Time and the Study of Conflict
Problems of Temporal Aggregation in Quantitative International Relations

Kevin A. Clarke [†]
H. E. Goemans [§]
Michael Peress [‡]

Abstract

International relations studies often have a temporal domain that stretches from 1816 (after the Napoleonic Wars) to the latest year for which the data are available. This breadth creates problems, which often go unrecognized, both for theory and data analysis. In order to make sense of both 1816 and 1990, theory must be so general as to lose any practical bite. Data analysis, on the other hand, must grow increasingly complex, both in terms of control variables and functional form, to account for local idiosyncrasies. We consider a log-likelihood ratio type test statistic suitable for the panel data that is ubiquitous in quantitative international relations.

April 20, 2010

[†]Corresponding author. Department of Political Science, University of Rochester, Rochester, NY 14627-0146. Email: kevin.clarke@rochester.edu.

[§]Department of Political Science, University of Rochester, Rochester, NY 14627-0146. Email: henk.goemans@rochester.edu.

[‡]Department of Political Science, University of Rochester, Rochester, NY 14627-0146. Email: michael.peress@rochester.edu.

Introduction

An almost innumerable number of quantitative international relations articles have temporal domains that start at, or soon after, the Congress of Vienna, which concluded in June of 1815. The justification for this starting point is the contention that the Congress of Vienna was the beginning of the modern international system, and that the system has maintained a "remarkable constancy" in the years to follow (Singer & Small 1968, 251). The other end of the temporal domain is often the latest year for which the data are available. Contemporary scholars, then, are working with nearly 200 years of data.

The problems created by this temporal breadth often go unrecognized. Making sense simultaneously of 1820 and 1990 requires theoretical commitments that must be so broad as to lose any practical bite. Thus, international relations scholars refer to states, democracies and non-democracies, leaders, war, conflict, and territory as if these terms have meanings immutable and absolute. Data analysis, on the other hand, must grow increasingly complex. Additional variables, measured with varied success, must be introduced to control for technological change, the emergence and disappearance of states in the international system, the emergence of non-state actors, world wars, and enduring rivalries. Typically, scholars are forced to employ complex functional forms to deal with interdependence, dynamics, and selection.

Those scholars that do recognize the problems created by extended tem-

poral domains often solve the problem by restricting the domain to one that they feel is more homogeneous. These determinations, however, are often made in an *ad hoc* fashion, without the benefit of quantitative evidence, and conditioned on the availability of data or the convenient placement of World War II. The downside of this strategy is that it leaves these structural changes unexamined and unexplained, and ignores the possibility that structural changes might exist at times other than those postulated by current conventional wisdom.

The goal of this project is to both identify these change points with data-based evidence and to explain theoretically why they fall where they do. The statistical problem we face is not a trivial one. The well-known tests for structural change, collectively known as the Chow test, are appropriate only for linear regression where pooling is acceptable. International relations data, however, are characterized both by a panel structure, often state/years or dyad/years, and by nonlinearity. The problem is further complicated by the lack of theory regarding structural change in international relations and possibility of the existence of multiple change points.

As a first cut at this problem, we consider the possible change points in a variety of dependent variables used in the literature. The approach we take is to define a likelihood ratio type test statistic that can identify a single change point in a panel data setting. As neither the finite nor asymptotic properties of this statistic are known, we use the bootstrap to generate an empirical

sampling distribution under the null hypothesis of no change. Our results demonstrate that change points do exist and that they only sometimes fall precisely where we might expect them.

The Argument for Long Time Spans

Arguments for a temporal domain starting with the Congress of Vienna go back to the inception of quantitative international relations. The persistence of that start date, however, is a function of the reliance on a single, influential data set, the rarity of international war, and a belief in the existence of social science "laws."

The Correlates of War is the single most influential data set in international relations. Early in the project, Singer & Small (1968, 251) argued for the "remarkable constancy" of the period between the Congress of Vienna and the Japanese surrender in Tokyo Bay,

> The national state was the dominant actor and the most relevant form of social organization; world politics were dominated by a handful of European powers; the Napoleonic reliance upon the citizen's army endured, with all of its implications for public involvement in diplomacy; the concept of state sovereignty remained relatively unchallenged; and while technological innovation went on apace, the period postdates the smoothbore and

> predates the nuclear-missile combination. In sum, it seems reasonable to conclude that this period provides an appropriate mixture of stability and transition from which generalization would be legitimate.

To their credit, Singer and Small left open the question whether their generalizations might hold in the period following 1945. Once the Correlates of War data set established the post-Congress of Vienna period as the start date, however, it was natural that others would seek to build on the original time period.

One reason for building on the original data set, of course, is the relative rarity of international war. Extending the temporal domain is one of way of ensuring an ever growing list of wars, which is particularly important given the panel structure of most international relations data sets. Similar reasoning has also led to the use of the Militarized Interstate Dispute (MIDs) data set as the number of MIDs is far larger than the number of international wars. The data from Russett & Oneal (2001), for example, are in year-dyad form. Covering a period from 1885-1992, the data set includes 39,996 observations of which the dependent variable, the presence of a militarized interstate dispute in that year for that dyad, takes a value of one only 1880 times, or 4.7% of the time. If the time period were further restricted, that percentage would obviously drop even lower.

Also contributing to the presumed non-problematic status of extensive

Figure 1: The Deductive-Nomological Model of Explanation

$$\left.\begin{array}{r} L_1, L_2, \ldots, L_n \\ C_1, C_2, \ldots, C_n \end{array}\right\} \quad \text{Explanans}$$
$$\overline{\qquad\qquad\qquad}$$
$$\left.\begin{array}{r} E \end{array}\right\} \quad \text{Explanandum}$$

temporal domains is the view of explanation adopted, sometimes unknowingly, by international relations scholars and more generally, political scientists. According to the tenets of this conception of explanation, to explain a phenomenon is to show that the phenomenon was to be expected (Hausman 1992). That is, an explanation consists of a set of conditions, which if true, are sufficient to produce the phenomenon. In other words, an adequate explanation consists of a logically valid argument where the event to be explained (called the *explanandum*) deductively follows from the premises of the argument (called the *explanans*). This account of explanation is known as the deductive-nomological, or D-N, or Covering Law model, and it holds that if the event to explained can be deduced from general laws and initial conditions, then the general laws and initial conditions form an explanation for the event. The model is depicted in Figure 1, where the L_i are general laws, the C_j are initial conditions, the solid line stands for a logical deduction, and E is the event to be explained.

The reason for this seeming digression is to point out that international relations scholars believe in general laws; that is, laws such as gravity and

electromagnetism that do not change across time or space. While cites to this effect can be found throughout the literature, for the moment we will focus on the Correlates of War whose data set is at the heart of the data sets considered in Section . In an overview and critique of the Correlates of War project, Dessler (1991) demonstrates that the structure of the project is an application of the deductive-nomological model of explanation. Singer, himself, would agree. (Singer 1977) approvingly cites Ernest Nagel to the effect that "reliable general laws" are the aim of science, and Singer (1979) argues that the differences between the natural sciences and the social sciences have been greatly exaggerated. If science requires general laws, then extensive temporal domains are necessary. The belief that such laws exist or that science requires them has long been a discredited philosophical notion, and political science has been behind other fields in recognizing it (Clarke & Primo 2007).

The Test Statistic

Testing for structural change in international relations data is complicated by the panel structure of the data, as well as by the nonlinear functional forms popular in the literature. Further complicating the analysis is that, in the absence of theory, we must treat the location of any potential structural change as unknown. In addition, we have no *a priori* reason to believe that a *single* change point exists, as opposed to multiple change points. Most of the

structural change tests widely available to political scientists address the case of a single change point in a linear regression. These tests are collectively known as "Chow" tests. The test we discuss remedies these problems. For the purposes of the current paper, however, we focus on the case where we suspect a single unknown change point.

Following (Andrews 1993), we are interested in null hypotheses of the form

$$H_0 : \beta_t = \beta_0 \text{ for all } t \geq 1 \text{ for some } \beta_0.$$

The alternative hypothesis that we are interested in concerns the case where the change point is unknown.

The test statistic we employ is a version of the standard likelihood ratio test, where a full sample maximum likelihood model is compared with a partial sample maximum likelihood model.[1] That is, we compare the full model against a model that is split in two at time t. As we do not know the location of the change point, we run the full model against a series of partial sample models, where t changes by year. The result is a series of likelihood ratio type test statistics of which we take the maximum as the test statistic. The statistic is then of form $supLR$.

The difficulty with this statistic is generating a distribution under the

[1]Andrews (1993) addresses full sample GMMs and partial sample GMMS, of which maximum likelihood is a special case.

null hypothesis with which to compare it. Andrews (1993) derives the large sample distribution of this test statistic under the null hypothesis when the model is based on a single time series. Since we seek to apply this test statistic to a large set of models (e.g. panel data, dyadic data), so we cannot rely on his result. Instead, our solution is to make use of the bootstrap following Hall & Yatchew (2005). Our procedure works as follows. We estimate a model without a breakpoint. We sample from this model using the parametric or nonparametric bootstrap, generating R replicated data sets. Note that by sampling from the estimated model with no breakpoints, we are sampling from a model which satisfies the null hypothesis. For each replicated data set, we compute the $supLR$ statistic. We obtain R values for this test statistic which form an approximation to the large sample distribution of the $supLR$ test statistic under the null hypothesis. We reject the null hypothesis of no break point if the observed value of the test statistic is greater than the 95% quantile of the bootstrap replications of the test statistic (if we consider the $\alpha = 5\%$ confidence level).

This principle can be extended to testing for multiple breakpoints by considering the difference between two $supLR$ statistics with a different number of breakpoints. Specifically, we define the $supLR_B$ test statistic to be the generalization of the $supLR$ to the case with B breakpoints. It is the largest value for the test statistic where the null hypothesis is the best fitting model with $B - 1$ breakpoints and the alternative hypothesis is a model with B

breakpoints. We then define the $\Delta supLR_B$ statistic by,

$$\Delta supLR_B = supLR_B - supLR_{B-1} \tag{1}$$

This generalizes the test statistic used by Bai & Perron (1998). Bai and Perron derive the large sample distribution for this test statistic for the linear model. Once again, we are interested in a more general set of models, so we employ the bootstrap to derive the large sample distribution of the test statistic under the null hypothesis.

In the case of multiple breakpoints, computational issues become more central. A naive approach would require evaluating and comparing $T * (T - 1) * ... * (T - B)$ models. We rely on techniques suggested by Bai & Perron (2003). Specifically, let $L_{r,s}$ be the value of the likelihood of the model when restricted to the sample, $r \leq t \leq s$. Let us store the value of all such likelihoods in a triangular matrix. Notice that this requires estimating $T * (T - 1)$ models rather than $T * (T - 1) * ... * (T - B)$ models. We can then for the likelihood for any model with B break points using the following procedure. If we consider the model with breakpoints at $(b_1, b_2, ..., b_B)$, we form the likelihood using,

$$L_b = T * \left(\frac{L_{1,b_1}}{b_1} + \frac{L_{b_1+1,b_2}}{b_2 - b_1 + 1} + ... + \frac{L_{b_B+1,T}}{T - b_B + 1} \right) \tag{2}$$

This construction shows how all the relevant likelihoods and likelihood

ratios can be formed from the elements of $L_{r,s}$. In addition, Bai and Perron demonstrate that computing the $\Delta supLR$ statistic does not require searching over all $T * (T-1) * ... * (T-B)$ models, but instead can be computed in $O(T^2)$ operations by employing dynamic programming.

The Data

The data we use come from three sources: Russett & Oneal (2001), Huth, Gelpi, & Bennett (1993), and Reiter & Stam (1998). At this stage of the project, we consider only the dependent variables from these data sets. Although the methods described in the last section are perfectly applicable to models with covariates, it is interesting to consider possible change points in the dependent variables before introducing covariates, such as the presence of nuclear weapons, that can proxy for time dummy variables.

The first data set we use comes from Russett & Oneal (2001), who are interested broadly in the democratic peace. As noted earlier, their dependent variable is a dichotomous variable indicating the presence of a militarized interstate dispute in a given year for a given dyad. The data set is explicitly based on the Correlates of War data base, and the authors count every conflicting pair of states that was involved in a dispute, giving "full weight to expanded, contagious, multistate disputes" [95].

The second data set we use comes from Huth, Gelpi, & Bennett (1993).

Huth and his co-authors are interested in testing structural realism against rational deterrence theory. To perform their test, they construct a data set of great-power extended and direct immediate deterrence encounters from 1816-1984. Again, the data set is based explicitly on the Correlates of War data base. The dependent variable is a dichotomous variable indicating the escalation of a dispute (the failure of the deterrent policies of the great-power defender). The data set includes 97 observations.

The third data set we use comes from Reiter & Stam (1998), whose interest is in determining why democracies win more wars than nondemocracies. The data set consists of all participants in interstate wars between 1816 and 1982. Again, the data set is based explicitly on the Correlates of War data base. The dependent variable is a dichotomous variable indicating whether a state participating in an international war experienced military victory or defeat. The data set consists of 197 observations.

Results

The results of our analyzes are in Table 1 and in Figures 2-4. Taking the Russett and Oneal data first, we found an observed value of the $supLR$ test statistic of 87.13 corresponding to the year 1936. This value corresponds to the tallest peak in Figure 2. The maximum value of the statistic from 200 bootstrapped replications of the model under the null hypothesis returned a

Table 1: Results from the Three Data Sets

	Russett and Oneal	Huth *et al.*	Reiter and Stam
Breakpoint	1936	1941	1942
Observed Value	87.130	19.170	2.980
Bootstrapped Maximum	12.930	16.100	20.640
P-Value	<0.005	<0.005	0.574
Mean before breakpoint	0.031	0.619	1.000
Mean after breakpoint	0.053	0.177	0.781

value of 12.93. Thus, we can reject the null hypothesis of no structural breaks with a p-value under .5%. The most telling evidence, however, is the difference in the mean probability of a dispute before and after the breakpoint. Before the breakpoint, the mean probability of a dispute is 31%. After the breakpoint, however, the mean probability jumps up to 53%. The finding is interesting because while we might expect a breakpoint during or after World War II, the breakpoint in these data occurs well before the start of the war.

The Huth *et al.* data show a more conventional breakpoint in 1941. The observed value of the *supLR* statistic is 19.17 corresponding to the tallest peak in Figure 3. The maximum value of the test statistic from 200 bootstrapped replications of the model under the null hypothesis is 16.1. Again, given 200 replications, we can reject the null hypothesis of no structural break with a p-value less than .5%. The breakpoint makes a large difference in this data. Before 1941, the mean probability of Great Power dispute escalation is 62%. After 1941, however, the mean probability drops to just under 18%. This change likely has something to do with the outbreak of World War II,

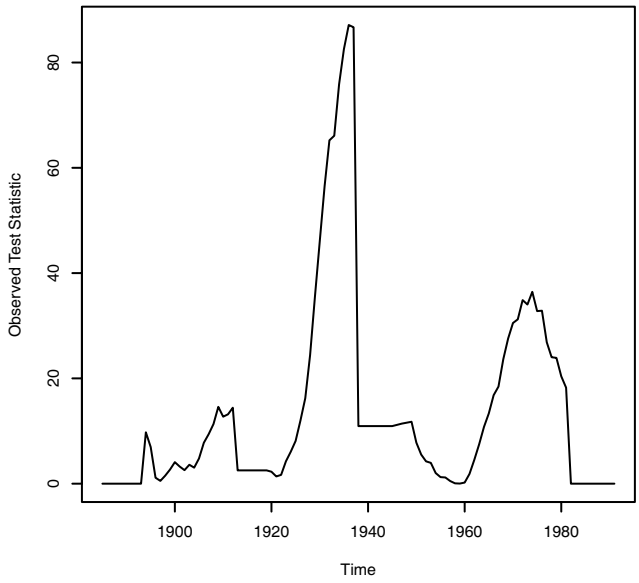

Figure 2: The observed test statistic for the Russett and Oneal data.

and the acquisition of nuclear weapons by the major powers in the late 1940s and early 1950s.

The Reiter and Stam data are of interest because although Figure 4 bears a striking concordance to Figure 3, the test statistic in this case is not statistically significant. The observed value of the test statistic is just under 3 (note that the scales for Figures 3 and 4 are quite different), which is swamped by the bootstrapped maximum value of the test statistic under the null hypothesis. Given 200 replications, the p-value in this case is 57.4%, which means that we fail to reject the null hypothesis of no structural change at conventional levels of statistical significance. We can only speculate at the reasons why the Reiter and Stam data may differ from the other data sets.

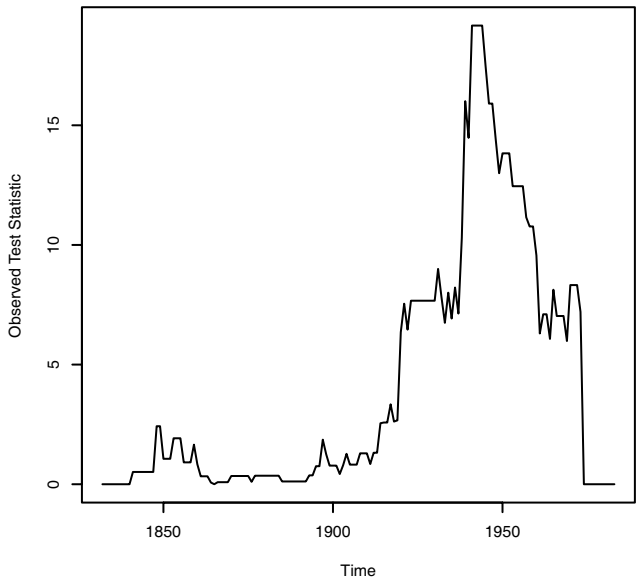

Figure 3: The observed test statistic for the Huth *et al.* data.

It seems that there exists a difference between the onset of a dispute and its escalation versus the probability of winning a dispute once it has already escalated. It is quite possible that the former are time-dependent while the latter is not.

Directions for Future Work

- add covariates to the analysis
- allow multiple changepoints
- a theory of structural change

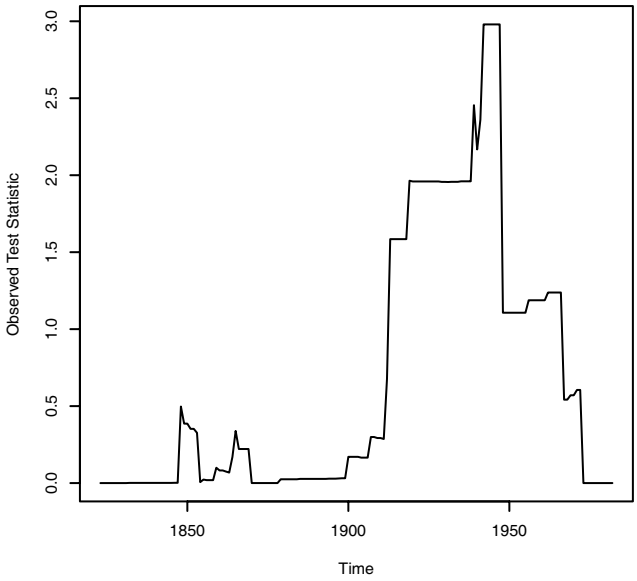

Figure 4: The observed test statistic for the Reiter and Stam data.

References

Andrews, Donald W. K. 1993. "Tests for Parameter Instability and Structural Change With Unknown Change Point." *Econometrica* 61 (July): 821-856.

Bai, Jushan, & Pierre Perron. 1998. "Estimating and Testing Linear Models with Multiple Structural Changes." *Econometrica* 66 (January): 47-78.

Bai, Jushan, & Pierre Perron. 2003. "Computation and Anlysis of Multiple Structural Change Models." *Journal of Applied Econometrics* 18 (Jan.-Feb.): 1-22.

Clarke, Kevin A., & David M. Primo. 2007. "Modernizing Political Science: A Model-Based Approach." *Perspectives on Politics* 5 (4): 741-753.

Dessler, David. 1991. "Beyond Correlations: Toward a Causal Theory of War." *International Studies Quarterly* 35 (September): 337-355.

Hall, Peter, & Adonis Yatchew. 2005. "Unified Approach to Testing Functional Hypotheses in Semiparametric Contexts." *Journal of Econometrics* 127 (August): 225-252.

Hausman, Daniel M. 1992. *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.

Huth, Paul, Christopher Gelpi, & D. Scott Bennett. 1993. "The Escalation of Great Power Militarized Disputes: Testing Rational Deterrence

Theory and Structural Realism." *American Political Science Review* 87 (September): 609-623.

Reiter, Dan, & Allan Stam. 1998. "Democracy, War Initiation, and Victory." *American Political Science Review* 92 (June): 377-389.

Russett, Bruce, & John R. Oneal. 2001. *Triangulating Peace: Democracy, Interdependence, and International Organizations*. New York: Norton.

Singer, J. David. 1977. "The Historical Experiment As a Research Strategy in the Study of World Politics." *Social Science History* 2 (Autumn): 1-22.

Singer, J. David. 1979. "Conflict Research, Political Action, and Epistemology." In *Handbook of Conflict Theory and Research*, ed. Ted R. Gurr. New York: Free Press.

Singer, J. David, & Melvin Small. 1968. "Alliance Aggregation and the Onset of War, 1815-1945." In *Quantitative International Politics: Insights and Evidence*, ed. J. David Singer. New York: Free Press.