A Quantitative Approach to Evaluating International Environmental Regimes

*Ronald B. Mitchell**

Introduction

To date, quantitative analysis has been largely absent from efforts to study the effects of international environmental regimes.[1] Yet, applying statistical procedures to relatively large sets of quantified data offers rich opportunities to address questions central to this research program. Quantitative analysis allows us to answer questions that either cannot be or usually are not answered by other methodologies as well as to reexamine (and buttress or refute) answers to questions already addressed by other methodologies. Quantitative techniques that use careful modeling and appropriate data could shed light on which features of a regime are responsible for a regime's success and which are superfluous, whether the effectiveness of a particular type of regime is problem-contingent, and how a regime's effectiveness varies with international and domestic contexts. Thus, quantitative analysis offers a valuable complement to qualitative techniques in evaluating the determinants of regime effects and effectiveness.

Consider some questions regarding regime effectiveness. Are sanctions always more effective at inducing behavioral change than rewards and, if not, under what conditions are rewards more effective?[2] Are pollution problems, on

* A revised version of this article will appear in Arild Underdal and Oran Young, *Regime Consequences: Methodological Challenges and Research Strategies*, Kluwer Academic Publishers, 2003. This article has benefited greatly from comments from Arild Underdal, Oran Young, Detlef Sprinz, and participants in a conference on "Regime Consequences: Methodological Challenges and Research Strategies" hosted by the Centre for Advanced Study of the Norwegian Academy of Science and Letters in June 2000. This article was completed with the generous support of a Summer Research Award from the University of Oregon

1. For some exceptions, see Meyer et al. 1997; and Downs, Rocke, and Barsoom 1998. Several scholars have put together data sets that code a variety of parameters for a range of environmental treaty regimes. The International Regimes Database (IRD) has begun pulling together an extensive set of data on thirty different treaties which, once completed, will constitute a significant advance in the data that will be available to the policy and scholarly community. Haas and Sundgren examined trends in environmental treaty making (Haas and Sundgren 1993). Dmitris Stevis has collected data on membership and characteristics of international environmental institutions (Stevis 1999).
2. Downs et al. 1996.

Global Environmental Politics 2:4, November 2002

average, more difficult or easier to resolve than wildlife preservation problems? Do demands for new behaviors generally work better or worse than bans on existing behavior?[3] Such questions are difficult to answer convincingly with case studies of single regimes, simply because most regimes do not employ both sanctions and rewards, address both pollution and wildlife problems, or both ban some behaviors and require others. We certainly want to analyze those valuable but rare regimes that exhibit variation on such variables, since they convincingly control many other variables. Yet, existing case studies of environmental regime effectiveness as well as future ones face inherent problems of generalizability. Even commendable recent efforts to draw conclusions across multiple regimes, each analyzed by a different scholar, face difficulties in ensuring convincing comparability across regimes.[4] Carefully designed case studies often generate compelling findings that fit the case studied quite well but usually do so by sacrificing the ability to map those findings convincingly to many, if any, other cases.[5]

Quantitative analysis involves an opposite trade-off. Quantitative analysis can identify general propositions that hold reasonably well across a range of cases, even as they fail to explain any particular case well.[6] Examining many regimes and their consequences can help identify what "tends to happen" in regimes in general and in regimes of particular types. It can tell us whether regimes generally have large effects on behavior, or none at all. It can help "fill in the blanks" left by qualitative analysis, using patterns *across regimes* to clarify why certain types of regimes address certain types of problems better than others, or why regimes in one issue area work better than otherwise-similar regimes in a different issue area. Such comparisons across regimes move us beyond case study insights that a particular type of regime *can* produce a desired outcome to the often more useful claim that such a design *is likely to* produce such an outcome in some other context. They help us move from statements of possibility to statements of probability. Large-N, cross-treaty, comparisons can help us develop claims from qualitative research, for example, refining the general claim that country capacity influences compliance by evaluating whether the lack of a particular capacity inhibits compliance with some types of regimes but not other types.[7] Quantitative techniques offer the promise of replacing claims that "this strategy worked in this historical case" with more convincing policy-relevant and contingent prescriptions of which strategy is likely to work best to address a given problem under given conditions. Although a variety of quantita-

3. Princen 1996.
4. Brown Weiss, and Jacobson collected extensive information on compliance and its determinants for ten countries and five different treaties (Brown Weiss and Jacobson 1998). Miles and Underdal have developed a database of 44 cases involving regime phases or components (Miles et al. 2001).
5. Mitchell and Bernauer 1998.
6. Thus, the convincing, if contested, quantitative finding that democratic states rarely go to war against each other proves unsatisfactory in explaining why any particular war occurs.
7. Haas et al. 1993; Brown Weiss and Jacobson 1998; and Victor et al. 1998.

tive techniques could be used to investigate regime consequences, in this article I delineate one quantitative approach, that of using regression analysis on panel data.[8]

Definitions

Recent work on qualitative methodology in general and counterfactuals in particular reminds us that any attempt to make causal claims requires comparing at least two cases.[9] Here I clarify some terms useful for discussing quantitative study design, generally avoiding the term "case" because of its multiple, often widely divergent, meanings.[10] *Units of analysis* are the entities or phenomena about which the researcher collects data.[11] Units of analysis, often called cases, are a sample from a population or class of all conceptually-similar units that could have been studied. *Variables* are the dimensions, characteristics, or parameters of these units of analysis, with any variable having at least two possible *values*. Quantitative studies seek to evaluate the relationships among the values of variables. Dependent variables (DVs) are those whose variation we seek to explain. Explanatory or independent variables (IVs) are those whose variation we look to as possible explanations of the variation in the DV, based on theoretical claims regarding their causal influence on that DV. Control variables (CVs) are IVs believed to influence the DV that are included in an analysis in order to separate their influence on the DV from that of the primary IV of interest. To avoid confusion, I distinguish between a unit of analysis and an observation. An *observation* is one set (or vector) of the observed values of all variables (IVs, CVs, and DV) for a given unit of analysis. Notice that this definition allows us to speak of multiple observations of a single unit of analysis, as when we observe a regime (the unit of analysis) at several points in time. In a spreadsheet analogy, each column corresponds to a different IV, CV, or DV; each row corresponds to a single observation; the first column would contain a name for each observation; the dataset could contain rows of multiple observations from each unit of analysis as well as observations from multiple units of analysis; and each cell would contain the value of a given variable for a given observation. A quantitative study of regime consequences requires defining some potential consequent of regimes as a dependent variable, the presence or absence of a regime or some regime characteristic as the independent variable of interest, and some set of other factors predicted to affect the dependent variable as control variables. The analyst would then seek out regimes (units of analysis) that allow relatively comparable observations across these IVs, CVs, and DVs.

These definitions suggest that research studies are most usefully distinguished by the number of units of analysis rather than the number of observa-

8. The application of this approach is currently underway and will be reported in future work.
9. King et al. 1994; Fearon 1991; and Biersteker 1993.
10. See, for example, Ragin and Becker 1992; Galtung 1967; King et al. 1994; and Yin 1994.
11. King et al. 1994, 52.

tions. The major virtues and limitations of qualitative "case study" research stem from a reliance on one or a relatively few units of analysis, even when multiple observations are made. Making multiple observations of the same unit of analysis holds many variables constant, but poses obstacles to the analyst's ability to generalize to units of analysis with different values for those variables. The major virtues and limitations of quantitative research stem from a reliance on many different units of analysis, whether or not there are many or few observations of each. Including multiple units of analysis in a study introduces considerable variation in variables we would prefer to hold constant, thereby posing obstacles to the analyst's ability to identify causal relationships that may exist. In short, qualitative studies must overcome serious obstacles to the external validity of their claims whereas quantitative studies must overcome serious obstacles to the internal validity of their claims.

Finally, although recognizing the value of a broader definition, I use regime here to refer to the governance structures surrounding international conventions and treaties, including the norms, rules, principles, and decision-making procedures as well as the numerous actors who bring those components to life.[12] I use the term "subregime," to refer to different rules, compliance strategies, or other features that provide the basis for making analytically useful distinctions among various components of a regime. For example, we can view the stratospheric ozone regime based on the Vienna Convention, the Montreal Protocol, and subsequent amendments as consisting of three subregimes: one related to the ozone depleting substances (ODSs) phaseout commitments of developed states, one related to the ODS phaseout commitments of the developing states, and one related to the commitments of developed states to finance the ODS phaseout of the developing states.

The Contributions and Limitations of Quantitative Analysis

Case-studies have provided considerable evidence that certain regimes have influenced behavior. Indeed, a fairly extensive list now exists of regimes deemed effective or ineffective by thoughtful, well-informed scholars.[13] A quantitative approach, however, allows us to answer more comparative questions that are central to the regime consequences research program but that have, as yet, gone unanswered. What types of regimes are most effective? Why does one regime induce significantly more behavior change than another, apparently comparable, regime? How do contextual factors condition the effectiveness of a particular type of regime or subregime? It is precisely such questions, which involve comparing across regimes and subregimes, that quantitative analysis is particularly well-suited to address.

12. Krasner 1983.
13. Haas et al. 1993; Brown Weiss and Jacobson 1998; Victor et al. 1998, Young 1997; Young 1999a; and Miles et al. 2001.

Quantitative Modeling to Evaluate a Single Regime's Effects

Although ultimately interested in using quantitative analysis to investigate such cross-regime questions, for expository purposes, I start by examining how we could use quantitative analysis to evaluate a single regime's effects. Consider the question of whether the European Convention on Long-Range Transboundary Air Pollution's (LRTAP) Sulfur Protocol was effective at altering the sulfur dioxide (SOx) emissions that contribute to acid precipitation.[14] Although various approaches could be taken to this problem, one approach would involve identifying an econometric model that examines how SOx emissions (the DV) covary with membership in the convention (the IV). Several years of SOx emission data for various countries could be regressed on corresponding data of when, if ever, those countries ratified the Sulfur Protocol.

How we specify the model depends on what we seek to accomplish. Although we might want to accurately estimate the variation in emissions, in the present context I assume the analytic goal is to accurately estimate the effect of regime membership on emissions. Given that, we need only include those variables on the right hand side of the equation that correlate with both membership and emissions. Failing to do so would violate the assumptions underlying regression analysis and lead to misestimating the effect of regime membership. In the current context, we want to include other influences on sulfur emissions, such as population, coal usage, and energy efficiency, to avoid introducing omitted variable bias into our estimates of the influence of membership. Thus, we could specify a model as follows:

$$\text{EMISS} = \alpha + \beta_1 * \text{MEMBER} + \beta_2 * \text{POPN} + \beta_3 * \text{COAL} + \beta_4 * \text{EFFIC} + \ldots + \beta_N * \text{OTHER} + \epsilon$$

where EMISS is annual emissions of sulfur dioxide, MEMBER is coded as 0 in years of nonmembership and 1 in years of membership, and POPN, COAL, and EFFIC are annual data for population, coal usage, and energy efficiency, with OTHER allowing the inclusion of other influences on sulfur emissions as well.

What would the results from such a model, or similar models for other regimes, tell us? β_1 represents the difference in average emissions that (if we have modeled emissions correctly) can be attributed to membership, a number we would predict to be negative, on the assumption that membership leads states to reduce their emissions. The t-statistic on β_1 indicates the likelihood that this difference in average emissions would have occurred by chance. Although good qualitative analysis also assesses the likelihood that the observed outcome could have occurred by chance, quantitative analysis encourages prior establishment of a criterion (by convention, a probability of 5%) of whether to interpret an observed covariation of an IV with the DV as random or as resulting from a systematic, and presumably causal, effect of the IV on the DV.[15] For IVs that have

14. Levy 1993; Levy 1995; and Sprinz 1998.
15. The criteria usually viewed as necessary to infer a causal relationship between A and B are demonstrating "relationship" (co-variation of the values of A with the values of B), "time prece-

such "statistically significant" t-statistics (and for which independent theoretical support exists for making causal claims), the β can be interpreted as the average magnitude of the "effect" the IV has on the DV, having controlled for all other IVs.[16] It is important to distinguish the *statistical significance* of the t-statistic from the meaningfulness or what we might call *policy significance* of that IV. Thus, a study might show that members emitted only slightly less pollution than nonmembers but provide convincing support that their lower emissions levels cannot be readily explained by factors other than their membership in the regime. A t-statistic indicates whether the co-variation of IV and DV was "real" (more precisely, whether it was likely to have occurred by chance) while the β indicates whether the co-variation of the IV and DV was "large."

The coefficient on membership, β_1, therefore, corresponds to the counterfactuals of qualitative analyses. Counterfactual emissions for a member state in a given year, i.e., its emissions had it not been a member, can be roughly estimated as its actual emissions for that year minus β_1.[17] Using the model in this way, or others, to estimate counterfactuals for specific countries could supplement qualitative efforts to generate counterfactuals in indices of regime influence.[18] The R^2 represents the fraction of all the variation in the DV, in this case EMISS, explained by the variation in all the IVs taken together. Thus, the R^2 provides an estimate of how well the analyst has captured the factors that influence the DV, or how complete the analyst's model of the DV is.[19]

Quantitative Modeling to Compare Regime's Effects

With this single regime model as background, how do we devise a more generalizable model that can use data from several regimes and subregimes to address comparative questions, such as what features make a regime effective? As an example, consider the debate over whether treaty "enforcement" involving sanctioning violation induces behavior change more effectively than "management" involving facilitating compliance.[20] Precisely because each of these approaches will be ineffective sometimes, this question requires identifying either the "average" effect or context-contingent effects of these approaches across a range of regimes. Indeed, case studies documenting compliance rates with ei-

dence" (changes in A precede changes in B), and "nonspuriousness" (the ability to rule out other possible causal variables) (Asher 1976, 11; and Kenny 1979, 3–5).

16. Of course, a fully accurate interpretation of the β in this way requires that the analyst has paid careful attention to multi-collinearity, heteroskedasticity, omitted variable bias, and a variety of additional statistical concerns.
17. A more refined counterfactual might subtract β1 from emissions forecast by the model using each states' actual values for all the IVs. The impact of regime membership for that state would then consist of the difference between its actual emissions and the emissions forecast by this method.
18. Helm and Sprinz 1999; and Sprinz and Helm 1999.
19. The adjusted R2 is conceptually identical but corrects this estimate to reflect the fact that adding more IVs to a regression equation can increase the R2 even if the additional IVs do not have any significant correlation with the DV.
20. Chayes and Chayes 1995; and Downs et al. 1996.

ther type of approach can contribute to, but not resolve, the debate since they cannot assure us whether high (or low) rates in one regime constitute a systematic tendency or a mere anomaly. Quantitative analysis is crucial to determine whether sanctions work better than rewards across a range of regimes and circumstances, after controlling for the degree or "depth" of cooperation.[21] This question also highlights the value of a subregime-based analysis, since many regimes use enforcement for some rules and management for others, as evident in the Montreal Protocol's sanctions for developed country noncompliance and assistance for developing country noncompliance.

This debate surrounds the hypothesis that member states make major or "deep" changes in behavior in response to regimes and subregimes backed by sanctions but not in response to those supported by rewards. How might we construct a regression model of such a hypothesis? Since we want to include data from different regimes, we must have a DV that is comparable across regimes. Consider the following model:

$$\text{CRB} = \alpha + \beta_1\text{*MEMBER} + \beta_2\text{*SANCTION} + \beta_3\text{*MEM-SANCT} + \beta_4\text{*DEPTH} + \beta_5\text{*CGNP} + \beta_6\text{*CPOP} + \ldots + \beta_N\text{*OTHER} + \epsilon$$

where CRB is some annual measure of Change in Regulated Behavior under various treaties, MEMBER is again coded as 1 in years during which a state is a member and 0 otherwise, SANCTION is coded as 1 for years in which rules supported by sanctions are in force and 0 otherwise (although any other regime feature could be similarly included), and MEM-SANCT is coded as 1 in years for which a sanction-based rule is in effect for a state and 0 otherwise. DEPTH is some indicator of the depth of cooperation, CGNP is the annual change in GNP, CPOP is the annual change in population, and OTHER represents a range of other factors believed to covary with CRB. Assuming that the operationalization of CRB under various regime rules allows convincing comparison across rules (discussed below) and, again, that omitted determinants of behaviors do not correlate with the included IVs, such a regression could provide us with considerable information on the extent and pathways by which regimes influence behavior. The value of β_1 and its t-statistic would document how much membership tends to influence behavior, holding "type" of treaty (defined as sanctions or not) constant. The coefficient of SANCTION, β_2, would appear to represent an estimate of the influence of sanctions on state behavior. And it does. But, it constitutes an estimate of the average change in regulated behaviors of both members *and nonmembers* of regimes that employ sanctions compared to those that do not, i.e., how all states in the sample (whether members or not) differ with respect to behaviors regulated by sanction-based regimes and behaviors regulated by other types of regimes.[22] Thus, β_2 tests the influence of sanc-

21. Downs et al. 1996.
22. $\beta 2$ represents the change in behavior that correlates with variation in whether a regime has sanctions or not, controlling for membership (i.e., comparing members of sanction-based regimes to members of other regimes, and nonmembers of sanction-based regimes to nonmembers of other regimes).

tions in a theoretical model in which all states (whether members or not) respond to regimes being introduced into the international system, an assumption that seems likely to underestimate the influence of sanctions by including the behavior of nonmembers in the estimate. Regimes can, of course, influence nonmember behavior, as evident in the non-proliferation regimes impact on the nuclear programs of states that are not party to it.

Yet, we have strong theoretical reasons to believe that sanctions have more, if not exclusive, influence on member states than on nonmembers, a view captured in the interaction term MEM-SANCT. The coefficient on MEM-SANCT, β_3, represents the additional change in behavior (CRB) induced among members of sanction-based regimes or subregimes. In this model, β_1 estimates the influence of becoming a member of a non-sanction regime, and $\beta_1 + \beta_3$ estimates the influence of becoming a member of a sanction regime. Although a somewhat complex model, simply constructing the model forces us to clarify underlying notions of how regimes may influence state behavior. Indeed, the model allows us to assess whether regimes wield influence over all states in the system (whether members or not), only over states that are members, or only over members if they employ sanctions.[23]

Our faith in these estimates, especially in whether to interpret them as the "influence" of a regime rather than mere correlation, depends on excluding other possible explanations of behavior change. Most importantly, the model must include (and thereby remove the variation in behavior that correlates with) "depth," i.e., how much was required of members, since it is central to the claims made in the enforcement-management debate.[24] We also want to include other influences on environmentally-harmful behaviors, such as the level of economic activity, population, and other factors. The most important benefit of including such indicators lies in increasing our confidence that our estimates of the influence of membership and sanctions (i.e., β_1, β_2, and β_3) more accurately reflect their real correlation with CRB rather than a spurious correlation driven by left out or omitted variables. However, the coefficients on these variables may prove of interest in their own right. Thus, β_4 represents the change in regulated behaviors induced by a 1% change in depth of cooperation (although we might also want to include a membership-depth interaction term), β_5 that induced by a 1% change in economic growth, and β_6 that induced by a 1% change in population.

Although illustrated in terms of evaluating the influence of sanctions while controlling for depth of cooperation, the foregoing discussion demonstrates a generic model useful for evaluating the influence of any regime feature while controlling for other features or to evaluate the influence of contextual variables on regime effectiveness. Answering any specific question requires specific, theoretically-informed, modeling that captures relevant variables of in-

23. $\beta 1$ represents the additional change in behavior induced in members of a regime, controlling for type of regime (i.e., comparing members of sanction-based regimes to nonmembers of those regimes, and members of non-sanction-based regimes to nonmembers of those regimes).
24. Downs et al. 1996.

terest, interactions among variables, and appropriate control variables. But the model delineated shows how such influences can be modeled.

Before proceeding, some caveats and limitations of quantitative analysis deserve mention. First, as already noted, quantitative analysis trades off accuracy for generalizability. Including more units of analysis and more observations means doing so with less knowledge and detail. We rightly place more confidence in a researcher's assessment of the Atlantic tuna regime's effects if she studied only that regime than if she studied it as one of ten fisheries regimes. However, we also rightly are more cautious in generalizing from an explanation of regime success derived solely from the Atlantic tuna regime than from one derived from a large set of regimes. Second, because quantitative analysis requires simplifying each observation to collect data on many observations, it depicts trends in influence across regimes better than stories of a particular regime's influence on a particular state. Thus, the single-regime LRTAP model above would be more useful at identifying the average emission reductions induced by LRTAP membership than which countries' behaviors were most influenced by LRTAP.[25] Claims from quantitative analyses, even if convincing, may be too probabilistic or vague for the desired purposes. Third, systematic and careful specification of variables and models can capture the presence or absence, strength, or quality of even quite subjective assessments of institutional and contextual features. But, the quantitative analyst must choose between capturing empirical richness by including and coding variables for a myriad of distinguishing features or using coarser coding schemes with corresponding simplification. Such simplifying of complex phenomena, by definition, ignores nuance and makes accurate mapping of findings to a given regime (internal validity) less compelling than their mapping to a large set of regimes (external validity).

Choosing Sample Size and the Unit of Analysis

Before describing how to quantitatively analyze regime effects and effectiveness, we must ask whether such an analysis is possible. Quantitative analysis requires that the analyst have multiple and comparable observations. Most statistical techniques require at least as many observations (remember the definitions above) as independent and control variables. Many more observations are needed to distinguish real effects from random covariation of the IV and DV, with at least 5 (and preferably 20) times as many observations as IVs usually recommended.[26] Even higher ratios are recommended when the IV of interest is expected to have only a small effect on the DV or if the measurement of variables is imprecise, two problems that seem particularly likely in the study of regimes.[27] If we assume that producing a reasonable regression model of any DV

25. Levy 1993.
26. Tabachnick and Fidell 1989, 129.
27. Tabachnick and Fidell 1989, 129.

of interest involves 5 to 10 IVs, this suggests that we need data sets of at least 50 and preferably a few hundred observations including observations from a range of different units of analysis.[28]

This simple calculation seems to confirm the common assumption that quantitative analysis is not possible because there are too few environmental regimes to compare. Recalling the definition above, if each unit of analysis corresponds to a regime or its absence, than we certainly have too few to run a regression. Of the several hundred extant multilateral environmental treaties, most have little reliable data on any conceivable dependent variable and fewer still have the needed comparative data for the period prior to regime formation. Indeed, reliable data collection often only starts upon regime formation! These problems preclude using quantitative analysis to assess regime consequences if we consider regimes as our unit of analysis. However, quantitative analysis can become possible and appropriate if we increase the number of observations by one of three methods, each already alluded to: examining "subregimes" rather than regimes, observing multiple years rather than one year before and one year after regime formation, and observing individual countries rather than all states as a group.

First, as noted, theoretical considerations recommend viewing regimes as composed of distinct sub-units. Evaluating the "regulatory effectiveness of regimes" seems as valuable as evaluating "the regulatory effectiveness of governments." Our understanding of domestic governance derives from examining variation in regulatory effectiveness within a government as well as between governments. Questions like "do governments induce compliance with their regulations" or "do citizens comply with laws" entail such high levels of aggregation that they are unlikely to identify particularly compelling relationships. Assessing whether hiring more police officers "reduces traffic violations," for example, requires either mindlessly aggregating running red lights with exceeding the speed limit or using one of these metrics in lieu of both. Yet, it is precisely the variance in traffic light vs. speeding violations that shed light on the conditions under which people comply with traffic laws. Likewise with regimes. Most regimes contain many proscriptions and prescriptions, each of which can be viewed as a distinct subregime. It is likely that a regime's effectiveness varies across these rules in ways that reflect variations in the rules themselves and in the strategies used to induce compliance with them. So long as each rule has a separate indicator of its effect, there is good reason to treat these as separate units of analysis. Comparing subregimes has the additional virtue of holding many variables constant across that subset of observations derived from the same regime.

28. Statistical power analysis confirms these general rules of thumb, suggesting that a regression model using 8 independent variables, a statistical significance test (i.e., a) of .05, and a power criterion of .80 would need a sample of 107 to detect a "medium" effect size and a sample of over 700 to detect a "small" effect size (Cohen 1992, 155–159).

Second, even most case studies split regimes into at least two observations. Whatever the dependent variable, making a convincing argument requires comparing observations of member state behavior under the regime to some pre-regime period, to their behavior in similar contexts without a regime, to corresponding counterfactual thought experiments, or to the behavior of comparable nonmembers. In all of these instances, however, each regime-year can be considered to be a separate observation. Data on many air, land, and water pollutants as well as catch and trade statistics for various species are often available for a range of years that span entry into force of corresponding regimes. There is no reason to simply average data for the five years before a regime enters into force and compare it to the average for five years thereafter, when non-aggregated annual data provides greater analytic leverage.[29]

Third, some states never join certain regimes and those that do, do so at different times. Such variation in membership in regimes, and in opting out of certain provisions, permits comparing the behavior of states for whom an international rule is binding to their own behavior before it became binding as well as to that of states who are not legally bound. Even allowing that regimes may have some influence on states that are not members, it seems reasonable to assume that effective regimes have different, if not greater, influences on members than nonmembers. When combined with the previous points, this suggests that the best approach involves defining units of analysis at the subregime level and recording data on behavior, membership, GNP, and other appropriate variables for all relevant countries and years. That is, each observation would be identified in terms of subregime, country, and year.

Conducting the Empirical Analysis

How, then, might we actually run an econometric model to assess the influence of different regime features? A model's appropriateness depends, of course, on the purpose of inquiry. But, all models require careful attention to defining the dependent variable for study, selecting independent variables to capture potential sources and pathways of regime influence, and identifying additional independent variables that explain variation in the dependent variable and for which we want to control. The data must be collected so it allows sensible comparison across regimes and subregimes, a particularly difficult problem.

A word of note is also in order. Underdal and Young have promoted the value of distinguishing regime effectiveness from regime effects, i.e., a regime's intended and direct effects from other unintended or indirect effects.[30] While recognizing the value of research on both of these phenomena, the following discussion uses the mainstream debate on simple effectiveness, defined as a regime's or subregime's success at achieving the goals that led to its creation, for

29. Murdoch et al. 1997.
30. Underdal and Young forthcoming.

expository purposes. There is no reason, however, why any potential intended or unintended effect of an environmental regime, such as influences on equity, economic growth, or the growth of other institutions, cannot be modeled in parallel ways.

Defining the Dependent Variable

Considerable scholarship has sought to define regime effectiveness. Much early literature used compliance as a metric of effectiveness.[31] Most recent work has argued for behavior change and environmental progress as more appropriate metrics.[32] A new phase of this debate has been opened recently by efforts to identify a common metric or index for cross-regime comparisons of effects and effectiveness. Helm and Sprinz have proposed defining effectiveness as the amount of progress induced by the regime toward a regime's collective optimum from the no-regime outcome.[33] Their strategy requires estimating both the no-regime counterfactual and the collective optimum using game-theory, optimization, or interviews of experts.[34] Miles and Underdal attack the same problem by using qualitative case studies to assess effectiveness on different scales (ranging from 0 to 4 for behavioral change and 1 to 3 for environmental improvement) and then normalizing them to a range from 0 to 1.[35]

Both approaches produce a common metric of effectiveness ranging from no improvement relative to the no-regime outcome to full achievement of the collective optimum. Despite the value of these efforts to allow comparison of effectiveness across regimes, they cannot serve as dependent variables in a regression model. Both scores are essentially qualitative assessments of regime effectiveness, but effectiveness is precisely what we seek to derive from the regression equation. It might seem tempting to use their effectiveness metrics as the DV in a regression equation. But, as already noted, a regression identifies both the magnitude (β) and likelihood (t-statistic) that the DV covaries systematically with some IV representing the presence or absence of some regime feature and estimates the counterfactual as actual performance minus β. Thus, using metrics similar to those proposed by Helm and Sprinz or Miles and Underdal as a DV would entail regressing a qualitative assessment of regime effect on some regime characteristic to see if the regime had an effect. Although this obviously makes little sense, other research programs have made such errors.[36] Thus, al-

31. Mitchell 1994; Mitchell 1996; Chayes and Chayes 1993; and Brown Weiss and Jacobson 1998.
32. Young 1999a; Young 1999b; Victor et al. 1998; Miles et al. 2001; Stokke 1997; and Wettestad 1999.
33. Sprinz and Helm 1999; and Helm and Sprinz 1999.
34. Sprinz and Helm 1999, 365.
35. Underdal 2001, 4.
36. Indeed, the seminal quantitative work on economic sanctions made precisely this mistake, regressing a DV that included "the contribution made by sanctions to a positive outcome" on whether sanctions were imposed or not to determine whether sanctions influence state behavior (Hufbauer, Schott, and Elliott 1990).

though neither pair of authors has suggested using their metric to quantitatively analyze regime effectiveness, the temptation for others to do so should be avoided.

Although not useful as DVs, these metrics highlight the need for some comparable measure of change in actual performance (whether involving behavior or environmental quality) as our DV. Yet, we need to avoid confusing creation of a *common* metric of effectiveness with creation of a *comparable* one. Denominating each regime's progress toward its collective optimum in similar units does not allow interpreting those units as reflecting meaningful differences across regimes. As many authors have noted, regimes address problems that vary significantly in their resistance to remedy.[37] We want to capture both components of success, i.e., how much change the regime induced *and* how hard that amount of change was to induce. Consider trying to evaluate whether the whaling or ozone protection regime was more effective. Assume, heuristically, their goals were recovery of whale stocks to historical levels and complete elimination of ozone depleting substances (ODSs). If we assume that both ODS emissions and whales killed are at least somewhat lower than they would be without the regimes, then both regimes should be assessed as "somewhat" effective. But, which was moreso? Based on the magnitude of behavior change alone, we might assess the ozone regime as quite effective for inducing additional progress toward complete ODS phaseout even after eliminating the influence of non-regime reasons for phaseout. We might view the whaling regime as completely unsuccessful since actual performance in terms of whale stocks fell so far short of complete recovery. And, indeed, it may be appropriate to view the whaling regime as far less effective than the ozone regime. However, the degree to which the ozone regime "bests" the whaling regime when measured against the collective optimum owes as much to differences in the difficulty of achieving the collective optimum as to differences in the regime's effectiveness at doing so.

How, then, do we devise a DV that can be used to compare convincingly across regimes? I believe a satisfactory DV requires capturing environmental effort as well as behavior change. We need to capture the difficulty of making progress toward the collective optimum as well as how much progress was made. Assessing relative effectiveness across regimes, as opposed to absolute effectiveness of a regime relative to the counterfactual, requires assessing both the *amount of change* and the *per unit effort* needed to make such change. Let us consider these components in turn. First, we want our measure of behavioral or environmental change to be comparable across analysis units. Although all can be expressed numerically, we clearly cannot include numbers of whales killed, acres of deforestation, and tons of pollutants emitted in a single regression. How should we address this problem? It seems unlikely that we can identify a single metric that allows convincing comparisons across *all* regime types. Re-

37. Miles et al. 2001; Young 1999b; and Wettestad 1999.

gime goals simply differ too much. However, it may be possible to identify a relatively few categories of regimes that have sufficiently similar goals to allow valid comparison among regimes within a category. Thus, one category might include regimes that target pollutant emissions, including the European regime addressing acid precipitation, the Montreal Protocol, the Climate Change Convention, and a variety of river pollution treaties. A second category might include wildlife regimes, such as those addressing trade in endangered species, whaling, polar bears, fur seals, and various fisheries. A third category might include habitat preservation regimes, such as the conventions on wetlands, world heritage sites, and desertification. One can imagine devising a single, comparable indicator of effectiveness for each category: for pollutant regimes, levels of emissions; for wildlife regimes, numbers of animals killed or changes in species population; and for habitat regimes, relevant acreage.

Even among regimes whose indicators can be expressed in similar units (for example, sulfur dioxide, nitrogen oxide, volatile organic compounds, ODSs, and CO2 emissions can all be expressed in tons), differences in the levels of emissions make a regression using absolute levels (raw data) inappropriate. To compare across regimes, or even across countries within a regime, requires normalizing data. One might consider using annual changes in those absolute levels (first differences), or an index based on setting a given year's level as 100. Calibrating across regimes, across countries, and across time, however, seems to recommend normalizing absolute levels into annual percentage change scores (APCs). Using percentage change makes otherwise disparate data relatively comparable by adjusting for the initial level of the underlying activities both at the regime and country levels. Calculating those percentage changes on an annual basis provides the additional benefit of re-calibrating (and thus allowing comparison across) every year, rather than the single normalization of indexing. The differences are shown in Table 1 and 2.

The second, usually neglected, component necessary to a notion of relative effectiveness is per unit effort (PUE). The challenging goal here is to capture the difficulty of inducing a given change in behavior in a way that allows comparison across regimes, countries, and time. If we accept annual percentage change (APC) as one part of our DV, then it makes sense to define PUE as the difficulty of achieving a 1% change in the relevant behavior, be it emission reduction, animals not killed, or acres protected. In pollution regimes, this corresponds to the abatement costs of a 1% emission change; in wildlife regimes, perhaps to the benefits foregone by not killing 1% of a given species or the costs of protecting an additional 1% of the population; and in habitat regimes, perhaps to the cost of protecting an additional 1% of the existing acreage. Such an approach has the virtue of not producing astronomically high scores at low absolute levels of an activity, since a 1% change from already low levels of an activity involves a small absolute reduction and therefore will counter the influence of the increasing marginal cost of environmental protection. We can imagine three levels of resolution in such figures. At the grossest level, one can imagine

Tables 1 and 2
Example of a Dependent Variable Measured in Absolute and Relative Terms

Dependent Variable as Absolute Metric
Example: SOx and NOx Emissions (000s of tonnes)

Subregime	*Country*	*1985*	*1986*	*1987*	*1988*	*1989*	*1990*	*1991*	*1992*
SOx	Austria	195	176	160	115	102	91	83	63
SOx	Belgium	400	377	367	354	325	372	334	318
SOx	Canada	3692	3627	3762	3838	3695	3236	3245	3117
SOx	Iceland	18	18	16	18	17	24	23	24
NOx	Austria	220	217	213	203	196	194	198	188
NOx	Belgium	325	317	338	345	357	339	335	343
NOx	Canada	2038	2043	2131	2204	2188	2104	2003	1997
NOx	Iceland	21	22	24	25	25	26	27	28

Dependent Variable as Annual Percentage Change (APC)
Example: SOx and NOx Emissions (% change from prior year)

Subregime	*Country*	*1985*	*1986*	*1987*	*1988*	*1989*	*1990*	*1991*	*1992*
SOx	Austria	−10.6%	−9.7%	−9.1%	−28.1%	−11.3%	−10.8%	−8.8%	−24.2%
SOx	Belgium	−20.0%	−5.8%	−2.7%	−3.5%	−8.2%	−14.5%	−10.2%	−4.8%
SOx	Canada	−6.6%	−1.8%	3.7%	2.0%	−3.7%	−12.4%	−0.3%	−3.9%
SOx	Iceland	−5.3%	0.0%	−11.1%	12.5%	−5.6%	−41.2%	−4.2%	4.3%
NOx	Austria	0.9%	−1.4%	−1.8%	−4.7%	−3.4%	−1.0%	−2.1%	−5.1%
NOx	Belgium	n/a	−2.5%	6.6%	2.1%	3.5%	−5.0%	−1.2%	2.4%
NOx	Canada	8.9%	0.2%	4.3%	3.4%	−0.7%	−3.8%	−4.8%	−0.3%
NOx	Iceland	−4.5%	4.8%	9.1%	4.2%	0.0%	−4.0%	3.8%	3.7%

each regime having a single PUE score designed simply to capture variation in costs across regimes. At the next level, one can imagine PUE scores varying both by regime and by country.[38] Finally, one can imagine PUE scores varying by regime and by country over time.

The product of these PUE and APC constructs creates a total effort score that has several virtues as a DV. Essentially, it represents the effort made at environmental protection in "regime effort units" or REUs. Regressing REUs on a set of IVs including at least one regime-related variable, would allow us to use the β on regime-related variables as a metric of regime effectiveness that would be comparable across analytic units. Thus, a well-specified regression model of environmental effort (in REUs) that produced a significant t-statistic on a mem-

38. Indeed, the assumption that abatement costs, and by implication PUE scores, vary by country underlies the flexibility mechanisms designed into the Climate Change Convention.

bership variable would allow interpretation of β as the change in environmental effort induced by membership.

The plausibility of using REUs for comparison across regimes depends, of course, on how convincingly we assign PUE scores across regimes. Using REUs to compare countries within a regime requires only estimating how the costs of making a 1% change *on a given environmental problem* vary by country. Comparing the effectiveness of two different regimes or the responses of individual countries across regimes requires determining how the costs of making a 1% change *on two different environmental problems* compare. How do we convincingly assess whether the costs of a 1% change in sulfur emissions are greater or less (both on average and for specific countries) than those of increasing elephant or whale populations by 1%? Though difficult, the problem may not be unresolvable. One approach involves limiting comparisons to regimes with relatively similar types of costs. Thus, we may feel confident comparing abatement costs for sulfur emissions and volatile organic compounds but far less confident comparing them for sulfur emissions and wetlands protection. Initial efforts may need to compare similar regimes and address more challenging comparisons after developing experience and methodologies for identifying PUE in different contexts. Despite these practical difficulties, in instances where PUEs are available, REUs based on them may offer opportunities to make the cross-regime comparisons we seek.

They also help us keep efficiency and effectiveness separate. Consider a regime that induced one state in a given year to reduce its sulfur emissions by 2% at a cost of $20 million per 1% reduction (or $40 million total), while another state in that same year reduced its sulfur emissions by 2% at a cost of $5 million per 1% reduction (or $10 million total). The REU appropriately reflects the common-sense view that the regime was more effective with the former state (it induced a more costly change in behavior) but that the latter state was more efficient in undertaking its commitments.

Identifying Independent Variables

Having chosen a dependent variable (whether defined in terms of environmental effort units or some other metric), we need a model of both regime and other potential determinants of variation in that DV. The earlier discussion sheds light on three different types of regime influence that can be modeled: membership, features, and membership-feature interactions. The most obvious element involves using membership (coded as above) as the primary independent variable of regime influence, with membership varying by country, year, and subregime. Intuitively, this corresponds to (and allows us to evaluate) a theory that holds that regimes only influence the behavior of those states legally bound by a given rule. Employing a membership variable allows estimating regime influence by comparing a country's behavior while a member to its behavior while a nonmember (eliminating cross-country effects) as well as by comparing

member behavior to nonmember behavior during the same time period (eliminating cross-time effects). Regimes may also influence behavior by establishing norms or other social pressure that influence nonmembers, albeit less so than members. This suggests including indicators for different regime features that vary by subregime and over time but whose values are the same for all countries, regardless of membership. Such features could be coded as 1 after the date a treaty was signed (or negotiations began or a provision entered into force) and 0 otherwise. Regime features also may influence members differently than nonmembers. This requires including membership-feature interaction terms if we are to assess how regime features influence members, how they influence nonmembers, and whether the influence differs across the groups.

We also need to identify other IVs as control variables to avoid introducing omitted variable bias. Just as we constructed a DV for environmental effort that would allow comparison across regimes, a model that produces interpretable explanations of changes in environmental effort must select and define independent variables in ways that allow comparison across regimes. The IVs we might employ to maximize our explanation of variance in sulfur emissions (i.e., to produce a large R^2) will tend to prove useless for evaluating volatile organic compound emissions, let alone fish catch data. We need a relatively generic and generalizable set of IVs with strong explanatory power over a wide range of regimes. In essence, we desire a model that balances explanatory power with generalizability, sacrificing model specificity up to a point at which doing so requires "too big" a loss in explanatory power.

To estimate the extent to which regimes increase the environmental protection efforts of states requires devising a set of "usual suspect" IVs, such as indicators of economic size and level of development, state of technological development, type of government, population, land area, and level of environmental concern. Although each of these IVs could be operationalized in different ways, at least for those that vary over time using annual percentage changes would provide the most useful mechanism for modeling the DV of REU, itself an annual percentage change. Thus, we would want to use annual percentage change in GNP rather than raw GNP figures.

Developing control variables for such a model may be best thought of as a collective activity among scholars. Those investigating "environmental Kuznets curves" have developed models that predict national pollution levels based on indicators of economic growth, population, trade, inequality, technology, and other factors.[39] Given a common and growing database of evidence on different regimes, such a model could be refined by adding regime-related variables to an initial specification derived from this prior work. Existing variables in the specification could be removed as more accurate proxies were found. A collective effort to evaluate, critique, and improve such a generic model could foster a

39. For a review of this literature, see Harbaugh et al. 2001.

research program that collectively and progressively produced a more explanatory and generalizable model.

An optimal approach to model specification may involve combining a generic specification of some base set of IVs that could be used in analyzing observations from all regimes, more extensive and regime-specific specifications for use in quantitative analyses of subregimes within a single regime, and intermediate models that include IVs that apply to a range, but not all, regimes relatively well. A set of intermediate models could be developed for different types of regimes. Thus, one might imagine a specification for pollution treaties with variables for development, technology, and intensity of resource use. Wildlife protection treaties might be modeled using demand for the species as exhibited by price, stock recruitment rates, and number of countries having access to the species. Further research could identify more useful distinctions, including modeling regimes that address, say, overappropriation separately from those that address underprovision problems, with indicators of administrative capacity playing a central role in the first and indicators of financial capacity playing a central role in the second.[40]

Using Panel Data

Armed with a model of regime influence and a level of analysis that produces enough data to distinguish the real effects of regimes from random covariation, we must consider the appropriate analytic techniques. Defining observations in terms of subregime, country and year allows use of panel or pooled time-series data. Although mathematically complex, the analytic techniques used to analyze panel data are conceptually easy to follow. Their major advantage lies in their ability to "take into account unobserved heterogeneity across individuals and/or through time."[41] Thus, panel data can identify the extent to which the dependent variable covaries with the regime-related independent variables after controlling both for differences across countries and for variation over time.

Conceptualized visually, the values of the DV fit in a matrix of rows of country-subregimes, columns of years, and cells of data, as shown in Tables 1 and 2 above. The values of IVs can be fit in corresponding matrices. An observation would consist of a single "slice" through these matrices, picking up the value of the DV and corresponding IVs and CVs for a subregime-country-year. Many IVs, for example, membership or annual percentage change in GNP, are what are called "individual time-varying variables"[42] that vary by both country and year (both columns and rows differ). Other "individual time-invariant variables" vary by country but only slowly by year, such as administrative capacity

40. Ostrom 1990; and Mitchell 1999.
41. Hamerle and Ronning 1995.
42. Hamerle and Ronning 1995; and Finkel 1995.

or level of development, and are captured in matrices in which the value for a given country is the same for all years but those values vary across countries (rows differ but columns do not). "Period individual-invariant variables" involve time-specific differences that affect all countries equally, such as changes in regime features and changes in world oil and coal prices, and can be captured in matrices in which all countries have the same values for a given year but values vary across years (columns differ but rows do not). Tables 3 through 6 provide examples of these variables.

What are the advantages of panel data for evaluating the effects and effectiveness of regimes? Consider efforts to estimate how membership influences state behavior. Cross-section data would estimate this effect by comparing member behavior to nonmember behavior, failing to address the likelihood that member countries differ in systematic ways from nonmembers. With cross-section data, it proves difficult to decipher whether "better" behavior by members reflects the influence of membership or the fact that those most willing and able to alter their behavior become members. Even with proxies for such willingness or ability included in the model, the possibility remains that member and nonmembers differ in some systematic but unobserved way. In contrast, time-series data estimates the membership effect by comparing the behavior of states as members to their behavior as nonmembers, controlling for other factors. This approach ignores the possibility that other influences that occur contemporaneously with becoming a member (for example, the end of the Cold War in the LRTAP sulfur case) explain the change in behavior. Regression using time-series data cannot distinguish whether membership or the other factor is responsible for any behavioral differences.

Panel data mitigates these problems by taking advantage of both types of variation simultaneously. Panel data uses changes in nonmember behavior over time to estimate how time-varying factors would have effected member behavior, thereby avoiding erroneously attributing those effects to membership. Panel data controls for country-specific factors by using changes in behavior during the period in which a country was not a regime member to estimate how its behavior would have been driven by non-regime factors when it was a member, thereby avoiding erroneously attributing those effects to membership. Thus, panel data improves our estimate of regime effects by more effectively separating regime effects from those due to time or country variables. Fortunately, panel data analysis can be undertaken with observations over only two or three time periods, although multi-year panels are certainly desirable.[43]

Finally, panel data analysis improves our ability to make causal inferences by permitting explicit evaluation of time-dependency through lagged variables.[44] Panel data can allow assessment and estimation of measurement error

43. Finkel 1995.
44. Finkel 1995.

Table 3
Example of a Independent Variable of Interest

Independent variable of interest that is individual, time-varying
Example: "Membership" based on entry into force

Subregime	*Country*	*1985*	*1986*	*1987*	*1988*	*1989*	*1990*	*1991*	*1992*
SOx	Austria	0	0	1	1	1	1	1	1
SOx	Belgium	0	0	0	0	1	1	1	1
SOx	Canada	0	0	1	1	1	1	1	1
SOx	Iceland	0	0	0	0	0	0	0	0
NOx	Austria	0	0	0	0	0	0	1	1
NOx	Belgium	0	0	0	0	0	0	1	1
NOx	Canada	0	0	0	0	0	0	1	1
NOx	Iceland	0	0	0	0	0	0	0	0

Tables 4, 5, and 6
Examples of Three Types of Control Variables

Control variables that are individual time-invariant
Example: Land Area (000s of sq. kilometers)

Subregime	*Country*	*1985*	*1986*	*1987*	*1988*	*1989*	*1990*	*1991*	*1992*
All	Austria	83.9	83.9	83.9	83.9	83.9	83.9	83.9	83.9
All	Belgium	30.5	30.5	30.5	30.5	30.5	30.5	30.5	30.5
All	Canada	9976.1	9976.1	9976.1	9976.1	9976.1	9976.1	9976.1	9976.1
All	Iceland	103	103	103	103	103	103	103	103

Control variables that are period, individual-invariant
Example: World Oil Price Index ($/bbl)

Subregime	*Country*	*1985*	*1986*	*1987*	*1988*	*1989*	*1990*	*1991*	*1992*
All	Austria	143.5	79	97.6	81.2	107.7	123.5	120.7	131.3
All	Belgium	143.5	79	97.6	81.2	107.7	123.5	120.7	131.3
All	Canada	143.5	79	97.6	81.2	107.7	123.5	120.7	131.3
All	Iceland	143.5	79	97.6	81.2	107.7	123.5	120.7	131.3

Control variables that are individual, time-varying
Example: GNP per capita (000s of constant 1995$)

Subregime	*Country*	*1985*	*1986*	*1987*	*1988*	*1989*	*1990*	*1991*	*1992*
All	Austria	23.6	24.1	24.5	25.2	26.2	27.1	27.6	27.7
All	Belgium	22.3	22.7	23.2	24.3	25.1	25.7	26.1	26.4
All	Canada	17.3	17.5	18.1	18.7	18.7	18.5	18.0	17.8
All	Iceland	23.0	24.4	26.4	25.5	25.5	25.6	25.7	24.7

in the variables in the model, a problem particularly likely with the social science data likely to be used in studies of regime effectiveness. It also allows evaluation and correction for auto-correlation induced if both the dependent variable and included independent variables are functions of an omitted variable, thereby permitting assessment of whether the model is properly specified.[45] Finally, panel data allows evaluation of heteroskedasticity across the observations in the data set.

Availability of Data

Given what I hope are theoretically-compelling strategies for selecting the units of analysis, identifying DVs and IVs, and conducting the analysis, the question remains whether enough regimes with enough data exist to make quantitative analysis possible? The answer is a resounding yes. First, several behavioral or environmental quality data sets exist that can provide the foundation for calculating APC figures. In the LRTAP case, data on sulfur dioxide, nitrogen oxides, and volatile organic compounds are available for an average of 30 countries per year for the period 1980–1997, representing almost 1600 analysis units (30 countries for 18 years for 3 protocols). Similar levels of detail are available for the Montreal Protocol and CFC production. Fisheries, whaling, and other marine mammal data sets include most countries of the world for periods spanning up to 50 years, allowing evaluation and comparison of the many corresponding agreements. Some, less detailed data is available on catch, and in some instances populations (such as annual bird counts), relevant to many agreements on bird and land animal preservation.

Grounds for guarded optimism also exist regarding PUE figures. Most fisheries regimes have historical catch per unit effort information.[46] Researchers at IIASA have estimated abatement costs based on different scenarios for a range

45. Finkel 1995, 81.
46. Peterson 1993.

of countries and years that could serve as PUE estimates for European and North American acid precipitant regimes.[47] Sprinz and Vaahtoranta generated country-specific abatement costs for the ozone and acid rain regimes.[48] Data sets that could serve as at least rudimentary PUE estimates may well exist for other regimes, since they correspond so closely to the costs of environmental control. The success of IIASA analysts and Sprinz at estimating abatement costs using both sophisticated modeling and more rudimentary proxy variables suggests that efforts in this direction are likely to bear at least some fruit.

On the IV side, data on treaty entry into force and country membership are readily available for all treaties. Several analysts have already coded particular regime features for a range of environmental regimes, including data on institutional structure, monitoring, and enforcement, data that could easily be further enhanced through more systematic coding of environmental treaties.[49] Country-year data are also available on a wide variety of political and economic variables that are central to any model of environmental behavior, including various permutations of GNP, population, energy use, type of government, and level of development, with most available in electronic format. Even acknowledging that the mismatch of data on DVs and IVs would further reduce the set of usable observations due to missing values for various observations, a carefully cleansed data set seems likely to yield enough country-year observations to produce a collective data set with several thousand observations.

Other regimes we want to evaluate, however, will have no data, data for only a few years or a few countries, or data of such poor quality that it would make little sense to use it. What can we do in such situations? The obvious answer is to recognize the inability to analyze such regimes in the short term and attempt to establish data collection systems that will allow such analysis in the future. An alternative possibility, however, involves a more careful and iterative search for data by identifying indicators relevant to the effectiveness of a given subregime and determining whether they are available and, reciprocally, identifying available data sets and determining to which subregimes they might be relevant. Such a process may uncover nonobvious variables that are both relevant and available to support the use of quantitative analysis to evaluate regimes. As most case study scholars know, extensive relevant data turns up for many regimes if sufficient research time is invested. A systematic attempt to work with such scholars could take advantage of their knowledge of individual data sets to create a meta-database of environmental indicators for analysis.[50] Indeed, the belief that relevant data sets do not exist may owe more to the as-

47. Alcamo et al. 1990.
48. Sprinz and Vaahtoranta 1994.
49. For example, databases created by Peter Haas, Dimitris Stevis, Edith Brown Weiss and Harold Jacobson, and the International Regimes Database all have systematic codings of several variables for various environmental treaties. See Haas and Sundgren 1993, 401–429; Brown Weiss and Jacobson 1998; Victor, Raustiala, and Skolnikoff 1998; and Stevis 1999.
50. The author is currently in the process of developing such a project.

sumption that quantitative analysis is not possible than to the real unavailability of such data.

Conclusion

Quantitative analysis offers the opportunity to investigate certain aspects of regime consequences in ways that are not easily examined using qualitative techniques. Although factor analysis, contingency tables, and other techniques are certainly possible and should be explored, the present article has investigated the contribution that regression analysis using panel data could make to determining whether, and which type of, regimes are most effective. Studies that collect data on a range of regimes and subregimes provide valuable means of identifying general trends across regimes (e.g., "are regimes usually effective?"), for evaluating whether some regimes are more effective than others, and for determining how non-regime factors condition the ability of a particular type of regime to be influential. Although these questions could be answered by qualitative case studies, they are questions that are particularly suitable to large-N quantitative techniques.

Stating that quantitative techniques can complement qualitative analyses and contribute to the regime consequences research project does not mean, however, that undertaking such analyses will be easy. Indeed, the foregoing argument has sought to identify and clarify the numerous theoretical and empirical obstacles to using quantitative analysis to answer questions central to the regime effectiveness research program. Devising a dependent variable that would allow meaningful comparison across regimes requires careful attention to creating a comparable metric of change and a comparable metric of difficulty or regime effort per unit change. Likewise, representing regime influence in the model requires careful specification if we are to determine how regimes influence members, how they influence nonmembers, and how their influences differ across the two. Comparing across regimes also requires careful attention to specification of non-regime control variables. A model designed to apply to all regimes is likely to produce weak and perhaps uninterpretable estimates of regime effects; one designed to apply well to a single regime precludes comparison across regimes. Intermediate models specified to explain the variation in the dependent variable across a set of regimes that are selected for similarity in their predicted impacts may reach the right balance between these too-generic and too-specific extremes. Applying such a model to panel data using subregime-country-years as our observations allows us to control for variables in ways that more aggregated analyses cannot. Such data appears to be available, at least for enough regimes to make the enterprise worth pursuing. A well-specified model and corresponding data would allow us to evaluate whether regimes influence states, whether they do so in ways that would be unlikely to have occurred by chance, which ones do so better than others, and a variety of as yet

unidentified but important questions. The opportunities, if not endless, are out there.

References

Alcamo, J., R. Shaw, and L. Hordijk, eds. 1990. *The RAINS Model of Acidification: Science and Strategies in Europe.* Dordrecht: Kluwer Academic Publishers.

Asher, H. B. 1976. *Causal Modeling.* Beverly Hills: Sage Publications.

Biersteker, T. 1993. Constructing Historical Counterfactuals to Assess the Consequences of International Regimes: The Global Debt Regime and the Course of the Debt Crisis of the 1980s. In *Regime Theory and International Relations,* edited by V. Rittberger, 315–338. New York: Oxford University Press.

Brown Weiss, E., and H. K. Jacobson, eds. 1998. *Engaging Countries: Strengthening Compliance with International Environmental Accords.* Cambridge, MA: MIT Press.

Chayes, A., and A. H. Chayes. 1993. On Compliance. *International Organization* 47: 175–205.

_____. 1995. *The New Sovereignty: Compliance with International Regulatory Agreements.* Cambridge, MA: Harvard University Press.

Cohen J. 1992. A Power Primer. *Psychological Bulletin* 112: 155–9.

Downs, G. W., D. M. Rocke, and P. N. Barsoom. 1996. Is the Good News about Compliance Good News about Cooperation? *International Organization* 50: 379–406.

_____. 1998. Managing the Evolution of Cooperation. *International Organization* 52: 397–419.

Eckstein, H. 1975. Case Study and Theory in Political Science. In *Handbook of Political Science, Vol. 7, Strategies of Inquiry,* edited by F. Greenstein and N. Polsby, 79–137. Reading, MA: Addison-Wesley Press.

Fearon, J. D. 1991. Counterfactuals and Hypothesis Testing in Political Science. *World Politics* 43: 169–95.

Finkel, S. E. 1995. *Causal Analysis with Panel Data.* Thousand Oaks, CA: Sage.

Galtung, J. 1967. *Theory and Methods of Social Research.* New York: Columbia University Press.

Haas, P. M., and J. Sundgren. 1993. Evolving International Environmental Law: Changing Practices of National Sovereignty. In *Global Accord: Environmental Challenges and International Responses,* edited by N. Choucri, 401–429. Cambridge, MA: MIT Press.

Haas, P. M., R. O. Keohane, and M. A. Levy, eds. 1993. *Institutions for The Earth: Sources of Effective International Environmental Protection.* Cambridge, MA: MIT Press.

Hamerle, A., and G. Ronning, G. 1995. Panel Analysis for Qualitative Variables, In *Handbook of Statistical Modeling for the Social and Behavioral Sciences,* edited by G. Arminger, C. C. Clogg and M. E. Sobel, 401–451. New York: Plenum Press.

Harbaugh, W., A. Levinson, and D. Wilson. 2000. *Re-examining the Empirical Evidence for an Environmental Kuznets Curve.* Cambridge, MA: National Bureau of Economic Research.

Helm, C., and D. Sprinz. 1999. *Measuring the Effectiveness of International Environmental Regimes.* Potsdam: Potsdam Institute for Climate Impact Research, May.

Hufbauer, G. C., J. J. Schott, and K. A. Elliott. 1990. *Economic Sanctions Reconsidered: History and Current Policy.* Washington, DC: Institute for International Economics.

Kenny, D. A. 1979. *Correlation and Causality.* New York: John Wiley and Sons.

King, G., R. O. Keohane, and S. Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton, NJ: Princeton University Press.

Krasner, S. 1983. *International Regimes.* Ithaca: Cornell University Press.

Levy, M. A. 1993. European Acid Rain: The Power of Tote-Board Diplomacy. In, *Institutions for the Earth: Sources of Effective International Environmental Protection,* edited by P. Haas, R. O. Keohane and M. Levy, 75–132. Cambridge, MA: MIT Press.

_____. 1995. International Cooperation to Combat Acid Rain. In *Green Globe Yearbook: An Independent Publication on Environment and Development,* edited by F. N. Institute, 59–68.

Meyer, J. W., D. J. Frank, A. Hironaka, E. Schofer, and N. B. Tuma. 1997. The Structuring of a World Environmental Regime, 1870–1990. *International Organization* 51: 623–629.

Miles, E. L., A. Underdal, S. Andresen, J. Wettestad, J. B. Skjaerseth, and E. M. Carlin, eds. 2001. *Environmental Regime Effectiveness: Confronting Theory with Evidence.* Cambridge, MA: MIT Press.

Mitchell, R. B. 1994. *Intentional Oil Pollution at Sea: Environmental Policy and Treaty Compliance.* Cambridge, MA: MIT Press.

_____. 1996. Compliance Theory: An Overview. In *Improving Compliance with International Environmental Law,* edited by J. Cameron, J. Werksman and P. Roderick, 3–28. London: Earthscan.

_____. 1999. International Environmental Common Pool Resources: More Common than Domestic but More Difficult to Manage. In *Anarchy and the Environment: The International Relations of Common Pool Resources,* edited by J. S. Barkin and G. Shambaugh. Albany, NY: SUNY Press.

Mitchell, R. B. and T. Bernauer. 1998. Empirical Research on International Environmental Policy: Designing Qualitative Case Studies. *Journal of Environment and Development* 7: 4–31.

Murdoch, J. C., T. Sandler, and K. Sargent. 1997. A Tale of Two Collectives: Sulphur Versus Nitrogen Oxides Emission Reduction in Europe. *Economica* 64: 281–301.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action.* Cambridge, England: Cambridge University Press.

Peterson, M. J. 1993. International Fisheries Management. In *Institutions For The Earth: Sources of Effective International Environmental Protection,* edited by P. Haas, R. O. Keohane, and M. Levy, 249–308. Cambridge, MA: MIT Press.

Princen, T. 1996. The Zero Option and Ecological Rationality in International Environmental Politics. *International Environmental Affairs* 8: 147–76.

Ragin, C. C., and H. S. Becker. 1992. *What is a Case? Exploring the Foundations of Social Inquiry.* Cambridge, England: Cambridge University Press.

Sprinz, D. 1998. Domestic Politics and European Acid Rain Regulation. In *The Politics of International Environmental Management,* edited by A. Underdal, 41–66. Dordrecht: Kluwer Academic Publishers.

Sprinz, D., and C. Helm. 1999. The Effect of Global Environmental Regimes: A Measurement Concept. *International Political Science Review* 20: 359–69.

Sprinz, D., and T. Vaahtoranta. 1994. The Interest-Based Explanation of International Environmental Policy. *International Organization* 48: 77–105.

Stevis, D. 1999. Email communication.

Stokke, O. S. 1997. Regimes as Governance Systems. In *Global Governance: Drawing Insights from the Environmental Experience*, edited by O. R. Young, 27–63. Cambridge, MA: MIT Press.

Tabachnick, B. G., and L. S. Fidell. 1989. *Using Multivariate Statistics.* New York: HarperCollins Publishers.

Underdal, A. 2001. Conclusions: Patterns of Regime Effectiveness. In *Environmental Regime Effectiveness: Confronting Theory with Evidence*, edited by E. L. Miles et al. Cambridge, MA: MIT Press.

Underdal, A., and O. R. Young, eds. Forthcoming. *Regime Consequences: Methodological Challenges and Research Strategies.* Dordrecht: Kluwer Academic Publishers.

Victor, D. G., K. Raustiala, and E. B. Skolnikoff, eds. 1998. *The Implementation and Effectiveness of International Environmental Commitments.* Cambridge, MA: MIT Press.

Wettestad, J. 1999. *Designing Effective Environmental Regimes: The Key Conditions.* Cheltenham, UK: Edward Elgar Publishing Company.

Yin, R. 1994. *Case Study Research: Design and Methods.* 2nd ed. Newbury Park: Sage Publications.

Young, O. R. ed. 1997. *Global Governance: Drawing Insights from the Environmental Experience. Cambridge,* MA: MIT Press.

_____, ed. 1999a. *Effectiveness of International Environmental Regimes: Causal Connections and Behavioral Mechanisms.* Cambridge, MA: MIT Press.

_____. 1999b. *Governance in World Affairs.* Ithaca, NY: Cornell University Press.