Nevertheless, this change is still likely to occur through the process of legal reasoning. "[C]hange in the rule . . . occurs because the scope of a rule of law, and therefore its meaning, depends upon a determination of what facts will be considered similar to those present when the rule was first announced" (Levi 1949:2).

a legal rule, leads them down a certain path of analysis. In equal protection cases, for instance, the Supreme Court's determination of whether strict, intermediate, or minimal scrutiny should be placed on a legislative classification will lead to three very different decision procedures. If a classification is based on race, it does not simply make it more likely that the Court will strike it down, compared to a classification based on age, all else equal. It may shift the Court's reasoning into an entirely different mode.

The second implication of the influence of legal reasoning is that it may not be sufficient to consider case facts as having simply an additive effect; rather, it is likely that the effect of certain case facts will be interactive: the presence of fact q alone may not matter all that much in the case's classification, but the presence of q along with fact r and fact p may matter a great deal.

To illustrate how legal reasoning might influence our conception of the mapping from facts to cases, consider Fourth Amendment doctrine. Moylan (2003) has outlined a "checklist" to determine whether, under the U.S. Supreme Court's doctrine, a search meets the threshold for Fourth Amendment protection and, if so, whether it in fact violates the Fourth Amendment. While Moylan presents his checklist in prose, it is useful to consider it as a decision tree that guides one through the facts of the case to the probable outcome.[5] I have converted the first part of Moylan's checklist—that dealing with the question of threshold—into a decision tree, presented in Figure 1. The structure of the tree is illuminating. Note that a "no" answer to any of the four questions on the left-hand side does not incrementally shift the case toward being classified as "threshold not met." Rather, it rules out entirely one way that it will be classified as such (unless miscellaneous applicability exists). Conversely, in order to reach the "threshold met" classification, all four of those questions must be answered "yes." Thus, there is an interactive effect in the mapping from case facts to outcomes.

III. The Quantitative Study of Legal Rules

Do scholars' attempts to study the relationship between case facts and case outcomes adequately reflect this decision-making process? On the one hand, the success of numerous empirical models demonstrates that in many areas of the law there is indeed a quantifiable and significant mapping from case facts to case outcomes.[6] On the other hand, it is important to recognize the methodology of these studies and their implications for our understanding of legal rules and judicial decision making. Although these studies differ in

----

[5]For another example, see the decision tree put forth by Perry (1991:278) to help explain the Supreme Court's certiorari decisions.

[6]Following the pioneering efforts of Kort (1957), Segal (1984) brought the study of fact-pattern analysis to the forefront. Since his path-breaking work on the Fourth Amendment, fact-pattern analyses have been applied to several other areas of the law, including sex discrimination (Segal & Reedy 1988); obscenity (McGuire 1990; Hagle 1991; Songer & Haire 1992); the death penalty (George & Epstein 1992; Hall & Brace 1996); the Establishment Clause (Ignagni 1994; Kritzer & Richards 2003); the law of confessions (Benesh 2002); freedom of expression (Richards & Kritzer 2002); judicial review (Emmert 1992); and school financing (Swinford 1991).

Figure 1: Fourth Amendment decision tree, based on U.S. Supreme Court doctrine.

START
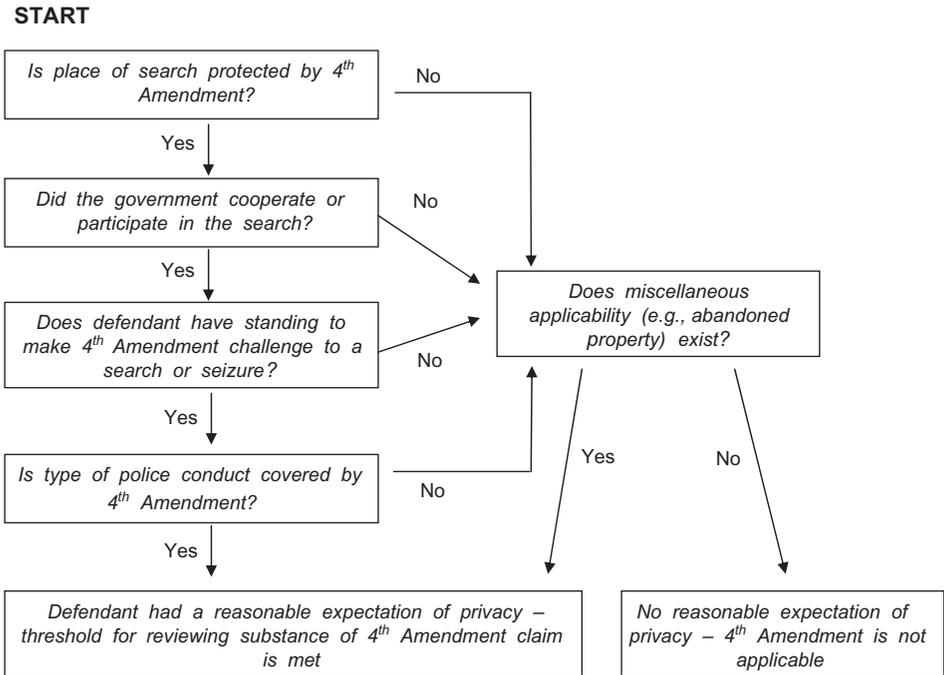
NOTE: The figure presents a decision tree for determining whether a search meets the threshold of Fourth Amendment protection according to the U.S. Supreme Court's doctrine, as adapted from Moylan (2003).

scope and goals, they all follow a similar procedure that involves coding cases for the presence or absence of case facts thought relevant to decisions being studied; performing logit or probit, regressing the outcome on the case facts; and using the coefficients from these regressions to assess the influence of various case facts on a court's decisions.

The output from this methodology indeed can and frequently does reveal a correlation between case facts and outcomes, but this correlation does not necessarily reveal a structural representation of the relationship between case facts and outcome, since logit induces a highly specific mapping from the higher-dimension fact space and the one-dimensional outcome space ("liberal" or "conservative"), in which a cutpoint separates one classification from the other.[7]

The methodology of fact-pattern analysis is thus based on at least two strong assumptions about judicial decision making that are frequently unrecognized in fact-pattern analysis. The first is the inherent assumption about the role of case facts in legal doctrine

———

[7]While more flexible than a linear probability model, which mandates a linear mapping from the case facts to the probability of a court finding the case to be of one class or the other, the logit specification remains highly restrictive and the mapping is constrained by the model's functional form (Beck et al. 2004:24).

Figure 2:   The influence of case facts on the Supreme Court's search and seizure decisions as assumed by logit.
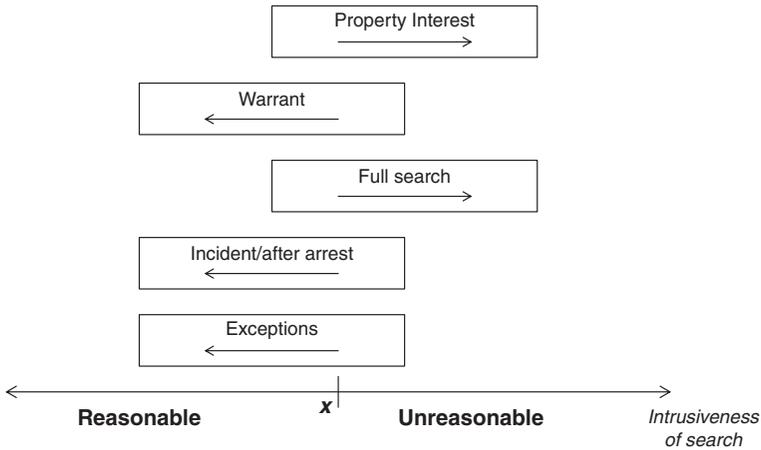
NOTE: The real line represents the intrusiveness of a search, while x partitions that line into two classifications: reasonable and unreasonable. Facts with arrows pointing to the left tend to support a finding that a search was reasonable under the Fourth Amendment, while facts with arrows pointing to the right tend to support a finding of unreasonable. The vertical line labeled "X" denotes a cutpoint dividing reasonable and unreasonable searches. Under this conception of legal rules, the presence or absence of facts simply "pushes" or "pulls" the case toward a liberal or conservative classification, with each fact having an additive effect.

implied by using a logit specification. Under this assumption, the weight of facts are simply added together based on their presence or absence in a given case; that is, the presence of certain facts will push a case toward one classification, while the presence of other facts will pull it toward another classification.

Consider Segal and Spaeth's (2002:312–20) analysis of the Supreme Court's search and seizure decisions. Using logit, they determine that the presence of the following facts is likely to push the Court toward the "liberal" classification (finding the search unreasonable): the person being searched has a property interest (i.e., the search is conducted in a home, business, car, or on one's person), and the police conducted a "full" search rather than a more limited one. On the other hand, the following facts are likely (i.e., statistically significant) to pull the Court toward the "conservative" classification, admitting the evidence in question: the police had a warrant, conducted the search incident to a lawful arrest or after such an arrest, and there existed exceptions to the warrant requirement. Figure 2 depicts graphically how the mapping from case facts to outcome space is occurring: the presence or absence of facts simply moves the classification of the case left or right on the real line, where *x* represents the cutpoint dividing the classifications.

The second, closely-related assumption is that the presence or absence of multiple case facts has an additive, rather than interactive, effect. To be sure, logit is capable of handling interaction terms, but they must be specified a priori by the analyst, something that is rarely if ever done (and only can be done if guided by strong theory). Thus, logit may be unable to reveal interactions that may be contained within the structure of the data.

To be sure, the logit approach and the legal reasoning approach each have their advantages. The former has the virtue of being subjected to systematic empirical testing and has uncovered a statistically significant mapping from case facts to outcomes, but suffers from the vice of oversimplifying that mapping. On the other hand, the latter has the benefit of reflecting more fully the hierarchical and interactive nature of legal decision making but, as Moylan's work illustrates, is based mainly on qualitative interpretations of judicial doctrine and has not yet been subjected to quantitative examination. I turn now to a method that unifies these virtues.

IV. Classification Trees: An Introduction

Classification problems come in many different forms across various disciplines. Doctors, for instance, want to determine whether a person who enters a hospital with chest pain is suffering a heart attack or not (Breiman et al. 1984). Botanists who discover a new plant seek to classify it properly (Gordon 1999). The multiplicity of classification issues has given rise to a wide range of methodological techniques—all of which can be said to fall under the broad category of "pattern recognition" (Ripley 1996)—that attempt to improve on classical regression methods. All these techniques, which include cluster analysis, neural networks, and nearest neighbor methods, involve searching inductively for patterns in data. Political scientists increasingly have begun to take advantage of these techniques, which are much more flexible than standard regression approaches. Neural networks, for instance, are now used to study international conflict, precisely because logit has been shown inadequate to capture the nonlinearities, interactions, and context dependence involved in the onset of conflict between states (see, e.g., Beck et al. 2000, 2004; Andreou & Zombanakis 2001).

To study the mapping from case facts to judicial outcomes, I employ a methodology—classification trees—that offers this flexibility and produces an output that is both easily interpretable and conforms to the hierarchical and dichotomous nature of judicial decision making.[8] A technical introduction is presented in the Appendix, but a brief summary of the method is helpful. Let C be the set of all possible classes, and let χ be the space of all possible values of predictor variables. Based on these predictors, a classification tree successively, or recursively, partitions χ into subregions such that the resulting partitions contain observations that are more similar to each other; each partition then contains a predicted class (Ripley 1996:213). More precisely, a classification tree will first split the data into regions based on the variable that minimizes the heterogeneity in the resulting two groups. The tree will continue to split the data, recursively, as long as doing so reduces the heterogeneity in the data. The result of this process results in a tree that is

---

[8]To my knowledge, classification trees have only been used extensively in one other judicial politics study. Ruger et al. (2004) attempted to predict every decision of the Supreme Court's 2002 term—before the Court handed them down—in two ways: first, by asking legal experts for their opinion; second, by using classification trees fit from the justices' voting behavior in the previous eight terms. Importantly, Ruger et al. were not interested in studying the substance of judicial decision making, only in predicting judicial outcomes.

possibly very large and may overfit the data. Thus, it is usually necessary to "prune" the tree, using a criterion that favors parsimonious trees (Sutton 2005:312–13).

The name "classification tree" comes from their traditional graphical presentations, which is in the form of an upside-down tree. The tree is composed of nodes and branches; observations are split at the former and proceed down the latter. The top node of the tree is known as the "root," at which the initial partition occurs. Successive nodes come in two forms: nonterminal (or internal) nodes, at which the data are further partitioned, and terminal nodes (or "leaves") at which no further partitioning of the data occurs and classification is made. Thus, observations are split at each node and follow the branches to successive nodes until they reach a terminal node.

The structure of classification trees differs from that of logit in several ways that are important for the study of judicial decision making, as discussed earlier. Most importantly, rather than assuming a particular mapping from the high-dimensional space to the low-dimensional outcome space, a classification tree takes a nonparametric approach and directly partitions the case fact space. As Eisenberg and Miller (2007:573) note, due to the fact that they are nonparametric, classification trees have "the advantage over logistic regression of . . . not depending on underlying assumptions about the distribution of the explanatory variables."[9] Second, the classification tree procedure will inherently reveal key interactions among all predictor variables entered into a tree-based model without the need for the analyst to specify them a priori. Third, a classification tree conforms to the hierarchical and dichotomous structure that often seems apparent in judicial opinions—hierarchical in that often the answer to an initial question (e.g., Did the police have a warrant?) will lead a judge down a certain path, and dichotomous in that the answers to the questions considered under the law frequently have a yes/no answer (Kornhauser 1992). Indeed, a classification tree strongly resembles the checklist advanced by Moylan. Finally, the resulting tree is easily interpretable; indeed, most people could predict the outcome of a process modeled by a classification simply by knowing the values of the variables used to construct the tree.[10] As Breiman (2001) notes: "On interpretability, trees rate an A+."

More generally, classification trees can enhance the use of more traditional regression models in several ways. As noted above, they inherently find interactions in the data of which a researcher might be unaware. Second, they enable a researcher to graphically analyze dichotomous outcomes in a flexible manner, which allows for exploration of the data before turning to a more complicated model. Finally, as Eisenberg and Miller (2007) demonstrate, classification trees can help resolve endogeneity problems, which can produce misleading regression results.

A stylized example involving a hypothetical legal rule and simulated data helps illustrate how classification trees are formed in practice, as well as how a tree represents a

------

[9]Three additional advantages of classification trees are that they handle missing data much better than traditional regression models, are invariant to monotone transformations of predictor variables (Clark & Pregibon 1992:378), and are highly resistant to outliers (Sutton 2005:303).

[10]Contrast this ease to the difficulty in interpreting logit coefficients, or even OLS regression coefficients for models with interactions (Brambor et al. 2006).

partitioning of the observational space. It also illustrates how the structure of trees differs from that of logit.[11] In the 1964 case of *New York Times v. Sullivan* (376 U.S. 254), the Supreme Court modified existing libel law and established the legal rule that in order for public officials and public figures to prove defamation, they had to show that (1) the statement made against them was false and (2) that the defendant made the statement with "actual malice." Assume a researcher could quantify falsity and malice and measure them on a scale of 1–10.[12] Further assume that the rule announced in *Sullivan* can be summarized as: *liable* if *falsity* ≥ 4 and *malice* ≥ 6; *not liable* otherwise. I generated 300 observations, or simulated cases, assuming that *falsity* and *malice* are distributed uniformly from 0 to 10, and that each was measured with error:

$$y_i = \begin{cases} 1 & \text{if } (\textit{falsity} + e_1) \ge 4 \ \& \ (\textit{malice} + e_2) \ge 6 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where e_1 and e_2 are independent error terms distributed $N(0,2)$. Thus, the true rule is that libel exists if and only if *falsity* ≥ 4 and *malice* ≥ 6, but we can observe both case factors imperfectly. This data-generating process resulted in *not liable* outcomes in 74 percent of the cases.

Figure 3a depicts the results of a logit analysis of the simulated cases, and Figure 3b presents a classification tree analysis of the cases.[13] The way to read the tree is as follows: starting with the root node, each branch gives a rule for splitting the data. Thus, cases in which *malice* ≤ 5 are predicted to proceed down the left branch to the next node, while those in which *malice* > 5 are predicted to proceed down the right branch.[14] Nodes surrounded by squares are terminal nodes; thus, cases in which *malice* ≤ 5 are predicted to result in a decision of *not liable*. On the other hand, cases that proceed rightward from the root node are further split based on the level of *falsity*; cases with *falsity* ≤ 4.8 are predicted to proceed to the root node of *not liable*, while cases with *falsity* > 4.8 are predicted to proceed to the root node of *liable*. The numbers at the bottom of each node give the absolute breakdown of the classification of the cases that reach that node, with the number of cases in the modal category listed first. Thus, at the root node, the tree can do no better than predicting the modal classification: 222 *not liable* outcomes.

According to conventional standards, the logit performs very well, as the coefficients on each predictor are highly statistically significant, and the model correctly classifies 82

[11]A similar example is presented in Cameron and Kornhauser (2005).

[12]In practice, case facts are almost always measured dichotomously. Constructing continuous measures, however, makes it easier to illustrate the use of classification trees and how they partition a case space.

[13]All trees shown in the article were generated using the *rpart* routine in *R* (Therneau & Atkinson 1997). Details of the procedure *rpart* uses to construct a tree are presented in the Appendix.

[14]Throughout the article, for ease of interpretation, I use the term "predict" to describe how a tree is splitting the data into different classes, even though the goal is to analyze and understand a set of already decided cases. This is analogous to the common interpretation of the "percent correctly predicted" statistic in logit or probit models, which is actually a measure of how well a model classifies observations in a data set rather than a measure of how well a model predicts future outcomes. As noted above, classification trees are frequently used to predict future outcomes, and the one application to date in judicial politics did in fact use them to this end.

Figure 3:   Logit results and classification tree predicting simulated liable case outcomes.

Falsity	0.5 (0.1)
Malice	0.6 (0.1)
Intercept	−7.5 (0.9)
% in Modal Category	74.0
% Correctly Predicted	82.0
Proportional Reduction	30.8

Is Malice ≤ 5?
NL
222/78
yes   no

NL
146/10

Is Falsity ≤ 4.8?
NL
76/68
yes   no

NL
58/13

L
18/55
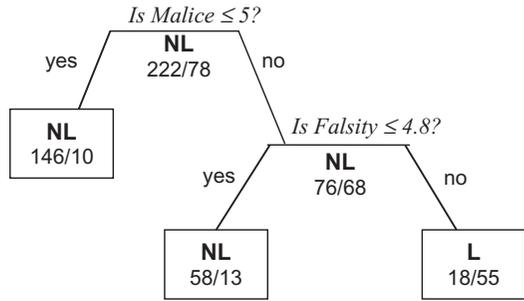
L = Liable (*N* = 78)
NL = Not liable (*N* = 222)
Modal Category: 78% Not liable
Resubstitution Classification Rate: 86%
    Proportional Reduction in Error: 38%
Cross-Validated Classification Rate: 83%
    Proportional Reduction in Error: 37%

NOTE: For each analysis, *falsity* and *malice* are both coded on a scale of 1–10 (see the text for the details of their creation), which is based on a simulation. For the logit, standard errors are given in parentheses. For the tree, the letters under each node give the modal classification of the cases that reach the node, which is the same as those cases' predicted class, assuming the tree ended at that node. The numbers under each node respectively denote the total number of cases that receive the "not liable" [NL] (which is the modal category) and "liable" [L] classification at that node—that is, the cases that have reached that particular node of the tree. For instance, the numbers "76/68" under the node "*Is Falsity ≤ 4.8?*" indicate that 76 cases that reached the node are classified "not liable," while the remaining 68 cases that reached the node are classified "liable." Thus, a total of 144 cases reached that node. Terminal nodes are indicated by squares. The tree predicts 86 percent of cases correctly. See footnote 18 for details on classification rates, and see the text for more details on interpreting the tree.

percent of the cases, for a proportional reduction in error of 30 percent. With these results, an analyst could confidently say that as *malice* and *falsity* increased, so would the probability that a court would find the defendant liable. The classification tree, however, performs even better, classifying 86 percent of cases correctly.

Of perhaps more interest is how each procedure captures the logical structure of the legal rule that generated the data. Figure 4 depicts the two-dimensional case space, along with the estimated partition implied by the logit and classiction tree results (the logit partition is based on the values of the predictors at which the probability of each outcome is 50 percent). While neither procedure perfectly partitions the case space, the tree comes much closer to capturing the structure of the rule than does the logit estimate.[15] With the flexibility to directly partition the case fact space, the tree can closely mimic the actual partition set forth in the rule, whereas the logit structure misrepresents the rule.

This simple example helps illustrate how classification trees may be applied to the study of legal rules. Once one moves to a case fact space with more than two or three

[15]Note that if we could observe the data-generating process without error, the classification tree would in fact produce a perfect partition.

Figure 4: A two-dimensional case fact space with the estimated partitions for the simulated liable cases generated with error.
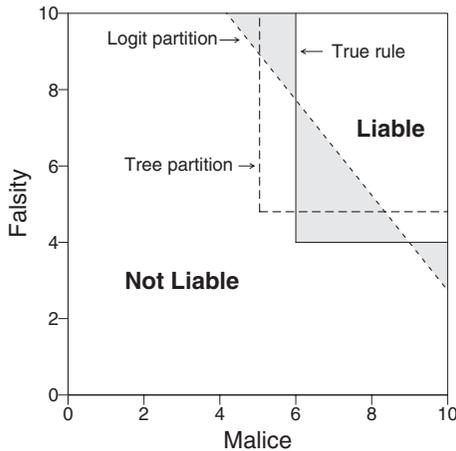
NOTE: The solid line indicated by the "True Rule" labels shows the true partition according to the simulated legal rule, while the dotted line represents the estimated partition from the classification tree and the dotted line marked by "Logit Partition" gives the estimate from the logit. The shaded regions are those in which the logit partition incorrectly classifies cases; that is, areas in which the logit model would predict a *liable* classification when the true classification is *not liable*, and vice versa. *Falsity* and *malice* are both coded on a scale of 1–10 (see the text for the details of their creation).

dimensions, the partitioning of the space becomes impossible to depict. A classification tree, however, can be depicted no matter how many dimensions are relevant, making it an attractive option for interpreting complex data (Hastie et al. 2001:268). Moreover, classification trees' comparative advantage in capturing the structure of legal rules is likely to increase as one moves to the study of actual judicial decisions, for which multiple case facts are likely to be relevant.

To be sure, the classification tree method is not without its disadvantages, compared to maximum likelihood procedures, including logit and probit. The estimates from the latter inherently capture uncertainty in their estimates, whereas uncertainty in classification trees must be measured either through cross-validation procedures or by examining out-of-sample predictions. In addition, trees can be more unstable than more traditional methods. As such, they should be seen as a complement to existing methods, rather than a replacement. Because logit and probit analysis of judicial decisions is already more commonly familiar, in the applications below I focus exclusively on tree-based methods and their potential to increase our understanding of judicial decisions and legal rules.

V. APPLYING CLASSIFICATION TREES

To explore whether classification trees can improve our understanding of legal doctrine, I first examine the Supreme Court's search and seizure cases from 1962 to 1984. I then

examine decisions of the courts of appeals in confession cases from 1946 to 1981. The former allows for a study of an important and much-analyzed Supreme Court doctrine. This application, however, is limited somewhat by the fact that the Court hears relatively few cases, even in one of the more active areas of the law such as search and seizure. Studying lower court decisions mitigates this problem and allows for a cleaner study of legal rules and case partitioning since the majority of the circuit court dockets are made up of less difficult or unusual cases than those heard by the Supreme Court.

With respect to the Supreme Court cases, I use as the unit of analysis the Court's final vote on the merits. If a legal doctrine is to be effective in guiding lower courts and other actors, it must be revealed through the Court's majority opinions (Friedman 2006).

A. *Search and Seizure Cases*

In analyzing the Court's decisions, I begin with the earliest period both for which data are available and there exists a sufficient number of cases to undertake a reliable empirical analysis (the 1962 to 1972 terms) and then sequentially include more decisions for later years up through the 1984 term. This research strategy allows for a detailed and structural analysis of how the Court's Fourth Amendment doctrine evolved over the course of this period. Another option, of course, would be to split the data into mutually exclusive groups and compare trees across each group. This approach would result in much smaller sample sizes, limiting the efficiency of the analysis. Substantively, this approach would also ignore the fact that past cases influence the development of a legal rule, even if a shift in doctrine dramatically alters the rule. Using a sequential approach, I can visually depict how the Burger Court gradually chipped away at the liberal search and seizure regime of the Warren Court.[16]

I follow the coding of Segal and Spaeth (2002:318) in analyzing the Court's decisions. The response variable is the direction of the Court's decision in each case: it can either find the search in question "reasonable" (a conservative decision) or "unreasonable" (a liberal decision). Thus, reasonable and unreasonable constitute the two classes in search and seizure decisions. I use most of the same case facts as predictors employed by Segal and Spaeth, all of which are dichotomous. If the description of each is satisfied, the variable is coded 1, 0 otherwise:[17]

- If the search was conducted in a home.
- If the search was conducted in a business.

-----

[16]I end my analysis at the 1985 term because an inspection of Segal's data (updated through the 2003 term) reveals that his model fails to achieve statistical significance when only cases decided after 1985 are considered. (Similarly, a tree analysis finds no structure in the cases from the later period.) This insignificance could result from the fact that the model, which was originally devised in the early 1980s, has possibly become anachronistic due to the Court's increasingly conservative decisions in this area of the law.

[17]I do not, however, include the following variables used by Segal and Spaeth: the lower court's determination of probable cause, because this is a judicial determination made ex post rather than a case fact; whether the search was conducted incident to an unlawful arrest, because there is little theoretical reason to think it has an effect on the Court's determination of reasonableness and has not been statistically shown to have one (Segal 1984:893); and the lower court's decision, since this is a control variable and not a case fact.

- If the search was conducted on one's person.
- If the search was conducted in a car.
- If the search was a full search, as opposed to a less extensive intrusion.
- If the search was conducted incident to arrest.
- If the search was conducted after a lawful arrest.
- If an exception to the warrant requirement existed (beyond that of search incident to a lawful arrest).

1. 1962–1972 Terms

To study the Court's overall votes, I begin with the Court's decisions in the 1962–1972 terms, which comprises the latter period of the Warren Court and the first few years of the Burger Court. This period begins shortly after the Court extended the exclusionary rule to the states in *Mapp v. Ohio*, a decision that marks the departure point for modern-day Fourth Amendment jurisprudence. It ends shortly after the appointments of Justices Blackmun, Powell, and Rehnquist, at which time the Court had only just begun its shift to the right.

I present the classification tree for the Court's search and seizure decisions in the 1962–1972 terms in Figure 5. Note that not every variable is included in the tree; the model will only select enough splits up to a point where the tree will not overfit the data (Breiman et al. 1984:ch. 3). The presentation of the tree is the same as the example displayed earlier. The tree classifies cases according to the following procedure, from top to bottom, beginning at the root node: if the search was only partial, find the search reasonable. If a full search, was it in a business? If yes, find it unreasonable. If no, was it in a home? If so, find it unreasonable. If not, was it on one's person? If so, find it unreasonable. If not, find it reasonable. The numbers at each node give the breakdown of the true direction of all the cases that reach that node, with the number of searches where the Court found the search unreasonable given first, while the label "U" for unreasonable or "R" for "reasonable" give the predicted class at each node. For example, at the nonterminal node immediately after the right-hand branch proceeding from the first node, of the 56 cases that reach that node, 35 were actually found unreasonable, and 21 were actually found reasonable. Accordingly, the labels at each nonterminal node give the predicted class of each case that reaches that particular node, while the breakdowns in each terminal node indicate how well the tree is classifying cases that reach that node. For example, at the right-most terminal node, eight of the nine cases that reach it are correctly classified "unreasonable." The left terminal node following—"On a person?"—performs less well, as only 10 of the 16 cases that reach it are correctly classified as "reasonable."[18]

―――――

[18]Summing the results of this procedure at each terminal node gives an estimate of the classification rate, known as the resubstitution estimate. In this procedure, the resulting tree is used to classify the same sample that created the tree (Feldesman 2002:264). (This is analogous to generating the percent of cases predicted correctly in a logit by using predicted probabilities from the same in-sample cases.) This procedure generates an upward bias in the classification estimate because the same cases are used for both estimation and classification. Ideally, one would generate a classification tree from a training set and then generate a true and unbiased classification rate by running a test set, which is independent of the learning set, through the estimated tree. This procedure, however, is not feasible in small samples. A compromise option is to undertake a *V*-fold cross-validation, where the entire training set

Figure 5:   Classification tree analysis of the Supreme Court's search and seizure decisions, 1962–1972 terms.
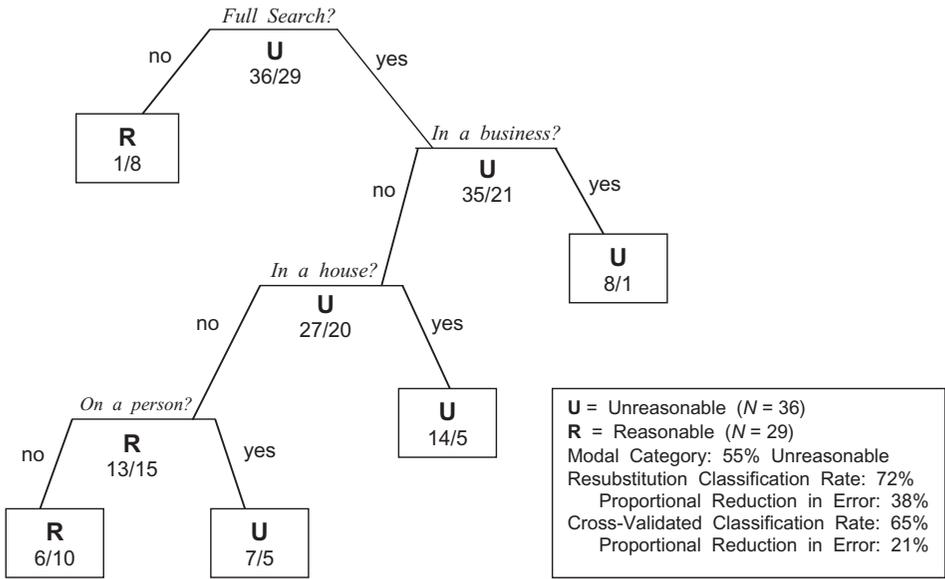

NOTE: At each split, the presence or absence of each variable listed at the top of the split sends a case down the tree, until it reaches a terminal node. Terminal nodes are marked by squares. The letters under each node give the modal classification of the cases that reach the node, which is the same as those cases' predicted class, assuming the tree ended at that node. The numbers under each node denote the total number of cases that receive the "unreasonable" classification (the modal category) and the "reasonable" classification, respectively, at that node. For instance, the numbers under the node marked "*Full Search?*" indicate that 36 cases in the data (since all cases reach the first node by definition) were classified as "Unreasonable," while 29 cases were classified as "Reasonable." The data are from Segal and Spaeth (2002).

The tree suggests that two dimensions were key in the Court's decision making during this period: the extent of the search and whether the defendant had a property interest, as the last three predictors that the tree uses all involve a search taking place on the defendant's property. First, if a search was only partial, the Court was all but certain to deem it reasonable. Conversely, full searches were given careful scrutiny, and if the search invoked a property interest, the Court was likely to find it unreasonable. The tree predicts that only full searches that did not invoke a property interest would be found reasonable.

———

is divided in *V* groups randomly. "[E]ach of the *V* groups is . . . set aside to serve temporarily as an independent test sample and a tree is grown . . . using the other V − 1 groups. In all, *V* trees are grown in this way, and for each tree the set aside portion of the data is used to obtain a test estimate of the misclassification rate for the tree grown from the entire learning sample using the same criteria" (Sutton 2005:312). The cross-validation estimate is an unbiased and independent estimate of the true classification rate (Feldesman 2002:264–65). For each tree, I report both the resubstitution rate and a 10-fold cross-validation estimated rate. The latter, however, will depend on the particular sampling of cases analyzed in each fold. I therefore simulate the cross-validation procedure 1,000 times—each time using a 10-fold validation—and report the median estimate for each tree.

Figure 6: Estimated case partitioning in Supreme Court's search and seizure decisions, 1962–1972 terms.
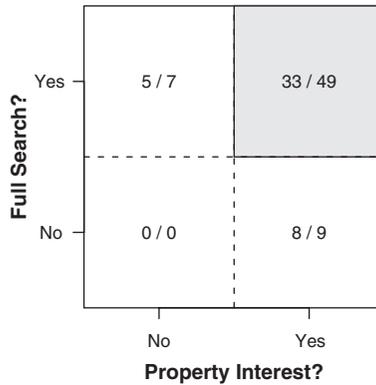
NOTE: The figure depicts a two-dimensional case partitioning resulting from a classification tree analysis in which only two predictors are used: the extent of the search and a property indicator. The shaded region indicates the "unreasonable" portion of the case space; that is, the area where the tree predicts that a case will be decided as unreasonable. For each region, the numbers show, respectively, the number of cases in each region that the tree classifies correctly and the total number of cases. For example, 8 of the 9 cases in the bottom-right quadrant were classified correctly. Using just two predictors results in 71 percent of cases being correctly classified. The data are from Segal and Spaeth (2002).

If we were to suggest a simple legal rule governing search and seizure law at the time, it would look something like the following. (1) If the search is a partial search, it is reasonable. (2) If it is a full search, it is only reasonable when not conducted in a home, business, or on one's person.

Accordingly, the tree suggests a generalized two-dimensional relevant fact space with partitions between full and partial searches, and searches with and without property involved. To illustrate this, I created a new variable indicating whether or not the defendant had a property interest, and then created a new tree using just that variable and whether the search was a full search as predictor. The resulting case partition is depicted in Figure 6, with the "unreasonable" space of the partition indicated with shading. For each region, the second number shows the number of cases that actually fall into that respective region, while the first number shows the number of those cases that the tree correctly classifies. Interestingly, the majority of the case space receives the reasonable classification under the estimated legal rule, suggesting a somewhat conservative rule. However, 75 percent of the cases the Court heard in this period fell into the unreasonable category, accounting for the Court's overall liberalism during this period.

2. 1962–1976 Terms

By the end of the 1976 term, Justices Blackmun, Powell, and Rehnquist had ruled in a half-decade's worth of search and seizure cases. In addition, Justice Stevens replaced Justice Douglas in 1975, further tilting the Court and its Fourth Amendment jurisprudence to the right. Figure 7 depicts the classification tree for the Court's decisions in the 1962 to 1976

Figure 7:   The Supreme Court's search and seizure decisions, 1962–1976 terms.
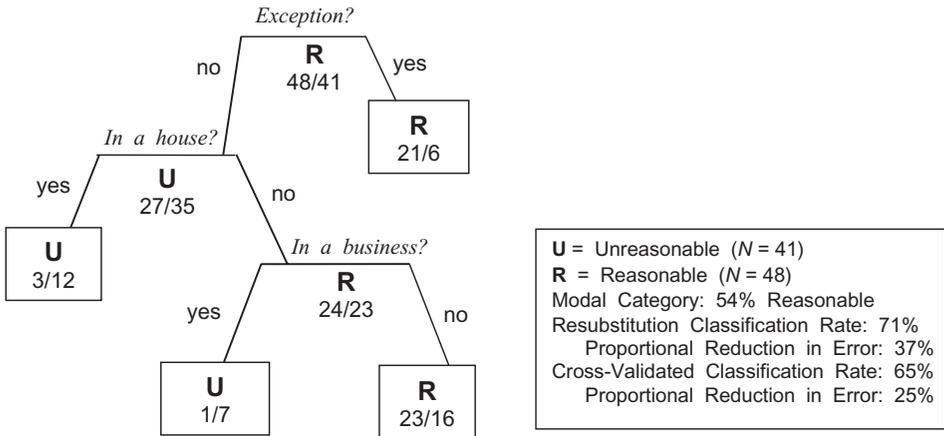

NOTE: At each split, the presence or absence of each variable listed at the top of the split sends a case down the tree, until it reaches a terminal node. Terminal nodes are marked by squares. The letters under each node give the modal classification of the cases that reach the node, which is the same as those cases' predicted class, assuming the tree ended at that node. The numbers under each node denote the total number of cases that receive the "reasonable" classification (the modal classification) and the "unreasonable" classification, respectively, at that node. For instance, the numbers under the node marked "*In a business?*" indicate that 24 cases that reached the node were classified as "Reasonable," while 23 cases that reached the node were classified as "Unreasonable." The data are from Segal and Spaeth (2002).

terms. The first split occurs at whether an exception existed to the warrant requirement; if it did, the tree predicts a conservative decision. Two property variables follow, with nonexcepted searches (either with warrants or without) of a home or business predicted to be deemed unreasonable.

This analysis adds only four terms' worth of decisions to the previous one, but the tree demonstrates the shift toward a more conservative legal rule for search and seizures. Notably, presaging the Burger Court's (and later the Rehnquist Court's) move toward carving greater exceptions to the exclusionary rule, whether an exception existed or not best splits the cases. Also of note, in this second tree many more cases (27, or 30 percent), fall into the first terminal node than in the first tree, suggesting a more categorical conservative rule in the former. Conversely, while searches of homes and businesses are still predicted to be granted as categorically unreasonable (if an exception does not exist), many fewer cases fall into these terminal nodes compared to those in the first tree. In addition, searches of persons receive less protection, shrinking the region of unreasonableness for searches of property. So while the structure of the second tree resembles that of the first, the former still reveals a significant shift in doctrine.

Figure 8 depicts the estimated two-dimensional case space from the 1962–1972 terms, using as the following case facts: whether an exception to the warrant requirement existed and whether the defendant had a property interest. As the partitioning indicates, only searches in which no exceptions existed and the defendant did not have a property interest fall in the "unreasonable" portion of the space (the shaded region).

Figure 8: Estimated case partitioning in Supreme Court's search and seizure decisions, 1962–1976 terms.
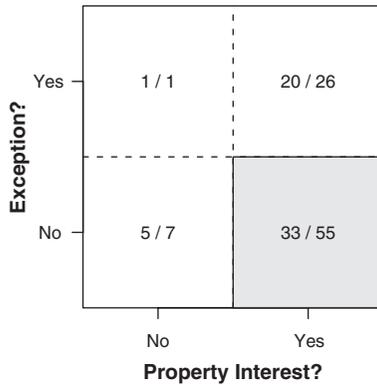
NOTE: The figure depicts a two-dimensional case partitioning resulting from a classification tree analysis in which only two predictors are used: whether an exception to the warrant requirement existed and a property indicator. The shaded region indicates the "unreasonable" portion of the case space; that is, the area where the tree predicts that a case will be decided as unreasonable. For each region, the numbers show, respectively, the number of cases in each region that the tree classifies correctly and the total number of cases. For example, 20 of the 26 cases in the top-right quadrant were classified correctly. Using just two predictors results in 66 percent of cases being correctly classified. The data are from Segal and Spaeth (2002).

3. 1962–1984 Terms

Following the appointment of Justice Stevens in 1975, the Court's membership remained stable until the replacement of Justice Stewart with Justice O'Connor in 1981. Sequentially adding an additional term following the 1976 term through the early 1980s reveals little to no change in the structure of the Court's decisions compared to the tree analyzed in the 1962–1976 period. Therefore, the last tree I present of the Court's search and seizure decisions analyzes the 1962 to 1984 terms; it is depicted in Figure 9.

Interestingly, the first split occurs on whether the search took place in a home. But whereas in the two previous trees, such a search was predicted to be unreasonable if the search was either a full search or no exceptions were present, respectively, here searches in a home are only predicted to be held unreasonable if there are no exceptions and a warrant was not obtained prior to the search. Thus, the tree illustrates how the scope of the Court's protection of the privacy interest inherent in having one's home searched continued to shrink in its decisions in the 1970s and early 1980s. Once again, all searches where exceptions are present are predicted to be held reasonable, as are all searches outside the home for which a warrant was obtained. Thus, the tree suggests that by the mid-1980s, for a search to be held unreasonable, at least three case facts had to be present, a stark change from the first tree, in which nearly all property-based searches were predicted to be unreasonable.

The number of predictors in this tree precludes a visual rendering of the case partition, but the changes in the structure of the three trees over a relatively short period of time illustrates how dramatically the Court's doctrine changed over this period.

Figure 9:   The Supreme Court's search and seizure decisions, 1962–1984 terms.
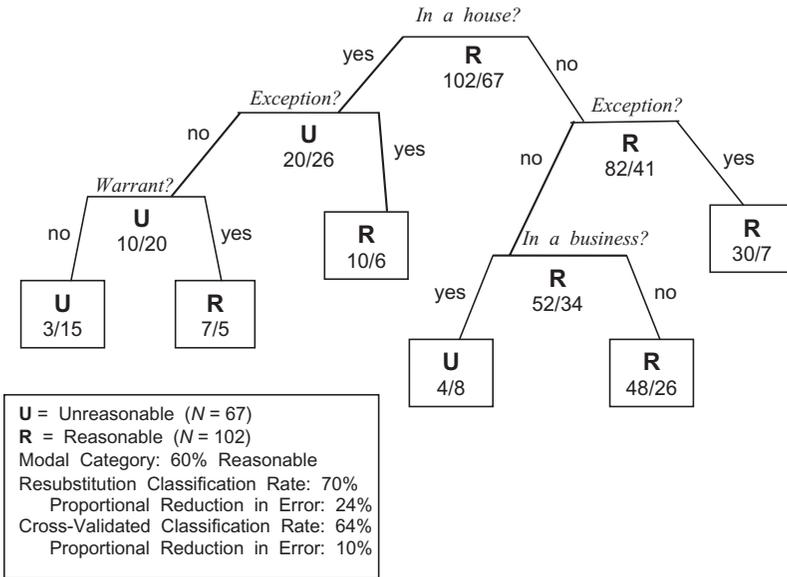
NOTE: At each split, the presence or absence of each variable listed at the top of the split sends a case down the tree, until it reaches a terminal node. Terminal nodes are marked by squares. The letters under each node give the modal classification of the cases that reach the node, which is the same as those cases' predicted class, assuming the tree ended at that node. The numbers under each node denote the total number of cases that receive the "reasonable" classification (the modal classification) and the "unreasonable" classification, respectively, at that node. For instance, the numbers under the node marked "*Warrant?*" indicate that 10 cases that reached the node were classified as "Reasonable," while 20 cases that reached the node were classified as "Unreasonable." The data are from Segal and Spaeth (2002).

B. Confession Cases in the U.S. Courts of Appeals

The Supreme Court creates and develops rules in many area of the law, but it is the lower courts that are tasked with implementing these rules. In recent years, the Court has heard less than 1 percent of the appeals that reach its doors, making the courts of appeals effectively the arbiter of last resort in thousands of cases at the federal level. In addition, while appeals courts do at times decide the "hard cases" that comprise much of the Supreme Court's discretionary docket, the majority of lower court cases are relatively routine, as appeals courts' nondiscretionary docket requires them to hear every appealed case. Thus, while most fact-pattern work has focused on the decisions of the Supreme Court, lower court decisions may lead to cleaner inductions of legal rules. In addition, from a practical standpoint, appeals courts hear many more cases than the Supreme Court, allowing for larger and more robust samples.

To apply classification trees to lower court decision making, I draw on the law of confession cases analyzed by Benesh (2002), which consists of the universe of courts of appeals decisions in confession cases from 1946 to 1981. In these cases, the court's task is to determine whether a confession obtained by police or other state officials was consistent

with the defendant's rights against self-incrimination as given by the Fifth Amendment to the U.S. Constitution. The response variable is whether the appeals court allowed (A) the use of the defendant's confession, or whether it excluded (E) it. Unlike with the search and seizure cases, where the Supreme Court shifted from a liberal to a conservative regime, the circuit courts tended to rule conservatively throughout the period of study, with the percent of liberal decisions never reaching higher than 50 percent in a given year. This tendency likely results from the fact that due to their nondiscretionary dockets, a large number of cases heard by the circuit courts are low probability appeals brought by defendants seeking to having their convictions overturned. Nevertheless, the fact that the circuits did exclude confessions in a nontrivial percentage of cases (22 percent over the entire period) allows for an estimation of legal rules.
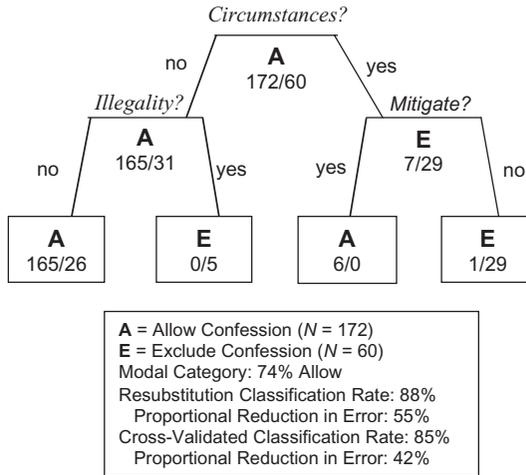
Benesh's analysis includes roughly 25 predictors; in order to avoid having correlated predictors and to make the analysis more generalizable, I chose to condense certain predictors within similar legal categories and create additive scales where appropriate.[19] The predictors I use are as follows.

- *Circumstances.* This variable measures the extent of the psychological and physical circumstances that contribute to the coerciveness of a detention. For each of the following present, 1 is added: if the police used psychological coercion in their questioning of the defendant; if they used physical coercion; if the detention was lengthy; if the place of detention contributed to coerciveness; if the defendant was deprived of basic needs; if the defendant was held incommunicado; and if the police used relays in questioning.
- *Defendant.* This measures any personal characteristics of the defendant that might contribute to or detract from the coerciveness of a confession. For each of the following present, 1 is added: the defendant had a mental deficiency; was a minority; or was a youth. If the defendant had previous experience with the law, 1 is subtracted.
- *Magistrate.* Coded 1 if the defendant was brought before a magistrate before confessing.
- *Procedure.* Coded 1 if the trial court followed adequate procedure for determining the voluntariness of the search.
- *Not Warned*
  - *Before* Miranda. Coded 1 if the police did not warn the defendant of the right to remain silent or to speak to an attorney.
  - *After* Miranda. Whether police failed to read the defendant his or her *Miranda* rights.[20]
- *Attorney.* Coded 1 if the defendant either did not have an attorney or requested one and the request was refused.

---

[19]For more information on the logic behind each predictor, which are based on the Supreme Court's jurisprudence in this area over the same period, see Benesh (2002:40–49).

[20]Due to high correlations between each type of warning, an additive scale is not used for this predictor.

Figure 10:   U.S. Courts of Appeals confession decisions, 1946–1966.

Circumstances?

A
172/60

no / yes

Illegality?

A
165/31

Mitigate?

E
7/29

no / yes   yes / no

A
165/26

E
0/5

A
6/0

E
1/29

A = Allow Confession (*N* = 172)
E = Exclude Confession (*N* = 60)
Modal Category: 74% Allow
Resubstitution Classification Rate: 88%
  Proportional Reduction in Error: 55%
Cross-Validated Classification Rate: 85%
  Proportional Reduction in Error: 42%

NOTE: At each split, the presence or absence of each variable listed at the top of the split sends a case down the tree, until it reaches a terminal node. Terminal nodes are marked by squares. The letters under each node give the modal classification of the cases that reach the node, which is the same as those cases' predicted class, assuming the tree ended at that node. The numbers under each node denote the total number of cases that receive the "allow" classification (the modal classification) and the "exclude" classification, respectively, at that node. For instance, the numbers under the node marked "*Illegality*?" indicate that 165 cases that reached the node were classified as "Allow Confession," while 31 cases that reached the node were classified as "Exclude Confession." The data are from Benesh (2002).

- *Mitigate.* This variable measures the numbers of circumstances that mitigate against the coerciveness of the confession, such as whether the accused waived his or her rights or volunteered the confession.
- *Illegality.* Whether the confession was the fruit of an illegal procedure, such as an illegal search.

In contrast to the Supreme Court search and seizure cases, the number of lower court confession cases across the period of study is large enough to perform separate analyses on nonoverlapping samples. I again analyze three periods that allow for an examination of how legal rules may change over time and, with respect to the lower courts, how they may respond to changes in Supreme Court doctrine: 1946–1966, the year *Miranda* was decided; 1967–1971, during which the courts of appeals heard a large number of cases in response to *Miranda*; and finally, 1972–1981, during which the Supreme Court under Chief Justice Burger rolled back some of the guarantees to defendants granted by the Warren Court (Benesh 2002:37).

1. 1946–1966

The tree for the circuit court confession decisions from 1946 to 1966 is presented in Figure 10. The controlling doctrine in this period, based on the Supreme Court's decisions, was a "totality of the circumstances" test—which one scholar believes "provided little

Figure 11:   U.S. Courts of Appeals confession decisions, 1967–1971.
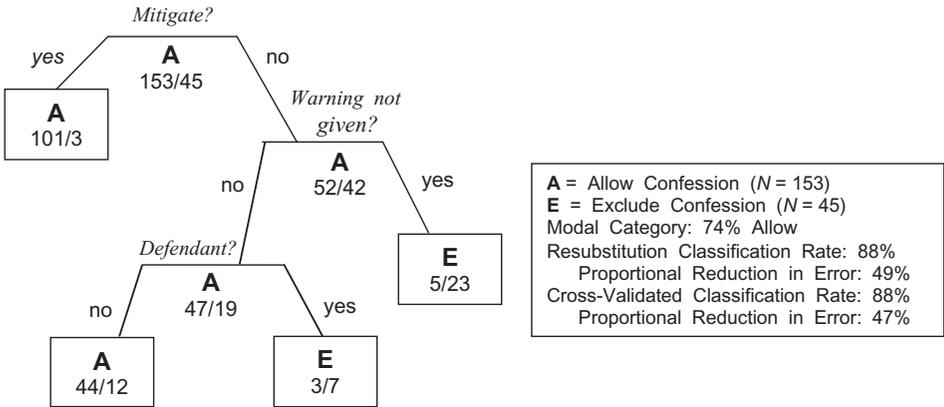
NOTE: At each split, the presence or absence of each variable listed at the top of the split sends a case down the tree, until it reaches a terminal node. Terminal nodes are marked by squares. The letters under each node give the modal classification of the cases that reach the node, which is the same as those cases' predicted class, assuming the tree ended at that node. The numbers under each node denote the total number of cases that receive the "allow" classification (the modal classification) and the "exclude" classification, respectively, at that node. For instance, the numbers under the node marked "*Warning not given*?" indicate that 52 cases that reached the node were classified as "Allow Confession," while 42 cases that reached the node were classified as "Exclude Confession." The data are from Benesh (2002).

guidance for police and lower courts" (O'Brien 2003:1009). The tree provides some support for this assessment, as the majority of cases—those involving confessions in which no coercive circumstances were present and that were not the fruit of an illegality—are clustered at the left-most terminal node, which does not predict outcomes as cleanly as the other terminal nodes. This suggests that in those cases where these elements were not present, the lower courts did undertake case-by-case reviews, while still allowing the confessions the vast majority of the time.

Still, using just three predictors, the tree predicts 88 percent of the cases correctly, for a proportional reduction in error of 55 percent. The first split occurs on whether the confession involved coercive circumstances; if so, and with only one exception, only the presence of mitigating factors prevented a confession from being excluded. If coercive circumstances were not present, then the tree predicts that only confessions that resulted from an illegality would be excluded, and indeed all five such cases did result in exclusions. Note, finally, that whether the police issued a warning to the defendant does not enter the tree. This is not surprising, as the necessity of doing so was not emphatically stated by the Supreme Court until 1966 in *Miranda*.

2. 1967–1971

The classification tree for the circuit court decisions from 1967 to 1971 is depicted in Figure 11. In contrast to the first period, in the years following *Miranda* the question of whether the defendant was warned now enters the tree. However, the first split occurs

Figure 12:   U.S. Courts of Appeals confession decisions, 1972–1981.
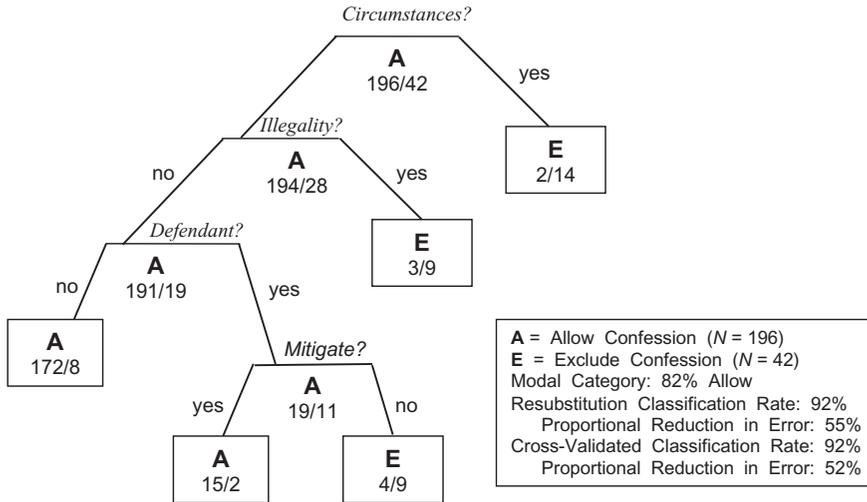
NOTE: At each split, the presence or absence of each variable listed at the top of the split sends a case down the tree, until it reaches a terminal node. Terminal nodes are marked by squares. The letters under each node give the modal classification of the cases that reach the node, which is the same as those cases' predicted class, assuming the tree ended at that node. The numbers under each node denote the total number of cases that receive the "allow" classification (the modal classification) and the "exclude" classification, respectively, at that node. For instance, the numbers under the node marked "*Defendant?*" indicate that 191 cases that reached the node were classified as "Allow Confession," while 19 cases that reached the node were classified as "Exclude Confession." The data are from Benesh (2002).

on whether mitigating factors existed, with the confession being allowed predicted in all cases where such factors were present. If they were not, then the tree predicts that all defendants who were not given their *Miranda* warnings would win their appeals (interestingly, though, in five such cases the confessions were allowed). Thus, while the requirements set forth in *Miranda* ostensibly increased the possibility that lower courts would be more likely to strike down confessions, the appeals courts seemed to focus on the presence or absence of mitigating circumstances, as the former all but ensured that the panel would uphold the confession.[21] The final split occurs on whether the personal characteristics of the defendant contributed to the coerciveness of his or her confession; if so, exclusion is predicted.

3. 1972–1981

The tree for the 1972 to 1981 period is presented in Figure 12. In contrast to the second-period tree, the presence or absence of mitigating factors appears only at the last

———

[21]Litigant strategies may also be at play here. If, as Songer et al. (1994:677) speculate, *Miranda* induced defense attorneys to challenge many more confessions than they would have previously in the hopes of stretching the categorization of involuntary searches, then lower courts not necessarily willing to go farther than the Supreme Court in the area of confessions might stress mitigating factors over factors likely to lead to a liberal decision.

split. In addition, the first split—whether coercive circumstances existed—leads to a categorical prediction of exclude if present. Similar to the first-period tree, if there were no coercive circumstances, a confession that was the fruit of an illegality is predicted to be excluded. The fact these first two splits occur on "liberal" variables and lead to categorical exclusions suggests a more liberal tree in this period. However, if neither of the facts were present, as occurred in the vast majority of cases, the tree becomes decidedly conservative: only confessions where the personal characteristics of the defendant contributed to the coerciveness of his or her confession *and* no mitigating factors existed are predicted to be excluded, a terminal node that only 13 cases (or 5 percent) reached. Finally, and interestingly, the presence or absence of *Miranda* warnings does not appear in the tree, suggesting that the lower courts moved toward a less strict interpretation of the ruling in parallel to the Burger Court's move toward a more conservative doctrine.

VI. Conclusion

This article has highlighted the ability of classification tree analysis to connect the foundational conception of legal rules and its theoretical underpinnings to a methodology suitable for studying them. While classification trees are not destined to replace more traditional statistical analyses of judicial decision making, the applications presented here illustrate how classification trees can increase our understanding of legal doctrine.

Beyond this primary benefit, I believe the classification tree approach has the potential to unify traditional political science and legal conceptions of the law in a way that encompasses the benefits of each. The classification trees presented in this article have shed light on legal doctrine and doctrinal evolution in a way that traditional regression analyses would be hard pressed to capture, while also revealing patterns that a traditional doctrinal analysis of case law could miss in the labyrinth of cases, opinions, and legal battles that characterize any given area of the law.

Finally, the potential for using classification trees extends well beyond the straightforward approach taken in this article. If trees can adequately capture doctrine, then an interesting test involving the judicial hierarchy would be whether lower courts, which generally comply with the Supreme Court, generate similar trees, which would suggest a more sophisticated understanding and following of precedent than traditional tests have shown to date (Songer et al. 1994; Benesh 2002). In addition, trees could be used to examine how the law develops differently across circuits, especially in areas where the Supreme Court has not weighed in (Klein 2002). Second, while I informally analyzed change in the Supreme Court's search and seizure jurisprudence and the courts of appeals' confession jurisprudence, trees may allow for more formal tests of structural change, such as Richards and Kritzer's (2002) jurisprudential regime theory (Capelli & Reale 2007). Finally, fact-pattern analyses typically ignore issues of preference aggregation that arise on collegial courts (Lax 2007). A justice-level analysis using classification trees could be used to examine how the rules of the individual members of a multimember court aggregated to achieve a single legal rule.

References

Amar, Akhil Reed (1994) "Fourth Amendment First Principles," 107 *Harvard Law Rev.* 757.

Andreou, A. S., & G. A. Zombanakis (2001) "A Neural Network Measurement of Relative Military Security: The Case of Greece and Cyprus," 12 *Defence & Peace Economics* 303.

Beck, Nathaniel, Gary King, & Langche Zeng (2000) "Improving Quantitative Studies of International Conflict," 94 *American Political Science Rev.* 21.

—— (2004) "Theory and Evidence in International Conflict: A Response to De Marchi, Gelpi, and Gyrnaviski," 98 *American Political Science Rev.* 379.

Benesh, Sara C. (2002) *The U.S. Court of Appeals and the Law of Confessions: Perspectives on the Hierarchy of Justice.* New York: LFB Scholarly Publishing.

Berch, Michael A., Rebecca White Berch, & Ralph S. Spritzer (2002) *Introduction to Legal Method and Process: Cases and Materials*, 3rd ed. St. Paul, MN: West Publishing Co.

Brambor, Thomas, William R. Clark, & Matt Golder (2006) "Understanding Interaction Models: Improving Empirical Analyses," 14 *Political Analysis* 63.

Breiman, Leo (2001) "Statistical Modeling: The Two Cultures," 16 *Statistical Science* 1999.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshern, & Charles L. Stone (1984) *Classification and Regression Trees.* Belmont, CA: Wadsworth International Group.

Cameron, Charles M., & Lewis A. Kornhauser (2005) "Modeling Law: Theoretical Implications of Empirical Methods," paper presented at the NYU Law School Conference on Modeling Law. New York.

Cameron, Charles M., Jeffrey A. Segal, & Donald R. Songer (2000) "Strategic Auditing in a Political Hierarchy: An Informational Model of the Supreme Court's Certiorari Decisions," 94 *American Political Science Rev.* 101.

Capelli, Carmela, & Marco Reale (2007) *Detecting Multiple Structural Breaks in the Mean with Atheoretical Regression Trees*, University of Canterbury working paper.

Clark, Linda A., & Darly Pregibon (1992) "Tree-Based Models," in John M. Chambers & Trevor J. Hastie, eds., *Statistical Models in S*, pp. 377–419. South Pacific Grove, CA: Wadsworth & Brooks.

Eisenberg, Theodore, & Geoffrey P. Miller (2007) "Do Juries Add Value? Evidence from an Empirical Study of Jury Trial Waiver Clauses in Large Corporate Contracts," 4 *J. of Empirical Legal Studies* 539.

Emmert, Craig F. (1992) "An Integrated Case-Related Model of Judicial Decision Making: Explaining State Supreme Court Decisions in Judicial Review Cases," 54 *J. of Politics* 543.

Feldesman, Marc R. (2002) "Classification Trees as an Alternative to Linear Discriminant Analysis," 119 *American J. of Physical Anthropology* 257.

Friedman, Barry (2006) "Taking Law Seriously," 4 *Perspectives on Politics* 261.

George, Tracey E., & Lee Epstein (1992) "On the Nature of Supreme Court Decision Making," 86 *American Political Science Rev.* 323.

Gordon, A. D. (1999) *Classification*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.

Hagle, Timothy M. (1991) "But Do They Have to See it to Know It: The Supreme Court's Obscenity and Pornography Decisions," 44 *Western Political Q.* 1039.

Hall, Melinda G., & Paul Brace (1996) "Justices' Response to Case Facts: An Interactive Model," 24 *American Politics Q.* 237.

Hastie, Trevor, Robert Tibshirani, & Jerome Friedman (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York: Springer-Verlag.

Holmes, Oliver Wendell Jr. (1897) "The Path of the Law," 10 *Harvard Law Rev.* 457.

Ignagni, Joseph A. (1994) "Explaining and Predicting Supreme Court Decision Making: The Burger Court's Establishment Clause Decisions," 36 *J. of Church & State* 301.

Klein, David E. (2002) *Making Law in the United States Courts of Appeal.* Cambridge: Cambridge Univ. Press.

Kornhauser, Lewis A. (1992) "Modeling Collegial Courts II: Legal Doctrine," 8 *J. of Law, Economics & Organization* 441.

Kort, Fred (1957) "Predicting Supreme Court Decisions Mathematically: A Quantitative Analysis of the 'Right to Counsel' Cases," 51 *American Political Science Rev.* 1.

Kritzer, Herbert M., & Mark J. Richards (2003) "Jurisprudential Regimes and Supreme Court Decisionmaking: The Lemon Regime & Establishment Clause Cases," 37 *Law & Society Rev.* 827.

—— (2005) "The Influence of Law in the Supreme Court's Search-and-Seizure Jurisprudence," 33 *American Politics Research* 33.

Lax, Jeffrey R. (2007) "Constructing Rules on Appellate Courts," 101 *American Political Science Rev.* 591.

Levi, Edward (1949) *An Introduction to Legal Reasoning.* Chicago, IL: Univ. of Chicago Press.

Llewellyn, K. N. (1951) *The Bramble Bush.* New York: Oceana Publications.

McGuire, Kevin T. (1990) "Obscenity, Libertarian Values, and Decision Making in the Supreme Court," 18 *American Politics Q.* 47.

Moylan, Charles E. Jr. (2003) "A Conceptualization of the Fourth Amendment," in William W. Greenlaugh, ed., *The Fourth Amendment Handbook: A Chronological Survey of Supreme Court Decisions* pp. 1–21. Chicago, IL: American Bar Association.

O'Brien, David M. (2003) *Constitutional Law and Politics: Volume Two, Civil Rights and Liberties*, 5th ed. New York: W.W. Norton & Co.

Perry, H. W. Jr. (1991) *Deciding to Decide: Agenda Setting in the United States Supreme Court.* Cambridge: Harvard Univ. Press.

Richards, Mark J., & Herbert M. Kritzer (2002) "Jurisprudential Regimes in Supreme Court Decision Making," 96 *American Political Science Rev.* 305.

Ripley, Brian D. (1996) *Pattern Recognition and Neural Networks.* Cambridge: Cambridge Univ. Press.

Ruger, Theodore W., Pauline T. Kim, Andrew D. Martin, & Kevin M. Quinn (2004) "The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking," 104 *Columbia Law Rev.* 1150.

Segal, Jeffrey A. (1984) "Predicting Supreme Court Cases Probabilistically: The Search and Seizure Cases, 1962–1981," 78 *American Political Science Rev.* 891.

Segal, Jeffrey A., & Cheryl D. Reedy (1988) "The Supreme Court and Sex Discrimination: The Role of the Solicitor General," 41 *Western Political Q.* 553.

Segal, Jeffrey A., & Harold J. Spaeth (2002) *The Supreme Court and the Attitudinal Model Revisited.* New York: Cambridge Univ. Press.

Songer, Donald R., & Susan Haire (1992) "Integrating Alternative Approaches to the Study of Judicial Voting: Obscenity Cases in the United States Courts of Appeals," 36 *American J. of Political Science* 963.

Songer, Donald R., Jeffrey A. Segal, & Charles M. Cameron (1994) "The Hierarchy of Justice: Testing a Principal-Agent Model of Supreme Court-Circuit Court Interactions," 38 *American J. of Political Science* 673.

Sunstein, Cass R. (1996) *Legal Reasoning and Political Conflict.* New York: Oxford Univ. Press.

Sutton, Clifton D. (2005) "Classification and Regression Trees, Bagging, and Boosting," in C. R. Rao, Edward Wegman, & Jeffrey Solka, eds., *Handbook of Statistics 24: Data Mining and Data Visualization*, pp. 303–29. Amsterdarn: North Holland.

Swinford, Bill (1991) "A Predictive Model of Decision Making in State Supreme Courts: The School Financing Cases," 19 *American Politics Q.* 336.

Therneau, Terry M., & Elizabeth J. Atkinson (1997) *An Introduction to Recursive Partitioning: Using the* Rpart *Routines.* Mayo Foundation Technical Report. Available at <http://mayoresearch.mayo.edu/mayo/research/biostat/upload/61.pdf>.

Twining, William, & David Miers (1999) *How to Do Things with Rules.* London: Butterworths.

Venables, William N., & Brian D. Ripley (2002). *Modern Applied Statistics with S: Fourth Edition.* New York: Springer-Verlag.

Wahlbeck, Paul J. (1997) "The Life of the Law: Judicial Politics and Legal Change," 59 *J. of Politics* 778.

—— (1998) "The Development of a Legal Rule: The Federal Common Law of Public Nuisance," 32 *Law & Society Rev.* 613.

Appendix: An Introduction to Classification Trees

In this appendix I provide a brief technical introduction to classification trees, as used in this article. This review draws heavily from the sources cited therein; see them for more detailed discussions. Note also that I omit any discussion of regression trees, which are used when the response variable is continuous. While regression trees are similar in structure to classification trees, the details of their construction differ. Finally, some of the procedural details may differ if a "test set" of data is available to test the tree produced by a "learning sample"; this requires having a large data set. Since all the trees presented in the article are based on cross-validated construction, in which the same data are used to both construct and test a tree, I only discuss that procedure below.

The process of constructing a classification tree is fairly intuitive: "[F]irst the single variable is found which best splits the data into two groups. . . . The data is separated, and then this process is applied *separately* to each sub-group, and so on recursively until the subgroups either reach a minimum size . . . or until no improvement is made" (Therneau & Atkinson 1997:4). If this tree is too large, steps are taken to prune the tree back to an acceptable size; this process is described below.

A. Splitting Rules and Criteria

The first issue in constructing the tree is to determine what splitting criterion should be used, and what type of splits to allow. With respect to the latter, while some programs allow more than two splits at each nonterminal node, *rpart* only allows binary splits, which avoids having to normalize splits by size to compare them (Venables & Ripley 2002:255). In addition, while some programs allow either linear combination splits for continuous variables and Boolean splits for categorical variables, *rpart* allows just single variable splits; this eases interpretation, decreases computing time, and keeps the resulting tree invariant to monotone transformations of the variables (Sutton 2005:309–10).

With respect to a splitting criterion, at each node the partitioning proceeds by choosing a split that minimizes the *impurity* of the resulting nodes (the node being split is referred to as the "parent" node, while the nodes resulting from the split are called "child" nodes). In other words, the split maximizes the homogeneity of χ. More formally, the process is governed by an impurity function, which is based on the proportions of the data belonging to all possible classes of the response variable; denote these proportions $p_1, p_2, \ldots, p_{J\bar{1}}, p_J$, where J is the number of classes (Sutton 2005:310). Intuitively, this function is minimized when a child node is completely pure, in which all observations are of the same class. Conversely, the function is maximized when a child node contains equal numbers of observations from each possible class. The default impurity function in *rpart*, and the one used for all the trees in this article, is known as the *Gini index of diversity*, which is defined as follows:

$$g(p_1, \ldots, p_J) = 1 - \sum_{j=1}^{J} p_j^2. \tag{A1}$$

Thus, for a two-class problem, such as those considered in this article, if at a node there are 10 observations from class 1 and 20 observations from Class 2, then the Gini index =

$1 - [(10/30)^2 + (20/30)^2] = 0.44$. Beginning with the complete set of observations, the tree selects the split that most decreases the Gini index. This process is repeated at every subsequent node until that node either contains observations all from the same class, or is too small to continue splitting.[22]

B. Pruning the Tree

The splitting, or recursive partitioning, process frequently results in a tree that is too large, thus overfitting the data. Each additional split results in a loss of parsimony as well; the goal is to find a tree that accurately describes the data as parsimoniously as possible. One option would be to employ some sort of "stop-splitting rule" that would govern the construction of the tree from the top down. Beginning with Breiman et al. (1984), it is generally viewed as preferable to completely grow out a tree as far as possible (based on one's splitting criterion) and then prune it back. The most common procedure, and the one employed by *rpart*, is known as cost-complexity pruning. Let $|T|$ equal the number of terminal nodes of a tree, T, and $R(T)$ be the resubstitution estimate of the misclassification rate of T—the rate at which observations are classified incorrectly after being sent through the decision rules in T. For every $\alpha \geq 0$, the cost-complexity measure is defined as:

$$R_\alpha(T) = R(T) + \alpha |T|. \tag{A2}$$

The goal of pruning is to produce a pruned tree that minimizes this measure (Breiman et al. 1984). α is called the "complexity" parameter; it measures the cost in complexity of adding additional leaves to the tree. This parameter captures the inherent tradeoff between increasing the classification rate of the tree while privileging smaller trees for their parsimony. When α is zero, the unpruned tree minimizes $R_\alpha(T)$; as $\alpha \to \infty$ it is minimized by a one-node "tree" (which is not really a tree, but just a prediction based on the modal class). "Since the resubstitution estimate of misclassification rate is generally overoptimistic and becomes unrealistically low as more nodes are added to a tree, it is hoped that there is some value of α that properly penalizes the overfitting of a tree which is too complex, so that the tree which minimizes $R_\alpha(T)$, for the proper value of α, will be a tree of about the right complexity" (Sutton 2005:313).

    Rpart uses the following cross-validation procedure to choose the proper value for α. First, an unpruned tree is grown using all the observations; it has been shown that one of the nested subtrees of this tree is the tree that minimizes the cost-complexity measure (Sutton 2005:314). Next, the data are divided into k parts of roughly equal size; the default in *rpart* is 10, which is used for all the trees in this article. Then, the first subset of the data is set aside, and a nested sequence of trees is grown using the $V - 1$ portion of the data (or 90 percent, when $v = 10$). Then, for every interval range of α, where the number of intervals is less than or equal to the number of terminal nodes in the unpruned tree, the observations from the k portion are run through the subtree corresponding to each possible value of α, and the classification errors are recorded, as well as the α that leads to the lowest misclassification rate. This process is repeated for all V groups. The results are then

———

[22]For each tree, the minimum number of observations in a node for which a split is computed is 15.

averaged, and the subtree selected is the one that has the smallest estimated cross-validation error rate. In practice, it is advised that one follow the "1-SE Rule," in which the final pruned tree selected is that with the smallest cross-validated error rate within the smallest absolute estimated cross-validated error plus the standard error of that estimate; this procedure has been followed for all the trees presented in this article.[23] Finally, once the 1-SE Rule has been implemented, the unbiased estimate of the true misclassification rate can be obtained from the cross-validated estimates. That figure is reported for each of the trees, along with the resubstitution estimation, which is biased upward.

C. Assigning a Predicted Class to Terminal Nodes

The final decision to make in constructing a classification tree is also the simplest: what procedure should be followed for making a prediction about observations in terminal nodes. The plurality rule is followed, in which the class with the greatest number of observations in a terminal node is predicted. Thus, if Class 1 has more observations than Class 2 in a terminal node, the former is predicted.[24]

─────

[23]Note, however, that the first-period and third-period trees for the search and seizure cases did not require pruning.

[24]This may not be the case if misclassifications are not weighted equally, however. See, for example, Ripley (1996).